Chapter 8
Data Quality in Rare Diseases Registries

Yllka Kodra, Manuel Posada de la Paz, Alessio Coi, Michele Santoro, Fabrizio Bianchi, Faisal Ahmed, Yaffa R. Rubinstein, Jérôme Weinbach, and Domenica Taruscio

Abstract In the field of rare diseases, registries are considered power tool to develop clinical research, to facilitate the planning of appropriate clinical trials, to improve patient care and healthcare planning. Therefore high quality data of rare diseases registries is considered to be one of the most important element in the establishment and maintenance of a registry. Data quality can be defined as the totality of features and characteristics of data set that bear on its ability to satisfy the needs that result from the intended use of the data. In the context of registries, the ‘product’ is data, and quality refers to data quality, meaning that the data coming into the registry have been validated, and ready for use for analysis and research. Determining the quality of data is possible through data assessment against a

Y. Kodra (✉) • D. Taruscio
Centro Nazionale Malattie Rare, Istituto Superiore di Sanità,
Viale Regina Elena 299, CAP, 00161 Rome, Italy
e-mail: yllka.kodra@iss.it; domenica.taruscio@iss.it

M. Posada de la Paz
RDR and CIBERER, Instituto de Salud Carlos III, Madrid, Spain
email: mposada@isciii.es

A. Coi • M. Santoro • F. Bianchi
Institute of Clinical Physiology, National Research Council, Pisa, Italy
email: acoi@ifc.cnr.it; michele.santoro@ifc.cnr.it; fabriepi@ifc.cnr.it

F. Ahmed
FRCPCH Samson Gemmell Chair of Child Health, Consultant in Paediatric Endocrinology
Royal Hospital for Children, Glasgow School of Medicine, University of Glasgow, Glasgow, Scotland, UK
e-mail: Faisal.Ahmed@glasgow.ac.uk

Y.R. Rubinstein
National Information Center of Health Services Research & Health Care Technology,
National Library of Medicine/National Institute of Health, Bethesda, MD, USA
e-mail: yaffa.rubinstein@nih.gov

J. Weinbach
French National Institute of Health and Medical Research, RaDiCo – Inserm UMR S 933,
Paris, France
e-mail: jerome.weinbach@radico.fr

© Springer International Publishing AG 2017
M. Posada de la Paz et al. (eds.), Rare Diseases Epidemiology: Update and Overview,
Advances in Experimental Medicine and Biology 1031,
https://doi.org/10.1007/978-3-319-67144-4_8
number of dimensions: completeness, validity; coherence and comparability; accessibility; usefulness; timeliness; prevention of duplicate records. Many others factors may influence the quality of a registry: development of standardized Case Report Form and security/safety controls of informatics infrastructure. With the growing number of rare diseases registries being established, there is a need to develop a quality validation process to evaluate the quality of each registry. A clear description of the registry is the first step when assessing data quality or the registry evaluation system. Here we report a template as a guide for helping registry owners to describe their registry.

**Keywords** Rare diseases registries • Quality assurance plan • Data quality indicators • Public health registry • Clinical research registry • Validity

### 8.1 Introduction

Patient registries are considered key instruments to develop clinical research in the field of rare diseases, to improve patient care and healthcare planning. They are the only way to collect a critical mass of data to increase the understanding of natural history of rare conditions, and to support the feasibility of the clinical trial. Therefore high quality data of rare diseases registries is considered to be one of the most important element in the establishment and maintenance of a registry. Quality is a much more complex term than it appears. Many definitions and interpretations exist depending on the goal, use and intent of the registry [9, 21, 36]. In broader terms, the term “quality” is defined as the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs [15] (ISO 8402:1994 2004). Quality evaluation of registry is considered to be one of the most important element in the establishment and maintenance of a registry. It is desirable that every registry should have a “builtin” a Quality Assurance Plan that should be implemented at every stage of the registry, from inclusion of new cases to dissemination of the final data analysis reports. As Brooke states, “every year an enormous quantity of medical statistics is compiled and published, and very little is known about the quality of the data on which these statistics are based” [4]. Before embarking on the quality evaluation of a registry, it should be determined whether the entire registry system (its total quality), or only part of it, will be assessed. The last guidelines on patient registry developed by the Cross-Border Patient Registries Initiative, a Joint Action Project funded by the European Union, identified numerous “quality influencing factors” that categorised the total quality of the registry into four groups. These categories should not be viewed separately when assessing the overall quality of a registry. Together these categories capture all the aspects of registry quality that are important to data end-users [38]. These categories are: (1) Registry governance; (2) Data quality; (3) Information quality; (4) Ethical issues (including security and privacy). The aim of this article is to focus on and address only the data quality aspects of a registry.
8.2 Data Quality

