Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-1

Detecting and Resolving Data Conflicts when Using International Claims Data for Research

Christian HAUX [a,1], Olivier KALMUS [b], Anna-Lena TRESCHER [b], Frank GABEL [b], Stefan LISTL [b,c] and Petra KNAUP [a]

[a] *University of Heidelberg, Institute of Medical Biometry and Informatics, Heidelberg, Germany*
[b] *University of Heidelberg, Department of Conservative Dentistry, Division of Translational Health Economics, Heidelberg, Germany*
[c] *Department of Quality and Safety of Oral Health Care, Radboud University, Nijmegen, the Netherlands*

Abstract. Using claims data for research is well established. However, most claims data analyses are focused on single countries. Multi-national approaches are scarce. The application of different anonymization techniques before data are shared for research as well as differences in the reimbursement systems hamper the use of claims data from multiple countries. This paper analyses data conflicts that occur when international claims data sets are used for research and develops a generic process to detect and resolve these conflicts. The approach was successfully applied in the EU-funded ADVOCATE (Added Value for Oral Care) project that acquired data from health insurance providers, health funds or health authorities in six European countries.

Keywords. Administrative data, secondary use, claims data, data conflicts

1. Introduction

Using claims data for research has great potential and is well established [1,2]. However, most claims data analyses are focused on single countries. Multi-national approaches, that use claims data from other countries, are scarce [3].

The European Union (EU)-funded project ADVOCATE ("Added Value for Oral Care") aims to use claims data on European level for dental care research by assessing the quality of dental health care services. The aim of the project is to analyze which national characteristics have a positive influence on these measures and to recommend successful strategies to other countries. The project involves the analysis of claims data from health insurance providers, health funds or health authorities in six European countries, namely Denmark, Germany, Hungary, Ireland, the Netherlands, and the

1 Corresponding Author: Christian Haux, University of Heidelberg, Institute of Medical Biometry and Informatics, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany, Email: christian.haux@med.uni-heidelberg.de

United Kingdom [4]. Oral health measures were defined by a group of experts [5]. The measures refer to topics regarding the access to dental care, symptoms and diagnosis, health behaviors, oral prevention and patient perception.

Differences in the underlying treatment codes and anonymization approaches that are applied to the data before they are shared for research purposes affect the feasibility of comparative analyses [6]. Hence, data quality and metadata descriptions must be assessed and data conflicts must be detected and resolved, prior to the analyses [7].

In this paper, we analyze on the example of the ADVOCATE project which data conflicts can occur when multiple international claims data sets are used for research and we develop a generic process to detect and resolve these conflicts.

2. Methods

A data cleaning process was developed in an evolutionary process. First, claims data were acquired from multiple data owners. In addition to the data sets, the data owners provided descriptions of their data (metadata). The descriptions specified semantic data characteristics, e.g. variables or attributes, columns and treatment codes of services. Characteristics of the data sets were analyzed and treatment codes corresponding to the respective oral health measures were identified. For each measure, a numerator and a denominator were defined that were available from the data.

A spreadsheet-based harmonization table as presented in Firnkorn et al. (2015) [8] was used to document the results. The table provides a comprehensive view for each claims data set compared to the corresponding general definition of the numerator and denominator of the oral health measure. For example, to identify patients that had at least one X-ray per year in the Danish data, billing events that involve one of the treatment codes 1150, 1151, 1152, or 1300 must be extracted. Table 1 depicts an extract of the harmonization table.

Various data conflicts were detected during the matching of treatment codes to the corresponding oral health measures that were classified according to the taxonomy of Spaccapietra et al. (1992) [9].

The matching was subsequently evaluated. Project partners in the respective countries were asked if the treatment codes that were identified from the metadata descriptions are suitable to calculate the particular oral health measure.

Table 1. Excerpt of the harmonization table that shows how to calculate the oral health measure "number of patients with at least one X-ray per year" with the Danish data.

Oral health measure		Claims data Denmark	
Numerator	**Denominator**	**Treatment codes**	**Denominator unit**
Number of patients with at least one X-ray	Total number of patients with claimable service per year	1150, 1151, 1152, 1300	Patients per year

After the feedback from the project partners was received, the oral health measures have been calculated with the available claims data sets.

3. Results

The development of the data cleaning workflow resulted in a three-step process: First, an initial matching of treatment codes to the corresponding oral health measures is developed based on the metadata that are provided by the data owners. The initial matching is subsequently validated by stakeholders from the respective countries. Data conflicts that are detected in these phases are resolved afterwards, if possible, and the final mapping is developed.

In the ADVOCATE project, the data cleaning process was performed with claims data sets and metadata from six data owners from Denmark, England, Germany, Hungary, Ireland, and the Netherlands. All data owners sent a description of their data and four additionally shared a data excerpt.

During the initial matching and the validation by project partners, various data conflicts were identified and classified:

Descriptive conflicts occur due to lack of information in claims data. This can occur due to differences in the reimbursement systems. If treatments are reimbursed as fixed sums, it is not possible to identify individual procedures from the data.

