Aus der Klinik für Neonatologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Die Anwendung von medizinischen Terminologien auf Freitext in Routinedatenbanken am Beispiel von Strategien zur Reduktion der Säuglingssterblichkeit

The application of medical terminologies to free-text in routine databases using the example of strategies to reduce infant mortality

zur Erlangung des akademischen Grades
Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

André Sander

aus Berlin

Datum der Promotion: 21.06.2020

Inhalt

Abstract (DE)

Hintergrund

Die Säuglingssterblichkeitsrate (IMR), ein wichtiger Indikator für die Qualität eines Gesundheitssystems, liegt in Deutschland seit 10 Jahren bei rund 3.5‰. Generische Qualitätsindikatoren (QIs), wie sie seit 2010 in Deutschland verwendet werden, tragen wesentlich zu einem so guten Wert bei, scheinen aber nicht in der Lage zu sein, den IMR weiter zu reduzieren. Die neonatale Sterblichkeitsrate (NMR) trägt zu 65-70% der IMR bei. Der vorgestellte Ansatz schlägt daher eine Einzelfallanalyse neonataler Todesfälle auf der Grundlage von Krankenakten vor. Die meisten elektronischen Krankenakten enthalten noch immer große Mengen an Freitextdaten. Die semantische Auswertung solcher Daten erfordert, dass die Daten mit ausreichenden Klassifizierungen kodiert oder in eine wissensbasierte Datenbank umgewandelt werden.

Methodik

Die Nordic-Baltic-Classification (NBC) wurde zur Erkennung vermeidbarer neonataler Todesfälle verwendet. Diese Klassifikation wurde auf eine Stichprobe von 1.968 neonatalen Todesfällen angewandt, die über 90% aller neonatalen Todesfälle in Ost-Berlin von 1973 bis 1989 darstellen. Alle Fälle wurden damals von einer speziellen Kommission verschiedener Experten auf der Grundlage der vollständigen perinatalen und klinischen Daten auf ihre Vermeidbarkeit hin analysiert.
Der entwickelte Ansatz ermöglicht es, Datenbanken, die über SQL (Structured Query Language) zugänglich sind, direkt über semantische Abfragen zu durchsuchen, ohne dass weitere Transformationen erforderlich sind. Dazu wurden 1.) eine Erweiterung von SQL „Ontology-SQL“ (O-SQL) entwickelt, die es ermöglicht, semantische Ausdrücke zu verwenden, 2.) ein Framework entwickelt, das einen Standardterminologieserver verwendet, um Freitext enthaltende Datenbanktabellen zu annotieren und 3.) ein Parser entwickelt, der O-SQL Ausdrücke in SQL konvertiert, so dass semantische Abfragen direkt an den Datenbankserver weitergeleitet werden können.

Ergebnisse

Die NBC wurde verwendet, um die Gruppe der Fälle auszuwählen, die ein hohes Vermeidungspotenzial hatten. Die ausgewählte Gruppe stellte 6,0% aller Fälle dar und 60,4% der Fälle innerhalb dieser Gruppe wurden tatsächlich als vermeidbar oder bedingt vermeidbar beurteilt. Die automatische Erkennung von Fehlbildungen ergab einen F1-Wert von 0,94. Darüber hinaus wurde die Verallgemeinerbarkeit des Ansatzes mit verschiedenen semantischen Abfragen nachgewiesen und dessen Güte mit F1-Werten von 0,91 bis 0,98 gemessen.

Zusammenfassung

Die Ergebnisse zeigen, dass die vorgestellte Methode automatisch anwendbar ist und ein leistungsfähiges und hochsensitives und -spezifisches Werkzeug zur Auswahl potenziell vermeidbarer neonataler Todesfälle und damit zur Unterstützung einer effizienten Einzelfallanalyse darstellt. Die nahtlose Verknüpfung von Ontologien und Standardtechnologien aus dem Datenbankbereich stellt einen wichtigen Bestandteil der unstrukturierten Datenanalyse dar. Die entwickelte Technologie lässt sich problemlos auf aktuelle Daten anwenden und unterstützt das immer wichtiger werdende Feld der translationalen Forschung.

Abstract (EN)

Background

The infant mortality rate (IMR), a key indicator of the quality of a healthcare system, has remained at approximately 3.5‰ for the past 10 years in Germany. Generic quality indicators (QIs), as used in Germany since 2010, greatly help to ensure such a good value but do not seem to be able to further reduce the IMR. The neonatal mortality rate (NMR) contributes to 65-70% of the IMR. The presented approach therefore proposes single-case analysis of neonatal deaths on base of medical records. Most electronic medical records still contain large amounts of free-text data. Semantic evaluation of such data requires the data to be encoded with sufficient classifications or transformed into a knowledge-based database.

Methods

The Nordic-Baltic classification (NBC) was used to detect avoidable neonatal deaths. This classification has been applied to a sample of 1,968 neonatal death records, which represent over 90% of all neonatal deaths in East Berlin from 1973 to 1989. All cases were analyzed as to their preventability based on the complete perinatal and clinical data by a special commission of different experts.
The developed approach allows databases accessible via SQL (Structured Query Language) to be searched directly through semantic queries without the need for further transformations. Therefore, I) an extension to SQL named Ontology-SQL (O-SQL) that allows to use semantic expressions, II) a framework that uses a standard terminology server to annotate free-text containing database tables and III) a parser that rewrites O-SQL to SQL, so that such queries can be passed to the database server, have been developed.

Results

The NBC was used to select the group of cases that had a high potential of avoidance. The selected group represented 6.0% of all cases, and 60.4% of the cases within that group were judged avoidable or conditionally avoidable. The automatic detection of malformations showed an F1 score of 0.94. Furthermore, the generability has been proved with different semantic queries and was measured with between 0.91 and 0.98.

Conclusion

The results show, that the presented method can be applied automatically and is a powerful and highly specific tool for selecting potentially avoidable neonatal deaths and thus for supporting efficient single case analysis. The seamless connection of ontologies and standard technologies from the database field represents an important constituent of unstructured data analysis. The developed technology can be readily applied to current data and supports the increasingly important field of translational research.

Background and related work

The infant mortality rate (IMR) is a key indicator of the quality of a healthcare system and heavily influences the overall pediatric mortality [1]. In Germany, the IMR has remained at approximately 3.5‰ for the past 10 years, whereas many other countries have managed to constantly further reduce that rate (see Figure 1). However, international comparisons are controversial, and more advantageous values of the IMR in other countries might be caused by different definitions of a stillbirth. Therefore, it is reasonable to focus on the further reduction of IMR within an established healthcare system such as Germany.

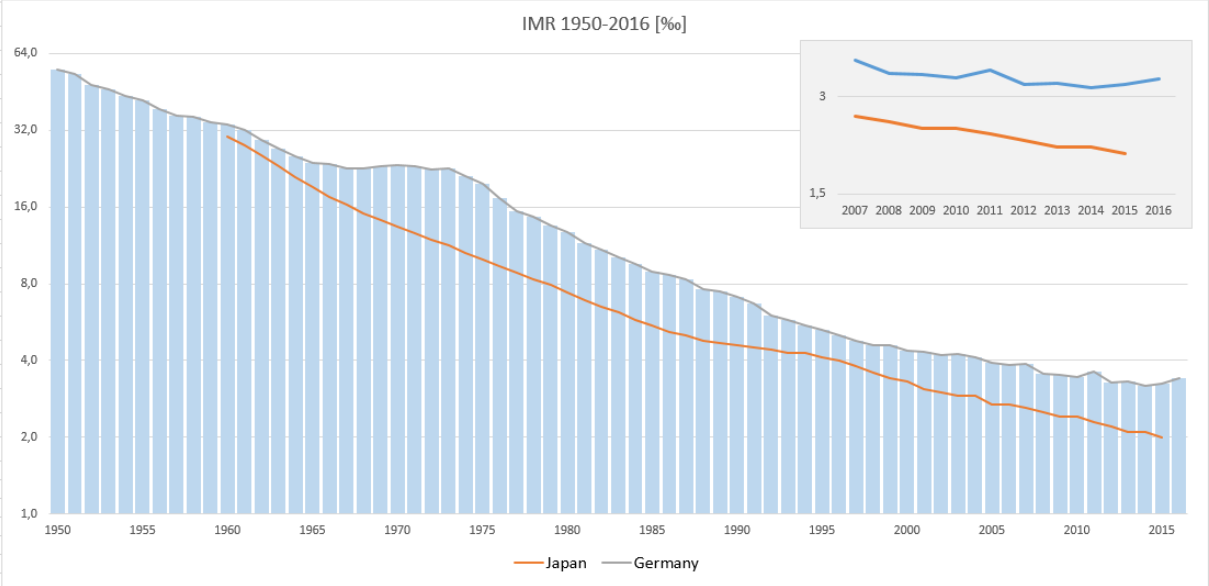

Figure 1 Development of the infant mortality rate (IMR): comparison between the IMRs in Germany and Japan 1950-2016. Detailed view shows the last 10 years (*Data before 1990 from West Germany) [Data from statistical yearbooks and knoema.de].*

The IMR is defined as the number of deaths of children under one year of age per 1,000 live births. A neonatal death is defined as a death during the first 28 days of life (0-27 days) and is further differentiated into an early death (0-6 days) and a late death (7-27 days). The neonatal mortality rate (NMR) is the number of neonatal deaths per 1,000 live births. The perinatal mortality rate (PMR) includes stillborn cases in the early NMR.

The current methods to measure and influence quality are based on calculated quality indicators (QIs) and perinatal surveys. Although Germany achieves very good values, the methods used have two major obstacles. First, these methods lack comparability due to different data and measures used, and they have a strong focus on morbidity instead of mortality. Second, neonatal and obstetric data must be merged to allow diagnostic data analysis. Diagnostic data include data regarding maternal anamnesis, delivery and the clinical course of the newborn. In reality, the datasets from perinatal and neonatal surveys and statistical offices are not adjusted to each other and thus cannot be merged [2]. In 2015, the "Association of the Scientific Medical Societies" (AWMF) stated: "In Germany, however, recording the quality of results with the help of neonatal survey is reaching its limits".

