COMPUTING HEALTHCARE QUALITY INDICATORS AUTOMATICALLY

Secondary Use of Patient Data and Semantic Interoperability

KATHRIN DENTLER

SIKS Dissertation Series No. 2014-17

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

The research reported in this thesis has been carried out in cooperation of the Department of Medical Informatics, Academic Medical Centre, University of Amsterdam, The Netherlands, and the Department of Computer Science, VU University Amsterdam, The Netherlands.

VRIJE UNIVERSITEIT

COMPUTING HEALTHCARE QUALITY INDICATORS AUTOMATICALLY

Secondary Use of Patient Data and Semantic Interoperability

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op maandag 19 mei 2014 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Kathrin Dentler
geboren te Karslruhe, Duitsland

promotor:	prof.dr. F.A.H. van Harmelen
copromotoren:	dr. N.F. de Keizer
	dr. R. Cornet
	dr. A.C.M. ten Teije

promotiecommissie:	prof.dr. A. Abu-Hanna
	prof.dr. R. Huijsman
	prof.dr. N. Klazinga
	dr. M. Peleg
	dr. Z. Huang

To the loving memory of my grandfather prof. dr. Hermann Füllenbach.
(Who used to brag that he did not even understand the title of this thesis.)

CONTENTS

ACKNOWLEDGMENTS

"Not everything that can be counted counts
and not everything that counts can be counted."
Albert Einstein

The past years as a PhD student in Amsterdam have probably been the most insightful, eventful and exciting time of my life, and I am immeasurably thankful to the many brilliant, enthusiastic and interesting people who supported me on the way.

During my master studies, many of my fellow students infected me with their ambition and enthusiasm for research. This enthusiasm was strengthened when working on my thesis about swarm reasoning, supervised by Christophe Guerét and Stefan Schlobach. Our ideas were so crazy that they caused strange dreams, but in the end we had a working prototype and an accepted ISWC workshop publication. Stefan believed in me from the beginning, and I owe both the idea to pursue a PhD as well as active support along the way to him.

My sincerest appreciation goes to my numerous (co-)promotors: Frank van Harmelen, Annette ten Teije, Nicolette de Keizer and Ronald Cornet. I truly admire how well they worked together as an interdisciplinary team from two research groups. They always found a balance between ensuring that I fulfil my research plan and leaving me the freedom to explore new ideas. Together, they motivated me to go the extra mile. They left no sentence unchecked and no idea undiscussed, always open for the best argument.

At the VU, I feel privileged that Frank is my promotor, as his Semantic Web Primer was one of the reasons that I moved to Amsterdam in the first place. The way how he captivates everybody surrounding him with his sparkling enthusiasm for science fascinates me. Annette regularly came up with outstanding ideas for our research, and always made me leave our weekly meetings full of inspiration and an urge to get things done. I also thank Elly Lammers, who never failed to provide valuable advice regarding administrative matters.

At the AMC, Ronald and Nicolette welcomed me to the world of medical informatics, and we soon shared a passion for SNOMED CT. Nicolette kept track on my research plan with a helicopter view, and patiently taught me scientific writing for medical informatics venues. As far as I

can judge, she succeeded. She also attempted to improve my Dutch, but here I am less optimistic about her achievements. Nonetheless, we never failed to understand each other, not least due to her great sense of empathy. Ronald contributed to our weekly meetings with his love for description logics, enthusiasm, sense of humour and ideas from beyond the box.

The AMC allowed me to ground my research in real care for real patients, which I owe primarily to the fruitful cooperation with Kristien Tytgat, Jean Klinkenbijl and Pieter Tanis from the GIOCA, the Gastro-Intestinal Oncology Centre Amsterdam. From the very beginning, they were enthusiastic and open, for example by allowing me to follow a patient for an entire day, and they always took the time to answer my countless questions. My gratitude also goes to the administrators of DIO and to Sean Heukels from the GIOCA, who - after a long quest - finally provided me with the patient data I needed. Likewise, I am indebted to Matthijs Numans. Our cooperation gave me the interesting opportunity to explore quality indicators not only from a hospital's, but also from a general practitioner's perspective.

Regarding my thesis and its defence, I thank the committee for having read and accepted it. I also thank my extraordinary paranymphs Krystyna Milian and Nicolas Höning, for always having been there to discuss and to enjoy the peculiarities of life as PhD students.

I thank the members of the wonderful groups of both the AMC and the VU, and my roommates of both J1b - 109 and U 3.03. Along the way, many of my colleagues became friends, so that Amsterdam felt like home. Cheers to them and to all my other friends who made the time worthwhile. I also thank everybody who travelled far to visit me, making it easy to maintain our friendships.

Last but foremost, I thank each member of my dear family for making me feel loved and special, and supporting me in all possible ways, particularly when I needed it most. I especially thank my open-minded, positive, adventurous and affectionate mother. My deepest gratitude goes to my fiancé Frederic for his unconditional love, and to our beautiful sons Atreju and Jonathan, who stayed with us far too short, but long enough to teach us what matters most.

Amsterdam, April 2014 Kathrin Dentler

1 INTRODUCTION

"Using data for secondary purposes is one of the most promising ways to improve health outcomes and costs"
PricewaterhouseCoopers, 2009

1.1 MOTIVATION AND MAIN RESEARCH QUESTION

Today, increasing volumes of healthcare data are routinely recorded and stored in Electronic Medical Records (EMRs)[1]. This rapid adoption of EMRs opens the door to large-scale secondary uses of patient data, with tremendous potential benefit both for individual patients and society in general. In fact, according to a report by PricewaterhouseCoopers [1], using data for secondary purposes is one of the most promising ways to improve health outcomes and costs. The question of trustworthy reuse of data [2] has become an important challenge and research question. Secondary purposes comprise clinical research, the recruitment of eligible patients for clinical trials, decision support, the early detection of epidemics, reimbursement, clinical audit, the generation or testing of medical hypotheses and quality monitoring or reporting based on healthcare quality indicators, which is the core subject matter of this thesis. Our main research question is:

> *Under which conditions can healthcare quality indicators be computed automatically by reusing data already collected during the clinical care process?*

More and more both legally mandatory and voluntary quality indicators are released by governments, patient associations, scientific associations and insurance companies to compare hospitals and general practitioners, and to monitor or improve the quality of their delivered care. Patients use quality indicators to select the care provider of their preference, and insurance companies to make informed choices regarding healthcare contracting. A quality indicator[2] is "a measurable element of practice performance

1 The term Electronic Medical Record is used interchangeably with Electronic Health Record (EHR) throughout this thesis.

2 The term quality indicator is used interchangeably with clinical / medical indicator / measure in this thesis. However, as most measures are only indicators of quality, we prefer the term indicator [3].

for which there is evidence or consensus that it can be used to assess the quality, and hence change in the quality, of care provided" [4]. According to Donabedian [5], quality indicators can be related to structure, process or outcome. Structure denotes the attributes of the settings in which care occurs. Process denotes what is actually done in giving and receiving care, and outcome denotes the effects of care on the health status of patients and populations. Process and outcome indicators typically focus on specific patient populations, and are often expressed as a percentage. The denominator consists of the relevant cohort of patients to whom the quality indicator applies, and the numerator of those patients contained in the denominator for whom criteria that indicate (high or low) quality of care are fulfilled.

The sample process indicator "Number of examined lymph nodes after resection of a primary colonic carcinoma" recurs in several chapters of this thesis. The indicator is defined by the Dutch Healthcare Inspectorate and has to be computed and reported by all Dutch hospitals that perform gastrointestinal surgical oncology. At least 10 lymph nodes should be examined after resection of a primary colonic carcinoma, and the indicator measures the proportion of patients for whom this is the case:

> **Number of examined lymph nodes after resection**
> *Numerator:* Number of patients who had 10 or more lymph nodes examined after resection of a primary colonic carcinoma.
>
> *Denominator*: Number of patients who had lymph nodes examined after resection of a primary colonic carcinoma.
>
> *Exclusion criteria*: Previous radiotherapy and recurrent colonic carcinoma.
>
> *Reporting year*: 2010

The indicator is partially based on evidence: Lymph node involvement is an essential part of colonic carcinoma staging systems, and correct staging of a resected colonic carcinoma is important both to assess a patient's prognosis and to make informed decisions regarding required postoperative adjuvant therapy. Patients with few retrieved and examined lymph nodes could possibly be misclassified as node-negative, as the probability to find metastatic nodes decreases for lower numbers of examined nodes. Thus, as many lymph nodes as possible should be examined to achieve a reliable staging. Several studies confirmed that the number of examined lymph nodes correlates with patient survival and searched for a cutoff value, i.e. a minimum number of lymph nodes to be examined [6–11].

Even though these studies do not agree on a specific cutoff value, the Dutch colonic carcinoma guideline recommends based on these studies that 10 or more lymph nodes should be examined in order to classify a tumor as node-negative [12].

A prominent problem is that most indicators are released in inherently ambiguous natural language. For instance, our sample indicator does not state explicitly which events (the diagnosis, the surgery, the pathology examination, a subset or all of them) should have taken place during the reporting year. The interpretation process typically takes place locally and in an ad-hoc manner, potentially leading to different interpretations in different institutions, and thereby to reduced indicator reliability[3]. Different interpretations can lead to significant differences in computed indicator results [13], which in turn causes their reduced validity and comparability.

Quality indicators might even be erroneous: during the formalisation process in cooperation with clinical domain experts, the question arose whether the denominator should indeed consist of patients who underwent a resection of a primary colonic carcinoma and had lymph nodes examined, as explicitly stated in the denominator, or rather of all patients who underwent a resection of a primary colonic carcinoma, which would make more sense from a clinical perspective, as for all patients who underwent a resection of a primary colonic carcinoma, lymph nodes should be examined. We reported this issue to those who release the indicator, and it has been corrected for the subsequent reporting year (2011).

Currently, quality indicators are often calculated *manually*, leading to a variety of problems: the process is time-consuming, expensive and error-prone, and the results are typically not timely, so that they can not be used to intervene directly to improve the quality of care.

For the automated computation of quality indicators, indicators need to be formalised, and the semantic gap between indicators and data sources needs to be bridged, and typically data from heterogeneous sources needs to be integrated. For instance, our sample indicator requires data concerning surgical procedures, diagnoses, radiotherapy sessions and pathology reports. Our approach to integrate "resources that were developed using different *vocabularies* and different *perspectives* on the data" [14] is *semantic interoperability*. According to Heflin and Hendler [14], "To achieve semantic interoperability, systems must be able to exchange data in such a way that the *precise meaning* of the data is readily accessible and the data itself can be translated by any system into a form that it understands". For this purpose, standards for both *vocabularies* and *perspectives* on the data

3 A reliable indicator is defined so precisely that it is measured in the same way on different occasions or by different observers [4].

are required. In our healthcare domain, *vocabularies* are also referred to as *terminologies*, *classifications*, *coding systems or ontologies*. Standards are e.g. ICD, LOINC and SNOMED CT. *Perspectives* are referred to as *information models*, which are agreed-upon clinical data structure definitions. Standards are e.g. *open*EHR, ISO / EN 13606 and HL7 CDA.

In an ideal setting, quality indicators are released in a precise, structured, machine-processable, standards-based, formal representation and can seamlessly be integrated with Electronic Medical Records, so that they can be computed automatically and in real-time, based on the vast amounts of valuable patient data collected during routine care, and so that the computed results are comparable across institutions.

Our research question of "*under which conditions healthcare quality indicators can be computed automatically by reusing data already collected during the clinical care process?*" can be divided into three main subject areas: I) indicators and their formalisation, which is required so that they can be computed automatically; II) patient data and its (re)usability, which implies data availability, data quality, as well as the use of well-established healthcare standards and finally III) semantic interoperability, which is required to integrate indicators and patient data as well as heterogeneous patient data sources. The eight main chapters of this thesis are arranged into three main parts according to these three subject areas. All chapters are published scientific papers.

1.2 OPERATIONALISATIONS OF THE MAIN RESEARCH QUESTION AND METHODOLOGY

This section presents operationalisations of our main research question for each of the main parts of this thesis, as well as our approach to answer them.

1.2.1 *Part I) Formalising and Automatically Computing Healthcare Quality Indicators*

To compute quality indicators automatically, they have to be formalised. Therefore, a *formalisation method* is required. As clinical quality indicators are often computed in a decentralised manner by the hospitals themselves, *reproducibility* of the formalisation method is essential to ensure the comparability of computed values. Another important measure for any scientific method is its *generalisability*, and therefore we aim to assess to what extent our method is applicable to a large set of heterogeneous indicators of different types and from various domains.

The first part of this thesis tackles the following research questions:

1. *How can quality indicators be formalised?*
2. *How reproducible are the results of our formalisation method, and which steps are particularly challenging?*
3. *How generalisable is the resulting method?*

Our approach to answer research question 1 was as follows: based on a literature study and a requirements analysis, we developed CLIF, a stepwise novel and flexible method to formalise healthcare quality indicators from natural language into an unambiguous, machine-processable, formal representation. Percentage-based quality indicators can be regarded as two queries that retrieve patients who fulfil certain constraints and criteria, one for the denominator and one for the numerator. Formalised indicators can be computed automatically by running these queries against an EMR. To prove the concept, we applied the method to formalise sample quality indicators manually from natural language into SPARQL queries, and ran them against self-generated synthetic patient data. Due to its large coverage and as it allows for meaning-based recording and retrieval of clinical information, we employed SNOMED CT to represent both the indicators and the patient data.

In order to assess the reproducibility of CLIF and to answer research question 2, we performed a case study to investigate whether several test persons who formalise the same quality indicator independently attain the same formalisation. For this study, we implemented a web-based indicator-authoring tool to facilitate the formalisation process by leading users through the method step by step. We analysed the results per step by comparison to a reference standard, which we developed in cooperation with clinical experts, to investigate which steps are particularly challenging and why.

To answer research question 3, we formalised the entire national set of 159 quality indicators for general practices with CLIF. The set of quality indicators is heterogeneous, as it contains indicators of various types (structure, process and outcome) and addresses 7 domains, such as asthma in adults and diabetes mellitus. Each domain contains a number of subdomains, such as HbA1c or smoking. Subsequently, we computed the formalised indicators based on a large database containing data related to more than 150,000 patients in the years between 2006 and 2011. Apart from the set of indicators for general practices, we also formalised and computed various colorectal cancer indicators for hospitals.

The first part of this thesis comprises the following chapters:

> **Chapter 2**: Kathrin Dentler, Annette ten Teije, Ronald Cornet, and Nicolette F. de Keizer. *Towards the automated calculation of clinical quality indicators.* In Knowledge Representation for Health-Care, LNCS 6924:51-64, Springer 2012.

Chapter 3: Kathrin Dentler, Ronald Cornet, Annette Ten Teije, Kristien Tytgat, Jean Klinkenbijl, and Nicolette F. de Keizer. *The Reproducibility of CLIF, a Method for Clinical Quality Indicator Formalisation.* Studies in Health Technology and Informatics, 180:113-7, 2012.

Chapter 4: Kathrin Dentler, Mattijs E. Numans, Annette ten Teije, Ronald Cornet, and Nicolette F. de Keizer. *Formalization and Computation of Quality Measures based on Electronic Medical Records.* Journal of the American Medical Informatics Association. Published Online First: 5 Nov 2013.

1.2.2 *Part II) Secondary Use of Patient Data*

Even though more and more data is recorded and stored during routine care, the mere existence of digitalised patient data is not the only requirement so that it can be used for *secondary purposes*. Therefore, our first objective regarding this part of the thesis was to assess possible *barriers* that impede the secondary use of patient data.

Secondary uses of data might require different degrees and dimensions of *data quality* than primary uses, and various authors suggested that reliable and valid quality indicator results are only achievable based on accessible and high-quality data [15–21]. Thus, our second objective was to assess whether the data quality of our hospital's EMR is sufficient to reliably compute colorectal cancer surgery indicators.

*Open*EHR archetypes [22] have been proposed to standardise clinical data to achieve semantic interoperability, and they have been shown to facilitate the integration of data from several sources and thereby the (re)use of EHR data. Clinical information to compute quality indicators is often scattered among various heterogeneous information silos that use different information models, so that it needs to be integrated to be (re)usable. Also, standards-based indicators are computable across institutions that employ the respective standards, and otherwise easier to map to local data structures than implicit non-standard definitions as often used in free text quality indicators. Therefore, one of our prevalent research objectives was to establish whether *open*EHR archetypes are suitable to represent both EHR data and elements of patient data required by indicators, and thereby to semantically integrate routine clinical data and quality indicators.

The specific research questions that we aim to answer in this part are:

4. *What are the barriers that impede the secondary use of patient data, and how can they be prevented?*
5. *How does data quality influence the reliability of quality indicator results?*

6. *Can* openEHR *archetypes facilitate the semantic integration of quality indicators and routine patient data to automatically compute indicators?*

In this part, we analyse the problem of automated data reuse in a real clinical setting within the Gastro-Intestinal Oncology Centre Amsterdam (GIOCA). The GIOCA is a specialised outpatient clinic within the Academic Medical Centre (AMC) that was founded to improve the quality of care for patients with (suspected) cancer of the gastrointestinal tract. Patients who register at the GIOCA are scheduled for an appointment within seven days at most. During this appointment, examinations to diagnose the patient are carried out, the case is discussed in a multidisciplinary meeting, and a detailed treatment plan is established and communicated to the patient. As this patient-centred rapid diagnosis process reduces the time until treatment starts, which might positively influence patient outcomes, the founders of the GIOCA are motivated to measure its performance.

We chose the domain of colorectal cancer surgery because it is also the subject of the Dutch Surgical Colorectal Audit (DSCA)[4], which has been set up in 2009 to measure and to improve the quality of colorectal cancer surgery. The DSCA is a medical quality registry that collects the data items necessary to compute a set of colorectal quality indicators released by the Dutch government. All Dutch hospitals that perform colorectal cancer surgery submit data to the DSCA register. Due to the importance of high data quality in the DSCA and also due to various barriers that impede the reuse of data, these data items are entered manually by one of our surgeons, which is labour-intensive and might lead to the undesirable situation that the data in registers differs from source data in an EMR. The GIOCA uses the same information systems as other departments of our hospital, plus additional spreadsheets for internal administration and management.

To answer research question 4, our first goal was to gather all raw source data from our hospital that is required to compute the set of indicators relevant for the GIOCA. In the course of this process, we experienced a number of barriers, including data quality issues after we finally obtained a version of the required data. We categorised all encountered barriers according to Galster's framework of causes that impede the reuse of clinical data in clinical settings [23].

To answer research question 5, we compared a set of 10 quality indicators computed based on routinely collected data from our EMR to the same indicators computed based on manually collected data for the DSCA register as reference standard, and performed a data quality analysis to

4 http://www.dccg.nl/colorectalaudit

explain any differences. We assessed the computability of quality indicators, absolute percentages of indicator results, and data quality in terms of availability in a structured format, completeness and correctness.

Our approach to answer research question 6 was to express both data from the AMC's data warehouse and the DSCA as well as previously formalised quality indicators in terms of *open*EHR archetypes. We then constructed archetyped SPARQL queries and ran them against the archetyped patient data to compute the indicators.

The second part of this thesis comprises the following chapters:

> **Chapter 5**: Kathrin Dentler, Annette ten Teije, Nicolette F. de Keizer, and Ronald Cornet. *Barriers to the Reuse of Routinely Recorded Clinical Data: A Field Report.* Studies in Health Technology and Informatics, 192:313-317, IOS Press 2013.
>
> **Chapter 6**: Kathrin Dentler, Ronald Cornet, Annette ten Teije, Pieter Tanis, Jean Klinkenbijl, Kristien Tytgat, and Nicolette F. de Keizer. *Influence of data quality on computed Dutch hospital quality indicators: a case study in colorectal cancer surgery.* BMC Medical Informatics and Decision Making, 14:32, 2014.
>
> **Chapter 7**: Kathrin Dentler, Annette ten Teije, Ronald Cornet, and Nicolette F. de Keizer. *Semantic Integration of Patient Data and Quality Indicators Based on* open*EHR Archetypes.* In Process Support and Knowledge Representation in Health Care, LNAI 7738:85-97, Springer 2012.

1.2.3 *Part III) Reasoning and Ontologies for Semantic Interoperability*

Typically, patient data is very detailed, but quality indicators query for groups of patients on a less granular level. For example, rectum cancer patients are undergoing the procedures "Stapled transanal resection of rectum" or "Wedge resection of rectum", which are both subclasses of "Resection of rectum". When we query for all patients with a procedure of type "Resection of rectum" to compute an indicator, we want to retrieve all patients with this procedure or a subclass thereof. Clinical terminologies can help to bridge such gaps, as they enable meaning-based retrieval by aggregating data conceptually according to its meaning. Hence, the encoding of relevant concepts from an indicator by concepts from a terminology is an essential step of our formalisation method, and we worked with several non-standard and standard terminologies such as SNOMED CT throughout the thesis to encode both relevant concepts of the indicators and the patient data.

As SNOMED CT relies on a logics-based representation, *automated reasoning* can be employed to bridge differences in granularity by inferring subclass relationships. SNOMED CT can be represented as an OWL 2 EL ontology, which in turn is based on the lightweight Description Logic EL++ [24, 25]. OWL 2 EL trades expressive power for the efficiency of reasoning, and is therefore suitable for typically large biomedical ontologies such as SNOMED CT, which comprises approximately 300,000 classes. A reasoner is a program that infers logical consequences from a set of explicitly asserted facts or axioms and typically provides automated support for reasoning tasks such as classification, debugging and querying. In practice, reasoners might vary with regard to their characteristics. Therefore, our objective was to identify relevant characteristic properties, and to assess how a selection of reasoners performs with respect to these properties.

Redundant elements in SNOMED CT concept definitions are harmless from a logical point of view, but they make concept definitions unnecessarily hard to construct and to maintain. Moreover, redundant elements might lead to content-related problems when concepts drift. For example, the rolegroup in the subconcept *Thyroid_uptake_with_thyroid_stimulation* was redundant in the July 2012 version of SNOMED CT, as it repeated a rolegroup already contained in the definition of the superconcept *Non-imaging_thyroid_uptake_test*. In the subsequent version of SNOMED CT, the method Radionuclide_*imaging* was removed from the rolegroup in the superconcept, which makes sense for a concept with the name *Non-imaging*_thyroid_uptake_test. However, the method was not removed from the rolegroup in the subconcept, which might be incorrect. Here, our research objective was to develop a method to identify redundant elements, and to apply this method to get insights into the extent of the issue.

The two research questions we aim to answer in this part are:

7. *What are the characterising properties of reasoners for OWL 2 EL, and how does a selection of reasoners perform with respect to these properties?*
8. *How can redundant elements in concept definitions of SNOMED CT be identified? How many redundant elements are identifiable using the resulting method?*

Our approach to answer research question 7 was as follows. To identify characterising properties of reasoners for OWL 2 EL, we analysed papers that describe reasoners as well as short advertising descriptions of reasoners, which usually outline the respective reasoner's strong points. Additionally, some characteristics in the dimension of practical usability arose while the reasoning experiments were performed. Subsequently, we categorised eight state of the art reasoners along the defined characteristics and benchmarked them against well-known biomedical ontologies.

To tackle research question 8 and to identify redundant elements in SNOMED CT concept definitions, we adapted and extended the rules of redundancy elimination for concept definitions that contain rolegroups as defined by Spackman et al. [26]. We systematically analysed the completeness and soundness of the results of our method by examining the identified redundant elements.

The third part of this thesis comprises the following chapters:

Chapter 8: Kathrin Dentler, Ronald Cornet, Annette ten Teije, and Nicolette F. de Keizer. *Comparison of reasoners for large ontologies in the OWL 2 EL profile.* Semantic Web Journal, 2:71-87, 2011.

Chapter 9: Kathrin Dentler and Ronald Cornet. *Redundant Elements in SNOMED CT Concept Definitions.* In Artificial Intelligence in Medicine, LNAI 7885:186-195. Springer 2013.

1.3 CONTRIBUTIONS

Our research resulted in the following contributions:

Part I) Formalising and Automatically Computing Healthcare Quality Indicators. (Chapters 2, 3 and 4, research questions 1, 2, and 3.)

1. A method to formalise quality indicators.
2. Insights into the reproducibility of the formalisation method, and into which steps can be particularly challenging.
3. Insights into the generalisability of the formalisation method.
4. A web-based tool that implements the formalisation method to lead users through the formalisation process.[5]
5. Various sets of formalised quality indicators.[6]

Part II) Secondary Use of Patient Data. (Chapters 5, 6 and 7, research questions 4, 5, and 6.)

6. Identification of barriers that impede the secondary use of patient data for the computation of quality indicators, and recommendations on how to prevent them.
7. Results showing that data quality can have a significant influence on quality indicator results.
8. Results showing that archetypes can facilitate the semantic integration of quality indicators and routine patient data to automatically compute indicators.

Part III) Reasoning and Ontologies for Semantic Interoperability. (Chapters 8 and 9, research questions 7 and 8.)

9. Definition of characteristics of OWL 2 EL reasoning engines, and a categorisation of eight reasoners along these characteristics. Results showing that reasoners can vary substantially.
10. A method to identify redundant elements in SNOMED CT concept definitions, and results showing that 12% of the concepts in the employed SNOMED CT version contained redundant elements.

5 Available on github: https://github.com/kathrinrin/clif
6 Available on figshare: http://figshare.com/authors/Kathrin_Dentler/452665

Part I

FORMALISING AND AUTOMATICALLY COMPUTING HEALTHCARE QUALITY INDICATORS

2 TOWARDS THE AUTOMATED CALCULATION OF CLINICAL QUALITY INDICATORS

To measure the quality of care in order to identify whether and how it can be improved is of increasing importance, and several organisations define quality indicators as tools for such measurement. The values of these quality indicators should ideally be calculated automatically based on data that is being collected during the care process. The central idea behind this paper is that quality indicators can be regarded as semantic queries that retrieve patients who fulfil certain constraints, and that indicators that are formalised as semantic queries can be calculated automatically by being run against patient data. We report our experiences in manually formalising exemplary quality indicators from natural language into SPARQL queries, and prove the concept by running the resulting queries against self-generated synthetic patient data. Both the queries and the patient data make use of SNOMED CT to represent relevant concepts. Our experimental results are promising: we ran eight queries against a dataset of 300,000 synthetically generated patients, and retrieved consistent results within acceptable time.

2.1 INTRODUCTION

A quality indicator[1] is "a measurable element of practice performance for which there is evidence or consensus that it can be used to assess the quality, and hence change in the quality, of care provided" [4]. Quality indicators can be related to structure, process or outcome. According to Donabedian, structure denotes the attributes of the settings in which care occurs. Process denotes what is actually done in giving and receiving care, and outcome denotes the effects of care on the health status of patients and populations [5]. Process and outcome indicators typically average over specific populations, and are often expressed by a fraction. The denominator consists of the relevant cohort of patients to whom the indicator applies, and the numerator of those patients contained in the

1 The term quality indicator is used interchangeably with clinical / medical indicator / measure in this paper. However, as most measures are only indicators of quality, the term indicator is preferable [3].

denominator for which criteria that indicate (high or low) quality of care are fulfilled. Both for the population of the denominator and numerator, inclusion and exclusion criteria can apply.

Clinical quality indicators are typically being developed and released by governments, scientific associations, patient associations or insurance companies. They are calculated based on patient data within hospitals, and the obtained results are reported back to the indicator-releasing organisations. The increasing number of indicators makes their manual calculation difficult and time-consuming. Furthermore, indicators that are released in natural language need to be interpreted locally, which is error-prone due to the inherent ambiguity of natural language. Therefore, quality indicators should ideally be released in an unambiguous, machine-processable, formal representation in order to automatically calculate comparable values.

In this paper, we regard quality indicators as semantic queries against patient data, and propose a preliminary method for their formalisation into semantic queries. We prove the concept by applying exemplary formalised queries on self-generated coded data consisting of 300,000 patients. The next Section 2.2 presents our approach, and Section 2.3 our formalisation method. We detail the generation of synthetic patient data in Section 2.4, and present our experimental results in Section 2.5. We end the paper by discussing related work in Section 2.6, future work in Section 2.7 and our conclusions in Section 2.8.

2.2 APPROACH

Our test set of quality indicators (see appendix) contains four indicators that have been released in natural language and stem from the domain of gastrointestinal cancer surgery, but in principle, we aim for a domain-independent approach. We investigate the feasibility of formalising the set of indicators into SPARQL queries[2]. The exemplary SPARQL query below retrieves all instances of type patient (the SNOMED CT code for "patient" is SCT_116154003). The SELECT clause defines the only variable that is to be retrieved as result (i.e. ?patient), and the WHERE clause defines a triple pattern which contains the same variable and is to be matched against the data graph.

```
SELECT ?patient
WHERE {
  ?patient a sct:SCT_116154003 .
}
```

2 http://www.w3.org/TR/sparql11-query/

Our proposed formalisation method consists of 8 steps: 1) to encode relevant concepts from the indicator by concepts from a terminology, 2) to define the information model, and 3) to 5) to formalise temporal, numeric and boolean constraints as SPARQL FILTERs. Step 6) is to group constraints by boolean connectors, step 7) to identify exclusion criteria and step 8) to identify constraints that only aim at the numerator, in order to construct the denominator by removing these constraints. All steps are explained in Section 2.3.

To test the formalised queries, we synthetically generated patient data that is represented in OWL 2[3], allowing for automated reasoning and semantic interoperability. We employ SNOMED CT [27] concepts from the July 2010 version to describe both the query variables (step 1 of our method) and our patient data. Typically, patient data is very detailed, but quality indicators query for groups of patients on a less granular level. We employ Semantic Web reasoning to bridge this gap by inferring subclass relationships. For example, generated rectum cancer patients are undergoing the procedures "Stapled transanal resection of rectum" or "Wedge resection of rectum", which are both subclasses of "Resection of rectum". To calculate an indicator, we query for all patients with a procedure of type "Resection of rectum" and retrieve all patients with subclasses of this procedure by automated reasoning.

2.3 FORMALISATION OF QUALITY INDICATORS

This section describes our formalisation method. As the numerator is always a subset of the denominator, and is thus restricted by more constraints, we first formalise the numerator and afterwards construct the denominator from it by removing constraints. We formalised a set of four quality indicators (see appendix, referred to as I1 - I4). In the following, we present our method by formalising the exemplary process indicator "Number of examined lymph nodes after resection" (I1). The clinical background of the indicator is a colon cancer guideline that states: "A minimum of 10 lymph nodes is recommended to assess a negative lymph node status". The original version of the indicator is:

> **I1: Number of examined lymph nodes after resection (process indicator)**
> *Numerator: number of patients who had 10 or more lymph nodes examined after resection of a primary colon carcinoma.*
>
> *Denominator: number of patients who had lymph nodes examined after resection of a primary colon carcinoma.*
>
> Exclusion criteria: Previous radiotherapy and recurrent colon carcinomas

3 http://www.w3.org/TR/owl2-overview/

Step 1: Encoding of relevant concepts from the indicator by concepts from a terminology The first step of our method is to extract all required concepts from the indicator, and to find the corresponding concepts in a terminology, in our case SNOMED CT. We perform this step first because the concepts are the building blocks for further formalisation. In SPARQL, we encode the query variables based on those concepts:

```
?patient a sct:SCT_116154003 .
```

Step 2: Definition of the information model Subsequently, we define the information model, i.e. how the resources are related to each other. This step could be automated once a standard information model is employed. In SPARQL:

```
?patient ehrschema:hasDisease ?coloncancer .
```

Step 3: Formalisation of temporal constraints (FILTER) The next step is to formalise temporal constraints. This step helps us to discover an ambiguity: the indicator does not state explicitly what should be included the reporting year. It could be for example the resection of the carcinoma or the lymph node examination. Because the indicator aims at the number of examined lymph nodes, we assume the latter. One of the temporal relationships between two query variables in this indicator states that the lymph node examination has to follow the colectomy. These constraints are expressed as FILTERs in SPARQL. FILTERs restrict solutions to those for which the filter expressions evaluate to *true*:

```
FILTER ( ?lymphnodeexaminationdate > "2010-01-01T00:00:00+02:00"^^xsd:dateTime )
FILTER ( ?lymphnodeexaminationdate < "2011-01-01T00:00:00+02:00"^^xsd:dateTime )
FILTER ( ?lymphnodeexaminationdate > ?colectomydate)
```

Step 4: Formalisation of numeric constraints (FILTER) The only numeric constraint contained in the indicator is that the number of examined lymph nodes has to be 10 or more. In SPARQL:

```
FILTER ( ?numberexaminedlymphnodes >= 10 )
```

Step 5: Formalisation of boolean constraints (FILTER) The exemplary indicator does not contain boolean constraints. However, the indicator "Participation in Dutch Surgical Colorectal Audit" (DSCA, I2) asks for patients for which data has been delivered to the DSCA. In SPARQL:

```
FILTER ( ?dataDeliveredToDSCA = true)
```

Step 6: Grouping of constraints by boolean connectors All elements of the constructed SPARQL query are connected by logical conjunctions. However, some queries require logical disjunctions. An example is again I2, which asks for surgical resections of a colorectal carcinoma situated in colon *or* rectum:

```
{ ?cancer  a sct:coloncancer . ?operation a sct:colectomy }
UNION
{ ?cancer a sct:rectumcancer . ?operation  a sct:resectionrectum }
```

Step 7: Identification of exclusion criteria (FILTER) One of the exclusion criteria of the example indicator is "previous radiotherapy". Thus, we exclude all patients who underwent radiotherapy before the lymph node examination. All criteria that are not explicitly identified as exclusion criteria are inclusion criteria.

```
FILTER NOT EXISTS {
  ?radiotherapy a sct:SCT_108290001 .
  ?patient ehrschema:hasProcedure ?radiotherapy .
  ?radiotherapy ehrschema:procedureDate ?radiotherapydate .
  FILTER ( ?lymphnodeexaminationdate > ?radiotherapydate)
}
```

Step 8: Identification of constraints that only aim at the numerator In this step, the numerator is already formalised, and constraints are removed to construct the query for the denominator. In order to do so, it is important to be aware of the clinical intent of the indicator. Regarding the example indicator, it is considered good practice to examine 10 or more lymph nodes. Therefore, the only constraint that is removed to construct the denominator is: "number of examined lymph nodes $>= 10$".

Resulting SPARQL query (Numerator)

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ehrschema: <http://apdg.net/owl/schema/>
PREFIX sct: <http://www.ihtsdo.org/>

SELECT ?patient
WHERE {

# step 1)
?patient a sct:SCT_116154003 .
?coloncancer a sct:SCT_93761005 .
?colectomy a sct:SCT_23968004 .
?lymphnodeexamination a sct:SCT_284427004 .

# step 2)
?colectomy sct:SCT_47429007 ?coloncancer . # SCT_47429007 = associated with
?patient ehrschema:hasDisease ?coloncancer .
?patient ehrschema:hasProcedure ?colectomy .
?colectomy ehrschema:procedureDate ?colectomydate .
?patient ehrschema:hasProcedure ?lymphnodeexamination .
?lymphnodeexamination ehrschema:procedureDate ?lymphnodeexaminationdate .
```

```
?lymphnodeexamination ehrschema:hasNumber ?numberexaminedlymphnodes .