Data quality can be defined as the totality of features and characteristics of data set that bear on its ability to satisfy the needs that results from the intended use of the data [1]. The term “quality” refers to the degree of excellence, as in, “a quality product”. In the context of registries, the ‘product’ is data, and quality refers to data quality, meaning that the data coming into the registry have been validated, and ready for use for analysis and research. Data characteristics must altogether satisfy the intended needs of the registry purpose. In fact, the success of a registry will ultimately be judged on its ability to meet the goals it was created for.

Determining the quality of data is possible through data assessment against a number of dimensions. Data quality dimensions can be defined as “a set of data quality attributes that represent a single aspect or construct of data quality” [37]. By identifying different aspects or constructs of data quality it is then possible to measure the quality of data against those aspects or constructs identified.

Some dimensions of quality have been well discussed and defined in other area of disease registries [3, 22].

The dimensions provided are applicable for different registry types with different purposes, however not all may be equally important. The importance of a particular quality dimensions depends on the set objectives of the registry.

According to the objectives they are interested in, Registries are classified in the following categories:

- Public health registry/surveillance registry (disease registry): focus on disease occurrence (estimate incidence prevalence, temporal trends geographical distribution in relation to person, place, time); source of cases could be various and multiple; data collected are “basic” and refer to demographics, outcomes such as mortality; non longitudinal data are collected and tempestive information is required; the principal uses of data are disease burden measure, disease descriptive epidemiology, disease aetiology and risk factors, public health surveillance; health planning generate hypothesis for epidemiological research; the advantage and disadvantage are that data are “basic” but representative and can provide population disease occurrence; the denominator is well defined and the population or geographical coverage is comprehensive (population based registry). Example of public health registry/surveillance rare diseases registry are the Italian National Rare Diseases Registry [33, 34], Spanish Rare Diseases Registry [39], French National Rare Diseases Data Bank [20].

- Clinical/genetic research registry (patient registry): focus on the study of natural history of disease, understand cause of disease, risk factor, prognosis or treatment effect; sources of cases are clinical units; data collected are “clinically rich” and refer to diagnosis, prognosis, clinical outcome measures; the principal uses of these types of data are for clinical research, patient care and disease management. The follow-up is essential and tempestive data information is not
fundamental; the advantage and disadvantage are that data are “clinically rich” but not representative of the residing population, thus cannot provide epidemiological estimates of disease at population level; the denominator is not well defined and the population or geographical coverage may not be comprehensive (non population based registry). Furthermore, depending on the initial research question posed, there will be clear inclusion/non-inclusion and exclusion criteria defined before starting collecting data, which will exclude cases. Example of such clinical registry are TREAT-NMD DMD Global Database [2] and RaDiCo cohort databases (RaDiCo is the French Programme on Rare Disease Cohorts coordinated by Inserm is funded by the French National Research Agency under the specific programme “Investments for the Future”, Cohort grant agreement ANR-10-COHO-0003): www.radico.fr).

- Treatment registry focus on safety of monitoring for post-marketed drugs or devices products; services health technology assessment; mainly collect clinical and anthropometric data, information about medication, devices and health services, and Patient-Reported Outcomes [26].
- Combination registries

While each of the dimensions may be considered equally important, there may be instances where the relative importance of one dimension is greater than another.

