OPEN ACCESS

Case Report

Data validation for healthcare cost analysis in STROKE OWL

Datenvalidierung in der Krankheitskostenanalyse in STROKE OWL

Abstract

The STROKE OWL project applies a new approach of cross-sector care management for strokes and determines the costs of this new approach. An evaluator determines the current costs of stroke care for the health insurance companies involved in the project as a comparative figure. Seven health insurance companies provide all necessary data in a uniform format for the health care cost analysis, despite different systems and internal formats. One important aim is to prevent the evaluator from performing analyses specific to individual health insurance companies. Therefore, a mediator accepts the data from the health insurance companies, checks the conformity of the data with the agreed format, and tries to make the re-identification of health insurance companies as difficult as possible through suitable transformations before forwarding the data to the evaluator. The conformance checking is particularly challenging due to underestimated effort and communication overhead for every participant. We propose a process for the data validation and a system assisting the validation of the mediator and describe our experience with the data validation for the health care cost analysis.

Keywords: data science, electronic data processing, health information exchange, data management

Zusammenfassung

Im Projekt STROKE OWL müssen die Kosten des neuen Ansatzes des sektorenübergreifenden Pflegemanagements ermittelt werden. Dazu muss der aktuelle Stand der Kosten eines Schlaganfalls für die Krankenkassen im Vorfeld analysiert werden. Diese Analyse wird im Rahmen des Projektes mit sieben Krankenkassen durchgeführt, die trotz unterschiedlicher Systeme und interner Formate die notwendigen Daten in einem einheitlichen Format zur Verfügung stellen. Ein wichtiges Kriterium ist es, zu verhindern, dass der Evaluator krankenkassenspezifische Analysen durchführen kann. Dafür ist ein Mediator zwischen Krankenkassen und Evaluator erforderlich. Der Mediator prüft, ob die Daten der Krankenkassen dem einheitlichen Format entsprechen. Die Überprüfung erweist sich aufgrund des unterschätzten Aufwandes und des Kommunikationsaufwandes für jeden Teilnehmer als besonders herausfordernd. Daher wird in diesem Paper ein Verfahren für die Datenvalidierung und ein System vorgeschlagen, dass die Validierung des Mediators unterstützt, und Erfahrungen im Rahmen der Datenvalidierung für die Krankheitskostenanalyse werden beschrieben.

Timo Wolters[1]
Timo Michelsen[1]
Christian Lüpkes[1]
Andreas Hein[1]

1 OFFIS – Institute for Information Technology, Oldenburg, Germany

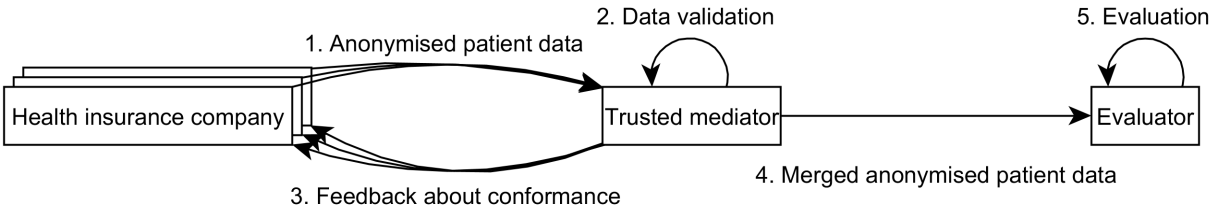

Figure 1: The process for the health care cost analysis

Introduction

The project STROKE OWL [1] (with its pilot region Ostwestfalen-Lippe (OWL), Germany) aims at a comprehensive implementation of cross-sector care management for patients after a stroke. So called *stroke pilots* carry out the actual care management process, who accompany patients in coordinating their lives after the stroke incident like case managers. The primary goal of the project is to measure the change of recurrence rates due to this new care management process.

A secondary goal is the evaluation of the impact on costs (e.g., for health insurance companies). Therefore, the evaluator in STROKE OWL needs data of health care costs of comparable patients in the timespan before the implementation of the new care management process. With this data, an evaluation of the status-quo of costs and region-specific cost structures is possible (so-called *health care cost analysis*).