# step 3)
FILTER ( ?lymphnodeexaminationdate > "2010-01-01T00:00:00+02:00"^^xsd:dateTime )
FILTER ( ?lymphnodeexaminationdate < "2011-01-01T00:00:00+02:00"^^xsd:dateTime )
FILTER ( ?lymphnodeexaminationdate > ?colectomydate)

# step 4); needs to be removed to construct the denominator (step 8)
FILTER ( ?numberexaminedlymphnodes >= 10 )

# step 7)
FILTER NOT EXISTS {
  ?radiotherapy a sct:SCT_108290001 .
  ?patient ehrschema:hasProcedure ?radiotherapy .
  ?radiotherapy ehrschema:procedureDate ?radiotherapydate .
  FILTER ( ?lymphnodeexaminationdate > ?radiotherapydate)
}}
```

Regarding the order of the steps, step 1) and 2) should be carried out first, because they formalise the building blocks that are used in subsequent steps. Steps 6) - 8) should be carried out last, because they build on previously defined constraints. Steps 3) to 5) can be performed in the preferred order of the user.

Experiences during formalisation We succeeded in formalising all four quality indicators included in our example set as SPARQL queries with the method as described above, and the formalisation process was relatively straightforward. The only construct that is not directly expressible in SPARQL is: "number of re-interventions during the same admission or during 30 days after the resection (choose longest interval)" (I4), because there is no function to subtract dates from each other in SPARQL. This is clearly an insufficiency. Two possible options to circumvent this problem are to implement a custom extension function or to first query for all patients who had a re-intervention and then to apply the filter on the retrieved results. Both solutions need to be implemented locally (extension functions have to be implemented for the triple store that is being queried, and results need to be filtered where the data is retrieved), and thus allow for the introduction of implementation errors and limit interoperability.

We found a high coverage of SNOMED CT with respect to the colorectal cancer surgery domain. The only concept that we could not encode was the exclusion criterion "Transanal Endoscopic Microsurgery (TEM)" (I3 and I4). We excluded "Stapled transanal resection of rectum", "Transanal disk excision of rectum" and "Transanal resection of rectum and anastomosis using staples" instead. None of these replacements are explicitly "endoscopic". Alternatives would have been to post-coordinate the concept or to employ a concept from another terminology.

We did not implement subtleties such as the presence of a radiologist, a radiotherapist, a surgeon, an oncologist, a colon, stomach and liver physician and a pathologist in a multidisciplinary meeting (I3). This would

Table 1: Concepts required to calculate quality indicators

Concept	**I1 (lymph nodes)**	**I2 (DSCA)**	**I3 (meeting)**	**I4 (reoperation)**
patient (SCT_116154003)	x	x	x	x
associated with (SCT_47429007)	x	x	x	x
lymph node exam. (SCT_284427004)	x			
lymph node examination date (date)	x			
number of examined lymph nodes (int)	x			
radiotherapy (SCT_108290001)	x			
radiotherapy date (date)	x			
pr. colon cancer (SCT_93761005)	x	x		x
pr. rectum cancer (SCT_93984006)		x	x	x
colectomy (SCT_23968004)	x	x		x
colectomy date (date)	x	x		x
resection rectum (SCT_87677003)		x	x (plus subconcepts)	x (plus subconcepts)
resection rectum date (date)		x	x	x
delivered to DSCA (boolean)		x		
multidisc. meeting (SCT_312384001)			x	
multidisc. meeting date (date)			x	
re-operation (SCT_261554009)				x
re-operation date (date)				x
polypectomy (SCT_82035006)				x
discharge date (date)				x

in principle be possible, but we argue that it is unrealistic to expect that meeting protocols document the presence of individual persons. Another concept that we did not implement is the definition of re-intervention. We employed the SNOMED CT concept "Reoperation" instead, and defined that it must be associated to the same carcinoma that the first operation was associated to.

We noticed a considerable variability in the natural language descriptions of the indicators contained in our test set. For example, all carcinomas should be primary and not recurrent. This is expressed in four different ways for four different indicators: I1) resection of a primary colon carcinoma (numerator and denominator); Exclusion criterion: recurrent colon carcinomas, I2) only count primary carcinomas (numerator and denominator), I3) Exclusion criterion: recurrent rectum carcinomas, I4) Inclusion criterion: Primary colorectal carcinoma = first presentation of a colorectal carcinoma (thus not recurrent); might be the second or next primary presentation.

We encountered several ambiguities and conclude that the expertise of a domain expert is indispensable during the formalisation process.

Table 2: Numbers of SPARQL filters required to calculate quality indicators

Filter	**I1 (lymph nodes)**	**I2 (DSCA)**	**I3 (meeting)**	**I4 (reoperation)**
Temporal Constraints (step 3)	4 (operation within reporting year; examination after colectomy; previous radiotherapy)	2 (operation within reporting year)	3 (operation within reporting year; meeting before resection)	5 (operation and reoperation within reporting year; operation before reoperation)
Numeric Constraints (step 4)	1 (number lymph nodes examined)	-	-	-
Boolean Constraints (step 5)	-	1 (data delivered to DSCA)	-	-
Exclusion Criteria (step 6)	1 (no previous radiotherapy)	-	3 (excluded TEM concepts)	4 (excluded TEM concepts and polypectomy)

Another observation is that many concepts occur in several indicators (e.g. colectomy), but there are also concepts that only occur in one indicator (e.g. lymph node examination). Table 1 shows the concepts and data items required to calculate the numerators (and thus also the denominators) contained in our quality indicator set. Similarly to the concepts, some filter patterns occur in all indicators, and others are indicator-specific. Table 2 gives an overview of the numbers of constraints that are required to calculate the numerators of the indicators. We conclude that many patterns can be re-used once they are created.

2.4 GENERATION OF DATA FOR ALL INDICATORS

We generated synthetic patient data in order to be able to test our formalised queries. It consists of an OWL schema that describes the data needed to calculate the exemplary indicators (TBox, i.e. terminological background knowledge), and the patient data (ABox, i.e. knowledge about individuals). We generated both the OWL schema and the patient data in OWL 2 with the OWL API [28]. Figure 1 shows the OWL schema. We deliberately kept this model as simple as possible (it consists of 25 axioms), and it reflects the information model as employed by the SPARQL queries. The OWL classes "Patient", "Procedure", "Disease" and "Examination of lymph nodes" are SNOMED CT concepts. In the schema, the classes are represented by their SNOMED CT identifiers, e.g. `sct:SCT_116154003` for "Patient". We also added the SNOMED CT concepts "Primary malignant neoplasm of colon" , "Secondary malignant neoplasm of colon", "Primary malignant neoplasm of rectum" and "Secondary malignant neoplasm of

rectum", which are all Diseases, and the Procedures "Colectomy", "Resection of rectum", "Radiation oncology AND/OR radiotherapy", "Multidisciplinary assessment" and "Reoperation".

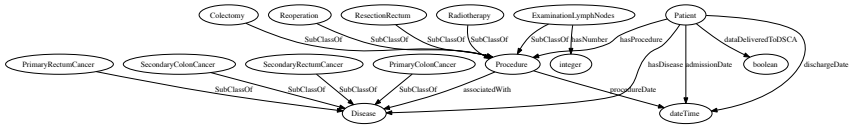

Figure 1: OWL Schema

The data generator generates an arbitrary number of patients as instances of the OWL Class "Patient". All generated patients are colon cancer (50 percent) or rectum cancer (50 percent) patients who underwent colectomy or resection of rectum during a random operation date within the years 2009 to 2011 (we assume that the reporting year is 2010). The malignant neoplasm is primary in 50 percent of the cases, otherwise it is secondary. All generated rectum cancer patients receive a random subclass of the SNOMED CT concept "Resection of rectum" as procedure. The data generator retrieves those subclasses with the help of FaCT++ [29]. Examples are "Stapled transanal resection of rectum" or "Wedge resection of rectum". Patients are admitted to the hospital one day before the operation and discharged between 1 and 60 days after the operation. 10 percent of the patients are re-operated between 1 and 60 days after the first operation. A patient has a lymph node examination with a probability of 50 percent at a random date within 60 days after the operation, with a random number (between 1 and 20) of examined lymph nodes. With a probability of 20 percent, the patient received radiotherapy at a random date within 60 days before the operation. Rectum cancer patients are discussed in a multidisciplinary meeting at a random date within 60 days before the operation with a probability of 80 percent and for all patients, data is sent to the DSCA with a probability of 90 percent. The defined temporal constraints result in radiotherapy always taking place before a lymph node examination, and a multidisciplinary meeting always before the operation. All probabilities are chosen arbitrarily.

Figure 2 shows an exemplary generated patient, and Figure 3 an extract of the same patient in OWL Functional Syntax. The data generator produces around 15 triples per patient, thus our ABox for 300,000 patients consists of over 4 million triples (4,530,578).

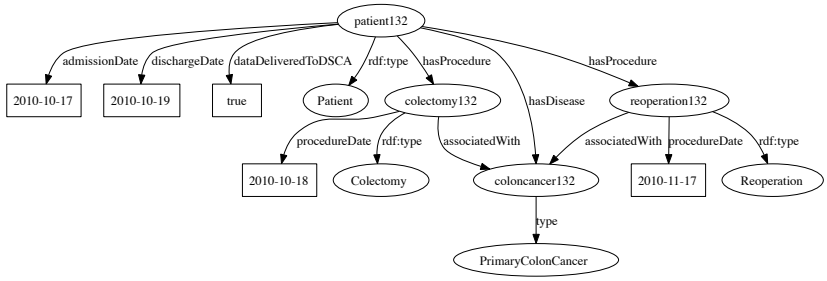

Figure 2: Synthetically Generated Patient Data

