Data quality

Yllka Kodra MD PhD
National Centre for Rare Diseases Rome, Italy
Topics

- Data quality dimensions
  - Completeness
  - Accuracy
  - Timeliness

- Factors influencing the data quality
  - Case Report Form
  - Data elements
  - Standards
• are we going to evaluate the quality of a registry in its totality or part of it?
Registry is complex system
DEMING SYSTEM DEFINITION

A system is a set of functions or activities within an organization that work together for the aim of the organization.

- **Input**
  - staff
  - IT
  - Sustainability

- **Process**
  - Data entry
  - Data management
  - Data analysis

- **Output**
What is the quality in the context of the registry?

• The term "quality" is degree of excellence, as in, "a quality product".

• In the context of registries, the ‘product’ is data, and quality refers to data quality.
DATA QUALITY DIMENSIONS
DIMENSIONS OF DATA QUALITY

Dimensions of data quality

- Accessibility
- Comparability
- Coherence
- Accuracy
- Consistency
- Usability
- Reliability
- Relevance
- Interpretability
- Timeliness
- Completeness
Completeness

• Completeness of case ascertainment (external completeness)

• Completeness of the items (internal completeness)
Completeness of case ascertainment

- **Definition**: is the extent to which all patients occurring in the population are included in the registry database

- The completeness of case ascertainment has implications for the conclusions and the extrapolations made to the general population.

- Assessing completeness of a register, on the other hand, is relatively a bit complicated process and becomes more difficult in the case of population based registers. One can try hard to maximise its coverage but there is no way to assure inclusion of all cases in a register.
Methods

1. Comparative approach to other similar study (crude and cheap)

2. Capture-recapture

3. Screening method (the alternative information source as a gold standard with which we compare the register)
### Screening method

<table>
<thead>
<tr>
<th></th>
<th>External source (gold standard)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
</tr>
<tr>
<td>Register Cases</td>
<td>a</td>
</tr>
<tr>
<td>Non cases</td>
<td>c</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
</tr>
</tbody>
</table>

Sensitivity = \( \frac{a}{a + c} \)

Positive predictive value = \( \frac{a}{a + b} \)
Capture Recapture

Capture-recapture methods have been advocated for use in estimating completeness of a register

- The first is that when there are two sources, they are assumed to be independent,
- The second is that all individuals have the same probability of being captured.
Capture Recapture

C - R estimate X those cases that are identified neither by the register nor by the alternative source.

Total number of cases in the community

\[ N = a + b + c + x \]
Completeness of the items

- Definition is the proportion of registered cases with missing values (or unknown) for different variables.

- % of missing values for each variable.

- It may be reasonable to focus on the objectives of full completeness on the essential items only because to get a high rate response in all variables involves high costs.
Accuracy of items

• Is the extent to which the data items measures what is intended to be measured

• In practice, it is the percentage of agreement between registry data and an independent source objectively measuring the same variable.
Examples of indicators are:

• proportion of cases complying with case definition;

• proportion of coding errors within a dataset;

• proportion of values that comply with a gold standard or reference value
Different steps are included in the coding process

I step Medical diagnosis
Medical diagnosis is the act of labeling something as disease through clinical, laboratory and pathological findings combined with clinical knowledge and judgment. The medical diagnosis generates a \textit{disease’s definition}.

II step Disease’s nomenclature
The medical diagnosis is conducted with a name or disease \textit{terminology}. The name of a disease is a convenient brief statement of the current conclusion of a diagnostic process;

III step Disease’s \textit{classification and coding}. The classification aims to put the disease to a unique category, which includes a set of similar disease, and to assign a code.
Orphanet nomenclature and classification of Rare Diseases

Classifications
Clinical signs
Mode of inheritance
Epidemiological Data:
Prevalence Age of onset Age of death
Clinical signs
(Orphanet Controlled Vocabulary)
Genes
PubMed query
MeSH descriptors
OMIM number
ICD-10
Rare Diseases nomenclature
Timeliness

**Definition**: refers to the rapidity at which a registry can entry collect, manage analyses and report sufficiently reliable and complete data, for producing results or outcome for action (report and/or research article and/or public health action).
Higher level of completeness or accuracy or timeliness should be evaluated in relation to the different registry types with different purposes.