Using routine data is desirable and the launching of a working group on the "use of electronic medical records for clinical research" in 2011 by the GMDS (Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie) is evidence of the enormous importance of medical records for research; however, it also underlines the difficulties that arise when trying to analyze these data.

The key to further increased quality is controlling and changing workflows. For that reason, the German Aqua Institute, responsible for quality measures in healthcare, started a discussion on alternatives to pure

statistical indicators. These indicators are oftentimes not sufficiently sensitive or even use questionable items [3]. Additionally, a QI typically measures the outcome (quality) and is not able to evaluate the underlying process. The actual cause of death can be traced back to various reasons, which come from different fields, such as obstetrics, postnatal treatment, transport, surgical treatment, care, and hygiene. It is proposed that the actual cause of death can be explained by only single-case analysis. A recent paper by Masson et al. described how such contributing factors can be identified from single-case analysis [4]. Another recently published paper showed how single-case analysis influences quality assurance [5].

That such analysis positively influences the IMR could even be observed in Germany, namely, before 1990 in the eastern part of Germany, in the former German Democratic Republic (GDR). The Commission on the Reduction of Infant Mortality was established beginning in 1958 in East Germany, and from 1970 on, the activities of this commission were carried out in accordance with uniform guidelines [6, 7]. The commission (Fachkommissionen zur Bekämpfung der Mütter- und Säuglingssterblichkeit auf Kreis-, Bezirks- und Landessebene) consisted of various specialists, such as pediatricians and neonatologists, obstetricians, pediatric and cardiac surgeons and most important, a pathologist whose responsibility was to determine the final cause of death. Thus, the data collected by this commission is unique in terms of medical precision. The primary goal was the critical analysis of single infant deaths in order to detect errors in healthcare management. The method of choice was the legally required single-case analysis. Every neonatal death had to be reported to and was discussed in front of that commission. The activity of this commission and its results were so successful that it continued until the end of the GDR in late 1989.

Datawise, two major problems are addressed with this approach: I) the semantic structuring and evaluation of free text and II) the querying of such information from medical records.

Many research papers have presented different approaches to the semantic structuring of free text. The outcomes vary in many aspects; however, in general, these approaches provide good to excellent results. Recent NLP (natural language processing)-based mapping systems were published by Friedman et al. in 2004 [9] and later by Savova et al. [10], both of which exhibited high accuracy. Elkin et al. analyzed mapping algorithms with the SNOMED (systematized nomenclature of medicine) ontology on chest X-ray reports with excellent results [11]. A similar task for pathology was recently presented by Allones et al. but required further improvement [12]. In the German language, a 2015 paper presented by Toepfer et al. showed very good results [13].

Many attempts to enable knowledge-based querying of information from medical records have been described; for example, Hogarth et al. suggested the so-called TQL (terminology query language), which encompassed SQL's idea of universality [14]. SPARQL (SPARQL Protocol and RDF Query Language) in particular has established itself in many areas as the de facto standard. However, for specific problems, such as mapping between ontologies, additional internal query languages have been developed, even in recent years [15]. There has been an evolution of integrated frameworks for the implementation of browser-based knowledge systems since early on [16, 17]. Today, SPARQL- and OWL (Web Ontology Language)-based systems are successfully implemented for defined applications, such as the management of blood pressure or hypertension [18]. In the area of infectious diseases, Kama et al. used the concept of “a semantic data warehouse” that integrates OLAP (online analytical processing) techniques [19]. Nevertheless, in this approach, specific query languages have remained unaltered in their respective domains. Epstein et al. therefore chose to implement and integrate needed subsystems (e.g., NLP pipelines) into SQL [20]. One of the most recent and interesting approaches came from Zheng et al., who have extended standard SQL with “semantic constructs” [21]. Nevertheless, for this purpose, numerous algorithms have been implemented in the middleware instead of consequently outsourcing them to a terminology server.

Classification-based approaches (e.g., ICD (International Classification of Diseases)-encoded data) are also used in current attempts to integrate heterogeneous data sources for cohort formation [22]. This approach,

however, is associated with three major problems: I) loss of specificity, II) low interoperability and III) low stability over time.

In this paper, single-case analysis is proposed as an additional method to reduce the IMR, and an efficient NLP-based way is presented to implement this proposal. The developed approach overcomes many of the described technical hurdles and still achieves comparable or even better results. Several algorithms are combined that automatically detect from electronic documentation the fatal cases that have the highest potential of avoidance. Analyzing that group of cases promises high knowledge gain, resulting in actions that can be taken and a reduced IMR.

Methods and material

Mortality prediction is a well-established method that can be performed based on many parameters (e.g., the Apgar score and maternal risk factors such as obesity and smoking), but the detection of deaths that could be avoided is much more difficult.

Avoidable are often deaths with selected causes of death, which are believed to could have been prevented (within a certain age range) in principle with appropriate treatment and prevention. There are a number of competing approaches to defining avoidable mortality but there is currently no international consensus on the definition [8].

The here used criteria of "avoidance" are based on common principles like

- Late, wrong or missing treatment
- Failures of organization (transportation, wrong assignment)
- Indicating diagnoses like nosocomial infections

In the presented approach, the “Nordic-Baltic classification” (NBC) is being used, which was developed to detect high-risk groups and groups with a high potential for avoidance [23]. This classification is relatively small and consists of only 13 classes (see Table 1). The most striking feature of this classification is that all cases are first divided into cases with and without malformations. Within the group of cases without malformations, further classification is done mainly based on the perinatal period (antenatal, intrapartum and neonatal) of the gestation week and on the Apgar score. Each group has a specific risk of death and thus a specific potential of avoidance (groups with a rather low risk have a naturally high potential of avoidance). The most interesting group was found in group “X”. The highest potential of avoidance of neonatal deaths is expected in that group because it contains cases that were born on time and that had a good Apgar score. Additionally, a new subgroup (“Xa”) was added that contained all cases from group “X” that survived at least 24 hours. An even higher potential of avoidance was expected in that group because deaths after 24 hours point to hygienic and other problems that are related to mismanagement and are thus avoidable.

Estimation of the completeness of the data set for East Berlin was performed based on historical documents and on numbers published by the Ministry of Health of the GDR. For some years, the exact number of neonatal deaths could be determined, while for other years, only the IMR was published, and the absolute number was calculated (see Table 2).

Class	Definition	Cases in that class [N]	Avoidable cases [N]	Avoidable cases [%]
I	Fetal malformation.	513	82	15.6
II	Antenatal death; single growth-retarded fetus ≥28 weeks of gestation.			
III	Antenatal death; single fetus ≥28 weeks of gestation.			
IV	Antenatal death; before 28 weeks of gestation.			
V	Antenatal death; multiple pregnancy.			
VI	Intrapartum death; after admission; ≥28 weeks of gestation.			
VII	Intrapartum death; after admission; before 28 weeks of gestation.			
VIII	Neonatal death; 28-33 weeks of gestation. Apgar score >6 after 5 min.	225	81	36.0
IX	Neonatal death; 28-33 weeks of gestation. Apgar score <7 after 5 min.	252	85	33.7
X	Neonatal death; ≥34 weeks of gestation. Apgar score >6 after 5 min.	125	70	**56.0**
Xa	Neonatal death; ≥34 weeks of gestation. Apgar score >6 after 5 min.; age >24 hours.	91	55	**60.4**
XI	Neonatal death; ≥34 weeks of gestation. Apgar score <7 after 5 min.	72	30	41.7
XIa	Neonatal death; ≥34 weeks of gestation. Apgar score <7 after 5 min.; age >24 hours.	37	15	40.5
XII	Neonatal death; before 28 weeks of gestation.	213	65	30.5
XIII	Unclassified.			
Σ		1,528	483	31.6

Table 1 Nordic-Baltic classification with extension (Xa and XIa) [23]. The "Number of cases in that class" column includes the fraction of the 1,528 used cases that was assigned to that class. The "Number of avoidable cases" column includes both avoidable and conditionally avoidable cases and was derived from a judgment made by a commission of experts (the gold standard). The blue sections show that classes "X" and "Xa" contain a very high number of avoidable cases and thus are good predictors of avoidance.

The gold standard consisted of 1,868 cases of live births that died in the first 28 days of life and therefore are neonatal deaths by definition. These cases were recorded between 1973 and 1989 in Berlin (East) by the Commission on the Reduction of Infant Mortality. By that time, the commission was led by the first European chair for neonatology [24, 25]. In preparation for each meeting, she prepared an index card that described the case and had a final judgment made by the commission if the case was avoidable, conditionally avoidable or not avoidable. In many cases, the index card also contained a revision of the initial judgment including an explanatory statement.

The index cards contained basic demographic data regarding the mother and child and free text regarding anamnesis, birth, postnatal treatment and course of death. Structured information could be retrieved for age (in hours), week of gestation, the Apgar score, weight and length and other parameters (see Table 3 for a complete list and statistics and Figure 2 for a sample).

Year	Births	Still births [N]	IM [N]	IMR [‰]	NM [N]	NMR [‰]	NMR/IMR [%]	Index cards* [N]	Coverage [%]	Avoidable [%]	Conditionally avoidable [%]
1	2	3	4	5	6	7	8	9	10	11	12
1973	10,907	86	179	16.4				62			
1974	11,054	81	176	15.9	*118*		*67.0*	108	*91.5*	1.9	16.7
1975	11,748	90	181	15.4	*121*		*67.0*	100	*82.6*	0.0	20.0
1976	13,365		202	15.1				119		0.0	21.0
1977				13.0				104		0.0	14.4
1978	15,664	116	182	11.6	*122*		*67.0*	120	*98.4*	1.7	20.8
1979	16,525	106	208	12.6	145	8.8	69.7	139	95.9	2.2	22.3
1980	17,526	103	233	13.3	166	9.5	71.2	156	94.0	3.2	26.3
1981				13.5		8.9		151		2.6	23.2
1982	17,725		234	13.2	160	9.0	68.4	161	100.0	1.9	32.3
1983	17,745	98	195	11.0	131	7.4	67.2	121	92.4	5.0	22.3
1984	16,885	75	159	9.4	91	5.4	57.2	84	92.3	15.5	33.3
1985	17,156	70	189	11.0	*127*		*67.0*	129	*100.0*	8.5	39.5
1986	17,467		168	9.6				95		14.7	30.5
1987	18,399	68	155	8.4	109	5.9	70.3	103	94.5	14.6	37.9
1988	17,965	85	148	8.3	99	5.5	66.9	88	88.9	21.6	38.6
1989	16,937	66	129	7.6				26			

*Table 2 Number of births, stillbirths, and deaths within the first year of life (IM: infant mortality) and the first month of life (NM) of all infants born in East Berlin between 1973 and 1989. The column "Index cards" represents the number of index cards available in the original data set. The column "Coverage" contains the ratio of available cards to all cases, which shows a high coverage of 93.7% on average. "Avoidable" and "conditionally avoidable" ratios are calculated based on the column "Index cards". The increase in the proportion of avoidable and conditionally avoidable deaths was statistically significant, with $p<0.0001$ and $p=0.0006$, respectively. Gray numbers are estimations based on the available data, blue numbers are calculated, and all other numbers are found in the private documentation of Prof. Rapoport and official data of the GDR Ministry of Health. IMR: infant mortality rate; NMR: neonatal mortality rate. *Two cases had no birthday documented.*

The index cards also contained a color coding, which denoted a premature birth (a red mark on top of the card) and a lethal malformation (yellow mark). The latter was used in the presented approach to evaluate the quality of the classification algorithm.