Semantic conflicts occur, when data are anonymized by the data owner before they are shared for research. Microaggregation or generalization are popular mechanisms to preserve the anonymity of the individuals. If the aggregation level is too high, specific measures cannot be calculated. For example, in the ADVOCATE project the calculation of measures was not possible on individual level for one county, because the data owner only provided data aggregated by the number of claimed services per quarter.

Additional semantic conflicts occur due to discrepancies between the meaning of treatment codes. In the ADVOCATE project, these conflicts were detected during the second phase. Project partners in the respective countries were asked which treatment code they would use to claim a specific procedure and if they see discrepancies to the codes that were selected from the data descriptions. If possible, these conflicts were resolved by choosing different treatment codes. This approach revealed some differences between the data descriptions and the actual application of treatment codes in practice. For example, although some treatment codes were listed in the data description, they were excluded by the project partners because they are not in use anymore for a long time or were not relevant to identify specific treatments. For example, to identify partial removable dentures from the Dutch data, the treatment codes F10, F15, F34, F35 were identified using the metadata description, which were removed later by the project partner, because these codes have not been in use anymore for a long time. For example, in case of the Danish data, it was possible to calculate twelve oral health measures with the available data. The feedback from the Danish project partner led to a correction of seven measures. No correction was recommended for the five remaining measures.

In the ADVOCATE project, 48 oral health measures were defined in total. Twenty three oral health measures could not be calculated for any of the countries. Twenty two measures have been calculated for single countries (e.g. number of teeth or X-rays). It has been possible to calculate three oral health measures for all six countries (periodontal examination, root canal treatment and tooth extraction).

4. Discussion

Using claims data for research has great potential. However, using claims data from multiple data owners is challenging because differences in data quality and underlying treatment codes can hamper the feasibility of analyses or the correctness of results.

This paper has described a data cleaning process as well as various data conflicts that can occur when international claims data from multiple data owners are used for research. The experiences made during the ADVOCATE project show that the feedback from the stakeholders facilitate the detection of semantic conflicts between the data definitions and the actual application in practices. Conflicts that concern differences between treatment codes can be resolved. However, other conflicts could not be solved. Furthermore, the content of claims data is limited because they are originally collected for administrative purposes. In the ADVOCATE project, measures that are related to health behaviors, e.g. tooth brushing could not be calculated using the claims data. Hence, alternative data sources must be found to gather missing information. Nevertheless, claims data are an important data source for research, because they include large numbers of observations and have a high representativeness [10].

The here presented approach is limited insofar, that it was developed pragmatically and the involvement of project partners was done during a later stage of the process. However, the involvement of the project partners was important to find discrepancies. To improve the process, domain experts such as dental practitioners or stakeholders from the health insurances should be involved already in the first phase of the process. The process can be extended towards a data harmonization approach that allows using claims data from multiple countries for comparative analyses.

5. Conclusion

Differences in benefit structures and the application of different anonymization techniques before data are shared for research hamper the use of claims data from multiple countries. By applying the data cleaning process, presented in this paper, it was possible to detect and classify data conflicts. The experiences made in the ADVOCATE project show that specific data conflicts can be resolved. The approach can be further standardized and used for other data integration projects that aim to use claims data from multiple countries.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 635183.

References

[1] H.M. Smeets, N.J. De Wit, A.W. Hoes, Routine health insurance data for scientific research: Potential and limitations of the Agis Health Database, *Journal of Clinical Epidemiology* **64** (2011), 424–430.

[2] J. Lane, C. Schur, Balancing access to health data and privacy: A review of the issues and approaches for the future, *Health Services Research* **45** (2010), 1456–1467.

[3] K. Kreis, S. Neubauer, M. Klora, A. Lange, et al., Status and perspectives of claims data analyses in Germany-A systematic review, *Health policy (Amsterdam, Netherlands)* **120** (2016), 213–226.
[4] H. Leggett, D. Duijster, G. Douglas, K. Eaton, et al., Toward More Patient-Centered and Prevention-Oriented Oral Health Care, *JDR Clinical & Translational Research* **2** (2017), 5–9.
[5] F. Baâdoudi, A. Trescher, D. Duijster, N. Maskrey, et al., A Consensus-Based Set of Measures for Oral Health Care, *Journal of Dental Research* **96** (2017), 856–890.
[6] P.T. Tyree, Challenges of Using Medical Insurance Claims Data for Utilization Analysis, *American Journal of Medical Quality* **4** (2006), 269–275.
[7] H. Adametz, A. Billig, Semantische Konflikte. White Paper: Semantic Interoperability, *Fraunhofer Institute ISST* **2** (2010).
[8] D. Firnkorn, M. Ganzinger, T. Muley, P. M. Thomas, P. Knaup, A generic data harmonization process for cross-linked research and network interaction, *Methods of information in medicine* **54** (2015), 455–460.
[9] S. Spaccapietra, C. Parent, Y. Dupont, Model independent assertions for integration of heterogeneous schemas, *The VLDB Journal* **1** (1992), 81–126.
[10] S. Schneeweis, J. Avorn, A review of uses of health care utilization databases for epidemiologic research on therapeutics, *Journal of Clinical Epidemiology* **58** (2005), 323–337.