For Public health registry/surveillance 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. On the other hand, in registries used for infectious disease, timeliness may be extremely important. For clinical registry, to satisfy the accuracy dimension, it may be necessary to sacrifice some elements of completeness or timeliness. In fact for clinical registries, exhaustive enrolment of all existing cases in the study and geographical coverage is not required, because only reaching an acceptable statistical power matters to perform the subsequent analyses.

Regardless of the type of registry, the high quality of the data is usually associated with a good oversight and governance mechanism, a secure and modern/adaptable information system, and with durable funding and would benefit from support in organizational aspects and management, innovation activities in information technology, epidemiology and statistics [6].

### 8.3 Dimensions of data quality and definitions

The data dimensions outlined in this article are: completeness of case ascertainment; completeness of the items; prevention of duplicate records; validity; comparability; accessibility; usefulness; timeliness.
8.4 Completeness of Case Ascertainment

Completeness of case ascertainment, known as external completeness, is the extent to which all patients occurring in the population are included in the registry database and applies to surveillance registries. A high degree of completeness of case ascertainment will ensure that the calculated incidence and prevalence rates are close to their “true value”. There are two kinds of methods to assess the case completeness: qualitative and quantitative [7]. The qualitative approach is to observe the trends in incidence/prevalence rates that can be a manifestation of changes in completeness of case registration. Implausible trends in incidence/prevalence may reflect incompleteness in recording events. Furthermore, failure to register deaths (and cause of death) will lead to overestimation of prevalence and of patient survival.

The quantitative methods may allow numerical evaluation of the completeness. Linkage with independent sources such as hospitals or national death certificates databases may be useful to estimate the number of cases missed by the registry [11, 16, 23]. These are less sensitive but inexpensive methods too. An independent survey with case ascertainment, however, gives a more accurate information on registry’ completeness [12, 27]. Besides, though it is expensive, it makes possible a subsequent examination of case selection bias, a point that needs to be examined particularly when registry are incomplete. Otherwise, one will never know if those registered cases are characteristically different from the missed ones.

Two statistical methods have been suggested by David H Stone to quantify completeness of a registry: pooling method and screening method [32]. In the latter case (Table 8.1), we just use the alternative information source as a gold standard with which we compare the registry.

Therefore, cases identified by both will be true positives, and we will have false negatives (sensitivity) and false positives (positive predictive value) depending on which one of the two sources has the cases. With the pooling method (Table 8.2), all cases identified by the registry and the alternative information source, excluding the repeat ones, are put together and the proportion of those identified by the registry is calculated as an estimate for completeness of the registry. Thus, by establishing a cut-off, it could be possible to see if a registry is reasonably complete.

<table>
<thead>
<tr>
<th>Table 8.1 Screening method</th>
</tr>
</thead>
<tbody>
<tr>
<td>External source (gold standard)</td>
</tr>
<tr>
<td>Registry</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Non cases</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Sensitivity = a/a + c
Positive predictive value = a/a + b
These methods are quite good particularly for assessing completeness of a population-based registry where the alternative information source might give us the possibility of identifying almost all cases diagnosed. However, in the case of population-based registry, it is too difficult to know all the individuals with that particular disease, and it is not easy to estimate how many of these are missed by both the registry and the alternative information source. Thus, the two methods mentioned above would not tell us how truly complete our registry is. This is an important point to be considered when we are interested in estimating in precise manner disease frequency in a target population.

Besides the methods mentioned above, there is a third and more accurate method to assess the completeness of a registry – a capture recapture method [5, 28, 29]. It is a relatively complex technique which requires a special software and the necessary know-how. It gives the opportunity to estimate the actual morbidity rate in the target population regardless of how complete the registry is. In brief, it is a method that helps to estimate those cases that are identified neither by the registry nor by the alternative source. By doing so, it completes the fourth cell of our 2 × 2 table and gives an estimate of the total number of cases in the target population.

Assessing completeness of a registry, is a relatively complicated process and becomes more difficult in the case of population-based registries. One can try hard to maximize its coverage but there is no way to assure inclusion of all cases in the registry. Complete case ascertainment mainly depends on peoples’ demand of medical care, accessibility of health care, health service utilization rate and health workers’ capacity to identify the illness (cases).

