PARENT
cross-border
PAtient REgistries iNiTiative

Methodological guidelines and recommendations for efficient and rational governance of patient registries – version 1.0

WORK PACKAGE	WP5
DOCUMENT NAME	Methodological guidelines and recommendations for efficient and rational governance of patient registries
DOCUMENT VERSION	1.0
DATE	3/2/2015

Project Title	**Cross-border Patient Registries Initiative**
Project Acronym	**PARENT**
PARENT Coordinator	Matic Meglič, NIJZ Marija Magajne, NIJZ
Work Package Leader	Metka Zaletel, NIJZ
Work Package	5
Deliverable	1
Title	Methodological guidelines and recommendations for efficient and rational governance of patient registries
Version	1.0
Dissemination level	Confidential

Authors of the Guidelines

Metka Zaletel, Matic Meglič, Vesna Lešnik Štefotič, Živa Rant, Ivan Drvarič, Marcel Kralj, Dmitri Wall, Alan Irvine, Eoin O'Brien, Marko Brkić, Borna Pleše, Vanja Pajić, Mladen Kostešić, Ranko Stevanović, Tamara Poljičanin, Ivan Pristaš, Óscar Zurriaga, Carmen López Briones, Miguel A. Martínez-Beneito, Clara Cavero-Carbonell, Rubén Amorós, Juan M. Signes, Alberto Amador, Carlos Sáez, Montserrat Robles, Juan M. García-Gómez, Carmen Navarro-Sánchez, María J., Sánchez-Pérez, Joan L. Vives-Corrons, María M. Mañú, Laura Olaya, György Surján, Károly Fogarassy, Haralampos Karanikas, Ioannis Skalkidis Persephone Doupi, Arto Vuori, Hanna Liukkunen, Antti Tuomi-Nikula

History changes			
Version	Date	Status Changes	Author
0.1	9.6.2014	First drafts of the chapters	NIJZ, HZJZ, THL, NKUA, GYEMSZI, Dmitri Wall, Vesna Lešnik Štefotič, CSISP-DGSP
0.2	11.9.2014	Some chapters/sections updated	NIJZ, HZJZ, THL, NKUA, GYEMSZI, Dmitri Wall, Vesna Lešnik Štefotič, CSISP-DGSP
0.2.2	3.10.2014	IT chapter updated	Vesna Lešnik Štefotič
0.3	15.10.2014	'Changing a registry' content added Chapter 'Quality dimensions of registry' moved	NIJZ
0.3.2	20.11.2014	IT chapter updated	Vesna Lešnik Štefotič, NIJZ
0.4	15.1.2014	IT chapter updated (EAV, temporal modelling) Chapter 'Changing and stopping registries' updated Chapter 'Registry data elements/dataset' updated	Vesna Lešnik Štefotič, NIJZ
1.0	3.2.2014	Chapter 'Registry Design' updated Chapter 'General requirements for cross-border use of patient registries' updated Chapter 'Re-use of registry data' updated Chapter 'Quality dimensions of registry' updated Chapter 'Types of patient registries' updated Introduction chapter updated Chapter 'Definition of a patient registry' updated Chapter 'Running a registry' updated in order to achieve coherence with the chapter 4 Chapter 'Planning a registry' updated in order to achieve coherence with the chapter 4 'Chapter Key issues arising within registries' updated	NIJZ, THL, GYEMSZI, HZJZ CSISP-DGSP Dmitri Wall

Table of Contents

1 INTRODUCTION

In today's world of aging populations and budgetary pressures, the healthcare systems in EU are daily forced to reinvent themselves to provide more and better care at less cost. In order to make informed decisions about introduction and funding of new and existing therapies, policy makers can only make as good decisions as the quality and availability of data they use in the process.

For patients with sharing characteristics (e.g. rare disease, implanted device or therapy, risk to develop a chronic disease), patient registries have for decades served as a key tool for assessing clinical performance as well as health technology assessment and policy implications on local, regional, national and in some cases international level. As a result, hundreds of registries have been set up, ranging from paper based spreadsheets in a physician's office to international rare disease initiatives coupling clinical and genetic data as well as biobanks. In the last 15 years information technology has given us an opportunity to greatly redesign the way we make informed decisions about individual patients as well as entire populations by – among others - enabling clinicians to collect, share, compare and analyse large amounts of patient data.

Where we still fall short is harnessing information and new knowledge from the wealth of data across registries – be it from one country to another or between/across registries with overlapping characteristics or patient pools. Researchers, HTA organisations and policy creators are wasting valuable time acquiring data from different sources and painstakingly pairing it in order to extract new knowledge. Also, setting up a new patient registry sets the clinicians on a high risk journey where a number of decisions need to be made about methodologies, processes, technologies and governance of the registry with little available guidance.

To provide guidance and tools on EU level to solve the above issues is likely the largest near-term opportunity towards data and information driven public health decision making, policy creation and research.

At PARENT we are proud to present the Guidelines – we created them to provide a practical and 'hands on' advice to set up and manage patient registries as well as to enable secondary use for public health policy and research. Accompanying it in a couple of months is also the PARENT Framework, supporting the guidelines with tools to make life easier for those setting up new registries and those exchanging data across registries; as well as pilot registries set up using the new tools to demonstrate the value of the Framework.

Getting to this point has been a challenging journey but we have made it as a result of commitment and passion of a number of PARENT experts from across EU as well as continuous support from numerous EU funded bodies and projects, and US Agency for Healthcare Research and Quality – all of whom generously contributed their knowledge and insights into the topic.

While the Guidelines are a first step towards greater interoperability of patient registries, a number of exciting and complex challenges still lie ahead, requiring continuous efforts to ensure that we utilise the full value of patient registries.

May the Guidelines serve you well.

2 PATIENT REGISTRIES

2.1 Definition of a patient registry

The terms “register” and “registry” are often used interchangeably, therefore, terminology in that field can be confusing. However, the registry is the organisation and process that supports a register and should be distinguished from the register itself. One registry may support a number of individual registers (4).

In the field of health, several definitions of the term register/registry have been provided. In 1949, Bellows (6) defined register as “system of recording frequently used in the general field of public health which serves as a device for the administration of programs concerned with the long-term care, follow-up or observation of individual cases.” In 1974, the WHO (5) defined a register as a “file of documents containing uniform information about individual persons, collected in a systematic and comprehensive way, in order to serve a predetermined purpose.” Another more broader definition was provided by Solomon et al. (8) who defined a registry as a “database of identifiable persons containing a clearly defined set of health and demographic data collected for a specific public health purpose.” Slightly different definition of a registry is proposed by ISPOR (3), which describes a registry as a “prospective observational study of subjects with certain shared characteristics, which collects ongoing and supporting data over time on well-defined outcomes of interest for analysis and reporting.” More specific definition is provided by the US National Committee on Vital and Health Statistics (1), which defines a registry as “an organized system for the collection, storage, retrieval, analysis, and dissemination of information on individual persons who have either a particular disease, a condition (e.g., a risk factor) that predisposes (them) to the occurrence of a health-related event, or prior exposure to substances (or circumstances) known or suspected to cause adverse health effects.” Despite variations in definition, it is clear that a registry involves a long-term, systematic and organized processes of collecting data, which is driven by specific, predefined aims.

Nowadays the term “patient registry” is often used in the health domain. The use of the term “patient” in combination with ‘registry’ (i.e. patient registry) is mainly used to distinguish the focus of the dataset on health information (9). The AHRQ (2) provides the definition of the patient registry, which is “an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes”.

For the purpose of the PARENT work, patient registry is defined as…

> ... an organized system that collects, analyses, and disseminates the data and information on group of people defined by a particular disease, condition, exposure, or health-related service, and that serves a predetermined scientific, clinical or/and public health (policy) purposes.

References

1. Available at: Frequently Asked Questions about Medical and Public Health Registries. The National Committee on Vital and Health Statistics http://ncvhs.hhs.gov/9701138b.htm.
2. Gliklich RE, Dreyer NA, eds. Registries for evaluating patient outcomes: A User's Guide. 3rd ed.2014.
3. Polygenis D, ed. ISPOR Taxonomy of Patient Registries: Classification, Characteristics and Terms. Lawrenceville, NJ; 2013.
4. Newton J, Garner S. Disease Registers in England. A report commissioned by the Department of Health Policy Research Programme in support of the White Paper entitled Saving Lives: Our Healthier Nation. Institute of Health Sciences. University of Oxford. 2002
5. Eileen M. Brooke, (WHO). The current and future use of registers in health information systems. Geneva, World Health Organization.1974. Available from: https://extranet.who.int/iris/restricted/handle/10665/36936
6. Bellows, Marjorie T. Public Health Reports, Vol. 64, No. 36, pp. 1148-1158. 1949. Available from:
7. http://www.jstor.org/discover/10.2307/4587080?sid=21105208469701&uid=70&uid=4&uid=3739008&uid=2&uid=2129
8. Solomon, D. J., R C Henry, J G Hogan, G H Van Amburg, and J Taylor. Evaluation and implementation of public health registries. Public Health Rep. 1991 Mar-Apr; 106(2): 142–150. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1580226/pdf/pubhealthrep00191-0040.pdf
9. Workman TA. Engaging Patients in Information Sharing and Data Collection: The Role of Patient - Powered Registries and Research Networks. AHRQ Community Forum White Paper. AHRQ Publication No. 13-EHC124-EF. Rockville, MD: Agency for Healthcare Research and Quality; September 2013. Available from: http://www.ncbi.nlm.nih.gov/books/NBK164513/pdf/TOC.pdf

2.2 Types of Patient Registries

Registries should be designed and evaluated with respect to their intended purpose(s) which can be broadly described in terms of patient outcomes. Some of the major general purposes for establishing and running a patient registry are to **describe the natural history of disease**, to **determine clinical and/or cost-effectiveness**, to **assess safety or harm**, and to **measure quality of care**, as well as to serve **public health surveillance and disease control**. In broad terms, patient registries should contribute to the improvement of patient care and healthcare planning as well as social, economical and quality of life outcomes and other health indicators (e.g. access to healthcare, health status, subjective and objective quality, health financing etc.) By following patients lengthwise and location-wise, medium- and long-term outcomes can be observed. A fine differentiation of types, sub-types, main and secondary purposes of each patient registry is essential. For example, a diabetes registry comparing to a surgical procedure registry share many common datasets that are achieving completely different purposes. Although the logic of tracking how patients do over time and factors that contribute to outcomes applies to both, there is a clear difference between the two example registries, as eligibility is pointed by a diagnosis and by an intervention, respectively.

The majority of patient registries can be divided into three general categories with multiple subcategories and combinations. These categories include observational studies in which the patient has a particular disease or condition, has had an exposure to a product or service, or various combinations of those.

The multitude of possible combinations of categories and subcategories can sometimes lead to difficulties in determining the taxonomic position of a particular registry (e.g. Is a registry of treated drug addicts primarily a disease, product or service registry, or a mixture of equally important purposes: disease surveillance, outcomes, natural history of disease, national intervention programmes evaluation). Furthermore, in some countries a very clearly defined chronic disease registry (such as a cancer registry) very often serves many secondary purposes, some of which could eventually become its primary purposes. Therefore, in order to establish an appropriate between registries and/or other sources data exchange (sharing) framework for particular secondary data use (i.e. research question), an extensive in-depth context analysis of each registry's content unit (data set, data element with properties and classes, value domains and property) should be performed. Such analysis would enable correct interpretation of the results and transparent disclosure of methodological restrictions.

One of the most important registries' quality indicators that contribute to the applied concept above is the amount and frequency of registry-related scientific publishing (meta-analysis and/or systematic review-like approach). See subchapter 4.3.

With the help of information gathered through the literature review, as well as with the insights gathered through the construction of the questionnaire and subsequent survey of registries for the RoR pilot, and with concern to the above stated complicacy of taxonomy of registries, several level classification of patient registries is offered[1] (Table 2.1).

Registries are classified according to how their populations are defined. For example, product registries include patients who have been exposed to biopharmaceutical products, medical devices

[1] This classification is by no means definite or indisputable but subject to change and modification.

or diagnostic/therapeutic equipment. Health services registries consist of patients who have had a common procedure, clinical encounter, or hospitalization. Disease or condition registries are defined by patients having the same diagnosis, such as cystic fibrosis or heart failure, or same group of conditions such as disability (1).

Table 2.1. Patient registry classification

Category	Diseases and conditions	Products	Services, events
Object type	chronic, acute communicable, rare diseases, disabilities, cause of death	medicines, devices, equipment	diagnostic, curative, preventive, discharges, births, abortions
Purposes / objectives (primary and secondary)	disease surveillance, control, natural course of disease	post-market surveillance	intervention evaluation, quality of care
	health outcomes (objective, patient reported)		
	effectiveness (clinical, comparative, financial)		
	safety and harm (HTA, vigilance)		
	intervention (planning, guidelines, reminders)		
Coverage (geographical and organizational)	health care unit (GP, hospital)		
	local (counties, districts, insurers, professional associations, NGOs)		
	national (MS, non-MS)		
	international (regional, EU, European region, global)		
Population definition	population (geographically based)[2]		
	population based (exposition dependent)[3]		
Observation unit	patient (user, client, insuree)		
	person with a characteristic of observation	person related device, equipment item	person related event (birth, death, service)

2.2.1 Disease or Condition Registries

The main inclusion criterion which disease or condition registries use is the state of a particular disease or condition. That state varies, as the patient may have a lifetime disease (e.g. rare disease such as cystic fibrosis, chronic condition such as disability) or for a more limited amount of time (e.g.

[2] A "population registry" is a registry that intends to cover all residents in a given geographic area within a given time period. The coverage of the specific registry may, however, be incomplete, but it is nevertheless a population registry if the aim is to include all the individuals in the target population. A population is defined by geographical boundaries, but usually only residents (or citizens) within a given time period are included in the definition. (Porta M (ed), A dictionary of Epidemiology. New York. Oxford University Press. 5th edition 2008)

[3] The term "population-based registry" should be used when all with a given trait, exposure or event, are intended to be included in the registry. If the registry includes all in the population (even the oldest), it becomes a population registry. Intention rather than performance defines the terms. A population-based disease registry aims at including all with the disease in the population, be it self-reported, clinically diagnosed or detected at screening. Population and population-based registries may be further classified as good or bad quality depending on coverage or other characteristics. (Olsen J, Basso O and Sørensen HT. What is a population-based registry?, Scand J Public Health 1999;27:78)

short-term infectious disease). The disease registry could be hospital/clinic-based or population based. The former is used for a specific disease irrespective of the location of the case. Alternately, a population based registry is used to compile information on specified diseases by region, community, and state in which they are diagnosed. The aims of disease or condition registries are most often primarily descriptive, such as describing the typical clinical features of individuals with a disease, variations in phenotype, and the clinical progression of the disease over time (i.e. natural course of disease). Value of disease registries is increasingly recognized as they are able to provide historically comparable data and long-term evaluation, potentially serving as an addition to randomized clinical trials, and thus providing insights about real-sites outcomes that could not be addressed in the limited controlled studies. These registries become even more important to regulators (and other parties involved) when the disease cases are rare or require highly specialised health intervention. Here registries may be the only means by which data can be obtained.

Disease or condition registries are defined by patients having the same diagnosis, such as cystic fibrosis or heart failure, or same group of conditions such as disability (1).

As an example of an EU project/initiative concerning improving disease registries in terms of defining purposes, legal context, semantic and technical aspects, **EUBIROD** ("European Best Information through Regional Outcomes in Diabetes") (http://www.eubirod.eu) is mentioned here. The project aims at sharing knowledge about prevention, treatment and patient care. Although there is a large amount of data and reports available, the information on diabetes in Europe is scattered and underutilized. For this reason, the objective of the EUBIROD project was to improve information to the public and formulate appropriate strategies, policies and actions and targeting appropriate sustainable coordination, in the area of health information, collection of data and information, comparability issues, exchange of data and information within and between Member States, continuing development of databases, analyses, and wider dissemination of information, and in fact to build a common European infrastructure for standardized information exchange in diabetes care. The main outcome of the project is a permanent and sustainable online standardised exchange of data and knowledge between EU countries.[4] Production of information is primarily enabled through the use of a common dataset[5], automatically achieving results that can later be harmonised to produce global indicators. EUBIROD aims to implement a sustainable European Diabetes Register through the coordination of existing national/regional frameworks and the systematic use of the BIRO technology. The EUBIROD project delivered also a complete platform for the accurate registration of diabetes data sources, defined "Meta Register", which includes criteria identified by the project for different areas of interest: a) to share definitions in diabetes; b) to report on data completeness and accuracy; and c) to apply common principles in the evaluation of privacy and data protection according to European legislation. Overall, EUBIROD can serve as a good example and model to be re-used for other chronic diseases as well (2).

There is also the European Academy of Allergy and Clinical Immunology (**EAACI**) (http://www.eaaci.org) as a next example of taking effort in improving disease registries. It is an association of clinicians, researchers and allied health professionals, dedicated to improving the health of people affected by allergic diseases. These diseases (e.g. asthma, hay fever, food allergies etc.) affect around 20% of children in European countries and early onset severe disease often persists into adulthood. They have an impact on quality of life in some degree, comparable to

[4] http://www.eubirod.eu/documents/downloads/BIRO_Monograph.pdf

[5] EUROBIROD Deliverable D5.1: Common dataset. Available at: http://www.eubirod.eu/documents/downloads/D5_1_Common_Dataset.pdf

diabetes and rheumatoid arthritis. Furthermore, immediate allergic reactions, for instance to foods and drugs, can be life threatening. EAACI is the primary source of expertise in Europe for all aspects of allergy and is also aimed at immunologic diseases (food and drug allergy, severe anaphylactic reactions, rheumatic and autoimmune diseases, AIDS). The EEACI association includes 41 European National Societies, and with over 7,400 members from 121 countries, research investigators and clinicians aimed at promoting basic and clinical research, collecting, assessing and disseminating scientific information, functioning as a scientific reference body for other scientific, health and political organizations, encouraging and providing training and continuous education, promoting good patient care in this important area of medicine.

The EAACI project also has goals to help standardization of data collection on allergic diseases, diagnosis and treatment and ultimately improve allergic disease and allergen exposure management. In order to achieve those goals EAACI is setting up a Task Force on allergic disease registries, whose overall objective is to provide a platform for the formation of allergic disease registries across EU country borders to develop suitable monitoring tools for use in both clinical practice and research. The initial A-reg project is focused on two national allergic disease registries that are planned to grow into a pan-European registry, namely anaphylaxis and drug allergy. Two further therapy-related projects were planned to be started de novo, one on cutaneous and systemic side effects of immunotherapy and one on immunosuppressive therapies in patients with severe atopic dermatitis. EAACI states that the main advantage of starting a registry in several European reference centres at the same time is that the same methodology ensures direct comparability (see chapter x.x.) from the start. EAACI also plans to incorporate bio banking in all of these registries for research purposes. It is anticipated that these four projects will inform the development of further allergic disease/therapy registries, especially with regard to methodology (data collection, software use, data analysis and ethics).

Since PARENT's main aim is to support EU MS in developing comparable and coherent patient registries, EAACI recognized this effort and has joined forces with the PARENT project as an official Partner Organization (3, 4).

Regarding cancer, the policymaking institutions of the EU identified cancer control as a major public health priority, and consequently many EU projects/initiative were started. As one of those projects, European Partnership for Action Against Cancer (**EPAAC**) (http://www.epaac.eu) was established for the period 2009–2013, and conceived as a framework for identifying and sharing information, capacity and expertise in cancer prevention and control, in order to avoid scattered actions and the duplication of efforts. The main objective was to assist countries in developing National Cancer Control Programmes (NCCPs), but also with goals in health promotion and prevention regarding cancer, screening and early diagnosis, research support, and mapping the existence of various data and information sources for cancer in Europe as well as checking on the availability and the quality of these data (5). Cancer registries are the main (often the only) source for incidence, survival and prevalence indicators, and can therefore be considered the cornerstone of cancer data in Europe. They are usually used in aetiological research as a means of enhancing knowledge on risk factors but also provide statistics on incidence for the purposes of assessing and controlling the impact of cancer in a given community. Given the importance of cancer registries, much effort has been made to monitor and improve the quality, type and coverage of the information they gather. With the goal to enhance comparability of cancer incidence data, promote cancer registration in the European region, and foster the use of cancer information for research and planning, the European Network of Cancer

Registries (**ENCR**) (http://www.encr.eu) was established[6] in 1989 under the EC's Europe Against Cancer programme. Today, more than 200 cancer registries are active under ENCR in Europe. Data collection systems in the EU reflect the specific organisation of national health systems, and barriers in data access persist. The move from the national to the European scale is still difficult as not all indicators are comparable across the EU (and the existence of these difficulties has been identified) through projects such as EUROCOURSE[7], EUROCHIP[8] and EUROCARE[9]). Registries presently provide most epidemiologic data on cancer, yet they are underfunded, mostly understaffed, struggling with national and European laws on protection data, or launched without proper planning (6).

In the area of cancer control, information and data are precious resources for researchers, health professionals and policymakers alike. Potential advantages in the cross-border exchange of cancer data are numerous, but achieving this goal is by no means simple. Cancer registries, being the main repository of data, vary widely in terms of geographical coverage and data quality. European initiatives, such as EUROCARE and EUROCOURSE mentioned above, have insufficient links to each other and to national databases. Moreover, data holders may be hesitant to release data due to privacy concerns, intellectual property rights or other reasons.

EPAAC project gathered insights about these issues and has, through an official proposal[10], put to attention the need of creating an integrated and comprehensive European Cancer Information System (**ECIS**). The main tasks of an ECIS should not imply collection of new data, but rather reorganisation and better coordination of existing activities. Five main types of tasks which should be carried under ECIS, have been identified: data management (each dataset flowing into ECIS must be organised according to a unique and coherent structure); data quality control (continuous improvement of quality and data standardisation as the only way for obtaining reliable data; datasets organisation (a user-friendly pathway should be implemented to structurally connect different datasets) (such cancer incidence and risk factors distribution across populations); data analysis (a plan of analysis for the main outcomes should be systematically and periodically laid down); data dissemination (the ECIS would be a key epidemiologic infrastructure for the European Research Area and results should be dissemination through general and specialised publications, press, leaflets, and web-based tools) (6).

As best suited for the role of creating ECIS, EPAAC's proposal identifies the Joint Research Centre (**JRC**) (http://ihcp.jrc.ec.europa.eu/our_activities/public-health/cancer_policy_support), which is the EC's in-house science service and with experience in harmonization and standardization of scientific/technical processes and systems,. The European Network of Cancer Registries (ENCR), from 2012 hosted by JRC, could also play a crucial role, specifically in maintaining a strict connection between the ECIS activities and those of the participating cancer registries, and therefore would be partially included in management of ECIS.

[6] Resolution of the Council and the Representatives of the Governments of the Member States, meeting within the Council, of 7 July 1986, on a programme of action of the European Communities against cancer Brussels: Official Journal of the European Union 3/07/1986 pp. 0019–0020; 1986. Available at:
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:41986X0723%2804%29:EN:HTML

[7] http://www.eurocourse.org/index.htm?do_id=947&mi_id=1338

[8] http://www.tumori.net/eurochip

[9] European Cancer Registry based study on survival and care of cancer patients; the widest collaborative research project on cancer survival in Europe. Available at: http://www.eurocare.it

[10] Available at: http://www.epaac.eu/from_heidi_wiki/WP9_proposal_EU_Cancer_Information_System.pdf

When discussing disease or condition registries, rare disease registries are given a special overview, due to their specificity. By EU definition (7), a disease or disorder is defined as rare when it affects fewer than 5 individuals in every 10000 citizens. Yet, because there are so many different RDs – between 6 000 and 8 000 – taken together they affect a significant share of the population. Between 30 and 40 million people in the EU, many of whom are children, suffer from rare diseases. Most rare diseases have genetic origins while others are the result of infections, allergies and environmental causes. They are usually chronically debilitating or even life-threatening.

Just to list a few examples, there are registries for: Niemann-Pick disease (8), Fabry disease (9) and organic acidurias and urea cycle defects (10). Common aim of rare disease registries is to contribute to better understanding of natural course/history of rare diseases, through pooling cases of rare diseases, and studying their outcomes. Additional objectives of rare disease registries are to connect affected patients[11], families and clinicians, and to support research on various (genetic, molecular, physiological) basis of rare diseases.

In the case of rare disease registries and due to low individual prevalence and the scarcity of information, knowledge and experience related to each rare disease, research is often conducted on a widest geographic scope possible (i.e. multi-nationally and/or across the continent), as benefits of international collaboration, sharing efficiencies and maximization of limited resources should be most obvious here. Also, when resources are combined, identifying standards (i.e. common data elements) becomes more important to allow data to be compared and shared across registries.

Considering the specific nature of rare disease registries another thing may come to mind – creating a single global registry for each disease (or a certain group of diseases). That however isn't always feasible, for a multitude of practical reasons and, most importantly, a single global registry wouldn't always be in the best interest of researchers.

On the EU level, much is being done to increase research, funding, and public awareness of RD (rare diseases). In 2009 the European approved a Council Recommendation (11) (2009/C 151/02) on an action in the field of RD, which covers several key areas, such as: “quality standards, including development of strategies and tools for periodical monitoring of the quality of databases and for database upkeep; a minimum common set of data to be collected for epidemiological and public health purposes; attention to user-friendliness, transparency and connectivity of databases; intellectual property, communication between databases/registries (genetic, more generically diagnostic, clinical, surveillance-driven, etc.). Importance should be given to linking international (European) databases to national and/or regional databases, when existing”[12], committed to foster exchanges of relevant experience, policies and practices between MS.

To aid the EC with the preparation and implementation of Community activities in the field of rare diseases, The European Union Committee of Experts on Rare Diseases (**EUCERD**) (http://www.eucerd.eu) was formally established in 2009.[13]

[11] **EURORDIS** (http://www.eurordis.org), as a non-governmental patient-driven alliance of patient organisations, is also bridging the gap between patients, addresses their needs and is active in promoting health policies and services and research policies and actions related to RD .

[12] Council of the European Union. Council recommendation of 8 June 2009 on an action in the field of rare diseases. Official Journal of the European Union. 2009/C 151/02. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2009:151:0007:0010:EN:PDF

[13] via the European Commission Decision (2009/872/EC). Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:315:0018:0021:EN:PDF

The EUCERD issued Recommendations on national/regional RD patient registration and data collection, which summarize the guiding principles that future actions on RD registration will rely upon and upon which harmonisation and standardisation procedures should be based across national and regional registries in Europe.[14]

From March 2012 until February 2015, the activities of the EUCERD are supported by a dedicated Joint Action, which comprises five main areas of work:

1) The implementation of plans and strategies for rare diseases at national level;
2) The standardisation of rare disease nomenclature at international level;
3) Mapping the provision of specialised social services and integration of rare diseases into mainstream social policies and services;
4) The leveraging of the value of EU networking for improving the quality of care for rare diseases;
5) The integration of RD initiatives across thematic areas and across member states (12).

Orphanet (http://www.orpha.net/consor/cgi-bin/index.php) is another initiative related to RD, and considered here a good practice. It is a reference portal and database for information on rare diseases and orphan drugs, run by a large consortium of European partners, with an aim to help improve the diagnosis, care and treatment of patients with rare diseases. Some of Orphanet's services include: an inventory of RD and its classification; an encyclopaedia of RD; a list of European RD registries (13). One of the benefits of the listed services is assistance in identification of potential data sources and collaborators.

EPIRARE (http://www.epirare.eu) (The European Platform for Rare Disease Registries) project is another important action in RD field on the EU level. Its wide-ranging mission includes several areas such as: to provide RD methods and guides for EU researchers and policy makers, while also aimed at agreeing on a common RD data set, disease-specific data collection and data validation, simultaneously addressing legal and ethical issues associated with the registration of RD projects. In order to accomplish these objectives EPIRARE has, among other things, conceived a central website platform which would share information and resources (data repository function and predefined output production), and hence increase the sustainability, networking and interoperability of registries, promote the use of standards and of registry quality procedures (common data set and quality assurance function) and provide an effective way of dissemination of the results.[15] EPIRARE has produced guidelines for data sources and quality[16] and by working on the existing registries it is attempting to formulate the core data elements, which then might be shared in a useful manner within the registry platform. The types of datasets being studied are: a minimum set of common data elements to be collected by all registries (necessary to interlink registries and to selectively extract basic data), other purpose-specific sets of common data elements (selected depending on the predefined outputs to be achieved by the platform), and project-specific sets of data elements (agreed by registries collaborating in ad hoc studies and/or in research on specific diseases).[17]

[14] List of core recommendations is available at: http://www.eucerd.eu/wp-content/uploads/2013/06/EUCERD_Recommendations_RDRegistryDataCollection_adopted.pdf

[15] EPIRARE Deliverable D5: Delivering a European Platform for Rare Disease Registries. Available at: http://www.epirare.eu/_down/del/D5_DevelopingaEuropeanPlatformforRareDiseaseRegistries%20FINAL.pdf

[16] EPIRARE Deliverable D4: Guidelines for data sources and quality for RD Registries in Europe. Available at: http://www.epirare.eu/_down/del/D4_GuidelinesfordatasourcesandqualityforRDRegistriesinEurope.pdf

[17] EPIRARE Deliverable D9.3: Common Data Set and disease-, treatment and other specific modules. Available at: http://www.epirare.eu/_down/del/D9.3_ProposalforCDE_FINAL.pdf

The PARENT project and its Registry of Registries (RoR) component is, although not RD specific, sharing several common goals with the EPIRARE project and has also envisaged a cross border platform to support research in various ways.

Another example in the rare disease category is The European Register for Multiple Sclerosis (**EUReMS**) (http://www.eurems.eu). European MS Platform (EMSP) is developing this tool to assess, compare and enhance the status of people with MS throughout the EU, enabling better data for better outcomes.

With regard to Persons with Multiple Sclerosis, across European countries there is currently a widely recognized lack of data at EU and national level on treatment and care for people with multiple sclerosis. A comprehensive approach to data collection in MS is hence needed in addressing these issues, building on existing systems for MS data collection, but bridging their gaps and limitations by incorporating additional purposes, standardized methodological strategies and ensuring its validity across different European populations. Such an effort should be built on existing national/regional data collections, registries or cohorts and using the expertise of clinicians, researchers and patient organizations involved. EUReMS was launched in 2011 and is run by a consortium of academic institutions and NGOs. The project has already made important progress (identification of MS registries and databases that are currently in use in Europe and detailed knowledge of their content and structure; producing a core dataset) on the road to providing a cross-border infrastructure for data collection, analysis, interpretation and dissemination of results in the MS field.

When concluded, through comparing data from different countries and through answering valuable scientific questions, EUReMS project should help all relevant stakeholders in Europe better understand the nature and impact of MS, and shape research and policy action towards improving the quality of life of those affected (14).

The **OSSE** project (Open Source-Registersystem für Seltene Erkrankungen in der EU / Open Source Registry System for Rare Diseases in the EU) (http://www.unimedizin-mainz.de/imbei/informatik/ag-verbundforschung/osse.html), funded by the German Federal Ministry of Health, provides a reusable software for RD registries. The aim of the project is to provide patient organizations, physicians, scientists and other parties with open-source software for the creation of patient registries. As a result, the national registry landscape would be improved to comply with European principles regarding minimum data set, data quality etc. (summarized in the EUCERD Recommendation on RD registries mentioned above) along with achieving necessary interoperability to allow federation of registries on a national and international level (e.g. distributed searches designed to comply with data protection requirements and preserve data sovereignty).

OSSE's backbone is a registry toolkit that enables scientists to build a registry for a specific rare disease even without special IT knowledge. A registry editor allows for the definition of forms for longitudinal and basic medical data and of the corresponding data schema, including an ID management/pseudonymization service (called "Mainzelliste"). ID Registry field (including, inter alia, data type, ranges, measurement units and value sets) are defined within the metadata repository (MDR) which is another integral part of the OSSE architecture, providing semantic interoperability and data quality. It is envisaged that all harmonized data sets for rare diseases would be available through the MDR. Also, each user of the OSSE registry toolkit should register with a registry of registries (RoR). Exchanging data among (national and regional) rare disease registries on the OSSE architecture is achieved taking into account data ownership and privacy aspects, through a search function with specified search queries based on the existing MDR. Depending on the search exposé which contains the description of the research question along with contact information, the data owner decides if and what to reply. Also, the OSSE architecture is not restricted to a single registry software solution but also enables integration of registries built on different software.

Another initiative which aims to develop (a) global registry/ies for a certain rare disease/s is **TREAT-NMD** (http://www.treat-nmd.eu). It is a network for the neuromuscular field that provides an infrastructure to ensure that the most promising new therapies reach patients as quickly as possible. Since its launch in January 2007 the network's focus has been on the development of tools that industry, clinicians and scientists need to bring novel therapeutic approaches through preclinical development and into the clinic, and on establishing best-practice care for neuromuscular patients worldwide. When a clinical trial is being planned, it is very important that patients suitable for that trial can be found and contacted quickly and the best way of ensuring this can happen is to make sure that patients' details are all collected together in a single database or "registry". That registry then contains all the information that researchers will need, including each patient's particular genetic defect and other key information about their disease.

The TREAT-NMD network is creating this kind of registry in countries across Europe and is also linking with other national registry efforts worldwide. The national registries all feed into a single global registry which combines the information from each of the national registries (with a pre-agreed internationally mandatory dataset), and this ensures that patients who register in their national registry can be contacted if their profile fits a clinical trial. In addition, these registries can help researchers to answer questions such as how common the individual diseases are across the world and will support other activities to improve patient care, such as the assessment of care standards in different countries. TREAT-NMD has so far created a Global Registry for Duchenne muscular dystrophy (DMD) and spinal muscular atrophy (SMA), with trials and development of innovative therapies in mind while Global registries for other conditions are in preparation. The network has also, as experienced in the creation and implementation of registries for neuromuscular conditions, issued a registries tool kit as a useful concise guide for creating a registry (be it general or NMD-specific)[18]. Some benefits of the TREAT-NMD registries include (15): one single entry point for access to patient data worldwide; registries contain accurate, verified genetic diagnosis together with key clinical data items including medication use and ambulation status; not anonymous since they contain detailed information on individual patients and thus are not simply a statistical tool but a recruitment tool; patient data is updated at least once a year; powerful feasibility tool as it can filter patients by precise mutation, age, ambulation status, medication type and location; and finally a powerful recruitment tool since patients have consented to being contacted about trials for which they may be eligible.

2.2.2 Product Registries

Once a drug or device passes the stage where it is approved for use by a regulatory authority (depending on the national state legislation) the user base becomes much bigger and from a more diverse population than the one in the stage of clinical trials, when the population is narrowly defined and only a small segment of the overall population. To address a need for quality assessment during this important post approval phase is where using a registry for identifying and enhanced understanding of product safety (acute as well as chronic use) should, as one of the available tools, come into consideration. Registries that aim to **assess safety or harm** associated with the use of

[18] Guide is available at: http://www.treat-nmd.eu/downloads/file/registries_toolkit/UK_SMA_registry_protocol.pdf

various products (drugs) or devices need to anticipate and assess the need for adverse event (AE) detection, processing, and reporting and registry sponsors are encouraged to discuss plans for AE collection and processing with local health authorities when planning a registry.

It is important to note that medical devices are significantly different from pharmaceuticals in the manner in which AEs and product problems present themselves, in the aetiology of their occurrence, and in the regulation governing the defining and reporting of these occurrences, as well as post approval study requirements.[19]

Also, compared with drugs, device technologies change more rapidly over a shorter time span, requiring device registries to adapt accordingly to the changes. In addition, healthcare providers may have different levels of experience with the device, which then may influence patient outcomes (especially with devices considered implants). Medical device registries should attempt to classify all parts of a device with as much identifying information as possible. All mentioned above special characteristics of medical devices should be thus taken into consideration when developing a device registry.

Product registries include patients who have been exposed to biopharmaceutical products, medical devices or diagnostic/therapeutic equipment (1).

Device registries can be designed for a variety of purposes, such as providing helpful information on long-term effectiveness of devices and their safety, combined with keeping track of the impact of factors such as type of surgical technique, surgeon, hospital, and patient characteristics. There are also some limitations of medical device registries, as is common amongst all observational studies. They are unable to control for often complex confounding variables as well as to take into account device version changes, surgical technique, and other unique factors all of which can lead to erroneous conclusions. However, analysis of medical device registries can often provide important information for decision making by clinicians, patients and policymakers.[20]

Post marketing vigilance of medical devices and drugs is needed as too much is unknown about the safety of the product when it's approved, and spontaneous AE reporting is a traditional (and legally binding) method through which this need is addressed. In comparison with spontaneous reporting of AE the safety/harm registers provide certain advantages, which are here considered. There are two main characteristics of these registries that are extremely important. Firstly, we know from other science fields that any choice data architecture that demands an active and non-systematic effort by the clinician to report an adverse event is inferior (in terms of under-reporting, rather than the quality of reporting) to a systematic follow-up of those events. Secondly, and related to this, in a non-systematic reporting of adverse events we usually do not know the denominator (the exposed population) and are therefore not able to provide any epidemiological measures of disease occurrence. In a structured safety/harm registry with a defined population we can calculate the incidence of adverse events and these registries are becoming increasingly more common in the area of medical products and medical devices[21].

[19] Other sources provide more information about defining and reporting of device-related AEs and product problems, and about post marketing studies (including those involving registries), such as: Baim DS MR, Kereiakes DJ, et al. Postmarket surveillance for drug-eluting coronary stents: a comprehensive approach. Circulation 2006; (113):891–7.

[20] (AHRQ) Registries for Evaluating Patient Outcomes: A User’s Guide, 3Ed, Volume 1. In: Guide, editor, 2012.

[21] the term 'medical device' covers all products, except medicines, used in healthcare for the diagnosis, prevention, monitoring or treatment of illness or disability

Thus, depending on the need to comply with a post-marketing requirement or out of a desire to complement spontaneous AE reporting, the proposed product and disease registries should also be considered as a resource. The registries could be used for examining unresolved safety issues and/or as a tool for (proactive) risk assessment in the post approval stage. Once again, the advantage of registries is that their observational method and non-restrictive design may allow for surveillance of a diverse patient population that can include sensitive subgroups and other groups not typically included in initial clinical trials (such as children or patients with multiple co-morbidities). In contrast to clinical trials, registry populations are generally more representative of the population actually using a product or undergoing a procedure. To list just a few advantages that those registry features provide: data collection may lead to insights about provider prescribing, and also follow-up duration can be long to identify consequences of long-term use (1).

Legislation on the EU level regarding pharmacovigilance for medicines marketed within the EU is provided for in: Regulation (EC) No 726/2004[22] with respect to centrally authorised medicinal products and in Directive 2001/83/EC[23] with respect to nationally authorised medicinal products (including those authorised through the mutual recognition and decentralised systems). There is also a central European medicine agency (EMA) (http://www.ema.europa.eu/ema), which could be roughly compared to the U.S. Food and Drug Administration (FDA), although not centralized and with a lesser level of authority. According to the EC Regulation[24], the EMA should take care about: the coordination of safety announcements by the MS, provision to the public of information regarding safety issues, functioning of the Pharmacovigilance Risk Assessment Committee, who are competent in the safety of medicines including the detection, assessment and communication of risk and in the design of post approval safety studies. The EMA has issued Guideline on good pharmacovigilance practices[25] (GVP) in order to facilitate the performance of pharmacovigilance activities. Finally, the EMA is responsible for the management of the EudraVigilance (https://eudravigilance.ema.europa.eu/human/index.asp) – an EU data processing network and management system for reporting and evaluating suspected adverse reactions during the development and after the market approval of medicinal products in the European Economic area (EEA). It consists of the modules, the Clinical Trial Module (EVCTM) and the Post Authorisation Module (EVPM), and is also of the main pillars of the European Risk Management Strategy[26] – a joint effort between the EMA and national Competent Authorities to strengthen the conduct of pharmacovigilance in the EEA.

The current system for medical devices is defined by European Medical Device Directive 93/42/EC[27], which sets and describes harmonized standards[28] for device manufacturing, labelling, and expected performance and safety profiles to be met. Any medical device placed on the European market must comply with the relevant legislation' where there are three types of medical devices outlined: general medical devices, active implantable medical devices, In-vitro diagnostic medical device.

[22] Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:136:0001:0033:en:PDF
[23] Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:311:0067:0128:en:PDF
[24] Regulation (EC) No 726/2004
[25] Available at: http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000345.jsp
[26] http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000306.jsp&mid=WC0b01ac058017e7fc
[27] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1993L0042:20071011:en:PDF
[28] e.g. ISO 14971 – Risk management for medical devices

Also, the EC has issued guidelines which aim at promoting a common approach by manufacturers and Notified Bodies[29] involved.[30] EC Directives also describe the basic standards for manufacturing quality-control systems and responsibilities for AE reporting.

The EC guidance documents are non-binding and offer great detail regarding handling AE and communicating safety concerns. Guidance documents also contain templates for data collection and reports, including ''clinical evaluation reports,'' which are intended to provide an outline of the technology underlying a specific device and current clinical data supporting its use, ideally in reference to established standards or similar devices. In practice, each country variously interprets the requirements for quality assurance and AE reporting.

Competent Authorities, which oversee NBs in each MS, submit AE and recall data to the European Databank on Medical Devices (EUDAMED)[31] (http://ec.europa.eu/idabc/en/document/2256/5637.html), a central database run by the EC. Since the database is non-public the basis for device approval and any post marketing commitments are largely unknown and EU-wide adverse event data is not accessible, though some MS post market surveillance events in non-systematic manner. For example, the UK Medicine and Healthcare Products Regulatory Agency (MHRA)[32] keeps AE and recall information in a searchable web portal. Another potential issue is that the clinical data forming the basis for approved devices and the post-approval studies aren't systematically published, as there is no requirement to do so (neither for NBs, manufacturers or CAs). Manufacturers and regulators are obligated to manage AE and safety problems with marketed devices through Field Safety Corrective Actions (FSCA)[33]. Recent EC Recommendations[34] for post market actions include the use of unique device identification (UDI) and connecting UDIs to EUDAMED.

As we have here first discussed certain EU principles, initiatives and legislation, we are still left with a mention of some of the best practices for product/device registries. There are numerous product/device registries in the EU, and differing in objectives, scope, field of medical expertise etc.

EU-ADR (http://www.euadr-project.org) was an EC funded project (FP7 programme) with an objective to design, develop and validate a computerized system that exploits data from electronic healthcare records and biomedical databases for the early detection of adverse drug reactions (ADRs). In this project, electronic healthcare records (EHRs) comprising demographics, drug use and clinical data of over 30 million patients from several European countries were available. EHR databases also form the foundation of the project, insofar as they supply the patient data on top of which the system is built. The EU-ADR system then intended to generate signals (drug-event pairs of pharmacovigilance interest) through the use of data mining, and epidemiological, computational and text mining techniques. Subsequently, the system was designed to substantiate these signals in the light of current knowledge and understanding of biological mechanisms, what was essentially

[29] private, for-profit third party bodies that are devices certified for marketing approval

[30] MEDDEV 2.12-1 Rev8 "Guidelines on a Medical Devices Vigilance System" and MEDDEV 2.12-2 Rev2 "Post-Market Clinical Follow-up (PMCF) Studies". Available at: http://ec.europa.eu/health/medical-devices/documents/guidelines/index_en.htm

[31] Evaluation of EUDAMED from 2012 is available at: http://ec.europa.eu/health/medical-devices/files/pdfdocs/eudamed_evaluation_en.pdf

[32] http://www.mhra.gov.uk/Devicesindustry/Vigilanceandadverseeventreporting/index.htm

[33] „action taken by a manufacturer to reduce a risk of death or serious deterioration in the state of health associated with the use of a medical device that is already placed on the market" (MEDDEV 2.12-1 Rev8 "Guidelines on a Medical Devices Vigilance System" Available at: http://ec.europa.eu/health/medical-devices/documents/guidelines/index_en.htm)

[34] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:099:0017:0024:EN:PDF

searching for evidence that would support causal inference of the signal. Finally, ultimate objective of the project was to demonstrate that an earlier detection of ADRs is possible through using EHRs (16).

EuraHS (European Registry for Abdominal Wall Hernias) (http://www.eurahs.eu) is a registry which observes hernia operations and not patients. Its mission is to develop and provide for all members of the EHS (European Hernia Society): an international platform for registration and outcome measurement; an online platform for reporting early or late mesh complications (as a survey of implant materials); a set of definitions and classifications for use in clinical research on abdominal wall hernias; a uniform method of presenting outcome results in clinical studies of its repair. It is also trying to convince existing European hernia databases to join the EuraHS, in order to collect their data on the same Internet platform, and to fulfil the goal of the registry as being a good instrument to acquire data for post marketing surveillance, increasing quality and quantity of outcome reports in hernia devices related surgeries (17).

EAR EFORT - The European Arthroplasty Register (EAR) (www.ear.efort.org) is a major activity of the European Federation on National Associations of Orthopaedics and Traumatology (EFORT). It is organized as a scientific non-profit society located in Austria, inside the EFORT organization. It was founded in 2001 as a voluntary network of national arthroplasty registries. Main aims of the EAR EFORT project are: to support national orthopaedic societies to establish national arthroplasty registries based on the EU-level standardization and harmonization of processes, basic research on comprehensive patient registries, the support for scientific activities related to arthroplasty registries and also the support of registries via EFORT. Currently, 30 projects in 26 countries in Europe, Israel and Saudi Arabia are linked to the EAR-network. The projects are in different stages of development. The work is organized on national level as a cooperation of orthopaedic societies and national public health authorities. EAR's main focus of activities is on outcome research and methodological research in context of arthroplasty registries. Arthroplasty Registries are considered as a powerful instrument to assess the performance of arthroplasty procedures, and a major source for scientific discussion. EAR supports the development of Arthroplasty Registries and register documentation, and aims to enhance the comparability of reports by standardization. EFORT also produces minimal datasets which are included in all national arthroplasty registers upon which EAR's evaluation are based (18).

As already discussed above, a product/device registry may possess great potential and effectiveness in areas of post market surveillance, adverse effect reporting, assessing safety and harm but also in improving quality of care, depending on the registry's objectives.

As an illustrative example of registry effectiveness, the **Swedish Hip Arthroplasty Register** (SHAR) (http://www.shpr.se/en) is presented. The registry started in 1979, web-based reporting system is in place since 1999, and since 2002 it measures patient reported variables.[35] In 2005 the registry also started collecting data about partial arthroplasty. The registry has excellent coverage (patient coverage 98% and hospital coverage 100% in year 2009) (19). The registry is governmentally funded, and no device manufacturing industry funding is present (although the registry sells data to industry, without identifiers). The Swedish legal context enables undisturbed data collection. The data is collected after surgery and reported to the SHAR through the internet. In accordance with the

[35] The patients complete a question form about satisfaction and health-related quality of life, serving as a baseline for comparisons one, six, and ten years after surgery.

Swedish Data Act, all patients are informed about the registry and are free to renounce their participation in the registration at any point. Analyses of register data focus primarily on re-operations, short-term complications revisions (surgeries to replace devices) and patient-reported outcomes. Revision rates for hip implants in Sweden declined substantially over the years, which is largely due to the registry's success in detecting devices for hip replacement surgery which have longer survival rates. Judging to the registry's success, for instance in comparison with other countries such as the U.S., the survival of hip replacement implants among Medicare patients in the United States (1997-2005) and patients age 65 and older in Sweden, the failure rate is about three times higher in the U.S. (20).

From its original focus on the devices themselves, the registry has moved to analyze the whole process surrounding hip implant surgery to find predictors of good and poor outcomes (21). Also, beyond the registry's quality improvement purpose, the data in SHAR have been used for research, including several doctoral dissertations and a stream of publications on outcomes associated with different prostheses and surgical techniques; age, ethnic, and socioeconomic predictors of outcomes of hip replacement surgery; the occurrence of rare adverse events; and patient-reported outcomes. The creation of the Nordic Arthroplasty Register Association that pools data from the Swedish, Norwegian, Danish, and Finnish registers is creating additional research opportunities both because of larger numbers and because the countries have different use profiles (22).

2.2.3 Health Services Registries

Finally, registries may be created in order to **measure and improve the quality of care**, defined as "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge" (23). This kind of registry usually compares patients or, sometimes, providers based on the provided treatments or achieved outcomes in relation to performance measures with "gold standards" or some other benchmarks for specific health outcomes (e.g. infection rates). These registries may be used for various purposes, such as: to monitor trends in the use of certain procedures and to evaluate trends in healthcare usage; to examine provider adherence to safety protocols and best practice guidelines; to monitor the impact of prevention efforts and public health awareness campaigns; to survey the quality of care patients receive.

These registries may identify disparities in the availability of care, identify and investigate sub-optimum practice and processes, as well as demonstrate potential improvement opportunities. The steadily increasing costs of health care (for OECD countries annual health expenditure averaging almost 5% growth rate over the period 2000-2009) (24) imply the need to justify health care interventions and plans with accurate cost/benefit measures and by showing the impact of interventions on relevant outcomes. Without a valid system for monitoring outcomes within institutions there is little space for management to be aware of how their services truly compare with services elsewhere or with pre-determined quality standards. Since a registry can continuously record data, it has the potential to identify unnecessary or inappropriate variation of healthcare quality and incite its improvement by creating a feedback loop which can pinpoint areas of poor quality (25). Longitudinal data also provides the needed understanding in order to act as an early warning system if quality declines.

Health services registries consist of patients who have had a common procedure, clinical encounter, or hospitalization (1).

Indeed, recent experience suggests that, when introducing registries on the state or national level, they become one of the most clinically valued tools for quality improvement (26). And the registries can drive quality in a variety of ways, be it indirectly – through stimulating competition, or directly – through evaluating adherence with best practices and through affecting healthcare policy (pricing and regulation). In order to improve measurement of quality health care indicators, one should fully exploit the potential of (national) registries, particularly through the implementation of unique patient identifiers, secondary diagnostic coding and present-on-admission flags.[36]

As creating and maintaining a register of this type may require (considerable) resources, narrower focus of quality registries should thus be concentrated on conditions and procedures where outcomes are thought to vary and where improvements in quality which they provide actually have the greatest capacity to improve quality of life, and reduce costs (27) (i.e. monitoring renal transplantation outcomes, as poor outcome from this procedure forces patients to revert to haemodialysis, with subsequent consequences of lower quality of life much higher costs to the society). Finally, the ultimate purpose of data from quality of care registries should be to inform clinical practice, policy development and resource allocation.

Although these registries share common objectives in improving quality and can prove a powerful tool in improving health care value, their usefulness on carrying out the objectives varies depending on the registry's stakeholders (research or health policy oriented), scope of the registry, quality of registry data and finally on the crucial issue - usage of registry data by policy makers. Simply put, in order to carry out their more widespread and systematic usage championing quality registries by key stakeholders (such as governments and professional medical societies) is needed.

The EC is supporting such initiatives, and one such is example is **HoNCAB** (http://honcab.eu) – a pilot network of hospitals related to payment of care for cross-border patients, with the main objectives to determine the patients' rights in terms of access to cross-border health assistance and entitlement to reimbursement of such treatment, to ensure access and provision of safe, high-quality, efficient and quantitatively adequate healthcare abroad, to support collaboration between Member States regarding healthcare, and finally to obtain a better understanding of the financial and organisational requirements that may arise as a result of a patient receiving healthcare outside the MS of affiliation. The network of hospitals features a functioning organisational structure and established means of communication, supported by a web-based database to collect and exchange information, all with the aim to share between MS practical experiences, problems and solutions related to cross-border care.

Benefits of quality of care registries are apparent. For an example, the registry set up by the Danish Lung Cancer Group through feedback of indicators of high-quality care derived from registry data to those delivering care has been largely responsible for improvement in 30-day, 1-year and 2-year survival rates for people with lung cancer of 1.6%, 8% and 10%, respectively (28).

There is an international momentum gathering to develop new clinical registries as quality-improvement measures.

[36] OECD Health Ministerial Meeting. Improving Value in Health Care: Measuring Quality. Forum on Quality of Care, Paris, 7-8 October 2010. Available at: http://www.oecd.org/health/ministerial/46098506.pdf

On the EU MS level the biggest effort in developing quality of care registries is in Sweden, where a system of **national quality registers** (http://www.kvalitetsregister.se) has been established (since 1970s). The system is recently on the rise, going hand in hand with a number of initiatives at both the national and local Swedish government levels and suggesting that governance of health care services is guided by an emerging performance paradigm. Although the traditional objectives of distributive justice and cost control are still valid, they have been complemented by objectives concerning efficiency and value for the money spent on health care services (29).

Today, Sweden boasts about 90 national quality registries of various types: interventions or procedures (e.g. hip fracture repair and cardiac surgery); diagnoses and episodes of care (e.g. myocardial infarction and stroke); and chronic disease (e.g. diabetes and leukaemia). National quality registers cover more than 25% of total national health expenditures, about one third of the registries collect patient data on more than 90% of all Swedish patients diagnosed with a given condition or undergoing a particular procedure, and many have been in place long enough to provide unique longitudinal data.

Thus, besides representing a comprehensive primary data source for comparative studies, the high percentage of coverage of health services enables Swedish registry data not only to play an important role in work related to monitoring and evaluation of health care quality, but also represent a significant tool to help in developing nationwide health care policy[37], while constantly and in a powerful manner enabling one of registries' common feats – resource for research.

The vision for the quality registries and the competence centres is to constitute an overall knowledge system actively used on all levels (health provider, hospital, regional, state) for continuous learning, and evaluation, development, quality improvement and management of all health care services (30).

A national quality registry contains individualized data concerning patient problems, medical interventions and outcomes after treatment; within all healthcare production. It is annually monitored (quality control) and approved for financial support by an executive committee. Funding comes from the central state level and is allocated to a few competence centres, where several registries share the costs for staff and systems that it would not be possible for a single registry to fund. The successful development of the Swedish National Quality Registries is explained largely by their decentralized nature. Caregivers that have the greatest use for the data also have the main responsibility for developing the system and its contents, and the databases are spread out among different clinical departments throughout Sweden. Another potential reason for success could be relatively liberal legal provisions concerning personal data in Sweden, where i.e. special permission can be obtained that allows national personal data to be recorded and processed[38] (even in universities) (31).

Also, data quality of registries in the national quality list is quite high and as a result sufficient for use in clinical research (32).

[37] A recent assessment of quality in Swedish health care, including the country's register system, made by the Boston Consulting group, found that the registers are improving quality and efficiency in health care. The report from BCG recognized the potential of registers to increase value in health expenditures, and they estimated that investing in registries in the Swedish context would generate a significant cumulative return over the next years because of improvements in quality.
Available at: http://www.bcg.com/documents/file64538.pdf

[38] Possible importance of privacy legislation for success of a register – see subchapter x.x. regarding the Swedish Hip Arthroplasty Register.

Outside of the EU, Australia is also trying to establish a national base of clinical quality registries with goals similar to that of Sweden, and with certain advantages (national level policy) and disadvantages (existent registries lack nationwide coverage). Clinical quality registries in Australia are envisioned as indicated in Figure 2.1Figure 2.1 (33).

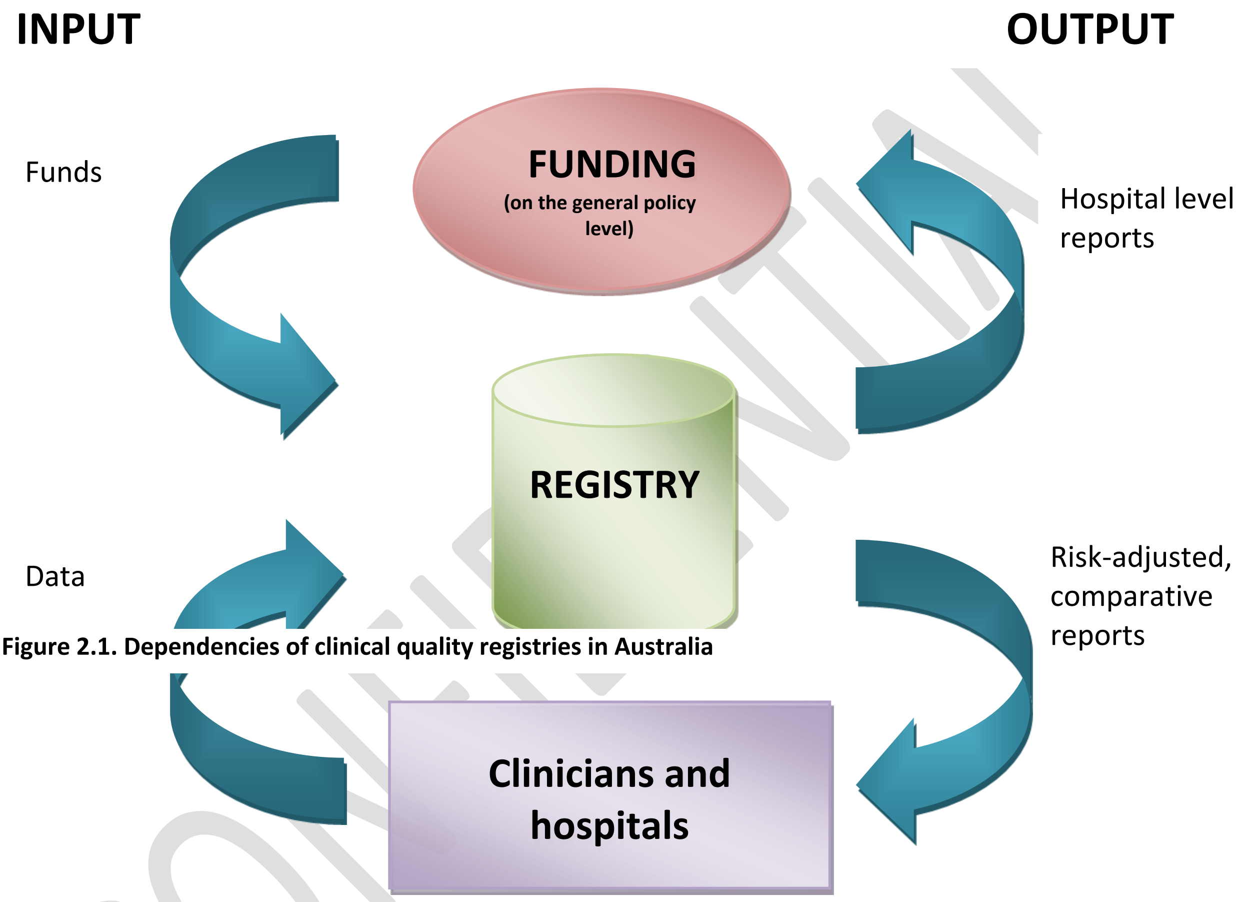


Figure 2.1. Dependencies of clinical quality registries in Australia

To summarize, when considering a clinical quality registry collection and feedback of data must be based on an effective central governance structure for the registry, with strong clinical leadership, and a regulatory framework providing incentives for quality improvement and dedicated approaches for managing poor performance. Local clinical leaders should ensure that registry outcomes drive quality improvement.

2.3 Diversity in Use of Patient Registries

As illustrated in the previous chapter, a patient registry can be a powerful tool for a number of potential needs: to understand variations in treatment and outcomes, to examine factors that influence prognosis and quality of life, to describe care patterns, to assess effectiveness, to monitor safety and harm, and to measure quality of care. Through functionalities such as feedback of data, registries are also being used to study quality improvement (34).

Registries today vary by organization, condition and type, and they exhibit different strengths and limitations accordingly. Different stakeholders perceive and may benefit from the value of registries in different ways. For a clinician, registries can collect data about disease presentation and outcomes on large numbers of patients rapidly, thereby producing a real-world picture of disease, current treatment practices, and outcomes. For a physician organization, a registry might provide data that can be used to assess the degree to which clinicians are managing a disease in accordance with evidence-based guidelines, focus attention on specific aspects of a particular disease that might otherwise be overlooked, or provide data for clinicians to compare themselves with their peers (35). From a private payer's perspective, registries can provide detailed information from large numbers of patients on how procedures, devices, or pharmaceuticals are actually used including data for evaluating their effectiveness in different populations. This information may be useful for determining coverage policies (36). Furthermore, for a drug or device manufacturer, a registry-based study might demonstrate the performance of a product in the real world, develop hypotheses, or identify patient populations that will be useful for product development, clinical trials design, and to identify individuals eligible to participate in research. The use of patient registries varies by priority condition, with cancer and cardiovascular disease having a large number of registries and areas such as developmental delays or dementia, far fewer. Overall, the use of patient registries appears to be active and growing (1).

2.4 Overview of European Registries

The current European registry landscape is often viewed as a collection of divergent registries. Design, development, and maintenance of patient registries revolve around registry platforms (software tools for managing registries' data). This approach leads to creation of segregated silos, resulting in expensive and inflexible IT systems. Often, registries are built for a single purpose, with its own data stores and for limited user profiles. Furthermore, registries have different legislative and governance rules and obligations and are spread across different European countries and types of organizations. As a result, patient registries implement only a subset of the registry functions, using and producing only a subset of the registry data, and often not applying existing interoperability approaches (standards, best practices). Thus these registries manifest themselves as islands of data and governance rules.

However, some efforts are being made to improve the situation. Through performing a literature review numerous such projects have been identified and recognized as best practices (briefly presented in chapter 2.2.).