Multiple health insurance companies (HICs) participate in STROKE OWL and agreed to deliver the needed data. On the one hand, it is important that all participating HICs send anonymous patient data to the evaluator in a uniform format. On the other hand, the evaluator should not be able to assign (parts of) data to individual HIC, thus disabling company specific evaluations.

Therefore, we propose the following 5-step-process for preventing health insurance specific evaluations (see Figure 1). In step 1, the HICs send anonymized patient data to the trusted mediator. The trusted mediator validates the data by comparing with the uniform data format (step 2) and reports to the HICs any technical deviations from the data format or acceptance of the data (step 3). If the data of all HICs are valid, the mediator merges the data so that it is no longer possible to trace which HIC has supplied which data. Finally, the evaluator receives the merged data (step 4) and performs the content check (step 5). After this process, the health care cost analysis is possible.

In this context, the health care cost analysis depends significantly on the conformity of the anonymized patient data to the data format, since, according to Romeo and Thoresen [2], analyses with highly noisy data require considerable effort to perform and have a high error rate. Data validation is crucial for the success of the health care cost analysis.

Due to the feedback loop of steps 1, 2, and 3, the validation is potentially iterative for each HIC until the delivered data matches the format. The different size of HICs leads to varying amounts of data. For this reason, the validation should be automatic, efficient, scalable, and should also remove or replace HIC specific characteristics. In this paper, we propose such a validation system to support the 5-step process mentioned before.

The rest of the work is structured as follows: in section 2, we describe the preconditions such as the data format and processes prior to data delivery as well as the structure of the associated system for data validation from the mediators' point of view. Section 3 summarizes the application of the system at the beginning and the end of feedback loop while giving an overview of frequently occurring deviations across HICs. Section 4 discusses the frequent deviations together with possible changes to the validation system and data format that arose in the feedback loop and future work. Section 5 concludes with a summary.

Case description

In STROKE OWL, seven HICs participate in the health care cost analysis. The number of patients per HIC ranges from a few patients to 40% of all insured persons in the study area. The HICs have similar databases, but they differ significantly in terms of internal formats for standards such as EBM numbers or aid and cure numbers. Analysis of these unadapted data requires integration of autonomous data sources similar to the scenario in [3]. Unlike Cabibbo et al. in [3], there are no data marts or data warehouses, but databases, making data integration more difficult. In the STROKE OWL project, integration is also completely avoidable.

In order to avoid the integration of heterogeneous data, the HICs, mediator and evaluator have created a uniform data format according to which the HICs supply data, the mediator checks the syntactic correctness and the evaluator sets up the health care cost analysis. The data format also allows other HICs to contribute data to future analyses too without affecting other parties outside the mediator. For the mediator himself only the composition of the data package for the evaluator changes, but everything else is unaffected.

In order to agree on a jointly usable data format with so many HICs with individual data records and internal data formats, many iterative discussions were necessary. Together we created several versions of the data format, discussed them, and continuously refined them. The result is a data format for the health care cost analysis, which consists of 18 tables and 115 columns. Insured person numbers and internal case numbers allow connec-

tions between tables or groupings. However, in different contexts, the data format has the same characteristics like money values, ICD codes for diagnoses or dates. This results in 35 different column formats across all tables, which the mediator has to check for conformance with the data format.
The tasks of the trusted mediator are two-fold:

1. check conformance with the agreed data format and
2. combine all data of every HIC in such a way that no conclusions about individual HICs are possible. Plausibility checks, completeness checks and other content checks are tasks of the evaluator and are not part of this paper.

However, transparency of the trusted mediator is an important criterion: the received data should be traceable, the validation time documented, and – if necessary – the mediator should communicate changes made to the data in the process. In particular, in case of inquiries from the HICs or the evaluator, the trusted mediator must be able to "translate" used key identifiers and pseudonyms.
The amount of data of the HICs can become very large. Therefore, a manual validation is too error-prone and not feasible in a reasonable time. An automatic check of the data is more useful at this point. In addition, this also enables precise feedback. The automatic check can also perform possible pseudonymization. Nevertheless, after the automatic check, the mediator should carry out a manual check of the deviations found in order to identify possible errors of the automatic check and to convert the deviations found into a more easily understandable format.
This leads to the architecture for a program that performs the automated checks and pseudonymization in Figure 2. In addition to the main process consisting of validation, transformation, and storage, other tasks such as configuration and the creation of in-memory logs for the transparency of the process are included. The validation extends the architecture of Ao et al. [4] with a declarative approach (see Figure 3).