### 8.5 Completeness of the Items

Completeness of the items known as internal completeness is the proportion of registered cases with missing values (or unknown) for different variables.

When registries collect large amounts of variables, it is important, in the perspective of data quality, to take into account the specific purpose of the analysis and to distinguish for this specific analysis items deemed to be ‘essential’ from those deemed to be ‘non-essential’. It may be reasonable to focus the objective of full completeness on the essential items only [7]. The missing value must be very low for variables which are critical to a specific analysis; for population based cancer registry the gold standard for missing value for critical variables <= 2% [13].

<table>
<thead>
<tr>
<th>Table 8.2</th>
<th>Pooling method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative information source = A</td>
<td></td>
</tr>
<tr>
<td>Registry = B</td>
<td></td>
</tr>
<tr>
<td>Pooled data = C</td>
<td>( C = (A \cup B) - (A \cap B) )</td>
</tr>
</tbody>
</table>
A registry of good quality should have a high percentage of item completeness throughout the course of its existence. If the collection of data is based on electronic data capture and data management tools (e.g., eCRF and e-query systems), simple automatic rules, professional and continuous data management support for making sure that critical items are completed and may help to support completeness of the registry.

8.6 Validity

Validity refers to the proportion of cases in a dataset with a given characteristic, which truly have the attribute. Lack of validity is referred to a bias or systematic error [24]. Validity depends on the precision of source documentations on the level of expertise in data classification and coding; on the registry “protocol” (explicit definitions, good coding systems, documented rules limitation of free text fields, preference for pre-defined list of possible information items; data coherence rules, continuous data management for validating entered data before final integration in the registry). Validity has both an internal and an external dimension [8].

- **Internal validity** relates to the extent of errors within the system. It depends on the following errors: – misdiagnosis: health outcomes with unspecified symptoms in the absence of laboratory confirmation; – miscoding: health outcomes which were not reported because the coding system doesn’t include a specific and appropriate code; – misclassification: health outcomes reported with inappropriate case definition category; no clear case inclusion/exclusion criteria including diagnosis criteria. Moreover limitation of free text fields, preference for pre-defined list of possible information items; data coherence rules, continuous data management for validating entered data before final integration in the registry improve the internal validity of item.

- **External validity** is the ability to generalize study results to a more universal population. It is the degree to which the conclusions in a study would hold for other persons in other places and at other times. One indication that a study lacks external validity is if the sample is not representative. Evaluating validity implies a registry indicator measured against a ‘gold standard’ value. An agreement between the registry under examination and an alternative information source in all the items of a single case is recommended. Cases can be selected at random from the registry and entries of each item can be compared with the alternative information source. This can be patients’ clinical record, laboratory records, etc. We may quantify their agreement in terms of percentage, and depending on the purpose of our interest we can establish cut-offs. Some authors use also the kappa coefficient, a statistical measure commonly used in testing the reliability of diagnostic tests, to see the agreement between the two information sources on specific variables [14, 35].
8.7 Coherence and Comparability

Coherence reflects the degree to which data can be successfully brought together with other statistical information within a broad analytic framework and over time. Coherence covers the internal consistency of data collection as well as its comparability both over time and with other data sources [38]. Comparability is the extent to which the data collected can be analyzed to make a comparison with other registries or over time. This is very important in the analysis of geographical and temporal distribution. Standardization of definitions, use of standard clinical vocabularies, terminologies, classifications and ontology, is the only sure way to achieve the international comparability [31].

8.8 Timeliness

Timeliness refers to the rapidity at which a registry can collect, process and report sufficiently reliable and complete data, for producing results or outcome for action (report and/or research article and/or public health action) [3]. This timeliness is determined by the time between the various steps in the registry information chain and depends, also on the aims of the registry. If a registry has a role in quality improvement of health care or immediate public health action, the time period needed to produce results for feedback to clinical centres is a crucial point.