```
Declaration(NamedIndividual(data:patient132))
ClassAssertion(sct:SCT_116154003 data:patient132)
ObjectPropertyAssertion(ehrschema:hasDisease data:patient132 data:coloncancer132)
ObjectPropertyAssertion(ehrschema:hasProcedure data:patient132 data:reoperation132)
ObjectPropertyAssertion(ehrschema:hasProcedure data:patient132 data:colectomy132)
DataPropertyAssertion(ehrschema:admissionDate data:patient132 "2010-10-17T05:49:20+02:00"^^xsd:dateTime)
DataPropertyAssertion(ehrschema:dataDeliveredToDSCA data:patient132 "true"^^xsd:boolean)
DataPropertyAssertion(ehrschema:dischargeDate data:patient132 "2010-10-19T05:49:20+02:00"^^xsd:dateTime)
```

Figure 3: Synthetically Generated Patient Data in OWL Functional Syntax

2.5 EXPERIMENTAL RESULTS

In this section, we present our experimental results with respect to the calculation of the formalised indicators, i.e. the execution of the SPARQL queries against the generated patient data. We derived the closure of SNOMED CT with CB [30], the fastest reasoner currently available for this nomenclature [31]. Then, we loaded the closure, our OWL schema and the patient data into BigOWLIM 3.5 [32], which is optimised for fast SPARQL evaluation and was allowed a maximum of 6GB memory. We employed openRDF Sesame 2.4 [33], which supports SPARQL 1.1[4] query features such as expressions, aggregates and negation.

We ran two queries per indicator: one for the numerator and one for the denominator. For the construct "number of re-interventions during the same admission or during 30 days after the resection (choose longest interval)" (I4), we chose to filter the final results from the results returned by the query and measured the runtime including this filtering. Table 3 shows the number of retrieved patients for the numerators and denominators of our queries, and the calculated percentage for each indicator. The last two rows of the table contain the runtimes for the queries, averaged over 100 runs. All queries are processed within seconds. As the calculation of quality indicators is not time-critical, the runtimes are acceptable.

4 http://www.w3.org/TR/sparql11-query/

Table 3: Number of results and runtimes in seconds

Data Item	I1 (lymph nodes)	I2 (DSCA)	I3 (meeting)	I4 (reoperation)
numerator	5,449	44,878	17,439	2,713
denominator	9,898	49,848	21,807	49,848
percent	55%	90%	80%	0.5%
runtime numerator	14.28	25.12	17.74	9.88
runtime denominator	15.90	25.71	15.43	41.36

We checked whether the experimental results are correct by comparing them to the results that we expected based on the probabilities that were used for data generation. For example, the DSCA indicator applies to primary colon and rectum cancer patients, i.e. 50% of our population (150,000). One third of these patients (50,000) is expected to have been operated in 2010, and 90% of the data is sent to the DSCA (45,000). The corresponding query retrieved 44,878 patients, which is comparable. Also the percentages are consistent: for example, the data generator produced a random number between 1 and 20 examined lymph nodes, and 55% of the examinations inspected 10 or more lymph nodes. The fact that we obtained consistent results within acceptable time based on the formalised SPARQL queries and synthetically generated patient data proves the concept and shows that the queries are well-formalised.

2.6 RELATED WORK

2.6.1 *Formalisation of Quality Indicators*

In the following, we discuss a method to formalise goals [34], and a formalisation method for clinical rules [35]. As they do not consider numerators and denominators and in- and exclusion criteria, which are the core elements of quality indicators, neither of the methods is directly applicable to our use case. Thus, we follow our own approach (Section 2.3) that re-uses steps of these methods wherever applicable. Both methods are gradual, and we believe that this is essential in order to preserve the clinical intent of indicators during their formalisation.

Stegers et al. [34] propose a 5-step method to translate goals (e.g. quality indicators) from natural language to the formalism of a verification tool. A domain expert is involved to guarantee the correctness of the result. The authors contribute a conceptual goal model, which serves as a common frame of reference for all involved experts and can be expressed

in a formal language. Their method consists of the following steps: 1) Reduction: explicitly describe the clinical intent of the indicator. 2) Normalisation: rewrite the goal in terms of the goal model. This disambiguates temporal constraints. 3) Formalisation: transform the structured natural language version to a formalised version in GDL (Goal Definition Language). 4) Attachment: formalise the natural language parts with concepts available in the process model. 5) Translation: transform GDL to the logic of the verification tool. This step should be strictly mechanical.

Elements of the method that we re-use are "Reduction" to make the clinical intent of the indicator explicit, which is needed to construct the denominator from the numerator in step 8) of our method, "Normalisation" in order to disambiguate temporal constraints in step 3) of our method and "Attachment", to encode relevant concepts and define the information model in step 1) and 2) of our method. "Formalisation" and "Translation" are not applicable.

Medlock et al. [35] propose the Logical Elements Rule Method, a 7-step method to transform clinical rules for use in decision support: (1) restate the rule proactively; (2) restate the rule as a logical statement (preserving key phrases); (3) assess for conflict between rules; (4) identify concepts which are not needed; (5) classify concepts as crisp or fuzzy, find crisp definitions corresponding to fuzzy concepts, and extract data elements from crisp concepts; (6) identify rules which are related by sharing patients, actions, etc.; (7) determine availability of data in local systems.

We re-use step (1) "restate the rule proactively" to make the clinical intent of the indicator explicit in step 8) of our method and step (5) "classify concepts as crisp or fuzzy, ..." to encode concepts, although we do not differentiate between crisp and fuzzy concepts, in step 1) of our method. Steps (3) "assess for conflict between rules" and (6) "identify rules which are related by sharing patients, actions, etc." relate several indicators. Because indicators are typically calculated independently from each other, these steps are not needed for our application scenario. Step (2) "restate the rule as a logical statement" is similar to step 6) of our method, which groups constraints by boolean connectors. Additionally, exclusion criteria are negated, and the elements of our SPARQL query are connected by logical conjunctions. Our method does not contain a step (4) "identify concepts which are not needed", as non-needed concepts do not need to be encoded. We consider step (7) "determine availability of data in local systems" to be part of the calculation of an indicator.

2.6.2 *Calculation of Quality Indicators*

Once an indicator has been formalised, it can be calculated based on patient data. Previous attempts to automatically calculate quality indicators

include [36] and [37]. The main conclusion of [36] is that for automated chart reviews, more fully-structured and coded data would have to be entered by physicians. As we generate synthetic patients, we do not encounter this problem. The authors of [37] present a rule-based Analytics Engine that is capable of interpreting documents in the Health Quality Measures Format (HQMF)[5] and generating reports. HQMF is a machine-processable standard for representing health quality measures as electronic documents (eMeasures).

2.6.3 *Indicators and Eligibility Criteria*

In- and exclusion criteria are referred to as eligibility criteria [38] and are commonly employed not only for quality indicators, but also for protocols, guidelines, and clinical studies and trials. In the following, we describe two methods for clinical trial recruitment [39], [40] that are based on Semantic Web technologies. Similar to our approach, both methods employ a terminology. In contrast to our approach, they rely on SWRL or description logic queries instead of SPARQL. Besana et al. [39] showed that the automatic recruitment of patients who meet eligibility criteria of clinical trials is possible based on OWL and SWRL, the Semantic Web Rule Language[6]. They use the NCI ontology to represent both patient data and the eligibility criteria. Patel et al. [40] demonstrated that clinical trial criteria can be formulated as description logic queries, which a reasoner can use together with SNOMED CT to infer implicit information that results in retrieving eligible patients.

2.7 FUTURE WORK

As we worked with arbitrary probabilities, the data produced by our data generator is not representative. With the help of a domain expert, it might have been possible to generate more meaningful clinical data. Furthermore, the use of self-generated data leads to avoiding common problems such as insufficient data quality and missing as well as irrelevant data items, but with respect to the difficulty of obtaining (large amounts of) real patient data we consider it to be useful to calculate first indicators as a proof of concept. In the future, we will work with real patient data that stems from several sources.

Our set of four exemplary quality indicators is not representative either. We will work with a larger, more diverse set of indicators in the future

5 http://www.hl7.org/v3ballot/html/domains/uvqm/uvqm.html
6 http://www.w3.org/Submission/SWRL/

in order to further investigate the generalisability of our method. Another open question is whether quality indicators released in natural language are precise enough to be formalised. We will cooperate with domain experts in order to answer this question and to ensure that the clinical intent of the quality indicator is preserved during its formalisation.

2.8 CONCLUSIONS

We presented a 8-step method that is inspired by previously proposed methods [34], [35] to formalise quality indicators as SPARQL queries. The steps are: 1) to encode relevant concepts from the indicator by concepts from a terminology, 2) to define the information model, and 3) to 5) to formalise temporal, numeric and boolean constraints as SPARQL FILTERs. Step 6) is to group constraints by boolean connectors, step 7) to identify exclusion criteria and step 8) to identify constraints that only aim at the numerator, in order to construct the denominator by removing these constraints. Applying this method, we succeeded in formalising a set of four quality indicators into SPARQL queries.

We encountered one construct that is not directly expressible in SPARQL. Although this limits interoperability, the problem can be circumvented. We found a high coverage of SNOMED CT with respect to the colorectal cancer domain. We noticed variability and ambiguity in the original descriptions of the quality indicators and conclude that a domain expert is indispensable to ensure the clinical correctness of the formalised indicators. Finally, we observed that many concepts and filter patterns can be reused once they are formalised.

We proved the concept by running the SPARQL queries that resulted from the formalisation process against self-generated data that consisted of 300,000 synthetically generated patients, and retrieved results that are consistent with the generated data in acceptable time. We conclude that semantic queries are a promising step towards the automated calculation of clinical quality indicators.

Appendix: Set of Quality Indicators

The indicators are released by the Dutch healthcare inspectorate and contained in the indicator set for 2011.

I1: Number of examined lymph nodes after resection (process indicator)
Numerator: number of patients who had 10 or more lymph nodes examined after resection of a primary colon carcinoma.

Denominator: number of patients who had lymph nodes examined after resection of a primary colon carcinoma.

Exclusion criteria: Previous radiotherapy and recurrent colon carcinomas

I2: Participation in Dutch Surgical Colorectal Audit (DSCA) (process indicator)
Numerator: number of surgical resections of a colorectal carcinoma situated in colon or rectum (only count primary carcinomas) for which data has been submitted to the Dutch Surgical Colorectal Audit.

Denominator: total number of surgical resections of a colorectal carcinoma situated in colon or rectum (only count primary carcinomas).

I3: Patients with rectum carcinoma who have been discussed in a preoperative multidisciplinary meeting (process indicator)
Numerator: Number of patients with rectum carcinoma who have been discussed in a preoperative multidisciplinary meeting.

Denominator: Number of patients with rectum carcinoma operated in the reporting year.

Inclusion criterion: Patients who have been operated in the reporting year due to a rectum carcinoma.

Exclusion criteria: Transanal Endoscopic Microsurgery (TEM) resections and recurrent rectum carcinomas.

The Dutch Surgical Colorectal Audit states that the presence of a radiologist, a radiotherapist, a surgeon, an oncologist, a colon, stomach and liver physician and a pathologist are required for a preoperative multidisciplinary meeting.

I4: Unplanned re-interventions after resection of a primary colorectal carcinoma (outcome indicator)
Numerator: number of re-interventions during the same admission or during 30 days after the resection (choose longest interval) in the reporting year.

Denominator: total number of primary resections of a colorectal carcinoma during the reporting year.

Inclusion criteria: Primary colorectal carcinoma = first presentation of a colorectal carcinoma (thus not recurrent); might be the second or next primary presentation.

Exclusion criteria: Transanal Endoscopic Microsurgery (TEM); Endoscopic and open polypectomy

This indicator comes with a list of definitions: Resection: surgical removal of colon segment where the colorectal carcinoma is situated. Re-intervention: re-operation in the abdomen or an intervention (possibly radiological) during which a complication in the abdomen is being treated (inclusive percutaneous incision and drainage, drainage via rectum, embolisations of bleedings in the abdomen, etcetera). Admission: the time which the patient spends in a hospital directly after the operation (the same hospital or another one where the patient has been referred to); can be longer than 30 days.

3 THE REPRODUCIBILITY OF CLIF, A METHOD FOR CLINICAL QUALITY INDICATOR FORMALISATION

In order to be able to automatically calculate clinical quality indicators, we have proposed CLIF, a stepwise method for clinical quality indicator formalisation. Quality indicators are used for external accountability and hospital comparison. As clinical quality indicators are computed in a decentralised manner by the hospitals themselves, reproducibility of the formalisation method is essential to ensure the comparability of calculated values. Thus, we performed a case study to investigate the reproducibility of CLIF. Eight participants formalised the same sample quality indicator with the help of a web-based indicator-authoring tool that facilitates the application of CLIF. We analysed the results per step and concluded that the method itself leads to reproducible results. To further improve reproducibility, ambiguities in the indicator text must be clarified and trained experts are needed to encode clinical concepts and to specify the relations between concepts.

3.1 INTRODUCTION

A quality indicator is "a measurable element of practice performance for which there is evidence or consensus that it can be used to assess the quality, and hence change in the quality, of care provided" [4]. Calculated values are used internally to monitor and to improve the quality of delivered care, and externally to support patients and insurance companies in selecting hospitals of high performance. Ideally, clinical quality indicators are published in an unambiguous, standard representation, so that they can be computed automatically and are comparable among different institutions. We have presented CLIF, a stepwise method to formalise quality indicators into queries in [41]. In this paper, we report on a case study that we performed in order to investigate the reproducibility of CLIF. Our main research question was whether several persons who formalise the same quality indicator independently arrive at the same formalisation. We answered this question for each of CLIF's steps. Any discrepancies were analysed to find the underlying cause.

3.2 METHODS

The case study is based on our previously proposed indicator formalisation method CLIF [41], which consists of eight steps. CLIF is applicable to process and outcome indicators expressed as proportions in general, but for testing its reproducibility, we focused on only one evidence-based process indicator defined by the Dutch healthcare inspectorate: "Number of examined lymph nodes after resection of a primary colonic carcinoma". We chose this indicator because it is important in the domain of gastrointestinal oncology and because it is time-consuming to calculate manually as it requires data from several sources. When lymph nodes are examined after resection of a primary colonic carcinoma, at least 10 lymph nodes should be examined, and the indicator measures the proportion of patients for whom this is the case:

> **Number of examined lymph nodes after resection**
> *Numerator:* Number of patients who had 10 or more lymph nodes examined after resection of a primary colonic carcinoma.
>
> *Denominator*: Number of patients who had lymph nodes examined after resection of a primary colonic carcinoma.
>
> *Exclusion criteria*: Previous radiotherapy and recurrent colonic carcinomas.
>
> *Reporting year*: 2010

We created a *web-based indicator-authoring tool* to facilitate the formalisation process by leading users through the method step by step. The formalisation is performed against a problem-oriented information model with the central concepts "diagnosis" and "procedure". The final result of the formalisation process is a query that is based on the information model. Our *test group* consisted of eight Master students in Medical Informatics. In an initial session, they were introduced to quality indicators, CLIF, the information model of our problem-oriented patient record and to SNOMED CT. They were trained on how to use the web-based tool and on how to search for SNOMED CT concepts in Snow Owl[1].

Reference Standard. We developed a reference standard to measure the quality of the results of our participants. We studied the literature on which the indicator is based, consulted the institution that developed the indicator and organised a consensus meeting with medical informatics experts and clinical domain experts. Table 4 shows the steps of CLIF and the developed reference standard.

1 http://www.b2international.com/portal/snow-owl

Table 4: Steps of CLIF and the reference standard for the sample indicator.

Step	CLIF	Reference standard for the sample indicator
1)	Extract clinical concepts (e.g., diagnoses, procedures) from the indicator text. Search for matching concepts in a medical terminology using standard terminology browsing tools.	Table 5 shows the five relevant concepts from the indicator text with their correct encodings (emphasised). For example, the procedure "lymph nodes examined" from the indicator text is encoded by the SNOMED CT concept "Examination of lymph nodes".
2)	The SNOMED CT concepts from *step 1* need to be related to the concepts of the information model. Finally, the relations between assigned concepts of the information model are defined.	All five SNOMED CT concepts encoded in *step 1* have to be assigned to the correct concepts of the information model, i.e. SNOMED CT finding/disease concepts to the database table "diagnosis", and procedure concepts to the database table "procedure". To maintain a problem-oriented information model, all procedures should be related to diagnoses: lymph node examination, colectomy and radiotherapy have to be related to the diagnosis primary colonic carcinoma. The concept containing the number of examined lymph nodes should be related to the procedure lymph node examination.
3)	Temporal constraints are formalised.	The reporting year 2010 needs to be defined and related to "lymph node examination", as this is the central procedure of the indicator. We expect two constraints to define that the lymph node examination has been after the start and before the end of the reporting year. We also expect two constraints that formalise the constructs "lymph nodes examined *after* resection" and "previous (i.e. *before* the colectomy) radiotherapy".
4)	Numeric constraints are formalised.	The only numeric constraint in the indicator is that the number of examined lymph nodes must be greater than or equal to 10.
5) & 6)	In *step 5*, Boolean constraints are formalised. In *step 6*, Boolean connectors can be used to group constraints.	There are no Boolean constraints in this indicator, and no constraints that have to be grouped by Boolean connectors.
7)	Exclusion criteria are defined.	Here, "radiotherapy", "recurrent colonic carcinoma" and the temporal constraint for "previous radiotherapy" have to be excluded.
8)	Constraints that only aim at the numerator are identified.	There is one constraint that only aims at the numerator: the numeric constraint that expresses that the number of examined lymph nodes should be higher than or equal to 10.

The reliability of agreement between the participants for encoding the concepts in SNOMED CT is measured as Fleiss' kappa and calculated in R.

3.3 RESULTS

Figure 4 visualises the quality of the participants' solutions in terms of adherence to the reference standard per step.

Step 1) All participants intended to encode exactly the five concepts contained in the reference standard. Seven of the eight participants entered five SNOMED CT concepts, and one entered four. The participants entered

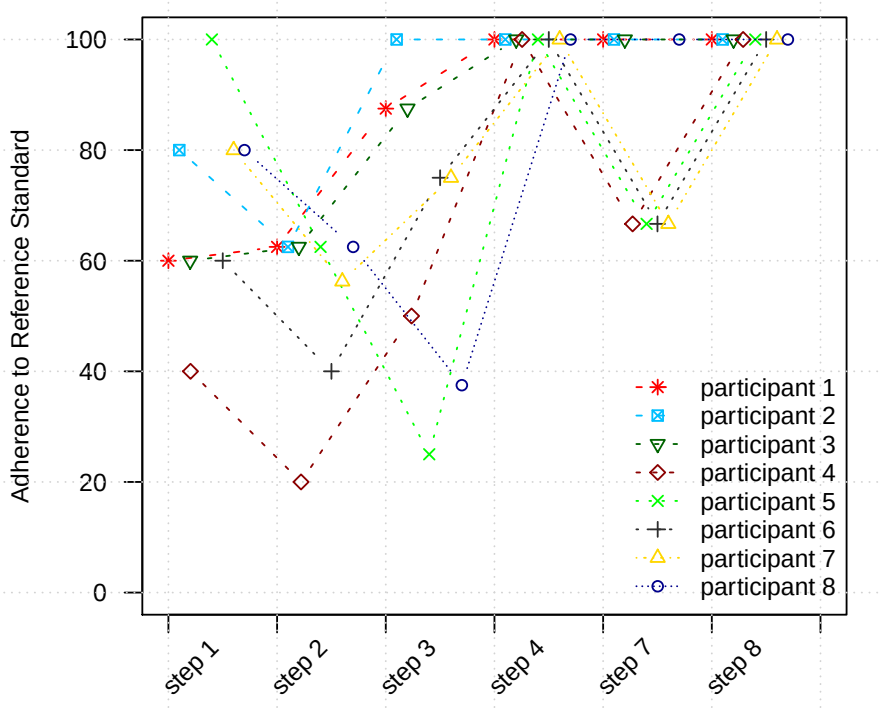

Figure 4: Quality of participants' solutions in terms of adherence to the reference standard. Each participant can reach up to 100 per cent for each step: To quantify the participants' solutions in *step 1*, each encoded concept that meets the reference standard receives 20%. In *step 2*, each correctly assigned concept receives 10%. Correct relations receive 12.5%, and solutions that use un-assigned concepts of the information model receive 6.25%. We do not penalise unnecessary relations. For *step 3*, all correct constraints receive 25% and the questionable ones 12.5%. Participants who formalised the numeric constraint in *step 4* reach 100%. Each constraint correctly excluded in *step 7* receives 33.33%, and participants who identified the constraint that only aims at the numerator in *step 8* receive 100%.

9 different SNOMED CT concepts to encode these 39 concepts. All entered SNOMED CT concepts are subclasses of the two SNOMED CT concepts disease and procedure. Table 5 gives an overview. The reliability of agreement between the participants for encoding these concepts in SNOMED CT is 0.754 ($p < 0.01$) according to Fleiss' kappa. This can be interpreted as substantial agreement.

Step 2) 36 out of 39 SNOMED CT concepts from *step 1* have been related to the correct concepts of the information model. Regarding the second substep, six of the eight participants related the colectomy to the primary colonic carcinoma. No participant has entered the three remaining relations contained in the reference standard.

Table 5: SNOMED CT concepts encoded by test persons.

Indicator Text	**(Number of Participants) SNOMED CT Concept**	**Comment**
lymph nodes examined	(8) *Examination of lymph nodes*	Correct according to reference standard.
resection of a primary colonic carcinoma	(8) *Colectomy*	Correct according to reference standard.
radiotherapy	(7) *Radiation oncology AND/OR radiotherapy*	Correct according to reference standard.
	(1) Radiation therapy procedure or service	Subconcept of correct concept. Contains only unreasonable subconcepts (e.g. "Disposal of radioactive source").
primary colon carcinoma	(5) Carcinoma of colon	Subconcept of correct concept. Defined via the associated morphology "Carcinoma, no subtype", which does not include specific carcinomas (e.g. adenocarcinoma) that should be included.
	(3) *Primary malignant neoplasm of colon*	Correct according to reference standard.
recurrent colonic carcinoma	(4) Secondary malignant neoplasm of colon	Sibling of correct concept. Synonym of metastasis in SNOMED CT; not related to recurrence.
	(2) *Local recurrence of malignant tumor of colon*	Correct according to reference standard.
	(1) Recurrent basal cell carcinoma	Skin carcinoma and thus not correct.

Step 3) Six participants defined the reporting year. Five of them related it to the lymph node examination and one to an undefined procedure. Seven participants formalised the construct "lymph nodes examined after resection", while only four participants formalised the "previous radiotherapy". "Previous" was interpreted two times as having been carried out before the lymph node examination and one time before the colectomy. Another participant defined "previous radiotherapy" as having been performed before the start of the reporting year. This is questionable, as the radiotherapy might have taken place in the reporting year and before the colectomy.

Step 4) Each of our eight participants defined the numeric constraint that we identified in the reference standard.

Step 7) All of the eight participants excluded the assigned concept "radiotherapy". Seven participants excluded "recurrent colonic carcinoma", and one excluded "carcinoma of colon". The participants also excluded all four temporal constraints that refer to "previous radiotherapy" and that have been formalised in *step 3*.

Step 8) All participants correctly identified the numeric constraint as only aiming at the numerator.

3.4 DISCUSSION

We found that our eight participants could use CLIF reproducibly to formalise a sample quality indicator. For *step 1*, we concluded that detecting diagnoses and procedures in natural language text is a reproducible task. In contrast, encoding these concepts can lead to varying results. This task is complex due to the large size of medical terminologies. For example, SNOMED CT contains more than 311,000 hierarchically organised concepts, with many similar, interrelated concepts, making it hard to choose among them. Tools are required to support users in selecting the correct concepts. For *step 2*, we concluded that assigning the concepts to the information model is reproducible. However, our participants did not relate the assigned concepts of the information model as intended. This is due to insufficient knowledge of the employed information model. The reproducibility of *step 3* was lower than expected. This can be ascribed to the ambiguities of the indicator: it is not clear which events should occur in the reporting year and which event(s) the radiotherapy should precede. *Step 4* and *step 8* were reproducible in our case study. In *step 7*, all participants who defined the constraint for "previous radiotherapy" also excluded it. In conclusion, CLIF itself leads to reproducible results, but the difficulty of encoding clinical concepts, defining relations between assigned concepts of the information model and ambiguities in the indicator text have a negative impact on its reproducibility.

Limitations. The main limitation of our study is that we only worked with one quality indicator, which did not require two (steps 5 and 6) out of CLIF's eight steps. Likewise, more participants would have been preferable. Finally, our results might have been biased by the choice of participants.

Related Work. Four clinical guidelines have been encoded into an early version of GLIF, the GuideLine Interchange Format, by two encoders each. The authors found that "different individuals produced different encodings as a result of different modeling choices, different representations of criteria given the use of narrative text in the current version of GLIF, and selection of different terminology for data elements in the absence of standards for clinical vocabulary and data models" [42]. We removed some of these obstacles in our case study: the authoring tool restricts possible formalisations, and we employed a standard terminology together with a common basic problem-oriented information model. Please note that later versions of GLIF adopt both standard terminologies and data models. An evaluation of the cognitive processes used in encoding guidelines in GLIF led to the conclusion that teams consisting of both clinicians and experts in computer-based representations produce better formalisations than individuals of either type working alone [43]. Medlock et al. [35] propose the

7-step Logical Elements Rule Method LERM to assess and formalise clinical rules, which are derived from quality indicators, for decision support. LERM has been validated empirically for inter-user reliability by comparing the results of two assessors who independently applied LERM on 16 rules. LERM was shown to be reliable provided that the users agree on a terminology and on when the rule will be evaluated.

Our main recommendations to increase the reproducibility of CLIF are: institutions that develop quality indicators should publish them together with sets of well-defined concepts from a standard terminology. Likewise, indicators have to be formulated as unambiguously and precisely as possible, so that they can be formalised and computed automatically. This is especially important with regard to temporal relations. The application of CLIF requires the cooperation of clinical domain experts to resolve ambiguities and medical informatics experts who are trained in clinical encoding and in the employed information model.

4 FORMALIZATION AND COMPUTATION OF QUALITY MEASURES BASED ON ELECTRONIC MEDICAL RECORDS

Objective Ambiguous definitions of quality measures in natural language impede their automated computability and also the reproducibility, validity, timeliness, traceability, comparability, and interpretability of computed results. Therefore, quality measures should be formalized before their release. We have previously developed and successfully applied a method for clinical indicator formalization (CLIF). The objective of our present study is to test whether CLIF is generalizable - that is, applicable to a large set of heterogeneous measures of different types and from various domains.

Materials and methods We formalized the entire set of 159 Dutch quality measures for general practice, which contains structure, process, and outcome measures and covers seven domains. We relied on a web-based tool to facilitate the application of our method. Subsequently, we computed the measures on the basis of a large database of real patient data.

Results Our CLIF method enabled us to fully formalize 100% of the measures. Owing to missing functionality, the accompanying tool could support full formalization of only 86% of the quality measures into Structured Query Language (SQL) queries. The remaining 14% of the measures required manual application of our CLIF method by directly translating the respective criteria into SQL. The results obtained by computing the measures show a strong correlation with results computed independently by two other parties.

Conclusions The CLIF method covers all quality measures after having been extended by an additional step. Our web tool requires further refinement for CLIF to be applied completely automatically. We therefore conclude that CLIF is sufficiently generalizable to be able to formalize the entire set of Dutch quality measures for general practice.

4.1 OBJECTIVE

We have previously developed [41, 44, 45] a method for clinical indicator (better known as quality measure) formalization (CLIF). CLIF supports its users in transforming quality measures - which are typically described in unstructured text - into precise queries that can be computed on the basis of patient data. The main envisioned users of CLIF are quality measure developers, but also those responsible for reporting measure results, as well as general practitioners and hospital physicians who are interested in the quality of care they deliver.

CLIF was originally inspired by the Logical Elements Rule Method (LERM) [35], a method used to assess and formalize clinical rules for decision support, as well as a method proposed by Stegers et al [34] to transform natural language into formal proof goals. We have successfully applied CLIF in the limited domain of colorectal cancer surgery to formalize a relatively small set of quality measures. In one of our previous studies [44], we tested whether our method leads to reproducible results. We did this by having eight test subjects - who were previously unacquainted with the problem - formalize a sample measure, and by comparing their results with a reference standard that we developed together with domain experts. The study showed that CLIF can lead to reproducible results, but that unambiguous measures and the cooperation of trained experts with clinical as well as medical informatics expertise are required. The objective of the present study was to test whether CLIF is generalizable - that is, whether it is applicable to a variety of different types of quality measures in various domains.

4.2 SIGNIFICANCE AND BACKGROUND

In recent years, automated reporting of quality measures based on data collected during routine care has become a necessity. The sheer amount of quality measures demanded by governments, patient associations, accreditation organizations, and insurance companies to measure, compare, and improve the quality of delivered care has increased dramatically at a rate that makes their manual calculation unfeasible. Besides being time-intensive, manual calculation is also error-prone and can jeopardize the reproducibility, validity, interpretability, traceability, timeliness, and comparability of quality measure results.

For these reasons, the automated computation and reporting of quality measures is included in the meaningful use of electronic medical records (EMRs), which is currently being put forward by the USA as a national

goal [46]. The non-profit National Quality Forum (NQF) developed the Quality Data Model (QDM[1]), an information model that defines concepts used in quality measures to automate their computation. The Centers for Medicare & Medicaid Services provide a web-based and QDM-driven measure authoring tool (MAT[2]) for quality measure developers to create so-called 'eMeasures'. The MAT is a powerful tool that supports its users by offering a broad variety of functions and features. However, it is not based on a structured method that divides the highly complex task into clear, ordered subtasks.

A formalization method can help to guide users who were previously unacquainted with the problem of measure formalization through the formalization process, and thereby help to ensure that the formalizations obtained faithfully represent the measure's intended meaning. Therefore, we propose our method, CLIF, as a complementary contribution.

4.3 MATERIALS AND METHODS

4.3.1 *Set of quality measures*

To answer our research question, we formalized the entire national set of 159 quality measures for general practice. This set is published in Dutch free text[3]. The quality measures are defined on a national level, so that software providers can support the registration of required data and the reporting of measure results. The set of quality measures is heterogeneous, as it contains measures of various types, and addresses seven domains, such as 'asthma in adults' or 'diabetes mellitus'. Each domain contains a number of subdomains, such as 'HbA_{1c}' or 'smoking'.

Table 6 provides an overview of the quality measures, categorized according to Donabedian's trilogy: structure, process, and outcome [5]. Some measures have complementary measures, which we include. For example, the measure 'Diabetes patients for whom HbA_{1c} has been measured' has the complementary measure 'Diabetes patients for whom HbA_{1c} has NOT been measured'.

The quality measures are released in a narrative-based pseudoformal format, and contain definitions such as 'age >40 and <80' and 'registration date <(reporting date - 1 year)'. The reporting date is defined as the

1 http://www.qualityforum.org/QualityDataModel.aspx
2 https://www.emeasuretool.cms.gov/
3 http://www.nhg.org/themas/artikelen/download-indicatoren, last accessed October 2, 2013

Table 6: Overview of the set of quality measures used. S, P, O and N stand for structure, process, outcome and not specified.

Domain	Measures	Subdomains	Type			
			S	P	O	N
Asthma in adults	13 (8%)	3	3	9	1	0
COPD	14 (9%)	3	3	9	2	0
Cardiovascular risk	23 (14%)	6	3	16	4	0
Diabetes mellitus	50 (31%)	10	0	27	18	5
Depression & anxiety	12 (8%)	0	10	2	0	0
Prevention	15 (9%)	2	4	0	11	0
Prescription	32 (20%)	9	0	27	0	5
All	159 (100%)	33 (26 distinct)	23 (14%)	90 (57%)	36 (23%)	10 (6%)

end of the reporting period, which is typically one reporting year. All quality measures are accompanied by relatively short lists of codes from the classification systems used (between one and 24 concepts per measure; approximately five on average). Two sample quality measures are presented in example 1. Appendix 1 contains additional sample measures.

Example 1 [Two sample quality measures (one process and one outcome measure)]

Process measure 'Percentage of diabetes patients whose HbA_{1c} value has been measured within the previous 12 months'. Definitions:

- Patients younger than 80 years
- International Classification of Primary Care (ICPC) codes for diabetes mellitus: T90, T90.01 or T90.02
- ICPC codes for diabetes mellitus recorded before the end of the reporting period
- Patients registered with general practitioner for 12 months or longer ($\geqslant$ 12 months)
- Code 2206 (main caregiver for diabetes mellitus); latest value for this code must be 48 (for general practitioner); $\geqslant$ 12 months
- HbA_{1c} measurement (code 2816) within the previous 12 months

This process measure is the basis for the outcome measure 'Percentage of diabetes patients whose latest measured HbA_{1c} value was below 53 mmol/mol', which only differs by one additional definition:

- HbA_{1c} value of last measurement below 53 mmol/mol (<53)

4.3.2 *Patient data*

We used an extract of anonymized routine healthcare data from the Julius General Practitioners' Network Database, which consists of administrative routine healthcare data extracted from the information systems of more than 60 primary healthcare centers (one to eight general practitioners per center) in the region of Utrecht, the Netherlands [47]. The administrative routine healthcare data were extracted locally from the general practitioner's EMRs by making use of the Mondriaan Client[4] and anonymized locally through a trusted third party (Custodix). This way, medical information cannot be used outside the practice location to identify individual patients by researchers or anyone else not directly involved in the treatment of the patients. Consultations, episodes, and diagnoses are encoded with International Classification of Primary Care (ICPC) codes, prescribed medications in the Anatomical Therapeutic Chemical (ATC) Classification System, and (laboratory) test results in a national coding system. We used data extracted from the 22 practices that use Promedico, a software system for general practices in the Netherlands. The other practices sharing data in the Julius General Practitioners' Network Database use other EMR software systems. Our database contains data related to 156,176 patients in the years between 2006 and 2011.

4.3.3 *CLIF*

CLIF is a method for formalizing natural-language quality measures as computable queries based on formally defined concepts, information model, and selection criteria. The original version of CLIF [41] consists of eight steps, which are presented together with the formalization of the sample outcome measure in Table 7.

4.3.4 *Web tool*

We have built a web tool[5] that implements CLIF to guide its users through all eight steps and stores the formalized criteria in a dedicated database. To test CLIF's generalizability, we use this web tool as a starting point to formalize the set of quality measures. Importantly, the user can record comments for each step, which is indispensable in cases when a measure is ambiguous and the user needs to decide on how to operationalize it. The user can create so-called query variables (aliases) for database tables,

4 http://www.projectmondriaan.nl/
5 http://clif.mash-it.net; login and password are both 'test'

Table 7: Steps of the original version of CLIF.
CLIF, clinical indicator formalization;
ICPC, International Classification of Primary Care;
QDM, Quality Data Model.

Step	CLIF	Reference standard for the sample indicator
1. Concepts	Extraction of clinical concepts (eg, diagnoses, procedures) from the quality measure text. Depending on the measures and the patient data, standard terminologies such as SNOMED CT [27], ICD or ICPC, or local/national coding systems can be used	The ICPC codes for diabetes mellitus (T90, T90.01, or T90.02), and the national codes for the main caregiver (2206) and HbA_{1c} (2816) are elaborated in the quality measure definition
2. Information model	Binding of concepts from the previous step to the concepts of the information model. Depending on the measures and the patient data, standard information models such as the QDM or openEHR archetypes, or local database schemas can be used	Here, we define query variables (aliases), such as diabetes, for the local database table that stores ICPC entries, and we bind this variable to the three diabetes mellitus concepts identified in the previous step
3. Temporal criteria	Formalization of temporal criteria	The sample measure contains various temporal criteria. Patients must be below 80 years to be included. They must be registered for 12 months or longer, and the general practitioner must have been the main caregiver for 12 months or longer, the diagnosis must be present at the reporting date or before, and the HbA_{1c} value must have been measured within the previous 12 months. Finally, the values of both the code for the main caregiver and the HbA_{1c} measurement must be the latest within the specified time frames
4. Numeric criteria	Formalization of numeric criteria	The sample measure contains two numeric criteria: the value of the code for the main caregiver must be 48 for general practitioner, and the HbA_{1c} value must be below 53 mmol/mol
5. Boolean criteria	Formalization of Boolean criteria	Our sample measure does not contain any Boolean criteria
6. Boolean connectors	Grouping of criteria by Boolean connectors	The three different codes for diabetes are connected by OR. Other criteria are connected by AND
7. Exclusion criteria / negations	Definition of exclusion criteria/negations	Our sample measure does not contain any exclusion criteria/negation
8. Numerator only	Identification of criteria that only aim at the numerator	The difference between the numerator and the denominator is not explicitly defined for this measure. We define it as the HbA_{1c} value being below 53 mmol/mol

and then attach one or more codes (eg, those specified for diabetes) to these variables. In subsequent steps, the user defines which criteria need to be valid for a patient to be included in the quality measure result. To increase usability, the underlying database schema or information model is used to populate options for each step. For example, for temporal criteria,

only database fields that have temporal data types are preselected. Also, criteria are colored (eg, red for exclusion criteria/ negation), and can be deactivated so that they are not included in the automatically constructed Structured Query Language (SQL) query. During or after the formalization process, users can run the query (if the tool is connected to a database). SQL was chosen because of the format of our underlying patient database, but other query languages, such as the SPARQL Protocol and RDF Query Language, or standards-based output formats, such as the Health Quality Measures Format (HQMF), which is used for eMeasures, could also be an option. The screenshots contained in Appendix 2 show how the sample measure is formalized step by step.

4.3.5 *Computation of quality measures*

To compute the formalized quality measures based on our patient data, we automatically constructed SQL queries based on the criteria that are stored in the database of CLIF's web tool. When our web tool did not support a construct, we applied CLIF manually by directly translating the respective construct into SQL. Subsequently, for every measure and reporting year (2007-2011), one query for the numerator and one for the denominator were constructed automatically. These queries were used to compute the measures, and to generate plots for all computed measures to visualize how the percentages develop over the course of the reporting years. The query in Appendix 3 represents the numerator of our sample outcome measure 'Percentage of diabetes patients whose latest measured HbA_{1c} value was below 53 mmol/mol' for the reporting year 2011.

4.3.6 *Evaluation of results*

Apart from assessing the face validity of our computed measure results based on the generated plots, we evaluated our result set computed for the reporting year 2011 for the quality measures in the domain diabetes mellitus, which is the largest domain contained in the measure set. We compared our result set with the result sets computed independently by two other parties for the same reporting year based on a large subgroup of general practices of the Julius General Practitioners Network that are working together in diabetes care. At the request of these practices, one of the measure result sets was provided by an academic institution specializing in the reuse of routine primary care data for research purposes (Integrated Primary Care Information (IPCI)[6]), Rotterdam, the Netherlands.

6 http://www.erasmusmc.nl/med_informatica/research/555688/?lang=en

The other measure result set was provided by a software company specializing in generating management reports based on extracted routine primary care data that are used to support reimbursement of diabetes care with the healthcare insurance companies that pay for it (Proigia[7], Ede, the Netherlands). In both cases, as well as in our own procedure, all data are anonymized at source, in the practices, before it is shared.

The comparison gives a first indication of the comparability of our computed results. However, a strict evaluation of the computed results is not possible because of the absence of a gold standard. As ambiguous quality measure definitions allow different interpretations, it is hard to distinguish right and wrong formalizations. Ultimately, definitions should be based on a broad consensus, and formalization helps to identify open issues and make them explicit. Another hindrance is that we only computed results based on 22 general practices which use Promedico, whereas the other parties computed results based on twice as many practices which use various information systems for general practitioners.

4.4 RESULTS

We formalized the entire set of 159 quality measures, which took, on average, 10 min per measure and resulted in a total of 849 concepts and 1283 criteria, with the help of the web tool. This way, 86% of the quality measures could be formalized fully, while the remaining measures required the manual application of CLIF. We computed all measures except for two numerators that combined a number of other quality measures and had to be canceled because of excessive run times. In the following, we quantify our results according to CLIF's steps. Note that one new step 'Textual criteria' has been added to the original version of CLIF. Table 8 presents an overview of the results.

Step 1: Concepts. In this step, we entered 849 (148 distinct) concepts.

Step 2: Information model. In this step, we defined 465 query variables and connected them to the concepts entered in the previous step. Of the 106 distinct variables, 60 were related to measured (laboratory) values, 33 to ATC medications, and 13 to ICPC diagnoses. In the set of quality measures, entire variables can be negated. For example, one measure asks for all patients whose HbA_{1c} value has not been measured. We implemented this option in the web tool and made use of it 10 times.

Step 3: Temporal criteria. 1068 criteria were temporal. Many quality measures pertain to the latest value of a measurement before the reporting date.

7 http://www.proigia.nl/

Table 8: Overview of results per step

Step	Used	Additionally implemented	Manual formalization
1. Concepts	849 (148 distinct)	-	-
2. Information model	465 (106 distinct) variables	Negated query variables (used 10 times)	-
3. Temporal criteria	1068 (83% of criteria)	Latest value (used 145 times)	-
4. Numeric criteria	206 (16% of criteria)	-	Numeric quantification: 9 (5% of measures)
5. Boolean criteria	-	-	-
6. Textual criteria	-	9 (1% of criteria)	-
7. Boolean connectors	1914 AND; 567 OR (only counted occurrences in numerators)	-	Custom connectors and nesting: 17 (11% of measures)
8. Exclusion criteria	66 (5% of criteria)	-	-
9. Numerator only	714 (56% of criteria)	-	-

We implemented this temporal abstraction in the web tool. This functionality was required 145 times.

Step 4: Numeric criteria. We formalized 206 numeric criteria, all of them simple value comparisons. A number of quality measures included numeric quantification over (temporal) criteria. For example, the 'chronic' intake of a prescribed drug is defined as 'at least three prescriptions or one prescription with a duration of 6 months or longer during the previous 12 months'. Even harder to formalize is the construct 'multiple' chronic intake, which is defined as chronic intake of five or more different drugs. Another criterion that comprises numeric quantification is 'at least two resurgences during the previous 12 months'. As we did not implement this option in our web tool, we manually formalized the respective parts of the nine measures that required numeric quantification over (temporal) criteria.

Step 5: Boolean criteria. Owing to the schema of our database, no Boolean criteria were required.

Step 6: Textual criteria. This step had to be added to CLIF because some data elements - for example, gender and smoking behavior - were stored in a text field in the patient database. Textual criteria were currently required for nine measures. Note that one needs to consider whether textual data elements can be transformed into coded form, in which case the step 'Concepts' would be adequate.

Step 7: Boolean connectors. The step for Boolean connectors is not implemented in CLIF's web tool in a way that users can manipulate them. In the

query generation, the standard connector is 'AND', and we automatically detect groups of criteria that must be combined by 'OR'. This is the case whenever only one value at a time is possible. For example, one entry in the medication database can have only one ATC code. Therefore, when a query variable is assigned to two or more ATC codes, they are automatically combined by 'OR'. The same is the case for value comparisons that are based on mutually exclusive categories. For example, the smoking status cannot be 'yes' and 'never' at the same time. This simple mechanism covered most of the required Boolean connectors. However, exceptions occurred: for example, one of the asthma measures covers patients with persistent asthma OR patients who smoke. Two different query variables must be defined, as these entities must fulfill different criteria. Also, criteria for patients with a valid reason for an absent cervical screening - such as refusal or pregnancy - are to be connected by 'OR'. Finally, custom Boolean connectors can also be applicable to values of codes. For example, because the smoking status must be updated yearly only for (ex-) smokers, the quality measure 'smoking habits known' measures the percentage of patients whose last recorded value for smoking was 'never' regardless of the registration date, OR 'previously' OR 'yes' during the reporting year. Likewise, there is a need to nest previously defined criteria, as required for the construct 'at least three prescriptions OR one prescription with a duration of 6 months or longer during the previous 12 months'. Therefore, the step to combine criteria by Boolean connectors has been performed manually for 17 measures.

Step 8: Exclusion criteria/negations. 66 criteria were marked as exclusion criteria/negations.

Step 9: Numerator only: 714 previously defined criteria only aim at the numerator.

To summarize, we added the step 'Textual criteria' to our method, and extended our web tool by this step as well as the possibility to negate entire query variables in the step 'Information model' and to specify the latest value of a measurement in the step 'Temporal criteria'. The functionalities of numeric quantification over (temporal) criteria and custom grouping of Boolean connectors have not been implemented, and we therefore applied them manually. This enabled us to formalize 100% of the quality measures.

4.4.1 *Evaluation of results*

Figure 5 shows the results for the 43 of the 50 diabetes measures that have been computed by all three parties (the other parties did not compute

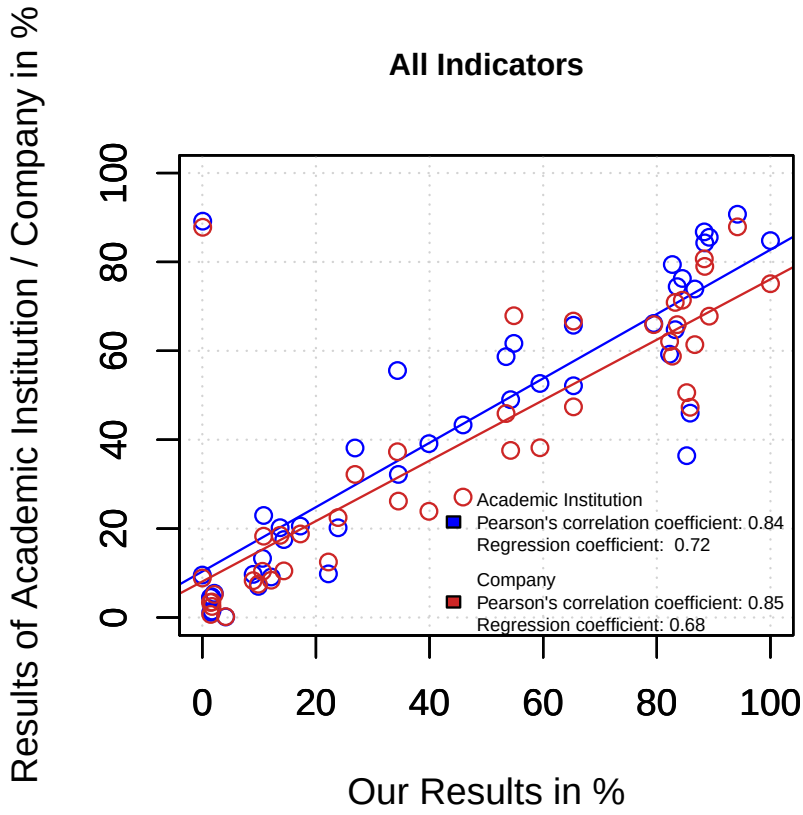

Figure 5: Comparison of our results (in percentages) with the results computed by an academic institution and a commercial company.

complementary measures). Our results are generally higher than the ones computed by the other two parties, with strong correlations according to Pearson's correlation coefficients.

The observed differences are explainable by differences in the approach to computing the quality measures: while we preserved the original measure definitions, the other parties adapted the measure definitions to match the data as much as possible. An example of this is the outlier on the top left, which is due to the fact that, in our dataset, the code specified in the measure narrative for diabetes mellitus type 1 only occurs twice. The problem is probably caused by versioning differences in the ICPC codes used to describe the data in the Promedico system. The other parties adapted the ICPC code from T90.01 to T90.1 to match the data, and thereby included many more patients with diabetes mellitus type 1, while we did

not. This difference may also have influenced the results of other measures, as patients who have diabetes mellitus, or diabetes mellitus type 1 or 2, are the basis for the denominators of the subsequent measures. Also different interpretations and definitions, such as alcohol usage registered as 'ever' instead of 'within the past 5 years', can influence the results. Further differences may be explainable by different approaches to handling missing data, and by different decisions on defining the denominator.

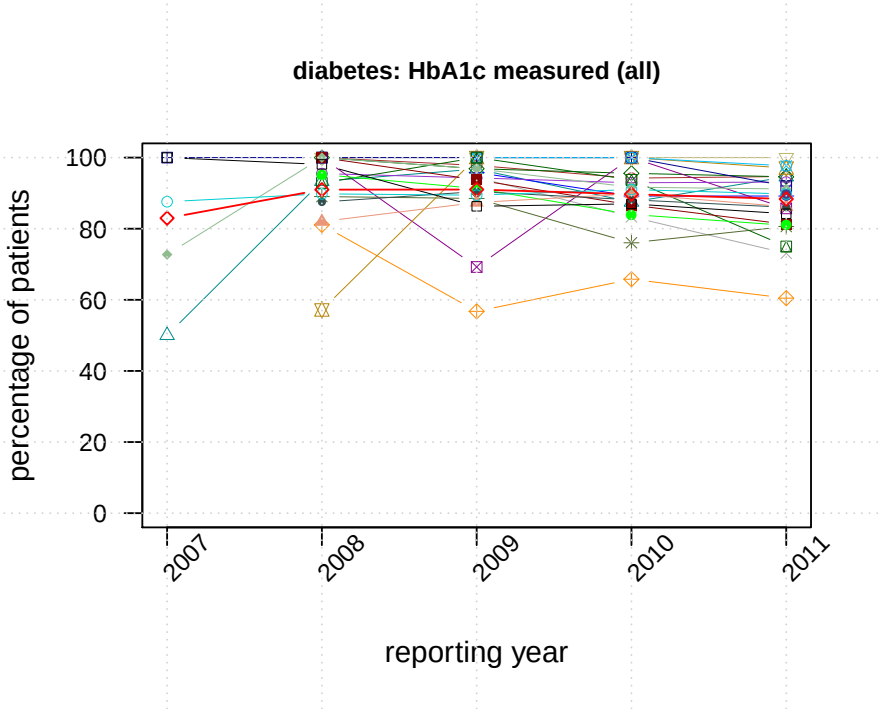

Figure 6: The target value desired by insurance companies is an HbA_{1c} measurement for 95% of the patients with diabetes.

Figures 6 and 7 show plots for the process measure 'Percentage of patients whose HbA_{1c} has been measured' and the outcome measure 'Percentage of diabetes patients whose latest measured HbA_{1c} value was below 53 mmol/mol'. The red line depicts the aggregated percentages for all included practices, and the other lines represent the individual practices. A high but decreasing variability can be observed. The lines in the plots do not suggest a trend but only connect the measurements per reporting year to increase readability.

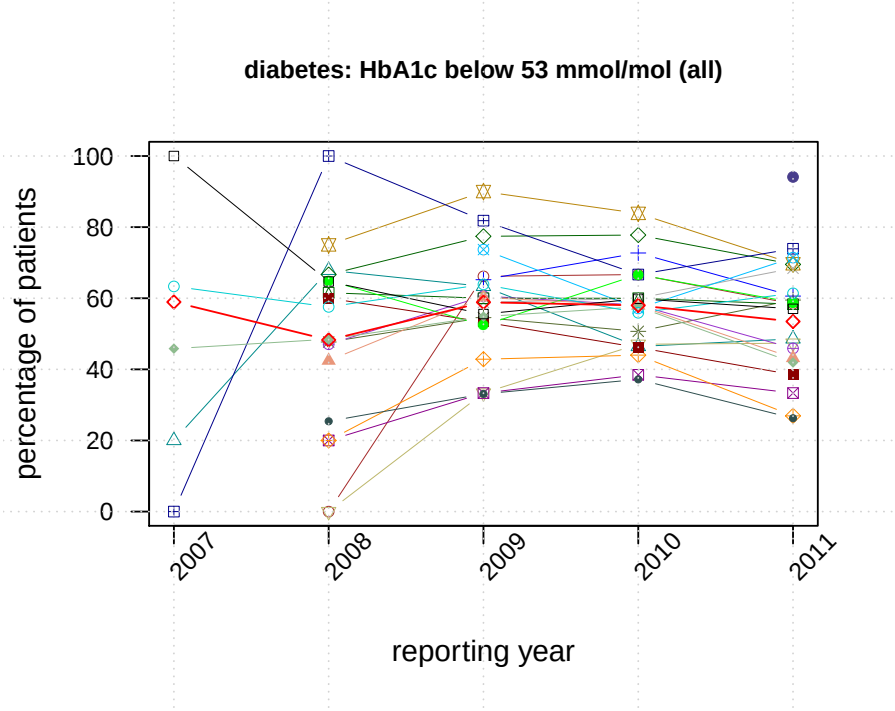

Figure 7: The measured value is best when it is below 53 mmol/mol. However, values up to 69 mmol/mol are acceptable.

4.5 DISCUSSION

After we extended CLIF by the additional step to formalize textual criteria, which was not available in its original version [41], our method covered the formalization of all quality measures. Our web tool, however, required additional functionality. Even though the web tool was not complete, we could apply CLIF manually by directly translating the missing constructs into SQL, enabling us to fully formalize 100% of the measures. This leads us to conclude that CLIF is sufficiently generalizable to be able to formalize the entire set of Dutch quality measures for general practice.

4.5.1 *Observations during our study*

4.5.1.1 *Quality measure definitions*

REPETITION AND REUSABILITY We observed considerable repetition. Many quality measures shared the same denominator, numerators were

used as denominators for subsequent measures, and measures of a number of subdomains such as smoking and body mass index were applicable to a number of domains such as diabetes mellitus and cardiovascular risk. This is advantageous, as the respective criteria only have to be formalized once and can be reused thereafter. Also concepts and query variables can be reused - for example, our measure set made use of 849 concepts, but only 148 distinct ones, so that 701 (82%) have been reused.

AMBIGUITIES Ambiguities in quality measure narratives leave freedom for interpretation, which is especially problematic when values of locally computed measures are compared. Also, results computed for ambiguous measure definitions are hard to assess, as it is unclear what exactly has been computed. Here, a structured formalization method, ideally with tool support, can help users to resolve ambiguities, and to document all steps and decisions. A major issue during the formalization was that in the set of quality measures, it is generally not explicitly stated how the denominators are defined. For example, it was unclear whether the denominator of the outcome measure 'Percentage of diabetes patients whose latest measured HbA_{1c} value was below 53 mmol/mol' is 'Diabetes patients for whom HbA_{1c} has been measured' or 'Diabetes patients'. Discussions with several experts showed that opinions vary about which denominator would be the correct one.

Another problem is the absent definition of qualitative terms such as 'high' dosage, as well as constructs such as 'indication for cervical screening' and 'gynecological intervention affecting the cervix'. Likewise, it is not clear whether medication dates in the measures refer to the prescription date, start date, or dispense date. We presented our findings to those responsible for the measure definitions.

4.5.1.2 *Mismatches between quality measures and data*

We detected several mismatches between quality measure definitions and our data. Some codes that were specified in the measures did not occur in the data. Examples include the ICPC code T90.01 for diabetes mellitus type 1 and the measurement code PAP XP BV for cervical screenings. Likewise, because of the absence of standardized codes, reasons for a cervical screening not being carried out were encoded depending on the underlying EMR, impeding an EMR-independent formalization a priori. Finally, some values should be encoded by 1 for 'yes', but they were stored as 'yes' in the database. The use of a standard information model for both measures and data might help to bridge mismatches between quality measures and data.

4.5.1.3 *Quality measure results*

Quality measure results should be treated with caution, especially when they are used to compare healthcare institutions. Percentages can be similar even if the numbers used to compute them are very different, affecting statistical significance. In our case, the numbers on which the percentages were based typically increased with the reporting years, but this is not evident in the plots. Similarly, although data quality improved over time in our dataset, data quality and missing values can influence the results.

4.5.2 *Related work*

The complexity of eligibility criteria in clinical trials has been analyzed in previous studies [48,49] and seems to be comparable to criteria for quality measures. For example, Conway et al [49] report a 'heavy reliance on nested Boolean logic, complex temporality and ubiquitous (...) codes'. Weng et al [50], as well as Tu et al [51], proposed semi-automated approaches to transforming free-text eligibility criteria into computable criteria. Milian et al [52] addressed the problem of formalizing eligibility criteria and derived a set of patterns that are the basis for a semi-formal representation. A pattern that Ross et al [48] also detected is the 'if-then' construct. Quality measures themselves can be rewritten into 'if denominator then numerator' constructs, and LERM [35] is applicable for such scenarios. With regard to phenotyping algorithms, Thompson et al [53] encountered 'non-Boolean logic' - for example, 'at least two of four criteria must be true', which did not occur in our measure set.

4.5.3 *Limitations*

One of the main limitations of our work is that the use of the tool presumably influenced our study. In retrospective, it is impossible to determine the actual influence of the tool on the formalization process and consequently on the obtained formalizations.

Another limitation of our work is that our results are limited to Dutch quality measures for general practice. However, in this and previous studies [41,44,45], we have formalized a variety of heterogeneous quality measures (structure, process, and outcome measures) for both hospitals and general practitioners in various domains, using a variety of standard and non-standard coding systems and information models. This experience suggests that CLIF might also be sufficiently generalizable to be able to formalize other sets of measures, but the level of complexity of Dutch measures may differ from sets in other countries.

4.5.4 *Future work*

More research may provide further insights into the generalizability of our method - for example, by formalizing international sets of quality measures, such as the meaningful use measures put forward by the USA.

We have shown that openEHR archetypes can facilitate the semantic integration of routine patient data from several sources and patient data and quality measures to automatically compute measures [45]. In the future, it would be interesting to analyze whether new information model standards, such as the QDM as currently used for eMeasures, could be integrated into our approach, and how they would affect its generalizability.

4.6 CONCLUSION

The formalization of quality measures with the help of CLIF forces the user to disambiguate unclear parts of quality measures that are documented in inherently ambiguous natural language, and to precisely define the difference between denominator and numerator. Additionally, a formalized measure ensures that it can be computed automatically, and that the same query is used across several locations to compute a measure, making the computed results reproducible, comparable, traceable, and interpretable. Therefore, we propose that quality measures should be released in a formalized form, and ideally based on standard information models and terminologies. CLIF has been shown to be a useful method for achieving this goal.

APPENDIX 1: ADDITIONAL SAMPLE QUALITY MEASURES FROM THE SET OF MEASURES FOR DUTCH GENERAL PRACTITIONERS

Asthma in adults: Percentage of patients aged 16 years and older, diagnosed with asthma and known to be smokers who received advice to stop smoking.

COPD: Percentage of patients diagnosed with COPD who had 2 or more (new) exacerbations within the previous 12 months.

Cardiovascular risk: Percentage of patients diagnosed with cardiovascular diseases who had a LDL-cholesterol measurement of 2,5 or higher who had no prescription for a lipid-lowering therapy in the group of patients diagnosed with cardiovascular diseases who had a LDL-cholesterol measurement of 2,5 or higher.

Depression & anxiety: Percentage of patients aged 18 years or older with a depressive disorder or depressive feelings who received a prescription for an antidepressant drug.

Prevention: Percentage of women eligible for cervical screening for whom a valid reason for postponement has been registered: refusal or pregnancy or birth and breastfeeding (until 6 months afterwards) or uterus extirpation or gynecological intervention on the cervix or last screening less than one year ago and normal result.

Prescription: Percentage of patients aged 75 years or older who are on multiple chronic medication. Multiple use means 5 or more distinct medications, and chronic more than three prescriptions within a year or a prescription with a length of 6 months or longer.

APPENDIX 2: SCREENSHOTS OF THE CLIF WEB TOOL THAT SHOW HOW THE SAMPLE MEASURE IS FORMALIZED

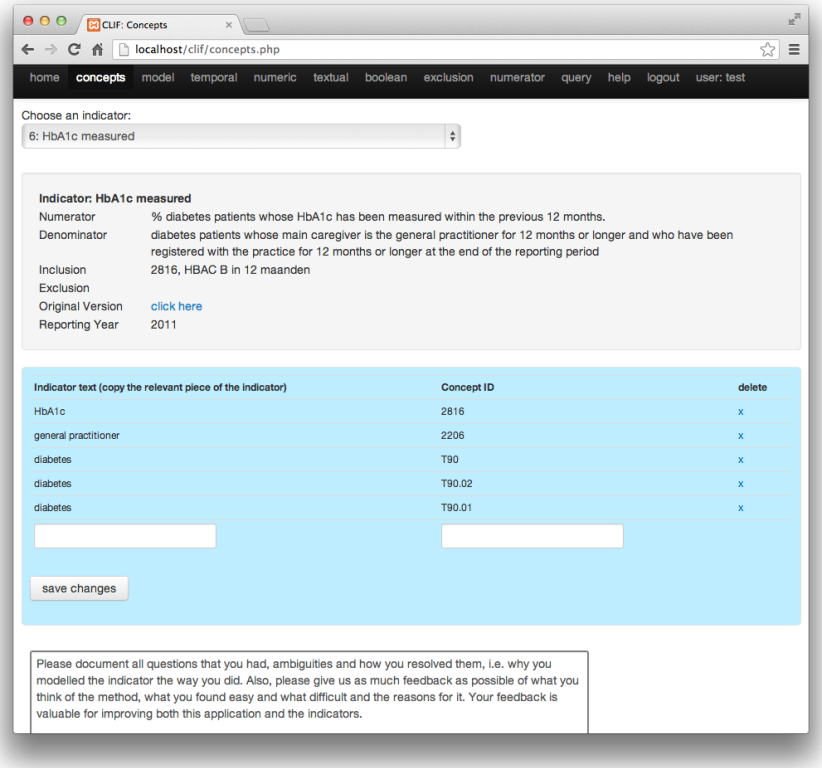

Figure 8: Concepts

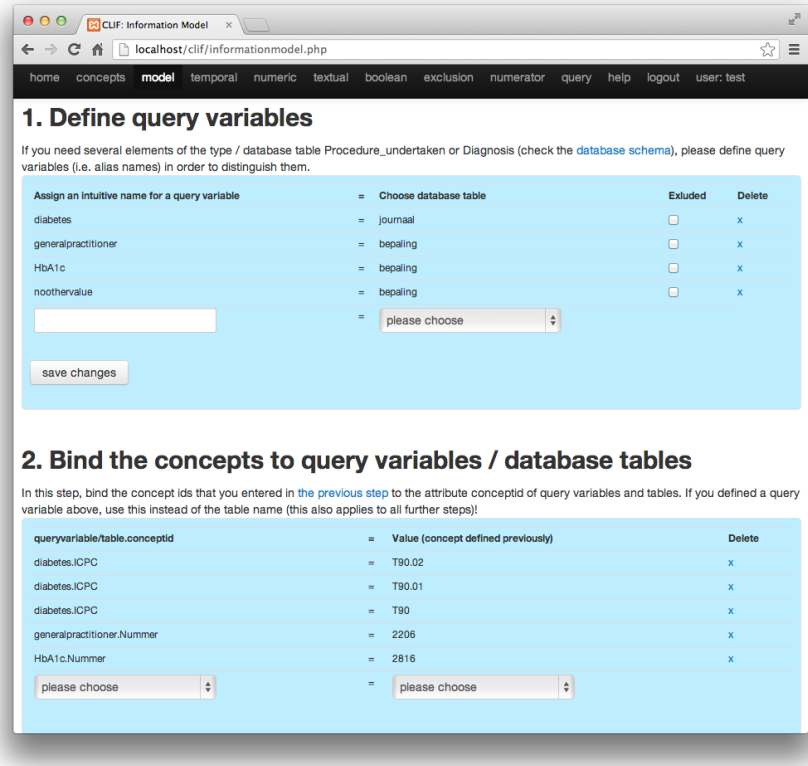

Figure 9: Information Model

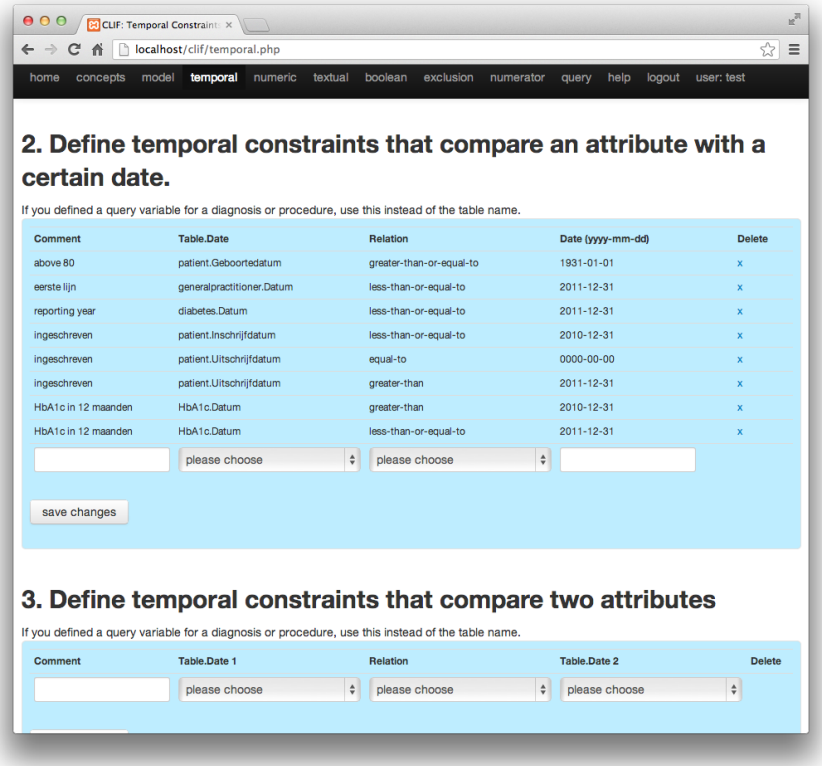

Figure 10: Temporal Criteria

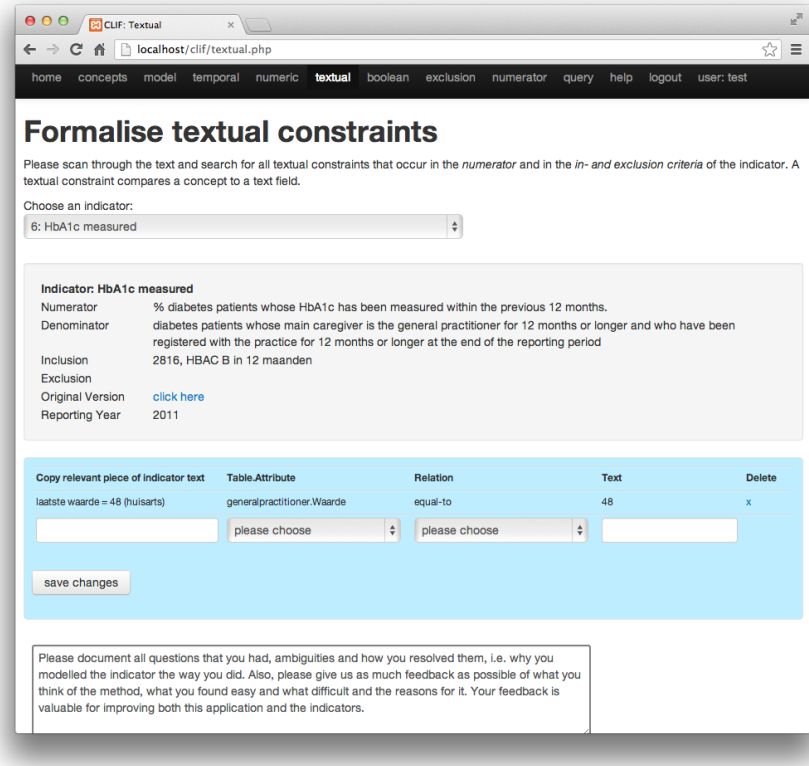

Figure 11: Textual Criteria

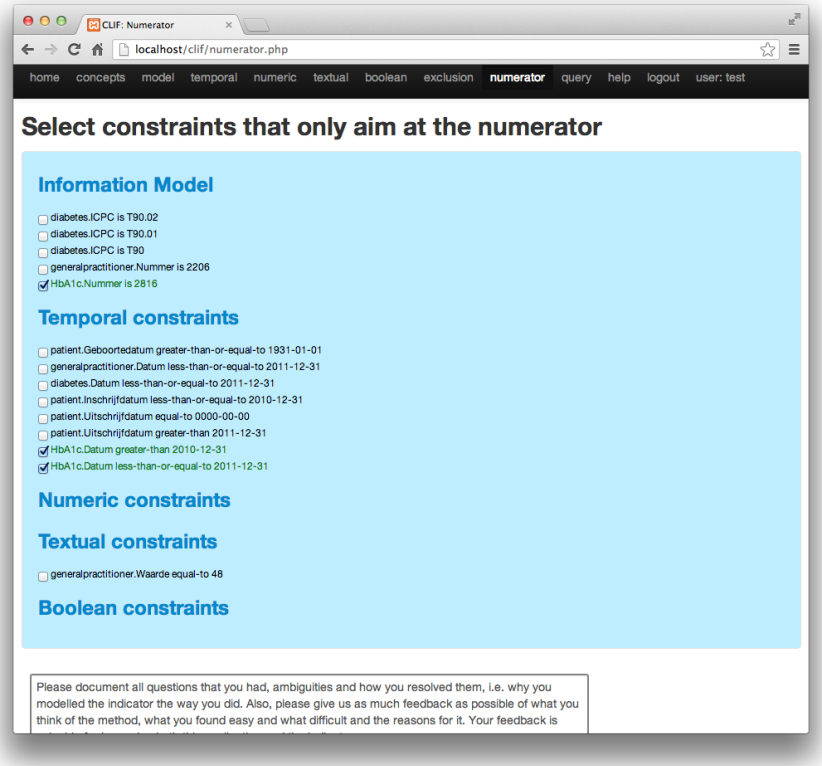

Figure 12: Numerator Only

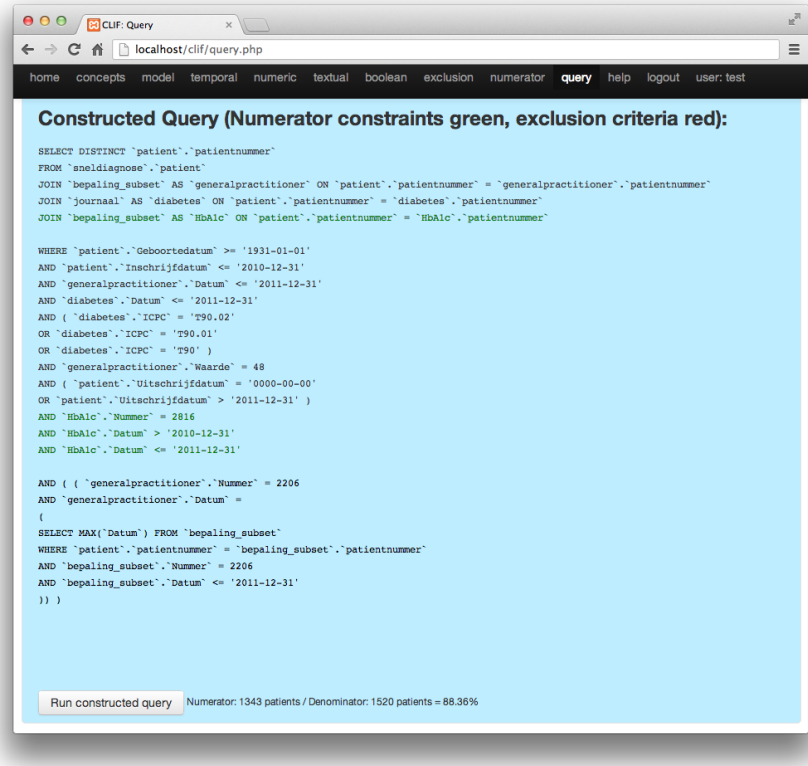

Figure 13: Query

APPENDIX 3: FORMALIZED SAMPLE MEASURE

