Paper PP07

Experiences of Managing Quality Registry Data for Effective Exploration

Catharina Dahlbo, Capish, Malmö, Sweden

ABSTRACT

Data analytics, the insight we gain from exploring data effectively and efficiently, is key in the quest for optimal health care and the development of drugs. As a process, data analytics can be divided into three main activities – data collection, data curation and data analysis.

Swedish national quality registries are used to continually evaluate and improve the quality of Swedish health care as well as form the basis for scientific research. This poster will focus on the experiences made while handling quality registry data in a data curation process, with the aim to improve the users' possibilities to explore and analyze their data. The data curation activity not only includes the modelling of data for optimal data exploration, but also quality control. While the quality registry data is commonly used for cross-sectional analysis, this data curation process adds new opportunities to explore data longitudinally.

INTRODUCTION

In Swedish health care, there are about 100 national quality registries containing individual patient data about e.g. diagnosis, interventions and outcomes in specific areas such as diabetes, breast cancer and cerebral palsy. Since the first Swedish registry started 1975, registries are used to evaluate and improve health care continually. The quality registries have been called a gold mine from a research perspective (1). The personal code number that is unique for each person in Sweden enables the integration of national quality registries, other registries and biobanks, providing unique opportunities in research (2).

The amount of data collected is different from registry to registry. The most comprehensive registries contain data for many different aspects of the condition and employ multiple data collection forms to fit different categories of staff, e.g. surgeons, neurologists, nurses, physiotherapists. Data is typically entered manually, normally as an add-on step to the documentation required for the patient's journal. This, in combination with a vast number of users, makes it difficult to achieve optimal data quality, impacting what can be done in data exploration, visualization and analysis.

A collaborative project has been carried out with one of the most comprehensive registries in Sweden, where data collection started already in the nineties. The representatives for the registry were looking for ways to widen their research projects, by interacting more dynamically with their registry data and enabling them to explore the data in new ways, e.g. longitudinally as opposed to the traditional cross-sectional analysis.

QUALITY REGISTRIES IN HEALTHCARE VS CLINICAL TRIALS

Differences have been revealed during this project between working with data from clinical trials versus data from quality registries. As an example, the objective of data collection is different, in clinical trials there is a clear objective and hypothesis and data is collected in the specific context of the proposed hypothesis. The initial purpose of a quality registry is to collect data to document the current status quo and the resulting research hypotheses are broad and change over time. Thus, data entered in the registry is often not collected in the context of the question at hand and might not be in the optimal format to answer a specific research question.

COLLABORATION PROJECT

EXPECTATIONS OF EXPLORATION

The aim of the collaboration for Capish was to demonstrate the possibilities that arise with an effective curation process and a software tool that provides enhanced capabilities for data exploration.

In addition, representatives of the Quality Registry were looking for a tool that could give researchers direct access to the data and allow for a more dynamic work process. Functionalities they were looking for were:

- To find connections and correlations
- To explore dependencies between events and episodes
- To explore the development of one patient over time in relation to others
- To identify patient cohorts and compare them

No detailed list of requirements was defined at the beginning of this collaborative project, but during the project a specification has been developed in collaboration, based on the requests of what the researchers wanted to do with the data. The approach can be said to have been agile, where new ideas were generated at subsequent demonstrations on how the data had been integrated and access was facilitated.

IDENTIFIED ISSUES OF THE QUALITY REGISTRY

Some of the issues preventing researchers to use the original registry data efficiently were as follows:

- Difficult to achieve data quality
- The structure of the data did not support the proposed analysis
- Limited ways of exploring data longitudinally
- Data was only accessible to a limited number of users
- Multiple stakeholders with different interests

COMPONENTS OF AN EFFECTIVE EXPLORATION

The following has been defined as key components to achieve an effective data exploration:

- Data can be
  - accessible both cross-sectional and longitudinally
  - searched freely using a free-text search
  - related in real or relative time
  - compared between different datasets
  - accessed by many users
- Users can
  - be situated anywhere
  - explore their own theories
  - save and share sessions with each other
  - get a holistic view of each patient combined with aggregate views of the complete dataset
  - easily identify cohorts, create and compare them to each other
  - easily find and follow an individual patient from a cohort

REQUIREMENTS FOR AN EFFECTIVE EXPLORATION

The following requirements need to be fulfilled for an effective exploration:

- Comparable data
- Standardized data
- Common measurement units
- Uniform terminology
- Data expressed explicitly not coded
- Metadata available together with data
- Data from multiple sources combined in one unified view
- Effective modelling of data
- A tool that can take advantage of the data model

DATA COLLECTION

The people involved in the data collection/entering process as well as analyzing/exploring data may include representatives from different disciplines, with different interests and it may not necessarily be the same people throughout the process. When only few people have the holistic view of the data, it is very difficult to grasp the importance of the data entry step for an effective exploration. It is not a surprise that data is not always correctly entered, but if enough effort is spent at this stage there is much to gain in data quality.