Couchoud et al. [7] propose four indicators to evaluate the timeliness. (1) Time until receipt: time from the clinical event to the record in the registry. (2) Process time: the time from the presence of the record to its availability for research (available in the ‘frozen’ database after quality control procedures). (3) Time to availability: sum of the two previous times. (4) Number of patients or data recorded in the registry after the database was ‘frozen’ to produce an annual report or a scientific paper. These cases or items are found the year after, in a new ‘frozen’ database [7]. An other indicator of timeliness is also timelines of patient visits and adherence to them in a given longitudinal study. If you consider a surveillance registry, you need only one capture of the “case”; if it is a clinical research registry, you need to make sure all planned visits (ex: 2 visits per year during 5 years are respected for all included cases.

Other prerequisites of data quality are accessibility, usefulness and prevention of duplicate records.

8.9 Accessibility

Accessibility is defined as availability of aggregate data, publication of periodic reports and/or literature in peer-reviewed journals, and clear framework and procedures (including at the technical level) for accessibility to external researchers of anonymized patient-level data. Registry data accessibility presents an opportunity
for sharing and more productive collaborations to collect relevant data, implement quality and standardization procedures, and provide broad access to comprehensive aggregate information and anonymized patient-level data to facilitate the advancement of research and improvement of patient care [19].

### 8.10 Usefulness

Usefulness refers to the extent to which an information system or its output provides any benefit or value. The usefulness of a registry can be perceived differently by different stakeholders. Government institutions are likely to value systems from public health point of view: for example, evaluate the population health status; planning health services; provide data on declining disease incidence. On the other hand, the scientific community will find it useful when disease registry data offer new insights in the discovery of disease knowledge and its natural history, or reveal new phenomena, which will help to generate new hypotheses. For clinical registry, the level of usefulness is intended for example how to use data registry in subsequent clinical trial and study design; participation in awarded grants; several publications through peer-reviewed publications.

One more feature closely linked to usefulness is the registry’s overall adaptability or its capacity to include new data items (eg to address specific research subprojects in partnerships with potential data end-users such as pharma companies).

### 8.11 Prevention of Duplicate Records

Duplicate records refer to the multiple registration of the same patients into the registry database. This might due to patient mobility, which often refer to more than one doctor and more than one hospital; jumping from paediatric care to adult care management or related to registration errors (spelling mistake in family name (or very long family names not entered the same way by two clinicians in the same of different hospitals). Specific methods to detect those duplicate should be in place, otherwise, incidence and prevalence rate may be overestimated. Identifying duplicate case records can be difficult, and a common set of criteria needs to be employed to prevent their generation. They can be detected with a series of deterministic/probabilistic matches using the personal identification number, or by a match in other identification variables such as name and surname if allowed, birth month and year, sex and etc. Records matching exactly in all of these fields are automatically assigned to the same patient. In some cases the diagnosis needs to be managed because a patient could have two different RD and the rest of the variables will match but this is not a real duplicate record. This only happens in wider registries where several diseases are registered. It is important that the registry needs to have in place procedures for handling duplicate registrations in order to avoid having duplicate patients entered into the registry and to calculate regulatory the percentage of duplicate records found in the whole database.
8.12 Factors Influencing the Data Quality Dimensions

Considering that data quality is part of a complex system, as many others factors may influence the quality of a registry: development of Standardized Case Report Form (CRF) and informatics infrastructure.

8.13 Development of Standardized Case Report Form

Case report form (CRF) (paper or electronic based) is the initial step in translating the protocol into standard questionnaires. The CRF must comprise all variables that are necessary to answer the research questions planned in the design phase and it has to use standard definitions of items and variables. Standard development of CRF using standard guidelines helps the collection of consistent and valid data [10]. Problematic CRF include: unclear questions (e.g. when acronyms, complex words or abbreviations are used; poor ergonomics and no use of branching logics and conditional fields systems resulting in too long reporting form); poor ergonomics and no use of branching logics and conditional fields systems (resulting in too long reporting form); no logical order of questions (e.g. clinical and laboratory sections not clearly separated or mixed-up); meaning of question is unclear [30]. In addition, scientific expert are encouraging the use of Patient-Centered Outcome Measures form (PCOM), as a relatively new concept, to be integrated with CRF. The use of PCOM form, which are potentially of relevance for rare diseases, are the instruments that can be used to measure real benefits for patients and from their perspective. The International Rare Diseases Research Consortium (IRDiRC) strongly recommend that the insertion of PROs into the design of rare diseases registries is necessary to fully evaluate their natural history [25].