Digitalization of the cards was performed in two steps: first, a professional service provider for archiving paper-based patient records scanned the index cards and delivered high-resolution images of the front and back sides of the cards. In the second step, the cards were manually transcribed into an SQL database. Spelling and grammar, however, were copied exactly. Quality assurance was implemented by a three-stage release process (involving three independent transcriptionists). In the final step, all values of the structured data items were analyzed through A-Z analysis for implausible data.

The central idea of the technical approach presented here consists of integrating a terminology server into an SQL-based RDBMS (Relational Database Management System) and extending the SQL language itself by adding the ability to formulate semantic criteria within the query with free text. Hence, the approach comprises two components:

I) Semantic structuring and annotation of the RDBMS tables, and

II) Syntactic extension of standard SQL ("Ontology-SQL").

Item	Cards with that item [N]	Coverage [%]
Admission date	855	45.8
Admission hospital	1,866	99.9
Number of pregnancy advisories	1,690	90.5
Age in hours	1,646	88.1
Age of mother at birth	1,832	98.1
Apgar score	1,677	89.8
Date of birth	1,866	99.9
Birth position	1,048	56.1
Number of children	1,807	96.7
Date of death	1,864	99.8
District	1,807	96.7
Week of first pregnancy advisory	1,284	68.7
Gender	1,868	100.0
Number of current pregnancy	1,850	99.0
Judgment	1,868	100.0
Revision of judgment	1,868	100.0
Body length	1,620	86.7
Number of multiples	1,868	100.0
Number of previous deliveries	1,847	98.9
Pathoanatomical diagnosis	1,843	98.7
Transfer hospital	1,804	96.6
Week of gestation	1,370	73.3
Body weight	1,862	99.7

Table 1 Extracted items on index cards: the number of cards on which the item could be found and the respective coverage when compared to all 1,868 cards.

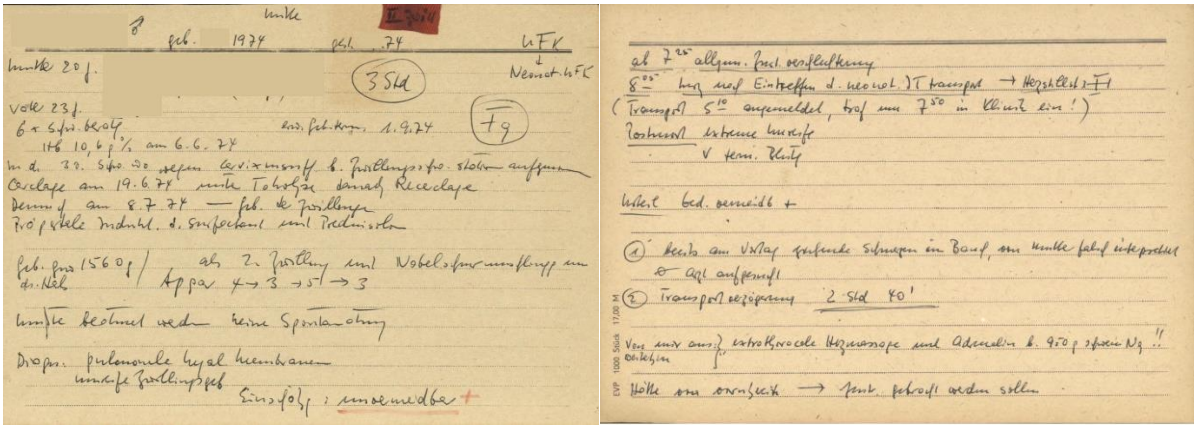

Figure 2 Original sample of a handwritten index card (anonymized). In preparation for each commission meeting on reducing infant mortality (IM), the chairman of that commission prepared such an index card (N=1,868). The index cards contain basic demographic data regarding the mother and child and free text regarding anamnesis, birth, postnatal treatment and course of death. Structured information could be retrieved for the values shown in Table 1. The index cards also contain a color coding system denoting a premature birth (a red mark on top of the card) and a (lethal) malformation (a yellow mark). Additionally, the index cards included the final judgment made by the commission if the case was avoidable, conditionally avoidable or not avoidable. In many cases, the index cards also contained a revision of the initial judgment including an explanatory statement.

Semantic structuring and annotation

The semantic annotation was stored in additional tables and the database schema was extended instead of changing existing tables. This approach offers the advantage that the data tables that store the annotations can be created in their own logical instance of an existing database server.

First, a single table was created for each source column that had to be annotated. The results are annotation tables that contain semantic representations of the content of the source columns. The precondition is that each table contained a unique primary key, which held true in practice. For each row of the given column, n rows were created in the annotation table that contained the semantic interpretations. These annotations are the specific concept identifiers in the selected ontology. The records are linked via the primary unique key. In addition to storing direct annotations, also the respective super classes were stored (derived from the "is a" hierarchy from the ontology) to enable a performance analysis. Figure 3 shows the designed structure, and Table 5 shows a sample data set.

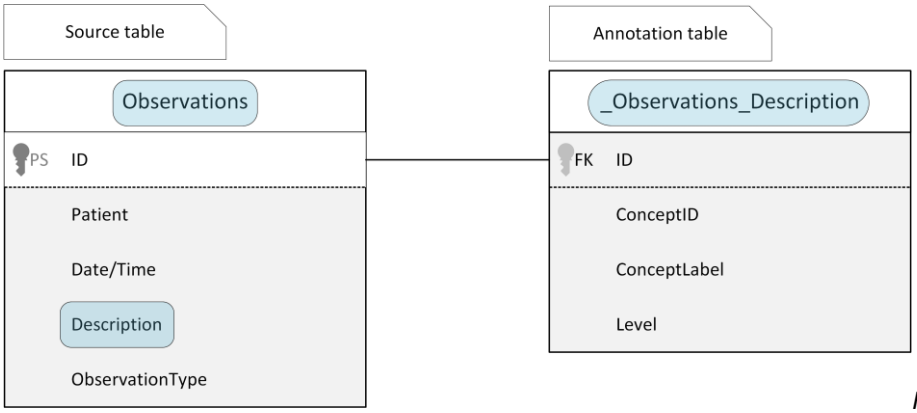

Figure 3 For each table ("aTableName") and each column ("aColumnName_x") enabled for semantic queries, an annotation table is created with the naming scheme "[_][aTableName][_][aColumnName_x]". The annotation table is linked to the source table via the unique primary key. Each row in the annotation table results in n rows in the annotation table, each holding the concept identifier (ConceptID) and a concept label (ConceptLabel) from the ontology and a level denoting the semantic distance (the super classes of the annotation concepts are also stored).

DiagnosisId	PatientId	CaseId	ICD-10	Description
1234124	4234324	234242	G44.0	cluster headache for two weeks

ID	ConceptId	ConceptLabel	Level
1234124	D0009F4	Bing-Horton syndrome	0
1234124	Z000002	two	0
1234124	GA000F8	week	0
1234124	GA00FE3	time unit	1
1234124	GA00FE3	measurement unit	2
1234124	GA0075D	number	1
1234124	D0009F1	migraine	1
1234124	M00F2C	unilateral continuous headache	2
1234124	D0011A0	headache syndrome	2
1234124	F00231C	chronic headache	3
1234124	F00FCE	cephalea	4

Table 5 Sample of the resulting data structure. The upper table represents a diagnosis table that has a primary unique key (DiagnosisId), a patient and case key, an ICD code and a description of the diagnosis. The diagnosis "cluster headache for two weeks" is annotated with "D0009F4 Bing-Horton syndrome Z000002 two GA000F8 week" and subsequently stored in the annotation table (lower table). The semantic

distance is "0" here because the concepts directly represent the narrative description of the diagnosis. Additionally, the parents of all concepts found are stored in the annotation table with the same diagnosis id. Therefore, "cephalea" is a parent of the 4th degree of "cluster headache". The parent concepts are retrieved with a function call from the terminology server that returns the taxonomy of a given concept. The tables are linked with the relation "DiagnosisId → Id".

The approach of using a generic table with metainformation for the description of the "source table" was discarded in favor of performance or respective costs. The annotation table schema is fairly simple with four columns (one for the key, one for the concept identifier, one for the concept label and one for the semantic distance (level)), and thus, modern RDBMS are expected to handle such tables extremely well.

The annotation of free text is accomplished by integrating a CTS2 (common terminology services)-compatible terminology server. The terminology server used here includes a complete NLP pipeline based on Gate and Jape in addition to numerous supporting algorithms, such as an extensive, discipline-specific list of abbreviations, collocation-based disambiguation, a typing error corrector, which can break up compounds and correct them step by step, and a function for German language-optimized word stemming (for further information, see Addendum 1).