The criteria for recognizing best practices are in accordance with overall PARENT aims, and include projects, organizations, initiatives, registries etc. working on national, regional or international level in the field of:

- Recognizing and converging similar sources of data (based on disease, device and/or service) in order to improve surveillance, quality, outcomes, safety and/or effectiveness.
- Tackling different levels of data exchange (individual or aggregated level, metadata) between similar (group of registries) or different sources of data (registries – EHR – insurance databases).
- Healthcare data exchange issues such as standards, interoperability, metadata, platform, common datasets etc.

- Defining and addressing needs for efficient health information exchange on different levels (patients, health care providers, researchers, payers, decision makers etc.) and ways to address those.
- Promoting collaboration, reducing redundancies, and improving transparency among patient registry holders.
- Patient registries classification, definitions, taxonomy, purpose, development and governance.
- Adding value through secondary use of health data.

In an attempt to provide insight and overview of the current European patient registry situation we have also included a compiled list of patient registries and a short descriptive analysis of some of their features, presented in the next subchapter.

2.4.1 Member States level registries overview

To the present date, PARENT WP4 team in collaboration with project partners compiled a list of registries that were identified as suitable for taking into consideration as regional/national/county and/or local level patient registries. The list currently contains 1028 registries and is continuously growing as additional information arrives (newly discovered data sources, literature, and information from project partners). It should be noted that the results presented in this chapter below are based on the response gathered from the project partners.

Looking into the distribution of registries across European countries, the most of them are represented in Spain, mainly due to specific organisational principle of Spanish healthcare registries. The vast majority of patient registries in Spain are county-based, which means that each contains equivalent registries (e.g. Basque Country Cancer Registry, Murcia cancer registry etc.), while many other countries may collect the same type of data on a national level (e.g. Polish national cancer registry). Other highly represented countries are also characterized by a comparatively high-level of organization of healthcare registries at a national level. These often provide meta-registries or registry lists (UK DocDat, IR HIQA, PT DIS, SE Nationella Kvalitetsregister etc.) which provide information on a large number of patient registries that are, or have been operating within a certain country. Less prominently featured countries often contain a smaller number of active patient registries in total, but may also be underrepresented due to a lower level of international visibility due to organizational issues, lack of connectedness between registries at a national level and/or lack of other specialized focal organizations at an international level which would connect them to a larger number of registries at a trans-national level (such as the European Arthroplasty registry, which consists of arthroplasty registries from across Europe). Although these international organization often contain comprehensive lists of patient registries, they are often characterized by a specific focus (like Orphanet, which contains the most comprehensive list of data on 641 patient registries, but consists only of rare diseases registries), which is why there would likely be immense benefits from an establishment of a general cross-border meta-registry organized around collecting data on all active patient registries. There are also already several multi-country registries in our list which collect data from several countries at once. These may be either international registries of specific conditions such as coronary events, or specialized international studies collecting patient data.

Table 2.2. Distribution of identified registries across European countries

Country	N	Country	N
Spain	191	Latvia	17
UK	139	Estonia	16
France	82	Slovenia	15
Portugal	66	Netherlands	14
Ireland	65	Multi-country	13
Germany	41	Czech Republic	11
Hungary	40	Switzerland	10
Austria	38	Malta	9
Italy	38	Cyprus	8
Finland	32	Greece	7
Sweden	29	Romania	6
Croatia	28	Lithuania	4
Poland	24	Serbia	2
Norway	23	Albania	1
Belgium	19	Bulgaria	1
Denmark	19	Georgia	1
Slovakia	18	Turkey	1
		Total	**1028**

Based on our general classification (primary purpose) we recognized that the majority (64%) of patient registries were disease/condition based, followed by service (26%) and product based patient registries (10%).

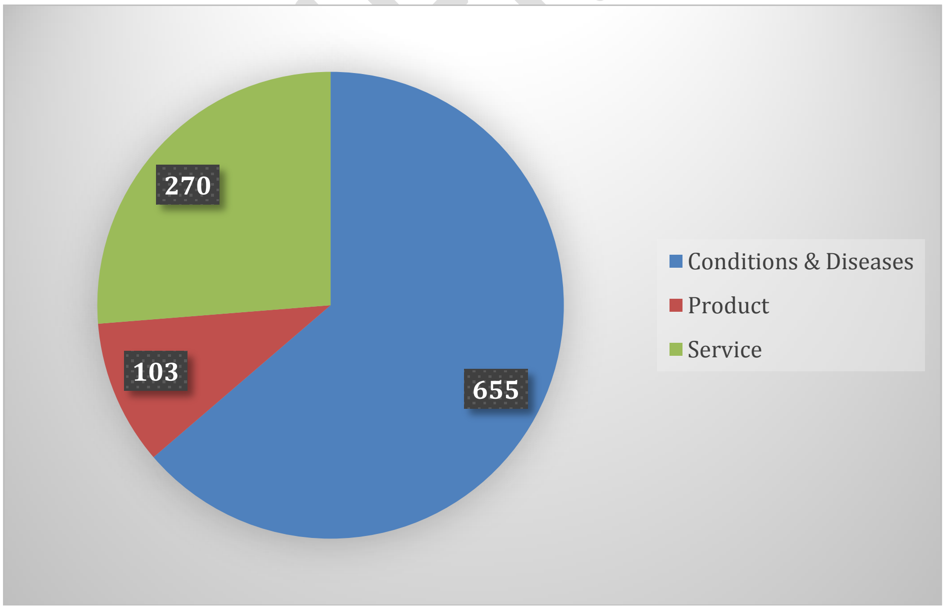


Figure 2.2. Breakdown of all registries based on primary purpose (N=1028)

Subcategorizing **disease/condition based patient registries** according to the entry criteria specific definition (particular disease or condition) we recognized several subdivisions based on organ system (cardiovascular, neuromuscular etc.) or clinical field (cancer, rare, congenital, occupational) irrespective of body part focus. The largest number of disease/condition based registries in our list falls under the coronary/vascular subcategory (27%)[39], followed by cancer/tumor/haematological (20%), infectious (9%), rheumatic (8%) and pulmonary (7%). Although rare diseases contribute only to 6% of registries in our list, the extended list[40] contains 641 rare disease registries in total (to be integrated as a joint activity of PARENT and Orphanet). All other subcategories account for 23% of total disease/condition registries in the list.

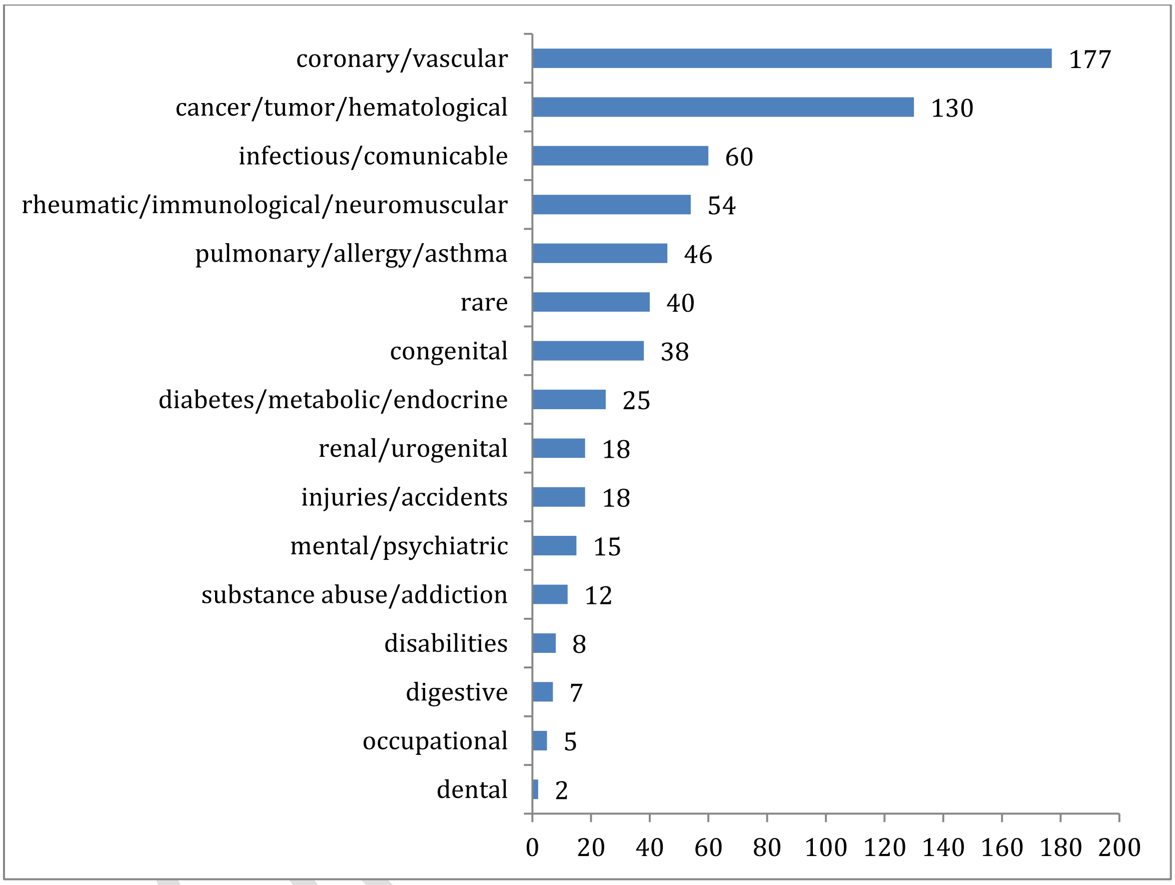


Figure 2.3. Breakdown of Condition and Disease based registries (N=655)

While the number of **product-based patient registries** represented a minority of all registries in our list, a further division can be identified within them defined by two subcategories: device registries (most prominently featuring devices such as pacemakers or arthroprostetics) and pharma registries (registries collecting data on pharmacological products). Less than 20% of product registries belong in the latter category, while the much larger proportion of product registries where recognized as medical device-based registries.

[39] mainly due to integration of European Society of Cardiology (ESC) meta registry data – dynamic portal (http://www.esc-crt.org/workstream/Pages/dynamic-portal.aspx Accessed: 9th June 2014)

[40] Orphanet list of rare diseases registries, January 2014 (http://www.orpha.net/orphacom/cahiers/docs/GB/Registries.pdf)

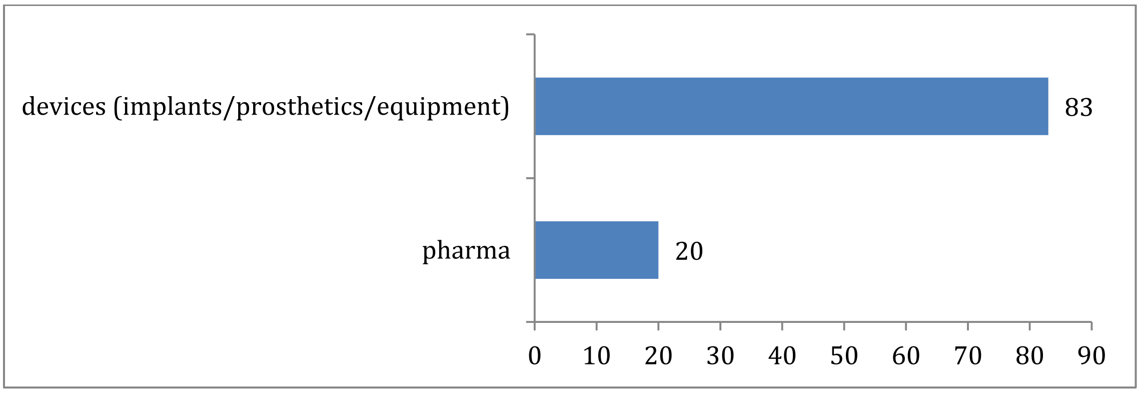


Figure 2.4. Breakdown of Product based registries (N=103)

Apart from condition/disease and product registries, our review of patient registries yielded a third category of registries which we categorize as **service-based patient registries**. This group is the most heterogeneous of all and consists of registries whose primary definition and focus is ostensibly based upon healthcare services. The largest identified subcategory contains registries evaluating preventative services, quality of care, and health monitoring. It accounts for exactly third of all service-based registries and includes population, permanent sample and vulnerable groups registries and registries used for evaluating preventative screening programs or monitoring population's health. Second biggest subgroup contains various specific medical procedures registries (24%) which monitor specialized surgical procedures, therapeutic or diagnostic services or emergency interventions. All other observed service based registries accounted for less than a half of this group and were subcategorized as registries of transplant procedures and/or donors (blood, bone marrow, organ etc.), various obstetric and gynaecological services registries (births, abortions, medically assisted fertilization), immunization, causes of deaths registries, hospital discharges registries and registries for health/social insurance purposes.

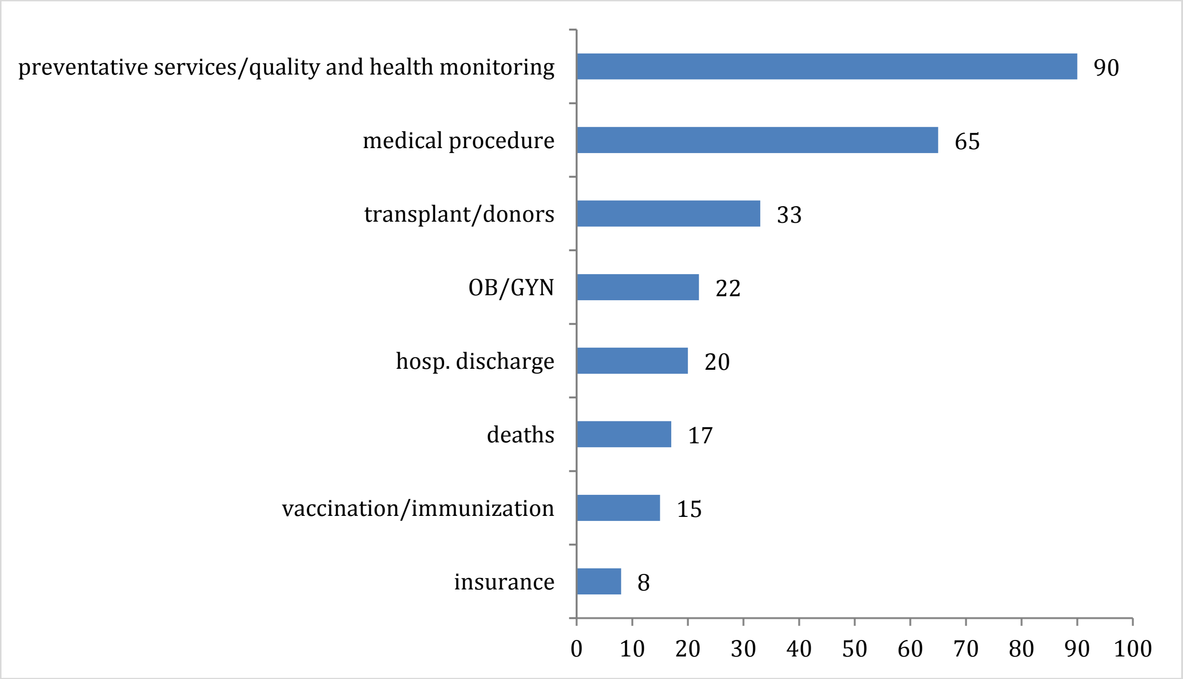


Figure 2.5. Breakdown of Service based registries (N=270)

2.5 Key issues arising within registries

Within this subchapter, only the most important and emerging issues arising within registries will be briefly explained. The possible solutions and proposals are listed in later chapters. At the beginning, one should distinguish between emerging issues at national / regional level and at EU level since the ways of setting-up and running a registry differs. At the same time, it is necessary to point out that majority of EU-level registries are based on secondary data sources.

Not all of below listed issues are relevant for national or EU registries, but when setting-up the registry, all of them should be considered.

1. The most important issues among EU registries is **unstable funding and** therefore **limited sustainability**. At this point, the differences among national (or regional) registries and EU-level registries. Funding of national registries by national authorities might not be stable; as mentioned in "PARENT - Deliverable 5: Registry analysis and Report", only half of the registries are currently funded by national government authority, about 16 % have "no specific funding". EU registries are funded either by an umbrella organisation or by a certain project what again introduces instability and limited sustainability.

Table 2.3. Funding source

Funding source Q7	Initial registry funding - set-up		Current registry funding	
	N	%	N	%
National government authority	58	36%	76	52%
No specific funding	27	17%	24	16%
Regional Authority	18	11%	18	12%
University/Research Institute	14	9%	1	1%
Foundation	12	8%	8	5%
EU commission agency	10	6%	3	2%
Hospital	10	6%	5	3%
Industry	9	6%	5	3%
Patient Association	2	1%	6	4%
Total	160*	100%	146	100%

2. There are many **legal issues** concerning registry set-up, data protection and re-use. Legal backgrounds in members states differ a lot. At this stage, the preparation of the new regulation on data protection should be pointed out as it might influence the future of majority of patient registries in EU. Much more on these issues is described in chapter 5.
3. Within the phase of development (or setting-up) the registry and also later on, **the roles of different stakeholders** are very important and, in many cases, not very clear. There are different possible roles: data owners, data holders, data users, etc. Much more on these issues is described in chapter 6.
4. **Modes of data collection**: almost half of the EU registries are still based on paper-and-pen mode (paper based questionnaires, paper based health records and laboratory results). The situation is burdensome for data providers and causes lower data quality. One should point

out that paper-and-pen data collection mode is nowadays not desirable since it is costly, time consuming and does not allow any controls of the data filled in.

Table 2.4. Sources of data for a registry

Sources of data Q17	N	%
Paper based questionnaires	67	22%
Electronic health care records	56	18%
Online questionnaires	53	17%
Paper based health records	44	14%
Paper based laboratory results	34	11%
Electronic laboratory results	26	8%
Directly from clinical examinations	17	5%
Interviews	14	5%
Total	311	100%

5. **Lack of awareness of existing standards and standard processes** when building or maintaining a patient registry. These standards are actually wanted by registry holders.
6. **Balance between accuracy and timeliness** is usually skewed in favour of accuracy, resulting in low timeliness. Comparability over time and/or space (as another quality component) is often limited due to set-up procedures, specific funding, etc.
7. **Data quality** (including completeness) is often compromised. There is low awareness of existing quality standards and there is also a lack of knowledge on quality assessment. On the other hand, only 20 % of registry holders would like to have a common quality control tool (see "PARENT - Deliverable 5: Registry analysis and Report").
8. Registry transparency and openness with the emphasis on **data access for research purposes**: majority of registries are closed for researchers from other institutions than data holder. There should be protocols enabling users / researchers to access the data under certain conditions (see project Data Without Boundaries: http://www.dwbproject.org/).
9. **Insufficient data dissemination**: minority of registries actually disseminates their aggregated data on the websites allowing users to get an easy access to the first results. The honourable exceptions are cancer registries with wide dissemination (see http://eu-cancer.iarc.fr/EUCAN/Default.aspx, http://eu-cancer.iarc.fr/EUREG/Default.aspx, http://eu-cancer.iarc.fr/EUREG/Default.aspx). These registries have established standards that should be followed by other registries.

References

1. Gliklich RE, Dreyer NA, eds. Registries for Evaluating Patient Outcomes: A User's Guide, 3Ed, Volume 1. In: Guide, editor, 2012.
2. http://www.eubirod.eu
3. http://www.eaaci.org/activities/task-forces/1904.html
4. http://www.eaaci.org
5. Available at: http://www.epaac.eu
6. Martin-Moreno JM, Albrecht Tit, Krnel Radoš S, eds. Boosting Innovation and Cooperation in European Cancer Control. Ljubljana, 2013. Available at: http://www.epaac.eu/images/OF_Ljubljana/Cancer_book_web_version.pdf
7. http://europa.eu/rapid/press-release_MEMO-14-141_en.htm
8. Patterson MC, Mengel E, Wijburg FA, Muller A, Schwierin B, Drevon H, Vanier MT, Pineda M. Disease and patient characteristics in NP-C patients: findings from an international disease registry. Orphanet J Rare Dis. 2013 Jan 16;8:12 http://www.ojrd.com/content/pdf/1750-1172-8-12.pdf
9. https://www.lsdregistry.net/fabryregistry/hcp/understd/freg_hc_u_aboutreg.asp
10. https://www.eimd-registry.org/
11. Council Recommendation of 8 June 2009 on an action in the field of rare diseases (2009/C 151/02) http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2009:151:0007:0010:EN:PDF
12. Aymé S., Rodwell C., eds., „2013 Report on the State of the Art of Rare Disease Activities in Europe", July 2013. Available at: http://www.eucerd.eu/upload/file/Reports/2013ReportStateofArtRDActivitiesIII.pdf
13. http://www.orpha.net/consor/cgi-bin/Education_AboutOrphanet.php?lng=EN Accessed: 30.5.2014.
14. http://www.eurems.eu
15. http://www.treat-nmd.eu/industry/trial-sites-and-patients/patient-registries
16. http://www.euadr-project.org/drupal/files/pdf/FinalPubishableSummary.pdf
17. Muysoms F, Campanelli G, Champault GG, DeBeaux C, Dietz UA, Jeekel J, Klinge U, Köckerling F, Mandala V, Montgomery A, Morales Conde S, Puppe F, Simmermacher RKJ, Śmietański M, Miserez M. EuraHS: the development of an international online platform for registration and outcome measurement of ventral abdominal wall hernia repair. Hernia. 2012 Jun; 16(3): 239–250.
18. http://www.ear.efort.org
19. Leonardsson O, Garellick G, Kärrholm J, Åkesson K, Rogmark C. Changes in implant choice and surgical technique for hemiarthroplasty: 21,346 procedures from the Swedish Hip Arthroplasty Register 2005–2009. Acta Orthop. 2012 February; 83(1): 7–13.
20. Kurtz SM et al. "Future Clinical and Economic Impact of Revision Total Hip and Knee Arthroplasty." Journal of Bone and Joint Surgery 2007; 89(Suppl 3): 144-151.
21. Swedish Hip Arthroplasty Register. Annual Report 2009. Available at: http://www.shpr.se/Libraries/Documents/AnnualReport-2009-EN.sflb.ashx
22. Gray BH. Registries as a Knowledge-Development Tool: The Experience of Sweden and England. Urban Institute: 2013. Available at: http://www.urban.org/UploadedPDF/412860-Registries-as-a-Knowledge-Development-Tool-The-Experience-of-Sweden-and-England.pdf
23. (IOM) http://www.iom.edu/Global/News%20Announcements/Crossing-the-Quality-Chasm-The-IOM-Health-Care-Quality-Initiative.aspx Accessed: 27.5.2014.
24. http://www.oecd.org/els/health-systems/health-expenditure.htm Accessed: 27.5.2014.

25. http://www.heart.org/idc/groups/heart-public/@wcm/@adv/documents/downloadable/ucm_438049.pdf Accessed: 27.5.2014.
26. Eyenet Sweden. Handbook for Establishing Quality Registries. Sweden: Eyenet Sweden, 2005.
27. Grosvenor Management Consulting. Australian Clinical Quality Registries Project - Final Report. 2010. Available at: http://www.safetyandquality.gov.au/wp-content/uploads/2012/02/27_External-Evaluation-Report-by-Grosvenor-Management-Consulting-REPORT-PDF-2935-KB.pdf
28. Jakobsen E, Palshof T, Osterlind K, Pilegaard H. Data from a national lung cancer registry contributes to improve outcome and quality of surgery: Danish results. Eur J Cardiothorac Surg 2009; 35: 348-352.
29. Anell A, Glenngård AH, Merkur S. Sweden: Health system review. Health Systems in Transition, 2012, 14(5):1–159.
30. Öien RF,Ovhed I. The Swedish Quality Registries and Primary Health Care. ImPrim Report, 2013. Available at: http://www.ltblekinge.se/download/18.588b0a5513a52b7563450e/WP3+Swed+Quality+Registries+ImPrim.pdf
31. Labek G, Janda W, Agreiter M, Schuh R, Böhler N. Organisation, data evaluation, interpretation and effect of arthroplasty register data on the outcome in terms of revision rate in total hip arthroplasty. Int Orthop. 2011; 35(2): 157–163.
32. Swedish Association of Local Authories and Regions (SALAR). National Healthcare Quality Registers in Sweden 2005, Stockholm, 2005. Available at: http://webbutik.skl.se/bilder/artiklar/pdf/7164-096-7.pdf
33. Adapted from: Swedish Agency for Growth Policy Analysis. Measurements for Improved Quality in Healthcare – Australia 2013.Available at: http://www.tillvaxtanalys.se/en/home/publications.html?state=view&skip=30&sv.url=12.6288e13b13a4f43c5882525
34. Labresh KA, Gliklich R, Liljestrand J, et al. Using “Get With The Guidelines” to improve cardiovascular secondary prevention. Jt Comm J Qual Patient Safety. 2003 Oct;29(10):539–50
35. Kennedy L, Craig AM. Global registries for measuring pharmacoeconomic and quality-of-life outcomes: focus on design and data collection, analysis and interpretation. Pharmacoeconomics. 2004;22(9):551–68.
36. Dhruva SS, Phurrough SE, Salive ME, et al. CMS’s landmark decision on CT colonography – examining the relevant data. N Engl J Med. 2009;360(26):2699–2701.

3 INTEROPERABILITY

3.1 Introduction

As a multi-stakeholder project and effort PARENT is a model environment in which interoperability is the key prerequisite for successful accomplishment of project objectives (including objectives and outcomes for each particular stakeholder), since all PARENT issues are, in essence, interoperability issues.

Interoperability, in the broadest sense, stands for “ability to operate with others”, thus can be applied to any situation where two or more entities achieve their goals or purpose by successfully interchanging services.

This principle can and should be applied in all aspects of registry establishment, development, operation, use and governance. This principle is also crucial for efficient cooperation with other national and EU registries and stakeholders and incorporating you registry in the connected Europe environment.

Since both registry and interoperability basic principles are by rule generic (bear no specifics regarding patient registries), these guidelines should be viewed as a brief introduction to the topic, necessary to understand, participate in and achieve the PARENT-specific goals as described in the rest of this document.

For the purpose of this chapter, the term *registry* is used for both a formal, current, verifiable, undisputable structured list of patient-related, medical or public records and for the whole organizational and technical mechanism needed to adequately maintain registry function and records and provide related services.

These guidelines will provide you with essentials for assessing and building your interoperability capabilities, and should be seen as a living reference which will be enhanced and supplemented through development of the PARENT framework.

For a deeper insight and advanced concepts behind these guidelines please consult and study the given references.

3.1.1 Contexts

3.1.1.1 EU context

Being an EU effort, the best way for PARENT to both support overall EU objectives and efficiently and effectively achieve goals is through compliance with EU interoperability context.

Interoperability is a key prerequisite for timely achievement of EU-level strategic plans by efficient and coordinated joint operative action of all stakeholders across all member states.

This is reflected throughout EU strategic and operational documents and activities. Getting acquainted with them can help you build your initial interoperability capacity and adequately position your role and goals in the context.

eHealth Action Plan 2012-2020[41] defines the overall operative context for the plan period. It is facing significant challenges of interoperability in eHealth, some of which detected by eHGI[42], and many directly reflecting as patient registry interoperability issues.

Success of EU-level action plans rely on interoperability of all involved parties. An excellent high level interoperability context overview is given in the eHealth EIF study[43].

For an insight in a service environment that demonstrates cross-border interoperability between electronic health record systems in Europe please review epSOS[44] project website.

Among the significant projects focusing on particular interoperability level issues are SemanticHealthNet[45] and EHR4CR[46].

For a practical context working model and coordination outcomes you can refer to CALLIOPE[47] EU thematic network site.

Patient registries are key healthcare information repositories, therefore interoperability of their stakeholders and users is crucial for eHealth Action Plan execution.

3.1.1.2 Registry interoperability context

In order to contextualize registry interoperability the following figures illustrate the intended approach. Figure 3.1 is shown for easier comparison of small, but significant differences between a simplified traditional registry context and the interoperability development context (Figure 3.2).

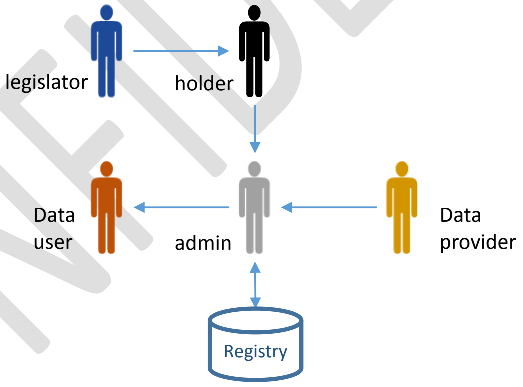


Figure 3.1. Generic single registry context

[41] eHealth Action Plan 2012-2020 – Innovative healthcare for the 21st century: http://ec.europa.eu/information_society/newsroom/cf/itemdetail.cfm?item_id=9156

[42] eHealth Governance Initiative Discussion paper on semantic and technical interoperability, page 2-3: http://www.ehgi.eu/Download/eHealth Network Paper eHGI Discussion Paper Semantic and Technical Interoperability-2012-10-22.pdf

[43] eHealth European Interoperability Framework: http://ec.europa.eu/digital-agenda/en/news/ehealth-interoperability-framework-study

[44] epSOS - Smart Open Services for European patients: http://www.epsos.eu

[45] SemanticHealthNet, a scalable and sustainable pan-European organisational and governance process for the semantic interoperability of clinical and biomedical knowledge: http://semantichealthnet.eu

[46] Electronic Health Records for Clinical Research, adaptable, reusable and scalable solutions for reusing data from Electronic Health Record systems for Clinical Research; http://ehr4cr.eu

[47] EU-funded Thematic Network "CALLIOPE - Creating a European coordination network for eHealth interoperability implementation: http://www.calliope-network.eu/

In the generic single registry context (no interoperability) all registry rules, administration and service provision is determined by the registry holder according to current legislation and business decisions. Dedicated administration (human actor) performs four-way interoperability actions when needed, as a part of business-as-usual job description. It is a closed system where each interaction is prescribed by the administration and, regardless of used IT solutions, requires human intervention in every use service.

When we want to optimize the relationship and enable parties to interoperate we need to establish a new, unprecedented joint business environment. If this is not mandated through legislation, the first prerequisite is a formal agreement/commitment of all stakeholders to jointly develop and use new functionalities (services). In eHealth EIF terms this formal mandate or formal agreement is called *political context*.

Once the political context is established, stakeholders need to make sure that their mutual interoperability is adequate to efficiently achieve the functionality determined on the political level, or what is needed to build the required capacity.

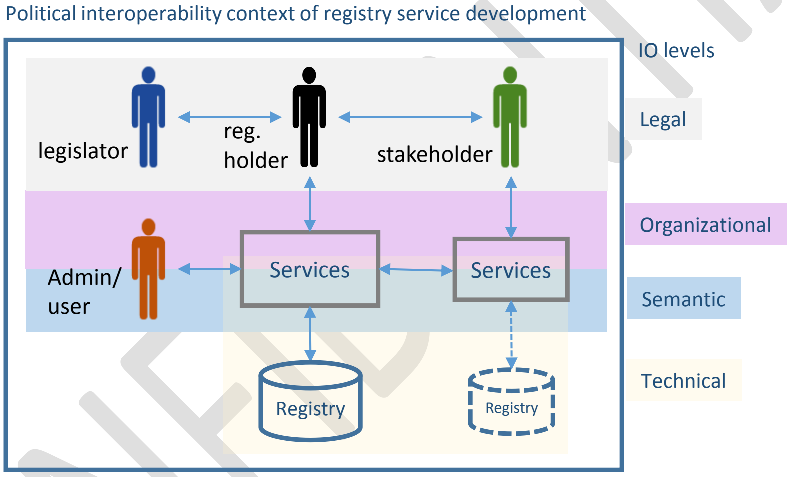


Figure 3.2. Interoperability development context

Competent and appointed stakeholder representatives need to review, verify and agree on requirements on four interdepended levels: legal (formal), organizational, semantic and technical level and adequately adjust their business systems and services. Elimination of any level from the process can result in inadequate solutions.

Once a number of stakeholders reach higher levels of interoperability we are going to be able to delegate more functions to a fully interoperable registry service context, as presented in Figure 3.3:

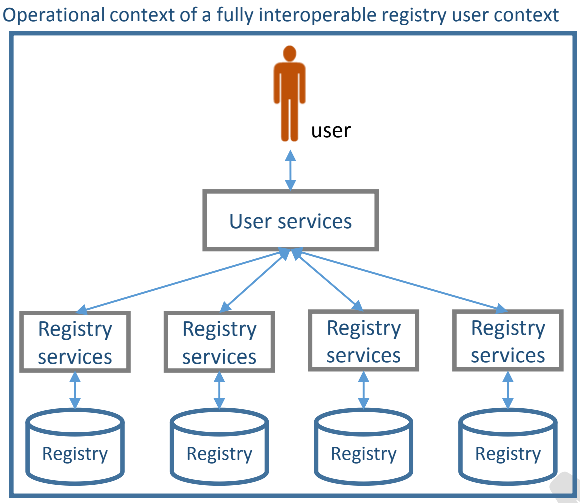


Figure 3.3. Advanced registry interoperability context

Figure 3.4 shows a functional PARENT framework (interoperability) context. The *user* now represents all stakeholder roles in previous contexts, while user services provide the interoperability environment. In this context registry holders ensure interoperability by developing and maintaining own registry services and using standardized and shared user services to interoperate either with human users or other technical systems and registries.

3.1.1.3 Generic use case context

Keeping in mind the previous contexts and disclaimers, a generic use case "*perform registry service"* covers key registry interoperability situations.

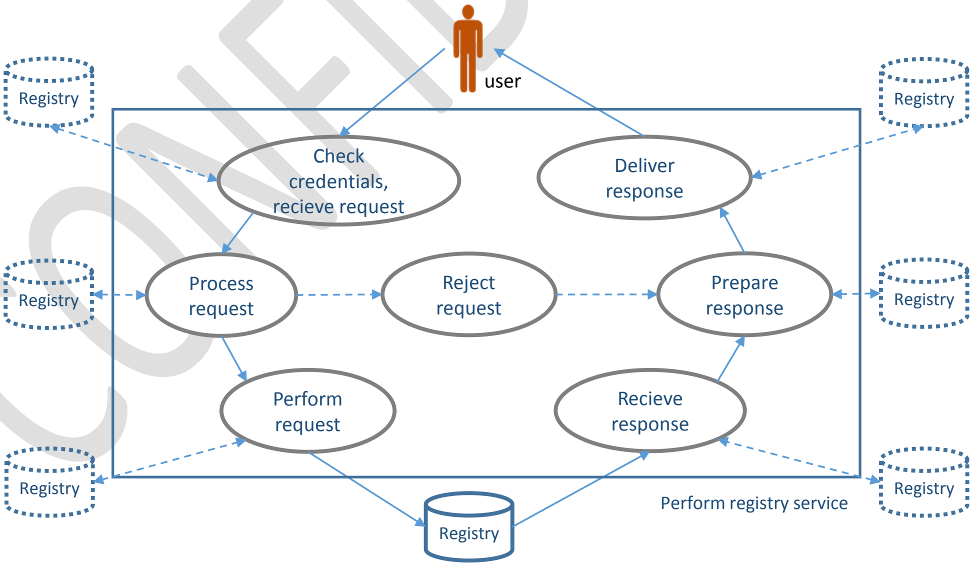


Figure 3.4. PARENT framework (interoperability) context

In this use case the actor *user* should also be viewed in the broadest sense. It can be anything, a human individual or a machine in any role: from a registry holder, physician, patient, government official to another service performed by your hospital or completely independent system.

If needed, each shown task can be also viewed as a separate interoperability use case and performed by any number of stakeholders that can provide the best result. Each task can also connect to one or more registries when the needed interoperability conditions are met.

Each task can and should be viewed form all interoperability level angles.

Each task can be performed either by a human or a machine. This enables modular development and can help a gradual interoperability capacity development for complex traditional registries. For example, at one point a nurse can perform all registry service tasks by hand and paper, while later you can automate one, several or all steps until the nurse takes position of the user and is relieved of administrative tasks.

From the interoperability point of view, the service or any particular task in this context could be delegated and performed by different independent stakeholders and systems located around the EU and supported by registries located somewhere else.

The value of this use case context is that it contains all key elements needed to perform create, read, write and delete actions (core of any registry operation) through services. This can help you map your particular registry situation to this generic model as a starting point in interoperability development that can be either your initiative or a task you participate in. It can also be a tool for conceptualizing a new registry that has all the interoperability prerequisites. The PARENT framework will provide tools to help you do that.

3.1.2 The PARENT Framework

At the initial stages of interoperability development most of the burden lays on individual stakeholders, but one of the most important interoperability possibilities is resource and capability sharing. Having a standardized system that could systematically build, incorporate and manage resource and capability sharing would relieve stakeholders and users of the interoperability and standardization overhead (the need to re-build and re-learn to be interoperable) and allow them to focus on their core business.

The PARENT framework is an objective-based interoperability framework, in other words its function is to provide a shared infrastructure for development of common interoperability support and functionalities to all stakeholders and projects joined around the PARENT objective. It should support the full interoperability range and functions as its integrator. It is an always active mechanism governed by key stakeholders on the political context level, providing support and service repositories for all interoperability levels.

On the strategic level it is intended to provide means to unify, standardize and deliver functions needed by all participating and potential stakeholders, as well as gather and disseminate information and knowledge that can generally speed up interoperability development among the target group.

On the operational level it is intended to develop functions for support of interoperability harmonization, project deployment and integration of project outcomes in the framework.

The PARENT framework' development and functionality will follow stakeholder requirements and priorities. Therefore it directly depends on the level of participation and involvement of participating stakeholders.

3.2 Registry interoperability guidelines

3.2.1 General

These guidelines primarily focus on providing initial orientation interoperability recommendations intended to aid patient registry stakeholders in grasping their interoperability environment, potential and issues, building own interoperability capacity and participating proactively in PARENT activities. The guidelines contain two viewpoints:

1. the stakeholder viewpoint focuses on interoperability issues you might directly face, and
2. the PARENT framework viewpoint, showing how the framework is envisioned to provide interoperability environment to PARENT participants, stakeholders and users.

The guidelines structure follows the interoperability structure as described in the eHealth EIF document.

3.2.2 The political (stakeholder) context

The political context, once well defined, is a simple and powerful overview tool. It can be compared to a letter of intent, defining a common goal, participants and their responsibilities in a multi-stakeholder development initiative.

The political context defines general initiative goals and measures. All later outcomes of the interoperability process on each level should be checked and verified against the context.
The political context can be defined in various ways:

- it can be prescribed through public legislation, strategies and planning,
- by a project or taskforce initiation documents,
- it is proposed by an initiative leader
- you define it with partners or by yourself.

Whenever you are required or want to participate/lead in one of the listed possibilities you need to assess your position in the context: who are the entities you need to interoperate with and how each relation reflects on your environment and operation.

The context simplifies problem analysis and solution drafting, since the relationship overview helps you to detect requirements, differences and interdependencies that define what changes you need to make (or propose) to achieve interoperability. It is recommended to make and maintain one even for simple situations (2-3 stakeholders). Often you'll quickly find that you initially omitted important stakeholders from the picture or that there are some valuable relationships that were not apparent at the beginning.

3.2.2.1 Context stakeholders

Key interoperability stakeholders are entities whose participation is required to achieve a goal, since it can be done only if mutually interoperable.

It is advisable to include all known issue stakeholders in the context, even if at a certain point you don't consider them essential for your cause. Awareness and early inclusion of the full context can help you in anticipating or orchestrating situations that might prove critical for success or solution to a broader issue that might arise later.

Continuously review your stakeholder list, propose and consult with them to avoid the most common mistake: to omit inclusion of indispensable stakeholders. This often happens with end users.

3.2.2.2 Context maintenance

Before moving on to harmonization by issue's interoperability levels you need to have at least all key stakeholders conclusively agree on the mutual purpose, commitment, responsibilities and a well-defined scope of your joint undertaking. Lacking a commitment of a key stakeholder means there is a high probability of failure and loss of time and resource investment. In that case you should either consider reducing the scope to the level you can have all key stakeholders on board, or postpone all further activities.

Interoperable development is an iterative process, allowing continuous adjustment and correction at all levels. If an issue emerges that challenges the political context (whether in outer environment, on political level or on lower levels, stakeholders must jointly review the issue and decide how to handle it. This can result in the context revision that requires revision of all lower interoperability level developments.

3.2.2.3 PARENT framework context

When defining or participating in a political context it is useful to compare it to the PARENT framework context, since it is a prototype of all patient registry contexts. All (current) general stakeholder groups are presented, and you can replace generic group names with stakeholder names in your context. It is most probable that your context doesn't contain all presented groups, so you simply exclude them.

You can observe that you can actually use and overlap the context for your different initiatives and projects, with the ability of transferring general models between them. You can also continuously develop your core business stack of information, helping you to be ready for joining future interoperable projects.

This context also describes stakeholder generic roles in an interoperable system. This might enable you to anticipate possible interoperability issues and prepare for them at early project consideration phases.

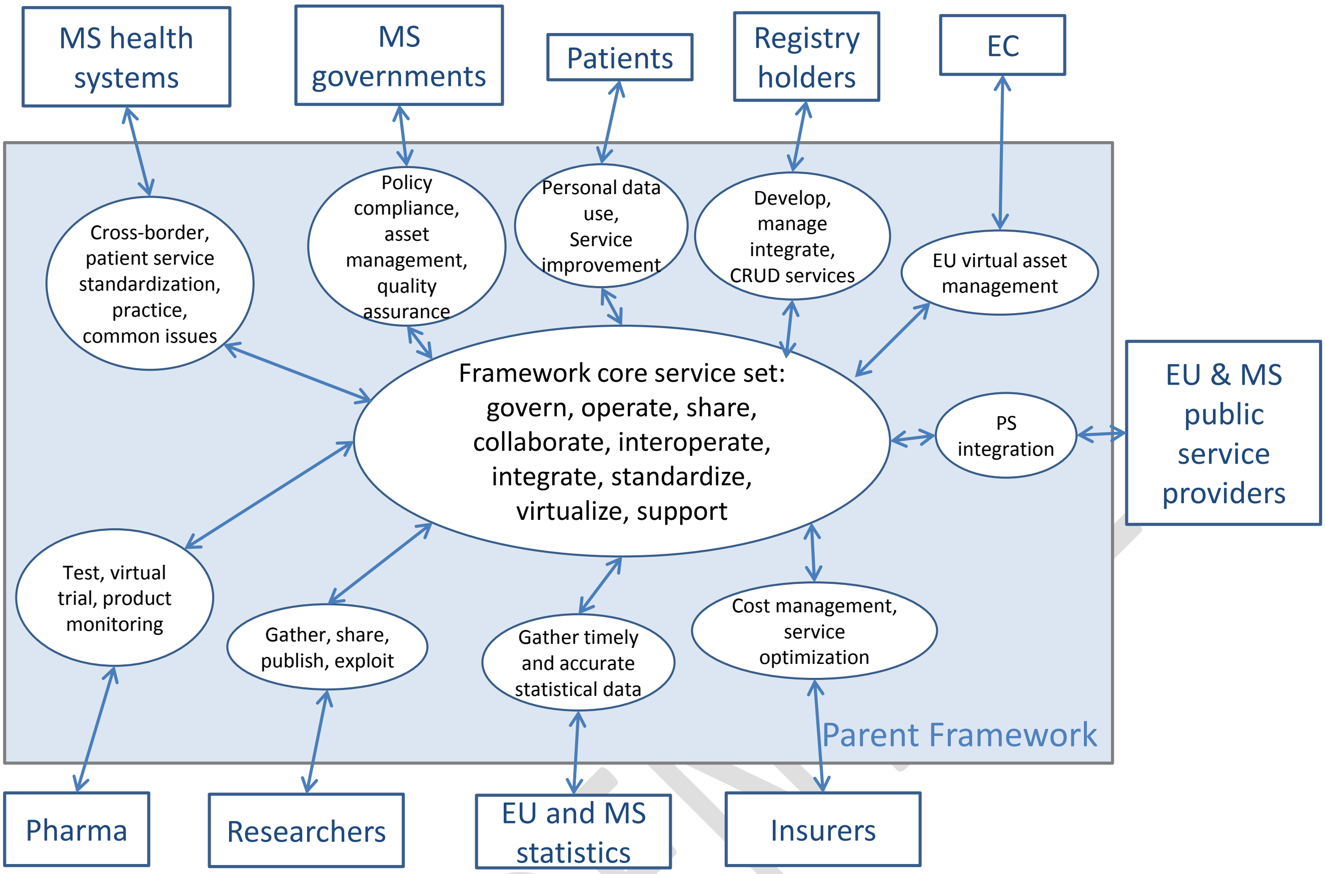

MS health systems
MS governments
Patients
Registry holders
EC
Cross-border, patient service standardization, practice, common issues
Policy compliance, asset management, quality assurance
Personal data use, Service improvement
Develop, manage integrate, CRUD services
EU virtual asset management
Framework core service set: govern, operate, share, collaborate, interoperate, integrate, standardize, virtualize, support
PS integration
EU & MS public service providers
Test, virtual trial, product monitoring
Gather, share, publish, exploit
Gather timely and accurate statistical data
Cost management, service optimization
Parent Framework
Pharma
Researchers
EU and MS statistics
Insurers

Figure 3.5. PARENT framework

On your side you might start building an interoperable service set that fits respective description and that could be universally used in many or all your contexts. When the PARENT framework prerequisites are met you might even delegate the services operation and maintenance to the framework and allow other stakeholders in your group to use them.

The central service set (envisioned to be incrementally provided by the PARENT framework) is a set of services required in all interoperability projects (person-driven or automated). For project risk management purposes actual political contexts should define stakeholder responsibilities for all of them.

3.2.3 Legal interoperability level

Patient registries need special attention to legal issues, since they contain very sensitive personal data are subject to frequent updates, and support multipoint and multi-stakeholder data exchange.

The first step after agreeing on the project political context is to review it through legal frames of the project and each key stakeholder. Besides compliance with official legislation (an important issue in cross-border projects due to legislation differences among participant countries) each stakeholder might be affected by particular legal restrictions or obligations (compliance to professional or sector rules, valid contract with other parties, constituent or owner-related issues, etc.).

If any of these presents an obstacle to the project parties must either propose a reduced project scope not obstructed by legal issues or propose feasible enabling measures or decisions for approval within the political context. Otherwise it would be wise to recommend postponement of further activity until the issues are solved or conditions met or termination of the initiative activities.

An example could be that a registry holder's country legislation forbids cross-border exchange of specific data and a research organization from another country is interested to use the data. If both stakeholders want to realize this they can work together to define a legally acceptable options to do it.

If all key stakeholders can agree on an acceptable initial legal frame that enables project continuation, project harmonization can continue to the next interoperability level.

Registry holders should pay particular attention both to domestic legal constraints and EU regulations for exchange of registry data content, storage and usage. Registries with no previous experience in exchanging data with any but traditional stakeholders, or where a part of registry processes and services are off-line should pay particular attention and start early. We strongly recommend a thorough review of the entire legal context of their registries and implications of the intended changes.
Special attention should also be taken in cases where a registry receives part of its content from other registries, in which case their holders must be included in the political context.

In some cases registry holders must also review legal situations where the intended interoperability solution might affect in any way the content delivered to users (for example where part of the delivered content now comes from other sources.

We encourage registry stakeholders to consult national personal and data exchange agencies, as well as healthcare and social security institutions, as they should have an overview of latest developments in this area.

On the EU level please take into account the following:

- Cross Border Healthcare Directive (CBHD)[48]
- Personal data protection regulation proposal[49]
- LIBE committee proposal[50]

3.2.3.1 PARENT framework formal implications

The PARENT framework should provide a general communication and harmonization platform for participation of stakeholder legal representatives on general legal review, recommendation and legislation change initiatives. At the beginning the scope will be limited to isolated cases and task force model.

The PARENT framework will enable tracking of legal environment in general and through single projects, both as an interactive reference resource and application support. The framework itself might initiate or develop own rules that would ensure optimization of the framework.

[48] CBHD: http://europa.eu/legislation_summaries/information_society/data_protection/l14012_en.htm
[49] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52012PC0011
[50] http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:52012PC0010

3.2.4 Organizational/process interoperability level

Based on the previously agreed project political context and the legal frame stakeholders need to review in detail operational responsibilities, roles, outcomes, service and data exchanges. Stakeholders should precisely define these elements for all required project processes, without overlaps or gaps.

This includes governance, quality control, and other issues pertinent to smooth, traceable, controllable, uninterruptable and conflict-free process execution, as well as risk management measures for predictable and other disrupting and undesired events.

Organizational and process interoperability is the most complex issue to tackle to achieve smooth interoperability, thus is hard to generalize or be described in a guidelines format. It can be taken as a rule that each stakeholder has a specific organizational approach, when complexity grows with the level of IT influence on the organization and processes.

Patient registries are used throughout the healthcare process, reflect on numerous national services and report to different EU level institutions. This makes this process extremely sensitive and need thorough analysis of each process step where data creation, recording, usage or exchange takes place.

This is the level where previously hidden project or legal issues might emerge. There might also be some unresolvable operational issues (e.g. inability to agree who should be responsible for a process). These should be documented and returned to the appropriate higher level for review and final decision.

An example could be a project where a patient’s treatment is carried out in a number of hospitals in different member states, so different medical records, practices and insurance schemas need to be harmonized.

A pragmatic approach on this interoperability level would be to avoid all cases that would involve complex organizational interventions in existing stakeholder business environments. It is better to develop a business case consisting of an integral set of services that can perform manageable, automatic and trackable functions. The case should reflect joint stakeholder effort to minimise the need for human intervention in the process and reduce cross-system data exchange and transformation.

A more long-term approach would be a full business implementation of SOA[51] with EDA[52], also known as SOA 2.0.

The business level is also the level where an IT business strategy and financial implications should be agreed on. It is recommended to maximize and optimize use of existing business and IT resources. It is up to stakeholders to determine which stakeholder will perform which function without jeopardizing functionality and sustainability of the whole system.

In general, use of a standardized business modelling is recommended, and BPMN encouraged, to facilitate articulation of the solution and its communication to the IT level.

[51] Services oriented architecture

[52] Event-driven architecture

When all key stakeholders agree on the acceptable organizational model and business process specification that enables project continuation, project harmonization can continue to the next interoperability level.

3.2.4.1 PARENT framework organizational and process implications

Although at the beginning the framework is not intended to go beyond resource provision and development support on this level, it is actually a potential provider for all shareable organizational and process models and building blocks resulting from successful interoperability projects by stakeholders.

PARENT should develop and provide the following organizational support on the framework level for participation of stakeholder organizational leaders, for the purpose of issue harmonization, project execution and deployment, and for local stakeholder application:

- A governance model and services,
- a collaboration model and services,
- a roles and responsibilities model and services,
- a quality assurance model and services.

3.2.5 Semantic interoperability level

Patient registries, particularly in the EU cross-border data, multi-lingual and information exchange and sharing, need a careful semantic consideration, analysis and harmonization.

After agreeing on a clear legal and business project definitions they should both pass the stakeholder semantic review and result in full agreement and implementation (where required). The main focus is on these general semantic review areas:

- Processes that generate or transform data exchanged between stakeholders,
- The data that is exchanged,
- Roles and identifiers of stakeholders present in the process,
- Information and instructions for general users given or exchanged on all system access points,
- Information system metadata, data structures and ontologies.

This doesn’t exclude any other aspect of semantic review and acceptance.

3.2.5.1 Standards, models and tools

Since the semantic interoperability is a highly structured, rule and standard-rich segment governing terminology, knowledge, standard and document interpretation, identifiers, etc. all agreements should aim to be compliant with standards or practices dominantly accepted for a particular area, particularly if determined on the EU level.

Naturally, the interoperability process requires initial assessment of stakeholders current compliance with semantic standards, models and tools, so you should be able to be acquainted and ready to exchange such information about your system with others and be aware of acceptable alternatives to be able to adjust, should you be find interest in doing so.

Here is an overview list of key semantic standards, tools and approaches for your reference. Their implementation and use closely depends on particular circumstances.

Metadata

- ISO/CEN Metadata standard 11179
- Dublin Core Metadata

Data structure/exchange

- OpenEHR
- HL7 RIM CDA, C-CDA
- HL7 FHIR
- I2b2
- ISO/CEN 13606
- IHE
- Clinical information modelling initiative

Terminologies

- CTS2 standard
- IHTSDO SNOMED-CT
- ICD10
- LOINC
- ATC

Ontologies

- OWL

Pharma and research

- C-DISC
- BRIDG

Semantic approach

- Archetypes
- Templates

3.2.5.2 PARENT framework organizational and process implications

PARENT should develop and provide the following organizational support on the framework level for participation of stakeholder semantic experts. The main areas:

- data harmonization, unification and standards,
- ontologies,
- data integration and reuse rules,
- multilingual support,
- archetypes,
- PARENT dictionary,
- data quality.

3.2.6 Technical interoperability level

This level of interoperability should be reviewed after all previous levels are fully harmonized and defined, since together they represent a detailed system specification. The most important part of the specification comes, from the IT point of view, from the organizational/process interoperability level.

Depending on the agreed business process and responsibilities division there are numerous possibilities regarding use, interconnection and sharing of existing stakeholder IT systems, using cloud capacities, building shared infrastructure, or EU modelled infrastructure, such as Connecting Europe Facilities (CEF), etc.

Before engaging in IT interoperability harmonization you should assess your IT system employed standards in the context of agreements reached on previous interoperability levels. You should review the following:

- The database solution,
- The business application solutions used,
- Web technologies used and supported,
- Web portal and interface used,
- Communications protocols supported.

In your interoperability projects you will probably discover that Patient registries currently operate on a highly diverse IT infrastructures, and it would be unreasonable to expect their major modification in foreseeable time, due to complexity, sensitivity and risk of such action. That's why, in general, your project technical interoperability efforts should focus more on solutions which rely on web technology based service and data exchange between existing IT systems wherever possible, in a way that uses existing systems without major modifications.

In each particular interoperability case stakeholders should review and decide on the most convenient suite of standards and protocols that best match their existing systems. All new development should adopt EU initiative models and standards as much as possible. A good reference point is the epSOS project and other references given in the Introduction.

XML is a well-established and universally accepted data exchange format that should be adopted whenever feasible, particularly in mixed health and public stakeholder environment.

HL7 is a data-communication protocol and format for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services. As it is particularly developed for health systems it should be reviewed as a possible choice in dominantly health-oriented cases.

Interoperability frameworks, such as eHealth EIF and its more general counterpart and predecessor EIF Annex II[53] give models for IT implementation of interoperable solutions. The development and implementation of new IT systems, as well as for more advanced cases, should be founded on these models.

[53] European Interoperability Framework: http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf

3.2.6.1 PARENT framework technical implications

PARENT framework fully implements the eHealth EIF and EIF Annex II in the SOA 2.0 environment, including the service model and adequate development and operational structure.

The technical implementation of its componentized model (fully compatible with the EIF service model) should provide PARENT you with continuous quick development, sharing and reuse of PARENT service IT solutions, reducing your effort in technical interoperability harmonization and development. Each component is a set of dedicated non-redundant services that comprise the whole framework service portfolio.

Incorporation of each of your interoperability projects into the PARENT framework will augment the PARENT stakeholder service portfolio and reduce the need to develop new IT solutions for each new interoperable business case.

Actual PARENT framework IT environment will implement EU regulations, guidelines project results, key standards and technologies and take into account your actual technical status, ensuring your and framework interoperability with systems and projects out of PARENT boundaries.

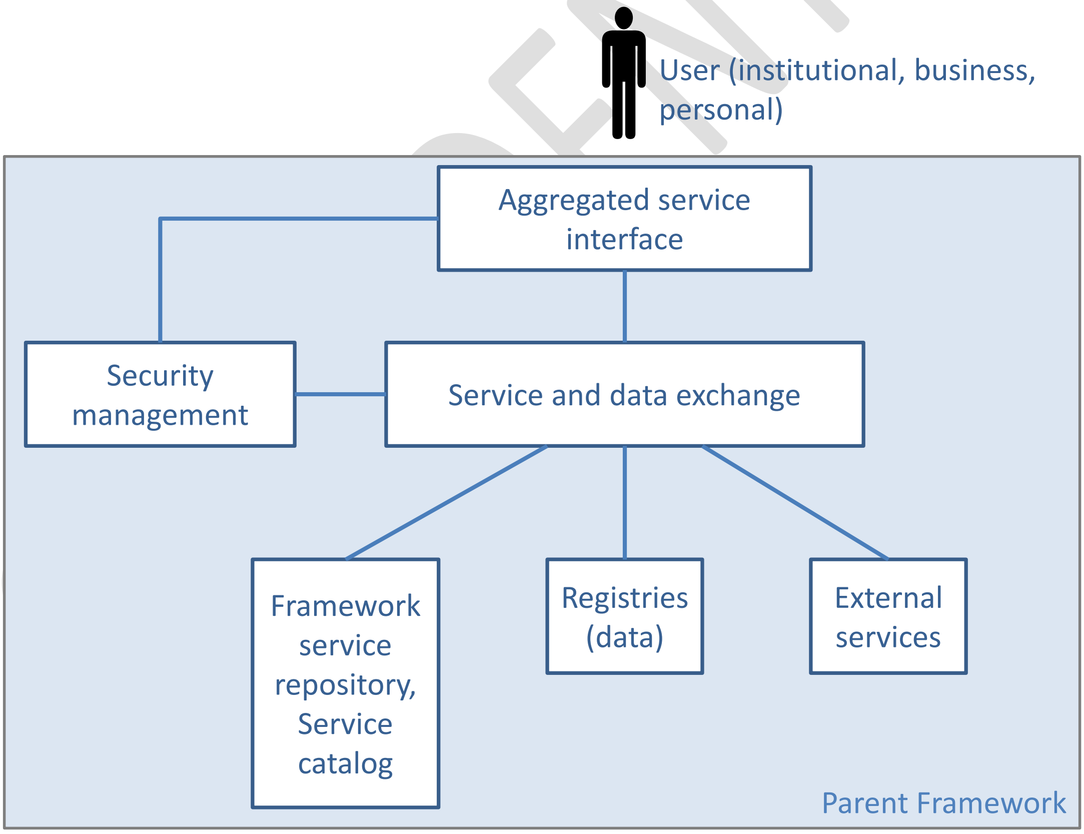

User (institutional, business, personal)
Aggregated service interface
Security management
Service and data exchange
Framework service repository, Service catalog
Registries (data)
External services
Parent Framework

Figure 3.6. PARENT framework technical infrastructure model

4 QUALITY DIMENSIONS OF REGISTRIES

In broader terms, quality can be defined as “the standard of something as measured against other things of a similar kind; the degree of excellence of something (1)”. In that light, other quality dimensions can also be (and should be) assessed. This way data quality remains the primary dimension within registry quality evaluation, but acknowledging that it is influenced by other identifiable registry features. Based on such rather holistic view and through conducting a literature review we have identified numerous “quality influencing factors” and categorized them into four groups, which are not to be viewed separately. These are: 1) governance; 2) data quality; 3) information; 4) ethical issues, security and privacy (Figure 4.1).

It is useful to consider these categories while planning and evaluating registries, since they should, rounded all together, provide a rough estimate basis for assessing registry performance.

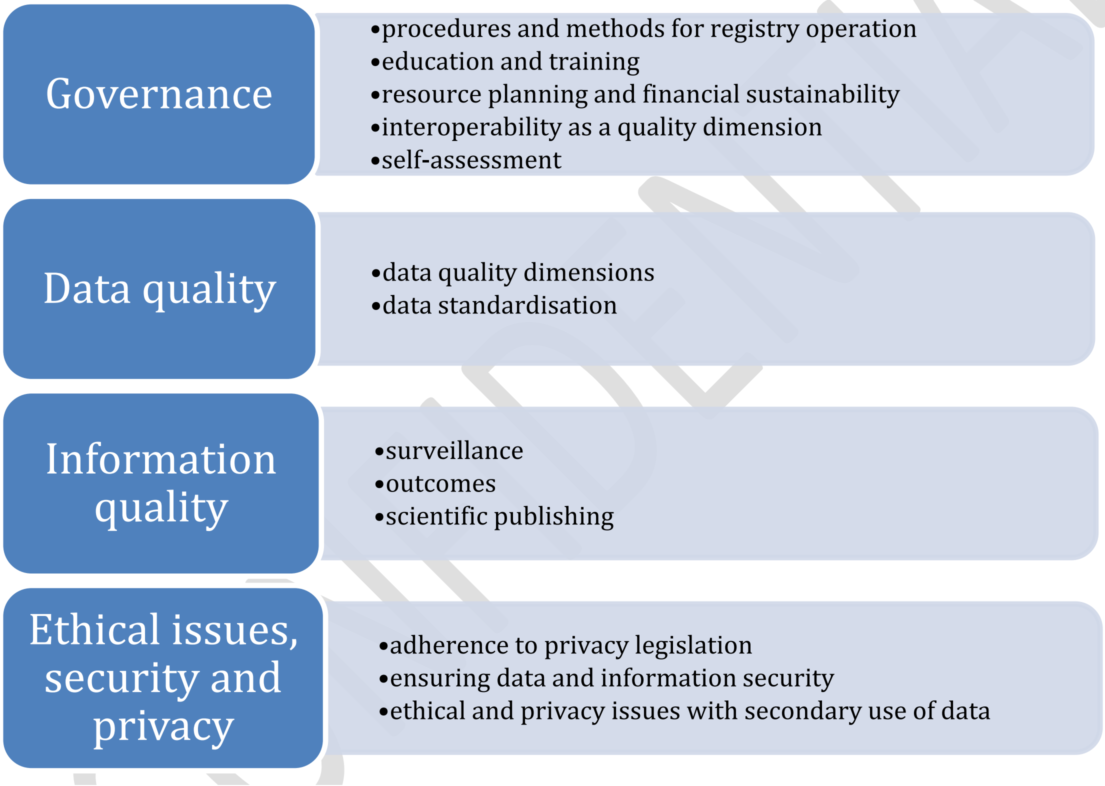

Governance
•procedures and methods for registry operation
•education and training
•resource planning and financial sustainability
•interoperability as a quality dimension
•self-assessment
Data quality
•data quality dimensions
•data standardisation
Information quality
•surveillance
•outcomes
•scientific publishing
Ethical issues, security and privacy
•adherence to privacy legislation
•ensuring data and information security
•ethical and privacy issues with secondary use of data

Figure 4.1. Quality dimensions of registries

4.1 Governance

Governance and management are the organisational foundations for patient registries, by:

- providing the framework to ensure that the registry achieves objectives set on its establishment
- driving the registry’s functioning in terms of securing resources (financial, human, technical), measuring performance and ensuring sustainability
- influencing data quality and registry outputs regarding dissemination of information
- accordance with legal aspects

Applying proper governance principles should ensure that robust operational procedures and processes are in place, clearly communicated, and easy to access for everyone involved in the data collection. Besides basic managerial and operative functions, the goal of apt governance should also be transparency to stakeholders in operations, decision making, and finally in reporting of results.