- For population-based registry: (include all existing patients cases) completeness is crucial.
- In registries used for infectious disease surveillance, timeliness may be extremely important.
Quality system

Data elements
Use of standards
Case Report Form

• Definition: A printed, optical or electronic document designed to record all of the protocol – required information to be reported to the sponsor on each trial subject.

• is the interview questionnaires used to collect registry data and include questions

• is the initial step in translating the protocol into standard questionnaires
Paper CRF vs eCRF
CRF module

Patient data

- Demographic data
  - Family history
  - Medical history
  - Behaviour history
  - Social status
  - Identifying information

Disease data

- PROs
  - Vital status
  - Treatment
  - Disease history
  - Follow-up
  - Examinations
  - Genetic data
  - Phenotype data
Standards for clinical research

Clinical Data Standards Interchange Consortium (CDISC)

Regulated Clinical Research (RCRIM) Technical Committee of Health Level Seven (HL7)
The NIH Common Data Elements (CDE) Repository has been designed to provide access to structured human and machine-readable definitions of data elements that have been recommended or required by NIH Institutes and Centers and other organizations for use in research and for other purposes. Visit the NIH CDE Resource Portal for contextual information about the repository.

The Repository is a platform for identifying related data elements in use across diverse areas, for harmonizing data elements, and for linking CDEs to other existing standards and terminologies, including the value sets in the Value Set Authority Center (VSAC).

**Search**
Search for individual data elements, by definition, users or sources. Search for sets of data elements ("boards") identified by a particular group for a particular use (e.g. particular research solicitation).

**Compare / Harmonize**
Analyze and resolve differences between data elements. Assure that your forms are using variables that will be usable by certified EHRs.

**Create**
Draw upon the experience of colleagues and others to design unique data elements and measures.
Data elements

- A DE refers to information that describes a piece of data to be collected in a study.

- A data element has a name, precise definition, and clear enumerated values (codes) if applicable.

http://www.nlm.nih.gov/cde/glossary.html#examples
## DATA element structure

<table>
<thead>
<tr>
<th>ID of data element</th>
<th>Name of data element</th>
<th>Description/definition</th>
<th>Data sources</th>
<th>Variable structure/Size</th>
<th>Response Value</th>
<th>data provider (patient/researchers/clinicians)</th>
<th>Source of standardization</th>
<th>Code standardisation</th>
<th>CRF question corresponding to data element</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>eg: disease name</td>
<td>Duchenne National Registry</td>
<td>date/free text/strings/numbers/integer/alphanumeric</td>
<td>1 – Yes/2 – No/3 – Don’t know</td>
<td>ORPHANET</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CDE Browser (1)
Standards

• Definition: are clinical vocabularies, terminologies, classifications and ontology that can be used for improving the comparability and interoperability with other database.

• Is the set of data elements collected using standards?
### Example of international standards

<table>
<thead>
<tr>
<th>Area</th>
<th>Standard</th>
<th>Developer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>ICD 10, ICD 9CM</td>
<td>WHO</td>
</tr>
<tr>
<td>Medical Nomenclature</td>
<td>SNOMED</td>
<td>International Health Terminology Satdards Development Organisations</td>
</tr>
<tr>
<td>Device</td>
<td>GMDN</td>
<td>GMDN Maintainance Agency</td>
</tr>
<tr>
<td>Drugs</td>
<td>ATC DDD index</td>
<td>WHO</td>
</tr>
<tr>
<td>Adverse reactions</td>
<td>WHO Art</td>
<td>WHO</td>
</tr>
<tr>
<td>Disability</td>
<td>ICF</td>
<td>WHO</td>
</tr>
<tr>
<td>Primary care</td>
<td>ICPC-2</td>
<td>WHO</td>
</tr>
<tr>
<td>Genes, genetic disorders</td>
<td>OMIM</td>
<td></td>
</tr>
<tr>
<td>Medical Laboratory</td>
<td>LOINC</td>
<td>Regenstrief Insitute</td>
</tr>
<tr>
<td>Observations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Recomandation

• Build quality into the system-do not add it later.
• Somebody must be responsible for quality control at each point in the system.
• Use standard guidelines for designing CRF
• There should be explicit standards and procedures for evaluating the system on a regular basis.

David J Solomon, 1991 Public Health
The word quality is now widely abused. It would be better to use it less or specify instead what you intend to do

Pier Luigi Morosini