As stated above, a central feature of the NBC is the exclusion of cases with malformations since these cases have the lowest potential of avoidance. In particular, malformations that are incompatible with life must be distinguished from those that only lead to treatable limitations. The goal of mapping the free-text portions of the index cards to the Wingert nomenclature was to use the underlying ontology to then detect such lethal malformations directly from the free text. For that reason, a semantic definition of what was assumed to be a lethal malformation was created, containing 16 categories (see Table 4). Then a semantic query based on this definition was formulated, which filtered all affected cases into class "I" of the NBC. All other cases were classified by simple SQL queries into classes "VIII" to "XII". Finally, all remaining cases were excluded from this analysis.

Semantic category
Developmental anomaly
Genetic disorder
Growth change and maturation change
Potter's syndrome
Hypoplastic left ventricle syndrome
Atresia
Heart defect
Caudal regression syndrome
Syndrome
Congenital disease
Neoplasm
Diaphragm
Ectopy
Dystrophy
Congenital deficiency/Congenital aplasia
Stenosis

Table 4 Semantic categories from the ontology that were used for the definition of a "lethal malformation".

Ontology-SQL syntax

Additionally, an extension to standard SQL was developed that enables the use of free-text and semantic relations within such a query. Expressions formulated through this extension can be transformed into standard SQL syntax using a preprocessor. The resulting query contains only standard SQL and thus can be directly passed on to the database server engine.

A basic O-SQL expression consists of a free-text part that is surrounded by square brackets and followed by a table and column name in which the free-text should be searched for. The latter is surrounded by round brackets. So the simpliest expression looks like this:

```
[free-text](tablename.columnname)
```

Furthermore, the table and column name can be a comma separated list. Since the free-text is mapped to the ontology, a semantic role (or in short: relation) can be given that will be applied to the expression. That relation is a keyword written before the O-SQL expression:

```
relation[free-text](tablename.columnname)
```

The default value for the relation is "isA", which would query all concepts subsumed under the concept described in the free text.

The generic relation “context” can be further specified by a modifier to extent the standard relations like “is a” and “part of” to relations like “has indication”. This context modifier is a free-text written in curly braces between the relation and the free-text query:

```
context{modifier}[free-text](tablename.columnname)
```

The approach of a context modifier was chosen to allow a generic, ontology independent syntax of O-SQL expressions.

As shown in Figure 4, the attributes of the semantic roles can be passed on through the isA-hierarchy, which allows inheritance of these attributes up to a specific depth. The inheritance depth is specified by a number separated from the context modifier by a colon. Finally, a leading prefix can be added to include the “is a” relation to the given relation.

Consequently, the complete syntax for O-SQL expressions is as follows:

```
[prefix][relation][{modifier}][:depth][[query]](table.column,…)
```

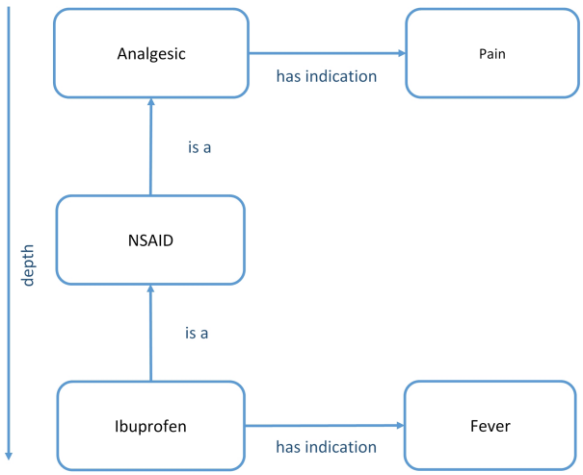

Figure 4 Inheritance via depth parameter: in this sample, indications of the concept “Analgesic” subsum indications of “Ibuprofen”; thus, a query for indications of “Analgesic” will find “Pain” and “Fever” if the depth is set to ≥ 2.

Conversation

An efficient conversion of O-SQL into standard SQL is crucial, as this step mainly affects the runtime of given queries. First, O-SQL expressions are extracted via regular expressions and are then converted into SQL subqueries.

The IN operator was used for generating the subqueries because the source table holds a primary unique key and the subclause can created as a simple enumeration of the found annotation rows. Modern RDBMS use a so-called "Clustered Index Scan" to process such queries efficiently.

For each O-SQL statement, a subclause was created by first mapping the free-text part of the O-SQL expression onto the terminology and then querying the annotation table for the found concept identifiers. From the results of these queries, the values of the column containing the key were used to build the subclause. Then, the O-SQL expression was replaced by that subclause. All other parts of the SQL query remained untouched. Therefore, all logical operators and language components – especially parentheses and the use of the logical operator NOT – function as usual so that structured discrete information can be directly linked to semantic information.

Given an O-SQL query such as the following:

```
SELECT * FROM table_name WHERE <O-SQL expression>
```

The free-text part of the O-SQL expression was transformed with the help of the terminology server to form concept identifiers and look them up in the annotation table (see Table 5):

```
SELECT ID FROM annotation_table WHERE conceptID = <concept identifier>
```

Then all the results were joined together and replaced the O-SQL expression with an "IN" subquery:

```
SELECT * FROM table_name WHERE ID in (…)
```

If a query contains multiple O-SQL expressions, each expression is converted separately. Therefore, one can use all standard SQL operators to combine O-SQL expressions. For an example, see Addendum 2.

Figure 5 illustrates the entire pipeline via which semantic queries can directly be integrated into standard SQL. The pipeline also provides feedback on the terminology concepts actually used to avoid undetected errors. Such errors can happen if an abbreviation is not known and is therefore misinterpreted.

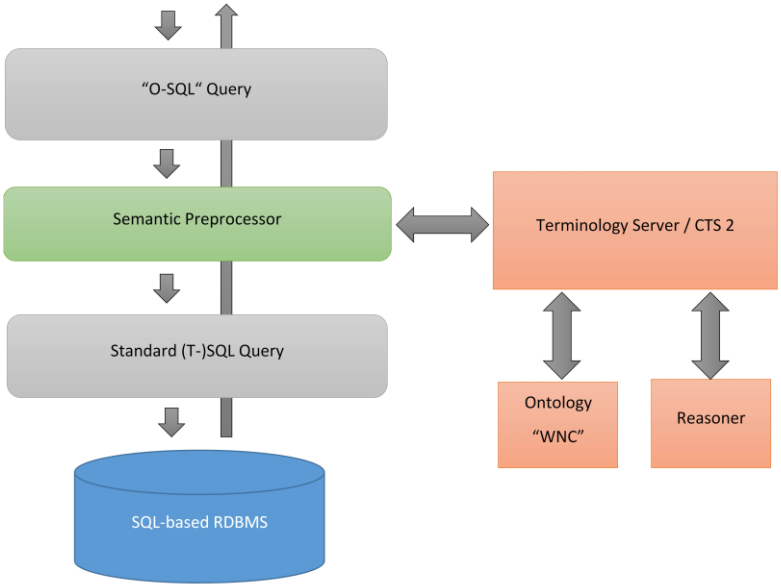

Figure 5 Schematic overview of transforming an "O-SQL" query into a standard SQL query using a standardized terminology server

Results

From the calculations and estimations shown in Table 1, it can be concluded that Rapoport's dataset represented well over 90% of all records of neonatal deaths that occurred in East Berlin during that period. The IMR is usually rounded up to one decimal place, so the maximum error must be below 0.05‰. This error results in a deviation of less than one individual when calculating the absolute numbers.

Annotation accuracy

The annotation algorithm of the terminology server could recognize abbreviations and had a spell check function optimized for the German language and a module for disambiguation of semantic interpretations. In particular, the spell check function was urgently needed since German allows the construction of so-called "compound nouns" and since many of these compound nouns had been shortened into their subwords and then reconstructed again. A typical example was the German term for pregnancy counselling, which is "Schwangerenberatung"; this term was often written as "Schw.brtg" in many variants.

A standardized procedure that included a manual annotation and an inter-annotator agreement to calculate precision and recall was not established for two reasons: I) it was especially interesting whether the annotation was correct and not whether the annotation was optimal (focus on precision and not on recall) and II) the architecture is independent of the terminology server and uses a standard interface for the integration. Thus, any CTS compliant terminology server can be easily used.

The automated annotation from 10% (N = 187) of the cards, namely, the section that contained the postmortem diagnoses was manually analyzed. 423 diagnoses were found, 304 of which were unique. Each annotation was classified as "completely correct", "partly correct" or "incorrect". The category "partly correct" contained items that could not be mapped better due to missing precoordinated concepts in the ontology, and thus, the expert partly disagreed with the interpretation (e.g., "aspiration of infected amniotic fluid" was annotated as "aspiration of amniotic fluid" and "infection"). The category "incorrect" contained all items that were incorrect. In total, 301 of the 304 (99.0%) items were correctly mapped or could not be mapped better. Twenty-three of these items were in the category "partly correct" and could be fixed by adding new (precoordinated) concepts to the terminology. The remaining three were "incorrect".

Furthermore, annotation quality was evaluated using a single concept from the terminology, namely, the anatomy concept "tentorium". All 1,868 cases were analyzed with a total of 9,080 postmortem descriptions and noted whether the concept "tentorium" was present. 50 different spellings were found in 77 cases. For all of these cases, it was verified whether the mapping algorithm used the correct concept. From these comparisons, a sensitivity of 97.62% (CI 95%: 91.66% - 99.71%) and a specificity of 100.00% (CI 95%: 98.87% - 100.00%) was found for the automated annotation.

Full-scope examination

Using the yellow color marking of the index cards as the gold standard for the class "malformation", a full-scope examination of all 1,868 index cards of the class "malformation" was carried out, thus adopting the yellow mark as the gold standard. The semantic query for malformations showed a sensitivity of 92.1% (95% CI: 89.5-94.3%) and a specificity of 96.3% (95% CI: 95.2-97.2%), resulting in an F1 score of 0.91. The positive predictive value was 90.6% (95% CI: 87.8-92.9%); thus, cases with lethal malformations could be correctly predicted in over 90% of the cases.

Finally, the semantic query for "malformation" was compared with further queries to verify the capabilities: (I) all of the of the index cards were manually reviewed to see whether the patient had a "vitium cordis", (II) 25% of the index cards were manually reviewed for the presence of a sign of "respiratory disorder". (see Table 6).