Governance is thus mostly concerned with guidance and decision making, which include the topics of registry concept, funding and dissemination of information. Governance plan is important at the registry's onset as it substantially determines future functioning. Therefore, the plan for registry governance and oversight should clearly address issues such as overall direction and operations, scientific content, ethics, safety, data access, publications, and change management. It is also helpful to plan for the entire lifespan of a registry, including how and when the registry will end and any plans for transition at that time (2).

Specific elements of the governance quality dimension are presented below.

4.1.1 Procedures and methods for registry operation

In order to justify the holding of personal health data the establishment of a register first requires a clearly stated purpose. The stated purpose should contain a brief description why is the registry established and what are its intended uses (e.g. program administration, service delivery or research). This purpose should also be subjected to review and change in case that the objectives and aims of the registry change. Statement of purpose should also contain information such as: full (legal) name of the registry, contact details, name of responsible registry holder, year started, overall function, objectives, list of data providers, legal basis for establishment, legislation and standards (privacy, national, international) that the registry must adhere to (3). Determining the appropriate scope of the registry, data set and target population, along with a study plan or protocol is fundamental to proper data collection and finally to the future quality functioning of the registry. At the registry's outset proper documentation managing should be upheld, in that the goals of the registry, its design, target population, all procedures related to data (methods and procedures for data collection, clearly defined data elements and items, data management, appropriate data analysis and reporting practice procedures), and how human subjects will be protected (privacy legislation) should all also be documented. It is very important for a patient registry to have a complete and detailed manual containing descriptions of protocols, policies, structures and procedures. Documenting policies and procedures of the registry enables it to become more process dependent than person dependent, this way potentially enabling an increase in data quality stability and reliability. Document management should be an active process, maintaining and updating documentation through the registry's further operational period. One more feature closely linked to document management is the registry's overall adaptability, as technical, regulatory and ethical frameworks of the registry should be periodically reviewed in order to address possible newly emerging issues.

4.1.2 Education and training

Staff education and training is another important aspect of registry quality. Inadequate registry staff training may cause data quality issues as well as security breaches and/or privacy violation. Sufficient staff qualification and training is thus necessary, and this can be achieved through training and education. All staff should receive training and education relating to their roles and specific job responsibilities, as well as proper training on the patient registry protocol and procedures, data sources, data collection systems and data definitions (with interpretation), accompanied by formal record of training and education (2).

For instance, if registry governance decides on applying standards, it doesn’t all by itself necessarily lead to enhancing any of the registry quality dimensions. Such action also demands achieving a certain satisfactory level of education and training of registry staff, in order to ensure straightforward implementation process of standards. For an example, the ICD-10 terminology, depending on the purpose of use (cause of death, cancer, discharges, DRG, infectious diseases), requires appropriate levels of education and training fit for and according to purpose of use.

The registry governance should have a training plan through which refresher training is to be provided on an ongoing basis. Training content should also regularly be subjected to review and updates, following potential changes in legislation, and national and international standards (3).

Training is not only important for registry staff, but also for the staff of the healthcare unit which provides data for the registry, in order to increase data quality. Training includes various methods, from providing manuals for proper data collection and data dictionaries to organizing training sessions with data providers (clinicians etc.) as participants where e.g. data extraction guidelines are discussed and practiced with patient cases (4).

4.1.3 Resource planning and financial sustainability

Since achieving objectives foremost relies on available resources (human, physical, financial), the managing organisation responsible for the registry should plan and manage its resources to ensure that they are used as efficiently and effectively. Resources should be adequate to ensure the sustainability, continual relevance and maximum impact of the data for which the registry holders are responsible. Considering that budgets are limited, careful planning and management of the use of resources is crucial to ensure they are used in the most efficient, useful and effective manner. How resources are used very much influences the quality of the information provided and the future sustainability of the registry. The allocation of resources is therefore also a fundamental factor in the delivery of quality data (3). One of the more promising ways to provide financial sustainability is collaboration amongst all the stakeholders involved in the registry, an approach which can reduce or avoid duplication of efforts, foster improved quality and robustness of data collected, and finally in a positive way to sustain registries as long-term ventures (5).

4.1.4 Interoperability as a quality dimension

Interoperability can be viewed as a quality dimension under the governance group in regard to the following:

Impact on any particular registry quality dimension cannot be pursued only within the registry holder (e.g. institution), as it is also necessary to influence the business processes and *modus operandi* of other registry stakeholders (data sources, identified data users, health information authorities etc.). Ideally, interoperability should be established through a gradual process of connecting internal processes of the stated stakeholders, therefore transforming the business processes towards convergence and making them mutual and public. Interoperability concepts and issues as well as interoperability as an envisioned common goal for patient registries across Europe is discussed further in (sub)chapter x.x.

4.1.5 Self-assessment

One of registry governance roles should be to consider how to ensure overall quality to a level sufficient for the intended purposes; therefore registries must pay careful attention to quality assurance issues. Quality assurance is important for any registry to ensure that appropriate patients are being enrolled and that the data being collected are accurate. Quality assurance activities can help to identify data quality issues resulting from inadequate training, incomplete case identification or sampling, misunderstanding or misapplication of inclusion/exclusion criteria, or misinterpretation of data elements and hence improve the overall quality of the registry data (2).

Self-assessment should perform quality control and serve to identify the sources of potential data quality issues and assess them by using indicators on data quality dimensions, developing measurements for evaluation, subsequently used to correct issues and track improvements. Use of quality assessments is also recommended to guide any decision on changing or modifying the registry's practices and procedures. Self-assessment can be an important registry governance feature as it is in fact, through using the quality assurance procedures and through using the results of quality control, a great self-propelling mechanism that ensures continual quality improvement.

Data quality improvements can be based on regular internal data quality audits including the quality of coding that incorporate clinician input (data source) as well as on external audits and external data quality reports. Self-assessment refers to periodically performing quality control through a data quality assurance programme and subsequent instituting of data quality improvements based on identified quality issues. However, self-assessment is here envisaged also as a governance responsibility, which should concern not only data quality checks but also overall registry functioning.

4.2 Data quality

In addition to a full understanding of study design and methodology, analysis of registry events and outputs requires an assessment of data quality. Requirements for data collection and quality assurance should be defined during the registry creation phase, and following the "collect once, use many" rule of data collection and management, it is paramount that the data is of sufficient quality, as the information and subsequent use for multiple potential purposes are all derived from that initial data. Data quality can be defined as the totality of features and characteristics of a data set that bear on its ability to satisfy the needs that result from the intended use of the data (6). High-quality data are then data that fit for use by data consumers, data that have sufficient usefulness and usability. This fact leads to viewing data quality as having many attributes, or in other words data quality is presented as a complex multidimensional concept.

4.2.1 Data quality dimensions and its assessment

Determining the quality of data is possible through data assessment against a list of dimensions which can be defined and measured. Data quality dimensions can be defined as a “set of data quality attributes that represent a single aspect or construct of data quality” (7). The dimensions are organized in a data quality framework, which attempts to capture all aspects of data quality that are important to data consumers.

Deciding on a list of quality dimensions is mainly dependent on the patient registry context (nation and/or region specific provisions, legal obligations etc.), type and purpose. When defining a data quality framework, in order to ensure subsequent appropriate measurements of data quality, the developer should take care to include all the context relevant data quality dimensions.

Through a review of the literature a large number of distinct data quality attributes that might determine usability was identified.[54]In an attempt to describe data quality most of the data quality dimensions were overlapping and had different interpretations, often with ambiguous definitions or completely lacking definitions, while the two most frequently cited were data „accuracy“ and „completeness“.

Trying to list all internationally used data quality dimensions and include their interpretations would prove a futile effort. Thus, the underlying principle for deciding on these dimensions and arranging them into a meaningful whole was providing comprehensive coverage while keeping dimensions organized in a collectively exhaustive way. Mutual exclusiveness was desired but is hardly achievable at the general level of description. Here is proposed a set of total 6 data quality dimensions (Table 4.1. Data quality dimensionsTable 4.1).

Table 4.1. Data quality dimensions

Data quality dimension	Description
1 Accuracy	How well information in or derived from the data reflects the reality it was designed to measure.[55] It is usually characterized in terms of error in statistical estimates. It may also be described in terms of the major sources of error that potentially cause inaccuracy (e.g., coverage, sampling, non-response, response).[56] ✓ How good are the data? ✓ What is done with the data?
2 Completeness	Extent to which all necessary data that could have registered have actually been registered.[57] It is usually described as a measure of the amount of available data from a data collection compared to the amount that was expected to be obtained[58] (e.g. coverage). ✓ Are all the appropriate data present?

[54] For a more detailed summary of the internationally commonly used data quality dimensions refer to a publication from HIQA: International Review of Data Quality. Dublin: HIQA, 2011. Available at: http://hiqa.ie/press-release/2011-04-28-international-review-data-quality

[55] Canadian Institute for Health Information, The CIHI Data Quality Framework, 2009 (Ottawa, Ont.: CIHI, 2009). Available at: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en

[56] Statistics Canada Quality Guidelines, Fourth Edition, October 2003. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2003001-eng.pdf

[57] Arts et al. Defining and Improving Data Quality in Medical Registries. J Am Med Inform Assoc. 2002;9:600-611

[58] Ehling, Manfred, Körner, Thomas. (eds.) Handbook on Data Quality Assessment Methods and Tools Eurostat, European Commission, Wiesbaden, 2007. Available at: http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf

3 Interpretability and Accessibility	Ease with which data may be understood and accessed.[59] This includes the ease with which the existence of information can be ascertained, the suitability of the form or medium through which the information can be accessed, whether data are accompanied with appropriate metadata and whether information on their quality is also available (including limitation in use etc.).[60] ✓ How readily accessible are the data? ✓ How well documented are the data? ✓ How easy is it to understand the data?
4 Relevance	The degree to which data meets the current and potential needs of users. The purpose is to assess how well a data collection can adapt to change and whether it is perceived to be valuable.[61] ✓ Can user needs be anticipated and planned for? ✓ How valuable are the data?
5 Timeliness	Refers primarily to how current or up to date the data is at the time of release, by measuring the gap between the end of the reference period to which the data pertains and the date on which the data becomes available to users.[62] It is typically involved in a trade-off against accuracy. The timeliness of information will influence its relevance.[63] ✓ Are data made available in a reasonable amount of time? ✓ Are key documents released on time?
6 Coherence	Reflects the degree to which it can be successfully brought together with other statistical information within a broad analytic framework and over time. Coherence covers the internal consistency of a data collection as well as its comparability both over time and with other data sources.[64] The use of standard concepts, classifications and target populations promotes coherence, as does the use of common methodology across surveys. Coherence does not necessarily imply full numerical consistency.[65] ✓ Does the database use standard definitions for data definitions? ✓ Can common groupings be derived from the data? ✓ Can databases be joined via a common data element? ✓ Are data values being converted correctly? ✓ Are data comparable with themselves over time?

The dimensions provided in the table are applicable for different registry types (and with different objectives), however not all may be equally important.

[59] Canadian Institute for Health Information, The CIHI Data Quality Framework, 2009 (Ottawa, Ont.: CIHI, 2009). Available at: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en

[60] Statistics Canada Quality Guidelines, Fourth Edition, October 2003. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2003001-eng.pdf

[61] Canadian Institute for Health Information, The CIHI Data Quality Framework, 2009 (Ottawa, Ont.: CIHI, 2009). Available at: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en

[62] Canadian Institute for Health Information, The CIHI Data Quality Framework, 2009 (Ottawa, Ont.: CIHI, 2009). Available at: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en

[63] Statistics Canada Quality Guidelines, Fourth Edition, October 2003. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2003001-eng.pdf

[64] Australian Bureau of Statistics. ABS Data Quality Framework [Online]. Available at: http://www.abs.gov.au/AusStats/ABS@.nsf/Latestproducts/5AFFD020BC4D1130CA25734700151AA5?opendocument

[65] Statistics Canada Quality Guidelines, Fourth Edition, October 2003. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2003001-eng.pdf

Assessing quality includes adequate management of each dimension, and additionally failure in one dimension can severely hinder the usefulness of the final registry report (i.e. when considering cancer registries insisting on the dimension of data completeness may ruin the demand for timely reporting). Likewise, each of the dimensions may possess equal importance, but also there may be instances where the relative importance of one dimension exceeds another. As stated previously, the importance of a particular quality dimension depends on the set objectives of the registry, its type, as well as its scope and methodology. Specifically, based on the definition of data quality provided above, the intended use of registry data actually determines the necessary properties and requirements of the data.

For instance, in a registry that is used to calculate incidence rates of diseases, it is essential to include all existing patient cases, therefore the completeness dimension is of critical importance.

Additionally, the need to explore different aspects of data quality is an accepted practice among patient-registries, and should be accentuated when not present.

For example, population-based cancer registries are considered particularly attentive to assessing data quality, as the value of the modern cancer registry and its ability to carry out cancer control activities rely heavily on the underlying quality of its data and the quality control procedures in place (8).

Data quality regarding cancer registries is usually assessed against the following three quality dimensions: comparability, validity, completeness, as well as sometimes timeliness as the fourth one. Factors influencing data quality and methods (both quantitative and qualitative) for measuring data quality within these dimensions have been devised and made available.[66]

Data quality dimensions are components that allow the user to quickly identify specific problematic aspects of data. Interrelatedness and overlapping are always necessarily present, the quality dimensions are not specific in regard of quality measuring, and for that to be possible, as exemplified with cancer registries, decisions are needed to identify which methods and indicators are to be used in order to successfully measure the registry's data quality against the dimensions. Data quality assessment programme should thus precisely define a data quality framework, preferably logically grouped what and how should be measured and monitored in the data domain, thus making data dimensions more specific by creating data characteristics and criteria, along with a rating method. Such an example of a comprehensive method for assessing data quality is the Data Quality Framework[67], by the Canadian Institute for Health Information (CIHI) issued with the purpose of improving data quality of national health data collections. The Data Quality Framework is based on Statistics Canada guidelines and methods and information quality literature. It is a highly developed hierarchical framework model, with established criteria useful for systemic data quality assessments.

In summary, efforts should be made to create various relevant data quality dimension groups dependent on type and objectives of the registry, and devise methods and indicators for assessing

[66] Reviews of these methods are presented in more detail here: Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer 2009;45(5):747–55 and Parkin DM, Bray F. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. Eur J Cancer 2009;45:756–64.

[67] Canadian Institute for Health Information, The CIHI Data Quality Framework, 2009 (Ottawa, Ont.: CIHI, 2009). Available at: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en

data quality, so that a registry can, potentially through an assessment tool for data quality, use those methods to measure and gradually improve their data quality.

Again, the importance of data quality should be highlighted as data quality is a pre-requisite for ensuring meaningful analyses and registry outputs.

4.2.2 Mode of data collection and impact on data quality

Considering data quality not in isolation but as part of a complex whole brings out another important and often neglected aspect which can influence data quality – the point where data is collected.[68]

The quality of initial data input from clinicians and health practitioners can vary. Quite frequently incorrect patients are registered or data items can be inaccurately recorded or not recorded at all.

A sustainable workflow model is an important element of a successful registry, a workflow that can be integrated into the everyday clinical practice of doctors, nurses, pharmacists, and patients (while respecting privacy legislation). Prior to the full launch of a registry pilot testing can be organized to gather preliminary input from health care workers and others included in the data collection.

Decision should be made on the mode of data collection, as there are a few ways to collect data, where the primary difference is whether it's collected in its conventional paper or the modern electronic form.[69]

4.2.3 Improving data quality

Since data quality is critical for any registry, in order for a patient registry to improve data quality, it should seek to implement and maintain a high standard in all other here identified quality dimensions of patient registries (governance, data quality, information quality, ethical issues, security and privacy). The governance dimension is crucial here (as discussed in subchapter 4.1.5.), as initiative within an organisation for improving data quality is based on managerial decisions, which set forth standards and channel staff efforts. In this light, the Health Information and Quality Authority of Ireland (HIQA) describe "seven essentials for improving data quality"[70], which are useful to consider in the context of a patient registry. These essential features are reproduced in the table below.

[68] This issue has been briefly discussed in subchapter x.x. concerning data provider training.
[69] Methods of data collecting (paper or electronic) are discussed in subchapter 6.1.2.1.3.
[70] Health Information and Quality Authority. What you should know about Data Quality. Dublin, Ireland: HIQA, 2012.

Table 4.2. Essentials for improving data quality

Leadership & Management	• What: involves having in place executive-level responsibility, accountability and leadership. • Why: knowing who does what (e.g. the establishment of a governance committee that will ensure the registry is committed to data quality). Decision-wise, this includes the selection of only essential data elements when datasets are established.
Policies and procedures	• What: developing and implementing clear policies and procedures on data quality for staff that are based on legislation and standards. • Why: can help ensure that a high level focus on data quality is translated into good practice amongst all those involved in data collection and handling within the registry.
Standardisation	• What: ensuring that data is collected and processed in a standardised fashion (e.g. use of minimal datasets, data dictionaries and the creation of standard templates for data collection), designing the registry with respect to national and international standards. • Why: facilitates data interoperability and making data available. Also can improve consistency and reduce error.
Data quality dimensions	• What: set of data quality attributes upon which data can be assessed, aligned with policies, procedures and training. • Why: measuring and monitoring level of data quality within a registry.
Training	• What: training of the staff in the requirements and importance of data quality. • Why: ensuring that policies and procedures adopted to generate high quality data are implemented and understood in practice.
Data quality audits	• What: independent systematic examination of data (internal or external). • Why: providing feedback to all staff, indicating the areas for improvement, highlighting good practice in order to facilitate learning (e.g. automation of data collection over manual collection where possible will reduce error rate, however, this will not be verified without planned audits of data quality).
Make data available	• What: availability of data when and where needed, in accordance with information governance safeguards (security, privacy). • Why: fulfilling the purpose for which the registry was created, increasing quality of registry data through its efficient utilization and dissemination.

4.3 Information quality

Information can be considered an output and the extension of the data collection process. Its quality is measured by the purpose of its use, which in the case of patient registries can be grouped into surveillance (including health statistics), outcomes, and scientific publishing.

Scientific publishing can be considered to control for methodological prerequisites including sufficient level of data quality. Therefore, it can serve as an indirect information quality indicator. Levels of measuring can be publication amount (total, yearly), subject relevance, up-to-date, impact factor, citation index.

Similarly, statistical data from registries focused on surveillance can be used as an indirect quality measure with regards to real-world decision making. Outcome based registries serve the same purpose in terms of indirect quality measurement albeit from a different viewpoint, i.e. using information from patient registries for influencing and improving treatment outcomes. Quality information gained from patient registries hopefully leads to informed healthcare management and better decision making.

4.4 Confidentiality, security, privacy, ethical issues, secondary use of information

This quality dimension is concerned with ethical issues and confidentiality and privacy regarding use of personal health information, as well as the need of proper patient registry data security and clear provisions regarding secondary use of information. Although actually concerning data and stemming from the wider dimension of (information) governance, it is here discussed separately as it involves privacy protection, a sensitive and seminal issue when discussing patient registries.[71]

Not meeting ethical and legal requirements concerning privacy renders the patient registry inoperable. Levels of data confidentiality, privacy and security also influence registry interoperability capacity as well as information dissemination.

Creating a balance between respecting individual privacy and providing high quality personal health information can, although very important, also be a difficult task faced by patient registries as well as other healthcare related stakeholders. Striving for cross-border interconnecting and interoperability of patient registries is accompanied by emerging security risks concerning privacy, judging to the fact that health information systems present technical challenges to existing privacy protection legal frameworks.

In order to maintain the privacy of participants enrolled in a registry and the data confidentiality, security measures should be implemented. All security measures should be contained in a document that describes in detail the data security risks, policies, and procedures specific to that registry. Physical and technical safeguards should be incorporated in the collection, storage, transmission of and access to data. These include data encryption, restriction of data access, data back-ups, methods (software) for de-identification of local data during potential transmission and storage etc. Also, implementation of safeguards shouldn't be done only once, but undergo continuous review and revision.

[71] Privacy, confidentiality and security are mentioned in more detail in chapters 5 and 6.1.4.

Considering data usage, we can distinguish between two types: 1) primary purpose; 2) secondary use of data.
This classification as primary or secondary is based on the relationship of the data to the registry purpose. Primary data sources include data collected and being kept by the registry holder (custodian) for direct purposes of the registry (i.e., primarily for the registry). The secondary use of health data considers use for purposes other than those for which it was originally collected. Secondary uses include using information for (further) research, performance monitoring, service planning, audit and quality assurance purposes etc. When thinking about secondary use of health data, it is needed to carefully balance between the public interest and the individual data subject. Since secondary use of data may violate patient privacy, precautions should be taken and conditions must be satisfied if proposing to use information for secondary purposes. Clear definitions of the circumstances where data is to be used for secondary reasons should be developed.

Legislative provisions concerning the secondary use of data are typically contained within general privacy or data protection legislation, which can differ depending on the specific MS context.

The important things with secondary data use are that patients should be made aware that their information may be used for this purpose and have the benefits of the practice clearly explained to them. Likewise, consent must be obtained for the collection, use or disclosure of information for purposes outside the direct registry's data outline plan. Efforts on data anonymisation should be performed as well as using data sharing agreements which offer an additional safeguard against inappropriate use of information.

To repeat and to conclude the subchapter, researchers and other data users should disclose clearly how and why personal information is being collected, used, and secured, and should be subject to legally enforceable obligations to ensure that personally identifiable information is used appropriately and securely. In this manner, privacy protection will help not only to ensure research participation, public trust and confidence in medical research, but also prompt cross-border registry cooperation. If registry holders are confident that their information is being appropriately protected and have trust in the system, then they are more likely to share information, which leads to improved safety and quality of care at an individual level.

4.4.1 Privacy impact assessment (PIA) – a method to assess privacy OVERLAP WITH 6.1.4.1.

A privacy impact assessment (PIAs) is a tool, process or method to identify, assess, mitigate or avoid privacy risks (9). PIAs are used internationally and across all sectors but are particularly useful for healthcare providers in assisting to identify potential risks around the collection and use of personal health information as this information is categorised as being sensitive. Used as practical solution, PIAs can help respond to the new privacy challenges in the design of cross-border health information systems. The primary purpose of undertaking a PIA is to protect the rights of service users. The process involves the evaluation of broad privacy implications of projects and relevant legislative compliance, through describing how data is collected, processed, disseminated and published. Where potential privacy risks are identified, a search is undertaken, in consultation with stakeholders, for ways to avoid or mitigate these risks and to facilitate solutions which help safeguard privacy. As PIA considers the future privacy consequences of a proposed project that involves the collection and use of personal health information, it is most beneficial when conducted in the early stages of a project, and ideally at the planning stage (3).

Related with the goals of the PARENT project, a very useful PIA initiative has been identified with the EUBIROD project. EUBIROD explored privacy issues at the level of systems' users, assessing the variability of data processing approaches in MS and their deviation from EU privacy standards and legislation, and by using the adapted version of the Canadian PIA Guidelines. Key elements of data protection (factors) were selected to ascertain the compliance/non-compliance with privacy principles/norms of data processing operations occurring in EUBIROD registries.

Registry privacy and data protection which should be investigated when conducting PIA are:

"accountability of personal information"; "collection of personal information"; "consent"; "use of personal information"; "disclosure and disposition of personal information"; "accuracy of personal information"; "safeguarding personal information"; "openness"; "individual access to personal information"; "challenging compliance"; "anonymisation process for secondary uses of health data" (10).

References

1. Kohn LT, Corrigan JM, Donaldson MS, eds. To Err Is Human: Building a Safer Health System. Institute of Medicine, Washington: Committee on Quality of Health Care in America, 2000.
2. (AHRQ) Registries for Evaluating Patient Outcomes: A User's Guide, 3Ed, Volume 2. In: Guide, editor, 2012.
3. HIQA) Guiding Principles for National Health and Social Care Data Collections. Dublin: HIQA, 2013. Available at: http://www.hiqa.ie/publications/guiding-principles-national-health-and-social-care-data-collections
4. Arts DGT,Bosman RJ ,de Jonge E, Joore JCA, de Keizer NF. Training in data definitions improves quality of intensive care data. Critical Care 2003;7:179-184.
5. EURORDIS-NORD-CORD Joint Declaration of 10 Key Principles for Rare Disease Patient Registries. Available at: http://download.eurordis.org/documents/pdf/EURORDIS_NORD_CORD_JointDec_Registries_FINAL.pdf
6. Arts DGT, de Keizer NF, Scheffer G-J. Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. Journal oft he AMercian Medical Informatics association. 2002;9(6):600-611.
7. Wang, RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems. 1996;12(4):5-33.
8. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methodsPart I: comparability, validity and timeliness. Eur J Cancer. 2009;45(5):747–755.
9. (NHS)Privacy Impact Assessment: care.data. Chief Data Officer, NHS England, 2014. Available at: http://www.england.nhs.uk/wp-content/uploads/2014/04/cd-pia.pdf
10. Privacy Impact Assessment Report. The EU.B.I.R.O.D. PIA Team, 2010. Available at: http://www.eubirod.eu/documents/downloads/D5_2_Privacy_Impact_Assessment.pdf

5 GENERAL REQUIREMENTS FOR CROSS-BORDER USE OF PATIENT REGISTRIES

Cross-border use of registries can take several different forms as the mapping work of PARENT has demonstrated, among others ***registry networks*** (e.g. the International Association of Cancer Registries, the Nordic Arthroplasty Register Association NARA), ***international clinical studies*** (GRACE – Global Registry of Acute Coronary Events) and ***international registries*** e.g. IBIR - International Breast Implant Registry). There are several strong drivers in using registry data across borders, such as the needs of studying differences between countries in morbidity, health system-level interventions' effectiveness and utility of procedures; the advantages of large international datasets vs. national ones in timely detection of rare, or previously unknown effects; gathering and promoting information on best practices worldwide.

Independent of the motive driving cross-border registries utilization, the success of the endeavour will always rely on the degree of achievement of certain prerequisites, the implementation of which starts at the local level - regional and/or national. PARENT Joint Action aims at the idea of establishing a continuous IT-assisted chain of health data capture, storage, processing, transmission and utilization. Therefore the purpose of fulfilling these prerequisites is the achievement of interoperability in the broadest understanding of the term, i.e. on legal, organizational, semantic and technical levels (see chapter 3) as well as the establishment of effective, sustainable solutions for cross-border registry collaboration. The focus of this chapter is primarily on the requirements imposed by legal and organizational interoperability aspects, and to a lesser extent on semantic and technical interoperability issues; these are in turn addressed in more detail in Chapters 3 and 10. An exception is the topic of metadata, which we briefly discuss here. A more detailed analysis of organizational interoperability aspects with regard to stakeholders and their roles will constitute part of the business models analysis of PARENT, an independent deliverable which is currently work in progress. It should be kept in mind that the allocation of requirements to respective interoperability aspects is at least to some extent artificial. Several requirements span many if not all levels of interoperability, even if they are discussed under a predominant heading.

Political context

The creation, maintenance and development of registries, as well as their preparedness for cross-border operations is largely dependent on the positioning (or lack thereof) of health data resources in national strategic prioritization regarding *scientific data resources and research infrastructures*. PARENT is analysing in a parallel activity national strategies and initiatives concerning Health Data and the ways in which they impact patient registry work.

Equally important is the question of whether registries are perceived as part of regional and/or national *eHealth infrastructure*. At the EU level, MS collaboration in the field of eHealth has until now focused primarily on the creation and exchange of health data at the point and for the purposes of patient care, as reflected in the work of the eHealth network on ePrescription and Patient Summary. The needs and requirements of secondary use of data, where the formation and utilization of registries also belongs have until recently remained unexplored. However, in order to achieve the vision of electronic collection, processing and re-use of health data throughout its lifetime cycle while ensuring the fulfillment of interoperability requirements, e-enabled registries need to be included as a target of national eHealth agendas, thereby establishing the link with ongoing EHR initiatives.

Organisational aspects – Registries' operations and procedures

Researchers' access to classified registry data has generally been quite complicated and time consuming starting with locating appropriate data, preparing research applications and on to requesting permissions and negotiating data transmissions or access rights. Each one of the steps in this process can take a variable length of time and incur widely differing costs, depending on the registry holding the data in question. Both elements though may turn into considerable barriers, particularly from the perspective of socially and politically urgent research work. New solutions of more straightforward application processes and remote access to data are being developed.

Procedures for granting access to or sharing data in a cross-border context must be in place, preferably including predefined response time targets. An organizational culture oriented towards, as well as appreciative of data utilization beyond its own remits, combined with appropriate resourcing are essential elements in achieving a high level of preparedness. Collaboration with other registries' holders is advisable, in order to exchange experiences, advice and ideas.

Open data is an overarching idea which stretches to cover parts of classified data in the form of metadata. Openly publishing the content information of limited access systems would boost the efficiency of scientific research, enhance the quality of results, increase transparency and help create new research ideas.

Legal aspects

The most important European law affecting patient registries' operations is the Data Protection Directive (95/46/EC) that regulates the collection, processing and distribution of personal data. Registry holders should always be aware of the basic notions and effective norms of Data Protection, as securing privacy of the research subjects is a fundamental task when establishing and maintaining a patient registry. Currently the implementations and interpretations of the Data Protection Directive vary between Member States. Additionally the roles of Ethical Committees and data protection authorities vary a lot. The legislative process toward the new harmonizing Data Protection Framework is still unfinished. Moreover, the European Union Directives and Regulations considering Medical Devises, Pharmacovigilance, Clinical Trials and Cross-Border Health Care induce new information needs that will increase demand for patient registry data. Registry holders should actively follow the ongoing overhauls of the aforementioned laws.

By and large a patient registry can be established using either of two legal instruments; by explicit consent of the data subject, or based on law. Current practices among the EU Member States registry holders' surveyed by PARENT appear to be almost equally divided between the two models. The final content of the forthcoming Data Protection Regulation will play a decisive role in the choices available for registry establishment and operations in the future.

Adoption of a consent model presumes thorough planning of the purposes of the registry. The required content of informed consent varies between Member States. That's why it is important to consult local data protection authorities or ethical committees in the process of formulating the consent model. The European Data Protection Working Party (WP 29) has given an opinion regarding the definition of consent (WP 29, 2011). The Opinion provides a thorough analysis of the concept of consent as currently used in the Data Protection Directive.

According to the existing Data protection Directive Article 2.h “the data subject’s consent” means any freely given, specific and informed indication of his/her wishes by which the data subject signifies his/her agreement to personal data relating to him/her being processed. The definition of the Directive implies an opt-in strategy of the consent. For the legal protection of the registry holder and patient, it is advisable that the consent is given in written form. If personal data is transferred abroad, this should be communicated in the context of acquiring informed consent.

Even though the European Commission has proposed a Regulation as a substitute for the existing Directive, it is also likely that some national legal variation regarding patient registries will continue to exist. These disparities reflect differences in Member States’ national health care systems, information infrastructures and legislations. Thus, it is always important to consult regional or national data protection authorities or ethical committees when establishing a registry.

The generalized interpretation has become that even encrypted and pseudonymous data are personal data. That is why it is pivotal to understand the basic notions regarding personal data in order to understand the areas where Data Protection Rules are applicable. Registry holders and data processors should always be able to differentiate clearly the notions of pseudonymous data, encrypted data, anonymised data and aggregated data. (*possible reference to PARENT glossary?*).

It is likely that the upcoming European Data Protection Framework will require more transparency and accountability from patient registry holders. Generally it is advisable to be open about the registration purposes and give clear information to maintain public trust and credibility of patient registries. This involves ethical and well-structured informed consent practices, as well as maintaining clear and open descriptions of the registry and its metadata online.

Securing privacy of the research subjects is a fundamental task when establishing and maintaining a patient registry. The generalized interpretation has become that even encrypted and pseudonymous data are personal data. That’s why it is pivotal to understand the basic notions regarding the personal data in order to understand the areas where Data Protection Rules are applicable.

According to the Data Protection Directive, *personal data* means any information relating to an identified or identifiable natural person (“data subject”). An identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

The current Data Protection Directive does not define the often used concepts of pseudonymous data nor encrypted data. According to European Parliament’s proposal given in March 2014, *pseudonymous data* means personal data that cannot be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organizational measures to ensure non-attribution. According to the same proposal e*ncrypted data* means personal data, which through technological protection measures is rendered unintelligible to any person who is not authorized to access it (European Parliament 2014).

It is notable that according to these definitions both pseudonymous data and encrypted data are considered to be personal data. Therefore the Data Protection Law applies to them.

Anonymised data means data in which all identifiers have been removed so that there is no reasonable possibility to link data back to individual persons to whom data relates and no code key exists to link the data to persons. Anonymised data are not personal data as the data has been altered so that the data subjects can no longer be identified. The possibility to re-identify data subjects must be considered on a case-by-case basis. For example, the deletion of names and personal identity numbers is often not sufficient to make data anonymous. Complete anonymity requires that the possibility for both direct and indirect identification is removed and that the code key is destroyed.

Aggregated data means statistical data on individuals that has been combined to show values without possibility to identify individuals within the aggregated data set. One practice has been to share and hand over *aggregated* or *anonymised data* in order to eschew the data protection norms. Often however, that is not possible as the analysis requires sharing of individual level data whether it is in encrypted or pseudonymous form.

The Data Controller of the Patient Registry should always be defined unequivocally. *Data Controller*, according to Data Protection Directive, means the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data. Where the purposes and means of processing are determined by national or Community laws or regulations, the controller or the specific criteria for his nomination may be designated by national or Community law.

Data Processor, according to the Data Protection Directive, means a natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller. *Third party* means any natural or legal person, public authority, agency or any other body other than the data subject, the controller, the processor and the persons who, under the direct authority of the controller or the processor, are authorized to process the data. *The recipient* means a natural or legal person, public authority, agency or any other body to whom data are disclosed, whether a third party or not.

Semantic aspects

Operating in an international environment or readiness to do so requires that solutions regarding linguistic barriers have been thought of and implemented – both at the level of data and at the level of generic information necessary for data sharing (e.g. information on procedures for access to data, application forms etc.)

The comparability and transferability of health data across languages and contexts of use is heavily dependent on the adoption and use of accepted coding standards (see chapters 3.2.5 and 10.11.2).

Metadata is “structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information source”. It is meant to describe the phenomenon it concerns, and also document its changes over time. Good quality metadata is vital for data utilization. To make datasets comparable and useful for other users and between registries, metadata should be standardized according to validated and widely used classifications. Another aspect of standardization is recording metadata elements in the register’s information model. That is, to make standardization as complete as possible, it must also cover data architecture and programming details.

When establishing and maintaining a registry, it is pivotal to identify the relevant stakeholders and generate a co-operation structure within them. The key stakeholders from the registry holders' perspective are usually health care professionals, patients, pharmaceutical and medical devices industry, ICT-suppliers, policy makers, researchers and other registries. If taken further, the opening of detailed metadata in standardized format would ease the registries' multi-stakeholder cooperation as well, particularly in the cross-border setting. The first step in opening registry metadata could include basic information about the data, such as description, owner, information content, target group, update intervals, dependencies from other data etc. preferably on the basis of agreed standards (*Reference to previous chapter)*. This kind of increased visibility and traceability of health data collections would benefit patient registries and lead to new ideas and innovations. Joining yellow-page type services like the PARENT Joint Action Registry of Registries (RoR), the AHRQ Registry of Patient Registries (RoPR) or other specialized "umbrella" registry is a concrete implementation step and opportunity for identifying further areas for targeting development efforts.

As open data has recently gained importance also on state administration level (e.g. the British government's "opening up government" initiative and the Finnish Ministry of Finances' "open data programme"), open data and the possibilities it may yield must be carefully considered in the patient registry environment. Firstly, a line must be drawn between the data which can be opened given the technical, and, above all, data security restrictions, and the data which cannot be opened (such as patient registries' microdata).

Technical aspects – Guaranteeing shareable data

There are different levels of implementation for to the technical solutions required, starting from the choices made on the level of an individual registry and up to the level of platforms for cross-border sharing of data. It is not the purpose of these guidelines to take a stand in advising for or against the use of specific technological solutions, since these are both context-specific and a constantly moving target as new technologies emerge. However, the technical layer is crucial in ensuring the 'shareability' of health data and hence adopted solutions must be such that take into account and support regional/national infrastructures and semantic requirements for patient data collected in the process of healthcare services provision.

On the level of technical operationalization of legal requirements, particularly in terms of data protection and safety, adopted solutions must be robust and reliably proven to perform the expected tasks.

Effective and sustainable solutions for cross-border registry collaboration

The creation of effective and sustainable solutions for the cross-border use of registry data is a process where all the aforementioned requirements must be concertedly brought to the play in order to serve clearly defined unique targets, such as those explored in the PARENT Joint Action Scenarios (Reference). The added value generated by achieving these targets will act as the key driver for the engagement of stakeholders who in turn can guarantee the sustainability of the required cross-border registry infrastructure and operation environment, a subject discussed in detail in the respective PARENT Joint Action report (Reference to Business Models and Sustainability).

6 CREATING A REGISTRY

6.1 Planning a registry

This guide assumes that a registry is designed to fulfil a need that can be met through the scientific analysis of predefined data, collected in a real-world setting. Though this data might ultimately be utilised to answer other questions, it is essential that registry establishment is an organised, well governed and purposeful scientific process rather than a purposeless exercise in data collection. This will ensure the creation of a resource that maximises resource allocation, efficiency and has well-defined, valuable outputs that can be measured, so that the quality and success of the registry can be verified.

Though this section evolves in what we believe is a logical, sequential process, components might be best addressed in tandem or may need to be revisited in an iterative fashion as further information becomes available. We do, however, feel that addressing each section will add value to the registry and increase the likelihood of developing a successful registry.

During each phase of planning, we advise considering how it may fit into the bigger picture, not just of the registry that is being created, but also with respect to the local, regional, national and international environment in which it is created. As the digital world becomes more connected we envisage the role of registries becoming progressively more valuable. This will only happen if they are developed in a manner that is cognisant of the importance of interoperability.

We also suggest that there is a wealth of experience to be gained from regulatory authorities, other registry groups and registry experts whose contributions could not only be helpful in the construction of a successful registry, but also critical to its implementation. We endorse in particular the creation of a resource such as PARENT's Registry of Registries or the AHRQ's Registry of Patient Registries (RoPR) which are helping to connect registries, while raising the standard of registries considerably (1, 2). We strongly advocate ensuring that any registry created forthwith joins such initiatives.

It should be noted that 2 resources in particular were of considerable help in informing the authors and in structuring this chapter. We highly recommended their utilisation as reference documents of extraordinarily high standard. These are "Registries for Evaluating Patient Outcomes: A User's Guide" and the "ISPOR Taxonomy of Patient Registries: Classification, Characteristics and Terms" (3, 4).

6.1.1 Defining the Purpose, Objectives and Outputs of the Registry

6.1.1.1 Purpose(s)

The first step is to clearly define the overarching aim(s) or purpose(s) for which the registry is being established. This may emerge from a clinical need, a post-marketing requirement, or an interest of patients or clinicians, but the purpose(s) should be capable of being realised through the prospective, non-interventional, scientific approach that a registry should adhere to.

In so much as possible, the purpose(s) should be limited in scope and number to ensure focus. As will become apparent, expansion of ideas is likely to occur rapidly once stakeholders become involved and, it is important to limit this at an early stage so as to prevent it becoming too unwieldy to manage.

As with any scientific endeavour, this process will be greatly facilitated by conducting a literature review to analyse what information already exists within the scope of the proposed registry. This might demonstrate that while the purpose and objectives are reasonable, a clinical trial or other study design might be a more appropriate means of delivering the required outputs. Furthermore, a literature review will highlight relevant experts and stakeholders in the field of interest who might be contacted as part of a stakeholder evaluation or for expert advice.

6.1.1.2 Objectives

To facilitate the generation of a valid scientific question, the registry's purpose(s) should be divided into specific objectives, which together will achieve that overarching purpose(s) of the registry. It is worth considering how each objective might translate into a dataset and to imagine whether a scientific methodology could be applied to help validate whether the objectives are achievable.

6.1.1.3 Outputs

Ultimately, a registry's findings are only valuable if the data they generate can be translated into information capable of improving health outcomes. This is more likely to occur if outputs are considered at an early stage, so as to achieve the following objectives:

- Ensure that a registry is purposeful as there will be measurable end-points against which it's success can be judged
- Identify potential experts required to advise the development of the registry
- Identify potential stakeholders
- Facilitate buy-in through the identification of outcomes of interest
- Identify the target audience for whom the information gained from a registry might be valuable. This will facilitate the most effective dissemination of results and will also help identification of unforeseen requirements. For example, the primary purpose of a drug registry might be to identify its effectiveness in a real world setting; however, mandatory reporting of adverse effects will also need to be considered.

6.1.1.4 Process outcome

Defining the purpose, objectives and outputs of a registry will typically clarify a registry as belonging to one of three groups (or a combination of these). Conversely, by considering your registry as belonging to one of these groups, may facilitate defining your registry's purpose, objectives and outputs.

- **<u>Condition based registry</u>**
    - **Purpose:** Though there are many listing registers, which identify patients suffering from a particular condition, 'condition-based registries' in this document refer to registries that aim to describe outcomes related to a particular condition.
    - **Purpose example**: description of the natural history of chronic obstructive pulmonary disease (COPD).
    - **Objective example**: identification of depression in patients with severe COPD.

    - **Output example**: Defining the prevalence of depression in COPD and examining how this might be more effectively be detected in the COPD care pathway.
    - **Registry example**: The Malta National Cancer Registry

- **<u>Product based registry</u>**
    - These registries typically focus on medical devices or pharmaceutical products.
    - **Example**: Arthroplasty Registries
        - There are multiple registries in the vast majority of EU member-states in different stages of development, which monitor approximately 4 million patients worldwide at present. Shortcuts to websites of relevant arthroplasty registries for further information are available at http://www.ear.efort.org/registers.aspx
        - **Purposes (Example):**
            - The assessment of:
                - Real-world effectiveness
                - Safety and cost effectiveness of a new device
                - Outcome monitoring of performance and potential safety issues over the entire life cycle
                - Early signal detection of inferior outcome of device and surgical techniques
                - The impact of patient profile/comorbidities/risk classes on patient side on the outcome
                - Market monitoring concerning implants and health care providers
                - Feedback to health care providers
                - Identification of fields for improvement and monitoring of effects.
        - **Objective (Example):**
            - Defining the number of post-operative complications related to the device insertion to facilitate feedback to stakeholders in order to support decision-making.
        - **Outputs (Example):**
            - Demonstrating that a device or surgical technique is associated with increased post-postoperative complications
            - Fulfilment of post-marketing obligations
            - Validation of realisation of expected value by innovations and/or premium products
            - Transparent ranking of quality achieved by implants and health care provider.

- **<u>Services based registry</u>**
    - These registries aim to evaluate the quality of service provision or correlate an intervention with outcomes.
    - **Registry example**: The Slovenian Hospital Discharge Registry
        - **Purposes (Example):**
            - The assessment of:
                - All hospital discharges (one day or longer) due to illnesses, injuries, poisonings, childbirths, stillbirths, sterilizations and new-borns in all Slovenian hospitals

        - Information for monitoring, planning, management and development of health care system
        - Health status of the population
        - Cost effectiveness
        - Patient safety quality
        - Other quality indicators.
    - **Objectives (Example):**
      - To assess the health status of the population and specific subgroups
      - Setting the priorities for developing national policies for improvements of health care system
      - To asses potential inequalities in health.
    - **Outputs (Example):**
      - Prevalence in the population for a certain disease or condition in a specific time period (e.g. year)
      - Determination of quality level for certain quality indicators:
        - patient safety indicators, e.g. postoperative complications, obstetric traumas,
        - quality indicators related to acute care, e.g. 30 day in-hospital and/or out of hospital mortality
      - Calculation of burden of specific diseases.

- **Combination**
  - As is obvious from the examples above, some registries may have aspects that belong to more than one registry type.

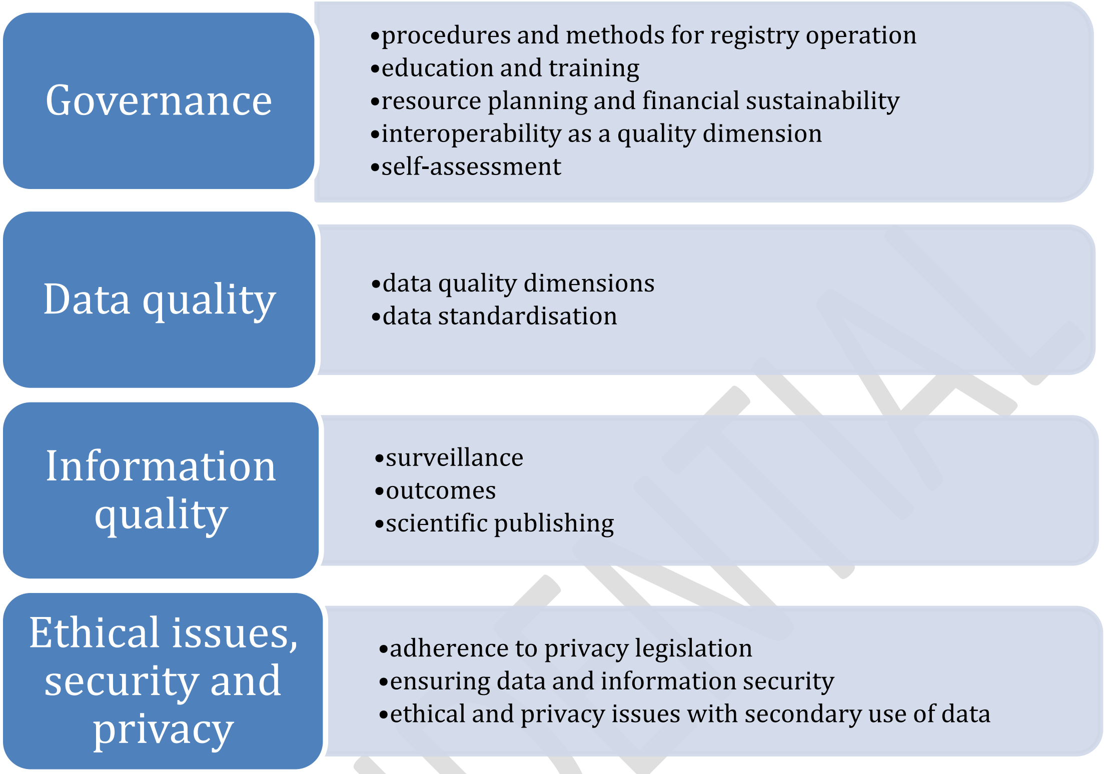
6.1.2 Data Considerations

The success of a registry will ultimately be judged on its ability to meet the goals it was created for. This requires the collection and analysis of sufficiently high quality, targeted data specified by research hypotheses and the dissemination of the results of these analyses. High-quality data is also a key component in enabling interoperability (discussed later in this chapter). Though this section might be expected to occur later in the sequence of registry planning, success-by-design warrants the consideration of the determinants of high-quality data at an early stage of planning a registry. This should result in a focus on instilling key building blocks of quality, making the process of verifying the quality of the registry much easier when audit and other quality assurance processes are conducted post-implementation.

Knowledge of key determinants of data quality and how to achieve it will raise awareness of possible obstacles that might threaten the creation of a registry, such as the absence of an electronic health record to provide useful data. This knowledge also serves to underline the importance of a considered stakeholder evaluation that avails of input from groups that the registry planners may not have considered.

6.1.2.1 Data Quality

Firstly it is worth recalling that data quality is influenced by a number of factors, categorized into four groups (Figure 6.1). These factors, including data quality, are considered in much greater depth in chapter 4, which the reader is strongly advised to review before proceeding. In this chapter, these components are integrated within a suggested sequence of steps in planning a registry rather than factor by factor. There is some repetition and re-use of figures to refresh the reader’s memory and for ease of reading.



Governance
- procedures and methods for registry operation
- education and training
- resource planning and financial sustainability
- interoperability as a quality dimension
- self-assessment

Data quality
- data quality dimensions
- data standardisation

Information quality
- surveillance
- outcomes
- scientific publishing

Ethical issues, security and privacy
- adherence to privacy legislation
- ensuring data and information security
- ethical and privacy issues with secondary use of data

Figure 6.1. Quality dimensions of registries

6.1.2.1.1 Data and Information types

Data or information may be considered primary or secondary. Primary data, or information, refers to data collected to “provide health or social care to the data subject” (5). “Secondary use of information relates to information collected in the course of providing care, being used for purposes other than direct service-user care” (5). The use of data for secondary purposes, such as research, is governed by legislation, which varies across member states. As such it is advised that legal expertise is sought. It might be prudent to adopt a position that secondary use of data requires explicit consent from patients or full anonymisation, which should be performed in keeping with local data protection regulation.

6.1.2.1.2 Data Quality Dimensions

“The delivery of safe and effective healthcare depends on access to, and use of information that is accurate, valid, reliable, timely, relevant, legible and complete” (5). Data quality dimensions are reproduced in Table 6.1, from Chapter 4, so that registry planners might consider what dimensions are significant in the context of the purpose and objectives of the registry that they are planning. “Seven essentials for improving data quality” are also reproduced here so that registry planners might consider at a high-level, how these will be addressed by their registry (5, 6).

Table 6.1. Data quality dimensions

Data quality dimension	Description
Accuracy	How well information in or derived from the data reflects the reality it was designed to measure.[72] It is usually characterized in terms of error in statistical estimates. It may also be described in terms of the major sources of error that potentially cause inaccuracy (e.g., coverage, sampling, non-response, response).[73] ✓ How good are the data? ✓ What is done with the data?
Completeness	Extent to which all necessary data that could have registered have actually been registered.[74] It is usually described as a measure of the amount of available data from a data collection compared to the amount that was expected to be obtained[75] (e.g. coverage). ✓ Are all the appropriate data present?
Interpretability and Accessibility	Ease with which data may be understood and accessed.[76] This includes the ease with which the existence of information can be ascertained, the suitability of the form or medium through which the information can be accessed, whether data are accompanied with appropriate metadata and whether information on their quality is also available (including limitation in use etc.).[77] ✓ How readily accessible are the data? ✓ How well documented are the data? ✓ How easy is it to understand the data?
Relevance	The degree to which data meets the current and potential needs of users. The purpose is to assess how well a data collection can adapt to change and whether it is perceived to be valuable.[78] ✓ Can user needs be anticipated and planned for? ✓ How valuable are the data?

[72] Canadian Institute for Health Information, The CIHI Data Quality Framework, 2009 (Ottawa, Ont.: CIHI, 2009). Available at: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en

[73] Statistics Canada Quality Guidelines, Fourth Edition, October 2003. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2003001-eng.pdf

[74] Arts et al. Defining and Improving Data Quality in Medical Registries. J Am Med Inform Assoc. 2002;9:600-611

[75] Ehling, Manfred, Körner, Thomas. (eds.) Handbook on Data Quality Assessment Methods and Tools Eurostat, European Commission, Wiesbaden, 2007. Available at: http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf

[76] Canadian Institute for Health Information, The CIHI Data Quality Framework, 2009 (Ottawa, Ont.: CIHI, 2009). Available at: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en

[77] Statistics Canada Quality Guidelines, Fourth Edition, October 2003. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2003001-eng.pdf

[78] Canadian Institute for Health Information, The CIHI Data Quality Framework, 2009 (Ottawa, Ont.: CIHI, 2009). Available at: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en

Timeliness	Refers primarily to how current or up to date the data is at the time of release, by measuring the gap between the end of the reference period to which the data pertains and the date on which the data becomes available to users.[79] It is typically involved in a trade-off against accuracy. The timeliness of information will influence its relevance.[80] ✓ Are data made available in a reasonable amount of time? ✓ Are key documents released on time?
Coherence	Reflects the degree to which it can be successfully brought together with other statistical information within a broad analytic framework and over time. Coherence covers the internal consistency of a data collection as well as its comparability both over time and with other data sources.[81] The use of standard concepts, classifications and target populations promotes coherence, as does the use of common methodology across surveys. Coherence does not necessarily imply full numerical consistency.[82] ✓ Does the database use standard definitions for data definitions? ✓ Can common groupings be derived from the data? ✓ Can databases be joined via a common data element? ✓ Are data values being converted correctly? ✓ Are data comparable with themselves over time?

[79] Canadian Institute for Health Information, The CIHI Data Quality Framework, 2009 (Ottawa, Ont.: CIHI, 2009). Available at: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en

[80] Statistics Canada Quality Guidelines, Fourth Edition, October 2003. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2003001-eng.pdf

[81] Australian Bureau of Statistics. ABS Data Quality Framework [Online]. Available at: http://www.abs.gov.au/AusStats/ABS@.nsf/Latestproducts/5AFFD020BC4D1130CA25734700151AA5?opendocument

[82] Statistics Canada Quality Guidelines, Fourth Edition, October 2003. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2003001-eng.pdf

Table 6.2. Seven essentials for improving data quality

Leadership & Management	• What: involves having in place executive-level responsibility, accountability and leadership. • Why: knowing who does what (e.g. the establishment of a governance committee that will ensure the registry is committed to data quality). Decision-wise, this includes the selection of only essential data elements when datasets are established.
Policies and procedures	• What: developing and implementing clear policies and procedures on data quality for staff that are based on legislation and standards. • Why: can help ensure that a high level focus on data quality is translated into good practice amongst all those involved in data collection and handling within the registry.
Standardisation	• What: ensuring that data is collected and processed in a standardised fashion (e.g. use of minimal datasets, data dictionaries and the creation of standard templates for data collection), designing the registry with respect to national and international standards. • Why: facilitates data interoperability and making data available. Also can improve consistency and reduce error.
Data quality dimensions	• What: set of data quality attributes upon which data can be assessed, aligned with policies, procedures and training. • Why: measuring and monitoring level of data quality within a registry.
Training	• What: training of the staff in the requirements and importance of data quality. • Why: ensuring that policies and procedures adopted to generate high quality data are implemented and understood in practice.
Data quality audits	• What: independent systematic examination of data (internal or external). • Why: providing feedback to all staff, indicating the areas for improvement, highlighting good practice in order to facilitate learning (e.g. automation of data collection over manual collection where possible will reduce error rate, however, this will not be verified without planned audits of data quality).
Make data available	• What: availability of data when and where needed, in accordance with information governance safeguards (security, privacy). • Why: fulfilling the purpose for which the registry was created, increasing quality of registry data through its efficient utilization and dissemination.

6.1.2.1.3 Method of data capture

The quality of data will be significantly affected by the manner in which data is collected. Data collection can be considered with respect to two major domains; data source and data provider.

- Data sources
  - Paper-based
    - Questionnaire
    - Paper health record review
    - Documentation review
    - Laboratory reports
    - Other
  - Electronic
    - Questionnaire
    - Electronic Health Record
    - Laboratory reports
    - Databases
    - Mobile applications
    - Health devices
    - Social media
    - Other
- Data provider
  - Clinical units
  - Laboratories/central services
  - Discharge registries
  - Patients and families
  - Patients user groups (associations/federations)
  - Disability registries
  - Centres of expertise
  - Birth registries
  - Cause of death registries
  - Insurance funds (public and private)
  - Other registries
  - Other

6.1.2.1.3.1 Paper-based methods

Paper-based records have the advantage of being relatively inexpensive to create and distribute. However, in an era where health is becoming progressively more connected, paper is potentially very restrictive and does not take advantage of many error avoidance techniques that electronic methods offer. It is also worth considering that at some point, the data will need to be collated electronically to facilitate analysis. Paper can still play a core role in registry design however. Questionnaires and process flows can be created using paper which can be far more accessible for primary stakeholders. Once a prototype has been created using this method, it can facilitate the development of an electronic solution.

6.1.2.1.3.2 Electronic-Based methods

Though the design of bespoke electronic solutions can be expensive, their advantage is that of connectivity, error minimization and reduction of duplication. While Electronic Health Records are still in evolution they are certainly not ubiquitous and they still have significant difficulties associated with their use. When they are available, and adhere to appropriate interoperability and terminology standards, they can offer an exceptional source of data for a registry. The list of potential electronic sources is large and it is for this reason that it is highly recommended that Registry designers make considerable effort to liaise with national, and possibly international, health, information and registry bodies to maximize resource utilisation.

6.1.2.1.3.3 Future developments

The recent explosion in mobile Health (mHealth) warrants consideration. As noted previously, there is an ever-increasing facility to utilize technology to connect data that has previously been unimaginable. Similarly, social media has established an almost ubiquitous presence and the extensive data networks that have resulted are of enormous potential to registries. It may be advisable to contact universities and connected health centres to consider what initiatives and ecosystems a registry could form part of to maximize the potential of mHealth and social media.

6.1.3 Overview of the Current State and the Importance of Interoperability

6.1.3.1 Overview of Current State

Having an appreciation for the organizational structure of registries and other healthcare information networks or ecosystems nationally and internationally is of vital importance in ensuring that a registry is best placed to make use of existing resources. This knowledge will also help orientate how a registry's role can best be positioned to fit into "the bigger picture" and contribute to the direction of health policy. There may be relevant data sources that could be integrated within your registry or vice versa. It may be that your proposal has previously been assessed, but was determined not to be feasible. Furthermore, significant resources might be spared through the identification of existing solutions the proposed registry might otherwise have replicated.

6.1.3.2 Interoperability

Interoperability is the means of ensuring that a registry will be able to integrate within "the bigger picture". Interoperability is defined by the Institute of Electrical and Electronics Engineers (IEEE) as the "ability of a system or a product to work with other systems or products without special effort on the part of the customer" (7). Interoperability is core component of good communication and as a result, effectiveness and safety.

Meta-analysis has demonstrated the importance of good communication within healthcare scenarios, suggesting that "interventions to improve the quality of information exchange increases effectiveness" (8). In addition, the value of improving information transfer has been noted by major organizations, such as the Agency for Healthcare Research and Quality (AHRQ) in the United States, as an "important patient safety practice" (9). Another US organization, the Institute of Medicine, having identified the extent of the risk posed by medical error in the landmark paper "To Err is Human", have suggested the development of improved communication systems as core components of modern healthcare systems (10, 11).

6.1.3.3 Planning for integration

As "interoperability is made possible by the implementation of standards", liaising with national regulation/quality improvement authorities, which can be a valuable source of advice regarding access to and appropriate use of relevant standards, is essential (7).

Of particular relevance from a registry development perspective, is the selection of standard datasets and terminology to facilitate local and cross-border interoperability. For general areas, such as demographics, PARENT is an excellent source of guidance with respect to standard datasets and terminology or facilitate contact with a registry in another state with a structure and composition that can be adapted or adopted for a new registry's needs. At a national level, regulatory bodies will typically be able to advise best use of classification systems such as the World Health Organisation's (WHO) International Classification of Diseases (ICD) or terminologies such as the International Health Terminology Standards Development Organisation's (IHTSDO) Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT®). For more specific areas, national or international professional clinical groups can be a rich source of information. It should only be a last resort that non-standardised terminology/datasets are developed by a registry team and this should only be considered after discussion with appropriate experts/standards bodies to advise on how the dataset is best developed.

An ultimate end point of achieving interoperability is to prevent potentially valuable data becoming trapped in information "silos" and facilitate more accurate representation of concepts and comparison of data across international borders.

The same connections that will facilitate interoperability are likely to be able to provide information regarding the current state of the art in registry design. In addition, we suggest contacting health authorities that are likely to have useful guidance regarding supportive ecosystems, including other registry groups. They may also be able to provide a clear picture of current and planned national health and information strategies and existing projects that could provide data sources for the registry, such as an electronic health record, or a national data architecture. PARENT will be able to offer further registry establishment advice, tools for registry development.