```
SELECT DISTINCT patient.patientnumber

-- information model
FROM   patient
       JOIN determined_values AS noothervalue
         ON patient.patientnumber = noothervalue.patientnumber
       JOIN determined_values AS generalpractitioner
         ON patient.patientnumber = generalpractitioner.patientnumber
       JOIN encounter AS diabetes
         ON patient.patientnumber = diabetes.patientnumber
       JOIN determined_values AS HbA1c
         ON patient.patientnumber = HbA1c.patientnumber
WHERE

  -- concepts
  ( diabetes.icpc = 'T90.02'
     OR diabetes.icpc = 'T90.01'
     OR diabetes.icpc = 'T90' )
  AND generalpractitioner.code = 2206
  AND HbA1c.code = 2816

  -- temporal
  AND patient.dateofbirth >= '1931-01-01'
  AND patient.registrationdate <= '2010-12-31'
  AND generalpractitioner.date <= '2011-12-31'
  AND diabetes.date <= '2011-12-31'
  AND ( patient.deregistrationdate = '0000-00-00'
        OR patient.deregistrationdate > '2011-12-31' )
  AND HbA1c.date > '2010-12-31'
  AND HbA1c.date <= '2011-12-31'
  AND generalpractitioner.date = (SELECT Max(date)
                                  FROM   determined_values
                                  WHERE  patient.patientnumber =
                                         determined_values.patientnumber
                                         AND determined_values.number = 2206
                                         AND determined_values.date <=
                                             '2011-12-31')
  AND HbA1c.date = (SELECT Max(date)
                    FROM   determined_values
                    WHERE  patient.patientnumber =
                           determined_values.patientnumber
                           AND determined_values.nummer = 2816)