A “library” of standard reporting form are elaborated by the National Institute of Neurological Disorders and Stroke (NINDS) (https://cde.nlm.nih.gov/home) with the aim of standardizing the collection of investigational data in order to facilitate comparison of results across studies and more effectively aggregate information into significant metadata results.

8.14 Informatics Infrastructure

The successful implementation and use of a registry depends on a thoroughly and accurate planning and construction of a suitable IT infrastructure [8].

The IT infrastructure for user authentication, data entry, data management, storage and subsequent analysis should be:

- Web-based for data entry through Secured-cloud-based for data storage and backup (information on a case or series of cases is entered into a data entry mask on a secured web page). Advantages are that this technology is common, cheap.
• Interface-mediated data retrieval before integration in the registry if data are initially collected or recorded in an external system. It may be possible to wholly automate data import/export by developing an interface (data warehouse) and machine readable forms (Extract, Transform, Load (ETL approach).

• Interface-mediated data management system, allowing for instance to implement pre-defined automated control rules in the forms as well as query messaging system between the data manager and the clinical unit participants for continuous (if not real time) control and validation of entered data.

• Open-source software. The great advantage of an open-source software that enables scientists to build a registry for a specific rare disease even without special IT knowledge. The downside is that the software is not supported in an enforceable way, i.e. by a legally binding contract.

• Secure-certified following regular security audits, to prevent from malicious/unauthorised interventions

• Adaptable to technological evolutions (resistant to obsolescence) and to the rise of “big data” needs.

Determining which information system architecture to use and how to design the system is an essential question when setting up a registry system [8].

The choice of server hardware and database solution can have a marked effect on data quality. Server hardware varies in levels of stability, maturity and speed and the choice of database software can affect data quality. To mitigate risks caused by the choice in soft- and hardware, the validity of data needs to be thoroughly monitored.

Based on a systematic review of the literature, Doris Lindoerfer et al. [17, 18], developed a checklist for patient registry software systems (CIPROS) which supports developers to assess requirements of an existing system. It also supports the reporting of patient registry software system descriptions in papers and it can be a first step to create standards for patient registry software systems.

8.15 Conclusion

With the growing number of registries being established, there is a need to develop a quality validation process to evaluate the quality of the each registry.

As stated earlier, the quality of a registry refers always to the objective for which it was meant.

It will be important to provide tools for registry managers and to elaborate on the quality indicators so they can conduct self-evaluation. This helps them to continue what they are doing if they are on the right track or to rethink and restructure their registry activity if they are having some problems. Accordingly, a questionnaire is developed as an initial tool for assessment of a registry. A clear description of the registry is the first step when monitoring data quality or the registry evaluation system. Here we report a template as a guide for helping registry owners to describe
their registry. It is necessary to update regularly this template description. Ideally all the questions of the questionnaire should be answered positively, before going ahead with the analysis of data quality (Table 8.3).

Table 8.3  Example template for registry description

<table>
<thead>
<tr>
<th>Indicate the date when you are filling out the template</th>
<th>date: dd/mm/yyyy</th>
</tr>
</thead>
</table>

1. **Contact information**

- Name of the registry (and acronym)
- Name of registry database owner (responsible legal entity for data management)
- Name of registry contact person
- Registry address
- Registry telephone number
- Registry fax number
- Registry email address
- Registry web home page

2. **Registry organisation**

- Year of establishment
- Registry language(s)
- Membership of other international networks (yes/no, if yes specify the name of the network)
- Indicate the registry funding source
- Describe staff working in the registry which may include: PI (e.g. management, financial sustainability), Registrar (e.g. collection, registration, data management and monitoring); Informatic personel (e.g. maintenance in operational condition, backoffice/helpdesk and bug resolutions, automation and output); Statistician/epidemiologist (methods, analysis, interpretation) Medical (e.g. pathology, coding), Administration (e.g. secretarial support) and etc.