Query	**"Vitium cordis"**	**"Respiration disorder"**	**"Lethal malformation"**
Number of query concepts	1	5	16
Sample size (Percentage of all cards)	1868 (100%)	467 (25%)	1868 (100%)
Sensitivity	93.22% (CI 95%: 89.22 - 96.08%)	97.76% (CI 95%: 95.79 - 98.97%)	92.13% (CI 95%: 89.48 - 94.29%)
Specificity	99.94% (CI 95%: 99.66 - 100.00%)	95.45% (CI 95%: 87.29 -99.05%)	96.30% (CI 95%: 95.16 - 97.24%)
Positive predictive value	99.55% (CI 95%: 96.87 -99.94%)	99.24% (CI 95%: 97.75 -99.75%)	90.57% (CI 95%: 87.75 - 92.92%)
Disease prevalence	12.57% (CI 95%: 11.10 - 14.15%)	85.90% (CI 95%: 82.41 -88.92%)	27.80% (CI 95%: 25.78 - 29.89%)
F score	0.96	0.98	0.91

Table 6 Statistical results of semantic queries of different complexity (complexity is represented by number of query concepts). All statistics are calculated with MedCalc® [29]. Raw data can be found in table A2 in the addendum.

Judgment of avoidance

The proportion of avoidable deaths increased from 1.9% in 1974 to 21.6% in 1988; additionally, the proportion of conditionally avoidable deaths increased from 16.7% in 1974 to 38.6% in 1988 (see Table). Both increases were statistically significant, with $p<0.0001$ for avoidable deaths and $p=0.0006$ for conditionally avoidable deaths. Trends of increasing proportion of avoidable deaths in this series are probably related to the commissions' criticism and objectivity rather than worsening standards of health care provision in East Germany.

For the judgment of avoidance, the decisions of the commission in the gold standard were analyzed and the following observations were made:

- 1384 cases were unavoidable
- 396 cases were conditionally avoidable
- 81 cases were fully avoidable
- 7 cases had no judgment

In some cases the commission revised its decision. This happened if a) new facts became known or b) an appeal was lodged against the decision. Analysis of this revision process revealed the following observations:

- 1666 cases had no revision
- 143 cases were conditionally avoidable
- 32 cases were fully avoidable
- 27 cases were unavoidable

In sum, 523 (31.0%) cases were conditionally or fully avoidable, including the revised cases. In the next step, 1528 (90.6%) cases were classified with the NBC, which included 483 (92.3%) avoidable cases (see Table).

As expected, classes “X” and “Xa” showed the highest potential of avoidance by far, as each of these classes included more than 50% of the avoidable cases within that class (see Table). Therefore, a measure was found that effectively supports an efficient single-case analysis. Additionally, the NBC was extended by two subclasses to gather significantly better results in predicting the avoidance of neonatal deaths.

Published queries

To judge the generalizability of the approach, published queries were analyzed and examined whether it was possible to express them in O-SQL and whether that expression was more compact and easier to create and understand. The following section demonstrates that these expectations held true for all examined samples.

In Lieberman et al., the request for patients with “coronary artery disease” resulted in the following partial expression [26]:

```
concept_id in (
   select concept_id from snomed_map
   where snmd_cncpt = 8957000 or snmd_cncpt in (
   select snmd_cncpt1 from snmd_relationship
   connect by snmd_cncpt2 = prior snmd_cncpt1 and
   relationship_type = 116680003
   start with snmd_cncpt2 = 8957000 and
   relationship_type = 116680003)
)
```

Its complexity is basically derived from the need to use nested subqueries to represent relationships within the ontology. The IDs are from SNOMED CT:

```
8957000 = Coronary artery disease (disorder)
116680003 = Is a (attribute)
```

This query can be represented in O-SQL by the following compact expression using a common abbreviation:

```
[chd](diagnosis_column)
```

The next sample shows that the complexity of SPARQL can also be greatly reduced. In [27], the filter criterion “Find all patients having a side effect of Prandin after administration” is defined. Pathak et al. transformed this criterion into the following SPARQL query (abbreviated):

```
SELECT DISTINCT ?MCLSS_KEY {
   { SERVICE <http://www4.wiwiss.fu-berlin.de/sider/sparql>
      { SELECT ?mySideEffect ?mySideEffectLabel WHERE {
         ?x rdf:type sider:drugs ;
         rdfs:label "Prandin" ;
         sider:sideEffect ?mySideEffect .
         ?mySideEffect rdfs:label ?mySideEffectLabel .
   }}}
   { SELECT DISTINCT ?rxnormCode WHERE {
         SERVICE <http://link.informatics.stonybrook.edu/sparql/> {
            ?rxAUIUrl rxnorm:hasRXCUI ?rxCUIUrl ;
            rdfs:label ?rxnormLabel .
            ?rxCUIUrl rxnorm:RXCUI ?rxnormCode .
            FILTER(regex(str(?rxnormLabel), "Prandin", "i")) .
   }}}
   { SELECT DISTINCT ?MCLSS_KEY WHERE {
         SERVICE <http://edison.mayo.edu/lss1p#> {
            ?icd9Url semr:dx_code ?icd9Code ;
            semr:dx_abbrev_desc ?diagnosis .
            FILTER(regex(str(?diagnosis),
```

```
            str(?mySideEffectLabel), "i")) .
            ?patientUrl semr:whkey ?MCLSS_KEY ;
            semr:diagnosis ?diagnosisCode ;
            semr:concept_id ?rxnormCode .
            FILTER(regex(str(?icd9Code),
            str(?diagnosisCode), "i")) .
}}}}
```

This query requires a deep understanding of SPARQL and the structure of external knowledge bases. In addition, this query requires that local diagnoses are encoded in ICD-9, as the medication database uses this classification to structure information on "side effects".

In contrast, the query can be drastically reduced to the following form using the presented approach:

```
select * from tableMed, tableDiag where
   tableMed.CID = tableDiag.CID and
   +partOf[Prandin](tableMed.Drug) and
   hasContext{side effect}[repaglinide](tableDiag.Diag)
```

The first partial expression searches the column "Drug" in the table "tableMed" for all occurrences of "Prandin" itself (note the "+") and for all concepts containing "Prandin". In doing so, the agents of Prandin are also found, and possible generic drugs are included. The second partial expression simply scans the ontology for a "side effect" of the agent and uses these results to search the column "Diag" in the table "tableDiag". Here, it must be ensured that specifications are also found in each case. Therefore, Prandin has "hypoglycemia" as a side effect and with the assistance of the ontology, the query will also identify patients in which "hyperinsulinism" is recorded because "hyperinsulinism" is a form of "hypoglycemia".

The last example was published by Leroux and Lefort, who queried "anti-diabetic drugs, such as Metformin," therefore defining the following request (abbreviated) [28]:

```
SELECT count (distinct ?subject) as ?count ?mp_med WHERE {
   SERVICE <http://wifo5-04.informatik.uni-mannheim.de/drugbank/sparql>
   {
      ?s drugbank:genericName "Metformin" .
      ?s drugbank:drugCategory ?category .
      ?drug drugbank:drugCategory ?category .
   }
   { SELECT distinct ?drug ?med ?subject ?mp_med WHERE {
      GRAPH <http://localhost/dataset/aibl/lcdc/clinical> {
      ?obs a lcdcobs:Observation .
      ?obs cm:medicinalProduct ?cm_mp .
      ?cm_mp skos:exactMatch ?drug .
      ?cm_mp amt:synonym ?mp_med .
      ?obs lcdccore:subject ?subject .
}}}}
```

Additionally, in this case, the query can be drastically reduced when using O-SQL:

```
select * from tableMed where hasContext{indication}:5[diabetes](Drug)
```

The indications for all children are determined recursively from "diabetes" up to the specified depth of "5" (semantic distance).

Discussion

Technology discussion

An approach that enables free-text queries stored in standard SQL-based RDBMS was developed. Therefore, the standard SQL syntax was extended and an architecture was created that allows to integrate a terminology server into existing databases.

The disadvantage of Zheng's approach was overcome[21], which needed an already annotated database, by integrating a terminology server into the analysis pipeline. In addition, the user is not required to have any knowledge about how the terminology works since the data annotation and free-text annotation of the query expressions use the same NLP-based terminology server.

The annotation results outperformed the approaches presented by Allones et al. [12] and Shah et al. [29] and compared very well to the ones presented by Topfer et al. [13], who obtained slightly better results, and Elkin et al. [11], whose results were slightly inferior to ours. This comparison of final metrics shows that the employed NLP pipeline and annotation algorithm deliver state-of-the-art results and provide an objective evaluation of the query results. Since the results come from a relatively small set of data, they can only be generalised if the terminology is complete and homogeneous, which needs to be validated. Maintaining terminology and ontology was not a specific topic of this research, as a standardized terminology server was used that can provide different terminologies and is easily interchangeable.

The slight loss of sensitivity when querying for heart defects was traced back to sixteen cases that were false negatives. The reasons for these false negatives were unrecognized abbreviations, missing precoordinated terminology concepts, missing ontology links and in one case, selection of an incorrect term-to-concept combination by the annotation algorithm. However, the results compare very well to those of Pakhomov et al. [30].

The imperfect specificity for respiration disorders can be explained by misinterpreted cases of "intrauterine hypoxia" and the reduced sensitivity was mainly due to cases with a "hyaline membrane" that were not correctly classified.

When comparing the complexity of SPARQL queries to O-SQL queries, it became clear that a considerable advantage was not only the ability to use free text but also the commonly available knowledge of the SQL syntax. It is assumed that committed physicians with existing knowledge of SQL can be trained to use O-SQL without issue. In particular, clinicians' calls for "secondary use" have caused big software companies to open up their databases to clients. The use of routine systems, however, bears certain limits, as the data must not be changed, and the stability of the database models is not provided per se. The first limit is completely circumvented by the framework presented here, as all annotations are stored in their own tables and the original tables remain untouched. Changes in the database schema with respect to the data model can be represented quite easily by re-generating the annotations. Also, when evaluating the differences, it is important to note that SPARQL is designed to query RDF data and SQL is designed to query relational data. Thus, the advantages of both languages directly reflect the data models on which they work.