  -- numeric
  AND generalpractitioner.value = 48

  -- numerator only (also numeric)
  AND HbA1c.value < 53
```

Part II

SECONDARY USE OF PATIENT DATA

5 BARRIERS TO THE REUSE OF ROUTINELY RECORDED CLINICAL DATA: A FIELD REPORT

Today, clinical data is routinely recorded in vast amounts, but its reuse can be challenging. A secondary use that should ideally be based on previously collected clinical data is the computation of clinical quality indicators. In the present study, we attempted to retrieve all data from our hospital that is required to compute a set of quality indicators in the domain of colorectal cancer surgery. We categorised the barriers that we encountered in the scope of this project according to an existing framework, and provide recommendations on how to prevent or surmount these barriers. Assuming that our case is not unique, these recommendations might be applicable for the design, evaluation and optimisation of Electronic Health Records.

5.1 INTRODUCTION

Today, increasing volumes of clinical data are being routinely recorded and stored in Electronic Health Records (EHRs). The potential benefit from reusing the resulting data sources is enormous, both for individual patients and society in general. In fact, according to a recent report by PricewaterhouseCoopers [1], using data for secondary purposes is one of the most promising ways to improve health outcomes and costs. Such purposes comprise clinical research, the recruitment of eligible patients for clinical trials, the early detection of epidemics, reimbursement, clinical audit, the generation or testing of medical hypotheses and quality monitoring or reporting based on clinical quality indicators. However, reusing clinical data is often challenging in practice.

The Dutch government releases sets of both legally mandatory and voluntary evidence-based quality indicators for various kinds of diseases and interventions. The government requests indicator results for entire reporting years to monitor and compare the quality of care. These indicators typically require data from several sources and are often computed manually, which is error-prone and time-consuming. To enable timely feedback and, where necessary, intervention inside hospitals, the indicators should be computed automatically and in real-time, based on routinely recorded

clinical data. For a recent study, we strove to gather all raw source data required to compute a set of indicators for the Gastrointestinal Oncology Centre Amsterdam (GIOCA).

The GIOCA is a specialised outpatient clinic that has been set up to improve the quality of care for patients with (suspected) cancer of the gastrointestinal tract. Patients who register at the GIOCA are scheduled for an appointment within only seven days at most. During this appointment, examinations to diagnose the patient are carried out, the case is discussed in a multidisciplinary meeting, and a detailed treatment plan is established and communicated to the patient. As this patient-centred rapid diagnosis process reduces the time until treatment starts, the founders of the GIOCA are motivated to measure its performance. We chose the domain of colorectal cancer surgery because it is also the subject of the recently founded Dutch Surgical Colorectal Audit (DSCA)[1]. The DSCA collects all data items necessary to compute the set of indicators. These data items are currently being entered manually by one of our surgeons, but ideally, they would be pre-populated from the underlying information systems, reviewed by the surgeon and then submitted to the DSCA. The GIOCA uses the same information systems as other departments of our hospital, plus additional spreadsheets for internal administration and management.

The goal of this paper is to report on the barriers that we encountered in the attempt to gather all raw source data required to compute the set of indicators. We categorised these barriers and provide recommendations on how they could be prevented or surmounted. A part of these recommendations can support the design of our hospital's new EHR. Supposing that our experiences are similar to data reuse projects in many hospitals, we assume that these recommendations might also help to design, evaluate and optimise systems in other hospitals.

5.2 METHODS

With the explicit consent and support of the management of the GIOCA, we cooperated with our hospitals' general ICT service in order to retrieve the data required to compute the set of four clinical quality indicators in the domain of colorectal cancer surgery (the same set as employed in [41]) for the reporting years 2010 and 2011. The indicators are contained in the sets released by the governmental program Zichtbare Zorg[2] and the Dutch Healthcare Inspectorate[3]:

1 http://www.clinicalaudit.nl/

2 http://www.zichtbarezorg.nl/mailings/FILES/htmlcontent/Ziekenhuizen/2011/Indicatoren/Verplicht/Indicatorengids%20Colorectaal%20Carcinoom%202011%20def%20nieuw.pdf

3 http://www.igz.nl/Images/2010-07%20Basisset%20kwaliteitsindicatoren%20ziekenhuizen%202011_tcm294-283436.pdf

I1: Number of examined lymph nodes after resection (process indicator)
Numerator: number of patients who had 10 or more lymph nodes examined after resection of a primary colon carcinoma.

Denominator: number of patients who had lymph nodes examined after resection of a primary colon carcinoma.

Exclusion criteria: Previous radiotherapy and recurrent colon carcinomas

I2: Participation in Dutch Surgical Colorectal Audit (DSCA) (process indicator)
Numerator: number of surgical resections of a colorectal carcinoma situated in colon or rectum (only count primary carcinomas) for which data has been submitted to the Dutch Surgical Colorectal Audit.

Denominator: total number of surgical resections of a colorectal carcinoma situated in colon or rectum (only count primary carcinomas).

I3: Patients with rectum carcinoma who have been discussed in a preoperative multidisciplinary meeting (process indicator)
Numerator: Number of patients with rectum carcinoma who have been discussed in a preoperative multidisciplinary meeting.

Denominator: Number of patients with rectum carcinoma operated in the reporting year.

Inclusion criterion: Patients who have been operated in the reporting year due to a rectum carcinoma.

Exclusion criteria: Transanal Endoscopic Microsurgery (TEM) resections and recurrent rectum carcinomas.

The Dutch Surgical Colorectal Audit states that the presence of a radiologist, a radiotherapist, a surgeon, an oncologist, a colon, stomach and liver physician and a pathologist are required for a preoperative multidisciplinary meeting.

I4: Unplanned re-interventions after resection of a primary colorectal carcinoma (outcome indicator)
Numerator: number of re-interventions during the same admission or during 30 days after the resection (choose longest interval) in the reporting year.

Denominator: total number of primary resections of a colorectal carcinoma during the reporting year.

Inclusion criteria: Primary colorectal carcinoma = first presentation of a colorectal carcinoma (thus not recurrent); might be the second or next primary presentation.

Exclusion criteria: Transanal Endoscopic Microsurgery (TEM); Endoscopic and open polypectomy

This indicator comes with a list of definitions: Resection: surgical removal of colon segment where the colorectal carcinoma is situated. Re-intervention: re-operation in the abdomen or an intervention (possibly radiological) during which a complication in the abdomen is being treated (inclusive percutaneous incision and drainage, drainage via rectum, embolisations of bleedings in the abdomen, etcetera). Admission: the time which the patient spends in a hospital directly after the operation (the same hospital or another one where the patient has been referred to); can be longer than 30 days.

Our hospital's general ICT service was currently in the process of investigating the requirements for setting up an operational data store (ODS), which integrates data from several operational databases. As the ICT service was especially interested in investigation of typical data collection requirements from a business intelligence perspective, we started a joint project to gather the required data.

In the absence of a central overview of the data available in our hospital, the goal of the *first phase* of the project was to identify the original sources of the required "raw" data elements, i.e. whether and in which systems these elements are stored, and who are the responsible contact persons. In order to do so, we interviewed the experts who are treating colorectal cancer patients, observed the work- and data flows, and interviewed those responsible for computation of the quality indicators as well as potentially responsible contact persons. In the *second phase* of the project, the team from our hospitals' general ICT service worked on the technical design of the ODS and on the actual data retrieval from the various databases. For each of the data elements established in the first phase, they identified its name, type, format and length in the database, and whether it was optional or mandatory. After this phase was completed, we obtained a version of the required data. In the *third phase* of the project, we analysed the data obtained and identified several quality issues that impeded its reuse.

We documented all barriers encountered in the course of this process, and - based on consensus - categorised them according to Galster's framework of causes that impede the reuse of clinical data in clinical settings [23]. Galster's categorisation is based on a literature review and shown in Figure 1. The causes are linked to underlying aspects that have been indicated in the Semantic Health Report [54], i.e. technical, organisational, legal and medical aspects. We also categorised the encountered barriers according to these underlying aspects, as well as the phases of our project.

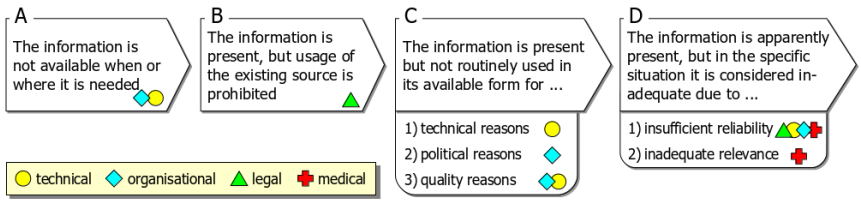

Figure 14: Categorisation of non-reuse of clinical information

Our hospital's Institutional Review Board waived the need for informed consent, as individual patients were not directly involved. The use of the

data is officially registered according to the Dutch Personal Data Protection Act.

5.3 RESULTS

5.3.1 *Required data*

In the first phase of our project, we identified 12 data elements required to compute the set of 4 quality indicators. After interviewing more than 15 people, including staff members of the GIOCA, those responsible to compute the set of quality indicators, and various database administrators, we identified 9 corresponding source systems as shown in Table 9. All required data elements were stored in a structured digital format, except for relations between diagnoses and procedures, which are essential to identify procedures that have been carried out for colorectal carcinomas and not for other reasons. These relations are often documented in free-text descriptions such as surgery reports. Most of the identified source systems were stand-alone systems for clinical and administration purposes, with data flows between them. For example, high-detail data from the surgical procedure system flows in less detail into the central procedure register. Two of the sources are external national registers. Please note that several items occurred in several databases, and in principle, one source per item should be sufficient to compute the quality indicators. As we strived to identify the source system with the highest data quality, our initial goal was to retrieve the required data from all identified systems.

The second phase of our project resulted in 5 delivered database tables after 8 months, which are underlined in Table 9. We analysed their quality in the third phase.

5.3.2 *Barriers to the reuse of routinely recorded clinical data*

In this section, the barriers we encountered are categorised according to Galster's framework as shown in Figure 14, and according to the three phases of our project.

5.3.2.1 *A) Data not available when or where it is needed*

HINDERED ACCESS TO DATA SOURCES (TECHNICAL AND ORGANISATIONAL REASONS), SECOND PHASE. The only way to obtain data from the nationwide histopathology and cytopathology archive was to request and receive it via email. Furthermore, some of the databases in our hospital are administered by external providers, which do not always guarantee structured and real-time access to our databases.

Table 9: Required data items and their source systems

Data Element	Source System	Indicator
Surgical procedure: date, type, anatomic location	surgical procedure system, procedure register	I1, I2, I3, I4
Diagnosis: anatomic location, type (primary, recurrent)	diagnosis register	I1, I2, I3, I4
Radiotherapy: date	radiotherapy system, appointment register, procedure register	I1
Lymph node examination: date, number of examined lymph nodes (in pathology report)	nationwide histopathology and cytopathology data archive	I1
Surgical procedure submitted to Dutch Surgical Colorectal Audit (DSCA)	national register	I2
Preoperative multidisciplinary meetings: date	EHR, appointment register	I3
Admission: admission and discharge date	admission register	I4

5.3.2.2 *B) Data present, but usage of the source is prohibited*

PATIENT NUMBERS (LEGAL REASONS), THIRD PHASE. The use of data was officially registered according to the Dutch Personal Data Protection Act, under the condition that it was de-identified. Therefore, we received the required data from our hospital's source systems with patient numbers hashed by the ICT service. In a later phase of our project, we could not use these patient numbers to match the patient data with data from our hospital's data warehouse, which uses other hashed patient numbers. Of course, this problem would not have existed if the ICT service and the administrators of the data warehouse had matched the data for us.

5.3.2.3 *C) Data present but not routinely used in its available form*

ORGANISATIONAL / CULTURAL BARRIERS (ORGANISATIONAL REASONS), FIRST AND SECOND PHASE. In our university hospital, we encountered various barriers in the attempt to obtain data for reuse that seemed to be due to insufficient prioritisation and culture of data reuse. First of all, no standard procedure existed to process data requests such as ours. Data can be requested via the ICT service, as in our project, or directly via the database administrators. In the busy environment of the ICT department responsible for critical IT systems, our project did not receive the highest priority, which may have caused some delay. Also, the composition of the team changed several times during the project, hampering smooth communication and progress. Once we identified the relevant source systems and the corresponding responsible persons, there were no clear guidelines and procedures on how to request a database extract;

rather, this issue had to be discussed with every database administrator individually.

Another major problem was that no central overview of the various data sources, their governance and content, including employed code systems and data dictionaries, existed, and that the management of our hospital did not envisage such an overview.

INSUFFICIENT QUALITY (ORGANISATIONAL REASONS), THIRD PHASE. Analysing the database tables that we received from the ICT service, we encountered the following quality issues.

- *Incompleteness on database level.* We did not receive any data on radiotherapy. For double-recorded elements (e.g. two sources for multidisciplinary meetings, which are recorded in the EHR, but due to the setup of the GIOCA can also be inferred from the patient's first visit), we received only one of the sources. Additionally, we encountered a problem that can probably not be generalised to other hospitals. The dataset did not include data for a complete reporting year, which would have been essential as quality indicators are computed per reporting year. We received data for 2010 and 2011, but for 2010 information on lymph node examinations and multidisciplinary meetings was missing, while for 2011, information on admission and discharge dates was missing.
- *Incompleteness on data element level.* While the surgical procedure data was delivered for both reporting years, it was probably incomplete, with the last surgical procedure of the year 2011 being on the 16th of December. Furthermore, around half of the multidisciplinary meetings had no date recorded, and were therefore unusable.
- *Incorrectness.* Some data elements were obviously incorrect, such as dates in the far future (e.g. year 2101).
- *Lack of interlinking of data in various sources.* To reuse data for indicator computation, it is essential to know which procedures have been carried out for which diagnoses, but these relations are not recorded in our hospital in a structured format. The only relation between the different data sources is the (hashed) patient number.
- *Missing provenance of data.* To reuse data, its provenance might be of interest, especially when data can stem from several sources. However, the data that we received did not contain any provenance information, so that we had to schedule further meetings to obtain a clear overview.

- *Lack of inside-knowledge of "meaning of data".* Our hospital employs national and local code systems instead of international standard terminologies, and a central metadata registry is absent, making it hard to identify the meaning of the respective data elements. For example, we encountered diagnosis-treatment codes such as *3314*, *554* or *11*, diagnosis codes such as *13862*, *29798*, *7155*, and specialisations such as *KGA*, *AUD* or *KEC*. Likewise, procedure codes such as *335127*, *989899* or *338533Y* were not interpretable without knowledge of the coding system.

PROBLEM OF SELECTING PATIENTS IN ONE SYSTEM AND QUERYING THEIR DATA IN ANOTHER SYSTEM (TECHNICAL REASONS), THIRD PHASE. Our hospital's data infrastructure is based on several small source systems instead of one large system, which makes the execution of queries - which are automatically optimised for integrated systems - harder. When querying several systems, one has to identify a suitable starting point to obtain a basic set of relevant patients, and then query other systems based on the identified patient numbers. However, it might not always be clear which one is the most suitable system to start with, and querying separate datasets can lead to a large number of irrelevant results. With regard to colorectal surgery indicators, for example, we would search for all patients who had a colectomy or a resection of rectum due to a colorectal carcinoma. In order to do this, we must query the surgical procedure database for all patients with relevant procedures and the diagnosis database for all patients with a colorectal carcinoma, and then construct the intersection of both query result sets.

5.3.2.4 *D) Data apparently present, but in the specific situation it is considered inadequate*

Because the data did not cover a complete reporting year, we did not attempt to compute the set of indicators, and therefore did not analyse the relevance and reliability of the data. We assume that all data was relevant, but that we might have encountered reliability issues, including obviously wrong procedure years such as 2101. Reliability issues are especially visible when data is recorded twice instead of being reused, such as in our hospital and in the DSCA, and double recorded items are inconsistent. We are currently investigating such issues in a subsequent study.

5.4 DISCUSSION

5.4.1 *Main Findings*

In our study, we identified a number of barriers that hinder the (timely) reuse of routinely collected clinical data. Even though all data that we required was in principle available in a digital format, and most of it within our hospital, it took a long time until we received a version of the requested data, and the data itself was of insufficient quality. The barriers that we identified cover all four of Galster's categories of why clinical information is not reused. However, category C, "Data present but not routinely used in its available form", contained the most problems, mainly due to underlying organisational/cultural and data quality reasons.

Due to the identified data quality issues, we proceeded in gathering the required data ourselves, with the explicit consent of both our hospital's ICT service and the management of the GIOCA. We started using our freshly launched hospital-internal data warehouse for research, which turned out to satisfy our requirements, with two exceptions: the data items radiotherapy and multidisciplinary meeting. We made contact with another data gathering and analysis initiative in the scope of the GIOCA, which already gathered radiotherapy and multidisciplinary meeting data and willingly shared it with us, so that we finally had a solid basis to compute first quality indicators.

5.4.2 *Related Work*

Holzer and Gall [55] compiled a catalogue of eight core requirements for secondary use of EHRs. Because the authors follow a document-oriented approach to data reuse (as opposed to our structured database-oriented approach), most of their core requirements, such as "possibility to formulate queries within the retrieved documents" cannot be related directly to our findings. However, other requirements, such as "use of standards and terminologies" fit ours.

Prokosch and Ganslandt [56] identify three challenges in the context of reusing EHRs for clinical research: to establish comprehensive clinical data warehouses that can be harvested with data mining methods, to establish an IT infrastructure that supports clinical research and to integrate and link medical record systems and clinical trial databases. We argue that data warehouses are advantageous but not imperative if source systems can be accessed directly, and also support the claim that a hospital's IT infrastructure should support clinical research. Their last challenge - to

link medical record systems and clinical trial databases - falls outside the scope of our study.

Ancker et al. [57] observed that secondary use of data might require a higher degree of data integrity than the original primary use. The Semantic Health Report [54] claims that to fully realise the potential of EHR systems, the data they contain should be of high quality, and that timely and secure access to those entitled has to be ensured. To reuse clinical data, systems must be able to exchange data, preserving its meaning. These requirements are also reflected in our recommendations that we compiled based on the barriers that we encountered.

5.4.3 *Strengths and limitations*

It might be regarded as a limitation of this study that it included only one hospital. However, we assume that our project can be seen as analogous to data reuse projects in many hospitals, which are likely to encounter similar problems and barriers, and therefore might profit from our recommendations.

It is also questionable to what extent the computation of quality indicators is typical for general data reuse. We argue that the general challenge that characterised our project was to retrieve high-quality data from several sources within and outside our hospital, and we assume that this challenge underlies most data reuse projects. Real-time access to clinical data and integration of feedback with EHRs is a further challenge, which would not only be desirable for the computation of quality indicators, but also indispensable for secondary uses such as clinical decision support.

5.4.4 *Recommendations*

5.4.4.1 *Ensure availability of data and accessibility of data sources*

When choosing external providers for EHR systems, data accessibility and reuse should be considered in order to avoid "data silos", in which it is easy to insert data, but hard to extract it. Likewise, only copies of high-quality local data should be submitted to external registers, ensuring the hospital's ownership of the data.

5.4.4.2 *Ensure patients' interests, privacy and security while allowing for reuse*

Even though this is not a direct finding of our study, it should be noted that the patients' rights, privacy and security must be protected. Patient data should always be de-identified, unless the patients' identity is absolutely necessary and of high value, such as in the recruitment of eligible patients to clinical trials, which might require informed consent.

5.4.4.3 *Set-up a reuse-friendly organisation and culture*

Especially in university hospitals, data reuse should be prioritised, and this prioritisation should be part of the hospital culture. Standard procedures should be set up to request data for reuse and financial resources should be made available to extract data for research that might benefit both the hospital and its patients. In order to facilitate reuse, a central overview of available data sources should be administered, including their governance and content as well as employed code systems and data dictionaries. Such a metadata registry could for example be based on ISO/IEC 11179, an international standard for representing metadata.

5.4.4.4 *Increase data quality*

Data quality comprises completeness and correctness, but also the recording of relations between diagnoses and procedures, and the use of standard terminologies and information models that enable meaning-based retrieval and facilitate the "Collect once, use many" paradigm [58]. In order to increase the quality of elements that are required for reuse, those responsible for recording the respective elements should be made aware of foreseeable secondary uses. Data quality also comprises metadata and provenance. In the scope of our project, it would have been helpful to know which systems the data stemmed from, as well as who recorded the data, and when and why it was recorded.

5.4.4.5 *Allow for cross-database querying*

While one monolithic overarching hospital-internal EHR might be desirable, in practice, the IT infrastructures of many hospitals consist of several dedicated source systems. In principle, this should not be a problem as long as data in all systems can be accessed and seamlessly integrated. However, querying several systems is harder and in our case required manual work. The ability to execute hospital-wide federated queries would alleviate this barrier.

5.4.5 *Conclusion*

In this paper, we categorised barriers encountered in the attempt to reuse data from our hospital for clinical indicator computation and provide recommendations that might support the design, evaluation and optimisation of EHRs. Patient data can be considered one of the most valuable resources that a hospital has at its disposal, and therefore its reuse should be facilitated while preserving the patient's privacy, security and interest.

6 INFLUENCE OF DATA QUALITY ON COMPUTED DUTCH HOSPITAL QUALITY INDICATORS: A CASE STUDY IN COLORECTAL CANCER SURGERY

Background. Our study aims to assess the influence of data quality on computed Dutch hospital quality indicators, and whether colorectal cancer surgery indicators can be computed reliably based on routinely recorded data from an electronic medical record (EMR).

Methods. Cross-sectional study in a department of gastrointestinal oncology in a university hospital, in which a set of 10 indicators is computed (1) based on data abstracted manually for the national quality register Dutch Surgical Colorectal Audit (DSCA) as reference standard and (2) based on routinely collected data from an EMR. All 75 patients for whom data has been submitted to the DSCA for the reporting year 2011 and all 79 patients who underwent a resection of a primary colorectal carcinoma in 2011 according to structured data in the EMR were included. Comparison of results, investigating the causes for any differences based on data quality analysis. Main outcome measures are the computability of quality indicators, absolute percentages of indicator results, data quality in terms of availability in a structured format, completeness and correctness.

Results. All indicators were fully computable based on the DSCA dataset, but only three based on EMR data, two of which were percentages. For both percentages, the difference in proportions computed based on the two datasets was significant.
All required data items were available in a structured format in the DSCA dataset. Their average completeness was 86%, while the average completeness of these items in the EMR was 50%. Their average correctness was 87%.

Conclusions. Our study showed that data quality can significantly influence indicator results, and that our EMR data was not suitable to reliably compute quality indicators. EMRs should be designed in a way so that the data required for audits can be entered directly in a structured and coded format.

Keywords: Data Quality, Clinical Quality Indicators, Electronic Medical Record, Clinical Audit, Patient Data, Reuse, Secondary Use

6.1 BACKGROUND

Over the last decades, it became possible and increasingly interesting to measure the quality of health care to implement quality improvement activities and to strengthen both transparency and accountability [59]. In this context, both legally mandatory and voluntary quality indicators [60] for various kinds of diseases and interventions have been released by governments, patient and scientific associations as well as insurance companies. The computed results are used for performance comparisons between health care institutions. As such comparisons have potentially serious implications, including influencing the choices of patients and insurance companies, indicator results should be reliable.

Ideally, clinical quality indicators are computed inside hospitals based on data recorded during the care process and stored in the Electronic Medical Record (EMR). In the United States, the meaningful use [46] of EMRs is put forward as a national goal, which includes the electronic exchange of health information as well as the computation and reporting of clinical quality measures [61]. This meaningful use reduces the registration burden for care providers and furthermore enables the unobtrusive measuring and monitoring of indicators in real-time, allowing for timely intervention.

Next to this development, national and international medical data registries proliferate [62], which are frequently used to quantitatively compare performance between health-care institutions. Due to various barriers that impede the reuse of data [63], many care organisations still collect the data for quality registers manually [64]. This labour-intensive process might lead to the undesirable situation that the data in registers differs from source data in an EMR.

In the Netherlands, "Zichtbare Zorg" [65] developed amongst others a set of 11 evidence-based colorectal cancer surgery indicators, which is computed based on the register of the Dutch Surgical Colorectal Audit (DSCA) [66]. The DSCA has been set up in 2009 to measure and to improve the quality of colorectal cancer surgery, serving as both national and international role model. All Dutch hospitals that perform colorectal cancer surgery submit data to the DSCA register. Ideally, data should be submitted (semi-)automatically, but in practice surgeons often enter it manually via a web form. The data is often submitted at the end of a reporting year, impeding timely feedback.

This study aims to assess whether the set of quality indicators can be computed automatically based on EMR data and to investigate barriers to succeed. Hence, we compared quality indicators computed based on our EMR data to the same indicators computed based on manually abstracted data for the DSCA register, and performed a data quality analysis to explain any differences.

6.2 METHODS

6.2.1 *Patient data*

We used two data sources of a department of colorectal cancer surgery in a university hospital: manually abstracted data for the DSCA register and structured data from the EMR.

The DSCA dataset consists of 212 variables, including demographic information, diagnoses, procedures, results of pathological examinations and clinical outcome. Attending surgeons enter the required data, either manually with the help of a web form, which takes 15 to 20 minutes per patient, or with a spreadsheet. In most hospitals the data is entered via the web form. In our hospital, the responsible surgeon preselects the patients for whom to submit data from the database containing all surgical procedures. He then browses structured and unstructured data such as pathology reports for the respective patients to identify as many of the required variables as possible. All patients of our hospital for whom data has been submitted to the DSCA in 2011 were included.

For this study, we regarded the DSCA dataset as the current reference standard. We deliberately do not refer to it as gold standard because we cannot exclude all possibility of errors due to manual data entry. However, surgeons have reported to enter the data carefully. Also, the data is monitored by the DSCA by an annual comparison to the dataset of the Dutch Cancer Registry. Its reliability seems to be high: A recent comparison showed that data has been submitted to the DSCA register for 94% of the patients in the Dutch Cancer Registry. Most data items correspond well, with discrepancies being mainly due to differing interpretations and definitions [67]. For example, anastomotic leakages are only registered in the DSCA if they caused a re-intervention, while the Dutch Cancer Registry handles a broader definition.

Regarding our EMR, several source systems that contain information on patients, diagnoses, operations, admissions, encounters, pathology reports, endoscopies and medications periodically insert data into our data warehouse. Diagnoses are encoded in ICD-9-CM, and surgical procedures in codes from a Dutch procedure classification consisting of nearly 40,000 codes. All patients who had an operation in 2011 have been extracted from the data warehouse. In the following, we refer to this dataset as EMR. All patients from the EMR who seemingly should have been submitted to the DSCA in the reporting year 2011 due to a recorded surgical resection of a primary colorectal carcinoma were included.

6.2.1.1 *Patient matching.*

In absence of patient identifiers, the patients for whom data has been submitted by our hospital to the DSCA in 2011 are matched with the patients from the EMR based on their gender, year of birth and operation date as well as sets of procedures that they underwent.

Our Institutional Review Board waived the need for informed consent, as individual patients were not directly involved. The use of the data is officially registered according to the Dutch Personal Data Protection Act.

6.2.2 *Quality indicators and their computation.*

We used the set of colorectal quality indicators released by a governmental quality of care program called "Zichtbare Zorg" for the reporting year 2011. The set consists of 8 thematic indicators, 3 of which comprise two related indicators denoted as e.g. 8a and 8b, resulting in a total of 11 indicators: 9 process indicators, 1 structure indicator and 1 outcome indicator (see Appendix). The process and outcome indicators are percentages computed based on the definitions for numerators and denominators of each indicator. The structure indicator 8a ("How many surgeons does the team include and how many of these surgeons carry out resections on primary colonic carcinoma patients?") is not designed to be computable based on the EMR. Therefore, we did not include it in our study. Of the remaining 10 indicators, the DSCA indicator 1 and the circumferential resection margin indicator 6a measure the percentage of patients for whom data has been submitted to the DSCA. As we do not expect submission of data to the DSCA to be recorded in the EMR, we exclude the numerators of these indicators. The 8 fully and 2 partially (i.e. only the denominator) included indicators have been formalised with our previously developed indicator formalisation method CLIF [41] to enable their automated computation, for which the obtained queries are run against the respective datasets. The queries are published on figshare [68].

6.2.3 *Outcome measures*

6.2.3.1 *Quality indicators.*

The first outcome measure is the *computability* of quality indicators, and the corresponding results. Numerators and denominators of indicators are computable if all required items are available in a structured format.

As in [69] and [61], we analysed the accuracy of quality indicator results computed based on EMR data by measuring *sensitivity* and *specificity*. We also measure the *positive predictive value* (PPV) and the *negative predictive value* (NPV) as well as the *positive likelihood ratio* (PLR) and the *negative likelihood ratio* (NLR).

Whether the difference in proportions was significant has been tested with Bland's and Butland's method to compare proportions in overlapping samples [70]. A p-value < 0.05 was considered significant.

6.2.3.2 *Data quality.*

We analysed the quality of the 14 data items required to compute the set of quality indicators (Operation date, Year of birth, Procedure, Operation urgency, Primary location / Diagnosis, cT score, pN stage, pM stage, Examined lymph nodes, Circumferential margin, Colonoscopy, Chemotherapy / Medication, Meeting date and Radiotherapy start date). The first quality dimension we analysed is *availability in a structured format*, as unstructured data cannot be used directly to automatically compute quality indicators. For data items that are available in a structured format, we focus on the quality dimensions *completeness* and *correctness* [71]. Completeness is measured as the percentage of items that should be recorded for each patient (such as the operation urgency, as all included patients have been operated) that are indeed available in the respective dataset. Items that do not necessarily apply to all patients, such as the start date of preoperative radiotherapy, are excluded, as a missing value might be due to the fact that the patient was indeed not treated with previous radiotherapy, but it might also be the case that the start date has not been recorded. Items explicitly recorded as 'unknown' are regarded as absent, diminishing completeness.

We measure *correctness* by checking whether data items recorded in the EMR are consistent with the corresponding items in the DSCA dataset with regard to the indicator definitions, i.e. whether they have the same effect on the indicator results. For example, a date for a multidisciplinary meeting is considered correct if both dates are before or both dates are after the operation.

Finally, encountered problems regarding data quality are categorised.

6.3 RESULTS

6.3.1 *Patient matching*

As shown in Figure 1, 75 patients are included for the reporting year 2011 in the DSCA dataset, and 79 in the EMR. Following the matching strategy, it was possible to match all 75 DSCA patients with patients in the EMR. Sixty-three of these patients were also selected by the query to compute the indicators based on the EMR dataset, while 12 patients were not selected. Manual inspection showed that 4 of these 12 patients had no relevant diagnosis recorded in the EMR. A fifth patient was recorded with a colonic carcinoma and a resection of rectum, but the query against the data warehouse selected patients with a colonic carcinoma and colectomy or a rectum carcinoma and resection of rectum. For the remaining 7 patients, the diagnosis date was after the (elective) operation date, so that a relationship between diagnosis and operation could not be assumed.

Sixteen patients from our EMR dataset could not be matched to the DSCA dataset because they were selected incorrectly due to incorrect (e.g. tumours that were classified as non-malignant based on the pathology examination) or imprecise (e.g. recurrent carcinomas) diagnosis codes or despite missing relations between the diagnosis and the procedure in the EMR dataset.

6.3.2 *Computation of quality indicators*

Table 10 shows the indicator results computed based on the DSCA dataset, as well as fully computable indicators and denominators based on the EMR data. The chemotherapy indicators 5a and 5b as well as the radiotherapy indicator 7 could not be computed, as the required carcinoma's stage was not available in a structured format.

Table 10: Indicator results based on both datasets. Percentages are denoted as % (numerator/denominator). The underlined indicators are those for which only the denominator has been included.

Indicator	DSCA	EMR	sensitivity	specificity	PPV	NPV	PLR	NLR
1 DSCA	(75/-)	(-/79)	-	-	-	-	-	-
2 lymph nodes	85% (39/46)	(-/36)	-	-	-	-	-	-
3 meeting	100% (29/29)	70% (23/33)	79% (23/29)	- (0/0)	100% (23/23)	0% (0/10)	-	-
4 imaging	88% (36/41)	58% (31/53)	58% (21/36)	60% (3/5)	67% (21/31)	14% (3/44)	1,45	0,7
5a chemotherapy	80% (8/10)	-	-	-	-	-	-	-
5b chemotherapy	17% (1/6)	-	-	-	-	-	-	-
6a CRM	62% (18/29)	(-/33)	-	-	-	-	-	-
6b CRM	14% (4/29)	(-/33)	-	-	-	-	-	-
7 radiotherapy	92% (22/24)	-	-	-	-	-	-	-
8b volume	46	37	-	-	-	-	-	-

6.3.2.1 *Comparison of selected patients.*

Table 11 shows the comparison of selected patients for all fully computable indicator elements.

Table 11: Patients selected based on the two datasets. TP stands for True Positives, FP for False Positives and FN for False Negatives. TP are DSCA and EMR, FP only DSCA and FN only EMR.

Indicator	Element	DSCA	EMR			
			EMR	TP	FP	FN
1 DSCA	num / denom	75	79	63	12	16
2 nodes	denominator	46	36	28	18	8
3 meeting	numerator	29	23	23	6	0
3 meeting	denominator	29	33	25	4	8
4 imaging	numerator	36	31	21	15	10
4 imaging	denominator	41	53	31	10	22
6a and 6b CRM	denominator	29	33	25	4	8
8b volume	-	46	37	28	18	9

6.3.3 *Outcome measures*

6.3.3.1 *Quality indicators.*

All 10 indicators were fully computable based on the DSCA dataset. Eight of these indicators should in principle be fully computable based on EMR data, but in practice this was the case for only three indicators. For the two indicators (multidisciplinary meeting and imaging) that are percentages, the difference in proportions computed based on the two datasets was significant.

For 4 indicators, only the denominators were fully computable, because the data items defining the quality of care measured in the numerator, such as the number of examined lymph nodes, were not available in a structured format.

6.3.3.2 *Data quality.*

The results of the data quality analysis are given in Table 12. Fourteen data items are required to compute the set of quality indicators. All of these items are available in the DSCA register, and 8 in the EMR, with

the remaining 6 only being available in free text. The pathology reports contained in the EMR comprise required data such as the number of examined lymph nodes, the circumferential margin and the pathological stage of the carcinoma only in free text. The clinical stage of the carcinoma is equally unavailable, although it might be present in free text sources that we did not have at our disposal, such as conclusions of physical or radiologic examinations or endoscopies, or contained in referral letters. It is contained in a structured format in the Dutch Cancer Registry, but the goal of our study was to focus on the data in our EMR.

For data items that should be recorded for each patient, the average completeness is 86% for the register's dataset and 50% for the EMR. The average correctness of data items in the EMR is 87%.

Table 12: Data Quality. Elements enclosed by square brackets are not supposed to be available for each patient.

Item	Completeness DSCA	Completeness EMR	Correctness
Operation date	100% (75)	100% (75)	100% (75)
Year of birth	100% (75)	100% (75)	100% (75)
Procedure	100% (75)	100% (75)	97% (73/75)
Operation urgency	100% (75)	100% (75)	95% (71/75)
Primary location / Diagnosis	100% (75)	100% (75)	91% (68/75)
cT score	39% (29)	0% (unavailable)	-
pN stage	100% (75)	0% (unavailable)	-
pM stage	100% (75)	0% (unavailable)	-
Examined lymph nodes	99% (74)	0% (unavailable)	-
Circumferential margin	24% (18)	0% (unavailable)	-
[Colonoscopy]	[100% (75)]	[80% (60)]	83% (50/60)
[Chemotherapy / Medication]	[99% (74)]	[97% (73)]	21% (15/73)
[Meeting date]	[85% (64)]	[79% (59)]	98% (57/58)
[Radiotherapy start date]	[33% (25)]	[24% (18)]	100% (18/18)
Average of available items	86%	50%	87%

6.3.4 *Catalogue of encountered problems*

In our case study, quality indicators could not be computed reliably based on the EMR data due to the general problems as enlisted in Table 13.

Table 13: Catalogue of encountered problems.

Problem	**Explanation**
Data not available in structured format	Data items required to compute many of the indicators, such as those contained in the pathology reports, were only available in non-structured free text, and therefore not directly (re)usable. Also structured data to exclude patients based on the exclusion criteria *recurrent carcinoma* and *TEM-resection* as well as *'resection' via colonoscopy* was not available in our EMR nor in the DSCA dataset. Non-recorded exclusion criteria can lead to lower indicator results, wrongly underestimating the quality of care for indicators whose percentages are to be maximised [72,73].
Incorrect data items	The double data entry in our case study helped us to discover incorrect data items. Furthermore, we identified imprecise and / or incorrect diagnosis codes in our EMR.
Incomplete view of patient history	Hospitals throughout the country refer patients to our hospital, which specialises in gastro-intestinal oncology. Some of these patients are only treated for a short time, and then referred back. Likewise, our hospital maintains an alliance with a nearby hospital. Referral letters are typically posted as physical letters, making a complete, consistent view on a patient's history difficult to obtain. For example, it is hard to retrace whether preoperative imaging of the colon has taken place in another hospital.
Lack of relations between data items	Our EMR does not store any relations between diagnoses and procedures, making it impossible to select the diagnosis that was the underlying reason for a procedure. For example, the lymph node indicator should only select lymph node examinations that have been carried out in the context of a primary colonic carcinoma, and not, for example, a previous mamma carcinoma. As a partial solution, we imposed the constraint that the diagnosis should have been established before the related operation was carried out, which resulted in some missed patients.
Lack of detail	None of the diagnoses in the EMR was detailed enough to meet the information required by the indicators, which include patients with *primary* colonic and rectum carcinomas. The only relevant diagnoses in the EMR were malignant neoplasm of colon, rectum and rectosigmoid junction. Therefore, the concepts employed in the queries to compute the indicators had to be generalised. Furthermore, only the type of endoscopies is registered, such as colonoscopy, but not whether the complete colon is affected.
Lack of standardisation	For example, the urgency of an operation is defined in the EMR according to 8 categories, but the DSCA dataset only differentiates urgencies according to 4 categories. It was not clear how these categories should be mapped, as their meaning was not unambiguously described (for example, one of the categories was called "extra").

6.4 DISCUSSION

Our results show that EMR-based indicator results significantly underestimate the quality of care compared to the same indicators computed based on manually abstracted data for a national quality register. Reasons were unavailable, incomplete and incorrect data items as well as missing relationships between diagnoses and procedures in the EMR. In particular, detailed data that reflects whether a patient's treatment met the ideal standard of care was often incomplete in the EMR.

6.4.1 *Comparison with other studies*

The use of EMRs has increased rapidly in the recent years, making trustworthy reuse of data [2] an important challenge and research question. Worldwide, EMR-based quality measures [74] are increasingly employed, and new standards [75] such as *eMeasures* [76] to automatically derive quality measures from EMRs are introduced.

Many researchers have compared results computed based on different data sources. Both Kerr et al. [77] and Parsons et al. [78] found that EMR-derived measures can underestimate performance in comparison to manual abstraction. Kern et al. [61] found that a "wide measure-by-measure variation in accuracy threatens the validity of electronic reporting". Likewise, results of quality indicators computed based on administrative data have been compared to results computed based on manually abstracted EMR data. MacLean et al. [79] found that the EMR allows for a greater spectrum of measurable quality indicators, while summary estimates computed based on both data sources did not differ substantially. Tang et al. [80] found a significantly higher percentage of patients that have been identified to be relevant by manual selection.

Ancker et al. observed that "secondary use of data [...] requires a generally higher degree of data integrity than required for the original primary use" [57]. It has been suggested that reliable and valid quality indicator results are only achievable based on accessible and high-quality data [13, 15–21]. Likewise, it has been shown that data quality issues are common in data warehouses and electronic patient records [81–83].

6.4.2 *Limitations of this study*

Our case study included one hospital and one year of data with a relatively small sample size, and it is questionable to what extent the situation in our hospital is generalisable to other hospitals. However, the sample size was sufficient to show that data quality can significantly influence computed quality indicator results, which should be independent from the respective location.

6.4.3 *Recommendations / Future Work*

Based on the encountered problems, we compiled a set of recommendations to improve the quality and (re)usability of EMR data.

6.4.3.1 *Availability of structured data.*

Data to determine the quality of care is particularly valuable, and hospital information systems should be set up in such a way that this data is available, accessible and usable for quality measurement and further use-cases. To obtain structured data, synoptic reports, i.e. predefined computer-based forms to record relevant procedures and findings in a structured, standardised format, have been shown to be advantageous [84–86]. A standard way to encode medical free text is the use of Natural Language Processing tools. However, as most tools are developed for English, further research is required to handle Dutch [87].

6.4.3.2 *Correctness of data items.*

Multiple data entry is unnecessary, error-prone, tedious and time-consuming. Data should be recorded only once, in an adequate quality. The quality might be risen by making those entering data aware of its possible reuses. Also, local quality improvement strategies from the literature [64, 88] could be applied. To submit data to the DSCA under such improved circumstances, required items could be preselected automatically from the EMR, checked by the one responsible and be submitted to quality registers or other authorised parties. If the data needs to be edited, changes should be applied locally before the data is shared with external parties.

6.4.3.3 *Longitudinal view of patient history.*

As patient referrals are common and hospital alliances are likewise to proliferate in the future, it must become common practice to exchange data securely and automatically. Patients are likely to become active managers of their health, increasingly enabled to share their data with their caregivers.

6.4.3.4 *Relations between diagnoses and procedures.*

To reuse clinical data, the relations between diagnoses and procedures must be traceable. To be able to automatically select only examinations that have been carried out in the context of a certain diagnosis, such relations should be recorded.

6.4.3.5 *Level of detail.*

Patient data should be recorded as detailed as necessary for quality indicator computation and further foreseeable use-cases, such as the recruitment of patients for clinical trials, decision support, the early detection of epidemics or general clinical research. This might seem time-consuming, but will likely reduce the workload in the long term, as each data item has to be recorded only once. To further reduce the workload, the process should be supported by advanced data entry methods and interfaces.

6.4.3.6 *Standardisation.*

Only data that is represented meaningfully - ideally in standard codes from comprehensive controlled clinical terminologies - can be reused automatically. Terminologies such as SNOMED CT can support the "Collect once - use many times" paradigm [58], which stands for the idea that data is captured only once and can be reused thereafter for a variety of purposes. Controlled terminologies can allow for meaning-based retrieval, for example by aggregation along hierarchical structures, or based on relationships between codes. An advantage of standard terminologies is that they are integrated in the National Library of Medicine's Unified Medical Language System Metathesaurus, which contains mappings between terms across multiple terminologies.

6.5 CONCLUSIONS

This study showed that data quality can significantly influence indicator results, and that our routinely recorded EMR data was not suitable to reliably compute quality indicators. To support primary and secondary uses of data, EMRs should be designed so that a core dataset consisting of relevant items is entered directly and timely in a structured, sufficiently detailed and standardised format. Furthermore, awareness about the (re)use of data could be risen to ensure the quality of required data, and local data quality improvement strategies could be applied. Data could then be aggregated for different uses, according to various definitions. This strategy likely leads to an increased volume of high-quality data, which can ultimately serve as a basis for physicians not only to monitor but also to deliver the best possible quality of care.

6.6 APPENDIX: ZICHTBARE ZORG INDICATORS FOR 2011 TRANSLATED FROM DUTCH TO ENGLISH

1.	**Dutch Surgical Colorectal Audit (Process)**
Numerator	Number of surgical resections of colorectal carcinomas located in colon or rectum (only count resections for primary carcinomas) for which data has been submitted to the Dutch Surgical Colorectal Audit
Denominator	Number of surgical resections of colorectal carcinomas located in colon or rectum (only count resections for primary carcinomas)
Inclusion	primary carcinomas
Exclusion	recurrent colorectal carcinomas; TEM-resection (transanal endoscopic microsurgery)

2.	**Number of lymph nodes examined after resection (Process)**
Numerator	Number of patients who had 10 or more lymph nodes examined after resection of a primary colonic carcinoma
Denominator	Number of patients who underwent resection of a primary colonic carcinoma
Inclusion	all primary carcinomas, for which a part of the colon has been resected via open or laparoscopic surgery
Exclusion	1) patients who had a 'resection' via colonoscopy; 2) patients with previous radiotherapy; 3) patients with a recurrent carcinoma

3.	**Patients with rectum carcinoma discussed in multidisciplinary meeting before surgery (Process)**
Numerator	Number of patients with rectum carcinoma who have been discussed in a multidisciplinary meeting before the surgery
Denominator	Number of patients with rectum carcinoma operated in reporting year
Inclusion	all patients who underwent a resection of rectum due to a primary rectum carcinoma in the reporting year, via open or laparoscopic surgery
Exclusion	TEM-resections and recurrent rectum carcinoma

4.	**Preoperative imaging colon (Process)**

Numerator	Number of patients with diagnosed colorectal carcinoma which has been resected electively en whose colon has been imaged completely before the surgery
Denominator	Number of patients with diagnosed colorectal carcinoma which has been resected electively
Inclusion	all primary carcinomas, for which a part of the colon has been resected via open or laparoscopic surgery
Exclusion	1) patients who had a 'resection' via colonoscopy; 2) patients with previous radiotherapy; 3) patients with a recurrent carcinoma

5.	**Adjuvant chemotherapy colonic carcinoma (Process)**
Numerator 5a	Number of patients < 75 years old with a resected stage III (N1-2 M0)colonic carcinoma who received adjuvant chemotherapy
Denominator 5a	Number of patients < 75 years old with a resected stage III colonic carcinoma
Numerator 5b	Number of patients $\geqslant 75$ years old with a resected stage III (N1-2 M0)colonic carcinoma who received adjuvant chemotherapy
Denominator 5b	Number of patients $\geqslant 75$ years old with a resected stage III colonic carcinoma
Inclusion	all primary carcinomas, for which a part of the colon has been resected via open or laparoscopic surgery, and which have been classified as stage III in an postoperative pathology examination
Exclusion	1) patients who had a 'resection' via colonoscopy; 2) patients with a recurrent carcinoma

6.	**CRM rectum carcinoma (6a: Process, 6b: Outcome)**
Numerator 6a	Number of patients with a resected primary rectum carcinoma for which the CRM (circumferential resection margin) has been included in the pathology report and registered in the DSCA
Denominator 6a	Number of patients with a resected primary rectum carcinoma
Numerator 6b	Number of patients with rectum carcinoma with a CRM of 1mm or less (tumor positive)
Denominator 6b	Number of patients with a resected primary rectum carcinoma

Inclusion	all patients who underwent a resection of rectum due to a primary rectum carcinoma in the reporting year, via open or laparoscopic surgery
Exclusion	TEM-resections and recurrent rectum carcinoma

7.	**Preoperative radiotherapy rectum carcinoma (Process)**
Numerator	Number of patients with T3 or T4 rectum carcinoma who received preoperative radiotherapy
Denominator	Number of patients with T3 or T4 rectum carcinoma
Inclusion	-
Exclusion	-

8.	**Volume (8a: Structure, 8b: Process)**
Indicator 8a	How many surgeons does the team include and how many of these surgeons carry out resections on primary colonic carcinoma patients?
Indicator 8b	Number of resections of primary colonic carcinomas
Inclusion	-
Exclusion	-

7 SEMANTIC INTEGRATION OF PATIENT DATA AND QUALITY INDICATORS BASED ON *OPEN*EHR ARCHETYPES

Electronic Health Records (EHRs) contain a wealth of information, but accessing and (re)using it is often difficult. Archetypes have been shown to facilitate the (re)use of EHR data, and may be useful with regard to clinical quality indicators. These indicators are often released centrally, but computed locally in several hospitals. They are typically expressed in natural language, which due to its inherent ambiguity does not guarantee comparable results. Thus, their information requirements should be formalised and expressed via standard terminologies such as SNOMED CT to represent concepts, and information models such as archetypes to represent their agreed-upon structure, and the relations between the concepts. The two-level methodology of the archetype paradigm allows domain experts to intuitively define indicators at the knowledge level, and the resulting queries are computable across institutions that employ the required archetypes. We tested whether *open*EHR archetypes can represent both elements of patient data required by indicators and EHR data for automated indicator computation. The relevant elements of the indicators and our hospital's database schema were mapped to (elements of) publicly available archetypes. The coverage of the public repository was high, and editing an archetype to fit our requirements was straightforward. Based on this mapping, a set of three indicators from the domain of gastrointestinal cancer surgery was formalised into archetyped SPARQL queries and run against archetyped patient data in OWL from our hospital's data warehouse to compute the indicators. The computed indicator results were comparable to centrally computed and publicly reported results, with differences likely to be due to differing indicator definitions and interpretations, insufficient data quality and insufficient and imprecise encoding. This paper shows that *open*EHR archetypes facilitate the semantic integration of quality indicators and routine patient data to automatically compute indicators.

7.1 INTRODUCTION

Today, increasing volumes of clinical data are being routinely recorded, and there is tremendous potential to benefit from reusing the resulting data sources both for individual patients and society in general. In fact, according to a recent report by PricewaterhouseCoopers, "using data for secondary purposes is one of the most promising ways to improve health outcomes and costs." [1]. Secondary purposes include research, the recruitment of eligible patients for clinical trials, the early detection of epidemics, reimbursement, clinical audit, the generation or testing of medical hypotheses and quality monitoring or reporting. Since patient data often resides in various heterogeneous systems, it needs to be integrated to be (re)usable. In addition, this patient data needs to be meaningful for applications that reuse it.

*Open*EHR archetypes [22] have been proposed to standardise clinical data to achieve semantic interoperability. They have been shown to facilitate the integration of data from several sources [89], to empower multicentre clinical research [90] and to be a solid basis for ubiquitous computing [91]. Also, archetypes have been shown to facilitate the reuse of patient data for clinical trials [92] and guideline systems [93], [94]. In this paper, we focus on the reuse of patient data for the automated computation of quality indicators, which are measurable elements of practice performance for which there is evidence or consensus that they can assess the quality of provided care, and thus also change in quality [4]. Our main objective was to represent both patient data from our hospital's data warehouse and national quality indicators in terms of *open*EHR archetypes to automatically compute quality indicators.

To apply formal representation to ensure semantic interoperability and to be able to perform automated reasoning with the archetypes and the patient data, we employ an OWL 2[1] representation of archetypes, representing the patient data as its instances. To the best of our knowledge, this is the first time that real patient data is being represented based on *open*EHR archetypes in OWL and used to compute clinical quality indicators.

The structure of this paper is as follows: Section 7.2 introduces quality indicators and archetypes, and Section 7.3 our methods and materials. We report on our case study in Section 7.4. Finally, lessons learned and future challenges are discussed in Section 7.5. Section 7.6 concludes this paper.

7.2 BACKGROUND: QUALITY INDICATORS AND ARCHETYPES

This section provides background information on quality indicators and archetypes.

1 http://www.w3.org/TR/owl2-overview/

7.2.1 *Quality Indicators*

Quality Indicators are employed internally by hospitals to measure and improve the quality of care and externally for accountability and hospital comparison. For the latter, it is essential that the same measurements are performed in each hospital. Quality indicators are often expressed as a fraction, where the denominator defines the criteria of patients to whom the indicator applies, and the numerator those criteria indicating whether the patients received high-quality care. Exclusion criteria can apply. These indicators can be computed automatically by running two queries against the required patient data: one for the denominator and another for the numerator. A sample indicator is the evidence-based process indicator "Number of examined lymph nodes after colon resection" as defined by the Dutch Healthcare Inspectorate[2] for the reporting year 2010:

> **Number of examined lymph nodes after resection**
> *Numerator:* Number of patients who had 10 or more lymph nodes examined after resection of a primary colonic carcinoma.
>
> *Denominator*: Number of patients who had lymph nodes examined after resection of a primary colonic carcinoma.
>
> *Exclusion criteria*: Previous radiotherapy and recurrent colonic carcinomas.
>
> *Reporting year*: 2010

7.2.2 *Archetypes*

Archetypes are knowledge-level models that represent clinical concepts and define the structure to record, exchange and integrate clinical data. *Open*EHR archetypes are created based on the consensus of domain experts, and are available via the public archetype repository *Clinical Knowledge Manager*[3]. They define occurrence and cardinality constraints, as well as constraints on the values to be entered. The main categories are Action (e.g. Procedure undertaken), Evaluation (e.g. Diagnosis), Observation (e.g. Blood Pressure) and Instruction (e.g. Medication order). Figure 15 depicts the publicly available archetype "Tumour - lymph node metastases"[4]. The optional archetype node "Number of nodes examined" constrains the number of examined lymph nodes to be greater than or equal to 0.

2 http://www.zichtbarezorg.nl/page/Ziekenhuizen-en-ZBC-s/Kwaliteitsindicatoren
3 http://www.openehr.org/knowledge
4 http://openehr.org/knowledge/OKM.html#showarchetype_1013.1.396_5

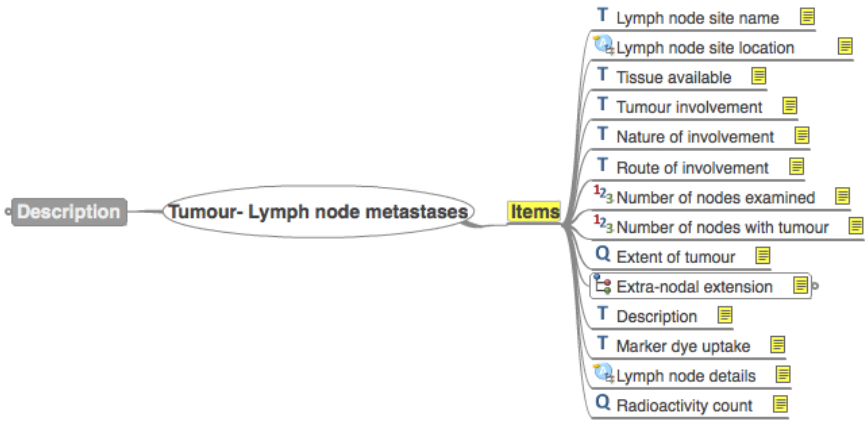

Figure 15: Archetype "Tumour - lymph node metastases". The icons depict the datatypes that are to be used. The *T* stands for free or coded text, *Q* for a quantity, *123* for a count, the globe depicts a slot (cluster) that can include other archetypes, such as "Precise anatomical location" for the node "Lymph node site location", and the tree icon depicts a cluster.

According to the Semantic Health Report [54], semantically interoperable EHR systems rely upon three layers to represent meaning: standard generic reference models such as the *open*EHR Reference Model or the Health Level 7 Clinical Document Architecture (HL7 CDA), agreed clinical structure definitions such as *open*EHR archetypes or HL7 templates, and clinical terminology systems such as LOINC[5] or SNOMED CT [27]. Archetype-enabled EHR architectures are based on the two-level methodology [22], which separates the knowledge level from the information level. Archetypes on the knowledge-level constrain the standardised stable and generic reference model on the information level that consists of few abstract classes. The reference model can either be implemented directly by EHR systems or mapped to a local data structure. Unlike the reference model, archetypes evolve together with medical knowledge. Here, we use the term *information model* to refer to both archetypes and their underlying reference model.

The two-level methodology allows queries against patient data to be constructed at the knowledge level, enabling clinical domain experts to contribute to the formalisation of quality indicators without having to know the underlying structure of the patient data. It also makes the resulting queries computable across systems that employ the required archetypes. If data is stored in proprietary systems or represented in competing standards, the required elements have to be mapped from the locally employed

5 http://loinc.org/

information model to the (elements of the) archetypes used to identify required elements to compute the quality indicator.

7.3 METHODS AND MATERIALS

This section describes the sample set of employed quality indicators, their formalisation, our patient data and how we related it to SNOMED CT codes, the translation of archetypes to OWL and how patient data was dealt with in OWL.

7.3.1 *Quality Indicators and their Formalisation*

Besides the sample indicator "Number of examined lymph nodes after resection" described above, the other two evidence-based indicators of the sample set are the process indicator "Patients with rectum carcinoma who have been discussed in a preoperative multidisciplinary meeting", and the outcome indicator "Unplanned re-interventions after resection of a primary colorectal carcinoma". We previously formalised the same sample set with our quality indicator formalisation method CLIF [41], employing a self-defined information model and self-generated patient data. CLIF consists of eight steps. In step 1), relevant concepts have to be identified in the indicator text and encoded in a terminology such as SNOMED CT. In step 2), the elements of the information model are defined and related to each other. In step 3) to 5), temporal, numeric and Boolean constraints are defined. In step 6, constraints can be grouped by Boolean connectors. Finally, exclusion criteria are defined in step 7), and the difference between denominator and numerator is made explicit in step 8). Of all steps, the second is the most relevant in the context of this paper, because we improve the formalisation by using public *open*EHR archetypes as information model. Besides, real patient data from our hospital's data warehouse is being used.

7.3.2 *Patient Data and SNOMED CT Codes*

We worked on a subset of our hospital's data warehouse, beginning from 2009.[6] The central patient table (1,672,104 entries) contains demographic information and patient IDs. Other relevant tables contain diagnoses (2,925,156), operations (144,860), admissions (259,005), encounters

6 In the Netherlands, there is no need for patient consent when, as in our study, individual patients are not directly involved. The use of the data is officially registered according to the Dutch Personal Data Protection Act.

(3,244,586) and pathology reports (92,870). The diagnosis table from the data warehouse contains ICD-9-CM codes for ca. half of the diagnoses, which we mapped to the latest SNOMED CT release (January 2012) via the SNOMED CT to ICD-9-CM crossmap included in the release. The procedures in the operation table contain codes from the Dutch procedure classification of nearly 40,000 codes that are not mapped to any other terminology. Therefore, we manually mapped a relevant subset that refers to "colorectal" procedures to SNOMED CT.

The sample set of quality indicators is computed centrally by the Dutch Surgical Colorectal Audit based on data submitted by Dutch hospitals for all operations on patients with a primary colorectal carcinoma. To extract a manageable but relevant set of patient data, we matched the data submitted by our hospital to the DSCA from 2009 to 2011 with the data stored in our data warehouse. In absence of a mapping between the patients in both systems, we searched DSCA patients in our data warehouse based on sex, year of birth, operation and discharge date as well as the procedures that they underwent. This strategy allowed us to match 192 of the 229 patients for whom data has been submitted to the DSCA.

In total, for the 192 patients 2,656 diagnoses have been recorded, of which 1,515 have (271 distinct) ICD-9 codes, the others are not encoded in ICD-9. 1,325 (239 distinct) of these codes are present in the SNOMED CT to ICD-9 crossmap, and related to 17,611 (3,878 distinct) SNOMED CT codes. 724 (201 distinct) procedures have been recorded, of which 287 (32 distinct) are present in our manually created mapping table, and related to 949 (50 distinct) SNOMED CT procedure codes. This results in 191 of the 192 patients being related to SNOMED CT procedure codes, and 190 patients to diagnosis codes. Data required that is recorded in our hospital but not contained in the data warehouse is information on radiotherapy and multidisciplinary meetings. However, it is present in the DSCA dataset and thus we retrieve it from there. We also retrieved the number of examined lymph nodes from the DSCA dataset, which is present in our data warehouse, but only in Dutch free text.

7.3.3 *Archetypes in OWL*

We reused the Archetype Ontologizer[7] [95] to create the OWL 2 ontologies for the archetypes required to represent the patient data and the quality indicators. The translated ontologies are based on the "*open*EHR Specific Data Structures and Data Types" ontology[8] [96] that represents the *open*EHR reference model, containing its data structures and data types

7 http://oe.dynalias.net:8080/JSPWebArchetypeOntologizer/
8 http://klt.inf.um.es/~cati/ontologies/OpenEHR-SP-v2.0.owl

with all their properties. We made minor adaptations to the translator so that the default namespace includes the ID of the respective archetype, and added the internal node IDs to class names for nodes. Furthermore, we made use of the OWL 2 reasoners HermiT and Pellet to check the consistency of the ontologies and the satisfiability of all classes. We used Pellet's explanation feature to identify the causes for unsatisfiable classes and improved the translator until all classes were satisfiable (for example, a datatype used in combination with a property from the "*open*EHR Specific Data Structures and Data Types" ontology had to be changed from integer to float to conform to the properties range). With this adapted translator, we translated the 5 archetypes needed to represent the patient data and the quality indicators from ADL, the Archetype Definition Language, to OWL. We then merged the resulting OWL 2 ontologies with the "*open*EHR Specific Data Structures and Data Types" ontology. The final ontology consists of 2,001 logical axioms, and has the expressivity $\mathcal{ALCHIQ}(\mathcal{D})$.

7.3.4 *Patient Data in OWL*

The patient data was originally stored in a MySQL database, and transformed into OWL using the OWL API. To run the queries against the patient data, we loaded the full closure of SNOMED CT (January 2012), the merged archetype ontology and the transformed patient data into OWLIM-SE 5.0 [97], and ran it in combination with Sesame 2.6.5[9], because it supports SPARQL 1.1[10].

7.4 CASE STUDY

To establish whether open*EHR archetypes are suitable to semantically integrate routine clinical data and quality indicators*, we first transformed patient data from our data warehouse into archetyped patient data (Section 7.4.1) and modelled the concepts of our sample set of quality indicators in terms of *open*EHR archetypes (Section 7.4.2). We then constructed archetyped SPARQL queries (Section 7.4.3) and ran them against the archetyped patient data to compute the indicators (Section 7.4.4).

9 http://www.openrdf.org/
10 http://www.w3.org/TR/sparql11-query/

7.4.1 *Transforming Patient Data into Archetyped Patient Data*

The first step of the transformation process is to map the data structure of our data warehouse and the DSCA dataset to *open*EHR archetypes. We make use of archetypes from the Clinical Knowledge Manager[11], as it can be assumed that publicly available archetypes are most widely employed.

Table 15: Mapping between the local data structure and *open*EHR archetypes. The added element is italicised. Database tables have been mapped to archetypes, and database columns to nodes of archetypes. Data warehouse is abbreviated by DWH, and SNOMED CT by SCT.

Table	**Column**	**Archetype**	**Node**
Patient (DWH)	Identifier (DWH)	Patient	Name
Admission (DWH)	Admission Date (DWH) Discharge Date (DWH)	Patient Admission	Admission Date *Discharge Date (added)*
Diagnosis (DWH)	ICD-9 (DWH) (SCT code via ICD-9 - SCT mapping)	Diagnosis	Diagnosis
Operation (DWH)	Dutch procedure code (DWH, SCT code via manual mapping)	Procedure undertaken	Procedure
(DSCA)	Radiotherapy	Procedure undertaken	Procedure with fixed SCT code (SCT_108290001)
(DSCA)	Multidisciplinary meeting	Procedure undertaken	Procedure with fixed SCT code (SCT_312384001)
Pathology (DWH, only lymph node examination)	Number of examined lymph nodes (DSCA)	Procedure undertaken Tumour- Lymph node metastases	Procedure with fixed SCT code (SCT_284427004) Number of nodes examined

Table 15 provides an overview of the mapping. Most database tables and their relevant columns can be mapped directly to (elements of) archetypes. The patient table is mapped to the demographic archetype “Patient”, and the patient ID to its mandatory node “Name”; SNOMED CT diagnosis codes are mapped to the node “Diagnosis" of the archetype “Diagnosis”, and operation codes to the node “Procedure" of the archetype “Procedure undertaken”. For radiotherapy, multidisciplinary meeting and pathology,

11 http://www.openehr.org/knowledge

exact codes are neither available nor required, as they are not specified by the indicators, so fixed codes were set. To represent the number of examined lymph nodes, we employ the archetype "Tumour - lymph node metastases", depicted in Figure 15, to record findings of lymph node metastases. While admissions and admission dates can be mapped directly, the admission archetype does not contain the required patient's discharge date, and at the time of writing, an archetype "Patient discharge" did not exist either. Consequently, we added the node "Discharge date/time" to the archetype "Patient admission". All procedure dates are represented via *open*EHR's reference model.

Based on the mapping, the patient data was transformed into OWL individuals of the archetype classes. Our program transforms every patient into an OWL individual of the archetype "Patient", with an arbitrary patient number represented in the obligatory archetype node "Name". Then, all SNOMED CT diagnoses and procedures with their corresponding dates, and all admissions are transformed into OWL 2 individuals. The number of examined lymph nodes is added, and the date of the first pathology report after the operation is set as lymph node examination date. Finally, data from the DSCA database table related to radiotherapy and multi-disciplinary meetings is added. The resulting dataset contains 52,495 logical axioms, and its expressivity is $\mathcal{AL}(\mathcal{D})$.

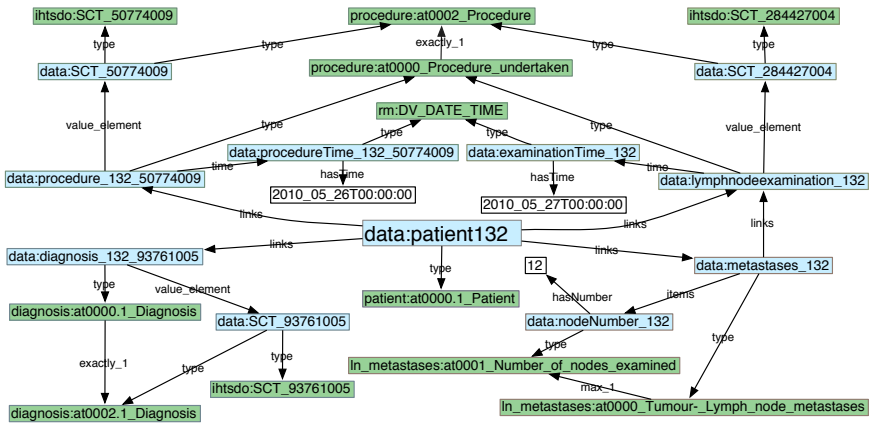

Figure 16: Example Patient. Green elements are the classes that stem from the OWL archetypes. Blue elements are instances of these classes, and white elements are literals. Note that not all relations defined between nodes of the archetypes are depicted. For example, "diagnosis:at0002.1_Diagnosis" is a node of the archetype "diagnosis:at0000.1_Diagnosis".

Let us consider the data for an example candidate patient for the lymph node indicator as depicted in Figure 16. The patient has an instance of a "Diagnosis" for the diagnosis primary colon carcinoma, an instance of a "Procedure" for a colectomy and one for lymph node examination, and an additional instance of "Tumour - lymph node metastases". The diagnosis and procedures are related to their respective SNOMED CT codes via the property "value_element". Relationships between a patient and other individuals are expressed by the "links" property.

7.4.2 *Modelling Quality Indicators in terms of* openEHR *Archetypes*

This section discusses the archetype-level modelling of quality indicators in terms of *open*EHR archetypes as employed information model.

7.4.2.1 *Bind concepts from a terminology to concepts of an information model*

For each SNOMED CT code that defines a diagnosis or procedure occurring in the indicator texts (as identified in step 1 of CLIF [41]), a corresponding archetype node has to be identified, and the code and the node have to be related to each other. This mapping is straightforward: all diagnosis codes are mapped to the node "Diagnosis" of the archetype "Diagnosis", and all procedure codes to the node "Procedure" of the archetype "Procedure undertaken".

7.4.2.2 *Defining relations between assigned concepts of the information model*

Subsequently, relations between the assigned concepts of the information model have to be defined, i.e. relations that concern the instances to be queried. As intra-archetype relations are part of the archetype definition, only inter-archetype relations need to be defined. According to the problem-oriented patient model paradigm, all procedures should be related to the diagnosis that they are associated with, and this should be feasible via the node "Reason/s for procedure" of the archetype "Procedure undertaken". Unfortunately, such relations are not present in our data warehouse, so that performing this substep was not possible. We related patients to their diagnoses and procedures, and the "number of examined lymph nodes" to the "lymph node examination" via the property "links" that stems from *open*EHR's reference model.

7.4.3 *Constructing Archetyped SPARQL Queries*

The SPARQL queries were constructed based on the mapping between relevant elements occurring in the quality indicators and their corresponding (elements of) *open*EHR archetypes. Defining quality indicators with the help of archetype elements makes them, in principle, computable across systems that make use of the required archetypes to store clinical data. We defined the graph patterns to be matched based on the translated OWL classes and properties and their inter- and intra-archetype relations.

Patients with SNOMED CT classes or subclasses identified in the indicator were retrieved with the help of the SNOMED CT closure. For brevity[12], the following query-extract shows only a query for archetyped patients with diagnoses of the SNOMED CT concept 93761005, i.e. "Primary malignant neoplasm of colon (disorder)":