Connected health is a developing concept which "encompasses terms such as wireless, digital, electronic, mobile, and tele-health and refers to a conceptual model for health management where devices, services or interventions are designed around the patient's needs, and health related data is shared, in such a way that the patient can receive care in the most proactive and efficient manner possible. All stakeholders in the process are 'connected' by means of timely sharing and presentation of accurate and pertinent information regarding patient status through smarter use of data, devices, communication platforms and people"(12). As connected health democratizes health information, there is significant potential for a registry to increase:

- Awareness and Interest
- Dissemination and impact of outputs
- Collaboration opportunities
- The volume and variety of data sources available
- Resource sharing

It is therefore worth liaising with centres promoting connected health, such as universities, or not-for-profit groups such as the ECHAlliance to establish how a registry might integrate in the process (13). Conversely, the considerable organization required to develop a registry may facilitate the development of an ecosystem that can drive connected health.

6.1.4 Considering Legal Aspects and Confidentiality

While there are many important aspects to planning a registry, ensuring compliance with data protection regulations are not only vital, but a legal requirement; breach of which may result in termination of the registry project. Furthermore, adopting a gold standard, transparent data protection practice is likely to increase the confidence registry participants will place in your registry and add to its value. As such it is essential to prioritise contacting the relevant national Data Protection Commissioner's Office early in the design of the registry. Contact details for EU member Data Protection Commissioners are outlined in Table 6.3. More information about the legal aspect is available in chapter 5.

EU Member State	Data Protection Authority	email
Austria	Österreichische Datenschutzbehörde	dsb@dsb.gv.at
Belgium	Commission de la protection de la vie privée	commission@privacycommission.be
Bulgaria	Commission for Personal Data Protection	kzld@cpdp.bg, kzld@cpdp.bg
Croatia	Croatian Personal Data Protection Agency	azop@azop.hr, info@azop.hr
Cyprus	Commissioner for Personal Data Protection	commissioner@dataprotection.gov.cy
Czech Republic	The Office for Personal Data Protection	posta@uoou.cz
Denmark	Datatilsynet	dt@datatilsynet.dk
Estonia	Estonian Data Protection Inspectorate	viljar.peep@aki.ee
Finland	Office of the Data Protection	tietosuoja@om.fi
France	Commission Nationale de l'Informatique et des Libertés	
Germany	Der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit	poststelle@bfdi.bund.de
Greece	Hellenic Data Protection Authority	contact@dpa.gr
Hungary	Data Protection Commissioner of Hungary	peterfalvi.attila@naih.hu
Ireland	Data Protection Commissioner	info@dataprotection.ie
Italy	Garante per la protezione dei dati personali	garante@garanteprivacy.it
Latvia	Data State Inspectorate	info@dvi.gov.lv
Lithuania	State Data Protection	ada@ada.lt
Luxembourg	Commission nationale pour la protection des données	info@cnpd.lu
Malta	Office of the Data Protection Commissioner	commissioner.dataprotection@gov.mt
Netherlands	College bescherming persoonsgegevens; Dutch Data Protection Authority	info@cbpweb.nl
Poland	The Bureau of the Inspector General for the Protection of Personal Data	sekretariat@giodo.gov.pl
Portugal	Comissão Nacional de Protecção de Dados	geral@cnpd.pt
Romania	The National Supervisory Authority for Personal Data Processing	anspdcp@dataprotection.ro
Slovakia	Office for Personal Data Protection of the Slovak Republic	statny.dozor@pdp.gov.sk
Slovenia	Information Commissioner	gp.ip@ip-rs.si
Spain	Agencia de Protección de Datos	internacional@agpd.es
Sweden	Datainspektionen	datainspektionen@datainspektionen.se
United Kingdom	Information Commissioner's Office	casework@ico.org.uk
	Information Commissioner's Office	casework@ico.org.uk
	Information Commissioner's Office	ni@ico.org.uk
	Information Commissioner's Office	Scotland@ico.org.uk
	Information Commissioner's Office	wales@ico.org.uk
EU	Data Protection Officer of the EU	DATA-PROTECTION-OFFICER@ec.europa.eu

Table 6.3: Data protection authorities and contact details for EU member states

6.1.4.1 Privacy and Privacy Impact Assessments

"Privacy is the right of individuals to keep information about themselves from being disclosed" (14, 15). A privacy impact assessment (PIA) is a process that "facilitates the protection and enhancement of the privacy of individuals" and is best conducted at a planning stage to protect the registry and its participants from potentially irreconcilable personal and organisational breaches that may be damaging at a later stage (15). This will facilitate the identification of risks to privacy breaches and examination of how these risks can be allayed. Detailing the process involved in a PIA is beyond the scope of this chapter, however the Health Information and Quality Authority of Ireland provide an excellent range of resources in this area, including a review of international PIAs, a tool to establish whether a PIA is required, details regarding how it should be conducted and a sample report (15-17).

6.1.4.2 Data Protection Policy

Even following a PIA, it is advisable to develop a data protection policy for the registry project and ensure that all involved with design and implementation of the registry are appropriately trained in this regard and regularly made aware of their responsibilities. A local Data Protection Commission Office or health authority may provide links with groups who have a policy that can be adapted for the purpose of the registry.

6.1.4.3 Data Ownership, Access and Intellectual Property

While considering data security it is prudent to consider data ownership, access and intellectual property. This is likely to require dedicated expert guidance and, to ensure transparency; it is advised that the outcome of this process is formalized in a policy document. This document should also consider the scenario in which the registry project is terminated so that it is clear how the data might best be protected.

6.1.5 Eliciting Expert Opinion & Generating an Advisory Board

Expert elicitation refers to the "solicited exchange of knowledge, information, or opinion from an expert" (18). If the initial planning processes suggest that there is a valid opportunity to establish a registry, further planning can be greatly facilitated by expert guidance. We suggest the establishment of an Advisory Board consisting of a knowledgeable panel with expertise relevant to the registry domain and those who are committed to the establishment of the registry. This will not only facilitate the implementation of best practice, it will also help identify stakeholders who might not be immediately apparent to the group establishing the registry. Finally, the selection of appropriate representatives for an advisory board is likely to increase the engagement of potential stakeholders with the project by virtue of their involvement, which can be vital to the success of the project.

6.1.6 Defining the Scope of the Registry & Building a Registry Development Team

It is advisable, at this point, to consider with the advisory board and funders what the scope of the registry will be. Though this may seem obvious once the purpose(s), objectives and outputs have been defined, these may be challenged by the open nature of stakeholder involvement and there is a significant risk of losing focus if clear limitations are not imposed. In addition, though an open stakeholder engagement process is likely to engage stakeholders' imaginations and promote innovative ideas and engagement with the project, false promises can lead to significant disappointment at a later stage in the project.

The scope should aim to highlight the value of achieving the purpose(s), objectives and outputs of the registry with the minimal complexity possible, and in a manner that is most likely to be successfully accepted by users. Financial resources should be defined and a rough timeframe be agreed to give invited stakeholders an opportunity to plan when they can engage.

As there will be considerable time and preparation involved in developing the registry from this point, it is advised that a project development team is established that is proportional in size to the level of resources available to the registry. This might be a registry champion or person with an interest in the area in question, but thought might also be given to involving a research fellow with an interest in registries, healthcare informatics or the area targeted by the registry. Ideally, this person would be a primary stakeholder with a long-term interest in the area the registry is focused on. This will facilitate development of skills that can improve the long-term success of the registry, while also ensuring that the registry is designed in a fashion cognizant of end-users requirements.

6.1.7 Performing Stakeholder Engagement and Analysis

The Health Information and Quality Authority of Ireland (HIQA) have produced a document entitled "Guidelines for Stakeholder Engagement in Health Technology Assessment" which provides a comprehensive overview of stakeholder assessment that is extremely relevant to registry planning (18).

HIQA note that "Stakeholder engagement is an iterative process of actively soliciting the knowledge, experience, judgment and values of individuals selected to represent a broad range of direct interests in a particular issue"(18, 19). Though stakeholder analysis might also involve the experts identified in 4.1.3, the aim in stakeholder analysis is to avoid solicited advice and instead facilitate wider engagement on the topic (18).

The process of stakeholder engagement should also be seen as an inclusive "hearts and minds" campaign. An effort to be inclusive and respectful of all stakeholders' contributions can significantly improve the registry's later adoption and success.

6.1.7.1 Identification of Stakeholders

Though the definition of what constitutes a stakeholder varies, for the purpose of a registry, two subtypes can be considered (3):

- Primary stakeholders are intrinsically involved in the design and funding of the registry, but may also include parties with a regulatory capacity.
- Secondary stakeholders may be affected by and involved in using and operating the registry, but do not have direct involvement in its design.

6.1.7.2 Engagement

As the stakeholders of a registry may be extremely diverse, it is recommended that a flexible approach is adopted to engagement. None-the-less, to facilitate transparency, consistency and relevance it is advised that a standard information document is prepared and distributed in advance, where feasible. Ideally, this document would support the conduct of a semi-structured interview.

6.1.7.3 Recording Stakeholder (& expert participation)

Even with a focused registry, the number of potential stakeholders and registry contacts can increase significantly beyond the expected scope. As such we would advocate using a tool to monitor involvement at a high level. Table 6.4 is a re-usable means of collecting information about possible registry stakeholders and recording high-level outcomes from meeting with them, relevant for the purpose of designing the registry.

High-level categories of contacts include:

- Clinical groups
- Public health and Regulatory groups
- Product and Device Manufacturers
- Health Care Service Providers
- Health Funding and Insurance groups
- Patient and advocacy groups
- Academia
- Relevant Experts
- Professional groups and societies
- Registry groups
- Registry Sponsor groups
- Development groups (Informatics and Management)
- Other International Groups

6.1.7.4 Content of the Stakeholder Evaluation

Though the content of the evaluation will vary greatly depending on the nature of the registry and its stakeholders, we suggest a process that has been adapted from Registries for Evaluation Patient Outcomes: A User's Guide" (3). We recommend providing the stakeholder with a document, striking a balance between delivering information and being concise, that consists of the following components:

1. Introduction to the group designing the registry, the current state and the motivation for developing a registry
2. A brief introduction to Registries
3. The purpose of the document
4. Engagement requesting input from stakeholders regarding
   a. Purpose & Objectives
   b. Available relevant information sources
   c. Key stakeholders
   d. Feasibility – barriers and motivators to establishing a registry
   e. Registry team membership
5. Description of the further steps required to establish the registry

Though direct feedback is likely to be limited given the time constraints of busy stakeholders, this process is likely to not only be an exercise in clarifying the registry design process, but can create a template for a semi-structured interview with stakeholders at a later stage and also helps develop awareness of the project and confidence in the design process.

6.1.7.5 Stakeholder Evaluation Output

Given the diverse nature of stakeholders it is difficult to ensure consistency and as such a scientific document is unlikely to be produced. None-the-less, by following the process described above, the opportunity has been presented for frank and honest engagement and useful information and requirements will be made apparent.

The analysis of the stakeholder evaluation also provides a useful opportunity to personalize further interactions with stakeholders and to provide relevant information at conferences or with other stakeholders to increase awareness and further engagement.

Table 6.4: Registry contact record

Stakeholder type	Super-Group/Region	Group	Suff	Firstname	Surname	Professional Role	Contact details	Relevance to Registry	Involved	Recommendations	Action
Clinician Groups											
	Public Consultants	Hospital 1									
		Hospital 2									
	Private Sector consultants	Hospital 3									
	Trainees										
	Primary Care physicians										
	Nursing Staff										
	Administrative Staff										
Public health/Regulatory											
	National	Health Information and Quality Authority									
		Data Protection Commissioner									
	International – EU	PARENT		Matic	Meglič	Co-ordinator PARENT EU joint action group and Manager at National Institute of Public Health Slovenia					
Product Manufacturers											
	National										
	International										
Health Care Service Providers											
	Health Service provider	National Pharmacoeconomic Centre									
		National Health Intelligence Unit									
		Information and Communication Technology									
		National Management									
		Clinical Governance									
		Medicines Board									
	Department of Health	Legal									
		Strategy									
	Hospital 1	Audit Manager									
Health Funding/Insurance groups											
	National										
Patient/advocacy groups											
	National										
	International										
Academia											
	National	University 1									
	International	University 2									
Relevant Experts											
	National	Database/Data									
		Epidemiology									
	International	Health Economist									
		Registry									
Professional groups and societies											
	National	Royal College of Physicians									
Registry groups											
	National	National Cancer Registry									
	International – EU	EAR (European Arthroplasty Register – Austria)									
		EUBIROD (European Best Information through Regional Outcomes in Diabetes)									
	International – non-EU										
Registry Sponsor Groups											
	National										
	International										
Development Groups											
	National										
	International – EU										
	International – non-EU										
Other International Groups											
		SNOMED-CT – IHTSDO									
		ICD-11 working group									
		International Organization for Standardization									
		Ecosystem support groups									
		EU Data Protection									

6.1.8 Re-defining the Scope of the Registry

Following stakeholder assessment it is advisable to reconsider the scope of the project. While factors likely to improve stakeholder engagement and ultimately increase the chance of the registry's success are important, these should be weighed against the considerable expense the extra scope is likely to add.

A final scoping document will facilitate the creation of a business case and will better inform selection of data elements of the registry and the registry data model.

From this point, changes to the scope may result in significant resource utilisation and, as such, a change management strategy should be created which outlines how further adaptations to the scope should occur in the future.

6.1.9 Governance, Oversight and Registry Teams

Before considering data elements for the registry and beginning to focus on the practical implementation of the registry, it is advisable to establish a governance plan and to develop teams that will facilitate design of the registry and maintenance following implementation.

This serves a number of purposes:

1. Creating teams can involve end-users, increasing buy-in.
2. Facilitating a better understanding of how the registry will operate and how intellectual property will be handled.
3. Creating the governance framework for data sharing and dissemination of data or information created by the registry.
4. Ensuring oversight and that the registry development is progressing as planned.

Particularly when the scope of the registry is small, there may need to be overlap, however, at a minimum, we suggest prioritisation of a project management team, scientific committee and a quality assurance committee. It is suggested that, though specific teams/committees will benefit from members with specific skill sets, that members be selected to ensure that all stakeholders groups are adequately represented. In particular, it is advised that patient groups should be asked to contribute to ensure that the patient’s voice is represented appropriately as their data is the subject of the registry process.

6.1.9.1 Project management team

The involvement of a person skilled and experienced in project management is advised. If this is not possible, it would be worthwhile considering training for a project manager and consideration given to the use of project management software. Table x outlines a tool to facilitate registry team organisation and selection.

6.1.9.2 Scientific Committee

The aim of the scientific committee should be to ensure that the registry is outcomes driven and that data collected is disseminated effectively. It is suggested that the committee aim to meet 4 main objectives:

- Question identification
- Data element identification and selection
- Dissemination of results
- External data access/study proposal adjudication

As such, this group should consist of subject matter experts, ideally with a track record in publication of scientific results. It would also be ideal to include members of the group with statistical/epidemiological and health outcomes analyses experience, so that these factors remain in focus throughout the design, implementation and life of the registry.

6.1.9.2.1 Question identification

Based on the scope identified by the advisory board and the input of the stakeholder evaluation, the committee should identify specific questions that the registry will address. These questions will inform the selection of data fields that the Registry will record.

6.1.9.2.2 Data element identification & selection

It is suggested that this process be considered an iterative one that considers the dimensions of data quality discussed previously.

6.1.9.2.2.1 Rough selection

In the first instance, it is advised that the Scientific committee consider a rough map of possible data fields. This should then be submitted for statistical analysis based on the scientific questions that have been proposed.

6.1.9.2.2.2 Statistical and Epidemiological analysis

This process is vital to ensure that the registry is developed to an appropriate scale that ensures the purpose and objectives it was created for are met.

Extra data fields add considerable complexity and cost because of data validation requirements. A statistical analysis can help highlight the essential fields for registry success and help maintain as much simplicity as possible; reducing the resources required ensuring completeness of data entry when the registry is implemented. It will also reduce the effort required to validate and analyse data.

It is advisable that this process is conducted by statisticians and epidemiologists trained in registry science. If the registry development group has no formal attachment with experts with skills in this area, it is worth checking with universities or other registry groups, who might identify relevant experts.

6.1.9.2.2.3 Health Outcomes/Pharmacoeconomics analysis

At the same time as a statistical analysis review of potential registry outcomes from a health outcomes and pharmacoeconomics perspective should be considered.

Increasingly, the relevance of real-world effectiveness is being prioritised and the relevance and attractiveness of a registry can be greatly increased by engraining it within national and international strategies. This is also a mechanism of scientifically establishing the potential economic worth of the registry and as a means of creating a benchmark against which the registry might later be evaluated as a marker of success. This can be of particular consequence when funding organisations are approached with a view to ensuring the long-term feasibility of the registry project.

6.1.9.2.2.4 Final Selection of elements

The final selection of data elements is only likely to occur at the time of implementation of the registry, or ideally, after a pilot project has been conducted and after a financial analysis has identified the scope that can be realistically be supported. The aim of the data selection process at the planning stage should therefore be to outline the data fields that will be required to a level adequate to conduct a feasibility study.

6.1.9.2.3 Dissemination of Results

Dissemination of registry data increases the potential impact of a registry and facilitates peer review. This process enables registry methods and data to be independently scrutinized, which in turn can validate the quality of the registry. Planning how registry data will be disseminated can help develop a timeline for implementation as well as ensuring that adequate funding is considered for this purpose.

6.1.9.2.4 External data access/study proposal adjudication

If a registry collects high-quality data, it is both likely and desirable that external requests will be received requesting access to data or proposing studies that can utilize registry data. To ensure transparency and facilitate best use of data, it is suggested that the scientific committee establish a formal plan to adjudicate on such requests. This might involve defining the grounds for collaborative agreements where external parties, in addition to gaining access to data, can benefit from the experience and expertise of committee members aware of the context in which the data was collected.

6.1.9.3 Quality assurance Committee

Ensuring that the registry's quality is validated will increase the value of the registry. Though the project management team and scientific committee will together increase a registry's quality, it is advisable to have an independent committee established to assess whether this is the case through the creation of a formal audit and quality assurance plan. In addition, this group might be well placed to handle complaints or to ensure that ethical and legal obligations are being met in the absence of a specific group to manage this.

This group would ideally comprise of experts familiar with registry analysis and who have experience of audit and quality assurance. There should also be consultation with regulatory groups to ensure that all regulatory requirements are met; this is of particular relevance when the registry is focused on safety assessment, such as devices.

Table 6.5: Re-usable table/tool to facilitate selection and recording of possible Registry team members.

Registry Teams	**Technical expertise required**	**Name**	**Group**	**Comments**
Project Management				
To oversee the management of the overall project				
Clinical/Subject Matter Board				
To determine the scope of the data captured	Subject Expert			
	Clinical Champion			
Scientific Committee				
To guide scientific utilisation of registry data assess external applications for utilisation of data	Health Outcomes			
	Epidemiology			
	Statistics			
	Data mining			
	Data standards			
	Social Media			
	eCommerce			
Data Collection & Database Management Board				
To guide with respect to attainment of best data standards	Data standards			
	Data linkage			
	Data quality			
	Databases			
	Data security			
	Data mining			
	Clinical Standards Manager			
Legal/patient privacy				
To ensure compliance with legal requirements	Health Information Act		Health Providers	
	Health law		Regulatory Bodies	
	Project management lead		Data Commissioner	
Quality Assurance & Liaison				
To ensure the quality of the registry is maintained	Data quality expert		Regulatory Bodies	
	Epidemiology			
	Patient representation			
	Health Outcomes			
Other Comments				

6.1.10 Resource requirements

Resource requirements will vary significantly depending on the scope of the registry project. The steps followed to this point should identify the extent of resourcing that will be required to meet the objectives outlined. Resources to consider include:

6.1.10.1 Human Resources

Though registry committees can perform a large quantity of work, there is likely to be either dedicated or intermittent need for staff to ensure proper set-up and maintenance of a high-quality registry. Depending on the scope of the registry project, this might include staff to meet needs in:

- Administration
- Project management
- Data management
- Data collection
- Study design, epidemiology & statistical support
- Data dissemination
- Programming
- Design (question and graphic)
- Training
- Financial
- Legal/data security & protection
- Clinical

6.1.10.2 Information Technology Resources

Depending on the environment in which the registry is to be established, requirements can range from analysis software to an extensive hardware and software budget. It should be stressed that information technology support with experience of registry design is extremely valuable. Gaining advice from other registries, registry groups such as PARENT and local regulatory bodies is invaluable and should be sought to ensure any system delivered is designed appropriately and with interoperability in mind.

6.1.10.3 Financial Resources

Though the outlay for the initial design and implementation of a registry is the most obvious requirement, consideration should be given to the long-term sustainability of the registry project. Financial resources will vary significantly depending on the scope of the registry; however, by following a planning process with an inclusive stakeholder assessment, it is more likely to identify appropriate funding avenues and collaborations that may maximise financial investments in addition to the financial value of registry outputs. Examples of funding avenues include public-private partnerships, governmental funding, patient-groups, and sponsorship from charities or pharmaceutical companies.

Finally, it is necessary to take account of the financial implications of closing down a registry and what arrangements would need to be made to ensure that data security is maintained in this scenario.

6.1.10.4 Other Resources

The list of potential other resources is extensive, however, particular note should be drawn to office space. It is ideal if the registry emerges from a group that can provide accommodation. If this is not possible, even from a data security perspective, it will need to be considered.

6.1.11 Funding Strategy

It is likely that for each group who propose a registry, there is a funding source that has helped bring the idea to this stage. By including a directed stakeholder evaluation, it is likely that further opportunities might present. Of particular significance, however, is the need to consider how funding might influence how the outputs of the registry are interpreted. At all times, funding should be arranged in a manner that is transparent and without conditions that might undermine the validity of the scientific study.

6.1.12 Risks and feasibility

Risks accompany each component of the registry establishment and maintenance process, from excessive dataset selection and lack of adherence to recognized standards through to a failure to consider a registry termination strategy. Of all these, however, we suggest that a failure to be aware of the extensive preparations required to develop and maintain a registry are the most significant. The process described in this section may seem overburdensome, but it can present a myriad of advantages, such as identifying collaborative opportunities and identifying funding opportunities. Apart from this, as registry science evolves, regulation is likely to follow and create obstacles that might threaten the creation and survival of registries that have already consumed significant resources. As such, we recommend that an extensive planning process be undertaken under the guidance of experts familiar with the process of registry design and with stakeholders. Once this has been completed, an informed feasibility assessment can be undertaken. This should review whether the objectives and purpose of the registry are likely to be met within the timeline considered, the budget available, the scientific model proposed and within the environment the proposed registry is due to be implemented in.

6.1.13 Developing an Implementation Plan

If the feasibility study reaches a positive conclusion, it is likely that most of the components will now be adequately developed to create an implementation plan. It is suggested that a further review of the steps involved in planning the registry is undertaken to develop an action plan and timeframe for each step in conjunction with the appropriate expert or stakeholder identified by the planning process. Within this, rate-limiting steps should be identified to help determine the “critical path” which will dictate how long the project is likely to take. It is suggested that at this point, particularly in the case of a large registry project, an experienced project manager is involved to help deliver the project on schedule.

As part of the implementation plan, it may be useful to consider a pilot project as a proof-of-concept model before proceeding with a full implementation. This can generate significant support for a registry, create useful outcomes and identify significant obstacles that may not have been initially obvious. It can also create a wealth of knowledge and experience at a manageable level that can increase the chances of ultimate success.

A project proposal should be formalized with firm time and budgetary constraints outlined to facilitate regular oversight by the project management committee (or similar). Though numerous measures of quality have been mentioned, ultimately, the registry will need to be regularly evaluated against the objectives and purpose it was designed to meet. This can facilitate review and adjustment of the registry that can further improve outcomes, efficiency & relevance is maintained.

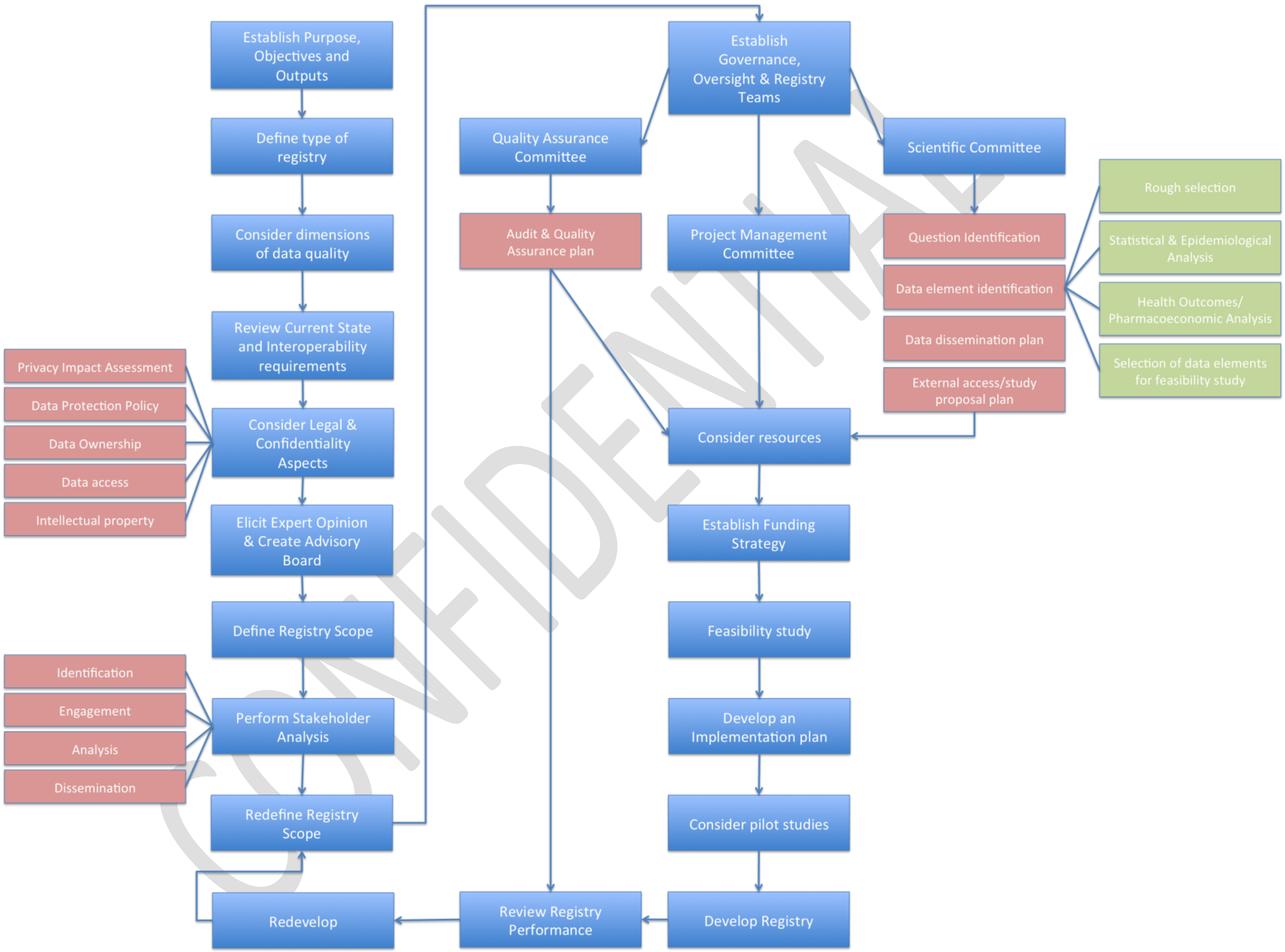

Establish Purpose, Objectives and Outputs
Define type of registry
Consider dimensions of data quality
Review Current State and Interoperability requirements
Consider Legal & Confidentiality Aspects
Privacy Impact Assessment
Data Protection Policy
Data Ownership
Data access
Intellectual property
Elicit Expert Opinion & Create Advisory Board
Define Registry Scope
Perform Stakeholder Analysis
Identification
Engagement
Analysis
Dissemination
Redefine Registry Scope
Redevelop
Review Registry Performance
Develop Registry
Establish Governance, Oversight & Registry Teams
Quality Assurance Committee
Audit & Quality Assurance plan
Project Management Committee
Scientific Committee
Question Identification
Data element identification
Data dissemination plan
External access/study proposal plan
Rough selection
Statistical & Epidemiological Analysis
Health Outcomes/ Pharmacoeconomic Analysis
Selection of data elements for feasibility study
Consider resources
Establish Funding Strategy
Feasibility study
Develop an Implementation plan
Consider pilot studies

Figure 6.2: Planning a Registry Process

References

1. PARENT (PAtient REgistries iNiTiative). PARENT Pilot Registry of Registries 2014 [24th May 2014]. Available from: http://www.parent-ror.eu/ - !state/list_all.

2. Agency for Healthcare Research and Quality. RoPR | Registry of Patient Registries: U.S. Department of Health & Human Services; 2014 [24th May 2014]. Available from: https://patientregistry.ahrq.gov/.

3. Gliklich R, Dreyer N, Leavy M, eds. Registries for Evaluating Patient Outcomes: A User's Guide. Third edition. Two volumes. (Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc., a Quintiles company] under Contract No. 290 2005 00351 TO7.) AHRQ Publication No. 13(14)-EHC111. Rockville, MD: Agency for Healthcare Research and Quality. April 2014. http://www.effectivehealthcare.ahrq.gov/registries-guide-3.cfm.

4. ISPOR Taxonomy of Patient Registries: Classification, Characteristics and Terms. Lawrenceville, NJ2013.

5. Health Information and Quality Authority. Guiding Principles for National Health and Social Care Data Collections. Dublin, Ireland: HIQA; 2013.

6. Health Information and Quality Authority. What you should know about Data Quality. Dublin, Ireland: HIQA, 2012.

7. IEEE Standards Glossary: IEEE; [22 May 2014]. Available from: http://www.ieee.org/education_careers/education/standards/standards_glossary.html.

8. Foy R, Hempel S, Rubenstein L, Suttorp M, Seelig M, Shanman R, et al. Meta-analysis: effect of interactive communication between collaborating primary care physicians and specialists. Annals of Internal Medicine. 2010;152(4):247-58.

9. Shojania KG, Duncan BW, McDonald KM, Wachter RM, Markowitz AJ. Making health care safer: a critical analysis of patient safety practices: Agency for Healthcare Research and Quality Rockville, MD; 2001.

10. Institute of Medicine. Crossing the quality chasm: A new health system for the 21st century. Washington: National Academies Press, 2001 0309072808.

11. Levit L, Balogh E, Nass S, Ganz PA. Delivering high-quality cancer care: charting a new course for a system in crisis. Institute of Medicine Washington, DC: Institute of Medicine. 2013.

12. Caulfield BM, Donnelly SC. What is Connected Health and why will it change your practice? QJM. 2013;106(8):703-7.

13. ECHAlliance | Delivering leadership for the development of Connected and MHealth market and practice across Europe 2014 [24th May 2014]. Available from: http://www.echalliance.com/.

14. Erickson JI. Caring for patients while respecting their privacy: renewing our commitment. Online journal of issues in nursing. 2005;10(2).

15. Health Information and Quality Authority. Guidance on Privacy Impact Assessment in Health and Social Care. Dublin, Ireland: HIQA; 2010.

16. Health Information and Quality Authority. International Review of Privacy Impact Assessments. Dublin, Ireland2010.

17. Health Information and Quality Authority. Sample Privacy Impact Assessment Report Project. Dublin, Ireland2010.

18. Health Information and Quality Authority. Guidelines for Stakeholder Engagement in Health Technology Assessment in Ireland. Dublin, Ireland: HIQA; 2014.

19. Deverka PA, Lavallee DC, Desai PJ, Esmail LC, Ramsey SD, Veenstra DL, et al. Stakeholder participation in comparative effectiveness research: defining a framework for effective engagement. Journal of comparative effectiveness research. 2012;1(2):181-94.

20. Irish Platform for Patients' Organisations, Science & Industry [07/12/2013]. Available from: http://www.ipposi.ie/.

21. Irish Platform for Patient Organisations, Science and Industry,. Towards a National Strategy for Patient Registries in Ireland. Dublin, Ireland: 2008.

22. Donohue F. Presentation IPPOSI Meeting Dublin 10 May. Dublin2011.

6.2 Registry design

When the purpose and main objectives of the registry are defined, the next step is to define the data to be collected, and determine the methodology/protocol with which the registry will try to achieve the defined goals. At this point, registry holder needs to consider many issues, including the defining of the registry target population, anticipated registry size and duration, study design, data sources for the registry, registry dataset and data collection methods/procedure. At the same time, the registry holder needs to look at the registry resources, costs and consider the quality aspect. This chapter describes those registry's elements and covers the important aspects that are necessary to take into account during that development stage.

6.2.1 The population covered by a registry

Enrolment of the patients for a registry starts with clear understanding of the target population, which is a population to which the registry would like to generalize its results and findings (e.g. patients with multiple sclerosis in Slovenia). When building a registry it is important to accurately define target population since it is a key factor in forming the registry population. Registry holder should understand and determine whether the registry is hospital-based registry[83], population-based registry or even population registry[84]. It is necessary to define the registry in terms of geographical and organisational coverage.

In addition to target population, it is recommended that a registry provides a case definition which is a detailed specification of the patient/cases that are going to be included in a registry. Registry team should specify so-called eligibility or inclusion criteria that are set of conditions that patient must meet to be eligible for inclusion in a registry, and generally include geographic (e.g. hospitals in a particular region of the country), demographic (e.g. age, gender, ethnicity), disease-specific (e.g. certain diagnosis, stage of disease), time-specific (e.g. specification of the included dates of hospital admission), laboratory-specific, and other criteria (e.g. size of the hospital in terms of number of patients) (2, 35). Exclusion criteria, oppositely, are those criteria that disqualify subjects from inclusion in the registry. Inclusion and exclusion criteria often reflect considerations such as cost and practical constraints (sometimes subjects are not included, not because they are out of interest, but due to additional cost/burden of including them), ethical concerns, people's ability to participate (e.g. the health condition may prevent participation), and design consideration (it is sometimes advantageous to have more homogeneous population as a means for reducing confounding, but in terms of generalizability, stringent inclusion criteria might reduce the generalizability of the registry findings to the target population) (24). Inclusion and exclusion criteria should therefore be defined carefully and many aspects need to be taken into account while defining those criteria, as the selection of inclusion/exclusion criteria can optimize the internal validity or generalizability of the registry, improve its feasibility (also in terms of follow-up and attrition), and lower its costs (25). Besides very clear definitions of the inclusion and exclusion criteria it is crucial that criteria are well documented, including the rationale for these criteria.

6.2.2 Anticipated size and duration

[83] Registry that aims to record information on all patients seen in a given hospital or group of hospitals irrespective of geographical areas.

[84] Description of the population registry and population-based registry is provided in chapter 2.2 Types of patient registries.

Estimation of anticipated registry size is an important part of the planning process. Some registries try to include all cases from the defined population, but often registries include only a sample of population. In that case it is recommended to prematurely estimate how many cases registry is planning to include. If the registry is too small, it may have insufficient analytical power, and it may not ensure adequate exploration of the objectives. On the other hand, registry that is too large may waste time, resources and money. Hence, it is important to adequately plan the registry size. Various components impact on estimating registry size and need to be considered, including (2, 30):

- the study outcome and its frequency/variability
- size of clinically important effects, the desired precision of estimates (e.g. the width of a confidence interval);
- timeframe (e.g. for analyses, dissemination of results);
- available resources and money, feasibility;
- support for regulatory decision-making (e.g. if registry is intended to support regulatory decision-making, the precision of estimate is important);
- anticipated drop-out rate

Many methods for sample size calculation exist and are described in general statistics textbooks (31, 32, 33). There are also different tools that can assist in sample size calculation. Besides software programs (e.g. G*Power, nQuery Advisor, PASS, STATA) there are also online tools that allow free sample size calculations, such as:

- Russ Lenth's Power and Sample Size
- David Schoenfeld's Statistical Considerations for Clinical Trials and Scientific Experiments
- UCLA Calculator Service
- The Survey System's sample size calculator
- Raosoft's sample size calculator

These tools should be used with caution, since they are not always reliable or suitable for any situation.

Although a patient registry is generally considered as a long-term and sustainable action, the anticipated duration of a registry (taking into account the enrolment and follow-up phase) should also be specified when developing a registry. The duration of a registry depends on what type of registry it is, what are the specific procedures in the registry and what objectives need to be met. Some registries collect data at only one time point and others collect data for the lifetime of the patient. A registry may be open-ended or it may have fixed end point when enough data to achieve the registry's objectives is expected to have accrued (3). If we neglect the funding as the biggest factor for registry duration and sustainability, the factors that registry holder, together with the key stakeholders, should consider when estimating registry duration, includes the induction period for desired outcomes, sufficient follow-up time for the exposure, data collection method; sample size, complexity of data being collected, anticipated accrual of enrolled subjects, and deadlines for dissemination of results (2, 35).

It is worthwhile to note that registry size can also refer to the number of sites included in a registry, and to the volume and complexity of data being collected (3). Hence, registry holder can consider these perspectives as well.

6.2.3 Registry dataset

Registry needs to develop the dataset that will serve the purpose and objectives of the registry. Although some key variables/data elements can be identified and determined soon in the developing process, this can be a very lengthy activity and should not be underestimated since it is the registry dataset who will eventually determine the usefulness and success of the registry. More information on developing a registry dataset is available in chapter 6.3).

6.2.4 Data collection procedure/protocol

The decision on how the registry will collect the data is affected by several factors, namely the characteristics of the registry's target population, the information that needs to be obtained and other specific goals of data collection, available data sources, registry resources and time limits. The registry data collection procedure must support the highest possible data quality, lowest possible burden for the reporting units and lowest possible costs for the registry. Registry holder needs to identify and evaluate all available data sources and determine which one will be used. Registry must make the agreement with the data providers and develop the technical protocol for the data acquisition. Before the first data acquisition, the instructions, procedures and tools for the acquisition and for the implementation of the controls (e.g. readability of data, adequacy of records and their number) has to be prepared. (More information on data sources is available in the chapter 6.4 Data sources for registries.)

When developing a registry data collection procedure/protocol, the registry should take into account the technological aspect of data collection (e.g. paper-based forms, web-based data entry, use of personal computers, handheld computers, scanners, mobile phones) and be aware of advantages and disadvantages of both, paper-based and electronic approaches. The choice of which system to use depends on where the data are captured, by whom and what resources are available for particular reporting unit. It is important that that the approach is practical and reliable. In addition, a registry designer needs to look also from the perspective that is especially well-covered in the field of survey methodology, where a great emphasis is placed on the modes of data collection, their characteristics and principles of good practice. This includes, for example, the consideration whether the case report forms or questionnaires are understandable and easy to use, are questions or instructions worded correctly, are they measuring the right things, will the presence of the interviewer/data collector (e.g. nurse) influence the patient's answers, will the self-administered mode yield more honest answers or will produce lower response rate, can the telephone data collection be used to obtain the data more cost-effectively, will this mode enable to get the response from all patients etc.

Registry data collection can be transversal, where all defined patients are registered once, or longitudinal, where the data is collected at different time points for the same patient. In case of the longitudinal design, the registry should carefully determine (a) which data needs to be (re)collected, (b) at what time points (e.g. every 6 months), (c) how long (e.g. for 10 years) and (d) with what means (e.g. with the telephone, by visiting a personal doctor, by a data linkage to other records). When developing the follow-up strategy it is important to consider the costs which can increase significantly when the follow-up is implemented via personal contact, the extra work that will be put on the data providers and the burden that will be imposed on the patient. The latter can quickly become an issue as the preparedness of patient to provide data is easily exhausted. This may result in loss to follow-up which can lead to the biased results, especially if these losses are not random. For example, if in a follow-up process only the data from the satisfied patients with encouraging outcomes are obtained, meanwhile the unsatisfied patients with the less promising outcomes do not want to participate in a follow-up then the registry does not reflect the true picture. The registry should therefore, develop a good patient retention plan that is suitable to the target population.

In all this, a registry should prepare thorough documentation for the entire data collection protocol and provide methodological guidelines, standard instructions and rules for data collectors/providers and other data users. It is often advisable to describe the typical data flow of the registry, where the information on how the data travels from the source to the registry, together with the other additional information (e.g. key persons/stuff included in the process, type of technology and data collection method used, access rights, data transmission, timetables) is clearly specified. The description of the data flow can help the registry team and other stakeholders (e.g. company that will provide the technical solution) to better understand the whole data collection protocol. Among the other things, it can serve also when performing evaluations of the data collection protocol (e.g. identification of potential sources of errors etc.)

Table 6.6: Example of the data flow description

- In the hospital a nurse collects the data from the patient via the paper-based form
- the data is then entered into the web-based system; the nurse and the doctor are the only one who have access to the data and can modify the data;
- the data is transmitted via web server to the central database;
- after the 6 months the patient is contacted by telephone and asked three additional questions; data is collected via paper-based form and then entered into the web-based system;
- …

6.2.5 Research-based registries - additional points to consider

Nowadays, many registries are being developed that are taking a more research approach. These study-oriented or research-based registries possess different characteristics, therefore some additional points need to be considered when developing this type of a registry. Hoverer this does not mean that points described below should be entirely ignored by a registry holders who aim to develop more 'classical', wide encompassing registries.

6.2.5.1 Research questions and hypotheses

When the purpose and main objectives of the registry are clearly defined the next step is to take that purpose or idea and shape it into a researchable question. Research questions and hypotheses narrow the purpose of the study and become major 'signposts' for guiding the overall study (1).

Research questions for registries range from purely descriptive questions aimed at understanding the characteristics of people who develop the disease and how the disease generally progress, to highly focused questions intended to support decision-making (2). Research questions in registry-based studies are generally hypothesis generating (i.e. developing hypotheses after the data are collected and new knowledge is gained) or evidence building, rather than hypothesis testing. However, registries focused on determining clinical effectiveness, cost-effectiveness or risk assessment are commonly hypothesis driven (2, 3, 4). Regardless of the nature of a research questions (or hypotheses) it is crucial for a registry planner to define them because all further decisions (e.g. registry population, what data will be collected and analysed) and work in a registry development process are guided by research questions of interest. Proper formulation of a research question or hypothesis is not an easy task and should not be underestimated. Not properly defined, unfocused or underdeveloped research question or hypothesis can generate a risk for not getting the right results and accomplished objectives of a registry. Accordingly, it is highly recommended that registry developer invests/spends required time to suitably develop a research question or hypothesis.

A (research) ideas as a foundation for developing a research questions or hypotheses are typically gathered by literature review, critical appraisal of the published clinical information, brainstorming with colleagues, seeking experts' opinions, and evaluating the expressed needs of the patients, health care providers (2, 5). The clinical questions of interest can also be defined by payers/sponsors of the registry. Thus, it is not uncommon that multiple questions are set as a result of interest of different stakeholders. In that case a registry planner should be aware that a higher number of research questions can increase the complexity of a registry study design and subsequent collection of a data and statistical analysis. Registry developers should therefore assess whether it is feasible to answer every question of interest.

When defining a research questions or more specific research hypotheses it is important that they are accurate, understandable and focused enough for a specific registry. The clinical epidemiology literature offers various instructions on research questions and hypotheses, such as, for example, FINER (6) and PICOT (7) criteria for a good research question. The example of a research question and hypothesis for a registry is presented in Table 6.7.

Table 6.7: The example of a research question and hypothesis for registries

Idea/Interest/Purpose	Research questions/hypotheses
Monitoring clinical effectiveness of hip implants	Hypothesis: In Europe, exchangeable neck hip stem implants have significantly higher revision rate than hip implants with un-exchangeable neck.
Natural history of patients with diabetes disease	Research question: What is the incidence and prevalence rate for diabetes type 1 disease among children and adults in Slovenia?

6.2.5.2 Key exposures and outcomes

In a simplified way we can describe the exposure and outcome as a relationship, where one event (i.e. exposure) affects the other (i.e. outcome). In the field of patient registries, the term 'exposure' refers to treatments and procedures, health care services, diseases, and conditions, while outcomes generally represent measures of health, onset of illness or adverse events, quality of life measures, measures of health care utilization and costs (2).

It is crucial to identify the key exposures and outcomes at the very beginning of a registry development, since the selection of exposures and outcomes will affect further registry development (e.g. registry study design, data collection process). The identification of key exposure and outcome variables is guided by the registry research questions that are defined at the registry's outset. When identifying the key exposures and outcomes it is important to know that sometimes more outcomes need to be selected (as a result of multiple questions of interest), and exposure often includes a collection of different information, such as dose, duration of exposure, route of exposure, and adherence (2, 8). For example, if we select smoking cigarettes as an exposure for measuring a particular outcome (e.g. heart disease) probably it would not be enough to have only one binary variable for exposure (i.e. smoking or non-smoking), but to include also other information, such as dose (e.g. how many cigarettes per day) and duration (i.e. how many years of smoking). During the identification of key exposure variables it is therefore necessary to consider also this aspect, and it is useful to take into account independent risk factors for the outcomes, and confounding variables as well. More information on selecting data elements for registry is provided in Chapter 6.3.

6.2.5.3 Study design

Registry studies are observational studies in which the researcher merely observes and systematically collects information, and, unlike in the experimental studies, does not assign specific interventions to the study subjects being observed. In observational studies the researcher choose what exposures to study, but does not influence them.

Although the patient registries are generally considered as prospective observational studies, the registries, from time perspective, could be both – the prospective and retrospective study. Prospective studies are designed to gather data about events that have not happened yet, while retrospective studies are designed to gather data about events that have already happened. Thus, prospective studies look forward in time and retrospective studies look backward (9).
It is not always simple to define which study design[85] the registry follows, using the traditional epidemiological terms. For example, in some situations study design for a registry can be considered as an opened cohort or simply a case series of patients under some specific diagnosis (22). Sometimes even the registry's nature itself does not require clear specification of its study design. However, it is necessary for a registry designer to understand which study model can be applied in a registry. Several study designs that are more commonly applied in registries are:

Cohort study

Cohort studies follow a group of persons with common characteristics to observe whether or not they develop a particular endpoint or outcome over time. Cohort studies are used for descriptive studies and for studies evaluating comparative effectiveness, safety or quality of care. Cohort study may include patients with specific exposures (e.g. to a specific drug or class of drugs) or disease of interest, and it may also include one or more comparison groups for which data are collected simultaneously using the same methods. A single cohort study may include multiple cohorts, each defined by a common disease or exposure (2). Cohort studies are particularly useful for studying rare exposures, as exposed patients are selectively enrolled in the study. Cohort studies allow the investigator to study multiple outcomes for a particular exposure, and since the individuals are followed over time, incidence rates and relative risk can be calculated (10). If the outcome of interest is rare, a very large study population must be followed to observe a number of outcomes that is sufficient to demonstrate a precise association between the exposure and outcome. Limitation of registry-based cohort studies is that they can be very expansive and time consuming due to large sample size and long follow-up period. The attrition of study population/loss to follow-up which can cause a serious bias in study’s results is also one of the weaknesses of the cohort studies.

Case-control study

[85] A study design is a specific plan or protocol for conducting the study, which allows the investigator to translate the conceptual hypothesis and research question into an operational one (21).

In a case-control study two groups are assembled: subjects (cases) with a given health outcome (e.g. disease, adverse events), and subjects (controls) without the outcome but otherwise come from a similar population. Case-control study then retrospectively compare these study groups (cases and controls) to identify any possible causal exposures. Case-control studies are well suited to investigate rare outcomes (therefore ideal for studying the etiology of rare diseases) or outcomes with a long latency period because subject are selected from the outset by their outcome status. In comparison to cohort studies, case-control studies can be more efficient way of studying outcome of interest as they do not require time and expanse of following patient over time to observe the outcome (3, 10, 11). Controls in case-control study can be selected from outside the registry as well. In that case, care must be taken that controls meet the requirement of arising from the same source population as the cases to which they will be compared. It is worth mentioning also, that matching in case-control studies, which ensures that characteristics (e.g. gender, age, race) are similar in the cases and their controls, is important element and can increase the study's efficiency (2, 12). However, the limitation of case-control studies is that they cannot measure incidence, prevalence and relative risk. As the exposure information is retrospectively collected case-control studies are also susceptible to recall bias, especially when exposure information is collected by self-report or interview. Another weakness in case-control studies is the risk for selection bias. In general, selecting a suitable control group is the most complex challenge in conducting a case-control study (10, 13, 14).

Nested case-control study

A nested case-control study is comprised of subjects sampled from an assembled cohort study in which the sampling depends on disease/outcome status. In a nested case-control study, cases of a disease/outcome that occur in a defined cohort are identified, and, for each, a specified number of matched controls is selected from among those in the cohort who have not developed the disease/outcome by the time of disease occurrence in the case. A nested case-control study offers significant reductions in cost and efforts of data collection and analysis, as extensive/additional data of interest are collected only for smaller number of subjects, instead of for entire cohort. Its high cost efficiency but relatively minor loss in statistical efficiency is the reason that for many research questions and in many patient registries a nested case-control study is applied. However, the reduced statistical power due to reduction of the sample size, is considered as a potential weakness of the nested case-control study (15, 16).

Case-cohort study

Like a nested case-control study, the case-cohort study runs within a cohort. It is basically the alternative to the nested case-control study with simpler selection of controls but with a more complex analysis. In case-cohort study all subjects who developed an outcome of interest in the cohort, are included as cases, while the controls are selected with random sampling of the entire cohort, which means that every subject in a cohort has an equal probability of being selected as a control. Unlike the nested case-control design, controls (often referred to as a subcohort) in the case-cohort study are not matched, but are randomly selected. Since controls are selected without matching, case-cohort studies have several advantages: controls can be easily selected at the beginning of the follow-up; multiple subcohorts can be selected, and the same subcohort can be used for multiple outcomes. Since only a sample of a total cohort is studied, this study model is (like nested case-control study) also very cost effective, especially when additional expensive data collection is needed (15, 16, 17).

Case series

A case series is a descriptive study, which can sample patients with both a specific outcome and a specific exposure, or includes patients with specific outcome regardless of the exposure. Case series drawn from the registry is an uncomplicated application that does not need complex analytical work. It is used to describe the characteristics of certain cases or events, and does not provide strong statistical evidence. Nevertheless, case series, which is dependent on the generalizability of the registry itself, should not be underestimated, since it is a useful means to introduce the specific event (e.g. unusual and rare conditions) and is often the basis for future research. Readers are encouraged to read also more about the self-controlled case series which are commonly applied in registries, and are especially useful for vaccine safety assessment (2, 18, 19, 20).

Other study designs that are also sometimes used, but are not covered here, include cross-sectional study and case-crossover design.

6.2.5.4 Comparison groups

Registry can include and collect also data on one or more comparison groups. Although registries usually do not use comparison groups, they are essential when it is important to distinguish between alternative decisions, to assess the magnitude of differences, or the strength of associations between groups. Based on the registry's objectives three types of comparison groups can be used:

- ***internal comparison group*** (data is collected simultaneously for patients who are similar to the focus of interest, but who do not have the condition or exposure of interest),
- ***external comparison group*** (data have been collected outside of the registry for patients who are similar to the focus of interest, but who do not have the condition or exposure of interest),
- ***historical comparison group*** (refers to patients who are similar to the focus of interest, but who do not have the condition or exposure of interest, and for whom information was collected in the past, for example, before the introduction of an exposure or treatment or development of a condition)

When deciding about including a comparison group in a registry, registry developer should consider also that adding a comparison group may add complexity, time, and cost to a registry (2).

6.2.5.5 Sampling frame and sampling method

Registries sometimes try to include all units of the target population, but often they include just a sample of the target population from which inferences about the whole population can be made. The need for including only a sample of the target population typically arises because of limitations of time and resources but also due to other constraints (26). The activity of selecting cases (i.e. patients, institutions, object or events) into sample from a larger collection of such cases, according to a specific procedure, is called sampling. Ideally the sample is drawn directly from the target population but usually this is not the case, because sample can be drawn only from cases to which registry/participating sites have access (i.e. accessible population). Hence, accessible population represents the sampling frame from which a sample is selected. Sometimes the accessible populations is the same as the target population, but usually is a subset of the target population. In terms of precision of registry's estimates/results, a registry planner should be aware of this issue, since a non-coverage of certain parts of target population can lead to the biased estimates (27, 28).

In other words, if cases of the target population who cannot be sampled (because there is no access to them) are different from those who can be drawn into a sample, the registry findings can be biased. During a sampling phase a registry planner need to asses what impact on the registry findings a sampling frame and its potential non-coverage issue could have.

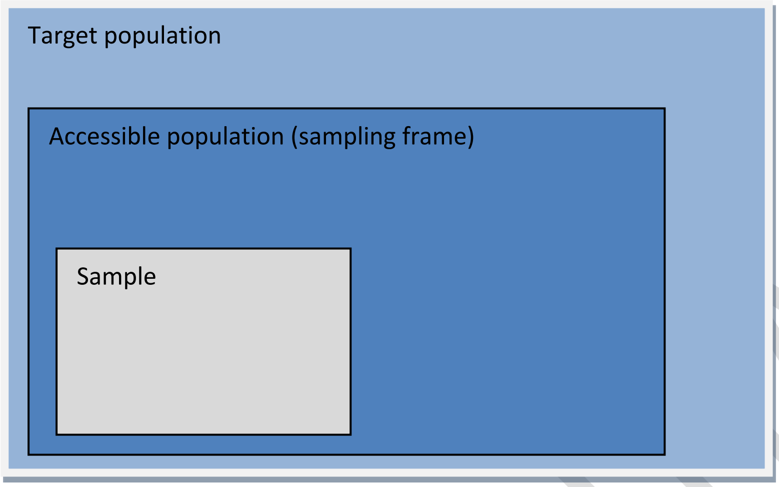


Figure 6.3. Key concepts in sampling (adapted from Sim J and Wright C, 2000)

Many different sampling methods can be used when selecting cases for a registry. Sampling designs are classified as either probability sampling or nonprobability sampling. In general, the probability sampling is preferred method in which the selection of individual cases (e.g. patients, events) is left to chance, rather than to the choice or judgement of the person. However, in some situations the probability sampling is not feasible and nonprobability sampling is more useful. Some sampling methods that are often used for generating sample include simple random sampling; stratified random sampling; systematic sampling; cluster sampling; multistage sampling; case series or consecutive (quota) sampling; haphazard, convenience, volunteer, or judgmental sampling; modal instance; purposive; and expert sampling (2).

6.2.5.6 Representativeness and generalizability

When selecting patients, hospitals or events it is important that consideration about representativeness is made, since the representativeness is essential component of a registry study. If the sample is not properly representative, conclusions/generalization may be incorrect. Registry developer should consider representativeness in terms of patients (e.g. men and women, children, the elderly, racial and ethnic groups), sites (e.g. geographic location, practice size, academic or private practice type) and events (e.g. type of events/services on a particular day) (2). Registry developer should critically assess how the potential lack of representativeness can affect the results of a registry. For example, suppose that the purpose of the registry is to monitor the clinical effectiveness of a specific surgeries. If a registry will include only academic centres/hospitals with high technical support, then the results probably would not reflect a true picture. On the other hand, for example, when registry is not representative in terms of gender (e.g. higher number of women in a registry), this would have no impact on the representativeness of the registry findings if the outcome that is observed (e.g. clinical effectiveness of a specific drug) does not vary with gender.

Associated with the representativeness, the generalizability concept is often used, which refers to the extent to which the conclusions of the registry study can be generalized/applied to populations other than those sampled and included in the registry. Strong generalizability or external validity is achieved by the inclusion of a typical patient sample which is often more heterogeneous (e.g. different demographic characteristic, comorbidity). Patient registries are generally designed to have strong external validity so that their population will be representative and relevant to decision makers. It is important to note that the way in which patients are included, classified and followed directly affects generalizability of (2, 3). In terms of data interpretability it is important to describe and document the representativeness and generalizability of a registry, and whether it covers the relevant patients, events and periods of interest.

References

1. Creswell John. Research Design: Qualitative, Quantitative and Mixed Method Approaches, 3rd Edition
2. Gliklich RE, Dreyer NA, eds. Registries for evaulating patient outcomes: A User's Guide. 3rd ed.
3. ISPOR: Taxonomy of patient registries: classification, characteristics and terms.
4. The Yeshiva Fatherhood Project. Introducing qualiative hypothesis-generating research.
5. Lobiondo-Woow G, Haber J. Chapter 2. Nursing Research - Methods and Critical Appraisal for Evidence-Based Practice.2013.
6. Hulley S, Cummings S, Browner W, et al. *Designing clinical research*. 3rd ed. Philadelphia (PA): Lippincott Williams and Wilkins; 2007.
7. Brian Haynes R. Forming research questions. *Journal of Clinical Epidemiology* 2006; 59:881-6.
8. International Agency for Research on Cancer. Cancer Epidemiology – Principles and Methods.1999.
9. Stark Nancy J. Registry Studies for Medical Devices – Whitepaper and Workshop Invitation. 2010.
10. Verhamme K. Study designs in Paediatric Pharmacoepidemiology. European Journal of Clinical Pharmacology 67, S1 (2010) 67-74
11. Song Jae W, Chung Kevin C. Observational Studies: Cohort and Case-Control Studies. Plast Reconstr Surg. 2010 December, 126(6): 2234–2242
12. Rose S, van der Laan M. J. Why Match? Investigating Matched Case-Control Study Designs with Causal Effect Estimation. The International Journal of Biostatistics,Volume 5, Issue 1 2009 Article 1.
13. Carlson D. A. M, Morrison S. R. A User’s Guide to Research in Palliative Care: Study Design, Precision, and Validity in Observational Studies. JOURNAL OF PALLIATIVE MEDICINE, Volume 12, Number 1, 2009
14. Jepsen P., Johnsen S. P., Gillman W. M., Sorensen H. T. Interpretation of observational studies heart 2004; 90:956–960.
15. Ernster, V. L. Nested case-control studies. Preventive medicine, 23, 587-590, (1994).
16. Langholtz, B. Case-Control Study, Nested. Volume 1, 646-655. In Encyclopedia of Biostatistics, 2nd Edition. Eds. Armitage, P. and Colton T. John Wiley & Sons, Ltd, Chichester, 2005.
17. Wachoider S, Silverman DT, McLaughlin JK, Mandel JS. Selection of Controls in Case-Control Studies. American JounaJ of Epidemiology Vol. 135, Mo. 9. 1992
18. Kooistra B, Dijkman B, Einhorn TA, Bhandari M. How to design a good case series. The Journal of bone and joint Surgery. 2009, 91 Suppl 3:21-6.

19. Dekkers OM, Egger M, Altman DG, Vandenbroucke JP. Distinguishing case series from cohort studies. Annals of Internal Medicine, 2012 Jan 3;156(1 Pt 1):37-40
20. Andrews N. Epidemiological designs for vaccine safety assessment: methods and pitfalls. Biologicals. 2012 Sep;40(5):389-92
21. WHO. Training manual for community-based initiaties: A practical tool for trainers and trainees. 2006.
22. EPIRARE. Deliverable D4: Guidelines for data sources and quality for RD Registries in Europe. 2014.
23. Laforet P et al. The French Pompe registry. Baseline characteristics of a cohort of 126 patients with adult Pompe disease. Revue Neurologique, Vol 169, Issues 8–9, pages 595–602. 2013.
24. Polit DF, Beck CT. Nursing Research: Generating and Assessing Evidence for Nursing Practice. 8th Edition. Lippincott Williams and Wilkins, a Wolters Kluwer business.
25. Eduardo Velasco. Inclusion Criteria. In Encyclopedia of Research Design, eds. Neil J. Salkind. 2010. SAGE Research methods. Available at: http://srmo.sagepub.com/view/encyc-of-research-design/n183.xml
26. Sim J, Wright C. Research in Health Care: Concepts, Designs and Methods. 2000. Stanley Thornes (Publishers) Ltd.
27. Brick, Michael J., Ismael Flores-Cervantes, Kevin Wang in Tom Hankins. 1999. Evaluation of the Use of Data on Interruptions in Telephone Service. Proceedings of the Survey Research Methods Section of the American Statistical Association.
28. Groves, Robert M. 1989. Survey Errors and Survey Costs. New York: Wiley.
29. Cochran WG. Sampling Techniques. Third ed. Wiley; 1977.
30. Noordzij M et al. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant (2010) 25: 1388–1393.
31. Altman DG. Practical Statistics for Medical Research. London, UK*:* Chapman & Hall*;* 1991.
32. Bland M. An Introduction to Medical Statistics. 3rd edn*.* Oxford, UK*:* Oxford University Press*;* 2000*.*
33. Lwanga SK, Lemeshow S. Sample size determination in health studies - A Practical Manual. World Health Organization 1991.
34. Carey TS, Sanders GD, Viswanathan M, et al. Methods Future Research Needs Reports, No. 8. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012 Mar.
35. Gregg, M.B. Field Epidemiology. New York: Oxford University Press, 2002
36. Statistični urad Republike Slovenije. Metodološki priporočniki: Smernice za zagotavljanje kakovosti, št 2. Ljubljana, 2012.