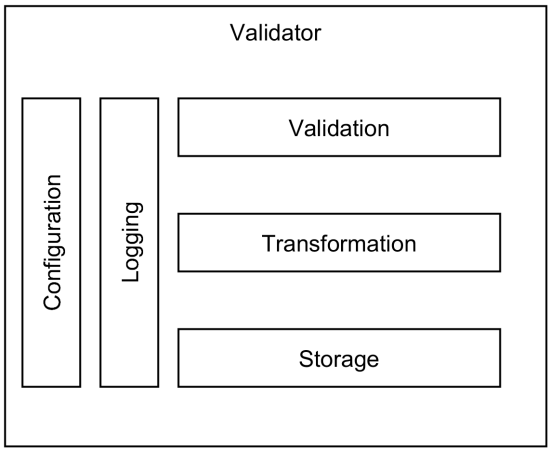

Figure 2: General architecture of the validation system

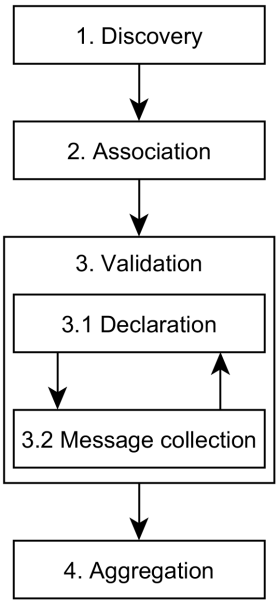

Figure 3: Procedure for the validation based on [4]

At first, there is an exploratory step where all possible files are collected (step 1). In the association step (step 2), found files are assigned to (1) the different HICs and (2) the corresponding table of the data format. This step filters files that are not assignable to any HIC or table. During validation (step 3), the table declares and describes the order and column formats that shall be used (step 3.1). This declaration step eases applying changes of the data format to the program as well as temporary changes of the declaration, if columns are swapped or additional columns are present.
Depending on their applicability, the checking of column formats ranges from simple list comparisons with valid values to regular expressions that check the structure of standards or ambiguous specifications to the calculation of check digits if the standards used contain check digits. For gender information, a simple selection list for male (m), female (w) and other (o) is sufficient, while the validation of money uses regular expressions, because the decimal separator can be either a dot or a comma and values may have a maximum of two digits after the decimal separator. A Pharmacy Product Number in Germany consist of seven digits and a check digit, which is calculated and checked during validation.
After validation of an entry, the (error) messages are collected and the next column is processed (step 3.2). When the validation is complete, in step 4 the data is aggregated and the messages are written into log files. Therefore, the log files guarantee full transparency and unique messages can be easily viewed per column and file. This is especially helpful for structural errors.
The transformation afterwards serves to remove characteristics, which could identify specific HICs. Identifying characteristics are the insured person numbers, internal case numbers and any deviations from the data format. The data format contains no specifications for the structure of the insured person numbers or internal case numbers.