```
PREFIX patient: <http://few.vu.nl/~kdr250/archetypes/openEHR-DEMOGRAPHIC-PERSON.person-patient.v1.owl#>
PREFIX diagnosis: <http://few.vu.nl/~kdr250/archetypes/openEHR-EHR-EVALUATION.problem-diagnosis.v1.owl#>
PREFIX schemarm: <http://klt.inf.um.es/~cati/ontologies/OpenEHR-SP-v2.0.owl#>
PREFIX sct: <http://www.ihtsdo.org/>
SELECT DISTINCT ?patient WHERE {
    ?patient a patient:at0000.1_Patient .
    ?patient schemarm:links ?diagnosis .
    ?diagnosis a diagnosis:at0000.1_Diagnosis .
    ?diagnosis schemarm:value_element ?diagnosiscode .
    ?diagnosiscode a diagnosis:at0002.1_Diagnosis .
    ?diagnosiscode a sct:SCT_93761005 .
} ORDER BY ?patient
```

7.4.4 *Calculating the Indicators by Running the Queries*

Table 16 compares our computed indicator results to the results contained in the report generated for our hospital by the DSCA, and the results publicly reported[13].

The results reported here are a first approximation, and a thorough analysis is required to determine their reliability, validity and all causes for differing results. As a first evaluation, we analysed the results for the denominator of the indicator "Patients with rectum carcinoma who have been discussed in a preoperative multidisciplinary meeting", which retrieves patients with rectum carincoma who have been operated in the reporting year. The query on the DSCA dataset retrieves 21 patients, whereas the query on the data warehouse retrieves 24. Three out of the 21 patients retrieved on the DSCA dataset were not mapped to patients of our data warehouse. Thus, the query on the data warehouse retrieved 6 patients

12 Translated archetypes, extract of synthetic patient data and constructed queries: http://www.few.vu.nl/~kdr250/archetypes/

13 http://www.ziekenhuizentransparant.nl/

Table 16: Comparison of our results to those reported by the DSCA and publicly reported results. Note that some of the indicator definitions and interpretations differ: For example, the re-operation indicator publicly reported includes all colorectal operations, and not only those due to a colorectal carcinoma. Also, it defines re-operations as having taken place within 30 days after the operation, while our indicator - as specified in the indicator description - in addition includes re-operations during the same admission.

Indicator / Results	Our Result	DSCA	Publicly Reported
Lymph nodes	85,71% (42/49)	80,00% (43/54)	-
Meeting	91,66% (22/24)	100% (21/21)	-
Re-operation	1,66% (1/60)	9% (7/75)	8,33% (20/240)

who were not retrieved by the other query. All of these patients are registered with a carcinoma located in the *Colon sigmoideum* in the DSCA dataset. In the data warehouse, this is represented with the ICD-9 code 154.00 (Malignant neoplasm of rectosigmoid junction) in all cases. Via the ICD-9 to SNOMED crossmap, this code is mapped to 4 different SNOMED CT concepts, some of which are subconcepts of "Primary malignant neoplasm of rectum (disorder)", which is employed in the indicator query. Thus, these patients are retrieved as rectum carcinoma patients, whereas they have been classified as colon carcinoma patients by the surgeon who entered the data.

7.5 DISCUSSION

This section discusses the most notable lessons learned during our case study.

7.5.1 *Differing Indicator Results and Encoded Data.*

Besides from differing indicator definitions and interpretations, differing indicator results are likely to be caused by missing patients, who could be not be mapped from the DSCA dataset to our data warehouse based on their properties. The fact that not all patients could be mapped indicates insufficient data quality. Another cause might be insufficient encoding: only a little more than half of the diagnoses in our data warehouse are encoded in ICD-9. Also, no mapping from the Dutch procedure classification to SNOMED CT exists. We detected that patients who have been classified to have a carcinoma in the colon sigmoideum by our surgeons are retrieved as rectum carcinoma patients. This might be due to an incorrect ICD-9 code in the data warehouse. If those patients are indeed sigmoid

colon carcinoma patients, the ICD-9 code 153.3 (Malignant neoplasm of sigmoid colon) would have been preferable. The general question remains whether ICD-9 is suitable for our use case, as it is not intended to support the secondary use of clinical data. Routine data must be of sufficient quality, structured, complete, and encoded in detailed, correct concepts from standard terminologies to be (re)usable. Clinical quality indicators must be well-formalised so that comparable results can be obtained. In future, we will investigate the effect of data quality on the reliability and validity of obtained indicator results.

7.5.2 *Coverage of the* openEHR *archetype repository.*

Modelling both the patient data and the quality indicators at the archetype level was intuitive. With regard to the coverage of the *open*EHR archetype repository, we were able to map nearly all required elements to (elements of) archetypes. A missing element was "discharge date", which we expected to be present as a node "discharge date" in the "admission" archetype, or as a separate archetype. Editing the "admission" archetype to fit our requirements was easy, and it would have been possible to contribute our addition to the public repository. In total, we made use of 6 nodes from 5 archetypes.

7.5.3 *Archetypes in OWL and Properties.*

The Archetype Ontologizer proved to be useful after some minor adaptions, and working with the OWL representation of archetypes was practical due to the wide range of Semantic Web tools available.

Regarding the archetyped patient data, all employed properties stem from the reference model, except from *hasTime*, *hasNumber* and *hasBoolean*. In OWL, XML Schema datatypes are used in typed literal values, while the reference model defines datatypes such as DV_DATE_TIME. As literals can not be instances of classes by definition, the relationship between the literals and the classes can not be expressed directly. Defining properties between OWL classes of the archetypes and their instances was complex, as it was unclear which properties would be the correct ones to use for inter-archetype relationships. We chose the property "links" from the reference model to relate patients to their diagnoses and procedures. The use of more meaningful alternatives will be explored in future work. In our data warehouse, procedures are not related to the diagnoses due to which they have been carried out. This forces us to employ heuristics (e.g., a procedure is typically being carried out after the corresponding diagnosis has been recorded), which might negatively impact the validity of indicator results.

7.5.4 *Automated Reasoning with Patient Data and Information Models in OWL: Past and Future*

Many researchers have demonstrated the added value of patient information models represented in OWL. Lezcano et al. [95] integrated archetypes in OWL 2 with SWRL (Semantic Web Rule Language) rules, which are then to be applied to instances of clinical data. Rector et al. [98] represented a set of information models and bindings to a coding system (i.e. allowed codes) in OWL and validated it with a reasoner. They also validated whether individual data structures conform to the information model with the help of added closure axioms. In a comparable study, the *open*EHR library of archetypes was translated into OWL classes and subsequently validated with OWL reasoners [99]. Heymans et al. [100] formalised a subset of the constraints in the implementation guide on *Using SNOMED CT in HL7 Version 3* as OWL Integrity Constraints and automatically validated CDA documents using the OWL 2 reasoner Pellet.

The OWL representation of archetypes and patient data opens new opportunities for automated reasoning: First, reasoning may be useful at a patient data level. The massive data sources currently locked up in EHRs contain a wealth of implicit knowledge that could be made explicit by formal reasoning. In addition, the OWL representation of archetypes could be used to validate whether the patient data fulfils the constraints defined in the corresponding archetypes. For example, it could be checked whether the number of examined lymph nodes is indeed greater than or equal to 0. Finally, it may be possible to infer archetype class memberships for patient data. Reasoning is also required at the archetype-level: It is unrealistic to expect publicly available archetypes to be expressive enough to cover all possible clinical concepts required for all kinds of use cases. Thus, users of the two-level methodology define their own archetypes, and it is important to be able to infer subsumption and equivalence relationships between self-defined and publicly available archetypes. Finally, as information models and terminologies are developed independently from each other, they may overlap, and different systems and users will make different modelling choices. It must be possible to detect semantically equivalent constructs.

7.6 CONCLUSION

Our research question for this paper was whether open*EHR archetypes are suitable to semantically integrate patient data and quality indicators*, with the goal to reuse routine patient data for secondary purposes such as the computation of indicators. Mapping both our local database schema and elements of patient data occurring in indicators to (elements of) archetypes

was intuitive. This can be attributed both to the two-level methodology, which also makes the resulting queries computable across institutions employing the required archetypes, and the high coverage of *open*EHR's public Clinical Knowledge Manager. We edited an existing archetype to fit our requirements. Based on our mappings, we archetyped the patient data and formalised our sample set of indicators as SPARQL queries with our indicator formalisation method CLIF. We ran the resulting queries against the archetyped patient data to prove the concept. Since *open*EHR archetypes are applicable to represent both patient data and elements of patient data required to compute clinical quality indicators, we conclude that they are suitable for semantic integration of patient data and quality indicators. Further research is required into the potential benefit of automated reasoning based on the OWL representation of archetyped patient data.

Part III

REASONING AND ONTOLOGIES FOR SEMANTIC INTEROPERABILITY

8 COMPARISON OF REASONERS FOR LARGE ONTOLOGIES IN THE OWL 2 EL PROFILE

This paper provides a survey to and a comparison of state-of-the-art Semantic Web reasoners that succeed in classifying large ontologies expressed in the tractable OWL 2 EL profile. Reasoners are characterized along several dimensions: The first dimension comprises underlying reasoning characteristics, such as the employed reasoning method and its correctness as well as the expressivity and worst-case computational complexity of its supported language and whether the reasoner supports incremental classification, rules, justifications for inconsistent concepts and ABox reasoning tasks. The second dimension is practical usability: whether the reasoner implements the OWL API and can be used via OWLlink, whether it is available as Protégé plugin, on which platforms it runs, whether its source is open or closed and which license it comes with. The last dimension contains performance indicators that can be evaluated empirically, such as classification, concept satisfiability, subsumption checking and consistency checking performance as well as required heap space and practical correctness, which is determined by comparing the computed concept hierarchies with each other. For the very large ontology SNOMED CT, which is released both in stated and inferred form, we test whether the computed concept hierarchies are correct by comparing them to the inferred form of the official distribution. The reasoners are categorized along the defined characteristics and benchmarked against well-known biomedical ontologies. The main conclusion from this study is that reasoners vary significantly with regard to all included characteristics, and therefore a critical assessment and evaluation of requirements is needed before selecting a reasoner for a real-life application.

8.1 INTRODUCTION

Ontologies are formal definitions of concepts and the relationships between them. The ontology language OWL 2[1] is a W3C Recommendation since 2009. It is based on Description Logics (DLs) [101], a family

1 http://www.w3.org/TR/owl2-overview/

of knowledge representation formalisms. OWL 2 has three tractable profiles[2], i.e. logical fragments that trade expressive power for the efficiency of reasoning. Each profile is restricted to a different sublanguage of OWL 2. Which profile to choose for a given application scenario depends on the structure of the employed ontology and on the required reasoning tasks. The three profiles are OWL 2 RL, OWL 2 QL and OWL 2 EL. OWL 2 RL (Rule Language) reasoning systems allow for rule-based reasoning. OWL 2 QL (Query Language) supports conjunctive query answering against large volumes of instance data that is stored in relational database systems. OWL 2 EL aims at applications that employ large ontologies. This profile is sufficiently expressive for many biomedical ontologies, such as the very large ontology SNOMED CT [27], and basic reasoning problems for OWL 2 EL can be decided in polynomial time. As indicated by its acronym EL, the profile is based on the $\mathcal{EL}$ family of description logics that provide only existential (and no universal) quantification.

A reasoner is a program that infers logical consequences from a set of explicitly asserted facts or axioms and typically provides automated support for reasoning tasks such as classification, debugging and querying. For OWL 2 EL, scalable implementations of dedicated reasoning algorithms are available. A question is whether these implementations perform better on OWL 2 EL ontologies than traditional reasoning engines, which have been designed for much more expressive languages. Tableau algorithms can be highly optimized [101], so that they are not necessarily outperformed by straightforward implementations of polynomial-time algorithms [102].

Ontologies consist of two different types of statements: TBox statements describe intensional knowledge, that is terminological background knowledge, while ABox statements describe extensional knowledge about individuals. The experiments of this study are limited to (very large) TBoxes.

The main contribution of this paper is the *identification of reasoner characteristics* that influence the choice of a particular reasoner for a given application scenario. A second contribution is the *categorization of dedicated OWL 2 EL and tableau-based reasoners along these characteristics*. To categorize the reasoners along performance indicators, a benchmark is employed that is based on three biomedical ontologies and comprises several TBox reasoning tasks. This categorization can be used to make a well-motivated choice for a particular application. The remainder of this paper is organized as follows: The next section summarizes related work. Section 8.3 gives an overview of characteristics that are relevant to compare reasoning engines. Those characteristics are grouped in the three dimensions reasoning characteristics, practical usability and performance indicators. Section

2 http://www.w3.org/TR/owl2-profiles/

8.4 presents the eight reasoners that are included in this study. Section 8.5 contains the classification of the reasoners as well as the experimental results on their performance. Section 8.6 discusses the results and their implications.

8.2 RELATED WORK

A benchmark typically comprises a selection of employed ontologies and a number of standard TBox and ABox reasoning tasks and serves as a basis to evaluate and compare reasoners. With the increasing availability of reasoners for OWL and OWL 2 EL, several benchmarks have been proposed.

The Lehigh University Benchmark (LUBM) [103] and the University Ontology Benchmark (UOBM) [104], which is an extension of the LUBM, are based on synthetically generated ontologies. LUBM evaluates the performance of answering conjunctive queries over an ABox of varying size that commits to an OWL Lite ontology. Additionally, LUBM measures correctness by examining query completeness and soundness.

A framework for an automated comparison of DL reasoners that focusses on TBox classification is presented in [105]. It is based on real-life ontologies and allows users to compare the classification performance of reasoners as well as to analyze the "correctness" of classification by comparing computed concept hierarchies. This benchmarking system is based on the DIG standard [106], which facilitates the comparison of DIG-compliant reasoners such as FaCT++ [29], KAON2 [107], Pellet [108] and RacerPro [109].

The authors of [110] aim at providing guidance for the nontrivial task of choosing an appropriate reasoner for a given application scenario. The paper surveys the ontology landscape and defines a benchmark, which includes classification as representative TBox reasoning task and conjunctive query answering as ABox reasoning task. Employed performance measures contain load time and response time for ontologies that are representative for identified language fragments. The OWL 2 EL fragment is not included in this study, but the authors state that the investigation of tractable fragments of OWL and the development of reasoners specialized for these fragments is an important research topic. Reasoners are grouped into three categories according to their underlying reasoning techniques: tableau-based algorithms (HermiT [111], RacerPro and Pellet), datalog engines (KAON2) and standard rule engines (Sesame [33] and OWLIM [32]).

A comprehensive survey of OWL reasoners that aims to serve as a decision help for Semantic Web application designers is provided by [112].

Reasoners are described in the categories "official OWL specification language conformity", correctness, efficiency, interface capabilities and inference services. Included reasoners are FaCT++, RacerPro, Pellet, KAON2 and Hoolet, an OWL DL reasoner that uses the first-order theorem prover Vampire [113]. The correctness of reasoners is evaluated by running inference test cases for selected language features.

The survey [114] employs a benchmark suite for large ABox data, as well as a selection of small but difficult T- and ABox test cases. It analyzes the correctness of the results of FaCT++, Pellet, RacerPro, KAON2 and HermiT. The authors extend the challenge of finding an optimal OWL reasoner for a specific application by finding an optimal service interface. The performance of several protocols is compared in different computing environments, which leads to the conclusion that those components may have a high impact. Additionally, the paper contains a feature matrix of selected system characteristics including available interfaces such as the OWL API, Jena [115], DIG and OWLlink, language support in terms of expressivity, retraction, incremental reasoning, SWRL support, query language and query entailment, as well as available licenses and implementation language.

The SEALS (Semantic Evaluation at Large Scale) project provides an infrastructure to evaluate semantic technologies. Its Storage and Reasoning Systems Evaluation Campaign 2010[3] includes evaluation scenarios for standard inference services such as classification, concept satisfiability, ontology satisfiability and logical entailment. In the scope of the 2010 campaign, the reasoning engines HermiT, FaCT++ and jcel[4] have been evaluated based on an OWL 2 repository and widely-used real-world ontologies.

Recently, Mishra et al. [116] presented an extensive survey of nineteen reasoners that have been released between 1975 and 2009. The authors compare these reasoners with respect to their inference support, completeness and algorithm, implementation language and supported Semantic Web languages.

Developers of dedicated OWL 2 EL reasoners have been comparing their classification performance to other reasoners. All these comparisons employ life-science ontologies in OWL 2 EL as benchmark ontologies: the Gene Ontology (GO), a large ontology from the US National Cancer Institute (NCI), the Foundational Model of Anatomy (FMA), the Generalized Architecture for Languages, Encyclopaedias and Nomenclatures in medicine (GALEN) and the Systematized Nomenclature of Medicine, Clinical

3 http://www.seals-project.eu/seals-evaluation-campaigns/storage-and-reasoning

4 http://jcel.sourceforge.net/

Terms (SNOMED CT). All mentioned experiments except [117] (that summarizes common characteristics of several life-science ontologies and suggests that the use of DLs in the $\mathcal{EL}$ family is beneficial both in terms of expressivity and of scalability, and also promotes CEL's reasoning services) have been performed with the goal to demonstrate that the respective newly introduced or re-introduced reasoner outperforms existing reasoners, with TBox classification performance as the only dimension for comparison. In the following, the classification performance with regard to SNOMED CT measured in these studies is briefly outlined.

CEL (Classifier for EL) [118] has been compared to FaCT and Racer in 2005 [102]. In this study, CEL was the only reasoner successful in classifying SNOMED CT, which took around 3.5 hours. FaCT and Racer failed due to memory exhaustion (the test machine had 2GB memory). The fact that CEL succeeded in classifying SNOMED CT motivated the DL community to investigate optimizations that exploit the simple structure of biomedical ontologies. In 2006, CEL was compared to FaCT++, RacerMaster and Pellet [119]. CEL and FaCT++ were successful in classifying SNOMED CT, while RacerMaster and Pellet failed (512MB; Java heap space set to 256MB). FaCT++ needed a little more than an hour and CEL completed just under half an hour. CEL was compared to FaCT++, HermiT, KAON2, Pellet and RacerPro in 2008 [117]. Here, CEL (around 20 minutes), FaCT++ (around 10 minutes) and RacerPro (around 20 minutes) succeeded. KAON2 failed due to a timeout after 24 hours, while HermiT and Pellet failed due to memory exhaustion (machine with 2GB memory, heap space set to 1.5GB). The same results are presented in [118]. The consequence-based reasoner CB has been compared to FaCT++, Pellet, HermiT and CEL in 2009 [30]. CB classified SNOMED CT in less than a minute, FaCT++ in around 10 minutes and CEL in around 20 minutes, while HermiT and Pellet failed to return a result (1.5GB RAM, 1GB heap space; timeout 1 hour). Finally, various Protégé Plugins (Snorocket, FaCT++, Pellet and CEL) have been compared in 2010 [120]. CEL and Pellet failed due to memory exhaustion (4GB memory, 1,900MB maximum heap space). Snorocket classified SNOMED CT in under a minute and FaCT++ in around 20 minutes. Figure 17 shows how the classification performance for SNOMED CT has been improving over the recent years. Only successful outcomes (i.e. no timeout or memory exhaustion results) are included in this figure.

Classification performance is indeed essential and the fact that very large ontologies such as GALEN or SNOMED CT can be classified at all and within a reasonable time is a remarkable achievement of the recent years. But when a reasoner is to be applied in a real-world setting, many more orthogonal aspects are relevant. In the following section, we will identify those characteristics and group them into three dimensions.

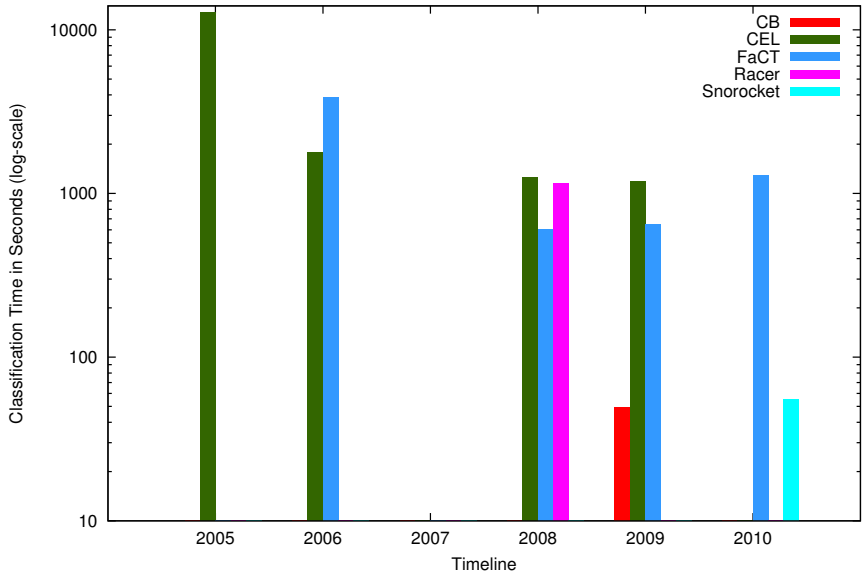

Figure 17: Classification Performance for SNOMED CT over Time

8.3 CHARACTERISTICS

The characteristics described in this section stem from a literature review that included the related work. We analyzed papers that describe the reasoners contained in this study (see Section 8.4) as well as short advertising descriptions of reasoners, which usually outline the respective reasoner's strong points. Additionally, many characteristics in the dimension of practical usability arose while the reasoning experiments have been performed. The characteristics are arranged in three dimensions: reasoning characteristics, practical usability and performance indicators.

8.3.1 *Dimension Reasoning Characteristics*

Methodology

Most DL reasoners are based on (hyper)tableau calculi [121, 122], which are sound and complete. Such procedures aim at large expressivity and, according to [30], classify an ontology by iterating over all necessary pairs of concepts and trying to build a model of the ontology that violates the subsumption relation between them. New kinds of reasoning procedures have been developed for less expressive, tractable DLs such as $\mathcal{EL}^{++}$ [24], [25]. The procedure for $\mathcal{EL}^{++}$ infers subsumption relations by using so-called completion rules in a goal-directed way.

Soundness and Completeness in Theory

This property evaluates whether the inferences of the employed reasoning methods are sound based on the underlying semantics and whether they are complete, i.e. whether all possible inferences are inferred. Soundness or completeness can be sacrificed for a significant speed-up of reasoning [123]. Thus, to employ a reasoner in a real-world application, it is not always important *that* its underlying reasoning method is sound and complete, but it is important to know *whether* it is sound and complete. Most of the methods underlying the reasoners included in this study have been proven to be sound and complete, i.e. correct. This does not imply that their implementations are correct.

Expressivity and Computational Complexity

For description logics, a tradeoff exists between logical expressivity and computational complexity: the more expressive a language, the higher its computational complexity. Reasoning problems in OWL DL and OWL 2 are, in the worst case, solvable in time that is (double) exponential with respect to the size of the input. However, hard cases that lead to worst-case behavior rarely occur in practice [124], [121]. When the input ontology is in a tractable profile such as OWL 2 EL, it is theoretically possible for reasoners that support a more expressive language to terminate in polynomial time. Table 17 lists several DLs with their corresponding worst-case complexities for concept satisfiability checking taken from the literature.

Table 17: Worst-Case Complexities of Concept Satisfiability Checking

Logic	Worst-Case Complexity
$\mathcal{EL}^{++}$	PTime [24], [25]
Horn $\mathcal{SHIQ}$	ExpTime [30]
$\mathcal{SHIQ}$	ExpTime [125]
$\mathcal{SHOIQ}$ (OWL DL)	NExpTime [125]
$\mathcal{SROIQ}$ (OWL 2)	N2ExpTime [126]

The expressivity of a particular DL is determined by the concept constructors it provides. The informal naming convention for DLs describes the constructors that can be used: $\mathcal{E}$ (existential restrictions), $\mathcal{Q}$ (qualified number restrictions), $\mathcal{O}$ (nominals, objects), $\mathcal{I}$ (inverse roles), $\mathcal{H}$ (role hierarchies) and $\mathcal{S}$ is the abbreviation for $\mathcal{ALC}$ with transitive roles. The basic description logic $\mathcal{ALC}$ uses the constructors $\neg C$ (negation), $C \sqcap D$ (conjunction), $C \sqcup D$ (disjunction), $\exists R.C$ (existential restriction) and $\forall R.C$ (value restriction).

THE DESCRIPTION LOGIC $\mathcal{EL}^{++}$

The language $\mathcal{EL}$ [127] allows for concepts constructed from atomic concepts A and the top concept $\top$ (i.e. owl:Thing) by using the constructors conjunction $C \sqcap D$ (i.e. owl:ObjectIntersectionOf) and existential restriction $\exists r.C$ (i.e. owl:ObjectSomeValuesFrom), where r is an atomic role. Axioms of $\mathcal{EL}$ are general concept inclusions (GCI) $C \sqsubseteq D$. A primitive concept definition (PCD, necessary but not sufficient) $A \sqsubseteq D$ is a GCI with a concept name on the left-hand side, while a full concept definition (FCD, necessary and sufficient) $A \equiv D$ can be expressed by the two GCIs $A \sqsubseteq D$ and $D \sqsubseteq A$. A finite set of GCIs is called a TBox. Concepts and roles correspond to OWL classes and properties. $A \sqsubseteq D$ corresponds to owl:SubClassOf(A, D) and $A \equiv D$ corresponds to owl:EquivalentClasses(A, D).

$\mathcal{EL}^+$ [102] extends $\mathcal{EL}$ by complex role inclusions which allow to express role hierarchies, transitive roles and right identities, such as: $r \circ s \sqsubseteq t$, e.g. has-parent $\circ$ has-sister $\sqsubseteq$ has-aunt. The example expresses that two individuals that are connected by a chain of roles (has-parent $\circ$ has-sister) are necessarily connected by the role on the right-hand side (has-aunt). $\mathcal{EL}^+$ is sufficiently expressive for many well-known biomedical ontologies such as the ones in our test-suite.

$\mathcal{EL}^{++}$ [24], [25] extends $\mathcal{EL}^+$ by nominals (and thus ABoxes), the bottom concept $\bot$ (and thus disjointness constraints on concepts in the form of $C \sqcap D \sqsubseteq \bot$), reflexive roles and range restrictions $range(r) \sqsubseteq C$ (a global syntactic restriction applies to guarantee polynomiality) and a restricted form of concrete domains (e.g. references to numbers and strings; datatypes in OWL). $\mathcal{EL}^{++}$ is one of the few description logics for which standard reasoning problems such as ontology consistency, concept subsumption, and instance checking are decidable in polynomial time. To gain this tractability, commonly-used constructors such as universal value restrictions and inverse and functional roles have been sacrificed. For most biomedical ontologies, scalable reasoning seems to be more important than the expressivity of the language [128]. A complete description of $\mathcal{EL}^{++}$ and its formal semantics is given in [25], and the structure of OWL 2 EL ontologies is specified in the OWL 2 EL profile specification[5]. Mappings between $\mathcal{EL}^{++}$ and OWL 2 EL can be found in [118]. Further tractable extensions on $\mathcal{EL}^{++}$ exist [129], [130].

As an example, let us consider the ontology in Figure 18. In SNOMED CT, concepts form logical groupings, which are expressed by nested existential restrictions. All concepts that are allowed to be grouped

5 http://www.w3.org/TR/owl2-profiles/#OWL_2_EL

$$
\begin{aligned}
&\textit{Fracture} \sqsubseteq \textit{Traumatic abnormality by morphology}\\
&\textit{Traumatic abnormality by morphology} \sqsubseteq \textit{Traumatic abnormality}\\
&\textit{Traumatic abnormality} \sqsubseteq \textit{Damage}
\end{aligned}
$$

$$
\begin{aligned}
&\textit{Bone structure of foot} \sqsubseteq \textit{Bone structure of ankle and/or foot}\\
&\textit{Bone structure of ankle and/or foot} \sqsubseteq \textit{Bone structure of lower limb}
\end{aligned}
$$

$$
\begin{aligned}
\textit{Fracture of bone} \equiv\ & \textit{Disorder of bone} \sqcap \exists \textit{rolegroup}.\\
& (\exists \textit{associated morphology.Fracture}\\
& \sqcap \exists \textit{finding site.Bone structure})
\end{aligned}
$$

$$
\begin{aligned}
\textit{Fracture of lower limb} \equiv\ & \textit{Fracture of bone} \sqcap \exists \textit{rolegroup}.\\
& (\exists \textit{associated morphology.Fracture}\\
& \sqcap \exists \textit{finding site.Bone structure of lower limb})
\end{aligned}
$$

$$
\begin{aligned}
\textit{Fracture of foot} \equiv\ & \textit{Fracture of lower limb} \sqcap \exists \textit{.rolegroup}\\
& (\exists \textit{associated morphology.Fracture}\\
& \sqcap \exists \textit{finding site.Bone structure of foot})
\end{aligned}
$$

Inferred:

$$
\begin{aligned}
&\textit{Fracture} \sqsubseteq \textit{Damage}\\
&\textit{Bone structure of foot} \sqsubseteq \textit{Bone structure of lower limb}\\
&\textit{Fracture of foot} \sqsubseteq \textit{Fracture of lower limb}
\end{aligned}
$$

Pellet's explanation for $\textit{Fracture of foot} \sqsubseteq \textit{Fracture of lower limb}$:

$$
\begin{aligned}
\textit{Fracture of foot} \equiv\ & \textit{Fracture of lower limb} \sqcap \exists \textit{.rolegroup}\\
& (\exists \textit{associated morphology.Fracture}\\
& \sqcap \exists \textit{finding site.Bone structure of foot})
\end{aligned}
$$

Figure 18: An example $\mathcal{EL}$ ontology (motivated by SNOMED CT)

are included under an existential restriction that represents the (potential) grouping. This restriction is labeled with an owl:ObjectProperty named rolegroup. Attributes (i.e. Object Properties or roles) are presented in lower case. The figure contains five partially defined concepts (Fracture, Traumatic abnormality by morphology and Traumatic abnormality, Bone structure of foot and Bone structure of ankle and/or foot) that form hierarchies. The three fully defined concepts are Fracture of bone, Fracture of lower limb and Fracture of foot. It can be inferred that a Fracture is a Damage, i.e. $\textit{Fracture} \sqsubseteq \textit{Damage}$, that $\textit{Bone structure of foot} \sqsubseteq \textit{Bone structure of lower limb}$ and that Fracture of foot is a Fracture of lower limb, i.e. $\textit{Fracture of foot} \sqsubseteq \textit{Fracture of lower limb}$.

Incremental Classification

When an ontology has been classified and is updated afterwards (by additions or removals), it makes sense for a reasoner to reuse the previous classification information together with the updated axioms to produce the new concept hierarchy. This is especially reasonable in typical ontology development scenarios that involve only minor modifications between classifications that are performed to check whether the developed ontology

is (still) consistent. Alternatively, the reasoner has to re-start the whole classification from scratch, which can be time-consuming.

Rule Support

Rule support enables the combination of ontologies with rules. Some reasoners support SWRL[6] rules. SWRL, the Semantic Web Rule Language, extends the set of OWL axioms to include Horn-like rules. A simple exemplary rule is to assert that the combination of the *hasParent* and *hasBrother* properties implies the *hasUncle* property: $hasParent(?x1,?x2) \wedge hasBrother(?x2,?x3) \Rightarrow hasUncle(?x1,?x3)$. In contrast to most other rules, this simple example can also be expressed by a complex role inclusion (i.e. owl:ObjectPropertyChain).

Justifications

Justifications are minimal entailing subsets of an ontology [131]. Given an ontology and an unclear consequence, may it be a subsumption relationship or an unsatisfiable concept, it can be very helpful if a reasoner computes a justification (or all justifications) for the consequence, which can subsequently be used to explain or debug that consequence. The OWL API contains a method that returns all explanations for a given unsatisfiable concept, or an empty set if the concept is satisfiable.

Support of ABox Reasoning Tasks

ABox reasoning is reasoning with individuals and comprises instance checking, (conjunctive) query answering and ABox consistency checking. Instance checking tests whether a knowledge base entails that an individual is an instance of a concept. It is the basis of query answering, which can be performed by iterating instance checking for all individuals in a knowledge base [132]. Whether a reasoner supports ABox reasoning tasks or not is a characteristic that, depending on the intended application, can be very relevant.

6 http://www.w3.org/Submission/SWRL/

8.3.2 *Dimension Practical Usability*

OWL API

The OWL API [28] is an Application Programming Interface (API) for working with OWL ontologies. It supports parsing and writing in the syntaxes that are defined in the OWL 2 specification. The open source reference implementation in Java includes validators for OWL 2 profiles. It also provides a standard interface to OWL reasoners, so that an application can embed different reasoners without having to change its implementation. A number of existing reasoners provide OWL API wrappers and are thus easily integrated into OWL API based applications such as Protégé 4.

OWLlink

OWLlink [133] provides an extensible, implementation-neutral protocol to interact with OWL 2 reasoners. It succeeds the DL-oriented DIG interface. OWLlink facilitates client applications to manage reasoners, to assert axioms and to access reasoning services via a set of standard queries. The OWL API based OWLlink API implements the OWLlink protocol. It allows to turn OWL API aware reasoners into OWLlink servers and to access remote OWLlink servers from OWL API based applications (such as Protégé).

Availability as Protégé Plugin

Protégé is an open source ontology editor. The new version 4.1 fully conforms with the OWL 2 language specification and is built on top of the OWL API. It is a common practice of reasoner developers to release a plugin for Protégé. OWL API aware reasoners can also be used from Protégé via OWLlink.

License

Many reasoners come with a dual license. This means that they are free under certain conditions, and that for different use, arrangements have to be made with their developers. The major distinguishing feature concerning licenses is whether the license is a recognized open source license[7] or not.

7 http://www.opensource.org/licenses

Further Characteristics

The remaining characteristics are self-explanatory and include whether the source of the reasoner is open or closed, the programming language the reasoner is implemented in, the supported platforms, whether the reasoner has a native Jena[8] interface and the kind of institution (academic, governmental or commercial) it has been developed in. Jena is a Java framework for building Semantic Web applications.

8.3.3 *Dimension Performance Indicators*

The dimension performance indicators contains characteristics that depend on the input ontology and that can be measured empirically. At the ontology level, fundamental reasoning services include classification and consistency checking. The two most important reasoning services at the concept level are satisfiability checking and subsumption. In our experiments, performance indicators are measured based on these reasoning tasks. The performance of ABox reasoning tasks is not included in this study. Another characteristic that is included is the minimum required amount of heap space for Java reasoners. Finally, we analyze classification results in order to check whether the reasoners' theoretical correctness is confirmed in practice.

Classification Performance

Classification, i.e. the computation of the concept hierarchy, is one of the most important reasoning services and supported by all modern DL systems. Thus, its duration is often used as a performance indicator to benchmark reasoning engines. From a practical perspective, an ontology should be classified regularly during its development and maintenance in order to detect unwanted subsumptions as soon as possible. To make this feasible also for large ontologies, classification should be fast.

TBox Consistency Checking Performance

An interpretation I is a model of an ontology O if the interpretation satisfies all implications in O. An ontology is consistent if it has a model [134].

8 http://jena.sourceforge.net/

Concept Satisfiability Checking Performance

A concept satisfiability check tests whether a concept C can have instances. According to [101], satisfiability is formally defined as follows: A concept C is satisfiable with respect to a TBox T if there exists a model I of T such that C^I is nonempty. Concept satisfiability checks are a special case of concept subsumption checks, because a concept C is unsatisfiable if, and only if, $C \sqsubseteq \bot$ [128].

Subsumption Query Performance

Subsumption queries check whether one concept subsumes another concept or return all concepts subsuming or subsumed by a concept. According to [101], a concept C is subsumed by a concept D with respect to T if $C^I \subseteq D^I$ for every model I of T. This is written as $T \vDash C \sqsubseteq D$.

Soundness and Completeness in Practice