6.3 Registry dataset

Before deciding on what data to collect in a registry, it is important to be clear about the purpose of the registry. Once the registry's purpose and goals are determined, the data that are required to meet those objectives can be identified. The selection of data elements[86] for a registry starts with the identification of the data domains which are collections of data elements that relate to a common topic. Data domains that are commonly used in registries include (1, 2, 3):

- ***Patient domain*** (Data that describe the person, such as demographic information, contact information, information about medical history, health status and patient identifiable information. The inclusion of patient identifier is often necessary and can bring many advantages. For example, it enables linkage to other data sources; tracking patient through time and place; it enables gathering information when the personal contact with patients is required; and data quality checks. However, the main issue on that field is the privacy aspect. More about that is described in the chapter 5.)
- ***Provider domain*** (Data that describe the characteristic of the individuals providing health care interventions to the patients included in the registry)
- ***Exposure domain** (*Data that describes the patient's experience with the product, disease, device, procedure, or service of interest to the registry.)
- ***Outcome domain*** (Data that are of main interest to the registry. Often this refers to measures of health, onset of illness or adverse events, quality of life measures, measures of health care utilization and cost.)
- ***Covariate/confounder domain*** (Data that are not of primary interest to the registry but their inclusion and measuring is still important, since they are related to the exposure or outcome or both. Inclusion of covariates allows controls/adjustments during analyses.)
- ***Administrative domain*** (Information related to registration process; for example, date of previous or next follow-up, date of reminder.)

After the identification of data domains for a registry, decisions about which data specifically will be collected by a registry need to be made. The process of selecting and building data elements is one of the most important and challenging task that often determines the final success of the registry. If the registry does not collect data that would fulfil its intended purpose and goals, it can turn out to be useless. On the other hand, if the registry sets too complex data collecting inducing higher costs and burden, it may jeopardize its sustainability. Hence, a careful approach is required and many aspects need to be taken into consideration when building dataset for a registry.

The process of building a dataset is undertaken by a team which typically includes clinical experts, health informatics, statisticians and epidemiologists. During the process various tools can be used, such as mind mapping (e.g. XMind, FreeMind) or/and spreadsheet (e.g. Microsoft Excel) tools.

[86] Data element is any named unit of data used to record information in a registry or database. It is characterized by a name, a definition, representation terms and the set, range and/or format values (8). The term is often used interchangeably with 'variable'. Patient's date of birth is the example of the data element.

6.3.1 General principles for building registry dataset

Minimalist approach in building a dataset

Data elements need to be carefully considered in relation to the purpose of the registry. Every data element must support the purpose and goals of the registry. If there is no strong argument for its collection, it should not be included.

The burden and costs for data collecting

The success or failure of a registry is often determined by the costs and burden of data collection. When building a dataset for a registry it is necessary to consider the burden of data collection that will be put on a patient, physician/health provider, and a registry team as well. The likelihood of loss to follow-up due to the burden of data collection should be also considered.

Availability of data sources for data elements

It is recommended to identify existing data sources and assess their usefulness. Linkage to other data sources can significantly lower the cost and burden of data collecting.

Privacy aspect

During the selecting and developing data elements, registry planners must take into account security policies and privacy issues. They must assess, whether dataset is complied with the information privacy principles, and how the inclusion of data elements that are private or confidential in nature will affect the patient's response.

Consideration of data quality for data elements

Data elements of uncertain quality or coverage should not be included in the registry dataset. Unless reliable information can be collected on majority of cases, the item should not be part of a registry dataset.

Use of data standards

The use of data standards is one of the most important aspects in building a registry. Standard data elements and definitions should be used when possible. Standards promote consistency, comparability, and common understanding of data elements. The use of existing data standards, such as classifications, clinical terminologies and common data sets enables comparison of results, data exchange and reuse – the activities that are nowadays invaluable and highly supported by European Union (see term semantic interoperability in the chapter 3.2.5).

Explicit definitions

When there is no suitable internationally standardized data elements or they cannot be used in specific registry, the registry team needs to define and select their own data elements. Definitions of data elements should be explicit and should ensure that there is no variation in concept, collection or format between institutions and individuals collecting and reporting on the data. ISO/IEC (10) specifies requirements and recommendations on the formulation of data definitions that are specified in Metadata Registries. According to the ISO/IEC 11179 – 4 (2004) a data definition should (a) be stated in the singular, (b) state what the concept is, not only what it is not, (c) be stated as a descriptive phrase or sentence(s), (d) contain only commonly understood abbreviations, (e) be expressed without embedding definitions of other data or underlying concepts, (f) state the essential meaning of the concept, (g) be precise and unambiguous, (h) be concise, (i) be able to stand alone, (j) be expressed without embedding rationale, functional usage, or procedural information, (k) avoid circular reasoning, (l) use the same terminology and consistent logical structure for related definitions, and (m) be appropriate for the type of metadata item being defined.[87]

Selecting value domains, setting validation rules

For each data element a set of permitted values (i.e. value domain) must be determined. A value domain can be enumerated, where the value domain is specified by a list of all its permissible values (e.g. 1=male, 2= female), or non-enumerated, where the value domain is specified by a description rather than a list of all permissible values (e.g. a value domain for person's age might be "18 years and older") (11). It is important that value domains are determined thoroughly and clearly. This can be achieved by the use of various attributes that are associated with a value domain (specifying, e.g., representation class, datatype, format, maximum character quantity, unit of measure). The Australian Institute of Health and Welfare (6) in its data development guide provides some of the key recommendations regarding the value domain, namely (a) ensure that the value domains are consistent and mappable to (inter)national data standard, where these exists, (b) where a classification or code set is used as value domain, the edition of the classification or code set must be clearly referenced to avoid ambiguities about which edition is in use, (c) permissible values must be exhaustive within the value domain and mutually exclusive, (d) consider the proper degree of granularity, (e) when using 'other', to ensure an exhaustive set of permissible values, using a code value that is contiguous with the last code in the permissible value sequence should be avoided since this allows adding another enumerated category to the list of permissible values without renumbering the codes, and (f) use the supplementary values to capture missing values in order to accommodate statistical analysis.[88]

Setting validation rules is another activity that is highly recommended to be done. Selecting possible ranges of the values (e.g. person's age cannot be above 120 years, body height in centimetres cannot contain more than 3 characters, date of injury cannot be date from the future) taking into account also internal consistence with regard other variables (e.g. if person is male or his/her age is higher than 55, he/she cannot be pregnant) or any other errors (e.g. empty cell) helps to reduce the number of errors and improve the data quality. This is especially in case of electronic data collection, where a mechanism can be established to automatically alert a user whether information entered are inconsistent, not within expected range of values, not given in the correct format etc.

Minimum dataset

Registry team should decide on the minimum/core dataset which is a list of variables that are essential to collect the data for any case/subject. It should be carefully considered and specified whether a data element is mandatory (i.e. always required to collect the data), conditional (i.e. required to collect the data when a certain condition is met) or optional. When a data element is of conditional type, the condition must be clearly documented (e.g. the number of cigarettes smoked daily is required if the patient is regular smoker).

Modifying data elements

Registry operating over a long period of time will be faced with the possibility that data elements will change. When changing the data elements a registry team should try to comply with the existing standards and to retain the longitudinal comparability. In any case, it is important that a registry

[87] Detailed explanation of the above-mentioned recommendations is provided at: http://standards.iso.org/ittf/licence.html.

[88] More information on value domain concept and its attributes can be found at:
http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=6442458038
http://standards.iso.org/ittf/licence.html

considers the impact that these changes will have on a collection and interpretation of findings. (See also the chapter 9.1 ‘Changing an existing registry’.)

Testing dataset

When the first version of the registry dataset is developed, it should be tested. Each data element should be checked separately whether its definition, value domain, any rules or other descriptions are properly determined, comprehensive and understandable. Looking on the entire dataset, registry team should check the overall consistency of the dataset, asses the data collection burden and evaluate the possibility of making errors in the data collection process.

Methodological guide

Normally, every dataset, together with the data collection process, requires the methodological guide that includes detailed information about what is collected and how. It is used to provide the user with the advice or interpretation on how to treat particular data element and successfully perform the data collection. The guide may include (a) the interpretation of data element’s definition and value domain, (b) the explanation of what exactly is collected/included in the observation and what is not, covering all unclear cases/situations, (c) the introduction of rules and restrictions for specific data elements, including the information about the data element’s format and about whether the data element is of mandatory type, and (d) the information about the data collection and data reporting, such as who should collect the data on the specific data element/variable, when he/she should collect the data and by which method/instrument, who is obliged to report the collected data and what is the dynamics of the reporting.

Well-documented and accessible data elements

Data elements should be well-documented and readily accessible to everyone who is interested in a registry’s dataset. Well-documented and transparent data elements gain an understanding of the collected data and ensure the consistency in the data collection process. Visibility and usability of the dataset are important characteristics, meaning that the dataset can be easily noticed and reused by others. This promotes standardisation and comparability. Hence, it is important that a registry establishes a data dictionary which is the inventory of all data elements/variables included in a registry (see chapter 6.5.6.6 ‘Data dictionary’).

(2, 4, 5, 6, 7, 8, 9, 10, 11)

6.3.2 International coding systems, terminologies and common data sets

As already mentioned, registry should use existing standards wherever possible since this facilitates consistency, comparability, data exchange and reuse. When developing a registry dataset, a registry developers and steering committee should together identify the existing standards that could potentially be used, and determine the most advisable standard to adopt. Table 6.8 presents several international standard coding systems and terminologies that are widely used in the health domain (see also chapter 3.2.5.1 ‘Standards, models and tools’ and 10.11.2 ‘eHealth standards’)

Table 6.8. International coding systems and terminologies

Area	Standard	Developer	Website
Diseases	ICD-10-CM ICD-9-CM ICD-O	WHO	www.who.int/classifications/icd/en
	ORPHA-codes	ORPHANET	www.orpha.net
Medical Nomenclature	SNOMED	International Health Terminology Standards Development Organization	www.ihtsdo.org/snomed-ct
Devices	Global Medical Device Nomenclature (GMDN)	GMDN Maintenance Agency	http://www.gmdnagency.com/
	Universal Medical Device Nomenclature System (UMDNS)	WHO Collaborating Centre ECRI	http://www.ecri.org.uk/umdns.htm
Drugs	ATC/DDD Index	WHO Collaborating Centre for Drug Statistics Methodology	http://www.whocc.no/atc_ddd_index/
	MedDRA (Medical Dictionary for Regulatory Activities)	International Conference on Harmonization (ICH)	http://www.meddra.org/
	WHO Drug Dictionary	WHO	http://www.umc-products.com/DynPage.aspx?id=73588&mn1=1107&mn2=1139
Adverse Reactions	WHO-ART	WHO, maintained by the Uppsala Monitoring Centre	http://www.umc-products.com/DynPage.aspx?id=73589&mn1=1107&mn2=1664
	EU SPC ADR database	EMA	http://www.imi-protect.eu/methodsRep.shtml
	MedDRA (Medical Dictionary for Regulatory Activities)	International Conference on Harmonization (ICH)	http://www.meddra.org/
Disability	ICF	WHO	http://www.who.int/classifications/icf/en/
External Causes of Injury	ICECI	WHO	http://www.who.int/classifications/icd/adaptations/iceci/en/
Primary care	ICPC-2	WHO	http://www.who.int/classifications/icd/adaptations/icpc2/en/
Procedures	ICD-10-PCS ICD-9-CM Vol. 3	WHO	www.who.int/classifications/icd/en
Medical Laboratory Observations	LOINC	Regenstrief Institute	http://loinc.org/
Genes, genetic disorders and traits	Online Mendelian Inheritance in Man (OMIM)	McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD)	http://www.omim.org/
Genes	HGNC	Human Genome Organization (HUGO)	http://www.genenames.org/about/overview

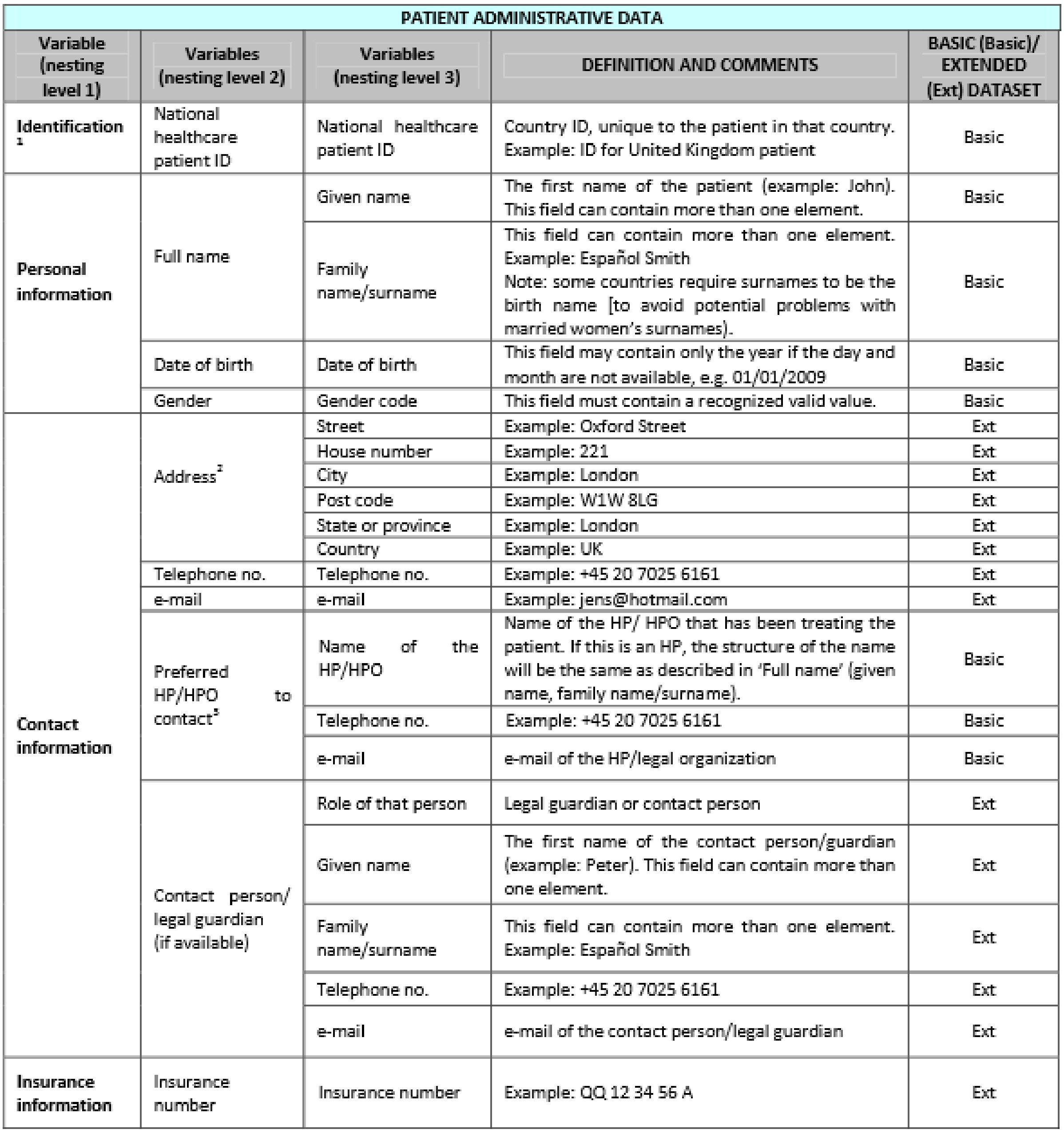
A registry team should also look for the existing data elements, or even more, common datasets. Before deciding on data elements, registry team should make an overview of the current state of the art on the domain that registry covers, and try to identify the already developed data elements and datasets that could be reused in their case. Reusing the commonly used and accepted data elements, in addition to the above mentioned advantages, could also mean saving effort that is needed for the development of new data elements. However, it should be noted that when the existing data element is not relevant or is too constraining for the needs of the registry, it should not be used (6).

In recent years, important steps have been made towards the harmonisation between registries and other data sources, when various organisations and projects started developing common datasets for their own domains. Here, it is certainly worth mentioning the epSOS project, which has done an important work in the field of sharing information about the patient. Its so-called Patient Summary dataset, which aims to support safe, high-quality cross-border care for emergency or unplanned care events, consists of approximately 70 variables and comprises patient administrative data and patient clinical data.
Table 6.9 shows a non-exhaustive list of common datasets that exist in the EU health domain.

Table 6.9. Existing common datasets

Area	Author	Common data set	Link to the dataset
Rare diseases	EPIRARE	EPIRARE common data set	http://www.epirare.eu/_down/del/D9.3_ProposalforCDE_FINAL.pdf
Arthroplasty	EFFORT-EAR	EFORT EAR Minimal datasets	http://www.ear.efort.org/
Patient summary	epSOS	epSOS Patient Summary	http://ec.europa.eu/health/ehealth/docs/guidelines_patient_summary_en.pdf
Cancer	ENCR	ENCR Standard dataset	http://www.encr.eu/images/docs/recommendations/recommendations.pdf
Cardiology	CARDS	CARDS Data Standards	http://www.escardio.org/Policy/Pages/data-standard-cards.aspx
Diabetes	B.I.R.O.	BIRO Common Dataset Summary	http://www.biro-project.eu/documents/downloads/D3_1_Common_Dataset_v1_7.pdf
Neuromuscular disorder	TREAT-NMD	FSHD Core Dataset	http://www.treat-nmd.eu/downloads/file/registries_toolkit/FSH_core_dataset.pdf
Multiple Sclerosis	EUReMS	EUReMS Core data set	http://eurems.eu/attachments/article/93/EUReMS%20Data%20Mask_August2014.pdf

Table 6.10. Example of the epSOS Patient Summary dataset



PATIENT ADMINISTRATIVE DATA				
Variable (nesting level 1)	Variables (nesting level 2)	Variables (nesting level 3)	DEFINITION AND COMMENTS	BASIC (Basic)/ EXTENDED (Ext) DATASET
Identification [1]	National healthcare patient ID	National healthcare patient ID	Country ID, unique to the patient in that country. Example: ID for United Kingdom patient	Basic
Personal information	Full name	Given name	The first name of the patient (example: John). This field can contain more than one element.	Basic
		Family name/surname	This field can contain more than one element. Example: Español Smith Note: some countries require surnames to be the birth name [to avoid potential problems with married women's surnames).	Basic
	Date of birth	Date of birth	This field may contain only the year if the day and month are not available, e.g. 01/01/2009	Basic
	Gender	Gender code	This field must contain a recognized valid value.	Basic
Contact information	Address[2]	Street	Example: Oxford Street	Ext
		House number	Example: 221	Ext
		City	Example: London	Ext
		Post code	Example: W1W 8LG	Ext
		State or province	Example: London	Ext
		Country	Example: UK	Ext
	Telephone no.	Telephone no.	Example: +45 20 7025 6161	Ext
	e-mail	e-mail	Example: jens@hotmail.com	Ext
	Preferred HP/HPO to contact[3]	Name of the HP/HPO	Name of the HP/ HPO that has been treating the patient. If this is an HP, the structure of the name will be the same as described in 'Full name' (given name, family name/surname).	Basic
		Telephone no.	Example: +45 20 7025 6161	Basic
		e-mail	e-mail of the HP/legal organization	Basic
	Contact person/ legal guardian (if available)	Role of that person	Legal guardian or contact person	Ext
		Given name	The first name of the contact person/guardian (example: Peter). This field can contain more than one element.	Ext
		Family name/surname	This field can contain more than one element. Example: Español Smith	Ext
		Telephone no.	Example: +45 20 7025 6161	Ext
		e-mail	e-mail of the contact person/legal guardian	Ext
Insurance information	Insurance number	Insurance number	Example: QQ 12 34 56 A	Ext

References

1. Australian commission on safety and quality in health care. Operating Principles and Technical Standards for Australian Clinical Quality Registries. 2008.
2. Gliklich RE, Dreyer NA, eds. Registries for evaulating patient outcomes: A User's Guide. 3rd ed.
3. ISPOR: Taxonomy of patient registries: classification, characteristics and terms.
4. Rare diseases task force. Patient registries in the field of rare diseases. 2011.
5. WHO. Planning and developing population-based cancer registration in low- and middle-income settings. International Agency for Research on Cancer, 2014
6. Australian Institute of Health and Welfare (AIHW) 2007. A guide to data development. AIHW Cat. no. HWI 94. Canberra: AIHW. Available from: http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=6442458038
7. Health Information and Quality Authority (HIQA). Guiding Principles for National Health and Social Care Data Collections. 2013.
8. EPIRARE. Deliverable D4: Guidelines for data sources and quality for RD Registries in Europe. 2014.
9. National Health Information Management Group (NHIMG) Minimum Guidelines for Health Registers for Statistical and Research Purposes. 2001
10. ISO/IEC 11179-4: 2004 (E). Formulation of data definitions. Available from: http://standards.iso.org/ittf/licence.html.
11. ISO/IEC 11179-3:2013(E). Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes. Available from: http://standards.iso.org/ittf/licence.html

6.4 Data sources for registries

6.4.1 Definition of Primary and Secondary Data Sources

The definition of primary and secondary sources is not strictly connected to the patient registries and could be commonly used also for other research or statistical purposes.

The short definition of the **primary data source** explains it as data collected from the individuals to create (or supplement) the patient registry. Individuals or data providers in these cases could be either patients or clinicians, caregivers, pharmacists or other persons involved in health care. When the registry is completely or partially built on the primary data sources, these sources are collected for the direct needs of the registry. When the primary data source is the only data source for a particular registry, the inclusion of the identifier is not necessary, but desirable also for the cases when the registry serves as the secondary source for another registry.

Secondary data sources are the sources that were established or collected previously for other purposes. Examples of these sources are EHRs, medical charts, different databases (e.g. hospital administration database, census database). If the register is built on two or more data sources, the identifier (e.g. personal identification number or some other unique identifier) must be included in all of the used data sources to enable merging the sources.

Among secondary data sources one should emphasize the importance of "non"-health data sources, especially statistical data sources. The later can serve as an important source of socio-demographic or socio-economic variables and therefore enable us to decrease the burden on patients or health care workers.

It is worth to emphasize that the primary data source for one registry can act later on as a secondary data source for another registry.

There are many pros and cons when using primary or secondary sources.

Primary data sources are in most cases costly, time consuming, but on the other hand can provide data of higher quality with the dimensions of completeness, validity and reliability. When collecting data by a questionnaire or other research instrument, we also create a burden for data providers (patients, clinicians, etc). The burden needs to be taken into account when planning a survey / data collection.

Secondary data sources are on the other hand less costly, they are easier to gather – providing that there is a sufficient legal background. There are the following considerations concerning usage of the secondary data sources:

- In most of the cases the secondary data sources are collected for other purposes (e.g. collected for insurance fund analyses, but later on used to perform different health care analyses)
- Data from secondary sources are usually used either by transfer into the registry or they are linked to other data sources to create a new, larger dataset for analysis.

Emerging challenges:

- Whenever we merge two or more data sources, the identifier of high quality is requested.
- Analyses of the data sources are quite often limited due to different purpose of primary data collection.
- If we merge two or more data sources, each of them carry its own level of quality and influences the final quality level.
- The sufficient legal background should be taken into account when we merge / link the data from different institutions.

6.4.2 Identification of Available Sources

When planning a new registry, all possible and available sources should be analysed. In most of the EU countries, there are legal acts defining registries, their ways of data collections, data providers and possible users. Based on these acts, the future registry holder can observe some of the potential data sources.

As mentioned earlier, when decision on usage of secondary sources is made, the quality and reliability of the source should be explored; on the other hand, the usability of the source for the "new" purpose should not be neglected. It is important to say that level of quality and reliability of one particular registry could be close to perfection for its own purpose, but could be quite unusable for some other purposes.

The most important data sources that the future registry holder should examine and analyse:

1. Primary data sources
   a. Patient reported data are – as described above – usually resource consuming in the sense of data collection, coding, keying, validating. There are many pros when using these data, especially gathering information not covered elsewhere, like opinions, life style, herbal supplements, etc.
   b. Clinician reported data could offer much more information since it is collected directly from the source with more clinical context. On the other hand, this value-added could be burdensome for data user since it is necessary to code the data or preform contextual analyses. Again, it is resource consuming task.

2. Secondary data sources:
   a. EHRs are information on routine medical care and practice. Usually they are structured, the information are coded according to different classification systems. Since these data were prepared for the patients' treatments, they can be very extensive, sometimes even in image formats, and the historical data could be hard to retrieve. But nevertheless, EHRs are the most valuable secondary data source for the patient registries.
   b. Human resource and financial databases could be used in some of the registries when the main purpose is evaluation of the staff or financial resources usage. The most important challenges are classifications since in many of these databases classifications are adapted to other financial systems and are hardly used for statistical or analytical purposes.
   c. Population databases or registries (like Census database, Central Registry of Population, National Patient Registry, etc) are in most cases valuable sources of socio-demographic or socio-economic variables. The user should bear in mind that these sources were

prepared either for administrative purposes (like Central Registry of Population) either for statistical purposes for the field of demographic statistics.

d. Other health registries are maybe the most important data source among above mentioned sources. Again, these registries were probably set up for different reasons and purposes which should be explored quite in details. On the other hand, we should reduce burden on reporting units (e.g. clinicians, hospitals). If there exists data reporting on particular issue (e.g. hospital admission), these data should be re-used as much as possible whenever legal framework allows it.

As written in previous subchapter, if one want to use one of the above mentioned secondary sources, the patient identifier is necessary. For statistical purposes, matching without identifier is also possible with certain probabilistic methods, but not recommended in the process of building the registry. Therefore, the prerequisites for using these sources are:

- Unique patient identifier which is used in all sources we are going to use
- Documentation explaining what is really the content of the source (e.g. target population, metadata on variables…)
- Adequate level of quality for selected variables.

At the end, it is worth to mention that every day new registries are born. The registry holder should take this into account by monitoring the newly registries (appearance and quality). Based on the new registries, the registry design could be changed if the new sources are of high quality, available and their inclusion would bring the burden decrease.

References

1. Wallgren A., Wallgren, B.: Register-based Statistics – Administrative Data for Statistical Purposes, Wiley, 2007
2. Polygenis D, ed. *ISPOR Taxonomy of Patient Registries: Classification, Characteristics and Terms.* Lawrenceville, NJ; 2013.
3. Health Information and Quality Authority (HIQA). Guiding Principles for National Health and Social Care Data Collections. 2013
4. Gliklich RE, Dreyer NA, eds. Registries for evaluating patient outcomes: A User's Guide. 3rd ed. 2014.

6.5 The role of information system methodologies and techniques in the phase of Patient registry creation

Patient registries can be **computerized to a different degree**. Some registries use IT tools only for processing, analysing and representing the data, some also for gathering information in electronic way directly from the information source (patient, clinicians etc.) or indirectly from other information systems (IS) such as electronic health records (EHR's).

Despite of the degree of computerization of the patient registry it is **very useful to take advantage of information system development methodologies, techniques and tools in the phase of development (design) of the patient registry content and functions.**

In this subchapter we will describe:

- how and why different modelling techniques (from the field of IS design) can be applied in patient registry creation;
- how important is to involve an IS expert (or other person with experience in IS modelling techniques) in the patient registry creation and to clearly understand the role of such expert;
- techniques for eliciting requirements / knowledge for patient registry and
- the importance of standard terminologies and code lists.

The main purpose of the following text is to briefly introduce some most used IS design techniques and diagramming notations to the reader. After reading this chapter the reader will:

- understand why modelling techniques are useful in patient registry creation;
- be familiar with some common used modelling notations[89] and terminology to be able to read a model;
- understand the role of IS expert (or other person with experience in IS modelling techniques) in the patient registry creation;
- understand the importance of using standardized terminologies and code lists if they exists.

For more information on this subject and described techniques the readers are encouraged to explore provided links to free tutorials and additional readings.

6.5.1 Why modelling?

"A picture is worth a thousand words."

A **model** is usually a **human construct** to help us **understand real world systems**. When we are **modelling** then we **construct an abstraction** of an existing real world system (or of the system we are envisioning). Modelling help us to see only the most important issues by preventing us from getting distracted by all the details which are not important at this time.

For example a world map is a model of a world. When we are looking only for continents, than the country borders on the map are not needed. But when we are searching the number of countries per continent, than we need a world map with a greater detail, including country borders.

[89] Notation = standardized way of presenting models (usually real world issues, like processes, things etc.)

As we will see we can look on real world issues from different perspectives using different modelling techniques and different standardized way of presenting it (=notation). For example we can explore collecting data for PR from process point of view (how the process is performed, which tasks are executed, who is participating in the tasks, what are inputs and outputs) with **process modelling techniques** such as **business process modelling** and prepare **process model** of collecting data using **business process management notation**. Then we can further explore collecting data for PR only from data perspective (what data elements are collected, how the data elements relate to each other etc.) with **data modelling techniques** such as **entity- relationship modelling** and prepare **data mode**l using **entity-relationship diagram**.

Models are excellent tool to communicate with others. Prerequisite is that all the participants understand at least basic notation standard of the presented model.

6.5.2 The role of IS expert (system analyst, process modeller, health informatics expert etc.)

For useful application of IS methodologies and techniques in PR creation **it is recommended to involve persons with knowledge and experience in IS methodologies and techniques** such as system analysts and/or business process modellers or other persons with the knowledge in this domain **as early as possible in the development of the patient registry** (see 6.1.10.1 Human Resources).

It is very important to clearly understand **the role** of this **IS experts** in the process of patient registry creation. They are only there **to facilitate the process of defining the right content and to provide guidance on** how to accomplish these important tasks with different IS techniques. **Health domain experts (usually clinicians) are those who define the content**, as they have the knowledge of the patient registry domain. IS experts cannot and should not define alone the scope, content, outcomes etc. of the patient registries. They are only facilitators of the PR creation process and responsible for proper modelling.

Communication across all team members and especially between health domain experts and IS experts is the key issue when modelling the PR. As we already stated IS expert is responsible for proper modelling and to be able to do so it is crucial to gather the right information from the right people. The most common way of gathering information is to conduct **guided interviews** with health domain experts and another option is to have an **interactive modelling workshop,** where model is prepared during the session. In both cases it is very important to **properly manage the process of information gathering** from preparation, execution to post execution phase.

In the following two subchapters we will describe some tips on how to conduct guided interview and what an interactive modelling workshop is and how to execute it.

6.5.2.1 Guided interview

Guided interview will provide IS expert with a wealth of information. It is usually divided in 3 phases: preparing interview, conducting interview and post interview phase.

A. In **preparation phase** – project team (or analyst) should:

- **define purpose and objectives** of the interview – why we will conduct interview, what do we want to achieve
- **select the right people** to be interviewed
- **prepare a set of questions** which will guide the interview, it is recommended also to prepare short checklists not to miss any important information
- **arrange a venue**
    - location
    - date / time
    - equipment
- **send out invitations; explain the participants in advance** the purpose/objectives of the interview, what kind of input is expected from them (documents, examples of reports, work instructions, etc.)

B. Execution phase – conducting of interview shall begin with an opening statement of purpose for the interview. This purpose statement is to ensure that the interview has a clear overall goal. It should also be used during the interview to ensure the interview stays on topic.

In the beginning it is recommended to use general and open ended questions. To clarify particular issue closed questions are recommended.

Active listening is very important. Some rules to be a good listener:

- focus on the speaker,
- be aware of non-verbal signs
- respond (verbally or non-verbally) to the speaker to encourage him or her to continue
    - using encouraging words and body language (head- nodding, smile etc.)
    - repeating the received sentence in your own words
    - reflecting your understanding of their position
    - ask questions to clarify his message.
    - summarizing etc.

During an interview it is very important to **provide feedback** to the speaker about his message. Provide feedback if you haven't understood the message with additional questions to clarify what was meant. Repeat sentence in your own words (paraphrasing) to show speaker that you are an active listener and in the same time, potential mistakes in the given message can be corrected.

It is strongly recommended to summarize periodically the speaker's messages.

At the interview end, the participants should be thanked and asked to review the results (written document and graphical models) of the interview on next session if necessary.

C. Post interview phase

After the interview is completed, the project team should review gathered information and prepare a written document and prepare model. Both should be reviewed together with the interviewees.

6.5.2.2 Interactive modelling workshop

Interactive modelling workshop is a guided interview combined with real time modelling. The result of such workshop is a model, confirmed by consensus, with resolved ambiguities.

Modeling workshop is usually led by system analyst/designer. He/she must involve all other participants in the discussion. Usually a whiteboard is used where models are drawn and notes taken. If it is possible, the addressed issue is also real-time modeled with some dedicated software and presented to the participants with overhead projector.

The preparation and execution phase of the process design workshop are similar to guided interview and all above mentioned recommendations should be followed.

6.5.3 Short description of frequently used modelling techniques and notations

There are a lot of different methodologies and techniques in the IS development which can be useful also in the process of patient registry creation, but we cannot list and describe all of them. So in the next subchapters we will present only some **useful IS methodologies and techniques** which can be applied in patient registry design. We will explore **Unified modelling Language (UML)** as a set of different modelling techniques notations used typically in software development and some of them (for example use case diagramming) very useful also in creation of PR. Then we will explore **process management techniques** which will help to **model processes** to fully understand tasks, roles, inputs and outputs, drivers / events of different processes related to patient registry design, patient registry execution and patient registry improvements. The next presented modelling techniques will deal with data- or in broader view knowledge modelling. In this subchapters we will explore “classical” **data modelling** using **entity - relationship modelling techniques** (part of single model approach) and also dual level approach to **knowledge modelling using archetypes**. Very important part of data modelling (especially in the context of semantic interoperability) represents also **terminologies and code lists**. Therefore separate subchapter will explain why it is recommended to use international approved terminologies and code lists when we are collecting health data.

The main purpose of this subchapter is to emphasize how useful are engineering methodologies and techniques of IS design in the patient registry creation and to show some examples how some of them can be used in this process. We expect that this subchapter will evolve / grow in the future and more examples of IS methodologies and techniques will be added.

The techniques are presented in random order. Where they can be applied in PR creation is described in the subchapters.

6.5.4 UML

Unified Modelling Language (UML) is a standardized (visual) modelling language consisting of an integrated set of diagrams. It was developed by Jim Rumbaugh, Grady Booch and Ivar Jacobson in 1994 to help system and software developers accomplish the following tasks: specification, visualization, architecture design, construction, simulation and testing, documentation (1).

Today UML is adopted by Object Management Group (OMG) a consortium of over 800 companies dedicated to developing vendor-independent specifications for the software industry (1).

UML is methodology independent, this mean that the process of gathering requirements, analyzing and modelling them is not formally defined; only diagramming notations of different diagrams are prescribed.

UML 2.0 defines **thirteen types of diagrams**, divided into three categories (2):

- **Structure Diagrams** include the Class Diagram, Object Diagram, Component Diagram, Composite Structure Diagram, Package Diagram, and Deployment Diagram.
- **Behaviour Diagrams** include the Use Case Diagram (used by some methodologies during requirements gathering); Activity Diagram, and State Machine Diagram.
- **Interaction Diagrams**, all derived from the more general Behaviour Diagram, include the Sequence Diagram, Communication Diagram, Timing Diagram, and Interaction Overview Diagram.

For the purpose of PR creation (and due to the limited space) we will present only **Use case diagram**. It is a very valuable tool to define the user requirement (goals) of the PR. We have to stress that also other types of diagrams can be used when creating a PR and especially developing software for PR (PR set-up).

Additional information on UML can be found on UML official web page http://www.uml.org .

6.5.4.1 Use case diagram

Use case analysis is a major technique used to find out the **functional requirements of a system**. **Use case**, an important concept in use case analysis, represents **an objective user wants to achieve** with a system. It can be in text form, or be visualized in a **use case diagram**.

Use case describes a system's actions from an external point of view (user's point of view). Use cases are named with verb or verb and noun phrase for example "Make data quality check".

Use case diagram provides a graphical overview of goals (represented by use cases) users (represented by actors) want to achieve by using the system (represented by system boundary but is often opt out in diagram). Use cases in a use case diagram can be organized and arranged according to their relevance, level of abstraction and impacts to users. They can be connected to show their dependency, inclusion and extension relationships.

An UML use case diagram is mainly formed by **actors**, **use cases** and **associations** (connectors); sometimes also by system boundaries.

An actor is any **person** (also organisational unit) or **external system** (machines, IT system, sensors) that interacts with the system in achieving a user goal. It is drawn as a named stick figure.

Questions to find all relevant actors of a use case:

- Who are the system's primary users?
- Who requires system support for daily tasks?
- Who are the system's secondary users?
- What hardware does the system handle?

- Which other (if any) systems interact with the system in question?
- Do any entities interacting with the system perform multiple roles as actors?
- Which other entities (human or otherwise) might have an interest in the system's output?

Use case is a **function of a system**. It is named by verb and drawn as an ellipse.
Associations are connections between actors and use cases; drawn by a line (sometimes with arrow).
System boundaries are drawn as a rectangle across use cases performed by a system.

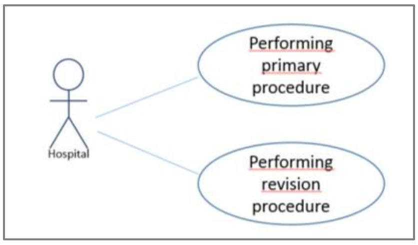


Figure 6.4. Example of use case diagram (»Hospital« as Actor, »Performing primary procedure« and »Performing revision procedures« as Use cases)

First we usually draw a **top level use case diagram** or a **context diagram**. It is a special kind of use case diagram, where the individual use cases are hidden and represented by the system of interest interacting with all the actors. It is very useful to define the context (environment of a system) in our case of a PR.

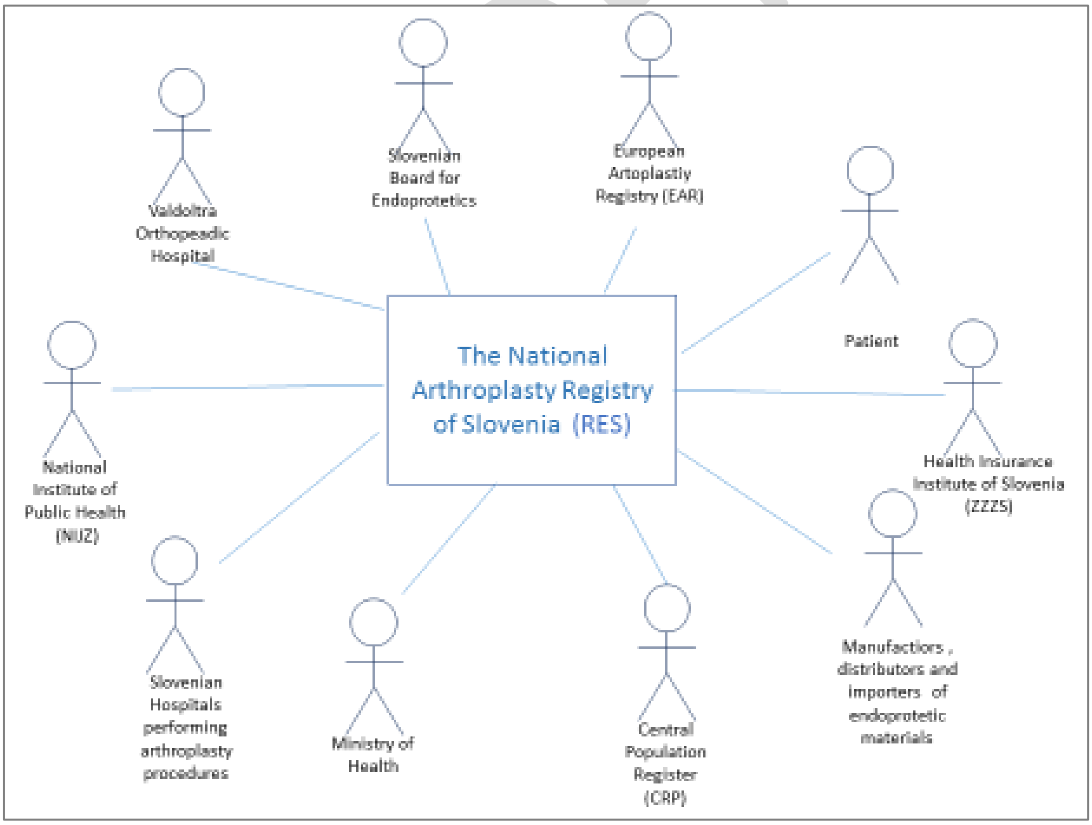

Valdoltra Orthopeadic Hospital
Slovenian Board for Endoprotetics
European Artoplastiy Registry (EAR)
Patient
The National Arthroplasty Registry of Slovenia (RES)
National Institute of Public Health (NIJZ)
Health Insurance Institute of Slovenia (ZZZS)
Slovenian Hospitals performing arthroplasty procedures
Ministry of Health
Central Population Register (CRP)
Manufactiors, distributors and importers of endoprotetic materials

Figure 6.5. Context diagram for The National Arthroplasty registry of Slovenia (RES)

Example of using Use Case diagrams in Patient registries:

- *TNPCR–AERRO Central Cancer Registry Business Use Case Diagram http://www.cdc.gov/cancer/npcr/informatics/aerro2/hospitals/h_business_use_case_diagram.htm*
- *NPCR–AERRO Central Cancer Registry Operations Use Case Diagram http://www.cdc.gov/cancer/npcr/informatics/aerro2/pdf/c_operations_use_case_diagram.pdf*

A useful tutorial on creating use cases can be seen on Pace University
http://csis.pace.edu/~marchese/CS389/L9/Use%20Case%20Diagrams.pdf

6.5.5 Process Modelling

Process model design is an engineering technique and part of **Business process management** (BPM), a disciplined approach to business processes. To fully understand the purpose and the ability to use process management in patient registries creation we will first quickly introduce Business process management as a whole and then **Business process modelling** as an important step to model existing or future processes using **Business process modelling notation** (BPMN).

6.5.5.1 Introduction to Business process management

In 1990's the focus on **processes** instead of functions was introduced in organisations. (3) Michael Hammer was the originator of reengineering and the process enterprise changed forever how businesses do business (5).

Typically, organizations (also healthcare organizations) are structured into divisions and departments based on the functionality of each division and department (for example IT department). Each division or department performs its own specific tasks and determines its own competency. Therefore, this organization structure tends to create 'Silo Thinking', each department stands alone with less or no interaction with other departments within the same organization. Compared to the Silo thinking - **processes**, on the other hand, **cut across** these **functional silos**. Where different activities in a process require different skills, the process is likely to involve a number of people and departments.

Business process thinking in organizations introduced new methodologies focusing on business processes such as **Business process management (BPM)**.

BPM is defined by Association of Business Process Management Professionals as »a disciplined approach to identify, design, execute, document, monitor, control, and measure both automated and non-automated **business processes** to achieve consistent, targeted results consistent with an organization's strategic goals. BPM involves the deliberate, collaborative and increasingly technology-aided definition, improvement, innovation, and management of end-to-end business processes that drive business results, create value, and enable an organization to meet its business objectives with more agility." (4)

Business process is a sequence of tasks / activities which transform inputs to outputs, that is of value to the customer, performed by human or machine (for example purchase process).
Business Process Management defines **process lifecycle** in eight steps:

1. Identify the process
2. Model the process
3. Discuss, audit, review the process
4. *Automate the process
5. Implement the process
6. Track (measure) the process
7. Optimize the process
8. Dismiss the process

*Automate the process – this step is not mandatory

From the above mentioned steps we can see that Business process modelling is an important part of Business process management.

Further reading:

- Hammer, Champy: Reengineering the Corporation: A Manifesto for Business Revolution (5),
- Keen, Knapp: Every Manager's Guide to Business Processes (6)
- Keen: The Process Edge, Creating Value Where It Counts (7)

6.5.5.2 Introduction to Business process modelling

Process modelling is a widely-used engineering approach to determine and **describe existing processes and future process scenarios.** We can say that process modelling comprises all the activities that need to be undertaken to document a process.

Process modelling considers the following elements:

- Process scope
- Process start
- Process activities and their interconnections
- Process numbers (measurable figures of its activities: e.g. duration, number of staff, maximum load, etc.)
- Process KPI (key performance indicators)
- Process end
- Process connections with other processes

All these elements need to be documented before starting implementation (and automation).

For business process modelling there exist a lot of modelling techniques. Graphical/visual representation of processes with process diagrams is a common way to describe a process. There are many diagramming techniques and notations to model processes, but the mostly used are **event based process chain (EPC)** – process modelling language invented by prof. Scheer and colleges at the University of Saarland in 1992 (8) and **Business process management notation (BPMN)** from OMG.

Business process management notation (BPMN) is a de- facto standard notation widely used in BPM community (9). In the following subchapters we will present where business process modelling could be useful in patient registries development and introduce EPC and BPMN.

6.5.5.3 Business process modelling and patient registries

Patient registries can benefit from BPM:

A. in phase of planning PR
   - **to better understand current state** (see Ch. 6.1.3.1 Overview of Current State; Ch. 6.1.6 Defining the Scope of the Registry & Building a Registry Development Team , Ch. 6.1.7 Performing Stakeholder Engagement and Analysis) we can model the current processes
   - **to identify all data sources** for Patient registry (Ch. 6.4 Data sources for registries)
   - **to model new process for PR operation** (collecting data, processing and analysing data, reporting data)
   - **to model PR supporting processes** such us for example *Perform annual audit*

B. in phase of PR set-up
   - **to model processes to be automated** (Ch. 7
   - Chonoles, Michael Jesse and Scardt, *James A.: UML 2 for Dummies, Hiongry Minds 2003.*

1. *Introduction to OMG's Unified Modeling Language™ (UML®), Web page http://www.omg.org/gettingstarted/what_is_uml.htm (access 25th June 2014).*
2. Asbjørn Rolstadås (1995). "Business process modeling and reengineering". in: *Performance Management: A Business Process Benchmarking Approach*. p. 148-150.
3. What is BPM Anyway? Business Process Management Explained, http://www.bpminstitute.org/resources/articles/what-bpm-anyway-business-process-management-explained [accessed 10th May 2014].
4. Hammer, M., Champy, J. (1993, 2003). Reengineering the Corporation: A Manifesto for Business Revolution, New York : Harper Business, 1993, New York : HarperBusiness Essentials, cop. 2003
5. Keen P.G.W., Knapp E.M (1996). Every Manager's Guide to Business Processes, Harvard Business School Press, 1996
6. Keen, P.G.W.: The Process Edge, Creating Value Where It Counts, Harvard Business School Press, 1997
7. Event based process chain (EPC), Aris community web page http://www.ariscommunity.com/event-driven-process-chain [accessed 23rd May 2014].
8. Chinosi,Michele, BPMN: An introduction to the standard, Computer Standards & Interfaces, Volume 34, Issue 1, January 2012, Pages 124–134.
9. *Davids, R., Brabänder, E. (2007): Aris Design Platform, Getting Started with BPM, Springer-Verlag London Limited 2007*
10. White, Stephen A., Introduction to BPMN, IBM Corporation,link, http://www.omg.org/bpmn/Documents/OMG_BPMN_Tutorial.pdf [accessed 10th May 2014].
11. Gliklich RE, Dreyer NA, eds. Registries for evaluating patient outcomes: A User's Guide. 3rd ed.2014.
12. *Hernandez, Michael J.: Database Design for Mere Mortals™: A Hands-On Guide to Relational Database Design, Second Edition, Addison Wesley Professional, 2003*
13. Everest, G.: Database management : objectives, system functions, and administration, New York, McGraw-Hill, 1986
14. Riordan, R.: Designing Effective Database Systems, Addison Wesley Professional, 2005
15. IIBA International Institute of Business Analysis: Business Analysis Body of Knowledge (Babok Guide V2.0),

IIBA 2009, ISBN-13: 978-0-9811292-1-1

16. Craig Larman (1998) : Applying UML and Pattern, An Introduction to OO Analysis and Design Prentice-Hall Int London ISBN: 0-13-748880-7
17. Richard Barker: Case Method-Entity Relationship Modelling Oracle Corporation 1990 UK Limited
18. Vojislav Ivetić, Janko Kersnik: Diagnostične preiskave za vsakdanjo uporabo, Združenje zdravnikov družinske medicine 2007, ISBN: 978-961-91889-5-8
19. Rafael J. Curbelo · Estíbaliz Loza · Maria Jesús García de Yébenes · Loreto Carmona:Databases and registers: useful tools for research, no studies Rheumatol Int (2014) 34:447–452
20. Nadkarni PM[1], Marenco L.,Easing the transition between attribute-value databases and conventional databases for medical data. Proc AMIA Symp. 2001:483-7.
21. European Arthroplasty Register: http://www.ear.efort.org/registers.aspx access [15.12.2014]
22. OB Valdoltra Register Artroplastike http://www.ob-valdoltra.si/raziskovalna-dejavnost/publikacije-register/register-artroplastike-ob-valdoltra access [15.12.2014]
23. Shahar Y[1], Combi C.: Timing is everything. Time-oriented clinical information systems. West J Med. 1998 Feb; 168(2):105-13.
24. Rafael J. Curbelo · Estíbaliz Loza · Maria Jesús García de Yébenes · Loreto Carmona:Databases and registers: useful tools for research, no studies.Rheumatol Int (2014) 34:447–452
25. Richard T. Snodgrass: Developing Time-Oriented Database Applications in SQL. Morgan Kaufman Publishers, 2000, ISBN: 1-55860-436-7
26. Richard T. Snodgrass: TSQL2 and SQL3 Interactions. Association for Computing Machinery / Special interest group on management of data
27. Beal, Thomas (2002), Archetypes: Constraint-based Domain Models for Future-proof Information Systems
28. Tapuria, Archana and all (2013): Contribution of clinical Archetypes, and the Challenges, Towards Achieving Semantic Interoperability for EHR, Healthcare Informatics Research, December 2013
29. Costa, Catalina Martinez and all (2011): Clinical data interoperability based on archetype transformation.
30. Garde S, Knaup P, Hovenga EJS, Heard S (2007): towards Semantic Interoperability for Electronic Health Records: Domain Knowledge Governance for OpenEHR Archetypes. Methods of Information in Medicine 46(3): 332-343.
31. OpenEHR web page, http://www.openehr.org/, access [25. 06. 2014]
32. Leslie, Heather & all: OpenEHR archetypes in detail, PPT, Ocean Informatics, 2012
33. The experience of selecting the code systems for the development of the epSOS master value catalogue (MVC), September 2013
34. Guidelines on minimum/non- exhaustive patient summary dataset for electronic exchange in accordance with the cross-border directive 2011/24/EU, eHalthNetwork, November 2013 http://ec.europa.eu/health/ehealth/docs/guidelines_patient_summary_en.pdf access [1st June 2014].
35. Stellman, Andrew; Greene, Jennifer (2005). Applied Software Project Management. O'Reilly Media. ISBN 978-0-596-00948-9.

- PATIENT REGISTRY INFORMATION SYSTEM DEVELOPMENT AND IMPLEMENTATION) – to gather user requirements

C. in phase of running a PR
  - **to improve processes** (Ch. 8.2 Overarching Processes)

6.5.5.4 Event-driven process chain (EPC)

Event-driven process chain (EPC) is the main ARIS model for representing processes. It is dynamic model bringing together the static resources of the business (systems. organization, data, etc.) and organizing them to deliver a sequence of tasks or activities ('the process') that adds business value (*10*).

An event "activates" an activity and activity will always "create" one or more new events (Figure 6.6 Example of the EPC diagram).

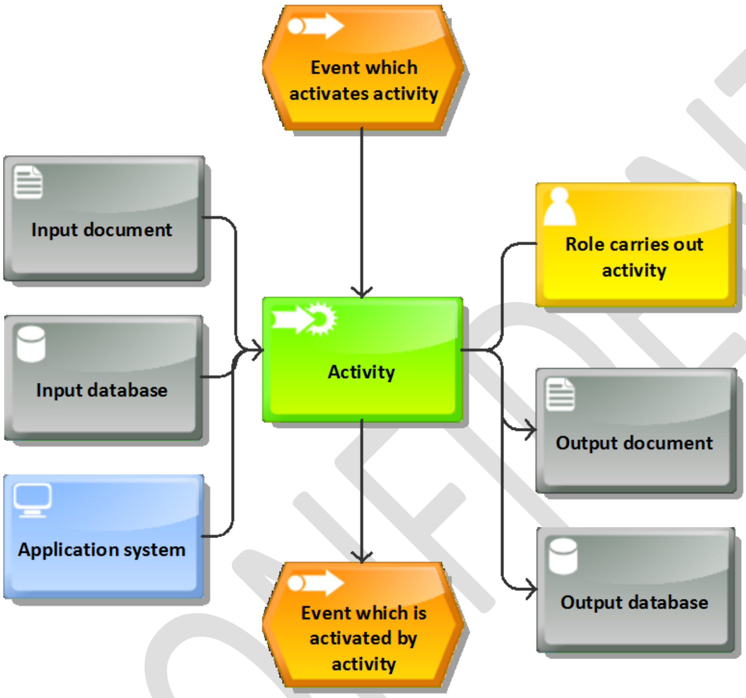

Event which activates activity
Input document
Role carries out activity
Input database
Activity
Output document
Application system
Output database
Event which is activated by activity

Figure 6.6 Example of the EPC diagram

6.5.5.4.1 The EPC Objects

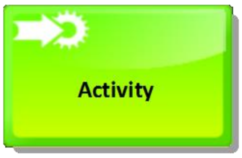
Essentially, there are four types of objects used in the EPC:

- Events,
- Functions,
- Rules,
- Resources (data, organisation, system, etc).

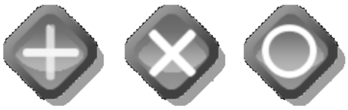
Event represent the changing state of the world as a process proceeds:

- External changes that trigger the start of the process
- Internal changes of state as the process proceeds
- The final outcome of the process that has an external effect

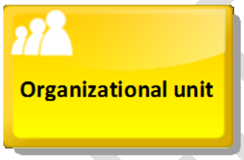
To describe events, we typically use the convention 'noun-verb'.



Functions / activities represent the activities or task carried out as part of a business process; ideally with each one adding some value to the business. Function max be carried by people or by IT systems. They have inputs (information or material), create outputs (different information or a product) and may consume resources.
To describe function/activity we use the convention 'verb-noun' or more specifically.

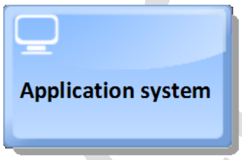


Rules: Real processes do not just consist of sequential steps. The need to cope with parallel paths, decisions, multiple triggers and complex flows is the reason we use modelling tools to represent processes. To model a process flow we add 'Rules' to the functions and events previous described. There are three basis types of rule: OR, XOR and AND.

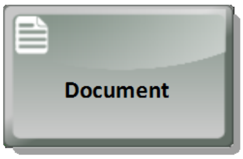
,


,

Organizational objects represent the people who perform the process tasks represented by functions/activities. We can represent specific people, departments, roles or teams, depending on the context and detail of the model.



Application Systems resources represent the computer and software applications used to support the business.



Activities have inputs (information or material) and create outputs (different information or a product). Usually they are data (in database) or documents.

Further reading:

- Aris online Academy http://cdn.ariscommunity.com/aris_online_academy/what_is_bpm3/50bfqndn/player.html
- Aris Architect & Aris Designer https://www.softwareag.com/corporate/images/SAG_ARIS_BusArchDesign_FS_4PG_Jan14-Web_tcm16-78556.pdf

6.5.5.5 Business process management notation (BPMN)

Introduction

Business process management notation (BPMN) principles originate from flowcharts, which were invented in 1946 by Goldstine and von Neuman. In the following years, many similar modelling techniques appeared. BPMN 1.0 was introduced in May 2004 by Business Process Management Initiative and was in 2005 acquired by Object Management Group (OMG). BPMN 2.0 was released in 2011 (11).

In the last ten years, BPMN has spread and has become the most widely used process modelling technique in the world, supported by the main IT companies.

BPMN is an open specification, therefore no royalty fees need to be paid.

BPMN is a standard for graphical modelling and also for transformation to execution model (XML code).

Where BPMN can be used?

BPMN has many fields of application. It can be used:

- To capture existing state of the process (AS-IS process);
- To gather requirements for the new information system (description of the system behaviour and interaction with the user);
- To optimize the process (describe the AS-IS and then TO BE process);
- To simulate behaviour using special software tools;
- To be automatically translated to execution mode (the program code), but it has to be defined in great detail.

6.5.5.5.1 BPMN Basics Elements

BPMN uses a set of graphical elements to define a **Business Process Diagram (BDP)**. BDP is a network of graphical objects which are activities and the flow controls that define their order of performance.

Using BPMN we can describe:

- **Processes and activities** (complex and atomic) - units of work (e.g.: project planning, product testing, floor cleaning, inputting user data, sending quote)
- **Events- impulses, conditions or business rules** that start or interrupt the process or activity (e.g.: receive document, Tuesday at 6:00, sales dropped below 20 units/month, out of stock)
- **States**

  - **Activity states**: e.g. the 'preparing order' activity can be in the following states: idle, starting, running, finished, interrupted
  - **Object states**: e.g. the ‘order' information object can be in the following states: prepared, draft, sent, deleted, archived
- **Decisions and conditions** that directly affect the flow of the process. Decisions can be used to split the process in two parallel or alternative branches or paths. (e.g.: is number of patients < 5, is purchase value > 500?)
- **Artefacts** - all objects (inputs, outputs) that are used within the process (e.g.: material, documents, information, user interfaces, reports, instructions, standards)
- **Roles, actors** - people, information systems or other organizations who perform the process activities (e.g.: employees, functions, information systems, databases, external systems)

BPMN has a small set of notation categories so the reader of a BDP can easily recognize the basic type of elements.

The four basic groups of elements are (8):

- Flow objects (activity, event, decision);
- Swimlanes (role, sub – role);
- Artefacts (documents, information) and
- Connecting Objects;

A. Flow objects are three core elements:

- **Activity -** represents performed work; we have two types of activity: **task** (atomic activity) and **sub- processes;** drawn as rounded corner rectangle with additions (+, II, loop)
- **Event -** represent what happens during the course of a business process, it affect the flow of the process; drawn as circle; we have 3 types of events:
  - **start event** (single circle), which is used on process start
  - **intermediate event** (double circle), which is used between the process start and end
  - **end event** (thick circle)
- **Gateway** – represents process decisions as well as the forking, merging and joining of paths; drawn as diamond shape, the centre of the shape represents the type of split; we distinguished 2 types of gateway:
  - (X) in the centre represents decision (choose only one process branch)
  - (+) in the centre represents parallel execution of all output branches

B. Swimlane in a diagramming technique is a mechanism to organize activities into separate visual categories. In BPMN swimlanes are representing **roles** (participants) in a process. Role is represented as a horizontal (or vertical) rectangle. Sub-roles or departments within the organization are represented as rectangles within another rectangle.

C. Documents and information are represented as a paper icon with a folded corner.

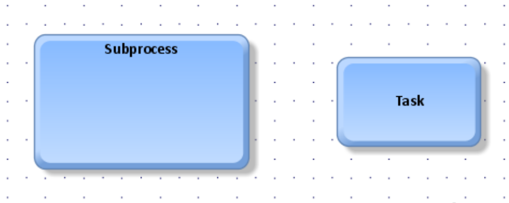
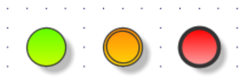
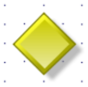
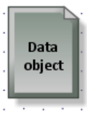
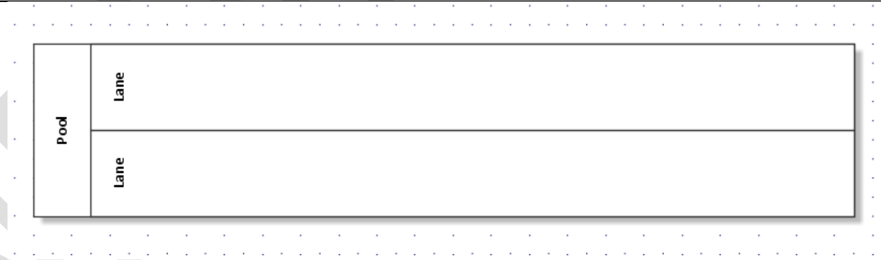
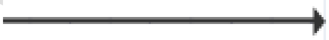
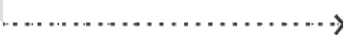
D. Process elements can be **connected** using:

- **Sequence flow** - this represents the logical sequence of activities, events and decisions (drawn as solid line with a solid arrowhead); connects activities, events and gateways

- **Information flow** - this element represents inputs/outputs of activities and triggers for message driven events. Usually, the document element is attached to the information flow. (drawn as dashed line with an open arrowhead)

Usually, sequence flows are drawn first; information flows are added later, in the process modelling phase.

Table 6.11. Basic elements of BPMN

Activity (Subprocess, Task)	
Event (Start, Intermediate, End)	
Gateway	
Documents	
Swimlane (Role)	
Sequence flow	
Information flow	

Additional BPMN elements

With the basic BPMN elements it is possible to model almost all real life scenarios. But there are additional types of elements:

- Specialized types of processes and tasks;
- Specialized types of events;
- Specialized types of gateways;
- Various information flow usage.

The various types of specialized processes, events, gateways and flow usage can be quickly viewed on Object Management Group - Business Process Model and Notation website http://www.bpmn.org under QuickGuide. Informative presentation of BPMN can be seen also on Camunda.org tutorial (http://camunda.org/bpmn/tutorial.html#tutorial).

It is recommended to download A1 Poster BPMN2_0_Poster.pdf (freely available on http://www.bpmn.org under Documents / BPMN 2.0 Poster). It is a very useful quick reference guide for using BPMN 2.0

The normative document on BPMN 2.0, can be viewed on http://www.omg.org/spec/BPMN/2.0/PDF.