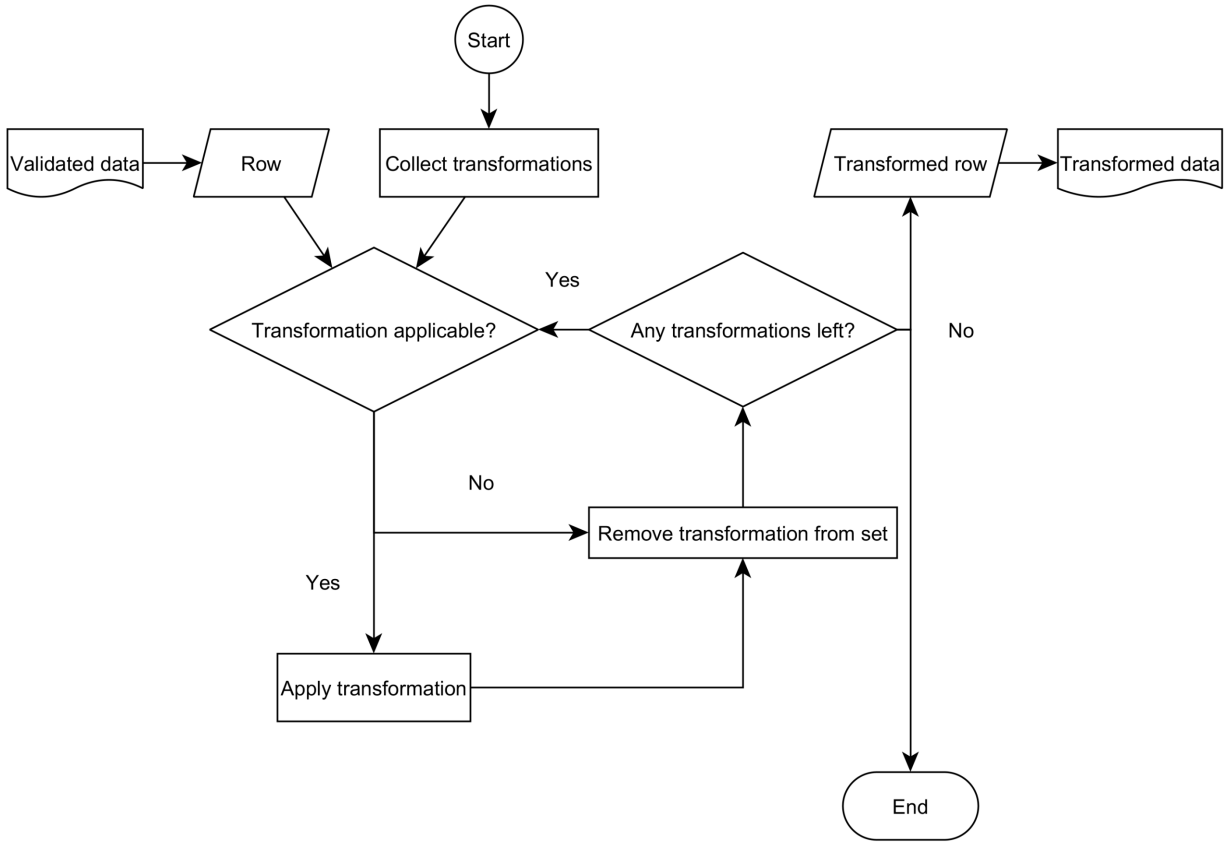

Figure 4: The transformation procedure after the validation

This means that all HICs form both numbers differently and the structure of these numbers can identify a HIC. Additionally, individual HICs cannot guarantee the uniqueness of their numbers over all HICs. Structural and content-related deviations allow grouping of the data and the mediator removes, together with the HIC – or the evaluator if the HIC does not remove the deviations – prior to forwarding the data these deviations as far as possible.

Overall, the transformation process proceeds as described in Figure 4. In order to be able to integrate the resolutions for deviations without adapting the systems, the transformation step uses a set of form-retaining transformations. A form-retaining transformation only changes the column contents, but not the number of columns or their order. For example, a row consisting of an insured person number, a diagnosis, and a diagnosis date will still contain the three columns after the transformation. However, a column transformation may have replaced the insured person number, a diagnosis, and a diagnosis date will still contain the three columns after the transformation. However, a column transformation may have replaced the insured person number with a pseudonym and/or removed special characters from the ICD codes of the diagnosis.

The transformation also defines its applicability to the given row. If it is applicable, the transformation step executes the transformation for the given row and removes it from the set of available transformations. If it is not applicable, the transformation step also removes it from the set and continues with the next transformation as long as transformations still exist in the set. In this way, the validated data changes gradually, but systematically with each applied transformation so that the revised rows are available at the end of the transformation. The program can then assemble the revised rows using the original files or the assigned tables, so that the mediator may inform the HICs about changes to their original data and send the evaluator the data of all HICs. The implemented transformations include, for example, the replacement of insured and internal case numbers, both of which implement a procedure for generating unique numbers from a finite set.

After validation and transformation, a lot of information is available in the form of persistent logs. For traceability purposes, a log file records the transformations of the insured and internal case numbers, and a log of the found deviations from the data format exists. In addition, the data revised by the transformation step are also stored, so that the mediator can send the data to the evaluator or HIC. The logs are stored as CSV files.