We analyze the output of reasoners in Section 8.5. For most ontologies, their closure, i.e. the set of all statements that follow from the underlying semantics, is not given. In such cases, the only way to evaluate the output of a reasoner is by comparing it to the output of other reasoners. All reasoners that correctly implement a sound and complete reasoning method that supports the expressivity of the input ontology should produce the same output for the same input. Thus, if the outputs of the reasoners differ from each other, we can infer that not all implemented methods are sound and complete in practice. For SNOMED CT, an advantageous situation applies: It is released both in stated and in inferred form, so that we can employ the inferred form as gold standard and compare it to the concept hierarchies computed by the reasoners. The inferred form has been generated with Apelon's[9] Ontylog DL classifier.

8.4 REASONERS

The reasoners which are compared based on the defined characteristics include the newly introduced reasoner TrOWL and all reasoners that occur in previous comparisons except KAON2, because it is not being maintained any longer. This section briefly describes each reasoner.

CB (Consequence-based reasoner, University of Oxford) is an implementation of a reasoning procedure [30] for Horn $\mathcal{SHIQ}$ ontologies, i.e. $\mathcal{SHIQ}$

9 http://www.apelon.com

ontologies that can be translated to the Horn fragment of first-order logic. CB's reasoning procedure can be regarded as an extension of the completion-based procedure for $\mathcal{EL}^{++}$ ontologies and works by deriving new consequent axioms. It is theoretically optimal for Horn $\mathcal{SHIQ}$ ontologies as well as for the common fragment of $\mathcal{EL}^{++}$ and $\mathcal{SHIQ}$ [30].

CEL (Classifier for $\mathcal{EL}$, TU Dresden) [119], [118] implements a refined polynomial-time algorithm [102], [24], [25] which allows it to process very large $\mathcal{EL}^{+}$ ontologies in reasonable time.

FaCT++ (Fast Classification of Terminologies, University of Manchester) [29] is the new generation of the OWL DL reasoner FaCT. It supports OWL DL and a subset of OWL 2 that is more expressive than the ontologies in our test suite. FaCT++ is implemented in C++ and based on optimized tableaux algorithms.

HermiT (University of Oxford) [111] can determine whether or not a given ontology is consistent and identify subsumption relationships between concepts, among other features. HermiT is based on a "hypertableau" calculus.

Pellet (Clark & Parsia) [108] was the first reasoner that supported all of OWL DL ($\mathcal{SHOIN}(\mathcal{D})$) and has been extended to OWL 2 ($\mathcal{SROIQ}(\mathcal{D})$). Pellet supports OWL 2 profiles including OWL 2 EL.

RacerPro (Renamed ABox and Concept Expression Reasoner, Racer Systems) [109] implements the description logic $\mathcal{SHIQ}$. Dedicated optimizations for OWL 2 EL have been added (structural subsumption tests [135]), enabling practical reasoning with SNOMED CT.

Snorocket (CSIRO) [120] is a high-performance implementation of the polynomial-time classification algorithm for $\mathcal{EL}^{+}$ [102]. It was primarily optimized for classifying SNOMED CT, and was licensed to the IHTSDO[10] for integration into the Workbench software used to maintain and produce SNOMED CT.

TrOWL (Tractable reasoning infrastructure for OWL 2, University of Aberdeen) [136] is the common interface to a number of reasoners. TrOWL Quill provides reasoning services over OWL 2 QL. TrOWL REL is an optimized implementation of the CEL algorithm that provides reasoning over OWL 2 EL. It employs a syntactic approximation from OWL 2 DL to OWL 2 EL to enable OWL 2 DL ontologies to be classified within polynomial time [137]. This approximation is soundness-preserving but sacrifices completeness. To support full DL reasoning, TrOWL allows for the use of heavyweight plugin reasoners, such as FaCT++, Pellet, HermiT and RacerPro.

10 http://www.ihtsdo.org/

8.5 CATEGORIZATION OF REASONERS

In this section, the eight reasoners are categorized along the defined characteristics. The evaluation of the performance indicators is based on our test suite, which comprises the three biomedical ontologies GO, NCI and SNOMED CT. Concluding, we analyze the tradeoff between a reasoner's supported expressivity and its classification performance.

8.5.1 *Dimension Reasoning Characteristics*

Table 18 summarizes the reasoning properties for the included reasoners.

Table 18: Reasoning Characteristics

	CB	CEL	FaCT++	HermiT	Pellet	RP	SR	TrOWL (REL)
Methodology	consequence-based	completion rules	tableau-based	hypertableau	tableau-based	tableau-based	completion rules	approximation (completion rules)
Soundness	+	+	+	+	+	+	+	+ (+)
Completeness	+	+	+	+	+	+	+	- (+)
Expressivity	Horn $\mathcal{SHIF}$	$\mathcal{EL}^+$	$\mathcal{SROIQ}(\mathcal{D})$	$\mathcal{SROIQ}(\mathcal{D})$	$\mathcal{SROIQ}(\mathcal{D})$	$\mathcal{SHIQ}(\mathcal{D}-)$	$\mathcal{EL}^+$	third-party reasoner (approximating SROIQ; subset of $\mathcal{EL}^{++}$)
Incremental Classification (addition/removal)	-/-	+/-	-/-	-/-	+/+	-/-	+/-	-/-
Rule Support	-	-	-	+ (SWRL)	+ (SWRL)	+ (SWRL, nRQL)	-	-
Justifications	-	+	-	-	+	+	-	-
ABox Reasoning	-	+	+	+	+ (SPARQL)	+ (SPARQL, nRQL)	-	+ (SPARQL)

RacerPro (RP) and Snorocket (SR) had to be abbreviated due to space limitations. + stands for yes and - for no.

Methodology

The first characteristic is the underlying reasoning methodology. Most reasoners rely on tableau-based methods or on (extensions of) completion rules for $\mathcal{EL}$. Pellet and TrOWL both implement optimized support for OWL 2 EL and their EL reasoners are activated based on the profile of the current ontology.

Soundness and Completeness in Theory

Most of the underlying reasoning methodologies have been proven to be sound and complete. The tableaux and hypertableaux calculi are sound and complete, the procedure for $\mathcal{EL}^{++}$ has been shown to be sound and complete in [24] and the procedure for Horn $\mathcal{SHIQ}$ is sound and complete according to [30]. TrOWL REL is based on the procedure for $\mathcal{EL}^{++}$. TrOWL REL's syntactic approximation from OWL 2 DL to OWL 2 EL is soundness-preserving but possibly incomplete.

Expressivity and Computational Complexity

Table 18 lists the languages that the reasoners support. CB's reasoning procedure described in [30] supports Horn $\mathcal{SHIQ}$, but its implementation supports only Horn $\mathcal{SHIF}$, which is a subset of Horn $\mathcal{SHIQ}$ that does not include cardinality restrictions. The expressivity of TrOWL depends on its configuration. TrOWL REL implements $\mathcal{EL}^{++}$ without datatypes and supports $\mathcal{SROIQ}$ by approximation. If a third-party reasoner is used, then its supported expressivity is the one of this reasoner.

Incremental Classification

CEL, Pellet and Snorocket support incremental reasoning. CEL and Snorocket both have a partial incremental classification functionality that only supports additions. Pellet supports incremental classification and incremental consistency checking. Pellet's incremental classification is based on module extraction: The first time an ontology is classified, Pellet computes modules for each concept. A module is a subset of an ontology which captures "everything" an ontology has to say about a particular sub-signature of the ontology [138]. By current methods, modules contain all justifications for all entailments expressible in their signature [138]. When the concept hierarchy of the ontology is changed, Pellet reclassifies only the affected module. Pellet's incremental reasoning supports axiom addition and removal [139].

The reasoner interfaces of the OWL API facilitate reasoners to expose incremental reasoning support. The API allows a reasoner to listen for ontology changes and to either immediately processes them or to queue them to processes them later [28].

Rule Support

Only HermiT, Pellet and RacerPro offer rule support. All of them support SWRL. HermiT and Pellet support SWRL in the DL-Safe Rules notion, which means rules will be applied only to named individuals in the ontology. RacerPro partially supports SWRL. SWRL is mapped to nRQL, which is RacerPro's native rule and query language.

Justifications

CEL, Pellet and RacerPro support justifications for inconsistent concepts. RacerPro allows to check an ontology for inconsistent (unsatisfiable) concepts and generates an explanation for each inconsistency. Pellet can give a justification for any inference which it can compute.

Support of ABox Reasoning Tasks

In contrast to all other reasoners, CB and Snorocket do not support ABox reasoning tasks. The reasoners that implement OWL API reasoner interfaces should (in theory) support all ABox reasoning tasks that are specified by the OWL API, such as retrieving the set of individuals that have been asserted to be an instance of a concept or retrieving the asserted types of an individual. RacerPro supports nRQL ABox queries and Pellet, RacerPro and TrOWL support SPARQL queries. SPARQL[11], the Query Language for RDF, is a W3C Recommendation since 2008. SPARQL allows to query required and optional graph patterns along with their conjunctions and disjunctions, to test values, to constrain queries by source RDF graph and to specify whether the result should be an RDF graph or a set.

8.5.2 *Dimension Practical Usability*

Table 19 shows how the reasoners are categorized along the defined usability characteristics.

11 http://www.w3.org/TR/rdf-sparql-query/

Table 19: Practical Usability. DuLi stands for dual license. + stands for yes and - for no. All platforms means that the reasoner is available for Windows, Linux, and Mac OS X. n/a abbreviates not applicable. Regarding the institution, a stands for academic, c for commercial and g for governmental.

	CB	CEL	FaCT++	HermiT	Pellet	RP	SR	TrOWL
OWL API	-	+	+	+	+	+	+	+
OWLlink API	-	+	+	+	+	+	-	-
Protégé Plugin	-	+	+	+	+	-	+	
License	DuLi: GLGPL	AP 2.0	GLGPL	GLGPL	DuLi: AGPL	own	own	DuLi: AGPL
Open Source	+	+	+	+	+	-	-	-
Language	OCaml	Common Lisp	C++	Java	Java	Lisp	Java	Java
Platforms	all	Linux	all	all	all	all	all	all
Jena	-	-	-	-	+	-	-	-
Institution	a	a	a	a	c	c	g	a

OWL API

All reasoners except CB are accessible via the OWL API, which is advantageous for applications that wish to access several reasoners via the same interface. The use of the OWL API highly facilitated the execution of our experiments.

OWLlink

Most reasoners are accessible via OWLlink, and future protocol bindings might ease the integration of further reasoners like CB.

Protégé Plugin

All reasoners except CB and RacerPro can be plugged into Protégé. The RacerPro engine can be used as back-end inference system for Protégé via the RACER Protégé Plugin[12] or via OWLlink.

License

CB can be redistributed and / or modified under the terms of the GNU Lesser General Public License (GLGPL) for non-commercial use. Pellet also comes with a dual license: software that is released under a recognized open source license can use Pellet under the terms of the Affero General Public License (AGPL), for other software, another license has to be arranged. This has the advantage that the community benefits from source code that uses Pellet under its open source license. TrOWL may be used under the terms of the AGPL for open source applications and is available under alternative license terms for proprietary, closed-source applications and other commercial applications. CEL comes with the Apache License 2.0 (AP 2.0), FaCT++ and HermiT with the GLGPL. Racer Systems offers several license types, including time-limited educational licenses, trial licenses and commercial licenses. Snorocket formulates its own license[13]. GLGPL, AGPL and AP 2.0 are open source licenses.

Further Characteristics

The remaining rows of Table 19 show further characteristics including whether the source of the reasoner is open or not and the programming language the reasoner is implemented in. Only Pellet has a native Jena interface. Further characteristics are the platforms the reasoner supports and the kind of institution (academic, governmental or commercial) it has been developed in.

12 http://www.uni-ulm.de/in/ki/semantics/owltools
13 http://research.ict.csiro.au/software/snorocket/LICENCE.txt

8.5.3 *Dimension Performance Indicators*

In this section, the biomedical ontology test suite and the experimental setup are being presented. Subsequently, we present the results of our experiments.

Biomedical Ontology Test Suite

The biomedical ontologies presented in this section are well-established and have been used in previous benchmarks. All ontologies are in the tractable OWL 2 profile EL, so that especially in the case of SNOMED CT, the challenge for the reasoners lies in the sheer size of the ontologies. Consult [117] for additional information. The ontologies mainly differ in size, but also in whether they employ fully defined concepts or not. Biomedical ontologies are a typical use-case for OWL 2 EL, but this profile is also applicable in other domains where fast reasoning outweighs expressivity.

GO The Gene Ontology[14] project is an initiative which aims to standardize the representation of genes and gene product attributes. The Gene Ontology (GO) is a controlled vocabulary to describe gene product characteristics and annotation data.

NCI The National Cancer Institute thesaurus[15] is a terminology that covers clinical care and research, as well as public information and administrative activities.

SNOMED CT SNOMED CT[16] consists of around 300,000 primitively and fully defined concepts. It is mainly used to represent clinical information in electronic health records. SNOMED CT contains one property chain (complex role inclusion) which is not used in the TBox.

14 http://www.geneontology.org/
15 http://ncit.nci.nih.gov/
16 http://www.ihtsdo.org/snomed-ct/

GO, NCI and SNOMED CT can be regarded as acyclic $\mathcal{EL}$ TBoxes, i.e. sets of concept definitions without cyclic dependencies. GO has one transitive role, which is a special case of a role inclusion [24]. Also SNOMED CT is extended with role inclusion axioms. Table 20 provides an overview of the properties of the benchmark ontologies. The 1,000 concepts of GO that are neither fully nor primitively defined are just declared as concepts and thus direct subclasses of the top concept, without further definitions. 997 of those concepts are annotated as being obsolete and the 3 remaining concepts are the top-level concepts biological process, molecular function and cellular component. The 17 concepts in the NCI ontology that are neither fully nor primitively defined are Kinds, i.e. the top-level superclasses for all of the concepts defined in the thesaurus. They represent the possible categories that concepts can belong to, such as Anatomy, Biological Processes, Chemicals and Drugs, and Diagnostic and Prognostic Factors. In SNOMED CT, the root concept is neither fully nor partially defined.

Table 20: Benchmark Ontologies

	$\lvert N_{LA} \rvert$	$\lvert N_R \rvert$	$\lvert N_C \rvert$	PDC	FDC
GO	28,897	1	20,465	19,465	0
NCI	46,940	70	27,652	27,635	0
SNOMED CT	292,023	62	292,012	227,315	64,696

$\lvert N_{LA} \rvert$ is the number of logical axioms, $\lvert N_R \rvert$ the number of roles and $\lvert N_C \rvert$ the number of concepts. PDC stands for the number of primitively defined concepts and FDC for the number of fully defined concepts.

Experimental Setup

For the experiments to measure performance indicators, the latest available versions of the included reasoners have been used: CB[17] build 6, CEL[18] plugin 0.4.0 for Protégé 4.1, FaCT++[19] 1.5.0, HermiT[20] 1.3.0, Pellet[21] 2.2.2, RacerPro[22] 2.0 preview, Snorocket Protégé plugin[23] version 1.3.2 and TrOWL[24] 0.5.1. We generated the SNOMED CT ontology with the OWL transformation script from the Stated Relationships Table of the latest (July

17 http://code.google.com/p/cb-reasoner/
18 http://lat.inf.tu-dresden.de/systems/cel/
19 http://owl.man.ac.uk/factplusplus/
20 http://www.hermit-reasoner.com/
21 http://clarkparsia.com/pellet/
22 http://www.racer-systems.com/
23 http://research.ict.csiro.au/software/snorocket
24 http://trowl.eu/

2010) SNOMED CT distribution. GO and NCI have been employed in other benchmarks and are available on http://reasonerben.ch.

For an ideal comparison, it would be desirable to run all reasoners via the same interface, such as the OWL API or OWLlink. Unfortunately, there is no interface that has been implemented by all reasoners. Thus, reasoners were tested separately: CB and RacerPro in batch mode and all other reasoners via the OWL API. For the reasoners which are called via the command line, their runtime-outputs are employed to measure the classification performance. For the reasoners which are used by a Java program via the OWL API, external time measurement is applied. All ontologies in the test suite are employed to compare the performance of the reasoners.

The experiments were performed under Linux 64-bit on a 4x AMD Opteron 8220 dual core, 2800 MHz CPU system with 16GB memory. For Java reasoners, Sun's Java Runtime Environment (JRE) version 1.6.0 was used with a Java HotSpot(TM) 64-Bit Server VM. We did not set a fixed maximum heap space but measured the minimum amount of heap space required to classify SNOMED CT in a separate experiment. The minimum required heap space has been approximated in steps of 0.5GB and the lowest heap space for which the reasoner classified SNOMED CT without crashing has been noted. All stated runtimes are averaged over 10 runs.

Classification Performance

As Table 21 shows, all tested reasoners succeeded in classifying SNOMED CT. For all input ontologies, CB took the least time to compute the subsumption hierarchy. FaCT++ and TrOWL REL are the only reasoners that are faster on NCI than on GO. FaCT++ needed longest to classify GO, RacerPro to classify the NCI ontology and HermiT to classify SNOMED CT (nearly 2 hours). The experiments measured only the classification time, and no loading and / or preprocessing times.

Table 21: Comparison of Classification Time in Seconds

	CB	CEL	FaCT++	HermiT	Pellet	RP	SR	TR
GO	0.34	3.19	20.75	6.48	3.41	10.67	1.54	2.43
NCI	0.65	7.52	11.10	11.75	14.84	52.87	4.31	1.83
S CT	28.08	1,112.23	700.87	6,793.76	1,345.65	3,652.03	101.16	344.93

TBox Consistency Checking Performance

CB has no support for consistency checking. For reasoners that implement OWL API reasoning interfaces, the time that the call `reasoner.isConsistent()` took was measured, before and after classification. Table 22 shows the duration of consistency checking. All stated times are before classification. After the first consistency check and after the classification, it is already known whether the ontology is consistent or not and the second consistency check takes less than one second for all tested reasoners. All input ontologies are consistent and all reasoners returned this result before and after classification.

Table 22: Comparison of Consistency Checking Time in Seconds

	CB	CEL	FaCT++	HermiT	Pellet	RP	SR	TR
GO	-	2.17	0.36	0.00	0.27	-	0.00	0.00
NCI	-	0.65	0.71	0.00	0.38	-	0.00	0.00
SNOMED CT	-	0.88	15.3	0.00	16.78	-	0.00	0.00

The CEL manual states that an ontology must be classified before it can be checked whether the TBox is consistent. TrOWL REL also needs to classify first. When checking for consistency before classification, Snorocket outputs a warning that the ontology is not classified. All reasoners check for consistency rather fast. For tableau-based reasoners, this is probably due to the fact that for consistency checking they construct the model only once.

Concept Satisfiability Checking Performance

To determine the performance of concept satisfiability checking, we measure the time it takes each reasoner to check the satisfiability of each concept in the ontology, before and after classification. CB has no support for concept satisfiability checking. For reasoners that implement OWL API reasoner interfaces, we checked for concept satisfiability with the method `reasoner.isSatisfiable (concept)`. RacerPro does not really distinguish between TBox consistency and satisfiability. It offers the function `check-tboxcoherence` which returns a list of inconsistent / unsatisfiable atomic concepts. If the top concept occurs in this list, all concepts are unsatisfiable. It does not compute the concept hierarchy, so that it is much faster than to classify the TBox.

A more direct and comparable (with respect to RacerPro) way to retrieve unsatisfiable concepts via the OWL API would have been to call

`reasoner.getInconsistentClasses()` for OWL API version 2 or respectively `reasoner.getUnsatisfiableClasses()` for the OWL API version 3, but as we wanted to measure the performance of concept satisfiability checking, we preferred to check the satisfiability of each single concept.

Table 23: Comparison of Concept Satisfiability Checking Time in Seconds

	CB	CEL	FaCT++	HermiT	Pellet	RP	SR	TR
GO BC	-	5.28	0.63	6.23	2.12	5.58	0.01	0.23
GO AC	-	2.08	0.07	0.06	0.12	0.00	0.01	0.03
NCI BC	-	4.46	1.26	11.73	3.47	37.73	0.01	0.06
NCI AC	-	3.19	0.14	0.09	0.18	0.00	0.01	0.04
S CT BC	-	38.42	22.37	5,276.85	56.91	273.45	0.07	5.17
S CT AC	-	34.59	1.76	1.36	6.07	0.00	0.06	0.46

BC stands for before classification and AC for after classification.

Table 23 shows that the reasoners vary significantly in their runtimes. Some runtimes are so low that it can be assumed that they do not support the method, but compute satisfiability during classification and thus return reliable results only after the ontology is classified. For example, TrOWL REL checks after the classification whether the concept is subsumed by owl:Nothing.

Subsumption Query Performance

To test subsumption query performance, we will query for all subclasses of the SNOMED CT concept "Fracture of lower limb", as presented in Figure 18. We will test it in two ways: by querying for direct subclasses of its concept name `SCT_46866001` and by querying for direct subclasses of its anonymous class definition as stated in Figure 18. The employed method of the OWL API is: `reasoner.getSubClasses()`, and the employed method of RacerPro `(tbox-retrieve (?x) (concept ?x has-child))`.

When performing this experiment, we noticed a correlation between the runtimes of the queries and the number of returned subclasses. Thus, Table 24 shows not only the runtimes but also the number of results. The four settings are tested sequentially in one run. CB does not support any subsumption querying. Snorocket returns a NullPointerException when the method is called before classification. FaCT++, HermiT, Pellet and RacerPro classify the ontology when they receive the first query. The number of returned subclasses varies. HermiT, Pellet and RacerPro return all

Table 24: Queries for subclasses of the SNOMED CT concept Fracture of lower limb: Comparison of subsumption query performance in seconds and number of results (i.e. returned subclasses)

	CB	CEL	FaCT++	HermiT	Pellet	RP	SR	TR
NC BC								
seconds	-	0.96	701.79	6,649.85	2,793.31	3,380.67	NPE	0.17
# results	-	1	20	20	20	20	-	0
AC BC								
seconds	-	0.00	0.06	16.94	0.49	0.74	NPE	0.00
# results	-	1	1	20	20	20	-	0
NC AC								
seconds	-	0.00	0.00	0.00	0.00	0.70	0.00	0.28
# results	-	20	20	20	20	20	20	20
AC AC								
seconds	-	0.00	0.06	17.12	0.00	0.92	6.97	0.00
# results	-	1	1	20	20	20	20	0

NC stands for named concept and AC for anonymous concept. BC stands for before classification and the second AC for after classification. NPE stands for NullPointerException.

20 subclasses in all settings. CEL returns the 20 subclasses only after classification and when the concept name is used. The other queries return owl:Nothing as the only result. FaCT++ returns the 20 subclasses when being queried for the concept name, while the query that contains the anonymous class definition returns the name of this concept.

Minimum Heap Space for Java Reasoners

Table 25: Minimum Heap Space for Java Reasoners

CB	CEL	FaCT++	HermiT	Pellet	RP	SR	TR
n/a	n/a	n/a	4.5	10	n/a	2.5	4

Table 25 shows the minimum required amount of heap space for Java reasoners with SNOMED CT as input ontology. Memory exhaustion is a known problem in tableau-based reasoners when processing large ontologies, and our experiments confirm this. The minimum heap space is just an indicator and might vary for other systems. It needs to be pointed out that our experiments have been performed on a Java HotSpot(TM) 64-Bit Server VM. It is known that the 64-bit mode consumes around 30% more memory as the 32-bit mode, and it makes sense to use the 64-bit mode mostly if 4GB heap space is not sufficient in the 32-bit mode. Running a 32-bit JVM is not supported on the system we performed our tests on.

Soundness and Completeness in Practice

In this paragraph, the computed concept hierarchies of the reasoners are analyzed to test whether the theoretical correctness of the tested reasoners can be confirmed in practice. For all included ontologies, we compared the output of each reasoner to the outputs of all other reasoners, with the rationale that if completely different reasoners generate the same output, the output is probably sound and complete. This does not exclude a scenario in which all reasoners output the same unsound statements and / or collectively do not produce inferences that should be produced. If the outputs differ from each other, we can infer that not all reasoners generate correct output. There is no standard specification on how to output the computed concept hierarchy, and thus we do not analyze the outputs line by line.

To summarize the insignificant differences that we found: In comparison to other reasoners, CB outputs less statements by omitting that top-level concepts are SubClassOf owl:Thing. CEL outputs additionally that every concept and every property is equivalent to itself. Also RacerPro generates additional axioms by stating for all leaf concepts (i.e. concepts that do not have subclasses) that they are superclasses of owl:Nothing. Apart from these differences and for SNOMED CT as input ontology, we found substantial differences: Pellet generates 386 inferences less than the other reasoners and 546 inferences that no other reasoner generates, while Snorocket misses 86 inferences that occur in the other outputs and generates 34 triples that no other reasoner generates. The missing and also the additional statements of Pellet and Snorocket have an empty intersection.

In the following, we will exploit the fact that SNOMED CT is released both in stated and in inferred form. The inferred form is the Relationships Table contained in the official SNOMED CT distribution and can be employed as gold standard to analyze computed concept hierarchies. The Stated Relationships Table differs from the Relationships Table in that it only contains those relationships that are directly asserted by authors or editors. When the generated OWL ontology is classified with a reasoner, the output should correspond to the Relationships Table.

As a first step, we successfully checked the accordance of the Stated Relationships Table and the generated OWL file. Then, to analyze the accordance of the computed concept hierarchies and the Relationships Table, we compared the computed concept hierarchies to the Relationships Table, and the Relationships Table to the computed concept hierarchies. First, we compared all subclass rows from the Relationships Table that only include active concepts to the outputs of the tested reasoners. All outputs are missing 50 concept model attribute statements, which is caused by the way in which the OWL file is generated. All outputs except the one of CEL are

missing 11 SubObjectPropertyOf statements. However, those statements are already present in the Stated Relationships Table and thus not really inferences. Pellet did not infer 386 SubClassOf relationships present in the Relationships Table and the Snorocket Protégé plugin did not infer 86 statements that are present in the Relationships Table. Table 26 summarizes the results. Both concept model attributes and SubObjectPropertyOf assertions are not counted in the table.

Table 26: Missing / Additional inferred SubClassOf statements in regard to the Relationships Table

	CB	CEL	FaCT++	HermiT	Pellet	RP	SR	TR
Missing	0	0	0	0	386	0	86	0
Additional	0	0	0	0	546	0	34	0

Finally, we checked for every inferred SubClassOf axiom of each of the outputs whether it exists in the Relationships Table to identify additional inferred statements. Ignoring tautological axioms such as the root node being SubClassOf owl:Thing and that owl:Nothing is SubClassOf all the leaf nodes, Pellet outputs 546 additional statements and Snorocket 34. Examples of missing and additional inferences are given in Figure 19. With regard to the SNOMED CT Relationships Table, the outputs of all other included reasoners neither missed inferences, nor did they contain additional inferences.

Pellet:

Drug-induced immunodeficiency $\sqsubseteq$ *Drug-related disorder (missing)*
Biological substance poisoning $\sqsubseteq$ *Drug-related disorder (additional)*

Snorocket:

Amiloride + hydrochlorothiazide 2.5mg/25mg tablet $\sqsubseteq$ *Oral dosage form product (missing)*
Betaxolol hydrochloride 20mg tablet $\sqsubseteq$ *Oral dosage form product (additional)*

Figure 19: Examples of missing and additional inferences with regard to the SNOMED CT Relationships Table

As a result of this study we found that the Snorocket Protégé plugin did give correct results when run with Java 1.5 but produced some incorrect

results when run with Java 1.6. The root cause of this problem has been fixed in the subsequent release (1.3.3). Furthermore, a new version (0.4.1) of CEL was released, which does not output the singletons for equivalent properties and concepts. The fix of Pellet's issue of missing and additional inferences will be part of its next release (i.e., 2.2.3). This shows that the analysis of computed concept hierarchies is valuable both to developers and to users of OWL reasoners. Another conclusion is that for SNOMED CT as input ontology, the comparison of the inferred concept hierarchies with each other delivered the same results as the comparison of the inferred concept hierarchies with the Relationships Table.

8.5.4 *Tradeoff between Expressivity and Classification Performance*

Figure 20 shows the classification performance for SNOMED CT, with the reasoners ordered by increasing expressivity (as displayed on the x2-axis). Pellet and TrOWL REL are in the EL section because both of them support OWL 2 EL with implementations that are based on [24]. Reasoners within the same expressivity category are ordered alphabetically.

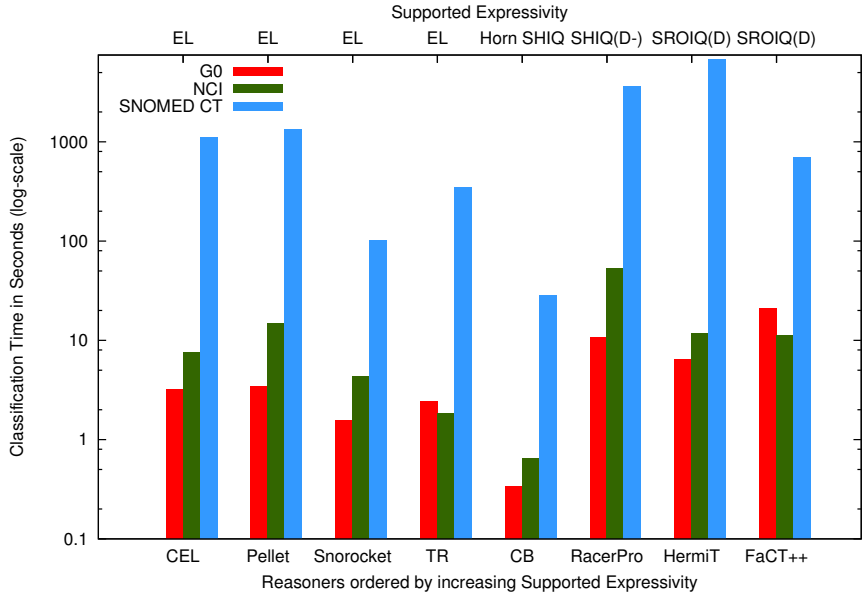

Figure 20: Classification Performance vs. Supported Expressivity

The time needed to classify SNOMED CT does not steadily rise with increasing expressivity. The very expressive reasoner FaCT++ is faster than

CEL and Pellet. CB is the fastest reasoner even though it supports a more expressive language than OWL 2 EL.

8.6 CONCLUSION, DISCUSSION AND FUTURE WORK

The main contribution of this paper is the definition of characteristics that are relevant to evaluate OWL reasoning engines in order to choose the most suitable reasoner for a given application, and the characterization of eight reasoning engines based on these properties. We showed that reasoners vary significantly with regard to all included characteristics. Therefore, a critical assessment and evaluation of core requirements is needed before selecting a reasoner for a real-life application. For example, let us consider a scenario in which a user chooses a reasoner for SNOMED CT. A crucial consideration is whether reasoning services such as incremental classification, rule support, justifications and ABox reasoning are required. Regarding practical usability, it needs to be decided which interfaces (OWL API, OWLlink, Jena) are needed, and whether the source of the reasoner should be open and come with a corresponding license. Also the platform on which the reasoner will run might influence the decision. With respect to the reasoner's performance, one of the aspects is how long the user is willing to wait for classification results, whether she wishes to query for anonymous concepts and concept satisfiability before classification.

Reasoning is an active field of research, and recent developments show that not only (the underlying methods of) established reasoners are being pushed further, but also new reasoners enter the field. Thus, this paper can only display a current snapshot of those rapid developments. Only dedicated OWL 2 EL and tableau-based reasoners have been taken into account for this comparison. Datalog engines, rule engines or reasoners that are based on theorem provers have not been included. The scope of this study is limited to ontologies in $\mathcal{EL}^+$. Performance results might be different for other ontologies, and when a reasoner is needed for a more expressive language, OWL 2 EL reasoners are not applicable. The selection of characteristics is not complete. Support for non-standard ontology features, such as description graphs, has not been included. Also, loading times or the different input and output formats that the reasoners can parse and write have not been evaluated. CB, for example, relies on the OWL functional syntax, while reasoners that integrate the OWL API are very flexible regarding serialization formats. Missing usability characteristics include support and documentation. Commercial reasoners generally offer more support, including support contracts. Also the level of documentation varies considerably for different reasoners. Regarding our experiments, it would have been fairer to measure subsumption checking performance for more than only one concept, as different reasoners

might be optimized for different structures. Also, different reasoners are optimized for different settings of retrievals of direct / indirect subclasses / superclasses, so that all those scenarios should be included in a balanced comparison. The time it takes to retrieve inconsistent / unsatisfiable concepts would be another interesting experiment. Our experiments heavily rely on the OWL API, which contains functions that do not have a tightly specified functionality, and this might be the source of some of the variations of our obtained results. Furthermore, we did not include the supported OWL API version in the dimension practical usability. Future work includes measuring incremental reasoning performance, reasoning with more expressive ontologies, such as GALEN, and benchmarks that involve ABox reasoning as well as inconsistent ontologies and unsatisfiable concepts.

A positive outcome is that all eight tested reasoners succeed in classifying the very large ontology SNOMED CT. The advantage of this ontology is that it is released both in stated and in inferred form, so that the concept hierarchies computed by the reasoners can not only be compared to each other but also to the inferred form, which can be employed as a gold standard to evaluate correctness. By comparing the outputs of the reasoners with each other and with the SNOMED CT Relationships Table, we found classification errors for Pellet and Snorocket that will be / have been fixed. Ongoing testing is necessary to evaluate the correctness of reasoners in practice. Our described characteristics can be applied to any reasoner and form a basis to evaluate reasoners not only on classification performance, but also on other aspects which can be relevant. We will present this study and future results on http://reasonerben.ch.

8.7 ACKNOWLEDGEMENTS

We like to thank the developers of all reasoners, especially (in alphabetical order of the corresponding reasoners) Yevgeny Kazakov, Julian Mendez and Boontawee Suntisrivaraporn, Dmitry Tsarkov, Boris Motik, Kendall Clark, Evren Sirin, Ralf Möller, Michael Wessel and Kay Hidde, Michael Lawley and Jeff Pan, for their active support, insights and interesting discussions. We also like to thank the reviewers for their constructive feedback.

9 REDUNDANT ELEMENTS IN SNOMED CT CONCEPT DEFINITIONS

While redundant elements in SNOMED CT concept definitions are harmless from a logical point of view, they unnecessarily make concept definitions of typically large ontologies such as SNOMED CT hard to construct and to maintain. In this paper, we apply a fully automated method to detect intra-axiom redundancies in SNOMED CT. We systematically analyse the completeness and soundness of the results of our method by examining the identified redundant elements. In absence of a gold standard, we check whether our method identifies concepts that are likely to contain redundant elements because they become equivalent to their stated subsumer when they are replaced by a fully defined concept with the same definition. To evaluate soundness, we remove all identified redundancies, and test whether the logical closure is preserved by comparing the concept hierarchy to the one of the official SNOMED CT distribution. We found that 35,010 of the 296,433 SNOMED CT concepts (12%) contain redundant elements in their definitions, and that the results of our method are sound and complete with respect to our partial evaluation. We recommend to free the stated form from these redundancies. In future, knowledge modellers should be supported by being pointed to newly introduced redundancies.

9.1 INTRODUCTION

SNOMED Clinical Terms (SNOMED CT) allows for meaning-based recording and retrieval of clinical information, which thereby becomes (re)usable. One of the advantages of SNOMED CT is its large size and coverage, which on the other hand makes defining new and maintaining existing concepts a challenging task.

Various (automated) auditing methods have been developed that can be applied to the content of controlled biomedical terminologies, amongst others to ensure the quality factor non-redundancy [140]. While such

methods mostly aim at detecting equivalent concepts, also parts or elements of concept definitions, i.e. intra-axiom redundancies, are problematic. The detection of intra-axiom redundancies is required during design time. In fact, Spackman et al. reported back in 2001 that "...during the concept definition process there has been confusion among modelers about which roles need to be explicitly modeled and which ones can be left unstated. Some of this confusion arises because of uncertainty about which roles and values are inherited from supertypes" [141]. And even though redundancies are harmless from a logical point of view, they impede the maintainability of a terminology [142], [143], as they misleadingly suggest that new information has been added to a concept, while in reality, this "new" information is more general than or equivalent to information that already has been stated in the definition of the same concept or a superconcept. In this paper, we make an inventory of redundant elements in SNOMED CT concept definitions.

9.2 BACKGROUND

9.2.1 *SNOMED CT concept definitions and rolegroups*

SNOMED CT is based on the lightweight Description Logic $\mathcal{EL}^+$ [24]. Its concepts are defined by conjunctions of other concepts as well as role-value pairs which are represented as exists restrictions ($\exists$), and can be either ungrouped or grouped in so-called *rolegroups* [144]. In SNOMED CT, rolegroups allow to nest or rather group existential restrictions within an existential restriction on a role named rolegroup. Concepts can be either *primitive*, i.e. specified by *necessary* conditions only (denoted by the subsumption operator $\sqsubseteq$) or *fully defined*, i.e. specified by both *necessary and sufficient* conditions (denoted by the equivalence operator $\equiv$). Example 2 presents a fully defined sample concept, which is defined by the conjunction of one concept and two rolegroups.

Example 2 [Brain stem contusion with open intracranial wound. RG stands for rolegroup]