DATA CURATION

Data collection, data curation, and data analysis are the three building blocks in the process of evaluating data. In this poster, we will focus on the impact of a good data curation process for an effective data exploration. Data curation is one of the most important processes that decides how well data exploration can be done. Sampled data undergoes a process once and there is no need to repeat it and this will prepare the data for analysis. If done well, the resulting structured data can be used for the analysis of the data in many different contexts, without the need for additional data manipulation/curation.

Data curation spans over a range of activities needed for adding value to data. Preparation, cleaning, translation, standardization, mapping, modelling are examples of activities needed to be done for an effective exploration. While not specifically aimed at data quality, these activities will automatically have a positive effect on data quality. Therefore, it is valuable to put effort into data quality during the data curation step.

DATA CURATION EXPERIENCES

As described above, reporting of data is done throughout time, at different locations and by different people. This has many implications for the collected data. Corrected data could end up as new records in the data set, with the consequence of having to handle duplicate data. Data is potentially also not entered in a standardized form, or might not be immediately comparable due to different ways of measuring the data at different healthcare centers. Other challenges include conflicting answers, comments in result fields, wrong dates or incomplete answers. The question then arises; how could we solve these challenges once for all for the specific data set?

When we received the data, it was provided in eight Excel files, which seemed to be easily manageable. But one file included more than 350 columns and was thus cumbersome to work with. The column names were coded and had to be resolved using a separate file providing the mapping between the coded and the explicit column names. The value in most of the cells were yes/no answers or the field was empty, not easy to analyze or search for.

The aim was therefore to restructure the data using Capish's ontology based approach to data modelling, providing a single structured data set for further analysis. In a first step, all codes were replaced with the column names. Subsequently the data was categorized into smaller blocks, holons, connecting each block to each other. Further data refinements had to be carried out to achieve this goal. Some of these steps had to be done manually, but others could be carried out automatically by programming procedures.

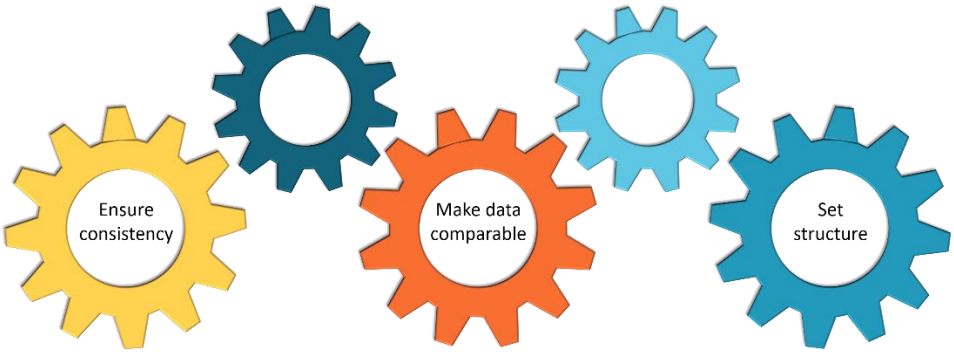

***Figure 1**. Data curation is a complex process that has great impact on data exploration.*

The first step included quality control of the data to ensure consistency.

- Duplicate entries – were adjusted to only one record.
- Conflicting answers – were handled by specific rules so only one answer was included.
- Comments in result fields – were moved into another field so the result field only contained relevant information.
- Values spelled in different ways – were mapped to one common spelling so it would be searchable.
- Date corrections – wrong dates entered were corrected and default dates were determined.

The second step included efforts to make the data comparable.

- Standardization – data was mapped according to e.g. ISO 8601 and ICD-10.
- Translation – data was translated into English which facilitates combining data from multiple countries.
- Terminology – data was mapped to a uniform terminology for better integration.
- Common units – measurement units were ensured to be common for the same measurement.
- Mapping – Yes/No was mapped to understandable terms within the medical concept (e.g. Have scoliosis = Yes was mapped to the searchable term 'Diagnosis Term' = Scoliosis) and coded values were mapped to explicit values (e.g. F/M to Female/Male).
- Reference comparison – values were compared to reference values, where the value could be mapped to normal or abnormal or mapped into different colors, dependent on the criteria.

The third step was to set a structure by effective data modelling.