Based on the logs, the trusted mediator creates a human readable protocol as feedback to the HIC, whereby the mediator checks the validity of the errors at this point. For example, if a HIC accidentally swaps two columns in a table, then the program will generate error messages for invalid values for both columns, although the actual deviation is in the swapped columns. For each table and column, it is possible to determine how often which error occurred and thus report the exact rows and columns. If

Table 1: The 10 most frequent deviations before the feedback loop

Type of deviation	# Messages initially	# Messages finally
money format	1,387,810	0
ICD format	1,043,630	803,640
date format	569,892	47,146
invalid PPN	220,937	220,937
invalid row length	212,493	0
invalid professional group key	194,562	120,949
foreign key does not exist	115,455	0
unknown diagnosis type	110,549	75,304
OPS format	90,404	44,882
unknown performance type	52,649	46,001

an error occurs many times, the mediator can formulate general recommendations for corrections.

Results

The result of the feedback loop divides itself into two parts: deviations found initially and deviations still present at the end of the feedback loop. To compare the two situations, the program checks the data initially sent by the HICs with the data last sent and compares it using various criteria.

The total number of log entries and the number of unique messages are considered. A unique message is defined here as an aggregate of a message that occurs multiple times in a single file and column. This means that structural errors in particular are only counted once and not for each row.

Initially, the HICs sent 102 files, containing over 6 million rows about approximately 8,200 patients, while 104 files with about 6.1 million lines about the same number of patients were sent at the end. In both cases, the program generated 243 MB of pseudonymized data from the approximately 308 MB of original data. The difference in size has its origin in the way the HICs form the insured person number and the internal case number and their usage. The insured person number occurs once in each line and the internal case number occurs in many lines. The pseudonyms are between 10 and 20 characters smaller than the numbers formed by the HICs, which results in less memory consumption.

Initially the program derived about 4.2 million log entries from the data and about 170,800 unique messages. On the one hand, this means that on average a unique message occurs 25 times. The maximum on the other hand is about 100,000, which suggests that a few errors occur very frequently. In total, this corresponds to a size of 1.09 GB of generated data. The generated data includes the log entries, the transformation logs, and the resulting pseudonymized data.

The program distinguishes 72 different deviations for the 35 column formats, of which 37 occurred initially and 30 are still present at the end. Table 1 shows the 10 most frequent deviations occurring initially and their reduction until the end of the feedback loop. The deviation types aggregate some types of errors as a complete presentation would go beyond the scope of this paper. For example, the program distinguishes between a valid ICD code without special characters (correct by data format), a valid ICD code with special characters (warning) and an invalid ICD code (error). Table 1 summarizes the warnings and errors under the item ICD format. These errors cover about 92% of all messages initially and about 71% of all messages at the end. Noticeable is the complete correction of money formats as well as the almost complete correction of date formats while HICs have only corrected about 20% of the deviations of ICD codes.

With the last sent data, there were still 1.9 million log entries with about 50,000 unique messages. This means that a unique message occurs 38 times on average, with the median being two and the third quantile seven. We can conclude that the HICs corrected many deviations that have occurred once, but the HICs fixed frequently occurring deviations only partially. This corresponds to a total size of 603 MB of generated data.

To provide a better overview of frequent deviations, we consider only the error types for the following. By far the most frequent retained deviation is in the ICD codes, with initially over 1 million and finally about 0.8 million occurrences. Here, the HICs have rarely corrected deviations, and the proportion of messages for this deviation is about 42%. The HICs reduced errors in the formatting of dates from initially over 569,000 to about 50,000. Other discrepancies such as too many or too few columns and missing relationships between the insured person number and the master data were removed completely. The professional group key '00' instead of an empty character string was initially reduced from about 194,000 occurrences to 120,000 occurrences, but still makes up 7% of all messages. The deviations in the ICD codes and the professional group key add up to about 50% of the remaining deviations.

Discussion

The question at this point is how to handle very frequently occurring deviations after the validation. Possibilities include

1. changing the data format,
2. changing the communication between HICs and the mediator or
3. changing the process.