Clinical discussion

Even though Rapoport's dataset is rather old, it is unique in terms of completeness and the large number of included cases [6]. In addition, all the deceased babies had a pathological-anatomical diagnosis based on an autopsy to determine the correct cause of death. Except for the first and the last year, the number of records in this data set represents on average over 90% of all reported neonatal deaths in that time In East Berlin. Thus, the number of cases that were used in this study can be assumed to be rather complete. The first and the last year were excluded from the completeness analysis because in 1973, the commission was not fully established yet, and in the second half of 1989, political issues arose; these situations resulted in incomplete datasets. The completeness of the information on the cards itself is very high. The rather low coverage of 88.1% for the item "age in hours" could be overcome by calculating this value from the items

“date of birth” and “date of death”, resulting in final coverage of 99.8% for age. The week of gestation was present in only 73.3% of all cases, which explains most of the 9.4% of cases that could not be classified into the NBC.

The classification of cases with the NBC was simple for the classes that are defined by structured and “crisp” items (“IX” to “XII”) but difficult for class “I”. Here, it is necessary to consider to what extent the diagnoses subsumed under the category of “malformation” ultimately led to death since this influenced the original judgment of avoidance. This situation becomes particularly clear in the case of congenital tumors. The systematic yellow marking in the case of teratomas showed that these tumors had been generally recognized as lethal malformations, whereas some neuroblastomas had not been included in this category. The overall quality was still excellent, as over 90% of the individuals could be correctly classified into NBC class “I”. Additionally, seven cards were found that had a yellow mark, but no malformation was found after autopsy.

The advantage of using a semantic query becomes striking clear because a definition, as shown in Table , can very easily be adapted to reflect scientific progress or specific local situations. In general, filtering of malformations makes sense since the focus is on lethal malformations, which are, by definition, not avoidable (or at least the potential of avoidance is very low). Additionally, one can see that such semantic definitions easily allow the inclusion of groups of patients who do not necessarily have a malformation but should be filtered as such and be excluded from the analysis of avoidance.

The use of the diagnosis malformation seemed to make sense to use it as a kind of index marker in Rapoport's dataset. Today, 30 years later, changed conditions exist for prenatal and postnatal diagnosis and effective therapy, which would have to be taken into account or adapted to current clinical databases when applying the developed methodology.

The NBC uses the Apgar score taken at 5 minutes after delivery. This score has been subject of some criticism lately, but a recent study [31] proved that this score still is a valid predictor of neonatal mortality. Moreover, the 1 minute score was recently proven to be highly predictive [32].

Extension of this analysis to intrapartum deaths and antenatal cases (NBC classes “I” to “VII”) is fairly easy, as the main challenge is the detection of lethal malformations, and this challenge has been solved. The parameters that define classes “II” to “VII” are usually available in a structured format or can be calculated easily (“growth retardation”); thus, complete classification can be implemented without further problems. This approach allows the potential of avoidance to be calculated and compared for further groups of cases.

The avoidance of transportation of the newborn is one of the key factors for a low NMR [33], scores that can be used already exist [33]. However, such methods should be combined with further methods, such as single-case analysis, which could detect problems even in rare transports.

Perinatal surveys seem to provide limited knowledge gain [35]; in the end, they are a highly useful tool for quality assurance but not for quality improvement [36]. Gmyrek et al. showed the pitfalls of these methods in [3, 37] and stated: “In Germany, however, recording the quality of results with the help of the neonatal survey is reaching its limits. While the sub-documentation on linking quality assurance to accounting data is decreasing, the problem of under-adjustment for allocation-controlling risk factors can only be solved by sufficiently small strata defined by birth weight or gestational age”. Thus, statistical measures alone are clearly not sufficient anymore.

Further questions, currently under investigation with the developed approach, are the monitoring of pain medication (in particular NSAID) with simultaneous administration of PPI with patients with osteoporosis or osteopenia [38] and the application of the „Frailty Index“ on routine data, like discharge letters [39].

Conclusion

Many of the difficulties in using ontologies for the semantic analysis of free text that were described in the introduction of this paper were overcome by the approach presented here. The overall concept of integrating an NLP-based terminology server into SQL syntax has been proven to be extremely viable.

In particular, the high complexity and diversity of the German language were processed at a quality that meets the requirements of medicine. The terminology server used here is multilingual, so the O-SQL syntax simply had to be extended to specify the language used. With this approach, it is now possible to compare databases on an international level, and country-specific annotations with only one uniform query language, such as German or English, could be implemented.

The ontology-SQL syntax will be extended to allow for nested expressions. For this purpose, the extent to which these expressions are relevant and whether they cannot perhaps be represented by concatenated expressions will have to be evaluated. One example would be a query for “patients showing symptoms of diseases that can be treated with specific agents”.

From a medical point of view, queries for rare diseases should be investigated. Here, selection will have to be carried out via O-SQL and will be followed by concrete case-by-case examinations.

In addition, the ontologies themselves (especially the standard ontologies) will become increasingly extensive as more and more “omics data” are represented. Genetic information that was unknown at the time of data collection can thus be considered in semantic queries. Deriving quality factors from historic data is of growing interest because it enables a comparison of historic procedures and treatments to the current medical state-of-the-art.

The approach presented here is independent of a specific ontology, on the contrary it allows access to any number of ontologies. However, also controlled vocabularies can be made applicable through this system. Therefore, physicians can access and explore data with specially developed ontologies that go beyond the spectrum of standard ontologies. This removes one of the typical limitations of standard ontologies, which cover a broad spectrum of knowledge but usually have a limited depth. This should significantly increase the acceptance of the system. The fact that routine data from many sources – as long as it is stored in SQL based databases - can be immediately used at an ontology-driven level without further transformation or integration demonstrates that the presented work makes an important contribution to translational research of routine data.

An automatic and efficient method to select cases of potentially avoidable neonatal deaths directly from routine data was developed. This method allows strategic implementation of single-case analysis into the process of quality management based on QIs. It is beliefed that this supports the process of further reduction of NMRs and IMRs. Even for hospitals that are already organizing regular perinatal conferences, this method can be highly supportive in that process.

This method can be even applied to free text narrative notes of physicians and determine within a very short amount of time if this particular text describes a potentially avoidable death.

Hospitals are encouraged to establish this method and propose to start a discussion regarding whether incentives or obligation would be the right way to increase the acceptance of such a method.

To ensure freedom of choice for physicians, it is strongly recommended always establishing an interdisciplinary commission for such single-case analysis. Not only obstetricians, neonatologists and primary care takers but also independent experts, such as pediatric pathologists, should ideally be involved.

Finally, it is assumed that by systematic application of this approach, a further reduction in the NMR and IMR can be expected.

Abbreviations

AWMF = Working Group of the Scientific Medical Societies of Germany

CI = Confidence Interval

CTS = Common Terminology Services

GMDS = Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie

ICD = International Classification of Diseases

ICD-10 = International Classification of Diseases Version 10

IMR = Infant mortality rate

NBC = Nordic-Baltic classification

NLP = Natural Language Processing

NMR = Neonatal mortality rate

NSAID = nonsteroidal anti-inflammatory drug

OLAP = Online Analytical Processing

OPS = Operationen- und Prozedurenschlüssel

QI = Quality indicator

PMR = Perinatal mortality rate

RDBMS = Relational Database Management System

RDF = Resource Description Framework

SPARQL = SPARQL Protocol and RDF Query Language

SQL = Structured Query Language

TNM = TNM Classification of Malignant Tumors (tumor, lymph node, metastasis)

OWL = Web Ontology Language

References

[1] Wauer RR: Säuglings- und Kindersterblichkeit in Deutschland und Berlin. Unterschiede in Ost und West. In A. Holzgreve & G.v.Cossel (Hrsg) Geschichte der Berliner Krankenhäuser. Medizinisch Wissenschaftliche Verlagsgesellschaft Berlin, 2018, S. 97-128

[2] Bührer C, Perinatalerhebung Berlin 2013, Geburtshilfe Frauenheilkd 2016; 76 - A4; DOI: 10.1055/s-0036-1571401

[3] Gmyrek D, Koch R, Vogtmann C, Kaiser A, Friedrich A, Warum Risikoadjustierung von Qualitätsmerkmalen?, demonstriert am Qualitätskriterium neonatale Spätinfektion. Sitzungsberichte der Leibniz-Sozietät der Wissenschaften zu Berlin 115(2013), 85–94

[4] Masson VL, Farquhar CM, Sadler LC: Validation of local review for the identification of contributory factors and potentially avoidable perinatal deaths., Aust N Z J Obstet Gynaecol. 2016 Jun;56(3):282-8. doi: 10.1111/ajo.12454.

[5] Merali HS, Lipsitz S, Hevelone N, Gawande AA, Lashoher A, Agrawal P, Spector J: Audit-identified avoidable factors in maternal and perinatal deaths in low resource settings: a systematic review., BMC Pregnancy and Childbirth 2014, 14:280 . doi: 10.1186/1471-2393-14-280.

[6] Ministerium für Gesundheitswesen der DDR 1970, Richtlinie für die Tätigkeit der Fachkommissionen zur Senkung der Säuglings- und Kindersterblichkeit in den Bezirken und Kreisen In:Ockel, E.: Gesundheitsschutz für Mutter und Kind. Beitrag zur Geschichte des Gesundheitswesens der Deutschen Demokratischen Republik, Berlin; trafo Verlagsgruppe, Berlin 1995

[7] Grube E, Lorenz K: Gezielte Bekämpfung der Säuglingssterblichkeit durch Analyse des einzelnen Säuglingstodesfalles. Das Deutsche Gesundheitswesen. 1958, 45:1450-1454

[8] Gaber E, Heft 52 - Sterblichkeit, Todesursachen und regionale Unterschiede, Robert Koch-Institut, Berlin, April 2011

[9] Friedman C, Shagina L, Lussier Y, Hripcsak G, Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004, 11(5):392-402.

[10] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13.

[11] Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, Asatryan AX, Tokars JI, Rosenbloom ST, Brown SH, NLP-based Identification of Pneumonia Cases from Free-Text Radiological Reports, AMIA Annu Symp Proc. 2008 Nov 6:172-6.

[12] Allones JL Martinez D, Taboada M, Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology. J Med Syst. 2014 Oct;38(10):134. doi: 10.1007/s10916-014-0134-x. Epub 2014 Sep 2.