6.5.5.5.2 Steps for graphical modelling of identified business process using BPMN

For creating graphical representation of the business process using BPMN we should follow the following steps (Figure 6.7)

1. Define roles
2. Define activities
3. Arrange activities in pools/lanes
4. Connect the activities using sequence flow
5. Add events
    - Start event (what triggers the process)
    - Intermediate events (pauses, exceptions)
6. Add documents, information and flows
    - Output activities, receiving events

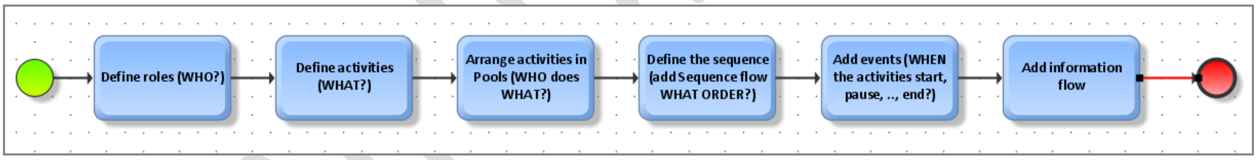

Define roles (WHO?)
Define activities (WHAT?)
Arrange activities in Pools (WHO does WHAT?)
Define the sequence (add Sequence flow WHAT ORDER?)
Add events (WHEN the activities start, pause, .., end?)
Add information flow

Figure 6.7: Steps for modelling graphical presentation of BPMN

Process model decomposition

Processes are usually not simple and their graphical models quickly become big, complex and unreadable. Therefore, they should be decomposed and not presented on a big A1 whiteboard. By modeling processes we should stick to the rule of 7-10, which means: maximum 10 activities per process model should be designed for maximum readability. If the process model contains events and gateways, this number should be lowered.

Two types of decomposition are recommended:

- **top down decomposition** and
- **role decomposition** or **pool focusing** (hiding other roles or pools).

Top-down or leveled process decomposition helps us to reduce cluttering on the process models. Different levels are also targeted for different audiences. For example, **level 1** should be read, defined and managed by top level management (CEO, general manager, quality manager, etc.) **Second level** should be targeted for process owners and performers, because it defines detailed activities, responsibilities, events, information flows etc. Usually 2 levels of detail are enough if the processes will be performed manually (it also depends on the complexity of processes). **Third level** usually contains technical details and it is targeted for developers, software architects, and performers.

Conceptual models (designed for business users) should not be mixed with technical process models, which include implementation details (designed for technical experts, software designers and programmers).

Pool focusing is the horizontal way of detailing processes. The main idea is to design details (activities, events, gateways etc.) in ONE pool and to represent other pools as black-boxes (without any details).

6.5.5.6 Documenting business processes

Usually, the process diagram alone is not enough to fully present and completely describe business process. There are some process properties like process scope, process goals, metrics of the process, etc., which can't be represented graphically. Therefore additional explanations are needed in text documents.

Documenting the processes includes:

- graphical presentation, where the flow of a process is presented using some graphical notation, and
- textual presentation providing more detailed descriptions of the process and document templates.

The **Process description document** is a document which may be provided for each process defined. Its creation is not mandatory, but it is strongly suggested to do so in order to provide the reference information about the process.

The content of the Process Description Document should define at least:

- aims of the process, objectives
- graphical representation in selected modelling notation (like BPMN), showing the sequence of activities, roles involved, documents used, etc.
- descriptions of activities providing more information about each activity defined in the graphical representation
- key performance indicators - defining how we can evaluate if the work has been performed correctly and efficiently,
- references to other processes; information which processes provide inputs to the process and which use the outputs of the process.

6.5.5.7 BPM tools

There a lot of BPM tools and platforms. We can group BPM tools by license type into 3 groups:

1. Open source, freely accessible tools: Intalio, Bonita
2. Free, but not open source: Aris Express

3. Commercial: Signavio BPM, Oryx BPM, Appian, RunMyProcess (Google Apps platform), ActiveVos – Socrates, ARIS, MS Visio, Lombardi Teamworks, Pegasystems BPMS etc.

6.5.6 Data modelling (using E-R diagram)

As we could see in the previous chapters there are many slightly different definitions of patient registries (for example see definitions listed in AHRQ – Registries for Evaluating Patient Outcomes: A User's Guide (12), p.35)., but almost all of them say that patient registries are "**an organized systems for the collection, processing and** storage of uniform health **data** on individual persons **in a systematic way for specific and defined purpose**."

According to Hernandez (13) we can call any data collected in systematic way and for specific purpose regardless of the collection method (electronically, paper-based) a database. From this definition we can conclude that the main ingredients of the patient registries are uniform health data about individual persons organised in **a database** (DB).

What is a database?

"A database is a collection of related data." (14)

"The database is a tool for efficient storage and manipulation of data." (15)
"A database is a collection of data that is used to model the organization or organizational process. It does not matter whether it is used for a computer program or it is on paper. As long as the data is collected and organized for a specific purpose, we have a database. "(13)

What are data?

Data are a representation of facts, concepts and instructions presented in a formalized manner suitable for communication, interpretation, or processing by humans or by automatic means. (ANSI, ISO)

"Data are facts presented by the values (numbers, signs, symbols) that have meaning in a particular context." (15)

"The data are static values stored in the database." (13)

What is information?

"Information is quantified data in a specific situation." (14)
"Information is data that is processed in such a way that meets the needs of the individual." (13)

6.5.6.1 Types of databases

Databases can be roughly divided in 2 categories: **operational databases** and **analytical databases**. First type – the **operational databases** are used primarily in transactional systems (**OLTP - On-Line Transaction Processing**), which are mainly intended for daily collection, modification and maintenance of data. The data stored in these databases are dynamic, which means that they change frequently. Operational databases always show the current status. An example of such systems is for example ATM.

The second type – the **analytical databases** are primary used in analytical systems (OLAP – On-Line Analytical Processing). In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas (usually star schema). This type of data is used for example in decision support systems, to analyse trends etc.

OLPT and OLAP are complementing technologies. OLTP runs your business day by day and analytical databases usually use data from operational databases as a main source. Both types of databases meet the specific tasks of data processing and therefore their development requires a different data modelling approach.

In the following section we will explore only modelling the first type of databases.

6.5.6.2 Data Model (E-R diagram)

Data modelling is originally part of software engineering discipline. The output of the data modelling is **data model** – a representation of a real world situation about which data is to be collected and stored in a database. A data model depicts logical relationships among different data elements.

There are a lot of different techniques of data modelling but we will focus on the most widespread technique – we will use **Entity Relationship Diagram** (ER diagram) to demonstrate data structure.

The ER model was introduced by Peter Pin Shan Chen in 1976 as a conceptual modelling approach that views real world data as systems of **entities** and **relationships**.
With ER diagram we can describe any system but E-R diagrams are most often associated with modelling databases that are used in software engineering. In particular, E-R diagrams are frequently used during the design stage of a development process in order to identify different system elements and their relationships with each other. **In patient registry creation we can use E-R modelling to identify required data elements** (See Ch. 6.3 Registry dataset) **and structure it properly** (to place data elements in a prominent and logical position).

E-R diagrams are very useful tool for data modelling and **visual presentation of data model**. They are **easy to understand** and do not require a person to undergo extensive training to be able to work with it efficiently and accurately. This means that it can be easily used in a communication among team members, developers and end users, regardless of their IT proficiency. E-R diagrams are also readily translatable into relational tables which can be used to quickly build databases.

Elements of E-R diagram

An E-R diagram is a visual presentation of data with the following elements: entities, attributes and relationships.

An **entity** is a thing (material or nonmaterial) that is relevant to a given system and on which the system must store data. It has to be recognized as being capable of an independent existence and which can be uniquely identified. It may be a **physical object or subject** such as patient or medical device, **an event** such as medical appointment, **a concept** such as an order. *For example, a patient registry may include entities: patients, diagnoses, interventions, outcomes, etc.* Entities are represented in ER diagrams by a rectangle and named using singular nouns (e.g., Patient).

Figure 6.8. Representation of an entity Patient

An **attribute** is a property, trait, or characteristic of an entity or relationship. The attributes describe entity or relationship. Attributes are named using singular names (e.g. Patient Name) and are represented in original notation by oval shapes, but in many other notations as a list inside the entity rectangle.

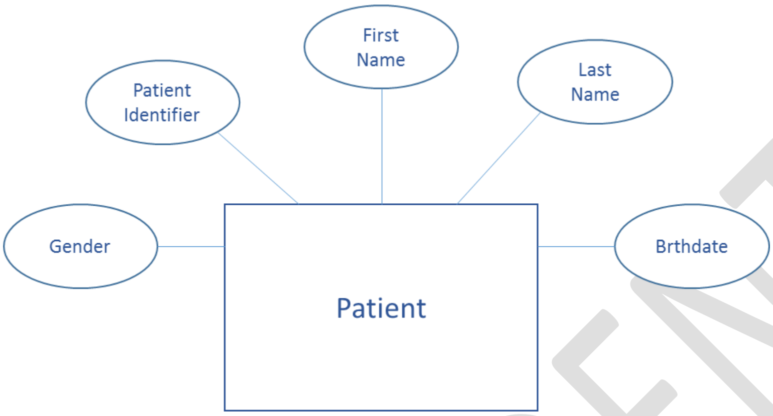


Figure 6.9. Representation of entity Patient with its attributes in original (Chen) notation

A **relationship** represents the interaction between the two entities. The phrase »The patient has diagnoses. » tells us that there is a relationship between the entities Patient and Diagnose. Usually, the relationships are binary (between two entities), but they can also be ternary (three entities), etc. Specific types of relationships are recursive relationships (entity is in relationship with it selves –also called self-referential relationship). Typical of such a relationship is: "Each employee can be a leader and has a leader. «

Relationships are in original notation represented by diamond shapes and are labelled using verbs. Usually they can be read in both directions, for example:

- »The Patient has diagnoses. «
- »The Diagnose is assigned to the Patient. «

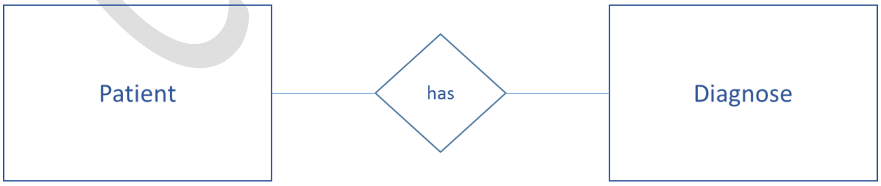


Figure 6.10. Representation of a relationship between entity Patient and Diagnose in original notation

Cardinality further defines relationships between entities by placing the relationship in the context of numbers. It depends on the rules in a modelled system. To define cardinality of the relationship we have to ask how often the entity occurs in conjunction with another entity. We are looking for a ceiling (the maximum number). For example we could ask: “How many diagnoses can have a patient?” and then check also the other direction: “How many patients can have a diagnose?”

There are relationships with different cardinalities :
- One-to- One (1 : 1),
- One -to-Many (1 : m), and
- Many-to- Many (m : n).

Hint: More on E-R diagramming technique can be read on ER Diagram tutorial website http://creately.com/blog/diagrams/er-diagrams-tutorial/ .

6.5.6.3 Building a Data Model

Data modelling is divided in 3 main activities:

- developing **conceptual data model**;
- developing **logical data model** and
- developing **physical data model**.

Conceptual data model is an abstract representation of problem domain. **Logical data model** describes the data in as much detail as possible. It is bound to the selected type of data model (for example relational, hierarchical, object – relational) but without regard to physical implementation. **Physical data model** is bounded to selected implementation platform and will not be presented in this document.

Bellow we will explore basics of conceptual modelling and present very important part of data models – data dictionary.

6.5.6.4 Conceptual data modelling and conceptual data model

A conceptual data model describes problem domain. The result is a general and abstract description of reality which helps team members to understand the data requirements.

Development of a conceptual model is usually divided into three phases (13):

1. Requirements gathering;
2. Designing E-R diagram;
3. Normalization.

Requirements gathering

To produce efficient data model we need to document and understand the requirements of the problem domain. At this stage, we need to define the **aims and objectives** of the Patient registry (See *6.1.1 Defining the Purpose, Objectives and Outputs of the Registry*) and also impose **limits on our system** (see *6.1.3 Defining the Scope of the Registry*), **review and evaluate the existing system of collection, storage and use of data and analyse the current operating environment** (See *6.1.2 Overview of the Current State*) and **predict future requirements**.

The outputs of this phase are:

1. Defined Mission statement and Goals of PR (see 6.1.1.)
2. Defined scope of the PR (see 6.1.3)

3. Preliminary list of entities (concepts/items/subjects about which we will store the data - e.g. Person) and Preliminary list of attributes (properties about which we will store data – e.g. Persons name, Persons address) (See 6.3)

Designing E-R diagram

Designing an E-R diagram can be divided into 3 parts:

1. identification of entities;
2. identification of attributes for each entity and
3. identification of the relationships between the entities.

The output of this phase is **graphical representation of data model** and **short descriptions of all identified entities**.

1. Identification of entities

At this stage, we will prepare a list of entities that will be used in the new PR.

1. First, on the basis of a preliminary list of attributes defined in the previous phase, identify entities.
2. Next, compare the obtained list of entities with the preliminary list of entities developed in the previous phase.
3. Compare the list of entities with the mission objectives for the database.
4. Add a description of the type of entity - a precise definition of the entity and why it is important for the patient registry project.

Rules to define name of an entity:

1. The name should be unique, descriptive name that is understandable by all the PR stakeholders.
2. The name should accurately, clearly and unequivocally identify the entity.
3. Use the minimum number of words that are necessary to describe the subject.
4. Do not use words that describe the physical characteristics (e.g., table, file format).
5. Do not use abbreviations or acronyms.
6. Do not use names that implicitly or explicitly identify more than one entity.
7. Use singular nouns.

2. Identification of attributes

At this stage, we will **add the attributes to the entities** from the preliminary attribute list and we will **define the type of attributes** (key, non-key).
Keys are special type of attributes and are extremely important because they:

- ensure that each record in the entity is accurately (unique) identified[90];
- ensure different levels of integrity;
- allow the creation of relationships between the entities.

3. Identification of the relationships between the entities.

[90] For example in the entity Patient, PIX (Patient Index) can serve as a primary key of this entity.

Identified entities have relationships between them. Relationships are an important part of the E-R diagram.

There are 3 types of relationships with different cardinalities:

- relationship one-to-one (1: 1);
- relationship one-to-many (1: m) and
- relationship many-to-many (m: n).

In the conceptual model we does not resolve the many-to-many relationships.

Normalization

Database normalization is the process of designing database with the desired properties. This optimizes the management of the database by eliminating redundant (duplicate) data and ensuring that only relevant data is stored. Normalization is based on so called normal forms and rules for their creation.

The outputs of the normalization process are **refined E-R model** and **refined descriptions of the entities**.

6.5.6.5 Logical data modelling and logical data model

A logical data model describes the data in as much detail as possible. It is bound to the selected type of data model (for example relational, hierarchical, object – relational) but without regard to physical implementation.

6.5.6.6 Data dictionary

An important part (especially for interoperability purposes) of the data model is also a **Data Dictionary** where all data elements (entities and attributes) are well defined. Typical description of data elements (or metadata[91] of data elements) includes (but it is not limited to this):

- Identification of the data element (name, short name, alias, ID)
- Definition of data element type (entity or attribute)
- Definition[92] of data element, where also clear purpose of the data element is described
- Logical representation of data element (Value set[93], permitted values, default values, data type etc.).

Data dictionaries are very often called also metadata repositories. Well known and also in health often used standard for metadata repositories is **ISO /IEC 11179 Metadata registries.** More on metadata registries can be found on the ISO/IEC 11179 website http://metadata-standards.org/11179 . Good example of metadata registry defined according to ISO/IEC 11179 is METeOR http://meteor.aihw.gov.au/content/index.phtml/itemId/181162 - Australian Health Metadata Registry.

[91] Data about data.

[92] Short and helpful guidance on data elements definition could be find in chapter 6.3 ‘Registry dataset’.

[93] See also Chapter 6.5.10 ‘The importance of terminologies and code lists’.

6.5.7 Entity-Attribute-Value Data Model in Medical databases

Widely used data modelling/design technique in clinical databases and clinical data repositories is Entity-Attribute-Value (EAV) design. Background component of EAV design is representing arbitrary information on some object as Attribute-Value list. An example of such representation in medical device implant database description would be:

Table 6.12. Units of implants as attribute-value lists

Hip Implant Unit	Ref no	Lot no
Acetabular Cup	9998-00-756	2582612
Inlay	8834-01-453	356321
Femoral Head	5632-01-234	234764
Femoral Stem	2345-03-234	234567

In Table 6.12 we have 3 lists with 4 attribute-value pairs. Conventional columnar data model is represented on Figure 6.11.

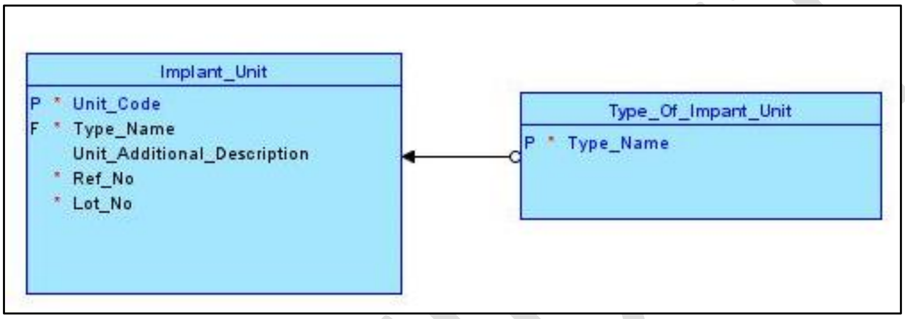


Figure 6.11. Example of conventional relational model with columnar form of attributes

Syntaxes of extensible markup languages like XML (16) are related to attribute-value pairs. XML elements that are delimited in open/close tags represent entities or attributes of entities.

However there are limitations in traditional relational oriented database modeling.
The conventional way to represent attributes for a class/entity is relational database modeling is in the form of columns in a table. That means one column per attribute:

- Hip Implant Component,
- Ref No,
- Lot No.

This approach is suitable when:

- there is fix number of attributes describing class or entity
- most or all attributes have values for given instance of class.

Such columnar representation is not optimal for classes:

- with potentially large number of attributes for given instance and
- there may be several instances with attributes having unknown or inapplicable value

This situation is analogy for computer science sparse-matrix problem using the term sparse data. The term sparse data denotes situation where there is discrepancy between numbers of potential versus actual attributes. It should be also considered that mainstream relational database engines are usually limited up to 1024 columns per table. And it is not rational strategy for partitioning such enormous set of attributes across several tables. Also volatility of data model should be considered, since number of attributes (parameters) continually increases as medical knowledge and standards advances. This requires continual modifications to the schema and user interface as well.

In following two Figures is represented case of transforming conventional columnar model for subset of attributes into EAV row level model.

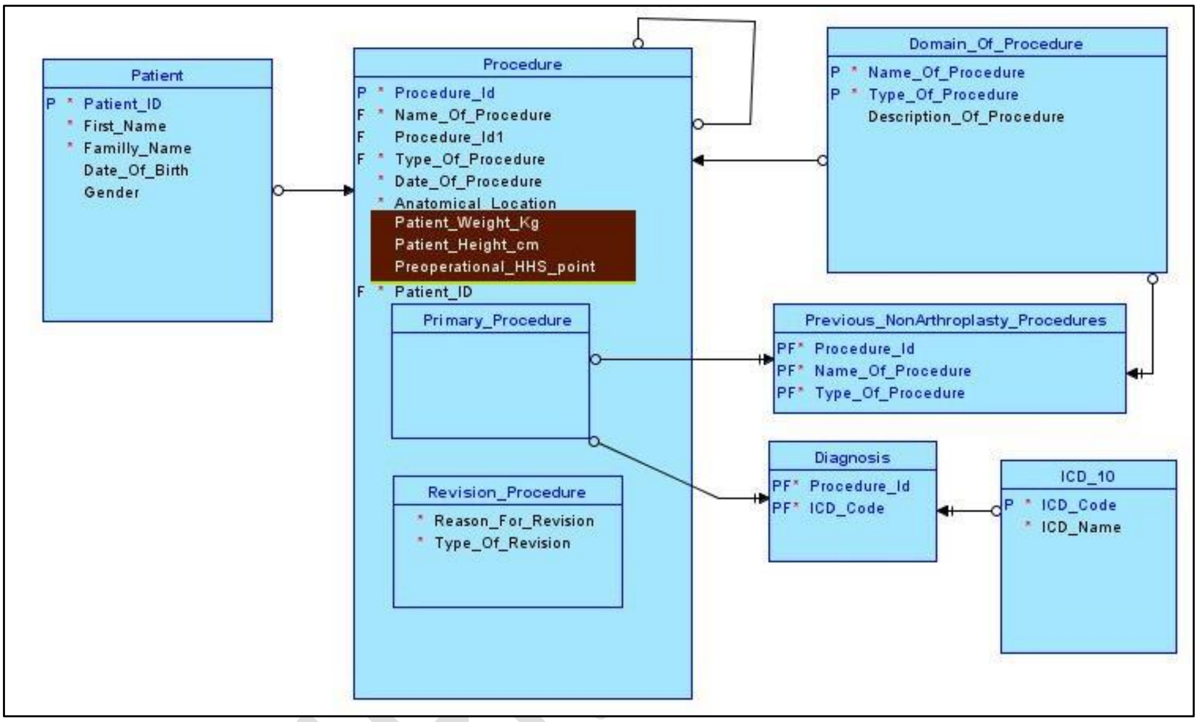


Figure 6.12. Example of conventional relational model before transformation to EAV entities

In Figure 6.12 is represented part of relational data model with entities that have conventional columnar presentation of attributes. However in entity “Procedure” there are 3 attributes (brown rectangle) with allowed “unknown” values:

- Patient_Weight_kg,
- Patient_Height_cm
- Preoperational_HHS_point – orthopaedic measure as assessment of patient’s hip disabilities

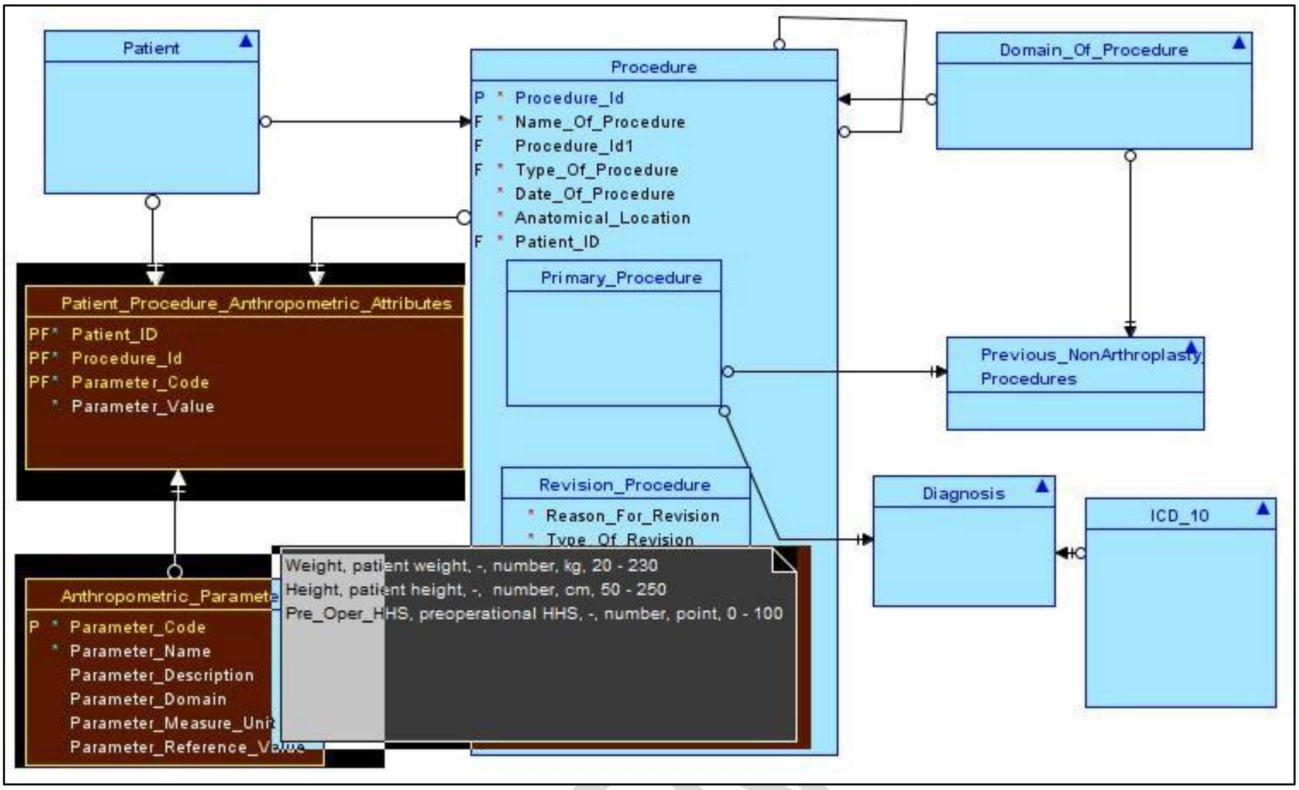


Figure 6.13. Data model with EAV entities

All attributes represent anthropometric measures of the patient. As such they are candidates for separated set of attributes that might be expandable with many other measures. Let us transform columnar representations of these attributes into model that considers EAV modelling/design technique. The result model is presented on Figure 6.13.

Model seems more complex at first glance. But provides more flexibility in management of changes on schema. Basis of the transformation to EAV is incorporation of the metadata describing nature of the attribute needed: business rules, data type, unit of measure.

6.5.8 Temporal modeling in medical databases

According to (24) three various type of databases are specific for medical database systems:

- **Administrative databases** that serves as operational support in organizational and economic aspect of information technology support: ERP accounting, ERP assets, ERP Human Resources, CRM.
  Databases in this realm are not prepared for clinical questions although there supposed to be connections (Patients, Organisation structure with Personnel, Accounting details) to other realms where more clinical orientation of data is structured. Database structure is complex and demands detailed knowledge of transactions to use the data for reference or analyzing.

- **Clinical databases** for tracking procedures and services, electronic patient records, more granular details of medical devices, cases, diagnoses. As such they represent daily operational tool in:
  - planning therapies, procedures,

- occupancy or availability of medical assets or human resources like therapists, physicians).

Clinical database can also serves as clinical research infrastructure. Data must be tracked in time with additional business rules on events (individually or on sequence of events in workflow) that control consistency of data as time series.

- **Registers** like disease registers, drug registers, incidents. Registers connect various sources resulting spatial-time oriented databases spanned above data clusters like patient, disease, medical device, physician, geo-location, date-time. Data from registers is serving risk-management processes, trend and survival statistical analysis, incident modeling. The data coming into registers is based on operational data in former two types of databases with additional data retrieved from quality assurance, data consolidation processes.

All three types of databases are included in consolidated data subsets (as data mart or data warehouse pivoted form) used in researches, cohort studies, data control in operational data tables with time as mandatory essential dimension.

Success of the process for data consolidation is significantly dependent on timing data consistency and clear simple presentation of time-dependency.

Technique in conventional modeling that considers TIME as entity specific it is called TEMPORAL DATA MODELING. And database that involves time constraints and controls is called TEMPORAL ORIENTED DATABASE.

In (25) are defined two main concepts of modeling TIME as attribute in relational or object model:

1. Time instants concept which considers time dependent entities (temporal entities) as series of events (changes) forming log/track or time series of changes (transactions).
   Records in such time series might be additive meaning we can perform aggregating functions like sum, count on different grouping criterias. Maintenance of time series might be more flexible. To provide consistency the rule of sequencing (linear orderliness) must be obeyed between various events in the sequence. Time attribute is represented as Dat_Time_Of_Change.

2. Time intervals concepts relating to the situations/states/statuses of the entity/object for specified time interval defined with upper (Date_Time_To) and lower limit (Date_Time_From).

   Both concepts are derivable from each other with pivoting techniques. But we must be aware that in second concepts other time dependent attributes cannot be additive. So aggregating techniques usually requires transformations of data of data.

Business rules on time attributes must consider to control risks of time interval overlapping of the same categories of data (temporal entities).

In the following example we present both concepts of modeling on the example of ER diagram in register database of patients, implants and procedures which are performed for implant insertions or removal by physician.

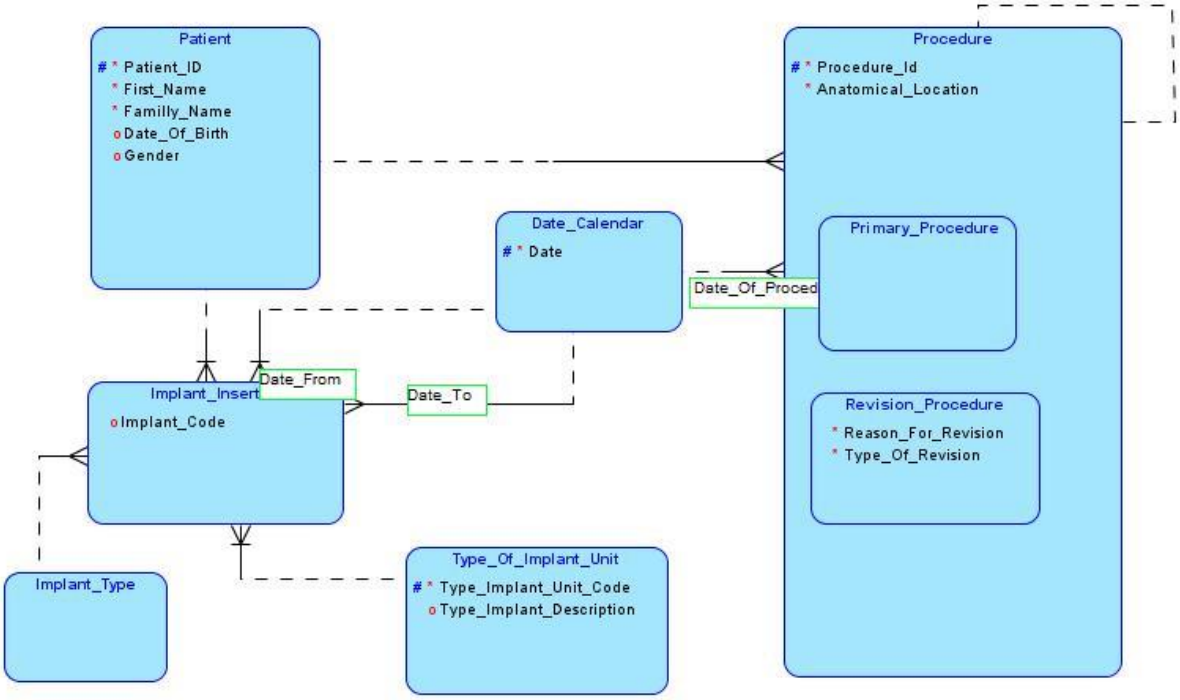


Figure 6.14. Data model with time interval temporal entities

In this example complete status of the implant that patient has is represented with time interval temporal entity Implant Inserted. Such concept is providing us complete history of all implant unit and how they form implant in patient. But time series of procedures and time implants are independent. So consolidation of these two time series must be performed through time interval operators for comparation, compliance of date of procedure with proper time interval of the implant unit.

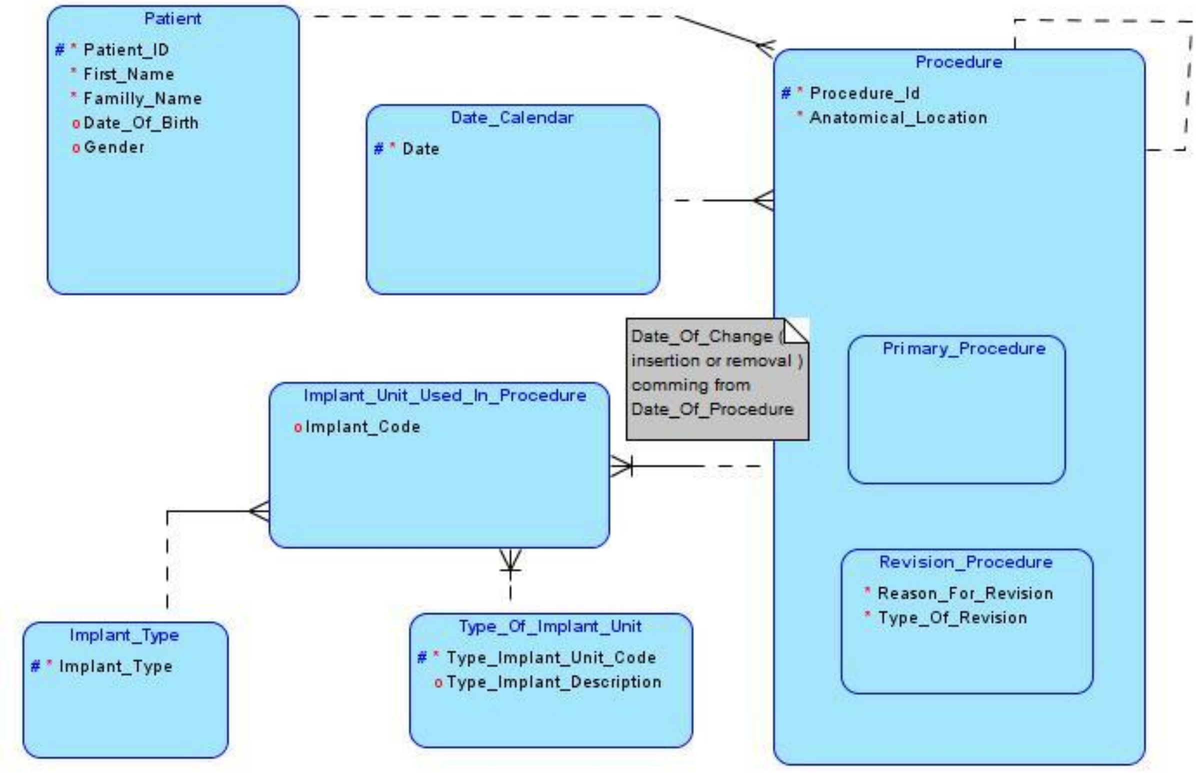


Figure 6.15. Time instant concept of time modeling

In second example we use time instant concept for temporal entity Implant_Unit_Used_In_Procedure. This way we track changes or transactions (insertions, removals) for each individual unit of implant. Time or Date of change is defined with Date_Of_Procedure. So Implant_Unit_Used is indirectly temporal dependent. The model is much more flexible and clear for performing business rules. But requires different query technique for retrieving consolidated implant (all units) in every moment of time.

6.5.9 Knowledge management using archetypes

Beal (28) states that the **health domain** is open-ended and there is a huge number of constantly changing concepts. Therefore he proposed instead of “classical” **single level approach** in which both information and knowledge concepts are hard-coded directly in software and database models, a **dual level approach** where information systems are built from information model only, and driven at runtime by knowledge-level concept definitions, known also as **archetypes**.

Nowadays health information systems are built using both approaches. But due to increased need for sharing of patient information across multiple settings and within diverse electronic health record repositories, **dual-level approach is in literature recognized as promising solution to ensure electronic health records interoperability** (28) (29) (30) (31). Therefore also PARENT is using this approach in defining new Artroplasty registry as a role model.

In the next subchapters we will present dual level approach methodology with emphasis on archetypes modelling using Open EHR.

6.5.9.1 OpenEHR

OpenEHR is a virtual community working on interoperability and computability in e-health. Its main focus is electronic patient records (EHRs) and systems (32). The success of openEHR is in no small part due to the formal acceptance of **CEN 13606** as a European and ISO standard. This standard is based on many aspects of the openEHR design approach, and part 2 of the standard is a snapshot of the openEHR Archetype specifications (32).

OpenEHR is based on multi-level modelling approach where clinical models (**archetypes**) built by domain experts are separated from data representation and sharing (**reference model**). The semantic architecture of OpenEHR is depicted on the picture (Figure 6.16) and in the following lines we will present the core elements of Open EHR.

Reference model is mostly technical infrastructure – generic technical artefacts for representing helath information (data structures and types, health record organisation, security etc.) and is hidden from content modelling for clinicians.

Archetypes are standardized computable models of discrete clinical concepts, for example: blood pressure, symptom, medication order, family history, Brest cancer histopathology result. They should capture as many clinical perspectives as possible to be universally applicable, but can be designed also for specific local use case.

Templates are used to create datasets as for example for discharge summary, Arthroplasty register etc. With templates we define data entry definitions and message definitions for a particular clinical context or purpose. Templates are an aggregation of archetypes according to specific use-case. In templates we constrain archetypes to make them practical– e. g. we remove unwanted items, set default values, bind the terminology etc.

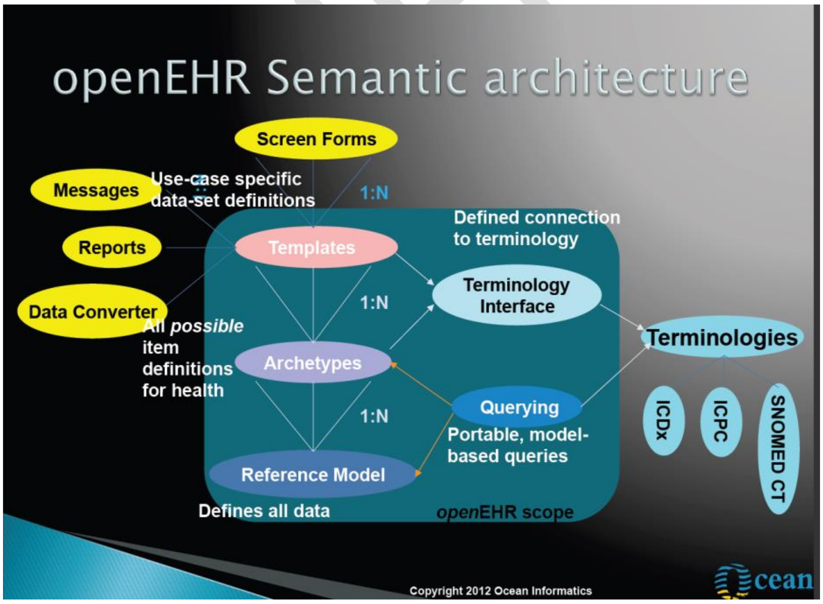


Figure 6.16. OpenEHR Semantic architecture (Beale, Thomas: Architecture Overview, Ocean Informatics, 2012)

6.5.9.2 Modelling Archetypes

To capture clinical knowledge we have to model (clinical) archetypes.

It is recommended that the archetype design is led by a person with experience in archetype modelling.

Process of modelling archetypes (33):

1. **Identify all clinical concepts**
2. **Explore if archetypes for the identified concepts already exists**

If yes: **3.a Use existing archetypes**
If no: **3.b Design new archetypes**

&1 Identify all clinical concepts

We have to research the subject, activity or task we want to model. For simple concepts as for example *body weight* we have to design one archetype, but for more complex concepts (for example *pregnancy*) we have to model multiple archetypes.

To visualize the researched subject, activity or task it is recommended to use some kind of Mindmap tool (for example XMind http://www.xmind.net/).

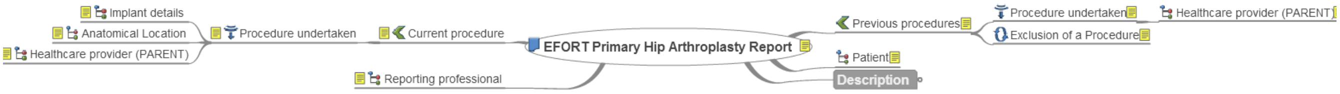


Figure 6.17. Example of EFORT Primary Hip Arthroplasty Report presented as a Mindmap

&2Explore if archetypes for the identified concepts already exists

Search for already defined archetypes in:

- openEHR CKM http://www.openehr.org/ckm/
- NEHTA CKM http://dcm.nehta.org.au/ckm/
- NHS archetypes http://www.connectingforhealth.nhs.uk/systemsandservices/clinrecords/nccr
- etc.

&3 Design new archetype

Designing a new archetype can be done in the following steps (33):

1. Gather content
2. Organise the content
3. Choose the archetype class
4. Build the archetype
    a. Name the archetype
    b. Select the structure

    c. Add data types
    d. Add constraints
    e. Add metadata
    f. Add terminology

5. Collaborate & Publish
6. Add to a Template

Gather content

Consider the clinical concept from all possible angles and point of views. Think about how different clinicians may record the identified data.

It is recommended to use guided interview technique (see chapters 6.5.2.1 and 6.5.2.2) or interactive modelling workshop (using mind map tool) to gather information on clinical concept. There are a lot of useful sources of information to model clinical content such as: different forms (also in paper), existing computer applications, clinical audit datasets, clinical trials datasets, patient registry datasets, reporting obligations etc. Look for similar projects locally and internationally, search for publications on the identified topic etc. To be as broad as possible research the clinical concept from different perspective: different medical specializations, nursing, researchers, public health, clinical decision support etc.

Organise the content

Organize the content using mindmap. Focus on identifying:

- Purpose – container or navigation
- Context
- Data elements
- Protocol
- State – context for interpretation
- Events
- Pathway steps
- Concepts needing coding/terminology

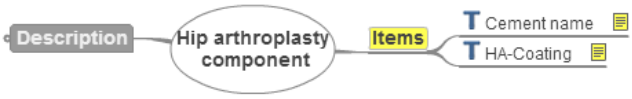
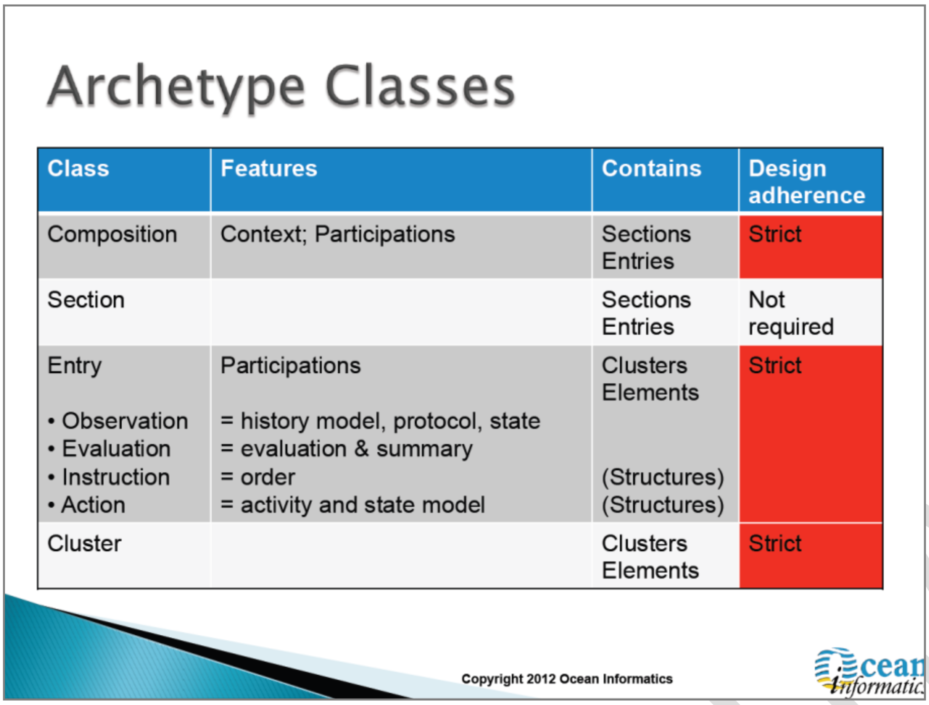


Figure 6.18. Example of newly defined Hip arthroplasty component archetype as a mindmap

Choose the archetype class

Although domain experts does not need to care about openEHR reference model (RM), the author of the archetype should know the possible “archetype types” called **archetype classess** presented in the RM.

As we can see (see Figure 6.19) the existing archetypes classes are: Composition, Section, Entry and Cluster. More on this topic can be read in OpenERH wiki. (www.openerh.org/wiki)

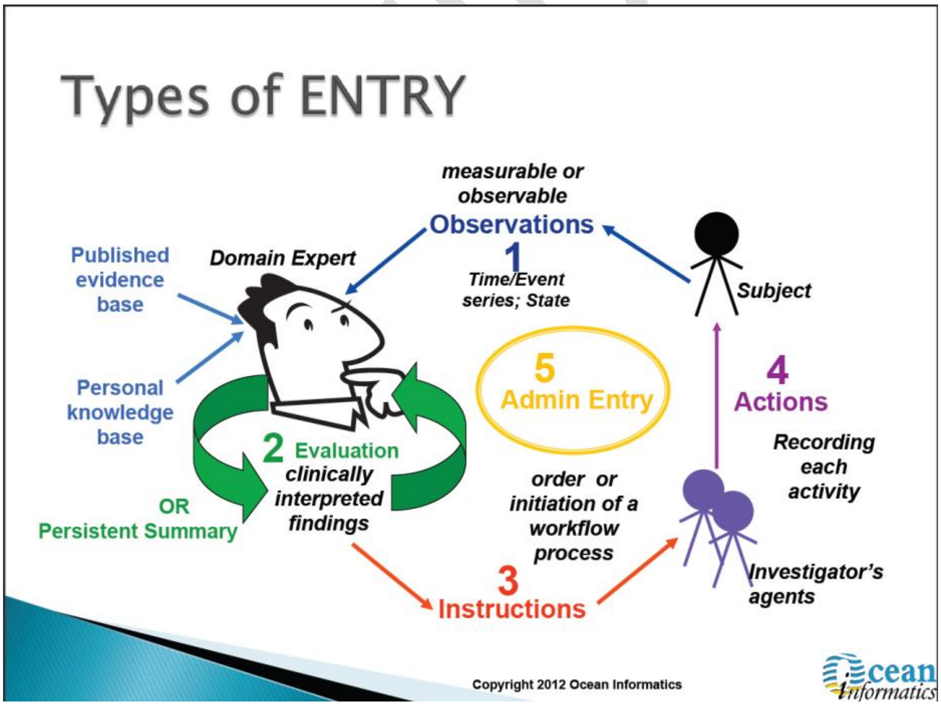


Archetype Classes

Class	Features	Contains	Design adherence
Composition	Context; Participations	Sections Entries	Strict
Section		Sections Entries	Not required
Entry • Observation • Evaluation • Instruction • Action	Participations = history model, protocol, state = evaluation & summary = order = activity and state model	Clusters Elements (Structures) (Structures)	Strict
Cluster		Clusters Elements	Strict

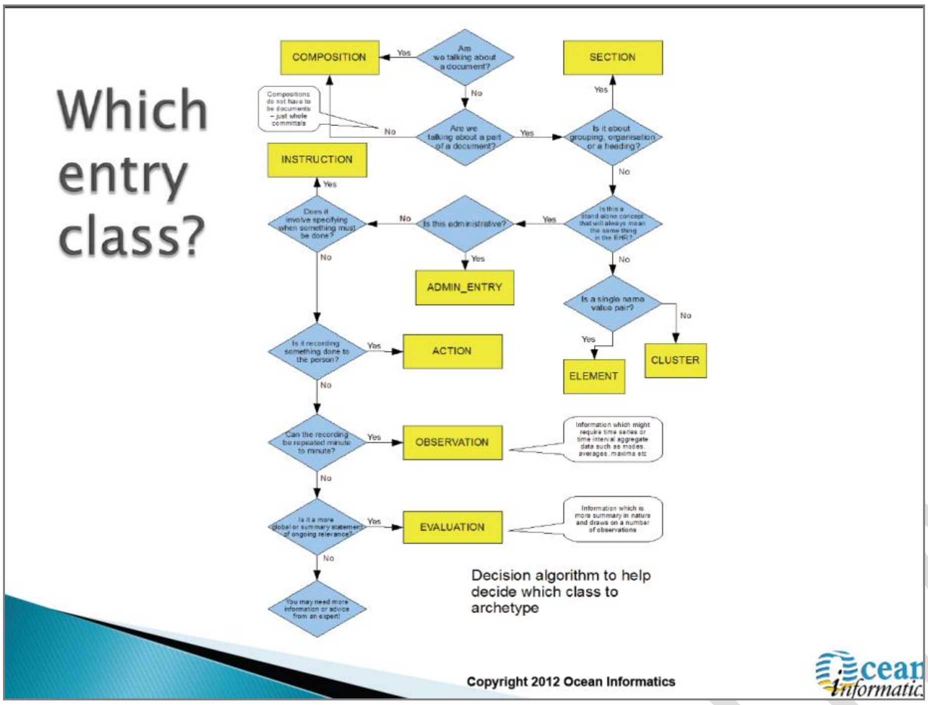
Figure 6.19. Types of ENTRY Archetype class (Leslie, Heather & all: OpenEHR archetypes in detail, Ocean Informatics, 2012)

Figure 6.20 demonstrate the different types of ENTRY class. To select the proper archetype class we can use tool presented by Ocean Informatics (
Figure 6.21)



Types of ENTRY

Figure 6.20. Types of ENTRY Archetype class (Leslie, Heather & all: OpenEHR archetypes in detail, Ocean Informatics, 2012)



Which entry class?

COMPOSITION
Yes
Are we talking about a document?
SECTION
No
Compositions do not have to be documents – just whole committals
No
Are we talking about a part of a document?
Yes
Is it about grouping, organisation or a heading?
Yes
INSTRUCTION
No
Yes
Does it involve specifying when something must be done?
No
Is this administrative?
Yes
Is this a stand alone concept that will always mean the same thing in the EHR?
No
Yes
No
ADMIN_ENTRY
Is a single name value pair?
No
Yes
CLUSTER
ELEMENT
Is it recording something done to the person?
Yes
ACTION
No
Can the recording be repeated minute to minute?
Yes
OBSERVATION
Information which might require time series or time interval aggregate data such as modes, averages, maxima etc
No
Is it a more global or summary statement of ongoing relevance?
Yes
EVALUATION
Information which is more summary in nature and draws on a number of observations
No
You may need more information or advice from an expert!

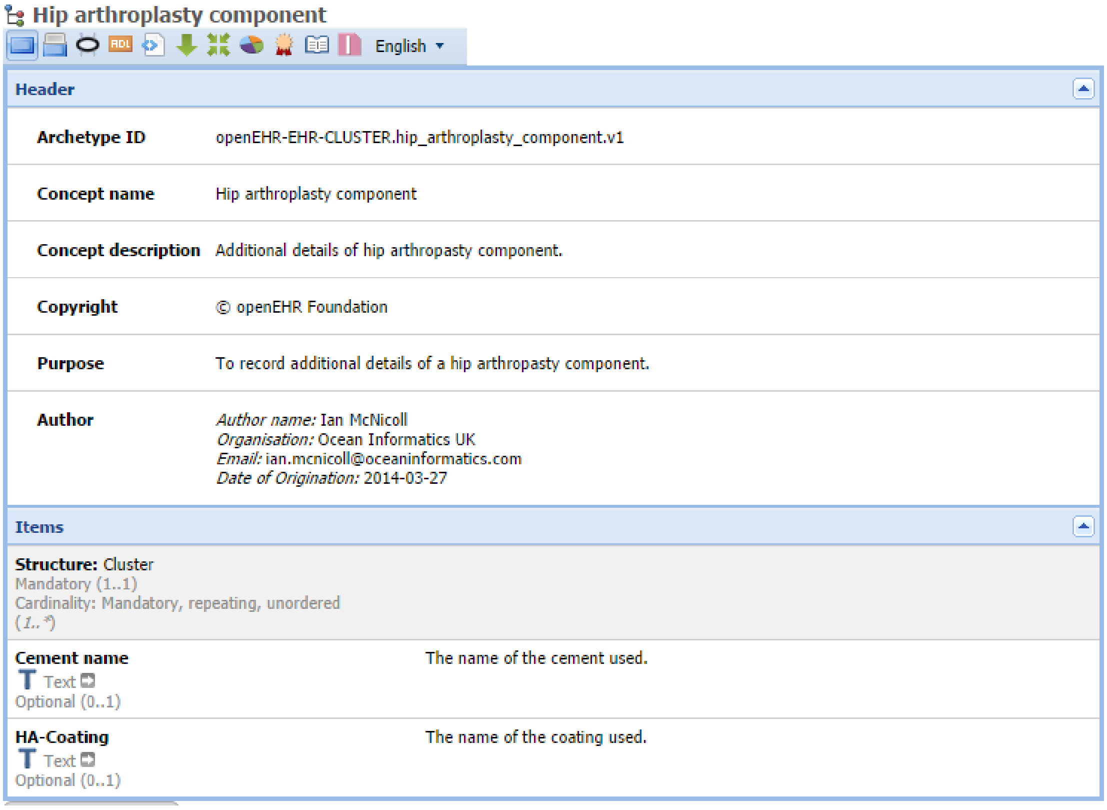
Decision algorithm to help decide which class to archetype

Copyright 2012 Ocean Informatics

Figure 6.21. Process of class type selection (Leslie, Heather & all: OpenEHR archetypes in detail, Ocean Informatics, 2012)

Build the archetype

To build actual archetype we have use computer application for archetype design (for example Archetype Editor). Tools for modelling archetypes can be found on the web page http://www.openehr.org/downloads/modellingtools .



Hip arthroplasty component

Header

Archetype ID	openEHR-EHR-CLUSTER.hip_arthroplasty_component.v1
Concept name	Hip arthroplasty component
Concept description	Additional details of hip arthropasty component.
Copyright	© openEHR Foundation
Purpose	To record additional details of a hip arthropasty component.
Author	*Author name:* Ian McNicoll *Organisation:* Ocean Informatics UK *Email:* ian.mcnicoll@oceaninformatics.com *Date of Origination:* 2014-03-27

Items

Structure: Cluster
Mandatory (1..1)
Cardinality: Mandatory, repeating, unordered
(*1..**)

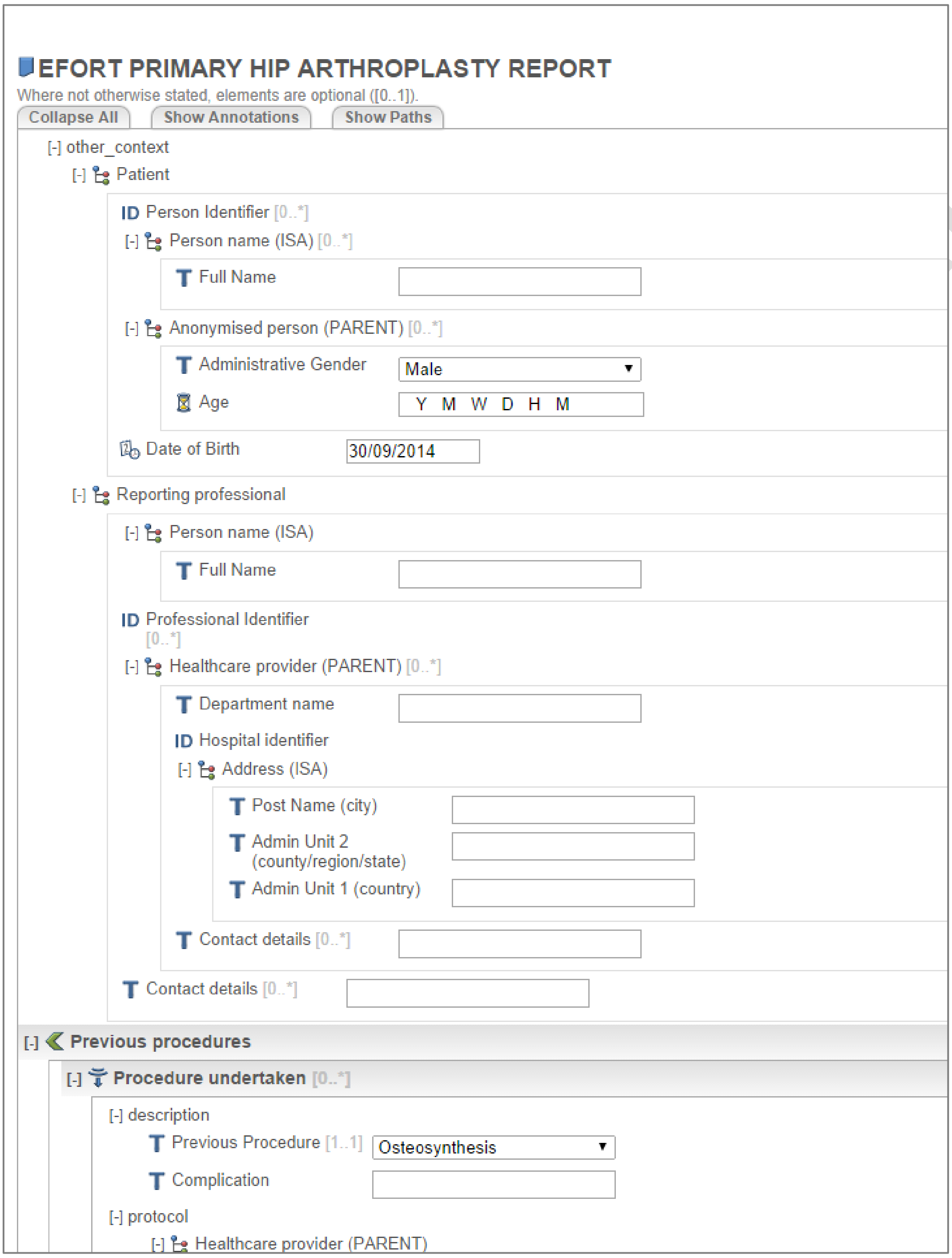
Cement name Text Optional (0..1)	The name of the cement used.
HA-Coating Text Optional (0..1)	The name of the coating used.

Figure 6.22. Hip arthroplasty component archetype

Collaborate & Publish

Archetypes should be published on openEHR CKM (http://www.openehr.org/ckm/).

Add to a Template



EFORT PRIMARY HIP ARTHROPLASTY REPORT

Where not otherwise stated, elements are optional ([0..1]).

[-] other_context

[-] Patient

ID Person Identifier [0..*]

[-] Person name (ISA) [0..*]

T Full Name

[-] Anonymised person (PARENT) [0..*]

T Administrative Gender: Male

Age: Y M W D H M

Date of Birth: 30/09/2014

[-] Reporting professional

[-] Person name (ISA)

T Full Name

ID Professional Identifier [0..*]

[-] Healthcare provider (PARENT) [0..*]

T Department name

ID Hospital identifier

[-] Address (ISA)

T Post Name (city)

T Admin Unit 2 (county/region/state)

T Admin Unit 1 (country)

T Contact details [0..*]

T Contact details [0..*]

[-] Previous procedures

[-] Procedure undertaken [0..*]

[-] description

T Previous Procedure [1..1]: Osteosynthesis

T Complication

[-] protocol

[-] Healthcare provider (PARENT)

Figure 6.23. Example of a part of (draft) template on EFORT hip arthroplasty report (tabular view)

6.5.10 The importance of Terminologies and Code lists

The **unambiguous definition**[94] **of PR data elements** and **proper selection of terminologies and code lists** for defined data elements are **prerequisites for semantic interoperability**. Representing clinical information in standardized ways allow humans and computers to understand clinical information correctly (see also hapters 3.2.5.1 'Standards, models and tools' and 10.11.2 'eHealth standards').

Therefore it is strongly recommended to use **standardized and internationally recognized terminologies and code lists** if they exists for the given item. For example diagnosis and conditions are most commonly coded using World Health Organisation (WHO) International Classifications of Diseases (ICD).