It is only possible to change the data format if the deviation is consistent. In the case of the deviations of the ICD codes, the HICs used human readable codes instead of the machine-readable codes and the data format could reflect this trend. Changes in the communication between HICs and the mediator are useful if, for example, parts of the data format were misunderstood or potentially not read. There might be a communication problem with the deviation of the specialist group since the HICs used a different value for unknown values instead of the value specified in the data format. One could also change the process for the individual steps. For example, the HICs could use an ETL process for creating data on the HIC side.

Lessons learned

In order to prevent the communication overheads, the evaluator and mediator clearly communicate the expected effort for the preparations to the HICs by highlighting the scope of the data, the data format, and the presumed size of the HIC itself.
Participants should understand the data format as a continuously evolving document in order to be able to make and communicate adjustments for uniform deviations or additional options with the least possible effort. The mediator can then translate the uniform deviations or options into the format required by the evaluator.
This approach is particularly useful if all HICs involved use the same internal standard for the presentation of certain contents. This can reduce the effort for the HICs, since they do not need to project data on their side, thus avoiding potential errors.
If no uniform formats are used, then the mediator and evaluator may suggest a process for data deduction to the HICs. For example, the first step is to extract data from the databases. In the second step, one selects the data relevant for the health care costs analysis and puts it into the correct table structure. In the third step, the HIC projects internal representations to the required values and finally saves them in the required format as the last step. This would correspond to an ETL process that is usually already available for internal evaluations of the HICs and the HIC can use it in an adapted form for the health care costs analysis.

Future work

In the future, we could expand the transformation step of the algorithm by adjusting the data in addition to pseudonymization. For example, the HICs deliver the ICD codes in the required machine format or in a human readable format and the program converts the codes to the required format during transformation.
The project should also revise the data format on the basis of experience within the health care cost analysis, so that the value for unknown is also accepted everywhere.
In addition, the project STROKE OWL carries out a further analysis approximately in 2021, which is very similar to the health care cost analysis.

Conclusion

Overall, there were many challenges in validating the data for the health care cost analysis. These include organizational challenges in order to integrate the data of seven health insurance companies (HICs) through a uniform data format, despite different data source systems. One should remove references to the HICs to make health insurance-specific evaluations more difficult. The challenges of designing the data format for the representation of standards such as ICD codes, groups of doctors or groups of persons also meant constant changes to the data format. The originator has communicated these changes to all parties involved, but the health insurance companies have still not implemented all data format specifications. In particular, deviations from ICD codes still account for more than 40% of all remaining deviations at the end of the validation. Since these deviations are uniform for all HICs, the validator can map them to the actual data format and must report this to the HICs. Overall, the addition of a program for automatic testing together with the creation of human-readable protocols was a successful approach to meet the concerns of the health insurance companies regarding health insurance-specific evaluations, while at the same time reducing the evaluators workload by providing all data in a uniform format.

Notes

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the Federal Joint Committee German Innovation Fund within the joint research project STROKE OWL (grant no. 01NVF17025).

References

1. Stiftung Deutsche Schlaganfallhilfe. STROKE OWL Project Website. [cited 2020 March 25]. Available from: https://stroke-owl.de/de/startseite/
2. Romeo G, Thoresen M. Model selection in high-dimensional noisy data: a simulation study. J Stat Comput Simul. 2019;89(11):2031-50. DOI: 10.1080/00949655.2019.1607345
3. Cabibbo L, Torlone R. On the integration of autonomous data marts. In: Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM); 2004 Jun 21-23; Santorini Island. p. 223-31. DOI: 10.1109/SSDM.2004.1311214
4. Ao KLB, Bull S. Architecture for data validation. US Patent 8,418,142. 2013.

Corresponding author:
Timo Wolters
OFFIS e.V., Escherweg 2, 26121 Oldenburg, Germany
timo.wolters@offis.de

Please cite as
Wolters T, Michelsen T, Lüpkes C, Hein A. Data validation for healthcare cost analysis in STROKE OWL. GMS Med Inform Biom Epidemiol. 2020;16(2):Doc06.
DOI: 10.3205/mibe000209, URN: urn:nbn:de:0183-mibe0002098

This article is freely available from
https://www.egms.de/en/journals/mibe/2020-16/mibe000209.shtml

Published: *2020-08-25*