[13] Toepfer M, Corovic H, Fette G, Kluegl P, Stoerk S, Puppe F, Fine-grained information extraction from German transthoracic echocardiography reports. BMC Med Inform Decis Mak. 2015 Nov 12;15:91. doi: 10.1186/s12911-015-0215-x.

[14] Hogarth MA, Gerz M, Gorin FA, Terminology Query Language: A Server Interface for Concept-Oriented Terminology Systems, Proc AMIA Symp. (2000):349-53.

[15] Travillian RS, Diatchkab K, Judgec TK, Wilamowskad K, Shapiro LG, An ontology-based comparative anatomy information system, Artificial Intelligence in Medicine 51 (2011) 1–15

[16] Riva A, Bellazzi R, Lanzola G, Stefanelli M, A development environment for knowledge-based medical applications on the world-wide web, Artif Intell Med 14 (1998) 279–293

[17] Bichindaritz I, Mémoire: A framework for semantic interoperability of case-based reasoning systems in biology and medicine, Artificial Intelligence in Medicine (2006) 36, 177—192

[18] Mabotuwana T, Warren J, An ontology-based approach to enhance querying capabilities of general practice medicine for better management of hypertension, Artificial Intelligence in Medicine (2009) 47, 87—103

[19] Kamaa AA, Choqueta R, Melsb G, Daniela C, Charleta J, Jaulent MC, An Ontological Approach for the Exploitation of Clinical Data, Stud Health Technol Inform. (2013);192:142-6.

[20] Epstein, R, Jacques, P, Stockin M, Rothman B, Ehrenfeld J, Denny J, Automated identification of drug and food allergies entered using non-standard terminology, J Am Med Inform Assoc (2013);20:962–968.

[21] Zheng S, Wang F, Lu J, Enabling Ontology Based Semantic Queries in Biomedical Database Systems, Int J Semant Comput. (2014) March ; 8(1): 67–83. doi:10.1142/S1793351X14500032.

[22] Bache R, Miles S, Taweel A, An adaptable architecture for patient cohort identification from diverse data sources, Am Med Inform Assoc (2013);20:e327–e333. doi:10.1136/amiajnl-2013-001858

[23] Borch-Christensen H, Langhoff-Roos J, Larsen S, Lindberg B, Wennergren M, The Nordic/Baltic perinatal death classification, Acta Obstet Gynecol Scand Suppl. 1997;164:40-2.

[24] Wauer RR.: Die Entwicklung der Neonatologie als Teil der Perinatologie an der Universitätsfrauenklinik der Charité in Berlin-Mitte. In: Geschichte der Berliner Universitäts-Frauenkliniken: Strukturen, Personen und Ereignisse in und außerhalb der Charité (Hrsg. David M, Ebert A. D.). Walter de Gruyter GmbH & Co. KG, Berlin 2009, S. 88 – 130

[25] Ingeborg Rapoport, [cited 2017 Dec.]; Available from: https://de.wikipedia.org/wiki/Ingeborg_Rapoport

[26] Lieberman MI, Ricciardi TN, Masarie FE, Spackman KA, The use of SNOMED CT simplifies querying of a clinical data warehouse., AMIA Annu Symp Proc. (2003):910.

[27] Pathak J, Kiefer RC, Chute CG, Using semantic web technologies for cohort identification from electronic health records for clinical research., AMIA Jt Summits Transl Sci Proc. (2012);2012:10-9. Epub 2012 Mar 19.

[28] Leroux H, Lefort L, Semantic enrichment of longitudinal clinical study data using the CDISC standards and the semantic statistics vocabularies, Journal of Biomedical Semantics (2015) 6:16 DOI 10.1186/s13326-015-0012-6

[29] Shah N H, Bhatia N, Jonquet C, Rubin D, Chiang A P, Musen M A, Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics. 2009; 10(Suppl 9): S14.

[30] Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care (2007):281-8.

[31] Cnattingius, S., Norman, M., Granath, F., Petersson, G., Stephansson, O. and Frisell, T., Apgar Score Components at 5 Minutes: Risks and Prediction of Neonatal Mortality. Paediatr. Perinat. Epidemiol 2017 Jul;31(4):328-337. doi: 10.1111/ppe.12360.

[32] Vogtmann C, Koch R, Gmyrek D, Kaiser A, Friedrich A: Risk-adjusted intraventricular hemorrhage rates in very premature infants—towards quality assur ance between neonatal units. Dtsch Arztebl Int 2012; 109(31–32): 527–33. DOI: 10.3238/arztebl.2012.0527

[33] Harris BA, Wirtschafter DD, Huddleston JF, Perlis HW: In utero versus neonatal transportation of high-risk perinates: a comparison., Obstetrics & Gynecology, 1981, Vol. 57, No. 4, 496-9

[34] Broughton SJ, Berry A, Jacobe S, Cheeseman P, Tarnow-Mordi WO, Greenough A: The Mortality Index for Neonatal Transportation Score: A New Mortality Prediction Model for Retrieved Neonates, Pediatrics 2004;114;e424, DOI: 10.1542/peds.2003-0960-L

[35] Selbmann HK, Münchner Perinatalstudie 1975-77, Deutscher Arzte-Verlag,1980, ISBN-10: 3769180135

[36] Bührer C, Perinatal- und Neonatalerhebung Berlin, [cited 2017 Dec.]; Available from: http://www.ggg-b.de/_download/unprotected/buehrer_c_perinatal_neonatalerhebung_berlin.pdf

[37] Rüdiger M, Paradigmenwechsel in der Neonatologie, Sitzungsberichte der Leibniz-Sozietät der Wissenschaften zu Berlin, 115(2013), 95–110

[38] Maes ML, Fixen DR, Linnebur SA. Adverse effects of proton-pump inhibitor use in older adults: a review of the evidence. Ther Adv Drug Saf 2017;8(9):273-297

[39] Clegg A, Bates C, Young J, Ryan R, Nichols L, Teale EA, Mohammed MA, Parry J, Marshall T, Development and validation of an electronic frailty index using routine primary care electronic health record data, Age and Ageing, Volume 47, Issue 2, March 2018, Page 319, https://doi.org/10.1093/ageing/afx001

Addendum 1

The origins of the here used mapping algorithm are going back to a paper by Wingert, who in 1987 described how annotation of ICD classifiation texts can be established using SNOMED concepts [A1]. However, as this approach was improved by modern NLP elements, the results are refered to and compared with results from mainly recently published papers, ensuring state of the art comparisons. The mapping basically matches a number of word stems from the input text to the word stems of the terminology concept. This process is optimized towards minimization of the number of such concepts. This way, it is ensured, that the concepts with the highest information density are used. Unlike stated by Allones et al [A2] who said, that there exist no mapping algorithms that are capable of using synonyms but the here used algorithm uses indeed synonyms. This is a crucial must, as it allows for compact modelling of the terminology:

Let concept C_1 be labeled by term T_1 (“heart infarction”) which consists of the sub-terms ST_1 (“heart”) and ST_2 (“infarction”)

Let concept C_2 be labeled by the terms ST_1 (“heart”) and a synonymous term ST_3 (“cardiac”)

Then concept C_1 has to be found by term ST_3 plus ST_2 (“cardiac infarction”), even so, it’s not labeled with these terms

Thus, the mapping algorithm first matches all input terms to “atomic” terminology concepts to retrieve synonyms, to then do the actual mapping with the above-described optimization. A very important addendum for languages with compounding is that since the algorithm works on base of word stems, it also works perfectly for compound nouns.

For disambiguation, the NLP engine combines heuristic approaches that use machine-learning algorithms with the well-established method of word collocations but also makes use of the ontology itself. If a word, abbrebivation or phrase has multiple meanings, the ontology is used for the calculation of semantic distances, which helps in guessing the correct meaning.

For the German-speaking area, there is at present only one terminology that both covers the medicine to the here required extent and is expressible in description logic. This terminology is the Wingert nomenclature (WNC) of the Friedrich Wingert Foundation. Its roots lie in SNOMED 2, although over the last decades and in analogy to SNOMED CT, its structure has been formalized and the Wingert nomenclature has been transformed and can now be expressed in dialects of description logic, like AL (attributive language) and EL (existential language) [A3].

In contrast to SNOMED, the WNC does not contain generic concepts that can be used as placeholders. Typically, such concepts are used in classifications to allow coding of items into categories, which are “not otherwise specified”. The disadvantage of such concepts is shown in [A4].

The particularity of the WNC is that all semantic roles are represented using a generic approach. The semantic bottom role is simply designated as "hasContext" and further defined through a parameter. This parameter in turn is an element of the terminology itself.

The representation of "`has_Topography`" or respectively "`has_FindingSite`" is consequently carried out using the expression "`hasContext{Topography}`". The general notation is "`hasContext{parameter}`". The advantage of this approach is that the roles themselves are represented by the ontology so that they can also be semantically "understood". This benefit becomes clear in the following example:

Let $X1 \equiv \exists hasContext\{R1\}.Y1$ and

$X2 \equiv \exists hasContext\{R2\}.Y2$ and

$R1 \sqsubset R2$, then follows

X2 ⊑ ∃hasContext{R1}.Y2

Transformation of that approach into standard notation and proof for the expressibility in AL and EL was done using Protegé [A5] and public available reasoners (ELK, PELLET) [A6, A7].

The use of certain perspectives in the maintenance of an ontology inevitably leads to contradictory models. This holds true in a general sense for aspects of the TCM (Traditional Chinese Medicine) as opposed to orthodox western medicine, which often makes subsumptions that do not reflect medical reality. These aspects can be represented in the WNC by using so-called "views". In so doing, basically the modeling of the ontology is parameterized. The approach of working with "views" is especially used for the implementation of negations, although it's only used to suppress inheritance along the "is-a" relation.