When selecting code lists we have to consider many things and for bigger (especially cross border) projects it is very valuable **to prepare a list of coding system selection criteria** in advance. As a start we can use coding system selection criteria defined by project European Patients – Smart Open Services (epSOS) (34) and also recommended in Guidelines on minimum/non- exhaustive patient summary dataset for electronic exchange in accordance with the cross-border directive 2011/24/EU (35):

- Be internationally used;
- Be in use by some project participants;
- Have translations in a number of different languages;
- Have a maintenance process;
- Have a number of transcoding systems/services, e.g. mapping facilities;
- Be easy to implement;
- Take account of the cost of licences, implementation and maintenance

[94] See ISO/IEC 11179 – 4 (1995) Good Data Definition

References

36. *Chonoles, Michael Jesse and Scardt, James A.: UML 2 for Dummies, Hiongry Minds 2003.*
37. *Introduction to OMG's Unified Modeling Language™ (UML®), Web page http://www.omg.org/gettingstarted/what_is_uml.htm (access 25th June 2014).*
38. Asbjørn Rolstadås (1995). "Business process modeling and reengineering". in: *Performance Management: A Business Process Benchmarking Approach*. p. 148-150.
39. What is BPM Anyway? Business Process Management Explained, http://www.bpminstitute.org/resources/articles/what-bpm-anyway-business-process-management-explained [accessed 10th May 2014].
40. Hammer, M., Champy, J. (1993, 2003). Reengineering the Corporation: A Manifesto for Business Revolution, New York : Harper Business, 1993, New York : HarperBusiness Essentials, cop. 2003
41. Keen P.G.W., Knapp E.M (1996). Every Manager's Guide to Business Processes, Harvard Business School Press, 1996
42. Keen, P.G.W.: The Process Edge, Creating Value Where It Counts, Harvard Business School Press, 1997
43. Event based process chain (EPC), Aris community web page http://www.ariscommunity.com/event-driven-process-chain [accessed 23rd May 2014].
44. Chinosi,Michele, BPMN: An introduction to the standard, Computer Standards & Interfaces, Volume 34, Issue 1, January 2012, Pages 124–134.
45. *Davids, R., Brabänder, E. (2007): Aris Design Platform, Getting Started with BPM, Springer-Verlag London Limited 2007*
46. White, Stephen A., Introduction to BPMN, IBM Corporation,link, http://www.omg.org/bpmn/Documents/OMG_BPMN_Tutorial.pdf [accessed 10th May 2014].
47. Gliklich RE, Dreyer NA, eds. Registries for evaluating patient outcomes: A User's Guide. 3rd ed.2014.
48. *Hernandez, Michael J.: Database Design for Mere Mortals™: A Hands-On Guide to Relational Database Design, Second Edition, Addison Wesley Professional, 2003*
49. Everest, G.: Database management : objectives, system functions, and administration, New York, McGraw-Hill, 1986
50. Riordan, R.: Designing Effective Database Systems, Addison Wesley Professional, 2005
51. IIBA International Institute of Business Analysis: Business Analysis Body of Knowledge (Babok Guide V2.0),
    IIBA 2009, ISBN-13: 978-0-9811292-1-1
52. Craig Larman (1998) : Applying UML and Pattern, An Introduction to OO Analysis and Design Prentice-Hall Int London ISBN: 0-13-748880-7
53. Richard Barker: Case Method-Entity Relationship Modelling Oracle Corporation 1990 UK Limited
54. Vojislav Ivetić, Janko Kersnik: Diagnostične preiskave za vsakdanjo uporabo, Združenje zdravnikov družinske medicine 2007, ISBN: 978-961-91889-5-8
55. Rafael J. Curbelo · Estíbaliz Loza · Maria Jesús García de Yébenes · Loreto Carmona:Databases and registers: useful tools for research, no studies Rheumatol Int (2014) 34:447–452
56. Nadkarni PM[1], Marenco L.,Easing the transition between attribute-value databases and conventional databases for medical data. Proc AMIA Symp. 2001:483-7.
57. European Arthroplasty Register: http://www.ear.efort.org/registers.aspx access [15.12.2014]

58. OB Valdoltra Register Artroplastike http://www.ob-valdoltra.si/raziskovalna-dejavnost/publikacije-register/register-artroplastike-ob-valdoltra access [15.12.2014]
59. Shahar Y[1], Combi C.: Timing is everything. Time-oriented clinical information systems. West J Med. 1998 Feb; 168(2):105-13.
60. Rafael J. Curbelo · Estíbaliz Loza · Maria Jesús García de Yébenes · Loreto Carmona:Databases and registers: useful tools for research, no studies.Rheumatol Int (2014) 34:447–452
61. Richard T. Snodgrass: Developing Time-Oriented Database Applications in SQL. Morgan Kaufman Publishers, 2000, ISBN: 1-55860-436-7
62. Richard T. Snodgrass: TSQL2 and SQL3 Interactions. Association for Computing Machinery / Special interest group on management of data
63. Beal, Thomas (2002), Archetypes: Constraint-based Domain Models for Future-proof Information Systems
64. Tapuria, Archana and all (2013): Contribution of clinical Archetypes, and the Challenges, Towards Achieving Semantic Interoperability for EHR, Healthcare Informatics Research, December 2013
65. Costa, Catalina Martinez and all (2011): Clinical data interoperability based on archetype transformation.
66. Garde S, Knaup P, Hovenga EJS, Heard S (2007): towards Semantic Interoperability for Electronic Health Records: Domain Knowledge Governance for OpenEHR Archetypes. Methods of Information in Medicine 46(3): 332-343.
67. OpenEHR web page, http://www.openehr.org/, access [25. 06. 2014]
68. Leslie, Heather & all: OpenEHR archetypes in detail, PPT, Ocean Informatics, 2012
69. The experience of selecting the code systems for the development of the epSOS master value catalogue (MVC), September 2013
70. Guidelines on minimum/non- exhaustive patient summary dataset for electronic exchange in accordance with the cross-border directive 2011/24/EU, eHalthNetwork, November 2013 http://ec.europa.eu/health/ehealth/docs/guidelines_patient_summary_en.pdf access [1st June 2014].
71. Stellman, Andrew; Greene, Jennifer (2005). Applied Software Project Management. O'Reilly Media. ISBN 978-0-596-00948-9.

7 PATIENT REGISTRY INFORMATION SYSTEM DEVELOPMENT AND IMPLEMENTATION

Healthcare is **information-intensive**, generating huge amount of data every day. It is estimated that up to 30% of the total health budget may be spent on handling information (1).

Also patient registries are dealing with data and information, collecting it, looking for it, storing it, analysing it. It is therefore imperative that **information in patient registries is designed and managed in the most effective way possible** in order to ensure high quality and reliable outcomes **using information technology (IT)**.

In this chapter we will describe the basics of patient registry information system development and implementation by presenting typical software development lifecycle and emphasize the importance of the (end)user in this process. We will address the different possibilities to obtain patient registry software (SW): in-house development, buying/using PR SW product or outsource the development of PR SW. Training in PR software is very essential to the proper and efficiently use of the application, therefore it will be covered as special topic.

After reading this chapter the reader will:

- understand the basics of SW development lifecycle and different SW development models
- understand the importance of user involvement in SW development
- be aware of different options to obtain PR SW
-

7.1 Computer based Patient Registry Information System

Laudon (2) defines an **information system technically** as a **set of interrelated components** that collect (or retrieve), process, store and distribute information to support decision making and control, helping people analyse the problems and visualize complex subjects, and create new products.

There are three main activities in each information system (see Figure 7.1): input, processing, and output which produce the needed information. Feedback in an information system is output returned to the information system with the aim to evaluate and refine the input.

Also a Patient Registry can be seen as an information system. Input is the data on patient health issues, processing is done by classifying, arranging or calculating the data, outputs are relevant reports, alerts etc. Feedback can be for example quality report on data collected which will initiate better data quality control.

Computer based patient registry information systems used for clinical data collections in research and operational settings are usually called **Electronic data capture (EDC)** applications or SW. They are used to collect clinical data in electronic forms. Modern EDC software applications are typically web-based and utilize a thin client. This allows end users to access the study database through the Web through a web browser without the need for the installation of an application on the local computer. Many of them also operate on a Software as a Service model (SaaS).

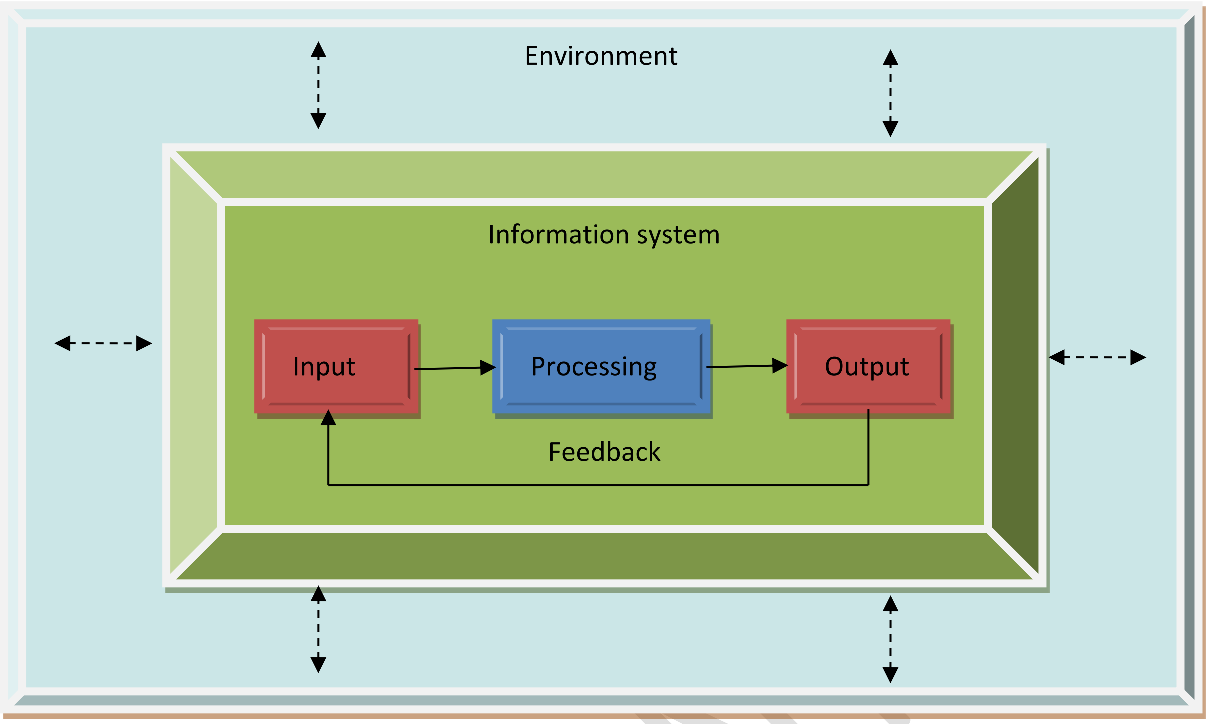


Figure 7.1. Main components of an information system

7.2 Development of Registry Information System

Computer based information system development (short software development) is an engineering approach following the so called Software Developing Life Cycle (SDLC). Typical stages of SDLC are (3):

1. Planning and Requirement Analysis
   - requirements gathering and system analysis, feasibility study (economical, operational & technical), planning of basic project approach
2. Defining software (SW) Requirements
   - preparing and approving (by customer/user) of *Software requirements specification document*
3. Designing the product architecture
   - preparing and approving the product architecture
4. Building or Developing the Product
   - building a product, code generation
5. Testing the Product
6. Deployment in the Market and Maintenance
   - formally release of the product, includes user training, regularly maintenance is required

There is also an international standard for software life-cycle processes - ISO/IEC 12207. Standard that defines all the tasks required for developing and maintaining software.

There are various software development life cycle models prescribing a series of steps to ensure success in SW development. The most important and popular SDLC models followed in the industry are (3):

- Waterfall Model
- Iterative Model
- Spiral Model
- V-Model
- Big Bang Model

The other related methodologies are Agile Model, RAD Model – Rapid Application Development and Prototyping Models.

More on SW development process models can be read in Tutorial on Software Development Life Cycle *http://www.tutorialspoint.com/sdlc/index.htm* .

7.2.1 Important role of users

Regardless of the selected model of SW development the **active user involvement** in the development cycle **is crucial for project success**. Users have the domain knowledge and have their expectation towards the functionalities of the SW application. **User requirements drive the entire system-building effort.** Users must have sufficient control over the design process to ensure that system reflects their priorities and information needs, not the biases of the technical staff. **Working on the design also increases users' understanding and acceptance of the system.** Insufficient user involvement in the development effort is a major cause of system failure. The required degree of users involvement is dependent on the nature of the system built and also on the selected SW development model. For example Agile model and prototyping require more intense user involvement then traditional approaches.

Users are typically involved in all stages of SDLC except in the product building stage. They are especially needed as a key members in **system analysis and system design process** (see chapter *6.5 The role of information system methodologies and techniques in the phase of Patient registry creation* where some modelling techniques are presented), in **testing** (user tests should be performed e.g. acceptance test) and especially in the deployment phase where **training** of the end users is very important.

7.2.2 Software testing

"Software testing is a process of analysing a software item to detect the differences between existing and required conditions (that is defects/errors/bugs) and to evaluate the features of the software item" ANSI/IEEE 1059 standard.

Testing is executing a SW system or its components with the intent to assess if SW satisfies agreed and specified requirements / functionalities or not.

In the process of testing usually software tester, software developer, project lead/manager and end user are involved (4). In SDLC testing can be started from the requirements gathering phase and lasts till the deployment of the software. However it also depends on the development model (*see Ch. 7.2 Development of Registry Information System*) that is being used.

Testing is performed in different forms like for example during *Requirements gathering phase* where analysis and verification of requirements is also considered as testing or code testing executed by developer (Unit type testing) .

Proper testing is undoubtable very important task in SW development. Therefore a lot of standards dealing with SW testing and quality assurance are used by SW developers: ISO

We can distinguish two testing types: **manual testing** where a system is tested manually without using any automated tools and **automation testing** where system is tested using special tools (for example scripts or another SW). Automation testing is used to re-run the test scenarios that were performed manually, quickly and repeatedly.

There exists different methods which can be used for SW testing: black box testing, white box testing, grey box testing.

There are two main levels of SW testing: **functional testing** and **non- functional testing**.

Functional testing assesses functionalities of the system. Examples of functional testing are:

- Unit testing – testing functional requirements of a unit;
- Integration testing – testing if different components work together correctly;
- System testing – testing the system as a whole in an environment close to the production environment;
- Regression testing – if any change is made to SW, then we have to test SW again and
- **Acceptance testing** – testing the SW system in production environment by end user, this testing are also legal and contractual requirements for acceptance of the system.

Non- functional testing:

- Performance testing – testing speed, capacity, stability, scalability; load testing, stress testing;
- Usability testing – testing efficiency of use, learnability, memorability etc.;
- Security testing – testing security and vulnerability of the system (confidentiality, authentication etc.);
- Portability testing – testing SW when it is moved in another environment (another computer with operating system).

All the tests should be well documented. Documentation usually includes: Test Plan, Test Scenario, Test Case and Traceability Matrix.

Further reading on SW testing: Software testing tutorial from tutorialspont.com
http://actoolkit.unprme.org/wp-content/resourcepdf/software_testing.pdf

7.2.3 Training

Proper user training is probably one of the most important aspects of successfully rolling out a new SW, but is often most poorly executed tasks. The training has to be planned in advance, tailored to the audience needs, and executed by lectures with knowledge on adult learning theory and experience in adult lecturing (not just IT experts).

Nowadays trainings could be executed also online as distance learning modules, when the audience is from different location or it is difficult to ensemble users in one place at the same time.

Training program related to PR should be combination of learning PR content and PR SW usage.

When we are preparing training program we have to have in mind the following questions (5):

- Who is the audience?
- What are learning objectives?
- What are the best mechanisms for disseminating the information?
- What is the best approach to ensure that learning has occurred?

7.3 Different options to obtain the Registry system

Information system for PR can be build **in-house**, we can **outsource** the development of PR application or we can **buy** some **product packages** (this options includes also buying SW as a service) from PR, which has usually be than tailored to our needs in-house or by an external partner.

All of the possibilities have their pros and cons. In the following lines we will describe the differences. The first step is to determine if there are any viable products on the market that will meet our business needs and we can **buy** it as an **off the shelf SW package**. If so, a careful analysis of all identified off-the-shelf products should be made. We have to analyse the features, functions, benefits and costs of the products. Usually the product will not fit our PR requirements completely and it will require customization. In cost estimation you have to include also the time and cost of this task. Another important cost factor is ongoing licensing and maintenance costs in the product lifecycle – it can make off-the shelf SW a very expensive option. Benefits of buying a SW package are: lower initial costs, reduced time to deployment, higher success rates, availability of training support, access to user manuals and documentation.

Interesting reading about Electronic data Capture SW is available on *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049639* .

PR can be built **in-house** when the organisation possess enough internal technical capacity. The major benefits to build SW in-house are: organisation has overall control of the development process, the IT experts can be involved already in planning PR (See Chapter 6.5), the organisation has clear understanding on how the SW works, the future development is in control of organization. There are also some common challenges to build PR in-house (6): Unrealistic deadline, vague definition of project deliverables, inadequate time allotted for SW design, little or no testing, lack of quality assurance process, lack of proper project management, insufficient resources for ongoing maintenance and support, documentation that is overlooked or avoided.

We usually choose **outsourcing** the PR SW development when there is no in-house development staff or they do not poses enough technical capacity. Outsourcing means that we will depend on an external company to complete SW for PR. Therefore a strong business relationship will be critical and effective communication between the organizations will be a critical success factor. To select a proper outsourcing SW company you should focus on its (7): experience, approach, infrastructure, quality, reputation, stability, culture.

The major benefits of outsourcing are: reduced project and financial risks, clearly defined requirements and deliverables, use of the most up-to-date design capabilities, reduced project timeline and budget and SW is easy to maintain and enhance.

Quick guide on Outsourcing SW projects guideline can be found on
http://www.bhmi.com/pdf/Outsourcing%20Guidelines.pdf.

References

1. Catalogue of National Health Information Sources in Ireland, HIQA, July 2010 http://www.hiqa.ie/system/files/HI-Catalogue.pdf [accessed 5th June 2014]
2. Laudon, C. Kenneth in Jane P. Laudon. 2012. Management in Information Systems – Managing the Digital Firm. 12th ed. New Jersey: Pearson Prentice Ltd.
3. Software Development Life Cycle Tutorial, *tutorialspoint.com*, http://www.tutorialspoint.com/sdlc/index.htm [accessed 5th June 2014]
4. *Software Testing Tutorial, tutorialspoint.com, http://actoolkit.unprme.org/wp-content/resourcepdf/software_testing.pdf [accessed 22nd October 2014]*
5. *Shinder, Deb, Plan your end-user training strategy before software roll-out, March 2006 http://www.techrepublic.com/article/plan-your-end-user-training-strategy-before-software-roll-out/* [accessed 5th June 2014]
6. Buy, Build or BHMI, BHMI; http://www.bhmi.com/buy_build_or_bhmi.html [accessed 5th June 2014]
7. Outsourcing guidelines, Best practices for outsourcing SW development projects, BHMI, http://www.bhmi.com/pdf/Outsourcing%20Guidelines.pdf [accessed 5th June 2014]

8 RUNNING A REGISTRY

8.1 Sequential Processes

8.1.1 Collecting data

Data collection is defined as the ongoing systematic collection, analysis, and interpretation of health data necessary for the patient registry. Data collection can be considered as regards two major domains; data source and data provider (see chapter 4.1.11.2. “Measures of Quality Data”).

The AMIA (the American Medical Informatics Association) has summarized the “Guiding principles for clinical data capture and documentation” that can be used to orient the implementation for collecting clinical data in a registry.

8.1.1.1 Modes of Data Collection

The way of collecting data for a registry is a crucial part, because it determines its feasibility. Regarding the data sources there are two main sources: paper based and electronic.

In the past, the paper based models were predominant but nowadays the electronic based methods are the main. However, paper can still play a core role in a registry.

Different paper based methods are listed and discussed in the chapter 6.1.2.1.3. Their important characteristics is that they are inexpensive and easy to create and develop, but in the registry whole process they imply a substantial cost because they need to be recorded in an electronic way and there is no a easy and cheap way to do that. The existing paper based processes are being adapted to an electronic environment, with the risk that the paradigm for electronic data capture would be determined by the historical model of paper based documentation.

The electronic based methods are the present and most probably the future ones (but almost half of the EU registries are still based on paper-and-pen mode). Electronic based methods can be computer based o mobile devices based (smart phones or tablets), but the main focus has to be that the data captured would be accurate, relevant, confidential, reliable, valid, and complete. Sometimes, the electronic based methods are focused to integrate several clinical data sources and to produce a new electronic form with the outcomes of the integration (see chapter 8.1.2. “Data Linkage”).

In the past, traditionally, a distinction was made between “passive” collection of data and “active” methods, and the difference was that the passive way is based on the notification and in the active one based on the personnel of the registry visiting the various sources to identify and collect the data. Nowadays the registries use a mixture of methods.

8.1.1.2 Case Report Form

A case report form (or CRF) is a paper or electronic questionnaire on which individual patient data, required by the registry, are recorded. The terminology is widely used in clinical trial research.

The CRF must include the common data elements planned in the design phase and it has to use standard definitions of items and variables (according to international recommendations).

The principles of a good CRF are: easy and friendly use, standard based, short, understandable and connected (if it is possible) with other potential sources.

An example of data to be included in a CRF can be accessed in the book "Cancer Registration: Principles and Methods" (Available from http://www.iarc.fr/en/publications/pdfs-online/epi/sp95/index.php). The EPIRARE project has worked to identify the common data elements for rare diseases registries across Europe and a questionnaire about it can be accessed in http://www.epirare.eu/del.html

8.1.1.3 Data entry/import

The data flow in a registry may include either the data entry (both paper o electronic based) or the capture, or it may import patients' data from clinical databases.

In both cases it is important to establish the next items:

-Who will enter the data?

-Does the data entry program allow certain data items to be entered automatically, or is the data recorder able to make any changes?

-Does the data entry program effectively validate the data?

Paper based:

If the CRF is paper based, a direct data entry can be used. A computer keyboard is used to enter data from the paper CRF into the registry database. It is the easiest way, however, it requires personnel specifically dedicated to record data. Another option is to capture the data from the paper CRF, by using a scanner as well as special software to extract the data from it. In this case, it requires specific CRF forms to avoid errors.

Electronic based:

The data entry can be carried out in a local computerized database, though usually this is an option only for localized registries with a few patients. It is more common to use central database servers using web based data entry forms. In this way the data entry for the registry can be shared in several places.

Mobile devices (smart phones, tables) can be also used as data entry tools, and it is specially indicated when the registry personnel goes to the clinical source.

Finally, a registry can get the data directly from the clinical databases. In this case, the data are captured or imported and require a data linkage process (see chapter 8.1.2.) with specific decision algorithms.

8.1.1.4 Patient/Data Provider Recruitment and Retention

A patient registry does not search completeness as a main goal, however, it is important to get enough patients to reach its objectives. In this way, it is need to develop a source study to know where the data about patients are and which type of data could be used. A plan to review each data source must be established (periodicity of review, type of data source, way to get the data, permissions needed,…). Sometimes, it will be necessary to contact the patients face to face and offer them to include their data in the registry. An informed consent form has to be ready to use.

There are some incentives to recruit patients to the registry, but the most effective is the prestige and outcomes of the registry. If a registry is scientifically well considered, that patient will be more willing to participate. If there are some advantages, like the access to some specific health care processes or the increasing of the visibility of some diseases (especially important in the rare diseases field), patients will be willing to collaborate with the registry.

The transparency and the reputation of the registry are especially important: any problem regarding data protection vulnerability, for example, will imply the lost of patients' confidence and will entail problems for their recruitment and retention.

If the cases are regularly followed up, it will be possible to produce outcomes like remission or survival. For this reason, a registry has to prepare strategies to get the patients' status data regularly. It will be important to maintain updated the registry database with date of each review. An active follow up process may be established by scanning different sources (mortality, treatment or drug prescriptions).

8.1.2 Data Linkage

The data linkage, or record linkage, process is referred to the task of identifying records in one or several datasets that correspond to the same individual or entity. This process may seem trivial if an identification code (ID) or a similar variable, unique for every entity, is available in the dataset(s) to be linked. Nevertheless, this setting is less usual in practice than could be expected or less usual than would be desirable.

Although it may seem obvious, it is worth to start mentioning the importance of a cleaning/purging phase on the dataset(s) of interest before proceeding to link them. This process should be done with particular attention to the variables used to link the databases. Dates in different formats or categorical variables with different codifications, such as {Male, Female} and {Man, Woman} for sex, are simple examples where this kind of issues may produce record linkage methods to fail dramatically. The preprocessing phase will also have to pay attention to string variables where different naming variants or nicknames could be used, such as Jim and James, and unify those variants to a single term.

When dealing with just a single dataset, it is very frequent that in case of having an ID variable included in the database, this is empty for a considerable number of registries. This is particularly frequent in health care registries, where sometimes urgent attention is required and do not having access to the ID of the patient is not enough reason to deny the attention requested. This problem would be particularly prevalent in foreigners who do not have an ID of the corresponding health system because either they require attention in a temporal visit to that country or they are already in process of getting their ID. In that case the ID corresponding to that record is forced to remain empty, with the problems that it may cause for identifying records corresponding to unique entities. A second problem when dealing with just a single dataset may come from records corresponding to children. In some health systems children do not have their own ID and they are recorded either with a missing ID or with the ID of one of his/her parents. This may cause that some records corresponding to children are linked to records of some of their parents (sometimes to one of them and sometimes to the other one) altering the results of analyses that could be subsequently done of that dataset. This may be particularly frequent in newborns, where administrative delays, as regards getting an ID, may make this setting as the general rule for this collective. Last, we should have also in mind the possibility that the ID code of some of the records in the database were wrongly introduced due to typeset errors or to some other reasons. All these circumstances will make naïve record linkage methods to fail and will make the use of more sophisticated methods necessary.

When dealing with more than one dataset this problem is even worse. Besides of the already mentioned problems, that will also be present in this case, the record linkage of two or more databases has some new particularities that we should also bear in mind.

Special care should be put to ensure that the linking fields of the databases are of exactly the same type and of the same length, since otherwise the linking process of the datasets could miss some records that should be matched. This is a particularly frequent setting when the databases to be linked come from different providers or institutions.

It is also a very common setting that the databases used in the linking process were not specifically devised to be linked and were designed for very different aims. Therefore, it is not rare to find that both databases do not share a common ID field that allows linking their records. This is also a very frequent setting when linking databases of different administrations, such as the health and economic authorities, since the identification codes used for any of them are usually different. Specific record linkage methods have been developed for these settings making use of several fields in the database instead of just one.

Record linkage methods can be divided into two sets: deterministic and probabilistic methods. Deterministic methods are used when the databases to be linked lack of a common ID field univocally identifying their individuals. However, if the datasets to be linked contain a set of variables whose combination could be an approximate ID, that combination could be used to link them. For example, the set of variables: name, surname, date of birth and city of residence, could be merged as a unique code univocally identifying any individual in the dataset. In that case, record linkage could be made attending to that code. Nevertheless, errors in the information recorded on these fields or simply because some of them contain missing values, would make this procedure to fail to detect some matching records. To make deterministic record linkage methods more robust to these scenarios, it is usual to include as much fields as possible in the linking process, and match only those records where the percentage of matching fields is above some threshold.

The second set of record linkage methods are those relying on probabilistic decision rules. Thus, not every field in the deterministic methods, such as sex on one hand and date of birth on the other hand, has the same probability of containing two matching records. Probabilistic methods take into account those probabilities to decide if two records belong, or not, to the same entity. It is common in probabilistic methods to build, for every pair of records, a score summarizing the probability of observing as many matching fields as they have, and compare them with a fixed threshold that separates those scores resulting just from chance, from those coming from records of a common entity.

Data linkage can be done with two main purposes: merging the records of several datasets of different providers (e. g. hospitals) in a unique dataset, or enhancing the information of the records in a dataset with those fields coming from a second dataset. In the first case, record linkage will identify records in the different databases to be merged to those that correspond to the same individual. This will avoid accounting for those individuals more times than it should, making it possible to derive reliable rates that otherwise would not be trusty at all if repeated records were not excluded from the analysis. In the second case, inaccurate record linkage methods will make the resulting database to be a riddled of missing data coming from unlinked records, making posterior analyses of that database either unreliable, or more difficult to be done.

Data linkage is one of the most important topics regarding the anonymization legal aspect, because an ID is needed, which is an obvious personal data. The individual right to integrity and protection of personal data has to be matched with the possibility of doing data linkage. There are several options to do it from the legal point of view, and currently a new regulation is in discussion in Europe. The perspectives of the new regulation in Europe are mentioned in the 6.1.4. point.

8.1.3 Controlling and Cleaning the Data

Data control and cleaning on patient registries involve the process by which erroneous data is removed or fixed and missing data is filled.

Three different phases in the cleaning process can be distinguished: screening, diagnosis and editing. All of them shall be applied not just as an independent step of the process, but also during the collection, linkage and analysis of the data.

The screening phase involves any action carried out to detect anomalies in the data. Several types of oddities can be found when screening data and each of them should be taken into account.

- Lack of data can be disguised when data sources use internal codes to declare a missing value, like filling a date field with '99/99/9999' or even literals like 'missing' or 'unknown'. A chart of these internal codes must be built and used as a filter.
- Duplicates can be detected by redundant identification code of the patient or by match in other identification variables as name, date of birth, sex or external identification codes. Algorithms of approximate matching can detect non exact duplicities.
- Format incoherence shall be scanned, detecting values that are incompatible with the preset format of the variable (if Sex is defined as 'M' or 'F', a field filled with 'Male' is erroneous, and shall be recoded).
- The nature of variables offer ranges of values that are improbable or impossible (Age must be non negative and is unlikely to be greater than 100). Thresholds must be defined to screen inconsistent and outlier values
- Joint distributions of variables present different and more restrictive improbable or impossible joint values, like some pathologies combined with sex or age (Sex='Female' and Disease='Testicle Cancer' are incompatible, though each value is coherent by itself). A particular case of this screening is the chronological coherence that dates and ages must have.

The diagnosis phase can classify each oddity detected as erroneous, correct or dubious. A 'hard cutoff' leaves outside logically or biologically impossible values of data that will automatically be classified as erroneous. Improbable but not impossible values are filtered through 'soft cutoffs', and declared as dubious. They should be crossed with external databases (like censuses or other registries) or checked with the primary sources.

Modification of the database in the editing phase can be done automatically or manually over erroneous data. Redundant data shall be merged or deleted. Erroneous values can be corrected or deleted. Linking external more databases provides a source to fill or correct missing or erroneous values. Special codes or flag variables can be set to distinguish corrected fields.

Proper documentation and transparency is required for good practice in data management. Procedures, criteria and actual modifications shall be documented. A good way to keep track of the modifications is to record in a different database the original entries of data before modification. Cleaning process can provide feedback to collection and linkage processes, so that future errors are prevented. It is important to encourage data users to report any anomalies they may find in the data, to improve the controlling and cleaning process.

8.1.4 Storing Data

Storing and retrieval of data are among the IT services giving support to registry operations. In addition to the general considerations about running these type of services, some specific remarks are worth to mention here (Refer to 6.2.7 Information System Management to complete the picture).

Data privacy is a major concern in European countries. At the moment of this writing (may, 2014) the legal framework of reference in this subject is still the Directive 95/46/EC, on the protection of individuals with regard to the processing of personal data and on the free movement of such data, and their different national implementations. In 2012 the European Commission announced a reform of this legal framework. After a lot of work and discussion, that reform is about to be completed.

Personal data about health are among the most sensitive issues. Accordingly, ethics, good corporate governance (transparency, responsibility, accountability, due diligence...) and regulations pose important restrictions to the processing and free movement of these data. Some restrictions have a direct impact on data storage and retrieval. Fines for noncompliance with regulatory requirements may be very important.

Access control (before). Procedures for proper user identification and authentication, as well as for granting and revoking access privileges have to be established. This also includes technical staff.

Access control (after). Logging procedures must keep track of every single access, even if it is only an attempt. Access logs must be kept safely, as they may become evidence, and be periodically examined. Any irregular event must be further investigated.

Data input/output. Any data input/output operation involving systems or facilities not under direct control of the registry owner must be previously approved and then recorded. Once again, these records must be periodically examined. This operations range from copying data to external devices to provide some sort of mobility, to data exports (or backups) to external facilities in order to provide data or operations recovery in case of disaster.

Cloud storage. Even when IT services, based on cloud computing, look interesting, they might be not appropriate at all. The registry owner and any potential provider of IT services (cloud based or not) must previously sign a detailed agreement. The following parts must be present in this document (among others):

1. the provider has to declare and assure his knowledge, will, and ability to fulfill all requirements posed by the aforementioned legal framework;
2. what the service provider has to do, what is not allowed to, and what must do when the engagement with the registry comes to an end;

3. the procedures or evidences available to the registry owner to make him sure that the service provider is running everything according to the terms of the agreement.

Many cloud services are provided out of the EU, where the legal framework mentioned above cannot be enforced (See also the safe harbor framework developed by the U.S. Department of Commerce). Besides, most big providers of cloud services have their own set of terms of service and operate on the basis of take it or leave it. Any of these two handicaps may be determinant to discard a provider.

Data integrity and availability. Power shortages, disk crashes, roof leaks, floods, fires, human errors... These things happen. If it is acceptable that they have an impact on the registry operations (or rather how much impact can be acceptable) is something to be determined by the registry owner, who will have to enable adequate countermeasures. Backup procedures should be conducted according to data recovery objectives and business continuity plans. The ability to recover from the backups is not something to give for granted, but to be tested on a regular basis.

Anonimization. For those purposes (i.e. research) where patient identity is not of primary relevance, dissociation of health data from identity data must be done. Privacy restrictions do not apply to data that can not be traced back to the identity of the patient. Therefore, adequate dissociation processes should be made available as an option for data retrieval. These processes may be either one way dissociation (=anonimization *sensu stricto*), or two way dissociation (=reversible dissociation). The difference is that, in the former one, it is virtually impossible to trace back to the identity of the patient. In the case of reversible dissociation, the keys and procedures to unveil patient identity must be kept under strict control.

8.1.5 Analysis of Registry Data

The analysis of registry data presents as much variety as we can find in the purpose and objectives of registries. Ideally, a detailed data analysis plan should be established beforehand, but flexibility is needed to deal with situations that registry planners could not originally foresee. Situations that call for unplanned analysis will often arise under two different circumstances: first, to address unexpected findings that can lead to new research questions, and second, to give answer to special requests set up by stakeholders. A planned analysis meets researchers' objectives, whereas the foundation of a study based on unexpected findings is developed after making the observation; on the other hand, ad hoc analyses are directed to satisfy registry user's specific needs.

Closely linked to the data analysis plan, statistical methods should be stated in as much detail as possible. Researchers need to be cautious when interpreting registry data, which often has inherent biases. Potential sources of bias should be addressed in advance and, to the extent that it is possible, also the procedures for handling missing data and controlling confounding.

8.1.5.1 Data Analysis Plan

The data analysis plan depends on the registry objectives, but registry planners should be aware that some relevant research questions could arise over time and may not be defined a priori.

Registry-based studies can be descriptive or analytical, but most of the times registries have aims that are primarily descriptive. Descriptive studies focus on disease frequency, distribution patterns (by examining the person, place, and time in relationship to health events), clinical features of patients and natural history of diseases; descriptive studies can suggest risk factors and can help to generate all kinds of hypotheses that could be later tested by analytical studies.

In the case of rare diseases, patient registries are often a first step to try to understand the number of people affected and the characteristics of the disease and the patients, though the scope of these registries may evolve over time.

Disease-specific health indicators (morbidity, mortality and disability indicators) should be made available for the total studied population and for age and sex subgroups. Absolute numbers, as well as crude and age-standardised rates should be calculated. To ensure comparability, standardization should be based on the European standard population.

The main measures of disease frequency are: incidence rate, cumulative incidence, point prevalence, period prevalence, lifetime prevalence and (for congenital diseases) prevalence at birth.

Incidence, often considered the most important measure in epidemiology, is usually expressed as incidence rate, which provides a measure of the occurrence of new disease cases per person-time unit; when incidence rate refers to one year, the denominator is the number of persons under surveillance. High mortality rate diseases, such as some cancers, are better measured in terms of incidence.

Point prevalence can be practically defined as the proportion of the population that has any given disease at some specific point in time, while period prevalence is the probability that an individual in a population will be a case, anytime during a given period of duration, often one year. Prevalence indicators are crucial in rare diseases, as prevalence itself constitutes the main criterion to define a disease as rare.

Mortality indicators, such as mortality rate and case fatality rate, provide a good measure of the burden of disease. Other health status indicators include premature mortality, measured by Years of Potential Life Lost (YPLL); disability-adjusted life year (DALY), a time-based measure that combines years of life lost due to premature mortality and years of life lost due to disability; and quality-adjusted life year (QALY), based on the number of quality years of life that would be added by an intervention.

Analytical studies, such as cohort studies and case-control studies, focus on examining causal associations between exposures and outcomes, or between characteristics of patients and treatment, and health outcomes of interest. Data quality requirements in analytical studies are much higher than in descriptive studies.

For analytical studies, the association between a risk factor and outcome may be expressed as attributable risk, relative risk, odds ratio, or hazard ratio, depending on the nature of the data collected, the duration of the study, and the frequency of the outcome. Attributable risk is defined as the proportion of disease incidence that can be attributed to a specific exposure, and it may be used to indicate the impact of a particular exposure at a population level.

For economic analysis, although not very common in registry-based studies, the analytic approaches encountered are cost-effectiveness analysis and cost-utility studies.

8.1.5.2 Statistical Analysis

Statistical analysis is used to summarize and transform the data stored in the registries into knowledge. This knowledge is the ending result of the registries, since it allows us to know the population covered by the registry and, if appropriate, to compare it with the general population. Besides this aim, registries have just an accounting aim for performing an administrative control of the registered people.

It is not easy to summarize a particular set of statistical tools of particular use in health registries, since these are devised for very different purposes and, depending on them, some statistical tools or some others will be needed. The first set of statistical tools to be used in the analysis of health registries are descriptive tools. Descriptive tools summarize the, sometimes overwhelming, information stored in these registries. For this aim, graphical tools, either depicting the distribution of the values of a single variable or relating the values of two or more of them, are of particular use. Descriptive statistics are also often used for summarizing information in the databases, thus, the mean, median and standard deviation are typical statistics used to summarize variables. If instead, we are pursuing some measure measuring the amount of dependence between two variables in one dataset, Pearson's correlation coefficient is frequently the most widespread tool.

In addition to the descriptive aims above, we will be often interested in making inference (learn) on some features of the population covered by the registry. In that case we could be firstly interested in contrasting some specific hypothesis in our own dataset. In that case, we should resort to statistical tests. There is a huge amount of statistical tests available for very different purposes and it is not within the scope of this Section to make even a brief description of their use. Nevertheless, we find it convenient to highlight chi-square and t-test as the most common tests for making data analysis. The t-test is usually an appropriate choice for comparing the mean of two different groups in the population, although it requires the variable to be studied to be Normal-shaped. If this condition is not achieved, some alternative non-parametric test should be used, such as the Wilcoxon's signed rank test. On the other hand the chi-square test is used to assess dependency between two categorical variables.

Instead of testing some particular hypothesis in our dataset, we could be interested in assuming a statistical model for our dataset and to learn about the parameters ruling that model. Thus, as an example, we could assume a linear shaped relationship between two variables and try to learn about the parameters defining that relationship. There are also several tools for achieving this goal. Thus, linear models (assuming a Normal outcome) are usually used for continuous variables, but if the outcome variable cannot be assumed to be Normal, Generalized Linear Models are the most usual tools to model this kind of settings. Logistic Regression and Poisson Regression models are just particular cases of Generalized Linear Models.

Finally, we would also like to mention Survival analysis as a statistical technique of particular use in health registries. Survival analysis is devised to the study of the time taken for an individual to develop an event of interest, such as the time survived before dying or developing a metastasis. The particularity of this kind of analysis is that many individuals in the dataset do not show the event of interest, maybe because that are not going to develop it in a future, because they have not developed it yet (although they will in a future) or because they have simply left out the study. This makes the variable of interest in these studies to be only partially known sometimes, and the analysis of this kind of data requires a particular treatment. If a more descriptive tool is wanted in this context Kaplan-Meier curves are the most usual tools, meanwhile, if we prefer to model the effect of some covariates on the time of survival, usually Cox regression models are the most widespread tools.

Finally, we also find it convenient to mention some available tools for carrying out this statistical analysis. Although we do not intend to be comprehensive with this list of statistical packages, we highlight SAS, Stata or SPSS as the commercial packages of more frequent use in the health sciences in general. Any of those packages could be perfectly suitable to carry out the above-mentioned analysis in the context of health registries. Nevertheless, R is nowadays an open source alternative with widespread use much beyond in health science. R is usually blamed of being a bit rough for non-statistical users. Nevertheless, some specific R packages, like Rcmdr, are intended to make the use of R for non-statistical users easier, so that they make R available for a wide community of users. The main advantage of R is that it will likely have specific state-of-the-art packages for hardly any task that you could wish to do in your registry, such as record linkage, dealing with "confounding by indication", missing values, …

There are lots of textbooks covering the statistical methods mentioned above; in fact, even specific monographs for any one of most of those methods have been published. For sure, the most appropriate book for any user will be that which illustrates their examples with the software habitually used for making the current statistical analysis. Thus, depending on the software used these could be the appropriate textbook choices: Le (2003) for SAS users, Cleophas and Zwinderman (2010) for SPSS users, Hills & De Stavola (2012) for Stata users and Lewis (2009) for R users. Although once again this is not intended to be a comprehensive list of possibilities, but just a collection of useful textbooks.

8.1.5.3 Analytical considerations

When undertaking the analysis of the information stored in health registries there are a series of issues that deserve particular attention and that should be always borne in mind. Following, we are going to introduce some of those issues to make the reader aware of their existence and their effects.

8.1.5.3.1 Potential sources of bias

There are numerous potential sources of bias when dealing with data providing from health registries. We are going to mention here 4 specific sources of bias in observational studies in general: selection bias, non-response bias, information bias and recall bias.

One of these sources is selection bias, which is the result of the selection mechanism in the inclusion of individuals in one registry. Thus, as an example, let us assume that a diabetic patient registry is composed by patients recruited from their visits to a hospital. By definition, only those patients who have visited at some moment the hospital have the opportunity to be included into the registry. Regretfully, patients visiting the hospital are not a random sample of the diabetic patients in the whole patients but, on the contrary, they are severe patients who have possibly had a complication related with his/her disease. This will make the results drawn from the registry to be non-representative of the diabetic patients in the whole population.

A related bias would be the non-response bias, in which all candidates to be included in the registry may have been previously recorded, however, some of them show missing values for some specific fields. These missing values can be rather innocuous if they are produced at random. Regretfully, quite often, the presence of missing values responses to a non-random mechanism, making those fields in the database not being representative of the population and, therefore, biasing the results if this potential bias is not taken into account during the statistical analysis.

The second source of bias that we consider convenient to mention is information bias. Information bias is the bias coming from inconsistencies in the way that information is introduced into the registry. Some artefact in the process of retrieving or coding the information into the database could make that information not to reflect the reality, but rather a biased and distorted image of that reality. This could be the case of a variable reflecting the vaccination status of the individuals in the database. By default, this variable could be set to non-vaccinated and changed to vaccinated if appropriate. Nevertheless, as the default value goes always in the same sense, it can often happen that vaccinated individuals are registered in the database as non-vaccinated for the reason that the person who administered the vaccine did not record it into the database. This systematic bias could introduce problems and further bias into posterior analyses of the information in the database.

Finally, the recall bias should also be born in mind when working with health registries, mainly when part or the entire database is retrieved from interviews or questionnaires. This source of bias is produced by differences in the accuracy of the information of the people included in the registry coming from their past. For example, people with familiar antecedents of some cancer could be more aware of previous expositions to some carcinogen agents than people without those antecedents. So, the information of both kinds of people could be systematically biased towards different directions, simply by their particular circumstances.

These biases are usually incorporated in the database since the very moment of introducing information. Registries professionals should be very aware about them, so that, even from the design phase of the registries they are prevented and, when possible, these biases are avoided by means of appropriate statistical analysis.

8.1.5.3.2 Confounding by indication

When analyzing data coming from health registries it is quite common to study a variable as a function of some covariates. Nevertheless, the distribution of the values of the covariates in data coming from registries is not done at random or following a specific and controlled design. On the contrary, these values in observational studies in general, and health registries in particular, are the result of some factors not registered and out of the control of the study. For example, the decision of administering a medicine to a patient may be taken by a practitioner as the result of a general assessment of his/her health. As a consequence those patients with a worse general health status will take the medicine and those who are better will not. When assessing the effect of the medicine on a final outcome, such as dying in the following year, we could conclude that taking the medicine could increase the probability of dying, when this would be an effect of the previous health status of the patients. This effect is known as confounding by indication, and it may take us to draw wrong conclusions on the effect of a variable because it is simply confused with other uncontrolled variable(s).

When interpreting the results of health-registry based analysis we should be very aware of this potential problem. In case of suspecting that it could have had an influence on the estimation of the effect of a variable in the study, we should resort to statistical techniques designed to control that effect, as for example, the inclusion of propensity scores into the analysis. Propensity scores will be auxiliary variables to be included into the analysis controlling the non-random mechanism that has generated the missing values in the dataset of the study.

8.1.5.3.3 Missing data

Health registries quite often contain missing data for some of their variables. These missing data are a real problem for data analysis and should be treated with care if we want to prevent them to induce some bias in the data analysis phase.

It is very convenient to know the reason why the missing data are produced. The best, although the less likely, scenario is that missing values occur at random. That is, no relationship can be found between their occurrence and any known variable. In that case, missing data are not very harmful, although they introduce some difficulties in data analysis. If the dataset at hand is large, those individuals containing missing values could be simply removed from the analysis and we would not expect big changes in the new results. If, on the contrary, we have knowledge or evidence from data, that missing data have not been produced at random, this should be born in mind because they could be much more harmful in the data analysis phase. In that case, removing these individuals from the analysis would mean to remove a particular part of the whole population, what could produce little or large biases depending on the degree of particularity of that sample. Therefore, in this case, a naïve removal of these individuals from the analysis does not seem to be an option. In this case imputing the missing values is the main option, although that imputation should be made taking into account the mechanism generating the missing data, e. g. if those individuals with particularly large (or low) values of a variable tend to show missing values in a second variable we should take this into account. If these two variables showed some correlation, we should consider the value of the first variable in order to impute the values in the second one, instead of doing it completely at random.

8.1.6 Data Dissemination

Well established, multicenter or population-based registries that held large data collections can be a rich source of information with many different users, while small locally-held registries have a limited number of potential users, but in both cases data should be made accessible to ensure that all information is used to the maximum benefit of the population it serves.

Data should be disseminated in different ways, depending of the addressee of the data. Thus, three different points of view should be taken into account concerning registry data dissemination: 1) registry holders or owners, 2) patients and general public, and 3) decision makers and researchers.

Patients and service users, researchers, health professionals and policymakers, as well as other stakeholders and even the general public, should have access to valid and properly presented information in order to make choices and decisions. By making outcome data transparent to stakeholders, well-managed registries enable medical professionals to engage in continuous learning and to identify and share best clinical practices. To identify potential stakeholders, it is important to consider to whom the research questions matter. It is useful to identify these stakeholders at an early stage of the registry planning process, as they may have a key role in disseminating the results of the registry.

Registry-based information can be made available in many different ways, such as periodical reports, extracts on request and specific tools provided to allow users to access the data themselves via online portals. The principles of good dissemination data have to be considered. An example would be the United Nations Good Practices on National Official Statistics (http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx).
Writing reports, presentations, tables, graphs or maps can be used to show the registry outcomes. The understanding is the main principle and it is very important to use the right type of tool for presenting the information. However, if a particular dissemination tool (represented, for example, by a table, graph or map) does not add to or support the analysis, it should be left out.

The dissemination reports should contain only information, only data or both data and information (data with a text explaining those data).

According to the addressee of the data, a good dissemination strategy (fixed in a dissemination plan) would have to consider the next features:

1. Registry holders or owners: dissemination requires actions to reinforce the acknowledgment of the people implied in the registry process, such as data providers, clinicians or managers
2. Patients and general public: it will be necessary to disseminate basic indicators, mainly in the form of basic tables, as they are more easily understood by them. On the other hand, graphics, maps and other sort of representations are also needed.
3. Decision makers and researchers: the dissemination has to be done in the form of aggregated data, but it is important to prepare individual anonymized data for researchers.

Every original finding and all scientifically significant information generated by disease registries should be communicated to the scientific community and finally rendered as scientific publications (on paper, online or both). Indeed publishing of results is inherently linked to the purpose of most if not all patient registries, as proper publishing can be considered an integral part of the scientific method.

Long-term population-based registries, an essential tool for public health surveillance, typically produce periodical descriptive analysis of data to be distributed to all potential users and especially to health professionals providing the data, as this feed-back enhances subsequent cooperation.

In clinical registries, data on disease progression or other long-term patient outcomes may not be available for many years, but safety data could be examined periodically over time. Studies based on patient registries, even short-term registries, may conduct intermediate analysis before all patients have been enrolled or all data collection has been completed, in order to document and monitor the progress of the project.

As the paradigm about health information goes, registry data should be collected once and used many times. Timeliness will be the key.

8.2 Overarching Processes

8.2.1 Data Quality Assurance

Electronic Health Records are generally designed for their primary use. As a consequence, when their data is collected with secondary, reuse purposes, such as for the construction of research repositories, their Data Quality (DQ) may not be optimal. Research repositories generally count with higher levels of DQ as specific, mostly due to manual curation and data profiling processes. However, DQ problems are still present. These can lead to suboptimal research processes, or even to inaccurate or wrong hypotheses. With the purpose to ensure the highest levels of quality, continuously improve DQ processes, and avoid further DQ problems, organizational DQ Assurance protocols should be established.

DQ Assurance protocols combine activities at different levels, from the design of the information system, the user training in DQ, to a continuous DQ control. To this end, many research and industrial DQ Assurance proposals have been related to the Total Quality Management Six Sigma process improvement methodology. Concretely, the DMAIC model can be used to improve the DQ and their related processes, involving the following cycle of steps: Define, Measure, Analyze, Improve and Control. We can affirm that defining what to measure and how to do it is the basis for the DQ Assurance, being them the initial steps to any DQ improvement. These steps, along with the DQ control, can be defined under a DQ Assessment framework.

8.2.2 Data Quality Assessment

The Data Quality (DQ) Assessment is managed according to DQ dimensions: attributes that represent a single aspect or construct of DQ. Dimensions can conform to data definitions or to user expectations. Thus, DQ Assessment concepts and methods can be defined according to specific domains or problems. We can establish a set of DQ dimensions to assess the DQ of cross-border patient registries based on different studies (see Chapter 4 ‘Quality dimensions of Registries).

There exist other dimensions which, rather than measured on data by themselves, can be measured on their related stakeholders. Data *Availability* refers to their degree of accessibility to users. Data *Security* refers to their degree of privacy and confidentiality. Finally, data *Reliability* refers to the degree of reputation and trust of the stakeholders and institutions involved in its acquisition.

DQ problems may affect to single or combined variables within an individual patient registry, e.g, inconsistent combination of variable values. Else, DQ problems may affect to a sample composed of a set of registries, e.g., a biased sample mean. For that reason, according to the purpose, methods should be considered to be applicable to large-scale big data repositories.

To conclude, we would like to remark that it is of upmost importance for the DQ Assessment to formally define what is to be measured and controlled according to the aforementioned dimensions. Based on that, strategies can be defined to correct or prevent DQ problems. DQ processes can be applied to off-line research datasets. However, continuously controlling (based on on-line methods or multi-site audits) DQ indicators within a DQ Assurance cycle, from which to obtain a feedback to improve processes, is a recommended strategy to continuously reduce the DQ problems and optimize resources.

8.2.3 Evaluation and Improvement of Registry Service

Quality assessment of a registry should be a continuous process integrated in the registry running. The dimensions needed to measure it (completeness, validity, opportunity,…) are common to different type of registries, but methods and indicators are related to the type of registry. The population cancer registries are one of the most advanced examples.

Regarding to their complexity and cost, some methods can be routinely implemented, while others - as contrasting it with an independent series of cases, which is one of the most used methods to asses completeness-, should only to be used in a sporadic way.

However, using an external audit for the registry is a good idea, though external audit and accreditation -used in the health sector since decades, and considered useful to promote high quality products and services with efficacy and reliability-, are less developed in the registry field, except in the United States of America.

For those registries in which the health administration is both the data supplier and data user (client), it is needed to incorporate in an active way the opinion of health planner and health management professionals.

An example could be the REDEPICAN (Latin America Network for Cancer Information Systems) Guide for the External Evaluation of Population-based Cancer Registries, used in several Spanish and Latin America cancer registries. It is a new tool inspired in the accreditation principles: voluntary process, standard and defined criteria, self-assessment, external verifying process, and independent organism report. The Guidelines assesses 7 dimensions (Structure, Procedures Manual, Registry Method, Comparability, Completeness, Validity, Outcome Dissemination, and Confidentiality and Ethical Aspects) through 68 criteria with three standard levels, allowing to assess the traditional indicators and procedures needed to make the changes needed, in order to offer the maximum efficiency. The final score, and the criteria with a low score, identify problems to be solved in the registry with concrete objectives for improvement. An external audit with a homogeneous measurement tool is useful as the starting point to measure quality improvements and to compare between registries.

8.2.4 Governance

Patient registries governance comprises the systems and procedures by which a registry is directed and managed. It refers to guidance and high-level decision making, including concept, funding, execution, and dissemination of information.

Good governance must include:

- Accomplishment of the normative (regional, national, international). In some countries a previous approval for the operation of a registry, by professional or health authorities is needed. The support and approval of the institution in which the registry is located is fundamental. The ethics committee approval is also needed.

- Principles in which the registry action is based. Some of them are: transparency, participation, accuracy, security and data protection.

- Operating rules definition. This is a document that specifies the rules, case definitions used, codes and classification used (assuring the semantic interoperability). All the operating procedures have to be elaborated and released to all the participants in the registry. The way in which the data of the registry may be accessed has to be clearly defined. A document for the consent and its procedures has to exist (see Chapter 4.1 'Governance').

- The structure of the governance board (and its role and responsibilities). According the governance plan (see chapter 6.1.9 'Governance, Oversight and Registry Teams') the governance board can be structured in several ways:

- The prioritization is to have a project management team, and scientific committee and a quality assurance committee.
- A scientific committee or expert group can be formed to guide the development of the registry and to ensure the scientific basis. Its role is as a consultant group.
- The project management team can be developed also as a steering committee. It has to ensure that the registry is running according to the principles and objectives marked and planned. Its composition has to be done taking into account the institution in which the registry is based, the organism which funds the registry, the professionals implied, the health authorities, the academic or scientific institutions related to the subject of the registry and the patients and their families affected. Its role is to assume the responsibility of the registry. The Chair of the steering committee assumes final responsibility.

An example of Registry Governance Document is the National Cancer Registry of Ireland.

8.2.5 Auditing

According to the Dictionary of Epidemiology, an audit is an examination or review that establishes the extent to which a condition, process or performance conforms the predetermined standards or criteria. In a registry, audits may be carried out on the quality of data or completeness of records. Depending on the purpose of the registry, several types of audit can be performed. The audit can assess: enrollment of eligible patients, data completeness, selection bias, or data quality. An example of quality assessment is shown in chapter 6.2.3. The audit can be conducted either on the whole set of data of the registry, or just for a selected (random or systematic) sample of patients, using sampling techniques.

For example, the Spanish National Rare Diseases Registry has performed an audit in a Spanish region to assess the validity of diagnosis of aplastic anaemia by the International Classification of Diseases codes in hospital discharge data and the mortality registry, in order to detect cases to be included in the rare diseases registry. After getting the data from both databases the patient medical records were reviewed to confirm true aplastic anaemia cases. Only 15% of the cases were confirmed[95].

The audit can be internal or external. Internal audit is carried out by the registry staff, using a concrete plan and specific indicators to assess the most significant sources of error as regards the purpose of the registry. External audit is performed by external personnel, in accordance with pre-established criteria.

[95] Ruiz E, Ramalle-Gómara E, Quiñones C, Rabasa P, Pisón C. Validation of diagnosis of aplastic anaemia in La Rioja (Spain) by International Classification of Diseases codes for case ascertainment for the Spanish National Rare Diseases Registry. Eur J Haematol. 2014 Aug 18. doi: 10.1111/ejh.12432. [Epub ahead of print]

8.2.6 Continuous Development

A registry is always in a continuous process of actualization. For example, the way of collecting data can experiment changes due to technological innovations, organizational modifications or new legal rules.

For that reason, a registry should be flexible and adaptive in all the facets of the registry process:

- For paper-based registries, it is crucial to move on to electronic based ones.
- New data elements could be added (new treatments or new disease stage for example).
- Definitions can be modified according to improved knowledge.
- Revisions of the classification systems happen and the registry has to be ready to be adapted to new one.
- It is needed to foresee any legal modifications regarding to ethical and protection data rules, as the personal identifier code can also change or may need to be encrypted.
- The technological innovations affect the way in which a registry operates.
- The methods of data quality processes should be adapted to the results achieved.
- The reports and the diffusion mechanisms need to be flexible, because new data users can be incorporated and the stakeholders' concerns may change.

The development of the registry has to be continuously and periodically tested, in order to progress and adapt to the potential changes.

All the modifications have to be done ensuring the quality and integrity of the data and planning the date of the beginning.

8.2.7 Information System Management

Running a registry requires dealing with a certain number of stakeholders (patients, providers, clients, partners, regulatory authorities...). Running a registry also takes a good deal of IT. A registry owner will therefore be interested in arising trust among the stakeholders, as well as in getting most value from his information systems. IT governance can provide both.

All of IT activities within the registry should not take place on an improvised, contingent or ad-hoc basis, but within an adequate governance and management framework[96]. This is the best way to:

- Maintain high-quality information to support business decisions.
- Achieve strategic goals and realize business benefits through the effective and innovative use of IT.
- Achieve operational excellence through reliable, efficient application of technology.
- Maintain IT-related risk at an acceptable level.
- Optimize the cost of IT services and technology.
- Support compliance with relevant laws, regulations, contractual agreements and policies.
- Provide trust to all stakeholders.

[96] ISACA's COBIT 5 is the most comprehensive business framework for the governance and management of enterprise IT. This framework provides with good practice and guidance from the knowledge and experience of a large (>115000 members) community of IT audit, security, risk and governance professionals worldwide. Alternatively, different sets of ISO standards address some of the main issues (e.g. ISO/IEC 38500, ISO/IEC 20000, ISO/IEC 27000 ...).
More at http://www.isaca.org/cobit/pages/default.aspx and http://www.iso.org.

Among IT activities there should be management processes related to deliver, service and support. In that area, the following processes have to be considered:

- Manage operations.
- Manage service requests and incidents.
- Manage problems.
- Manage continuity.
- Manage security services.
- Manage business process controls.

Other processes should be run to monitor, evaluate and assess performance and conformance, the system of internal controls, and compliance with external requirements. Key indicators are essential in this context, as they are a main source of knowledge and allow measuring variables like cost, risk, disruption, improvement, and others.

All of the above should discard/discourage anyone pretending to "take care of all IT stuff" around the registry without adequate knowledge or tools. Being proficient at making scrambled eggs at home does not qualify for a chef position.

References

1. National Health Information Management Group. Minimum Guidelines for Health Registers for Statistical and Research Purposes. Australian Institute of Health and Welfare. 2001.
2. Cruz-Correia RJ, Pereira Rodrigues P, Freitas A, Canario Almeida F, Chen R, Costa-Pereira A. Data Quality and Integration Issues in Electronic Health Records. In: Information Discovery On Electronic Health Records. V. Hristidis (ed.); 2010. p.55–96.
3. Cusack CM, Hripcsak G, Bloomrosen M, Rosenbloom ST, Weaver CA, Wright A, Vawdrey DK, Walker J, Mamykina L. The future state of clinical data capture and documentation: a report from AMIA's 2011 Policy Meeting J Am Med Inform Assoc. 2013;20(1):134-40.
4. ENERCA (European Network for Rare and Congenital Anaemias):
5. http://www.enerca.org
6. EPIRARE (Deliverable 1.4 Statistical Analysis of the EPIRARE survey data):
7. http://www.epirare.eu/_down/del/D1.4_StatisticalAnalysisofRegistrySurveyData.pdf
8. EUROCISS (Cardiovascular Indicators Surveillance Set): http://ec.europa.eu/health/ph_projects/2000/monitoring/fp_monitoring_2000_frep_10_en.pdf
9. Gliklich R, Dreyer N, editors. Registries for Evaluating Patient Outcomes: A User's Guide. 3rd ed. Rockville, MD: Agency for Healthcare Research and Quality; 2012 (Draft released for peer review).
10. Health information and Quality Authority. Guiding Principles for National Health and Social Care Data Collections. Dublin: Health information and Quality Authority; 2013.
11. Karr AF, Sanil AP, Banks DL. Data quality: A statistical perspective. Statistical Methodology. 2006, 3, 137:173.

12. Larsson S, Lawyer P, Garellick G, Lindahl B, Lundström M. Use Of 13 Disease Registries In 5 Countries Demonstrates The Potential To Use Outcome Data To Improve Health Care's Value. Health Aff. 2012;31:1220-227.
13. Lee YW, Strong DM, Kahn BK, Wang RY. AIMQ: a methodology for information quality assessment. Information & Management. 2002. 40, 133-146.
14. MacLennan R (1991). Items of patient information which may be collected by registries. In: Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, editors. *Cancer Registration*: *Principles and Methods*. Lyon: International Agency for Research on Cancer (IARC Scientific Publications, No. 95); pp. 43–63. Available from http://www.iarc.fr/en/publications/pdfs-online/ epi/sp95/index.php.
15. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. PLoS ONE. 2013 03;8(3):e55811.
16. National Cancer Registry of Ireland: http://www.ncri.ie/sites/ncri/files/documents/GovernanceFrameworkfortheNationalCancerRegistry24September2010.pdf
17. Navarro C, Molina JA, Barrios E, Izarzugaza I, Loria D, Cueva P, Sánchez MJ, Chirlaque MD, Fernández L. [External evaluation of population-based cancer registries: the REDEPICAN Guide for Latin America]. Rev Panam Salud Publica. 2013 Nov;34(5):336-42
18. Newton J, Gardner S. Disease registers in England. Institute of Health Sciences. University of Oxford. 2002.
19. Posada de la Paz M, Villaverde-Hueso A, Alonso V, János S, Zurriaga O, Pollán M, Abaitua-Borda I. Rare Diseases Epidemiology Research. In: Posada de la Paz M, Groft S, editors. Rare Diseases Epidemiology. Heildelberg, London, New York: Springer Science+Business Media B.V.; 2010.
20. Richesson R, Vehik K. Patient Registries: Utility, Validity and Inference. In: Posada de la Paz M, Groft S, editors. Rare Diseases Epidemiology. Heildelberg, London, New York: Springer Science+Business Media B.V.; 2010.
21. Röthlin M. Management of Data Quality in Enterprise Resource Planning Systems. Josef Eul Verlag GmbH (ed.). 2010.
22. Sáez C, Martínez-Miranda J, Robles M, García-Gómez JM. Organizing data quality assessment of shifting biomedical data. Stud Health Technol Inform. 2012;180:721–725.
23. Sáez C, Robles M, García-Gómez JM. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. Under review. 2014a.
24. Sáez C, Robles M, García-Gómez JM. Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. Under review. 2014b.
25. Sáez C, Martínez-Miranda J, Robles M, García-Gómez JM. Organizing data quality assessment of shifting biomedical data. Stud Health Technol Inform. 2012;180:721–725.
26. Sebastian-Coleman L. Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. Morgan Kaufmann (ed.). 2013.
27. Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K (2005) Data cleaning: Detecting, diagnosing, and editing data abnormalities. PLoS Med 2(10): e267.
28. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manage Inf Syst. 1996;12(4):5–33.
29. Wang RY. A Product Perspective on Total Data Quality Managements. Communications of the ACM. 1998;41(2):58–65.

30. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan;20(1):144–151.
31. Zurriaga O, Bosch A, García-Blasco MJ, Clèries M, Martínez-Benito MA, Vela E. [Methodological aspects of the registries for renal patients in replacement therapy]. Nefrología. 2000;20 Suppl 5:S23-31.

9 CHANGING AND STOPPING REGISTRIES

9.1 Changing an existing registry

A registry is a living system that evolves over time. In order to remain or become more useful and successful a registry sometimes needs to be modified. In general, it is important that a registry is flexible and adaptive, with a sense of continuous development. Regular checks and evaluations (e.g. internal or external reviews) of whether any of the registry's components needs to be modified are important factors that affect the sustainable success of a registry.

There are various reasons for a registry to undertake the modification or adaptation process. Unmet registry stakeholder needs, failure to meet certain standards, reducing the burden of the registry team or participants, new regulatory or legal requirements, innovations and changes in medicine and health care (i.e. new products, procedures, and services), innovations in information technology, or changes in the financing of the registry, for example, are one of them.

In connection with this a registry can undergo many different modifications. Maybe a registry needs to change its general purpose, goals or outputs, change the mode of data collection (e.g. moving from paper-based data collection to electronic data collection), introduce a new technology (e.g. web-based data entry), modify the target population or cohort (e.g. geographical expansion, the expansion of the age range, additional excluding criteria), adapt the outcomes or exposures (e.g. inclusion of the knee implants as an exposure in addition to hip implants), modify any of the data elements (e.g. removing a redundant data element, adapting the outdated one, or adding a new one), change the case report form (e.g. to develop more user-friendly form that is less subject to human errors), modify the data collection protocol (e.g. different time points for follow-up), improve data analysis or data dissemination (e.g. more appropriate analytical techniques, different graphical representation of data, or different frequency of the dissemination), adapt a registry team or governing board, change the funding source or find new stakeholders (e.g. a move to public-private partnership), improve the overarching processes (e.g. quality assessment, auditing) etc. Some minor changes in a registry can be implemented more easily and quickly, but modifying a registry can be also a complex task that requires more effort, time and money. Therefore, it is highly recommended that a registry team, wanting to modify a registry, considers various elements in order to implement changes successfully and run the transition smoothly.

This chapter does not provide guidance for each and every change that can occur and be implemented within a registry, and does not cover in detail every step of the registry modification. The modification of existing registry is in many ways similar to a process of establishing a new one, the latter already being covered in other chapters. Thus, a reader is encouraged to read other parts of the guidelines for additional information. For example, when reconsidering legal and ethical obligations during the modification process a reader is encouraged to see the chapter 5 'General requirements for cross-border use of patient registries'; or when modifying data elements a reader should read the chapters 6.3 'Registry dataset', and 6.5.6 'Data modelling', on how to adequately develop data elements for a registry. However, as it was already mentioned, there are various points that should be considered/emphasized while planning and/or implementing the modification of a registry:

1. It is important that there is a **clear rationale** for the registry change since not every change is a good one. A registry team needs to understand why the change is necessary, what exactly

needs to be changed, and what the change will bring. Thus, a clear purpose and goals of the change should be developed. If there is no solid reason for a registry change, it should not be implemented.

2. It is essential to know how major and complex the change will be. The fact that changing one element of a registry can lead to the changing of other elements here should be taken into account. For example, changing data set can lead to the adaptation of case report form, data collecting process, statistical analysis and data dissemination. When modifying a registry this can be a good opportunity to make some other changes that are necessary as well. Therefore, a registry team needs to **carefully determine and understand the scope** of the registry modification. Furthermore, an **assessment of the feasibility** of registry modification is a crucial part. Costs, time, effort, skills, and other resources are essential factors to consider. It is necessary to be aware of potential limitations and risks (e.g. technical breakdown or incompatibility, delays) as well. As for creating a new registry, **good planning** contributes to the successful implementation of a registry modification/transition. Thus, it is recommended to develop a thorough and realistic action plan and strategy for a registry modification/transition. It is worth mentioning here that piloting and testing are activities that should not be underestimated.
3. Once it is clear what modification in a registry will be implemented, a registry needs to have **a team for transition**. In that case it is important to consider skills and knowledge that members possess, and how the effort of the specific team member will be increased during the modification process, because a member will probably have to perform his or her regular work simultaneously. It is also important to establish continuous, honest and open communication between all members, because effective collaboration between them can identify some unexpected barriers or risks that can be suitable addressed during the planning phase. The **role of leadership** here cannot be overemphasized since, as in many other areas, it is one of the key factors for a successful implementation of the modification process (1, 2, 5).
4. The **elaboration of the consequences** before implementing a registry modification is a very important task and should be undertaken carefully. Knowing where and what differences will occur with the registry modification, and how this will affect the further registry operation, will help in making the right decisions during the registry modification. When thinking about these consequences a registry team should look at every step and part of the registry operation. Are the changes going to reduce the quality of data (e.g. greater number of errors, new biases, lower statistical power, incompatibility etc.), increase the burden on data providers or registry participants, cause delays in reporting of results, increase the operating costs etc. are just a few examples that need to be considered.
5. Registry team should **use experiences that were acquired with the existing registry**. Which things worked well and which did not (bearing in mind every component of the registry operation) represents an important feedback that can be used as an advantage when undertaking a registry change.
6. Registry needs to develop a **good notification protocol** for informing key stakeholders about the registry change. If the stakeholders are not engaged in a decision-making process they certainly have to be adequately informed to understand the rationale for the change, and its benefits. This is especially true for the participating sites/data providers which must be kept informed also about the timeline and implications that registry adaptation will have on the users. For any additional clarifications a registry team should be available, knowing that the change can take people out of their comfort zone and raise their stress and anxiety levels (1,

3, 5). If it proves necessary (e.g. in case of the additional follow-up), patients have to be provided with information about the change as well.

7. It is crucial to **reconsider the ethical and legal requirements**. Registry holder needs to be aware of how the changes in a registry affect the privacy, confidentiality and data access. It is necessary to consider whether the modifications require a new (or first) review from the ethics committee, if the inclusion of informed consent or change in informed consent form is needed, or if the re-consenting is required. In case of changes in stakeholder composition registry holder must also determine whether the previous stakeholders should have access to data and if so, to which one (1). We should look from the other point of view as well. A registry can be modified as a result of (new) legal requirements. It is therefore important that registry holders actively follow the potential changes on this area (e.g. regulation updates) and comply with them if necessary.
8. When implementing changes to a registry dataset (e.g. removing redundant data element, adding a new category/permissible value or modifying a whole value set, introducing a brand new data element, adapting data element's definition, changing a relationship between data elements etc.) a registry team should be aware that comparability over time (i.e. longitudinal comparability) can be great advantage in obtaining new information and knowledge. Therefore, it is advisable to try to **retain the comparability over time** as much as possible. If a registry team is changing a value set/categories of a specific data element, a **mapping** between the old and new value sets usually needs to be done and so-called conversion table designed to clearly show the link between the prior and new value set. It is important that the conversion table is accessible and understandable to every user. The mapping may be a lengthy and intensive process (e.g. problems with the equivalence of prior and new categories) which needs to involve well qualified personnel. Certain changes may make it difficult to match prior value set with the new value set which can result that missing ("unknown") data for subjects, on which data collection has already been done, appear. In that case, these subjects can be reviewed/re-evaluated to update the missing value with the valid one. When this is not feasible it means that the longitudinal comparability is not preserved. This is especially the case when significantly changing a definition on one of the key data element, where the reality often is that everything must start again, meaning there is no comparability with the previous registry period, unless some well-established and validated conversions exist that enable making approximate comparisons.
9. As a result of continuous development in technology, and also due to some other reasons (e.g. moving from one database vendor to another) a registry may go through the process of **data migration** which is a process of transferring of data between storage types, formats, or computer systems (4). Data migration is a complex process that should be carefully managed as, due to its iterative nature, it can easily lead to schedule and cost overruns. First, data on the old system needs to be mapped to the new system. Next, data is extracted from the old system, and, at this point, thorough data cleansing is recommended. If there is any redundant data, it should be removed. When the data is loaded/imported into the new system a data validation needs to be performed to check whether data was accurately transferred causing no errors or data loss. As already mentioned, mapping, loading, and validation steps will probably need to be repeated several times (6, 7, 8). Last but not least, a registry holder must ensure that data migration process complies with the legal requirements.
10. **Appropriately documenting** the registry modification will allow registry users (and other stakeholders) to understand changes that have been implemented in the registry, provide insights into the history of changes and increase the transparency of the registry. Rationale

for a registry change, a description of a change and its practical implications, were there any unexpected problems and how they were solved, are there any other changes that need to be done in a future as a result of a recent change are important items that should be documented.

11. Registry holder should think about whether a modification to a registry requires any **training or other support** for a successful implementation/application of this change. Changing a software for data entry or changing the analytical approach, for example, will probably require more comprehensive training than some other change, such as changing the data element's value set. A registry holder therefore needs to carefully consider how extensive should the training be, who has to be trained (e.g. data providers, registry's staff), what is the most appropriate way of training, and if any supporting material is needed.

9.2 Time to stop? - Stopping a registry

Partly also due to the fact that registries are often open-ended, the activity of stopping a registry does not seem so crucial, time and resource consuming as planning and setting up a registry. However, this does not mean that this activity should be neglected and that there are no important points that contribute to the correct and successful stop of the registry.

When stopping a registry (with this we mean stopping a data collection and ending all other sequential and overarching processes of a registry), first there must be a **clear decision on stopping** it. Setting the tangible and measurable goals/criteria for a registry stop in advance (in a registry planning phase) will help the registry holder to decide on whether the registry should continue with the operation, or if it is time to stop. Such criteria/goals might be, for example, to obtain a certain number of cases in the registry, achieve the desired precision of estimates and/or simply to fulfil the general purpose of the registry[97]. However, the registry is not stopped only when certain goals are accomplished but it should be looked from the opposite side as well - failure to meet registry's predefined objectives or a fact that a registry appears to be unable to meet them in a reasonable time, poor operating results, loss of registry's relevance, lack of a purpose for the continuation, or other serious problems (e.g. discontinued funding, lack of personnel, poor data quality, low patient accrual or significant withdrawal of the registry's participants, ethical issues) could also represent the rationale for ending a registry (1, 9).

When a registry holder together with other stakeholders involved in the decision-making process decides to stop the registry he or she should **establish the communication** with the data providers, and inform also registry users, personnel and, if necessary, any other stakeholders (e.g. patients that are enrolled in the registry) about the registry stop. It is important that the key stakeholders understand the rationale for the registry stop and the consequences that this decision will bring.

Furthermore, a registry team has to decide **what will happen to the registry data**. Will the registry aggregate and disseminate the collected data (as a kind of final report) and/or will archive the data, meaning the data will continue to be available in the future? If preserving a registry data brings important benefits (e.g. to have the insights into the historical data; possibility to perform additional analyses and address the questions that were not covered in prior registry's reports) then archiving might be the right decision. However, it is recommended that the decision about data archiving is discussed in a registry planning phase and not only when it comes to the registry stop.

[97] Planning and consideration of the registry's anticipated size and duration is covered in subchapter 6.2.2

When storing and archiving a registry data a registry holder should take into account several points of data preservation:

- ***Retention period*** (how long the registry will retain the data, considering regulatory requirements if they exist)
- ***Security*** (following the norms of data protection and confidentiality of information a registry should establish policies and procedures to safeguard all data against loss, destruction, unauthorized use, or inappropriate alteration, and if necessary, also policies for proper and secure destruction of data. Some practical procedures for the above issues are authentication of system users, firewalls, back-ups, use of appropriate technology/storage media, policies that address copying data, disaster preparedness, emergency response, disaster recovery and training (10).)
- ***Data for archiving*** (in addition to the main registry data that are usually obtained by the case report form, the registry should preserve also a data entry log that tracks changes and users who made them, allowing registry to find the sources of the potential errors easier. To ensure that data can be correctly (re)used in the future, especially by others, data that are selected to be preserved must be packed with sufficient metadata. According to ICPSR (10) preservation metadata include all the information that is required by an organisation to preserve data, namely descriptive, structural, administrative and technical metadata.)
- ***Monitoring and evaluation*** (monitoring and assessment of the quality and effectiveness of the data maintaining/archiving process enables controlling of the process, finding out if everything is going according to the plan, whether any system errors are occurring, and enabling the adaptation or improvement of internal operations themselves.)
- ***Costs*** (data preservation requires financial, human and IT resources; registry holder should assess whether the funding is and will be available for the long-term maintenance of the registry data).

Finally, it is recommended that a registry prepares **final report** in which its work, achievements, any encountered obstacles, rationale for the stop, and any implications for the future work are clearly described. Along with the report, a registry should provide all the necessary documentation that supports the potential future (re)use of collected data.

References

1. Gliklich R, Dreyer N, Leavy M, eds. Registries for Evaluating Patient Outcomes: A User's Guide. 3rd edition. 2014. Two volumes. Available from: http://www.effectivehealthcare.ahrq.gov/registries-guide-3.cfm.
2. Newton R. Managing Change Step by Step: All You Need to Build a Plan and Make It Happen. Pearson Education, 2007.
3. Harding P., Pooley J. Resource Efficiency and Corporate Responsibility – Managing Change. 2004. available from http://www.oursouthwest.com/SusBus/mggchange.pdf.
4. Bal Gupta S., Mittal A. Introduction to Database Management System. Laxmi Publications, 2009.
5. Queensland Government. Change Management Best Practices Guide: Five key factors common to success in managing organisational change. Available from: http://www.psc.qld.gov.au/publications/subject-specific-publications/assets/change-management-best-practice-guide.pdf.
6. AHIMA. "Data Mapping Best Practices." *Journal of AHIMA* 82, no.4 (April 2011): 46-52. Available from: http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_048788.hcsp?dDocName=bok1_048788#Notes.
7. Computer Economics. Ensuring Success of Data Migration (April 2008). Available from: http://www.computereconomics.com/article.cfm?id=1329
8. SAS Institute Inc. Enhancing Your Chance for Successful Data Migration - Critical steps for creating data migration solutions that balance cost and rapid delivery. 2009. Available from: http://www.sas.com/resources/whitepaper/wp_5969.pdf
9. Rothman J. K., Haas J. When Should a Patient Registry End? Draft White Paper for AHRQ Patient Registries Handbook II. 2009. Available from: http://www.effectivehealthcare.ahrq.gov/repFiles/draftDocuments/2009_0817StoppingARegistry.pdf
10. Inter-university Consortium for Political and Social Research (ICPSR). 2009. "Principles and Good Practice for Preserving Data", International Household Survey Network, IHSN Working Paper No 003, December 2009. Available from: http://www.ihsn.org/home/sites/default/files/resources/IHSN-WP003.pdf
11. NISO. Understanding Metadata. Bethesda, MD: NISO Press; 2004. p. 1-20. Available from: www.niso.org/ publications/press/UnderstandingMetadata.pdf.

10 RE-USE OF REGISTRY DATA

10.1 Background

Re-use of information in general is a current issue in informatics and for health informatics in particular. In 2012 the International Medical Informatics Association organised a summit in Brussels with the title "Trustworthy re-use of health data". The title in itself points out that re-use of health data is a sensitive issue and it is important to find ways, where re-use can be done in a trusted manner.
The conclusion of the summit has been published in the article referenced above. The participants considered various scenarios of re-use, with a focus on re-use of EHR data. In the following sections we try to show that re-use can be done at different levels, and all registries re-use clinical data in a certain sense, but on a higher level. Data stored in the registries can be re-used again for further purposes.

But first, in order to avoid confusion it is necessary to clearly define what is meant by re-use of registry data.

10.1.1 Definition of re-use

According to information theory, information is "something about something" i.e. a series of symbols that represents something else (1). For our purpose it is important to understand from this, that all information is only an abstraction of the thing (event or phenomenon) that is represented. No representation can completely describe the represented entity. Due to the abstraction, some features of reality are neglected, and only the relevant attributes of the real world entity are expressed. The best example for this is when we use identifiers to denote human beings. A "social security number" refers unanimously to a real person but nothing or only a very few attributes (e.g. gender, birthdate etc.) can be expressed by such a series of digits or characters.

As a consequence: **all reasonable representations are purpose dependent**. For a given purpose some features are relevant while others are not. The effective use of the information depends on appropriate selection of relevant features. Naturally, the relevancy depends on the purpose. A very good example for this is the different kinds of maps about the same territory. Maps for touristic purposes will be totally different e.g. from maps for public administration and these differences explain why a map created for some purpose is difficult or even impossible to use for another.

Re-use of information means cases where some information recorded for a given purpose is to be used for another one.

10.1.2 Re-use the in context of patient registries

The fact that all information is purpose dependent generates serious limitation of re-use, which of course does not mean that no information can be used for any other purpose but for which it was originally recorded.

Sometimes there is a temptation for purposeless data collection: i.e. trying to store everything without defining the goals and future usage of data. As data acquisition and storage costs decreases, this temptation could be larger and larger. In case of patient registries the privacy concerns prevent us from yielding to the temptation (moreover in most European countries legislation makes it impossible). But it is also important that purposeless data gathering is not a good way: it often leads to bad quality of the collected information.

Registries – often and preferably – are realisations of information re-use. Perhaps with the exception of registries created for public health purposes, it may be difficult to justify collecting data just for registry purposes, that are not relevant or not needed for clinical purposes (this is especially true for especially hospital-based registries). In this case the primary reason for storing some patient data is the clinical need, and **registries should store extracts and abstraction of clinical information**. This requirement will be addressed in Section 10.4

Summing up these considerations we may state the followings:

- Designing and operating registries should serve well defined purposes
- The normal way of using registry data is to serve the defined purpose
- Re-use of registry data is using data for any other purpose than originally planned for

The next two sections provide brief answers to the emerging questions on why to re-use data and whether re-use is possible.

10.2 Why to re-use?

If somebody understands why data collection is always purpose dependent, a reasonable inference is that any re-use is somehow a misuse of information. Sometimes it is really the case. Using ICD codes within the course of individual patient care for treatment purposes is a typical example, since ICD was planned for public health data collection, its granularity and reliability is far from the need of clinical care.

But we cannot state that re-use of data is always a mistake. While all information is an abstraction of reality, even abstract data can still preserve many important features of reality beyond what was considered in the design phase of data collection.

Practically, in many health systems a vast amount of information is collected and poorly utilised. If re-use is possible it is more advantageous than separate data collections for all different purposes. Re-use is a much more cost effective and straightforward way.

10.3 Is re-use possible?

In spite of the above mentioned concerns or limitation in many cases it is possible, however we always have to be careful. E.g. data, collected originally for health care reimbursement often can be used for quality assessment or capacity planning. But we have to know, that using some data for financial purposes always induce some distortion. Indeed, all observations distort somewhat the phenomenon that we want to observe. (It is a basic law in quantum physics, but also applies for many social phenomena). It is important to measure or at least estimate how large the distortion is, in order to draw right conclusions from noisy data.

10.4 Re-use of data

10.4.1 Re-use of clinical data in registries

It is a critical success factor for designing and implementing registries that the administrative burden of health care providers is minimised. Data collection systems should be as much automated as possible. The proper way is to extract all relevant data for a registry from the clinical documentation without much human workload. But this “extraction” is not always so straightforward. In Hungary there is a registry for premature newborns, and this registry stores information on administration of surfactants. In the data model of the registry this is just a YES/NO rubric. Naturally, there is no such a rubric in the patient records, but of course all drug administrations (including surfactants) are recorded. In order to automate the data submission to the registry, an abstraction process has to be implemented that is able to extract the information regarding which drugs are surfactants.

Therefore re-use of clinical data for registry (and other public health) purposes is usually and typically an abstraction process based on some sort of knowledge.

10.4.2 Re-use of spatial data

Using geographic data in different application domains has resulted in large amounts of data stored in spatial databases and these spatial data can be re-used for health purposes, sharing accurate geographic references to track communicable diseases by place and time, link various geo-referenced environmental factors such as air pollution, traffic, and built environment with geo-referenced health outcome data to analyze potential associations and identify risk factors. Such spatial data have been extensively used in health domain in recent years. However, re-use of spatial data collected outside of the health domain has still an enormous potential for re-use related to the health domain.

10.5 Types of re-use of registry data

10.5.1 Internal re-use

Whenever an authority establishes a patient registry, the tasks, roles and goals of the registry are defined. The data-model of the registry is ideally designed based on these task. It may happen however that later the collected data is used for further purposes. E.g., if the original task of a cancer registry is to measure cancer incidence, but later on the same data is used within the registry to estimate cancer prevalence, then this is a case of internal re-use. The term ‘internal’ refers to the fact that the re-use happens in the same organisation operating the registry. Sometimes such internal re-use requires additional data from different sources. In the mentioned example this could be cancer mortality data.

10.5.2 International comparison (same purpose, different context)

Patient registries for the same disease (same purpose) have been set up in different countries. Obviously, cancer registry is the best example, since most of the countries operate some sort of cancer registry. Evidently there is a benefit in cross-country comparison of their data. Due to lack of standardisation it is often not so easy. This applies not only for the standardisation of their data structure, but also for the aim, scope and organisation of the registries. E.g. data of population-based registries are difficult to compare with hospital-based registries. Comparison of national (one single registry for the whole country) with country level data aggregated from regional registries may raise methodological problems.

10.5.3 Cross-registry comparison (correlation between diseases)

Morbidity patterns are evergreen research topics. Correlation between disease incidences either from genetic or geographic aspect is a subject of tremendous number of studies. Using patient registry data for this purpose can be done on individual or aggregated level.

Cross-registry comparison of registry data at individual level implies the possibility to merge data about the same person from different registries. However, this does not necessarily require the use of personal data. Such investigations can be performed also on pseudo-anonymous data as well. Different scenarios are possible. Consider two registries for two different diseases. If we want make a comparison among them, the following options emerge:

a. When two registries use personalised data based on the same identifier (e.g. social security number), to make a comparison without infringing privacy, one possibility is to have both registries remove the IDs from the records, and replace by an artificial identifier, or pseudonym and merge them by this artificial identifier.
b. Another option is to aggregate the data in the two registries separately and compare them at aggregate level. This method necessarily has some limitation.
c. Datasets with common identifiers can be merged on a secure server with encrypted data transfers, and a de-identified dataset is generated on the server and provided back to researchers.

10.5.4 Comparison with information outside the health domain (e.g. environmental, economic, social etc. data)

With large amount of environmental, economic, social, spatial data generated and available in different databases and registries, these data can be linked with health data and secondary data analysis and comparisons are possible. Interactions of these factors could provide useful information for researchers, policy makers in both health and non-health domains.

10.6 Re-use of aggregations vs. re-use of elementary data

Patient registries typically store data about individual patients and create statistics from the individual data. Such statistical data can be used in many research or policy planning activities, and it can be integrated to other statistical data (e.g. comparing morbidity data with economical or social data etc.). Detailed studies, however, need to process the elementary data, when matching data from various sources is not possible on aggregated level. Re-use of elementary (individual) data is, of course, much more sensitive and problematic from the privacy perspective. Therefore, it is absolutely important to understand the nature of various kinds of elementary and aggregated information.

10.7 Definition of Possible Types of Data

10.7.1 Aggregated Data (Indicator Compilation)

Data about a single entity (legal or natural person, institution, etc.) is called individual data. Data aggregation is a process where data and information is searched, gathered and presented in a report-based, summarised format that is meaningful and useful for the end user or application. In statistics, aggregated data denotes data combined from several measurements. When data are aggregated, groups of observations are replaced with summary statistics based on those observations. Data aggregation may be performed manually or through specialised software.

Aggregated data are usually calculated from individual data by summing or averaging values of some data-type attribute of a set of individuals (population). For example, “body weight of John Smith is 76 kg” is an individual data. “The average body weight of adult citizens of London is 76 kg” is an aggregated data.

Health indicators such as community, public health, or occupational health indicators are typically aggregated data. Using aggregated data, various reports can be generated containing a compilation of selected indicators measuring health status, non-medical determinants of health, health system performance, and finally community and health system characteristics.

Patient registries can serve as a valuable source for health indicators such as morbidity and mortality rates.

Aggregated data are generally considered harmless from privacy perspective and hence can be used without any legal restriction in most cases. The normal way how most statistics works, is that we count a total amount of some phenomenon and then divide along some attributes. E.g., first, the total number of deaths is counted in a country then it is divided according to gender, age group, geography or cause of death. By combining of divisions along different attributes we often get very small numbers and run into the risk of possible identification of some individuals. For that reason in most countries legislation restricts the publication of aggregated data where there are less individuals than a certain limit behind each number. This limit varies typically between three and five.

It is reasonable, however, to make a distinction between publication (making data available to everybody, without any control of further use) and use of such kind of data e.g. for research purposes. In the latter case it is possible to control the proper use of data e.g. by supervision of an ethical committee.

An increasing number of global patient registries have been establised in recent years, which could especially be valuable for rare health conditions to help biomedical research. One example for a global patient registry coming from the US National Institute of Health (NIH), National Center for Advancing Translational Sciences (NCATS):

“The goal of the NIH/NCATS Global Rare Diseases Patient Registry Data Repository (GRDR) program is to serve as a central web-based global data repository that aggregates coded patient information and clinical data to be available to investigators to conduct various biomedical studies, including clinical trials. The aim of the program is to advance research for many rare diseases and apply to common diseases as well.

Data is collected and aggregated from rare disease registries in a standardized manner, linking the registry data to Common Data Elements (CDEs) using nationally accepted standards and standard terminologies. The aim is that through standardization, registries will be interoperable to enable exchange and sharing of data. Each registry will be free to develop its own survey questions according to patient preference and the nature of the disease."(2)

10.7.2 Anonymised Data

Anonymisation is a procedure to completely remove any information from the data that could lead to an individual being identified.

Oxford Redcliff Hospitals Confidentiality Guidelines states: "[Anonymous] data concerning an individual from which the identity of the individual cannot be determined"(3).

A Bristol University ethical document defines the following anonymous data types:

> "Anonymised data are data prepared from personal information but from which the person cannot be identified by the recipient of the information.
>
> 'Linked anonymised data' are anonymous to the people who receive and hold it (e.g. a research team) but contain information or codes that would allow the suppliers of the data, such as Social Services, to identify people from it.
>
> 'Unlinked anonymised data' contain no information that could reasonably be used by anyone to identify people. The link to individuals must be irreversibly broken. As a minimum, unlinked anonymised data must not contain any of the following, or codes traceable by you for the following (4):
>
> - name, address, phone/fax number, email address, full postcode
> - NHS number, any other identifying reference number
> - photograph, names of relatives"

The main difference between anonymous and pseudo-anonymous data is that the former does not contain any key to merge or collect different data about the same individuals. Both data are individual, i.e. contain information about a single person. For example, if all personal identifiers are stripped out from a death certificate (name, birth date, home address, social security number etc.) it is still about a single individual. However, such a document cannot be merged anymore with other (either anonymous, pseudo-anonymous or personal) data about the same person.

Using fully anonymised data is relatively safe from privacy perspective, however, if one is in possession of additional personal data that allows joining anonymous and personal data, than privacy concerns emerge.

On the other hand, usability of anonymous data is limited if multiple recording and counting is possible. If there is any chance to have more than one record about the same individual, then calculations will be incorrect (e.g., if we have salary data without personal identifier and one person can have multiple employments, then average incomes cannot be calculated.)

This is the main reason to use pseudo-anonymised (pseudonymised) data.

10.7.3 Pseudo-anonymised Data

Generally speaking, pseudoanonymisation (or pseudonymisation) is a procedure to break the link to the data subject by replacing the most identifying fields within a data record by one or more artificial identifiers, or pseudonyms.

Pseudo-anonymised (or pseudonymised) data means that information is represented in a way that allows collecting all data corresponding to the same person without the possibility to identify the real person. Such data cannot include personal identifiers such as names and addresses of the person.

However, there is a disagreement regarding the interpretation of what ‘possibility’ means. For example, according to recent Hungarian legislation, the possibility of re-identification exists if the handler of the data is in the possession of the technical tools necessary to re-identify the person. There are much stronger interpretations in some European countries that say if there is any chance to re-identify (e.g. by using additional information) then the data should be treated as personal. Again, other regulation considers the effort necessary to recognise the real persons, saying that data should be treated as personal only if reasonable effort is enough to re-identify.

There are other definitions of pseudo-anonymisation. For example, the National Health Service (NHS) in UK uses the following definition:

“The technical process of replacing person identifiers in a dataset with other values (pseudonyms) available to the data user, from which the identities of individuals cannot be intrinsically inferred, for example replacing an NHS number with another random number, replacing a name with a code or replacing an address with a location code.”(5)

This definition interprets the possibility of re-identification again in another way. It says that the data are pseudonymous if the real individuals cannot be "intrinsically" inferred, i.e. just by using the data. If data need to be merged with any other (extrinsic) information in order to infer to real persons, than it is not personal data.

Independently from which definition is worth to accept, it is clear, that the use of such data is extremely important and unavoidable for health research and evidence-based health policy.

On the other hand, it is clear that using such data requires special regulation. For example, current Hungarian legislation says that any data handled by governmental bodies are either public or personal. Pseudo-anonymous data is not mentioned in the legislation. Only the law of statistics mentions that statistical bodies must not publish data with less than three entities in any given cell presented. However, publication of data (i.e., making data available for everybody) and using data for research purposes are different.

A European directive on using pseudonymous data that defines this type of data and the conditions of use of them would be welcome.

Privacy concerns about using pseudo-anonymised data are not much more serious than using anonymous data. If we are able to join different records of the same individual then it is somewhat more likely to be able to join these data to some external personal information and re-identify the person. However, it is important to keep in mind, that in both cases it is necessary to obtain some personal data about the same individual, otherwise re-identification is not possible.

10.7.4 Personal Data

Several laws and regulations exist around the world, which include a definition for personal data. Personal data is defined in EU directive 95/46/EC, for the purposes of the directive, as the following:

> Article 2a: 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity (6).

Personal data, personally identifiable information may be categorised into two main groups:

1) Personal data, which are often used to identify the individual such as full name, home address, date of birth, birth place, national identification number, genetic information, telephone number, e-mail address, vehicle registration plate number, credit card numbers, biometric records, etc.
2) Personal data, which may be shared by many people and may identify the individual. Examples include city, county, state, country of residence, age, race/ethnicity, gender, salary, job position, etc.

However, it is important to keep in mind that sometimes multiple pieces of information, none sufficient by itself to uniquely identify an individual, may uniquely identify a person when combined.

Because a very rare disease itself could be personally identifiable information, collecting and publishing information about rare diseases in patient registries requires careful considerations.

10.8 Cross-border Use for Public Health

There are several initiatives and examples for cross-border use for both public health and research purposes of various data including patient registries' data. Sharing information, data exchange across the borders could serve several purposes.

10.8.1 Cross-country Data Comparison, Surveillance

Data exchange and information sharing accross the borders would allow cross-country surveillance, monitoring, and comparison of data. For example, disease rates, trends could be compared by various demographic and clinical characteristic. EUROCAT, European Surveillance of Congenital Anomalies, which collects data on birth defects from several regional and national birth defects programs to generate trends, is a good example for that, as well as the European Network of Cancer Registries (ENCR), which collects and regularly disseminates information on incidence and mortality from cancer in the European Union and Europe. The European Surveillance System (TESSy) is a highly flexible metadata-driven system for collection, validation, cleaning, analysis and dissemination of data on communicable diseases. Its key aims are data analysis and production of outputs for public health action (7).

10.8.2 Outbreak Alerting and Communication

Sharing cross-border information on communicable (infectious) diseases has great significance on the EU level or international alerting of outbreaks and potential pandemics. Severalinfectious diseases spread from human to human and these do not respect country borders. Therefore, effectively tracking and preventing, or at least minimising the consequence of an outbreak, to the extent it is possible, prompt information sharing and data reporting is extremely important. An example for this is the novel H1N1 influenza virus outbreak in recent years. However, these emerging diseases are usually not related to or part of patients' registries. Nevertheless, this information may be linked to special patient's registries (such as vulnerable patient groups) that could help alerting them and also better understanding the course and treatment of disease. In this highly globalised and mobile world, transmission of many diseases is more frequent and possible than ever before in recorded history.

10.8.3 Bioterrorism Threat

Sharing data among specific patient registries could even be helpful in the case of a bioterrorism threat to inform and protect vulnerable patients, groups in a timely manner (e.g., patients with immune deficiencies). The anthrax threat and infections in the United States a few years ago showed the potential danger and need to set up harmonised reporting systems. Patient registries may also benefit from sharing information if a functional cross-border data exchange system would be in place.

10.8.4 Identification of Best and Cost-effective Practices

Data sharing could help searching for and identifying best and cost-effective practices by health care provider such as timely diagnosis and treatment, professional recommendations. For example, identification of best practices for reducing hospital readmissions could lead to the implementation of such practice by other health care providers, which could lead to significant cost reduction, and reduce avoidable hospital readmissions.

10.8.5 Referral to Services, Establishing New Services

Mapping the distribution of patients by well-defined smaller geographical unit could help to refer these patients to the available services on a European level. At the same time, lack of services in certain geographical areas can also be identified and a new service may be established. Taking into account travel time and distance is very important from both the service providers' and the patients' point of view. The less time and distance is needed to travel, the better, especially in urgent care, to save life and also costs.

10.8.6 Public Health Research

Data exchange could provide information for basic and applied research (more information on this, see Section 8.3), and help also understand various demographic and clinical characteristics, long-term outcomes of specific diseases, comorbidities, and effective prevention and intervention efforts on a European or global level.

It is important to differentiate ad hoc, irregular cross-border data sharing, data communication, which could also have significant public health value, from public health surveillance, which is, by definition, an ongoing, systematic data collection in a timely manner.

10.9 Cross-border Use for Research Purposes

10.9.1 Issues

The use of registry data for public health and research purposes in cross-organisational and cross-border setting is becoming more and more important. For example:

- increasing mobility increases the risk of cross-country infections,
- for rare conditions setting up international databases or exchanging data is crucial to establish large enough cohorts to study a specific population or specific rare conditions such as genetic disorders, congenital malformations, and metabolic conditions.

Harmonisation of registry data could lead to reduced cost of managing and using these data, and better quality data would be available for analyses and various indicators.

10.9.2 Risk Factor Studies

Registry data could provide valuable information for epidemiologic studies to analyse potential risk factors for diseases. Sociodemographic data such as race/ethnicity, gender, age can help understand whether there is an increased risk among certain groups of people. Data on environmental factors like air pollutants, agricultural activities such as pesticide exposure can be linked and associations can be analysed. Natural disasters, neighbourhood effects on health can also be studied. Data on medication/drug use and adverse outcomes could be valuable information for drug safety studies.

10.9.3 Genetic Research

Registry data may include information on genetic analysis (molecular or cytogenetic), or the registry data may be linked with bio banks, biological samples that allow further genetic analyses. Gene mutations may be identified for rare genetic conditions. Registries could potentially contribute also to gene-environment correlation studies. Several genetic research initiatives are going on in Europe and researcher look for data from different sources including patient registries.

10.9.4 Clinical and Therapeutic Research

Registry data could also help clinical research studies to look at treatment options, may include data from clinical trials for new medications and medical devices. Using available data researchers can analyse clinical parameters, effectiveness, and outcomes. Inequalities and disparities in health outcomes by country or other factor could drive establishing new or improved clinical guidelines and recommendations, and inform policy makers.

10.10 Compatibility, comparability and interoperability

10.10.1 Data compatibility

The integration of multiple data sets from different sources requires that they be compatible. Methods used to create the data should be considered early in the process, to avoid problems later during attempts to integrate data sets.

“Compatibility is the capacity for two systems to work together without having to be altered to do so. Compatible software applications use the same data formats. For example, if word processor applications are compatible, the user should be able to open their document files in either product” (13).

“Another factor that should be considered is the compatibility of existing data sets. Frequently, a data search may reveal multiple sources of similar data types, but the metadata may reveal that the individual data sets are not compatible, as the data have not been collected in a consistent manner …” (14).

For registries it means that data created by one registry can be imported into another, without manual data manipulation. Such a scenario is reasonable and necessary e.g. when in a country data collection is carried out at regional level, and regional registry data is used to build up a national registry. Similarly if a European registry is built on member state registries. Data compatibility is usually considered at technical level (same data structure and format, character coding etc.) as in the mentioned example with word processors. In case of patient registries the issue is more complex however. If we want to compile a national registry from regional ones, this technical compatibility is a prerequisite only, but far not sufficient. Such compilation can be done on the level of elementary data (e.g. data of patients registered in each registry is to be sent to the national registry). But it also can be done at aggregated level, where only sums and (weighted) average numbers are sent. In both case we have to be sure e.g. that each patient is registered in one regional registry only, so double counting is excluded. It is also important, that there are no definitional of methodological differences among the regional registries, or at least we have to be aware of such differences.

Summing up, compatibility of registry data has the following requirements:

1. **Technical compatibility** of data (identical or convertible data structures, formats, coding schemes etc.)
2. **Comparability** see Section 10.10.2
3. **Double counting exclusion** See the problem of populations in Section 10.12

10.10.2 Comparability

Comparability is different from compatibility. Colloquially speaking, comparability means that we have to be sure to compare apples with apples and not peaches. Whenever we compare data form different registries we have to be sure that the observed differences are attributable to real differences in the thing we want to measure, not some artefacts that are consequences of external or irrelevant circumstances. Full comparability occurs exceptionally, i.e. raw data of registries are hardly comparable.

The more common situation is that we do know the differences that make raw data incomparable, and can find ways to resolve them. The most typical example is the standardised death rates. Raw mortality figures of different populations are practically never comparable due to the different age structure of different populations. By standardisation we can project our raw data to a standard age distribution that enables us to compare mortality data from very different countries.

In other cases comparability problems arise from different definitions and categorisations. Such entities like ‘hospital’, ‘hospital bed’, ‘long term care’, ‘community care’ are often interpreted differently, and data that are built on such entities are sometimes hard to compare. Contrary to the standardised death rate example, in such cases we cannot always fully resolve the problem. Sometime we have to settle for relative comparability. E.g. if we know that ‘number of hospital beds’ in country A covers more kinds (e.g. new-born incubators included) than in country B, but even so country A has less hospital beds than country B, then we can be sure that there is a real difference, but not in the reverse case.

The most important issue is to be aware of comparability issues. At least we have to know what is compared to what. To achieve the possible optimum, the following conditions have to be met:

1. **Sufficiently detailed metadata should be available**. Metadata should describe what is counted in a registry, with what exceptions, how the measured entities are defined, what data collection methodology was applied etc.
2. **Additional data necessary for standardisation should be available**. If there are known external or irrelevant factors that influence the thing to be measured (e.g. as age distribution influences mortality) then these data must be available in order to eliminate these effects.

10.10.3 Interoperability

Briefly and generally speaking, interoperability is the ability of systems to work together (Section 10.11 explains the technical aspects of interoperability in detail). In information technology it usually refers to information systems that are able to use each other’s data. Looking a bit more deeply, what ‘use’ means in this context, we realise that different levels of interoperability exist.

Interoperability has a huge literature and it is not the aim of this study to give a comprehensive overview of the various approaches, definitions and theories behind it. For the purpose of this chapter, we define – somewhat arbitrarily three levels of interoperability:

Technical interoperability means that data arising from a system can be technically read by the other one, and presented at least in human readable format to the users. Such a case is when in a hospital information system (HIS) discharge reports are created in pdf format. Such documents can be imported into another HIS that is able to store the report assigned to the particular patient and present them to the users of the second system.

Functional interoperability is a higher level, where data from one system can be imported to another in a way that allows using these data as if they were originated in the second system. The available functions depend entirely on the functionality of the recipient system.

Semantic interoperability is a stronger type of functional interoperability that hold between systems that are able to automatically interpret the information.

Semantic interoperability between registries implies that the recipient system is not only able to handle the received information but also able to automatically interpret it. It is possible that two registries that collect data for the same disease use different disease coding systems. Functional interoperability of such registries implies in such case that the disease codes can be imported, but does not imply that the semantically identical codes are recognised or codes from one scheme is converted to the other one. (See the problem of mapping in Section 10.11.3.2).

Semantic interoperability comes in question only if (at least one of) the systems are able to process information semantically: it makes inferences, or actions that depend entirely on the meaning of information, not on its syntax. Such semantic functions are hard to imagine without using some sort of ontology.

10.11 Interoperability Standards and Approaches for Data Exchange

10.11.1 General Concept

The concept of functional interoperability is to permit one system (sender) to transmit data to another system (receiver) to accomplish a specific communication in a precise and unambiguous manner. To achieve this, both systems have to know the format and content, and understand the terminology used. Using standard terminology can help database and system developers, and can facilitate exchange of data among various systems.

The recognition of the need to interconnect health related applications and exchange data led to the development and enforcement of interoperability standards. The following sections explain the standards used for structuring and encoding data.

10.11.2 eHealth standards

Exchanging and interchanging data in the health care domain in a seamless manner is becoming critically important. Lots of efforts have been made in this area to develop standards, which have obvious economic benefits as well. Here are a few examples of current standards developed and used for data exchange.

- Health Level 7 (HL7): HL7 and its members provide a framework (and related standards) for the exchange, integration, sharing, and retrieval of electronic health information. These standards define how information is packaged and communicated from one party to another, setting the language, structure and data types required for seamless integration between systems. HL7 standards were originally developed to exchange data among hospital computer systems. HL7 standards support clinical practice and the management, delivery, and evaluation of health services, and are recognized as the most commonly used in the world.
- The National Council for Prescription Drug Programs: The US National Council created data-interchange standards such as drug claims for the pharmacy services sector of the health care industry.
- Data Interchange Standards for Bioinformatics: These standards were developed to support data exchange among various databases in bioinformatics and have gained popularity.
- Health Informatics Service Architecture: The European Committee for Standardization (CEN) Standard Architecture for Healthcare Information Systems (ENV 12967), Health Informatics

Service Architecture or HISA is a standard that provides guidance on the development of modular open information technology (IT) systems in the healthcare sector.

- openEHR: It is a virtual community working on interoperability and computability in e-health. Its main focus is electronic patient records (EHRs) and systems. The openEHR Foundation has published a set of specifications defining a health information reference model, a language for building 'clinical models', or archetypes, which are separate from the software, and a query language. The architecture is designed to make use of external health terminologies, such as SNOMED CT, LOINC and ICDx. Components and systems conforming to openEHR are 'open' in terms of data (they obey the published openEHR XML Schemas), models (they are driven by archetypes, written in the published ADL formalism) and APIs. They share the key openEHR innovation of adaptability, due to the archetypes being external to the software, and significant parts of the software being machine-derived from the archetypes. The essential outcome is systems and tools for computing with health information at a semantic level, thus enabling true analytic functions like decision support, and research querying.
- EN/ISO 13606 - Electronic Health Record Communication: This European and ISO standard defines the means to communicate a part or all of the Electronic Health Record (EHR) of a single subject of care. The standard can be seen as a harmonisation of opeNEHR and HL7.
- ESRI developed spatial interoperability standards for public health and health care delivery (8).
- Extensible Markup Language (XML) is the most widespread markup languages used for data exchange. It defines a set of rules for encoding data structures (including documents) in a textual data format which is both human-readable and machine-readable. It is defined by the World Wide Web Consortium's (W3C) XML 1.0 Specification (23).
- The Resource Description Framework (RDF) and RDF-Schema (RDFS) are W3C recommendations used as a general method for conceptual description or modeling of information in web resources, using a variety of syntax notations and data serialization formats, the most used is XML. It is also used in knowledge management applications (24).
- The Web Ontology Language (OWL) is a family of knowledge representation languages for representing ontologies. The OWL languages are extensions of RDF by constructs allowing the representation of formal semantics and. OWL1 has been extended with aditional features in 2009, becoming OWL2. Both languages are supported by Protégé and DL reasoners such as FaCT++, HermiT, Pellet and RacerPro. OWL and RDF have attracted significant academic, medical and commercial interest (25).
- Simple Knowledge Organization System (SKOS) is a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, or any other type of structured controlled vocabulary. SKOS is part of the Semantic Web family of standards built upon RDF and RDFS, and its main objective is to enable easy publication and use of such vocabularies as linked data (26).
- Common Terminology Services, Release 2 (CTS2) is a Health Level-7 (HL7) and Object Management Group (OMG) specification for representing, accessing and disseminating terminological content (27). It is an extension of HL7 Version 3 Standard: Common Terminology Services, Release 1 (28).

In the United States the “Public Health Data Standards Consortium was invited by the Integrating the Healthcare Enterprise (IHE) to start a Public Health Domain at IHE. IHE is a collaborative of clinicians, administrators, standard development organizations and health information technology (HIT) vendors that drives the adoption of standards to address specific clinical needs through the development of the technical specifications for the software applications. PHDSC and IHE are collaborating to enable interoperability across clinical and public health enterprises." (9).

10.11.3 Coding schemes, terminologies

The idea of representing certain entities by codes instead of natural language descriptors goes back to many centuries. The original cause of using codes was twofold. An important aspect was the need of unambiguity, either across or within languages. The other reason was to represent the entirety of a domain by a limited number of concepts to conduct statistical studies. In the modern age the computational tractability became another point.

Most coding systems are based on some classification: entities of the given domain are arranged into a – usually hierarchical – structure. One of the earliest problems with classification was the problem of multiple hierarchies. For example, diseases can be classified by location (according to the primarily affected organ), by aetiology (infectious, acquired, hereditary etc.), by epidemiology (sporadic, epidemic, etc.), or by pathology (neoplastic, metabolic disorder, etc.). Therefore a certain disease can be a member of many different, partially overlapping classes. The problem of multiple hierarchy is quite ubiquitous, it applies for nearly all large classifications, not only in healthcare domain.

It depends again on the purpose, which dimension should be considered as the main aspect of classification. This is one of the most important reasons, why more than one classification exists for most of the medical domains. There are other reasons of course, like differences in granularity, content coverage, availability in different languages, etc.

For public health purposes, however, the *International Classification of Diseases* (ICD) is perhaps the most frequently used classification system, although different versions of it are in use.

The terms – terminology, nomenclature, and vocabulary – are often used interchangeably. However, there are differences in these terms. Terminology can be defined as a set of terms representing the system of concepts of a particular subject field. Nomenclature is a system of terms that is elaborated according to preestablished naming rules. Vocabulary refers to a dictionary containing the terminology of a subject field.

10.11.3.1 Most important terminologies

There are various terminologies used in the health domain. Here is a partial list of terminologies widely accepted and used either globally or by many countries.

- International Classification of Diseases and its clinical modifications: this is one of the best known terminologies, which was first published in 1893, and has been revised at roughly 10-year intervals, by WHO. The most recent version is the 10th revision (ICD-10). WHO has been working on the 11th revision for a few years. In the United States the National Center for Health Statistics published a clinical modification of ICD-9 and now ICD-10 by adding an extra digit to the codes to provide an extra level of detail (ICD-9-CM; ICD-10-CM). The Royal College of Paediatrics and Child Health (formerly British Paediatric Association) also created a modified and extended version of ICD-9 and ICD-10 codes for birth defects (congenital anomalies).

- International Classification of Primary Care: This classification includes over 1000 diagnostic concepts that are partially mapped into ICD.
- Systematized Nomenclature of Medicine (SNOMED): Originally called SNOP (Systematized Nomenclature of Pathology), it has been developed by the College of American Pathologists to describe pathological findings using topographic (anatomic), morphologic, etiologic and functional terms. The current version, SNOMED CT (SNOMED Clinical Terms) was created in 1999 by the merger, expansion and restructuring of SNOMED RT (SNOMED Reference Terminology) and the Clinical Terms Version 3 (formerly known as the Read codes), developed by the National Health Service of the United Kingdom. Since 2007, SNOMED CT ismaintained by the IHTSDO (International Health Terminology Standards Development Organisation).
- GALEN and GALEN-In-Use projects in Europe: the aim was to develop standards for representing coded patient information. The consortium developed the GRAIL concept modelling language, the structure and content of the GALEN Common Reference Model. It also created tools to enable the further development, scaling-up and maintenance of the model.
- Logical Observations, Identifiers, Names, and Codes (LOINC) in the US, and a similar EUCLIDES work in Europe: LOINC was created to represent laboratory tests and observations but later included also nonlaboratory observations such as vital signs. A similar work (EUCLIDES) has been done in Europe.
- WHO Drug Dictionary, ATC codes: The Drug Dictionary is an international classification of drugs by name, ingredient, and chemical substance). It is used by pharmaceutical companies, clinical trial organizations and drug regulatory authorities for identifying drug names in spontaneous ADR reporting (and pharmacovigilance) and in clinical trials. The dictionary was created in 1968 and it is regularly updated Since 2005 there have been major developments in the form of a WHO Drug Dictionary Enhanced (with considerably more fields and data entries) and a WHO Herbal Dictionary, which covers traditional and herbal medicines. Drugs are classified according to the Anatomical-Therapeutic-Chemical (ATC) classification.
- Unified Medical Language System (UMLS): started by US National Library of Medicine in 1986, it is a quarterly updated compendium (Metathesaurus) of biomedical terminologies, providing a mapping structure among these vocabularies and thus allows the transcoding among various terminology systems. Altogether, it contains over a million concepts and 5 million terms which stem from the over 100 incorporated terminlogies. Each concept in the Metathesaurus is assigned one or more semantic types, and they are linked with one another through semantic relationships. The Semantic Network provides these types and relations: there are 135 semantic types and 54 relationships in total. UMLS can be used to enhance or develop applications, such as electronic health records, classification tools, dictionaries and language translators. It can be also used for information retrieval, data mining, public health statistics reporting, and terminology research.

10.11.3.2 Mapping between classification systems

Whenever we are faced with the Babel of classification and coding systems, a trivial idea is the (automated) mapping (conversion) from one to another. At first insight, it can be done easily, e.g. by a simple cross-reference table, that contains the corresponding code pairs (triplets, etc.) Since coding systems are not just a set of code values, but – as we mentioned – most of them are built on a classification, the matter is not so easy. Usually the categories of one classification do not fit entirely in the categories of the other. Unless the underlying classifications are totally identical, no mapping is possible between two coding systems without distortion. Theoretically, a special case is also possible: if one classification is a mere subset of another, then there is an unambiguous mapping from the former to the latter but not vice-versa.

10.11.4 Ontologies and data structures

Computer-based patient records can be improved by the use of ontologies. "An ontology specifies the conceptualization of a domain and is often comprised of definitions of a hierarchy of concepts in the domain and restrictions on the relationships between them."(10)

An ontology representing the content of an electronic patient record may include (among others) the following:

- Clinical acts (health care flow, surgical and other procedures, etc.)
- Clinical findings
- Disease manifestation, etiology, pathophysiology
- Diagnosis

10.11.5 Mobile health delivery, personalized medicine, and social media applications

Mobile technology, social media, personalised medicine, remote diagnostics could transform health care. The number of e-health applications available for mobile devices steadily increases. Developing communication standards for information and communication technologies to facilitate interoperability among systems and devices, provide privacy and security, and address the needs of the developing world is timely and important.

Personalised medicine, "A form of medicine that uses information about a person's genes, proteins, and environment to prevent, diagnose, and treat disease", is a new area of e-health when personalised medical records are generated (11).

Social media applications related to health are on the rise. Patients often consult medical information online, and turn to social media communities for peer-to-peer support and information. Lot of information can be obtained but careful considerations are needed to filter out useful information (12).

10.12 Problem with populations

10.12.1 Definition of population

Comparability of data of population-based registries requires clear definition of the given population. Without such a clear definition we cannot be sure e.g. that there is no overlap between the populations of the registries. This is especially true within the EU, where free mobility of people increases the probability that the same person is registered in different registries.

The definition of population in general is in itself not without difficulties. Most often, population is defined as a group or collection of individuals inhabiting a certain territory or forming an interbreeding community. There is a proposed definition of population especially for public health purposes saying that "A population (in public health) is a group of persons sharing a common resource."(15)

10.12.2 Inclusion and exclusion criteria

To generate comparable data on a population level, it requires having the same set of inclusion and exclusion criteria (i.e., residency status, socio-demographic data, geographic area, etc), therefore using data from two or more systems or registries could be interpreted in a uniform fashion. For example, the definition of stillbirths (gestational age cut-off point) varies by country and collecting information on the stillbirths population and comparing characteristics and prevalence could lead to false interpretation of data. When comparing rates of population-based registries, the residency status criterion, whether including or excluding non-resident persons living in a defined geographic area, is very important.

10.12.3 Mobility

Free mobility within and across borders makes the establishment of population-based registries (especially in a smaller geographical area) and comparison of data between other registries without the risk of having the same person recorded in two or more databases challenging. National and EU level, or global registries could help eliminate this problem. Communication between systems and linking data on a regular basis could also help finding duplicate records and make data comparable.

10.12.4 Socio-demographic, genetic factors

Variations and differences in socio-demographic and genetic factors such as ethnicity, genetic mutations in certain populations could make it difficult or even nearly impossible to compare some specific data among populations.

10.13 Data Sharing Regulation

Data exchange is sometimes a complex process, and organisations, registries, data providers have to ensure compliance with cross-border restrictions, privacy and confidentiality rules. All member countries of the EU impose restrictions on the sharing of personal information outside the EU. Organisations sharing personal information collected in the EU with service providers based outside the EU need to find ways to comply with these laws (16).

Privacy generally applies to people, while confidentiality applies to information. There are many important reasons to protect privacy and confidentiality.

Privacy is the control over the extent, timing, and circumstances of sharing oneself (physically, behaviorally, or intellectually) with others. For example, persons may not want to be seen entering a place that might stigmatize them, such as a pregnancy counseling center clearly identified by signs on the front of the building. The evaluation of privacy also involves consideration of how the researcher accesses information from or about potential participants.

Confidentiality pertains to the treatment of information that an individual has disclosed in a relationship of trust and with the expectation that it will not be divulged to others in ways that are inconsistent with the understanding of the original disclosure.

Maintaining privacy and confidentiality helps to protect participants from potential harms including psychological harm such as embarrassment or distress; social harms such as loss of employment or damage to one‘s financial standing; and criminal or civil liability. Especially in social/behavioral research the primary risk to subjects is often an invasion of privacy or a breach of confidentiality.

The next sections present a few examples for data sharing policies and regulations related to health information in Europe and in the United States.

10.13.1 Policy on data submission, access, and use of data within TESSy

The European Centre for Disease Prevention and Control (ECDC) created the European Surveillance System (TESSy) to collect, analyse and disseminate surveillance data on notifiable infectious diseases in Europe. A procedure with a set of rules was developed for data submission, data storage and custody, data use and data access, and data protection. Relevant forms and notes are also available (17):

- Request for TESSy Data for Research Purposes
- Declaration Regarding Confidentiality and Data Use
- ECDC Data Disclaimer
- Conditions for Publishing Note
- Sample Agreement for Agencies and third parties
- Declaration on Data Destruction

10.13.2 European Commission’s proposal for a General Data Protection Regulation

The European Patients’ Forum, which is a not-for-profit, independent organisation and umbrella representative body for patient organisations throughout Europe, wrote a position statement on general data protection regulation, and made recommendations to the European Commission, the European Parliament and Member States to:

1) Ensure that the Regulation protects patients’ rights as data subjects and as owners of their health and genetic data, and contains measures to enable patients to benefit from these rights effectively (e.g. access to data, data portability, right to information and transparency). Any restriction due to the special nature of the data processed or legitimate reasons for processing of such data should be justified and limited to what is necessary for public health, or the patients’ vital interest.

2) Make the necessary adaptations to the Regulation in order not to hamper provision of care, the conduct of research and public health activities, including patient registries and activities carried out by patient organisations to advance research and patients’ rights, with clear and explicit provisions to ensure the good implementation of this Regulation in the health sector.

3) Put in place effective cooperation measures between Member States and minimum security requirements to ensure an equivalent level of protection of personal data shared by patients for healthcare and research purposes across the European Union, and facilitate cross-border healthcare and research.

4) Involve patient organisations in decision-making and activities at policy and programme level for questions that relate to the processing and sharing of patients’ personal data, transparency towards patients and informed consent, to ensure the processing is carried out ethically and in a transparent manner throughout the European Union (18).

10.13.3 European Data Protection Board, General Data Protection Regulation

The European Commission plans to unify data protection within the European Union (EU) with a single law, the General Data Protection Regulation (GDPR). The current EU Data Protection Directive 95/46/EC does not consider important aspects like globalisation and technological developments such as social networks and cloud computing sufficiently. New guidelines for data protection and privacy are required to address these issues. Therefore a proposal for a regulation was released in 2012. Subsequently numerous amendments have been proposed in the European Parliament and the Council of Ministers. The EU's European Council aims for adoption the GDPR in late 2014 and the regulation is planned to take effect after a transitional period of two years.

10.13.4 HIPAA Privacy and Security Rules for Public Health Data Exchange

In the United States the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy, Security and Breach Notification Rules were developed.

"The Office for Civil Rights enforces the HIPAA Privacy Rule, which protects the privacy of individually identifiable health information; the HIPAA Security Rule, which sets national standards for the security of electronic protected health information; the HIPAA Breach Notification Rule, which requires covered entities and business associates to provide notification following a breach of unsecured protected health information; and the confidentiality provisions of the Patient Safety Rule, which protect identifiable information being used to analyze patient safety events and improve patient safety."(19, 20, 21)

References

1. http://web.mit.edu/6.933/www/Fall2001/Shannon2.pdf
2. https://grdr.ncats.nih.gov/ last visited 20/07/2014
3. http://confidential.oxfordradcliffe.net/anondata last visited 26/05/2014
4. http://www.bristol.ac.uk/Depts/DeafStudiesTeaching/ethics/resource/anonymise.pdf last visited 26/05/2014
5. http://www.jiscdigitalmedia.ac.uk/clinical-recordings/storage_anonymisation.html last visited 26/05/2014
6. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML last visited 20/07/2014
7. http://www.ecdc.europa.eu/en/activities/surveillance/tessy/Pages/TESSy.aspx last visited 20/07/2014
8. http://www.esri.com/library/whitepapers/pdfs/hl7-spatial-interoperability.pdf last visited 20/7/2014]
9. http://www.phdsc.org/health_info/ihe-task-force.asp last visited 20/07/2014
10. A framework ontology for computer-based patient record systems http://ceur-ws.org/Vol-833/paper28.pdf last visited 20/07/2014
11. http://www.cancer.gov/dictionary?cdrid=561717 Last visited: 20/07/2014
12. https://itunews.itu.int/en/2472-E8209health-standards-and-interoperability.note.aspx last visited 20/07/2014
13. WhatIs.com Tech definitions http://whatis.techtarget.com/definition/compatibility]

14. Development of a framework for Mapping European Seabed Habitats (MESH) http://www.searchmesh.net/default.aspx?page=1826
15. Surjan G. Ontological definition of population for public health databases. Stud Health Technol Inform. 2005; 116:941-5.
16. http://media.mofo.com/files/Uploads/Images/130729-BNA-Cross-Border-Information-Sharing-for-Effective-Services.pdf last visited 20/07/2014
17. http://www.ecdc.europa.eu/en/activities/surveillance/tessy/documents/tessy-policy-data-submission-access-and-use-of-data-within-tessy-2011%20revision.pdf last visited 20/07/2014
18. http://www.eu-patient.eu/Documents/Policy/Data-protection/Data-protection_Position-statement_10-12-2012.pdf last visited 20/07/2014
19. http://www.cdc.gov/phin/resources/standards/data_interchange.html last visited 20/07/2014
20. http://www.hhs.gov/ocr/privacy/index.html last visited 20/07/2014
21. http://www.cdc.gov/phin/resources/standards/data_interchange.html
22. Int J Med Inform. 2013 Jan; 82(1):1-9. doi: 10.1016/j.ijmedinf.2012.11.003. Epub 2012 Nov 20. Trustworthy reuse of health data: a transnational perspective. Geissbuhler A1, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, Leese A, Richardson C, Mantas J, Murray P, De
23. http://www.w3.org/TR/REC-xml/] which is a free open standard
24. http://www.w3.org/TR/REC-rdf-syntax/, http://www.w3.org/TR/rdf-schema/
25. www.w3.org/TR/owl-features/, www.w3.org/TR/owl2-overview/]
26. http://www.w3.org/2004/02/skos/
27. http://wiki.hl7.org/index.php?title=CTS2, http://www.omg.org/spec/CTS2/
28. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=10