3. **Type of registry**

According to the objectives they are interested in, Registries are classified in the following categories:

1. Public health registry focus on disease occurrence (estimate incidence prevalence, temporal trends geographical distribution in relation to person, place, time); the principal uses of data are disease burden measure, disease descriptive epidemiology, disease aetiology and risk factors, health planning;
2. clinical/genetic research registries focus on the study of natural history of disease (understand cause of disease, risk factor, prognosis or treatment effect); the principal uses of data are for clinical research, patient care and disease management,
3. treatment registry focus on safety of monitoring for post-marketed drugs or devices products; services health technology assessment
4. combination registries.

4. **Objectives**

The objectives of the system indicates why the system exists

List the principal and secondary objectives of the registry

(continued)
5. List of diseases under registration

When preparing for the evaluation of registry system, all diseases covered by the system should be listed. The disease under registration could be specific (example: Prader-willy syndrome) or group of diseases (haemoglobinopathies, primary immunodeficiency). It is recommendable to use the list of diseases included to the Orphanet classification of rare diseases diseases (http://www.orpha.net/consor/cgi-bin/Disease_Classif.php?lng=EN).

6. Inclusion/exclusion criteria

The registry team should specify so-called eligibility or inclusion criteria that are a set of conditions that a 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. a 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). Exclusion criteria, on the opposite side, are those criteria that disqualify subjects from inclusion in the registry.

7. Data sources and data flow

A data source for a registry system can be defined as a place where the initial information on the disease to be reported is collected. Genetic laboratories and hospitals are the most common sources of information for registry. Other source (general practitioners electronic health record, administrative data, Patient Reported Outcomes PROs, connected devices generated data) may also be included in the registry system.

A clear flowchart for a generic case reporting system is necessary and the following elements should be considered in order to describe data flow: (1) Data providers or data sources as described in the previous section; (2) Processes for clinical diagnosis, case confirmation, and gathering of additional information; (3) Public health institutions (data recipients) that provide feedback information to participants of the case reporting process, public health professionals, and the general public. (4) data management entity.

8. Populations under surveillance of registry system

The population under surveillance can be defined as the general population or targeted groups. The targets can be based on specific age categories (e.g. children under five years of age) or other determinants.

9. Geographic coverage

The geographic coverage represents the geographic unit (municipality, region, country or any other pre-defined geographic area) where disease registry is conducted.

Based on geographical coverage registry could be classified in:

1. population-based registries, which refer to a geographically defined population and aim is to register all cases in that population. For public health registry this information is of critical importance.

2. non-population-based registries are based on selected bodies, clinical Centers or other types of structures where the population coverage may not be comprehensive. The majority of research clinical registry are non-population based as this information is less relevant. The geographical coverage of disease surveillance is linked to the concept of representativeness of the registry system.

10. Specification of the information to be reported

List all Variables included in the registry.

(continued)
Table 8.3 (continued)

11. Registry’s regulatory status
List all ethical and regulatory approvals obtained/country covered for the registry implementation.

12. Collaborative framework status
Is there a clear governance in place? yes/no if yes describe.
Please list any contract (eg consortium agreement) existing between all registry-participating institutions.

13. Informatics infrastructure (software and hardware)
Determining which system architecture to use and how to design the system is an essential question when setting up a registry. Give details of your computing system and software for data entry, data storage and data analysis, type of architecture; server selection.

14. Data management procedure/quality control
Is there a data management procedure in place? Yes/No if yes describe staff and procedures in place to manage registry data and its quality (Data management plan, (DMP), Data validation Plan (DVP) automatic rules in eCRF, continuous data management versus periodic controls, electronic data correction, query system, site visits and monitoring, etc.

15. Security standards and procedures
Describe any security measures in place (frequency of security audits, certifications obtained) and procedures (Active directory for user rights management, login authorisation procedures, history logs, back-ups, etc.).

References

36. United States Bureau of the Census (1998) Survey design and statistical methodology metadata, software and standards management branch, systems support division. Washington DC; Section 3.3.6, p 8