The modeling of the partOf relation is associated with huge problems, as outlined by Schulz and Hahn [A8]. The propagated "SEP (Structure Entity Part) approach" goes already back to the 80ies and was introduced and discarded by Young et al. [A9]. Although this approach had first been chosen for SNOMED CT, today alternatives are discussed. In the WNC, the partOf relation was not modeled transitively but especially used for the representation of anatomical relationships. Conceptual relations like, for instance, "drug ⊑ ∃hasPart.active_agent" or "bone_fracture ⊑ ∃hasPart.bone ∩ fracture" were modeled in the sense of conceptual relations using the generic "hasContext"-approach, leading, for example, to terms like "drug ⊑ ∃hasContext{concept}.active_agent". Thus this type of partonomic relation describes intrinsic parts.

The here used terminology server is CTS2 compliant [A10], which is a standard that has emerged from Open Terminology Services [A11]. The use of a standardized tool for the integration of the knowlegde allows for an uncomplicated exchange of the implementation and enables a flexible use of different ontologies. Among others, the backbone of the NCBO bioportal, which contains almost 600 terminologies and ontologies, is based on CTS2 [A12]. Therefore, all these ontologies can, in principle, be used when working with this approach. The CTS can be connected to a reasoner, which evaluates DL expressions in the chosen dialect.

Addendum 2

In the following, the conversion is explained on a simple example. So, it shall be queried for patients of a certain age with an allergy. For this purpose, the O-SQL query is formulated as follows:

```
select p.id as patient_id from patients p, allergies a where p.age > 65
and p.id = a.patient_id and [penicillin](a.allergy)
```

The tables "patients" and "allergies" are connected via the relationship "p.id = a.patient_id" and the free-text allergies are documented in the table "allergies" and column "allergy". The preprocessor then extracts the expression "[penicillin](a.allergy)", retrieves the concept-id for "penicillin" with the assistance of the integrated terminology server and resolves the references by means of the annotation tables. Thus, the resulting standard SQL is:

```
select p.id as patient_id from patients p, allergies a where p.age > 65
and p.id = a.patient_id and a.id in (23, 33, 45)
```

By means of this statement, now all patients are selected who are older than 65 years and for whom an allergy to "penicillin" has been recorded. This query includes, on the one hand, patients suffering from an allergy to one specific penicillin like, for example, Oxacillin or Amoxicillin, and on the other hand, it excludes those patients for whom such an allergy had been ruled out in the free-text documentation (e.g. "no allergy against pencillin")

Element	Values	Sample
Prefix	If the prefix is set to "+" the relation "isA" will be added to the given relation	"+partOf[bone]" equals to "partOf[bone] or isA[bone]"
Relation	• isA (default) • parentOf / isA • partOf / hasPart • isContextedBy / hasContext • context (equals to isContextedBy and hasContext) • siblingOf	"isA[inflammation]" will also find "phlegmon" and so on
Context	A free-text parameter further specifying the "context" relation	"hasIndication" is expressed by "context{Indication}"
Depth	Number specifying the inheritance depth; see explanation in the main text and in Figure 3	
Query	The free-text query; can be a single keyword or a complete phrase including negations	"chd", "diabetes without complication"
Table.Column	Name of the table and column to search in; can be a comma-separated list	

Table A1 Possible values of the elements of an O-SQL expression.

Query	TP	FP	TN	FN
Vitium cordis	220	1	1641	16
Respiration disorder	393	3	63	9
Lethal malformation	480	50	1303	41

Table A2 Raw data used for statistics calculation. TP = true positve; FP = false positive; TN = true negative; FN = false negative

Addendum references

[A1] Wingert F, Automated indexing of SNOMED statements into ICD., Methods Inf Med 1987 26: 93-98

[A2] Allones JL Martinez D, Taboada M, Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology. J Med Syst. 2014 Oct;38(10):134. doi: 10.1007/s10916-014-0134-x. Epub 2014 Sep 2.

[A3] Baader F, Calvanese D, Mcguiness DL, Nardi D, Patel-Schneider PF, The description Logik Handbook, Cambridge Press (2003)

[A4] Ruch P, Gobeill J, Lovis C, Geissbuehler A, Automatic medical encoding with SNOMED categories, BMC Medical Informatics and Decision Making 2008, 8(Suppl 1):S6

[A5] Musen, M.A. The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.

[A6] Kazakov Y, Kroetzsch M, Simancík F, Concurrent Classification of EL Ontologies. Technical report. University of Oxford 2011

[A7] Evren Sirin a , Bijan Parsia a , Bernardo Cuenca Grau a,b , Aditya Kalyanpur a , Yarden Katz, Pellet: A Practical OWL-DL Reasoner, Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 5, Issue 2, (2007):51-53

[A8] Schulz S, Hahn U, Part-whole representation and reasoning in formal biomedical ontologies, Artificial Intelligence in Medicine (2005) 34, 179—200

[A9] Yang Y, Patil R. KOLA: a knowledge organization language. In: Kingsland III LC, editor. SCAMC'89 — Proceedings of the 13th annual symposium on computer applications in medical care. New York, NY: IEEE Computer Society Press; (1989). p. 71—5.

[A10] Common Terminology Services 2 http://www.omg.org/spec/CTS2/ (Accessed: 06.11.2016)

[A11] Solbrig HR, Armbrust DC, Chute CG, The Open Terminology Services (OTS) project., AMIA Annu Symp Proc. (2003):1011.

[A12] NCBO Bioportal, http://bioportal.bioontology.org/ (Accessed: 07.11.2016)

Eidesstattliche Versicherung

„Ich, André Sander, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: „Die Anwendung von medizinischen Terminologien auf Freitext in Routinedatenbanken am Beispiel von Strategien zur Reduktion der Säuglingssterblichkeit“ (“The application of medical terminologies to free-text in routine databases using the example of strategies to reduce infant mortality”) selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen werden von mir verantwortet.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Betreuer/in, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass mir die Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis bekannt ist und ich mich zur Einhaltung dieser Satzung verpflichte.

Sämtliche Arbeiten an den primären Patientendaten wurden, entsprechend den Auflagen des Berliner Datenschutzes, auf dem Campus der Charité in der Klinik für Neonatologie durchgeführt.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

_____ _____

Datum Unterschrift

Anteilserklärung an den erfolgten Publikationen

André Sander hatte folgenden Anteil an den folgenden Publikationen:

Publikation 1

Sander A, Wauer R. Integrating terminologies into standard SQL: a new approach for research on routine data. J Biomed Semantics. 2019 Apr 24;10(1):7

Erstellung sämtlicher Inhalte, inklusive:

- Informationsmodell
- Datenmodell
- Implementierung eines Prototypen
- Auswertung
- Cloud basierte Bereitstellung des Prototypen
- Anfertigung der schriftlichen Ausarbeitung

Publikation 2

Sander A, Wauer R. From single-case analysis of neonatal deaths toward a further reduction of the neonatal mortality rate. J Perinat Med. 2018 Dec 19;47(1):125-133

- Erstellung der Studiendatenbank
- Korrektur aller Transkriptionen
- Recherche aller Vergleichszahlen (Abbildung 1 und Tabelle 1)
- Erstellung des Gold-Standard
- Kuration der benutzten Ontologie
- Analyse und Anwendung der Nordic-Baltic-Classification
- Statistiken und Berechnungen (Tabelle 1, 2, 3)
- Definition der medizinischen Abfragen (Tabelle 4)
- Anfertigung der schriftlichen Ausarbeitung

_____

Unterschrift, Datum und Stempel des betreuenden Hochschullehrers

_____

Unterschrift des Doktoranden

Publikationen

Integrating Terminologies into Standard SQL: A New Approach for Research on Routine Data

Journal of Biomedical Semantics, Springer, https://doi.org/10.1186/s13326-019-0199-z
Received July 23, 2018. Accepted march 26, 2019.
https://jbiomedsem.biomedcentral.com/about
2-year IMPACT FACTOR: 1.600
5-year IMPACT FACTOR: 1.883
SCImago Journal Rank (SJR): 0.952 (First quarter)
https://www.scimagojr.com/journalsearch.php?q=21100316001&tip=sid
Source Normalized Impact per Paper (SNIP) 2017: 1.225

Die Publikation ist in der elektronischen Version aus lizenztechnischen Gründen nicht enthalten.

From Single-Case Analysis of Neonatal Deaths toward a Further Reduction of the Neonatal Mortality Rate

Journal of Perinatal Medicine, De Gruyter, https://doi.org/10.1515/jpm-2018-0003
Received January 2, 2018. Accepted June 25, 2018.
https://www.degruyter.com/view/j/jpme
IMPACT FACTOR 2017: 1.558
5-year IMPACT FACTOR: 1.653
CiteScore 2017: 1.26
SCImago Journal Rank (SJR) 2017: 0.594 (Second quarter)
https://www.scimagojr.com/journalsearch.php?q=27622&tip=sid
Source Normalized Impact per Paper (SNIP) 2017: 0.684

Die Publikation ist in der elektronischen Version aus lizenztechnischen Gründen nicht enthalten.

Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Publikationsliste

- Sander A, Wauer R, From Single-Case Analysis of Neonatal Deaths toward a Further Reduction of the Neonatal Mortality Rate, Journal of Perinatal Medicine, De Gruyter, https://doi.org/10.1515/jpm-2018-0003

- Sander A, Wauer R, Integrating Terminologies into Standard SQL: A New Approach for Research on Routine Data, Journal of Biomedical Semantics, Springer, https://doi.org/10.1186/s13326-019-0199-z

- Deng Y, Sander A, Faulstich L, Denecke K: Towards automatic encoding of medical procedures using convolutional neural networks and autoencoders. Artificial Intelligence in Medicine; 2019: 93, 29-42; https://doi.org/10.1016/j.artmed.2018.10.001

Danksagung

Ich möchte Prof. Wauer für seine vor allem geduldige und zuweilen fordernde Begleitung, Betreuung und Beratung danken. Ich möchte Fr. Prof. Rapoport für die Bereitstellung der Daten und den daraus gewonnenen tiefen Einblick in ihr Lebenswerk danken. Prof. Bührer gebührt mein Dank für die Bereitstellung eines Arbeitsplatzes zur Analyse der Daten und generell der Möglichkeit an seinem Institut diese Arbeit zu schreiben.

Der Familie Diekmann möchte ich für diese außerordentliche Möglichkeit und Erfahrung danken – Fritz, der die Idee zu dieser Dissertation hatte, und Daniel, der mir den Freiraum gegeben hat, die Idee letztendlich auch umzusetzen.