- Classification – data was grouped into well-known concepts, called holons.
- Data was appended with metadata.
- Time was related to events for individual patients – in longitudinal data the same measuring points will be repeated during a patient's life, to allow for cross comparison at specific ages, a relative timeline was calculated for each patient and measuring point.
- Summaries were created in specific holons including:
    - Total durations – total episode durations were calculated in days, as well as minimum and maximum durations
    - Number of changes in levels of movement functionality
    - First and last date for episodes
    - Number of episodes
    - Worst case of reference evaluation
- Creation of episodes – repeated answers were combined as episodes rather than single events, for example making it possible to compare for how long a patient has felt pain.

Have you experienced pain?	*Yes*	*Yes*	*No*	*No*	*Yes*	*Yes*	*Yes*	*Yes*
Date	*2011-04-01*	*2011-10-01*	*2012-04-01*	*2012-10-01*	*2013-04-01*	*2013-10-01*	*2014-04-01*	*2014-10-01*
Original registrations	|	|	|	|	|	|	|	|
Pain location		*Knee*					*Knee*	
		|----------|			|----------|		
Reported by		*Custodian*			*Patient*		*Custodian*	
		|----------|			|----------|		|----------|

***Figure 2**. Example of a patient's registrations and how episodes can help summarizing the data. Each episode got a start date and a stop date.*

Although extensive data curation is potentially a time-consuming activity, it pays major dividend in the subsequent analysis of the data.

DATA ANALYSIS

The curated data with its metadata was explored in a tailor made intuitive application that takes advantage of the Capish way of modelling the data. This application allows users to move from detailed patient information to analysis of the entire population and vice versa. As an example, an outlier in a population graph can be chosen and further drilled-down to view everything that has happened to that patient. From the characteristics of that patient, cohorts of similar patients can be created to further explore and investigate a specific question.

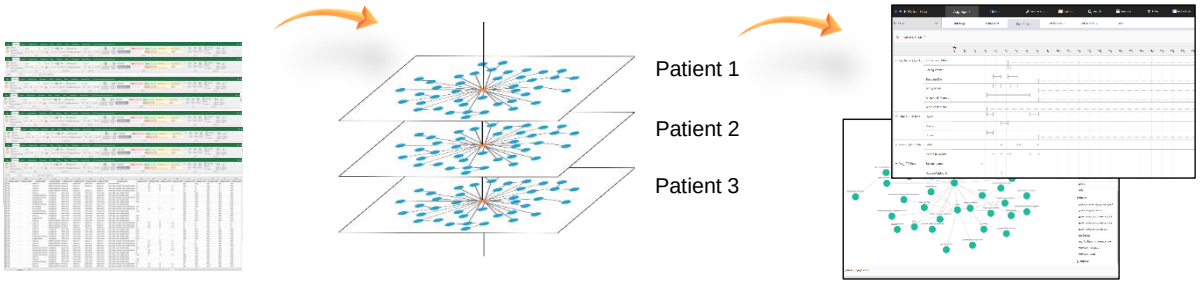

***Figure 3**. The data structure has a profound impact on how the data can be used in the best way. By modelling the data differently, the exploration possibilities increase.*

CONCLUSION

The source data was not collected, structured or cleaned in a way that facilitated exploration and optimal use of the data. Organizing the data using Capish's ontological approach to data curation and modelling enabled a more objective analysis of the data and also provided an additional data quality control step. The ability to view the data together with its metadata based on the underlying medical concept and not just as separate data points, optimizes the use of the data.

This collaboration was a great opportunity for Capish to provide proof of concept for its enhanced data curation process and its software tool aimed at streamlined, intuitive and efficient data exploration. For Capish, it is a confirmation that the chosen approach will fill a gap for the health care industry.

So,

Take Control of Data before Exploration!

REFERENCES

1. Rosén M. Översyn av de nationella kvalitetsregistren. Guldgruvan i hälso- och sjukvården. Förslag till gemensam satsning 2011–2015. Stockholm: Regeringskansliet/Sveriges Kommuner och landsting; 2010.
2. http://qrcstockholm.se/register/vad-ar-ett-kvalitetsregister/

ACKNOWLEDGMENTS

I would like to thank Gunnar Hägglund and Ann Alriksson-Schmidt at Skåne University Hospital for the collaboration, giving us the experiences of working with Quality Registries and especially CPUP (Cerebral Palsy Follow-Up Program).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Catharina Dahlbo
Capish Nordic AB
Carlsgatan 3
211 20 MALMÖ, Sweden

+46 (0)40 10 88 80
catharina.dahlbo@capish.com
www.capish.com