```
Brain stem contusion with open intracranial wound ≡
    Contusion of brain with open intracranial wound ⊓
    ∃RG(∃Associated morphology.Open wound ⊓
        ∃Finding site.Intracranial structure) ⊓
    ∃RG(∃Associated morphology.Open contusion ⊓
        ∃Finding site.Brainstem structure)
```

9.2.2 *Trivial and non-trivial primitive concepts*

For our evaluation, we distinguish *trivial primitive concepts*, that are primitive and subsumed by one concept only, and *non-trivial primitive concepts*, that are primitive and described by the conjunction of several concepts and optional additional exists restrictions. With regard to Example 3, we refer to the concept *Brain tissue structure* as trivial primitive, and to *Structure of lobe of brain* as non-trivial primitive.

Example 3 [Structure of lobe of brain]

```
Brain tissue structure ⊑ Brain part

Structure of lobe of brain ⊑
    Brain part ⊓ Brain tissue structure
```

9.2.3 *Redundant elements in SNOMED CT concept definitions*

An element that is part of a concept definition, i.e. a concept or an existential restriction, is redundant if it has been stated explicitly even though it is already implied by the definition of the same concept or a stated superconcept. Therefore, we define an element to be redundant if it is more general than or equivalent to an element that is contained in the definition of the same concept or a stated superconcept. Redundant elements can be eliminated without affecting the ontology's logical closure. For example, the concept *Brain part* in the definition of the concept *Structure of lobe of brain* in Example 3 is redundant as it subsumes the concept *Brain tissue structure*.

9.3 MATERIALS AND METHODS

We employed the July 2012 version of SNOMED CT in Release Format 2, which was transformed to OWL with the Perl script released in the same version. The script makes use of the released concept and stated relationships tables. The latter represents the faithful representation of the information entered by modellers.

We relied on the high-performance reasoner ELK [145] to classify SNOMED CT, and to check for subsumption and equivalence relationships between concepts and roles, while Pellet [108] was used in our evaluation to explain equivalence relationships that were hard to reproduce manually. We relied on the OWL API [28] to carry out all experiments.

9.3.1 *Method to detect redundant elements in SNOMED CT concept definitions*

We exploit the simple structure of SNOMED CT and its rolegroups to detect intra-axiom redundancies. Therefore, we adapted and extended the rules 1 to 3 of redundancy elimination for concept definitions that contain rolegroups as defined by Spackman et al. [26] (and adopted their original numbering). The rules are based on Definition 1.

Definition 1. *More general or equivalent exists restriction*. An exists restriction is more general than or equivalent to another exists restriction whenever both its role and its value concept subsume or are equivalent to the respective elements in the other exists restriction.

$$\exists R.C \sqsupseteq \exists S.D \iff (R \sqsupseteq S) \text{ and } (C \sqsupseteq D)$$

All concept definitions are merely conjunctions of ungrouped or grouped exists restrictions and superconcepts. Therefore, the rules define for each of these elements whether they are redundant:

1. An ungrouped exists restriction is redundant when it is more general than or equivalent to an ungrouped exists restriction within the definition of *the same concept or a superconcept*.

   $$(\exists R.C \sqcap \exists S.D \sqcap \exists T.E) \equiv (\exists S.D \sqcap \exists T.E) \iff \exists R.C \sqsupseteq \exists S.D$$

2. A rolegroup is redundant when all its exists restrictions are more general than or equivalent to those contained in another rolegroup in the definition of *the same concept or a superconcept*.

   $$(RG(\exists R_1.C_1 \sqcap \ldots \sqcap \exists R_n.C_n) \sqcap RG(\exists S_1.D_1 \sqcap \ldots \sqcap \exists S_m.D_m)) \equiv$$
   $$RG(\exists S_1.D_1 \sqcap \ldots \sqcap \exists S_m.D_m)$$
   $$\iff \forall i=1,\ldots,n\ \exists j=1,\ldots,m \mid \exists R_i.C_i \sqsupseteq \exists S_j.D_j$$

3. An exists restriction is redundant within a rolegroup when it is more general than or equivalent to another exists restriction in *the same rolegroup*.

   $$RG(\exists R.C \sqcap \exists S.D \sqcap \exists T.E) \equiv RG(\exists S.D \sqcap \exists T.E) \iff \exists R.C \sqsupseteq \exists S.D$$

4. A concept is redundant when it is more general than or equivalent to one of the other concepts in the definition of *the same concept or a superconcept*.

   $$(C \sqcap D) \equiv D \iff C \sqsupseteq D$$

Rule 3 is an exception with regard to our redundancy definition, as it does not concern an element of a concept definition, but an element within an element. To test whether a concept is defined redundantly, these four rules are applied to a concept and all its stated superconcepts. As the rules are independent from each other, their execution order should not influence the obtained results.

9.3.2 *Evaluation of our method*

To evaluate the results obtained by the application of the four rules of redundancy detection, we assess the completeness and soundness of its output. In absence of a gold standard, we measure completeness by matching our findings to definitions that are likely to be redundant according to Cornet's and Abu-Hanna's method [146], and soundness by checking whether the logical closure is preserved after classifying the manipulated version of the ontology.

9.3.2.1 *Completeness: Comparison of identified redundant concepts to redundant concepts according to Cornet's and Abu-Hanna's method.*

Cornet's and Abu-Hanna's method [146] detects concepts with equivalent definitions in terminological systems represented in a description logic, to addresses the problems of redundancy and underspecification. Concepts that become equivalent to any superconcept when applying this method are likely to be defined redundantly [147]. Let us regard Example 4, which presents a sample group of equivalent concepts that can be detected by applying this method.

Example 4 [Group of concepts with equivalent concept definitions]

```
Finding of volume of heart sounds ⊑
    Finding of heart sounds ⊓
        ∃RG(∃Interprets.Loudness of heart sounds)

Heart sounds diminished ⊑
    Finding of volume of heart sounds ⊓
        ∃RG(∃Finding site.Heart structure)

Heart sound volume variable ⊑
    Finding of volume of heart sounds ⊓
        ∃RG(∃Finding site.Heart structure)

Heart sound inaudible ⊑
    Finding of volume of heart sounds ⊓
        ∃RG(∃Finding site.Heart structure)
```

Here, we can make two interesting observations. First, we see three concepts with definitions that obviously become equivalent when making these concepts fully defined. Second, the three concepts become equivalent to their superconcept *Finding of volume of heart sounds*, and thus, they are likely to be defined redundantly. And indeed, four steps up the concept hierarchy, we encounter their common superconcept presented in Example 5, which already contains a rolegroup that defines the *Finding site* to be the *Heart structure*.

Example 5 [Explanation for redundancy]

```
Cardiac finding ⊑
    Cardiovascular finding ⊓
        ∃RG(∃Finding site.Heart structure)
```

We evaluate the results obtained by the application of the four rules of redundancy detection by checking whether the concepts that are likely to be redundant according to Cornet and Abu-Hanna are indeed contained in the identified set of redundant concepts. In order to detect redundant definitions, we apply the approach proposed by Cornet and Abu-Hanna as follows:

1. Replace each non-trivial primitive concept by a fully defined concept with the same definition.
2. Classify the ontology.
3. For each concept in the ontology, retrieve equivalent concepts from reasoner.
4. Identify concepts that have become equivalent to any stated superconcept, as those are likely to be defined redundantly.
5. Identify and exclude indirect redundancies that emerge due to concepts being subsumed by the conjunction of concepts with equivalent definitions such as in Example 6 and wrongly identified redundancies due to the propagation of equivalence such as in Example 7.[1]

Example 6 [Concepts without intra-axiom redundancy: Because *Midwifery personnel* and *Professional midwife* have the same definitions, they become equivalent. And because *Auxiliary midwife* is being subsumed by the two of them, it also becomes equivalent.]

```
Auxiliary midwife ⊑
    Professional midwife ⊓ Midwifery personnel

Professional midwife ⊑
    Medical, dental, veterinary/related worker ⊓
    Health visitor, nurse/midwife

Midwifery personnel ⊑
    Medical, dental, veterinary/related worker ⊓
    Health visitor, nurse/midwife
```

1 Please note that these cases could be prevented by applying the method only on one superconcept - subconcept pair at a time instead of the entire SNOMED CT. We did not apply this method because it is not feasible even with very fast classification times.

Please note that Cornet's and Abu-Hanna's method does not necessarily retrieve all redundant concepts. For example, a concept can refine its stated superconcept and additionally contain redundant elements. Likewise, redundant elements in fully defined concept definitions are not detected by Cornet's and Abu-Hanna's method. Therefore, the evaluation of the results of the four rules of redundancy detection can only be partial.

Example 7 [Example for wrongly identified redundancy. The concepts *Pancreatic function outside reference range* and *Measurement finding outside reference range* would be equivalent if all involved concepts were fully defined.]

```
Pancreatic function outside reference range ⊑
    Measurement finding outside reference range ⊓
    ∃RG(∃Has interpretation.Outside reference range ⊓
    ∃Interprets.Pancreatic function test)

Measurement finding outside reference range ≡
    Measurement finding ⊓
    ∃RG(∃Has interpretation.Outside reference range ⊓
    ∃Interprets.Measurement procedure)

Pancreatic function test ⊑
    Measurement procedure ⊓
    ∃RG(∃Has Method.Measurement - action)

Measurement procedure ≡
    Procedure by method ⊓
    ∃RG(∃Has Method.Measurement - action)
```

9.3.2.2 *Soundness: Preservation of logical closure.*

Deleting redundant parts of concept definitions should not affect the logical closure, and therefore a change in the concept hierarchy would indicate the removal of a non-redundant part of a concept definition. Thus, we delete all identified intra-axiom redundancies and check whether the computed concept hierarchy obtained from classifying the manipulated version is the same as the one obtained from classifying the original version by bi-directional comparison of both versions to the official SNOMED CT distribution.

9.4 RESULTS: REDUNDANT ELEMENTS IN CONCEPT DEFINITIONS

Applying the four rules of redundancy detection, 35,010 of the 296,433 SNOMED CT concepts (12%) were identified to contain redundant elements in their definitions. Table 27 gives an overview of the results, only regarding the first explanation for these redundancies (the rules were applied in the same order as they are presented in this paper). 11,858 of these concepts are fully defined, and 23,152 non-trivial primitive.

Example 8 [Parenteral form thymoxamine]

```
Parenteral form thymoxamine (product) ≡
    Thymoxamine (product) ⊓
        ∃Has active ingredient.Thymoxamine (substance)

Thymoxamine (product) ⊑
    Alpha blocking vasodilator ⊓ Alpha 1 adrenergic blocking agent ⊓
        ∃Has active ingredient.Thymoxamine (substance)
```

Table 27: Detected concepts with redundant elements. The examples in column 'example' refer to the examples disseminated along the paper.

Rule	Concepts	Example and Explanation
1 (ungrouped exists restriction)	7,874	Example 8: The ungrouped exists restriction ∃*Has active ingredient.Thymoxamine (substance)* is redundant, as it is already contained in the superconcept *Thymoxamine (product)*.
2 (rolegroup)	26,599	Example 2: The first rolegroup is redundant, as it is more general than the second one, because *open wound* subsumes *open contusion*, and *Intracranial structure* subsumes *Brainstem structure*.
3 (grouped exists restriction)	6	Example 9: The exists restriction ∃*Associated morphology.Traumatic abnormality* in the first rolegroup is redundant, as *Traumatic abnormality* subsumes *Closed traumatic abnormality*.
4 (concept)	531	Example 3: The concept *Brain part* is redundant as it subsumes the concept *Brain tissue structure*.

Example 9 [Closed skull fracture with intracranial injury]

```
Closed skull fracture with intracranial injury ≡
    Fracture of skull ⊓
        ∃RG(∃Finding site.Intracranial structure ⊓
        ∃Associated morphology.Traumatic abnormality ⊓
        ∃Associated morphology.Closed traumatic abnormality) ⊓
        ∃RG(∃Associated morphology.Fracture, closed ⊓
        ∃Finding site.Bone structure of cranium)

Explanation:
Closed traumatic abnormality ⊑ Traumatic abnormality
```

Figure 21 shows the SNOMED CT categories that the concepts with redundant elements belong to. Figure 22 depicts the distances between redundant concepts and the concepts containing the explanation for the redundancy. A distance of 0 is interesting as it makes a concept redundant with regard to its own definition. But also long distances are interesting: an element is introduced, not repeated for some concepts down the hierarchy, but then it is. The concept *Measurement of Human T-lymphotropic virus 1 recombinant glycoprotein 21 antibody and Human T-lymphotropic virus 2 recombinant glycoprotein 21 antibody* is among the concepts with the longest distance to its explanation (9 steps).

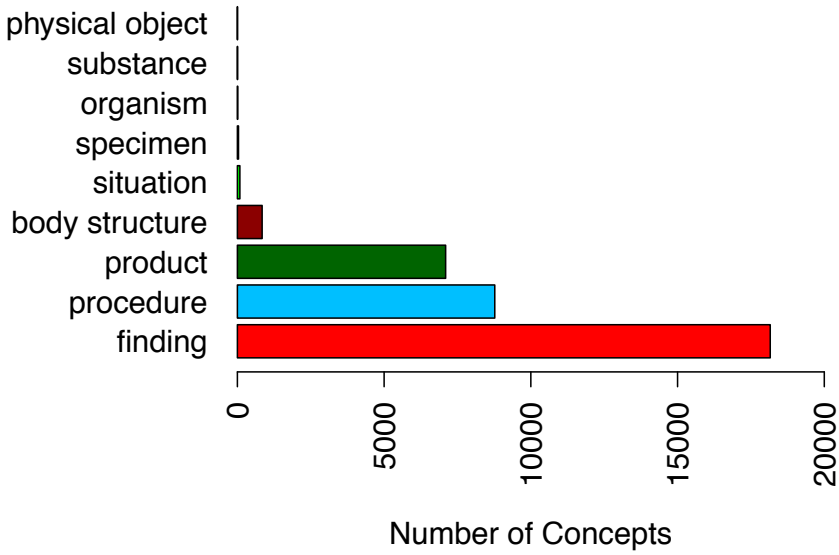

Figure 21: SNOMED CT categories of concepts with redundancies

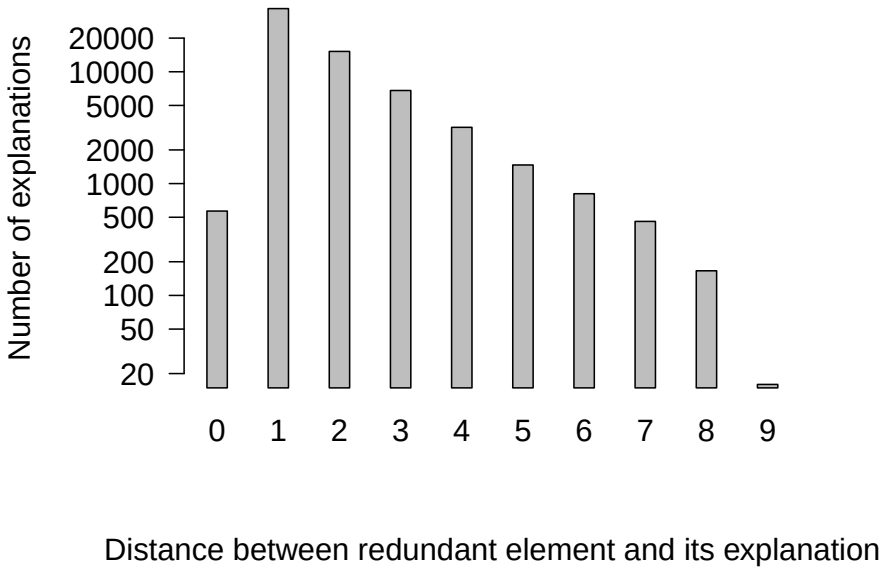

Figure 22: Distances between redundant concepts and the concepts containing the explanation

An exhaustive search for all redundant elements and all explanations results in 65,336 explanations: 13,808 for rule 1, 50,680 for rule 2, 6 for rule 3 and 842 for rule 4. The maximum number of explanations is 16 for

the concept *Late congenital syphilitic meningitis*. The concept with the most (6) redundant elements is *Diphtheria + tetanus + pertussis + poliomyelitis + recombinant hepatitis B virus + recombinant haemophilus influenzae type B vaccine*.

9.5 EVALUATION

9.5.1 *Completeness*

Applying Cornet's and Abu-Hanna's method, 45,975 concept definitions with at least one other concept with a logically equivalent definition have been identified, containing a total of 12,823 non-trivial primitive concepts with definitions that are equivalent to the definition of at least one of their stated superconcepts.

12,094 of these redundancies have been confirmed to be redundant by our method to detect intra-axiom redundancies. 698 out of the 729 non-confirmed redundancies were subsumed by the conjunction of concepts, such as the concept in Example 6. For the remaining 31 non-confirmed redundancies, we successfully generated explanations with Pellet based on the manipulated version of SNOMED CT. A manual revision confirmed that all of the explanations contained further axioms that have been re-defined from being primitive to fully defined, such as the explanation given in Example 7. Therefore, the results of our method are complete with regard to Cornet's and Abu-Hanna's method.

9.5.2 *Soundness*

We generated the logical closure of both the original and the manipulated OWL versions of SNOMED CT, and compared the computed class hierarchies to the one contained in the official distribution. The OWL versions and the database table contained exactly the same set of 438,554 subclass axioms or respectively "is-a" relations.

9.6 RELATED AND FUTURE WORK

In the past, most proposed methods focused at the detection of truly redundant, i.e. equivalent, concepts. Cimino has developed a method to identify multiple synonymous concepts and applied it to the 2001 UMLS Metathesaurus [148]. Grimm and Wissmann [143] provide methods to

compute irredundant ontologies, and Entendre [149] makes users aware of redundancies.

The IHTSDO[2] describes methods to convert concepts into normal forms, some of which imply the elimination of redundancies, and Peng et al. [150] have proposed a method to identify redundant classifications, i.e. unnecessary, simultaneous assignments to sub- and superconcepts. The Ecco tool [151] facilitates the analysis of ontology differences by applying methods to syntactically or semantically detect effectual changes as well as ineffectual changes such as adding or deleting intra-axiom redundancies.

An interesting direction of future work would be to generalise our method. In principle, our definition of a redundant element could be operationalised directly by checking whether an element is more general than or equivalent to an element that is contained in the definition of the same concept or a stated superconcept.

9.7 DISCUSSION AND CONCLUSIONS

Our results show that 35,010 of all 296,433 SNOMED CT concepts (12%) are defined redundantly. These redundancies unnecessarily impede the work of knowledge modellers, and our own experience confirms that manual search for the causes of redundancies can be a tedious task. Therefore, we suggest to remove them from the stated relationships. To reach this goal, the four rules of redundancy detection would have to be applied to the entire SNOMED CT once.[3] Further redundancies should be avoided by pointing knowledge modellers to newly introduced redundancies in the definitions of the concepts they are currently working on, and explaining why these elements are redundant. As shown by Figure 22, most redundant elements are so due to nearby superconcepts, so that the explanations will most probably be intuitive. For this task, the four rules of redundancy detection could be applied as a background process of terminology editing tools to the concepts that are currently being edited. In order to support these goals, we make both our tools and our results freely available[4].

2 http://www.ihtsdo.org/

3 It should be noted that applying the four rules of redundancy detection to the entire SNOMED CT is computationally expensive (ca. 6 hours on a laptop equipped with a 2.8 GHz Intel Core 2 Duo processor and 8 GB of physical memory). However, analysing only one concept is sufficiently fast to be executed as a background process.

4 https://github.com/kathrinrin/redundancies

10 CONCLUSIONS

"If you can not measure it, you can not improve it."
Lord Kelvin

This chapter summarises and discusses our main results and answers to the research questions, as well as directions for future work.

10.1 RESULTS AND ANSWERS TO RESEARCH QUESTIONS

The main goal of this thesis was to investigate *under which conditions healthcare quality indicators can be computed automatically by reusing data already collected during the clinical care process.* As the scope of our main research question is very broad, we operationalised it, resulting in several sub-questions formulated in the introduction (Chapter 1) of this thesis. In the following, we present our results and contributions in relation to these questions. The answer to the main research question is discussed subsequently.

1. How can quality indicators be formalised?

As no method existed to formalise healthcare quality indicators for their automated computation, we developed CLIF, inspired by LERM, the Logical Elements Rule Method [35], a method to assess and formalise clinical rules for decision support, as well as a method to transform natural language into formal proof goals proposed by Stegers et al. [34]. CLIF comprises 9 steps:

Step 1) Encode relevant concepts from the indicator by concepts from a terminology
Step 2) Define the information model
Step 3) Formalise temporal constraints
Step 4) Formalise numeric constraints
Step 5) Formalise textual constraints
Step 6) Formalise Boolean constraints
Step 7) Group constraints by Boolean connectors
Step 8) Identify exclusion criteria / negations
Step 9) Identify constraints that only aim at the numerator

Regarding the order of the steps, step 1) and 2) should be carried out first, because they formalise the building blocks that are used in subsequent steps. Steps 7) to 9) should be carried out last, because they build on previously defined constraints. Steps 3) to 6) can be performed in the preferred order of the user.

2. How reproducible are the results of our formalisation method, and which steps are particularly challenging?

In our case study presented in Chapter 3, we found that most steps led to reproducible results, but we also identified several problems that cause variability: Step 1), the encoding of relevant concepts from the indicator by concepts from a terminology, was challenging due to the size of the applied terminology SNOMED CT, and because medical expertise is required to encode certain concepts. Regarding step 2), the definition of the information model, our case study showed that all test persons had difficulties in applying the problem-oriented patient record, i.e. to relate procedures to their underlying diagnoses. The variability in step 3), the formalisation of temporal constraints, was due to ambiguities in the indicator text regarding temporal relations. For example, it was not clear what "previous radiotherapy" referred to.

As none of these issues is intrinsic to our method, we conclude that CLIF itself can lead to maximally reproducible results. To increase reproducibility, indicators have to be formulated as unambiguously and precisely as possible, also with regard to temporal relations (which, as results presented in Chapter 4 suggest, cover a large percentage of constraints). Ideally, quality indicators should be released in a formalised format, based on standard information models and terminologies, which is obtainable by applying our method. The application of CLIF requires (the cooperation of) trained experts with medical as well as medical informatics expertise to resolve remaining ambiguities, to encode clinical concepts and to specify the relations between concepts.

3. How generalisable is the resulting method?

In Chapter 4, we formalised the entire heterogeneous set of 159 Dutch quality indicators for general practices. This study showed that our web tool supported the full formalisation of 84% of the quality indicators into SQL queries. The remaining indicators could only be formalised partially due to missing functionality of the web tool. For these cases, we applied CLIF manually by directly translating the respective criteria into SQL, enabling us to fully formalise 100% of the indicators. To evaluate the formalised indicators, we computed them based on a large database of real patient data and obtained results that were comparable to results computed independently by two other parties. We conclude that even though our web tool requires refinement, the method itself covers all quality indicators. Therefore, the method is sufficiently generalisable to formalise the entire set of Dutch quality indicators for general practices.

While the study described in Chapter 4 focused exclusively on the generalisability of CLIF, we formalised and computed a relatively small set of colorectal cancer indicators for hospitals in several other chapters of this thesis, with an emphasis on detail and accuracy. This work further supports the generalisability of our method.

4. What are the barriers that impede the secondary use of patient data, and how can they be prevented?

In our study in Chapter 5, we identified a number of barriers that hinder the (timely) reuse of routinely collected clinical data. Even though all data that we required was in principle available in a digital format, and most of it within our hospital, it took a long time until we received a version of the requested data, and the data itself was of insufficient quality. The barriers that we identified covered all four of Galster's categories of why clinical information is not reused [23]. However, most problems were due to underlying organisational / cultural and data quality reasons.

We recommend the following measures to surmount the encountered barriers and to facilitate the reuse of patient data:

- Ensure availability of data and accessibility of data sources.
- Ensure patients' interests, privacy and security while allowing for reuse.
- Set-up a reuse-friendly organisation and culture.
- Increase data quality in terms of completeness both on database and data element level, as well as correctness, interlinking of diagnoses and procedures, provenance and "meaning of data".
- Allow for cross-database querying if a hospital's IT infrastructure consists of several dedicated source systems.

5. How does data quality influence the reliability of quality indicator results?

In Chapter 6, we examined the quality of manually collected data for the DSCA, which is a national quality register, and our EMR, and computed a set of indicators based on them. Our data quality analysis showed that all required data items were available in a structured format in the DSCA dataset, and their average completeness was 86%. The average completeness of these items in our EMR was 50%, and their average correctness 87%. All 10 indicators were fully computable based on the DSCA dataset,

but only 3 based on EMR data, two of which were percentages. Both percentages significantly underestimated the quality of care compared to the same indicators computed based on the DSCA dataset.

Reasons were unavailable, incomplete and incorrect data items as well as missing relationships between diagnoses and procedures in the EMR. If reliable indicator results are required, EMRs should be re-designed so that a core dataset consisting of variables requested for indicator computation is entered directly and timely in a structured, problem-oriented, sufficiently detailed and standardised format. Furthermore, awareness regarding the (re)use of data should be risen, and local data quality improvement strategies should be applied to ensure that required data is complete and correct.

6. Can open*EHR archetypes facilitate the semantic integration of quality indicators and routine patient data to automatically compute indicators?*

In our case study in Chapter 7, we successfully mapped both our local database schema and elements of patient data occurring in indicators to (elements of) publicly available archetypes. The coverage of the public repository was high, and editing an archetype to fit our requirements was straightforward. Based on our mappings, we computed a set of three indicators from the domain of gastrointestinal cancer surgery, and the results were comparable to centrally computed and publicly reported results. We conclude that *open*EHR archetypes can facilitate the semantic integration of patient data and quality indicators.

7. What are characterising properties of reasoners for OWL 2 EL, and how does a selection of reasoners perform with respect to these properties?

In Chapter 8, we identified several characterising properties and organised them along three dimensions: underlying reasoning characteristics, practical usability and empirically measured performance. We then categorised eight reasoners along the defined characteristics and benchmarked them against the well-known biomedical ontologies GO, NCI and SNOMED CT. The main conclusion from this study was that the included reasoners varied substantially with regard to all included characteristics. For example, our results showed that non-standard reasoning tasks were not widely supported, that classification times varied by up to two orders of magnitude and that they did not consistently increase with increasing supported expressivity of the reasoners. Also, soundness and completeness in theory did not necessarily imply soundness and completeness in practice. Hence, a critical assessment and evaluation of core requirements is needed before selecting a reasoner for a real-life application.

8. How can redundant elements in concept definitions of SNOMED CT be identified? How many redundant elements are identifiable using the resulting method?

In Chapter 9, we identified redundant elements in concept definitions of SNOMED CT. For this purpose, we adapted and extended Spackman's rules of redundancy elimination for concept definitions that contain rolegroups [26]. Applying these rules in a fully automated way, we found that 35,010 of the 296,433 SNOMED CT concepts (12%) contained redundant elements in their definitions, and that the results of our method are sound and complete with respect to our partial evaluation. We recommend to free the stated form from these redundancies. Our method can be applied to avoid further redundancies by pointing knowledge modellers to (explanations for) newly introduced redundancies in the definitions of the concepts they are currently working on.

10.2 ANSWER TO MAIN RESEARCH QUESTION

The experiences gained in this thesis lead to insights that enable us to answer our main research question:

> *Under which conditions can healthcare quality indicators be computed automatically by reusing data already collected during the clinical care process?*

First of all, *quality indicators* themselves should be released as precisely as possible, leaving no freedom for different interpretations, so that their computation leads to comparable results across institutions. Indicators need to be *formalised* to be automatically computable. Such a formalisation method should lead to *reproducible results*, and it should be *generalisable*. Our studies confirmed that our proposed formalisation method CLIF can lead to maximally reproducible results and that it is generalisable to a broad set of Dutch quality indicators, but that unambiguous indicators and the cooperation of trained experts are required. *Patient data* needs to be available and of high quality, as data quality can have a significant impact on quality indicator results. High quality implies the use of well-established healthcare standards, so that the meaning of data becomes machine-processable, which is indispensable for automated reuse. Finally, *semantic interoperability* is required to integrate data from various heterogeneous sources, and also to bridge the semantic gap between indicators and patient data. For this purpose, standard information models and large,

lightweight, logics-based terminologies such as SNOMED CT play an important role both for the formalisation and the computation of healthcare quality indicators. *Automated reasoners* provide support for tasks including classification, i.e. the computation of implied subclass / superclass relationships, and querying, which are especially important to compute quality indicators. As the characteristics of reasoners can vary substantially, they should be taken into account when choosing a reasoner for a specific application scenario. Finally, the quality of medical terminologies such as SNOMED CT must be ensured by automated auditing methods, as they are typically too large to be audited manually. A relevant quality factor is *non-reduncancy*. We developed a method to detect intra-axiom redundancies in SNOMED CT and showed that a large percentage of concepts of the current version contained this kind of redundancies. Our method and its evaluation made use of two reasoners, which we selected based on their classification performance, soundness and completeness in theory and practice as well as their ability to provide justifications for inferences.

We conclude that the automated computation of healthcare quality indicators by reusing data already collected during the clinical care process is feasible. However, data that is not recorded to be reused can not be expected to be of sufficient quality, so that the reuse of data for unforeseen use cases might come at the price of limited reliability, as shown in one of our studies. If reliable results are required, the variables needed should be integrated into the design of an EMR, so that high data quality can be ensured. Measures that increase the general quality and reusability of data independently from specific secondary uses are the implementation of standards, and the recording of relationships between variables, such as relationships between diagnoses and procedures in the healthcare domain.

10.2.1 *Generalisability of our Results*

Our results regarding the formalisation and automated computation of *healthcare quality indicators* might be transferable to other clinical (re)uses to some extent. Like quality indicators, reuses such as the real-time application of clinical guidelines in decision support systems, the recruitment of patients for clinical trials and various types of clinical studies, are based on eligibility criteria, comprising in- and exclusion criteria. All use cases have in common that patients who fulfill certain criteria and conditions need to be identified, a process which is also being referred to as *clinical phenotyping*. Also the formalisability of (the criteria for) other reuses depends on unambiguous descriptions. This potential transferability of research results might apply in the other direction too, so that the extensive

results that have been achieved in the domains of clinical guidelines and trials might be partially applicable to quality indicators.

Also the requirements concerning the *secondary use* of routinely recorded patient data, such as (real-time) data availability, data quality and standardisation for semantic interoperability might be generalisable to some extent, and our results and recommendations might be applicable to design, evaluate and optimise already existing or new systems in other hospitals with comparable information infrastructures. The results obtained in the last part of this thesis regarding reasoners and redundancies are relevant to (re)use data encoded in large, lightweight terminologies such as SNOMED CT that are typical for the healthcare domain, and to audit such terminologies.

On a high level, the secondary use of patient data for the automated computation of quality indicators is a specific case for a research question of general importance, namely *how data can be reused for other purposes than those that it was originally recorded for*. This research question lies at the core of the current trend towards analysing *Big Data*. Also here, challenges such as automation, semantic interoperability for integrating heterogeneous data from different sources as well as data quality play essential roles.

10.3 LIMITATIONS

This section discusses a number of limitations of the thesis.

Part I) Formalising and Automatically Computing Healthcare Quality Indicators

In two of our studies (Chapters 2 and 3), we worked with synthetic patient data generated based on arbitrary probabilities, which might not be representative. We recently developed a knowledge-based patient data generator, which produces more realistic data [152].

A limitation of Chapter 3, in which we analyse the reproducibility of CLIF's results, is that we only worked with one quality indicator, which did not require CLIF's steps for *Boolean constraints* and *Boolean connectors*. Besides, we added the step *textual constraints* to the method after analysing its generalisability. It would be interesting to repeat the experiment with an indicator that requires all steps.

An inherent limitation of the evaluation of formalised quality indicators is that due to the indicators' ambiguities, a perfect gold standard for one single "best" formalisation does not exist. Therefore, our way to proceed

was to develop a reference standard together with domain experts, but it is not unlikely that a different set of experts would have developed a different reference standard.

Our results regarding the generalisability of CLIF (Chapter 4) are limited to Dutch indicators. The set of indicators for general practices was representative, as we formalised it entirely. However, as respects indicators for hospitals, we only considered one domain, and we did not assess its representativeness. Even though our formalised indicators might be comparable to indicators in other countries, a thorough analysis of international sets would be desirable.

In Chapters 3 and 4, we studied the reproducibility and generalisability of CLIF, making use of a web-based tool to facilitate its application. A graphical user interface can influence a user's ability to accomplish a task, and the use of the tool presumably influenced our studies. In retrospect, it is impossible to determine the actual influence of the tool on the formalisation process and consequently on the obtained formalisations.

Part II) Secondary Use of Patient Data

Secondary use of patient data is a global challenge, affecting all healthcare institutions, and having many different use cases. The scope of the thesis and of the three chapters regarding the secondary use of patient data is limited to one single use case - quality indicators, and all studies took place in only one hospital. As the relationship between quality indicator computation and other reuses has not been studied systematically yet, we can only make educated guesses about commonalities. The same holds for similarities of the situations in our hospital and in others.

We have shown that *openEHR* archetypes can facilitate the semantic integration of routine patient data from several sources and quality indicators. During the time of writing, new information model standards emerged, such as the Quality Data Model and HQMF, the Health Quality Measures Format, which is a machine-processable standard for representing health quality indicators as electronic documents (eMeasures). In future, it would be interesting to analyse whether these new standards could be integrated into our approach.

Part III) Reasoning and Ontologies for Semantic Interoperability

Regarding our comparison of reasoners for large ontologies in the OWL 2 EL Profile as described in Chapter 8, ELK, a reasoner that can classify SNOMED CT in 5 seconds [145], has been released shortly after our study was completed. It would be interesting to include it in a future comparison.

Finally, in Chapter 9, we have evaluated the soundness and partial completeness of our method empirically. It would have been stronger to prove these properties in theory.

10.4 STRENGTHS

A major strength of this thesis is that we tackled the problem of automated healthcare quality indicator computation from multiple perspectives and in various settings.

To begin with, we employed indicators for several reporting years and domains: colorectal cancer indicators for hospitals as well as a heterogeneous set of indicators for general practices. The indicators covered all types of Donabedian's trilogy: structure, process and outcome.

Regarding our formalisation method, we employed several standard and non-standard terminologies such as SNOMED CT, ICD-9-CM, ICPC, ATC and various national as well as hospital-internal coding systems. Likewise, we worked with several standard and non-standard information models: self-generated custom models, the problem-oriented patient record, information models employed in our hospital, the registry and in general practitioner systems, and last but not least standard archetypes. This broad variety of employed terminologies and information models demonstrates our method's flexibility. We employed the method both manually and with the help of a web-based tool, and it proved to be sufficiently comprehensible to be used by a group of test persons who were previously unacquainted with the topic of indicator formalisation.

Also the data sources employed to compute the indicators varied: we worked with synthetic data as well as real patient data from several sources of our hospital, such as database tables from the clinical systems in routine use, our hospital's data warehouse and administrative data sources from the GIOCA. Moreover, we worked with real patient data from a medical quality registry and from general practices. The data was stored both in OWL ontologies and in SQL databases, and hence we employed both SPARQL and SQL as query languages. Finally, we evaluated formalised indicators and computed indicator results by checking their face validity and consistency, by comparison to a reference standard that we established together with domain experts and by comparison to results computed by others.

10.5 FUTURE WORK

Important research questions for future work include the following:

Part I) Formalising and Automatically Computing Healthcare Quality Indicators

The formalisability of a quality indicator is in itself a quality indicator, and one of our recommendations is to release quality indicators in an already formalised format. For this reason, it would be interesting to investigate whether, in addition to already existing research methods and recommendations [60, 153–156], *CLIF can contribute to support the design of quality indicators*. Related is the research question of *how patient preferences can be included in the design and computation of quality indicators?* For example, informed patients who deliberately opted out of a certain procedure should be excluded, as they otherwise distort the computed results.

Another very interesting research questions is *to what extent CLIF's steps could be (semi-)automated*. Here, we might be able to harvest the rich research results regarding the processing of natural language. For example, concept recognisers such as MetaMap [157] or the NCBO annotator [158] might support the (pre-)selection of concepts that occur in healthcare quality indicators. Also (semi-)automated approaches to transform free-text eligibility criteria into computable criteria such as proposed by Tu et al. [51] and Weng et al. [50] or the pattern-based approach as proposed by Milian et al. [52] might be suitable starting points.

Clinical guidelines, rules for decision support and quality indicators are all interconnected. Guidelines can be operationalised for real-time decision support, and quality indicators often measure whether a guideline has been followed. A promising arising research question is *how process indicators can be operationalised for clinical decision support and immediate feedback, so that better indicator results and better quality of care can be achieved* For example, our lymph node indicator could easily be transformed into the following rule: *If a patient underwent a resection of a primary colonic carcinoma, then 10 or more lymph nodes should be examined*. For this application scenario, a patient record could be scanned for fulfilled denominators, and whether actions regarding the numerator have already been carried out. If this is the case, and if the quality criteria defined in the numerator are fulfilled, the physician might receive positive feedback. Otherwise, he could be reminded, and / or be asked for the reason why the quality criteria have not been fulfilled. Such reasons might provide deep insights into the processes of care, and help to revise the indicators themselves. Reminders might be valuable given the ever growing amount of indicators. However, they should be personalisable to prevent alert fatigue, and they should be connected to the evidence underlying the indicators.

Part 2) Secondary Use of Patient Data

Regarding the secondary use of patient data for healthcare indicator computation, the most severe problem is inappropriate data quality. Hence, it should be investigated *how the collection of high-quality data can be supported, so that more reliable indicator results can be achieved*. Options include to exploit the explicit information needs of formalised indicators, to use archetypes to detect invalid data values and to apply real-time quality improvement strategies such as feedback methods to inform the one entering the data about the entered data's quality, and to raise his awareness of possible (re)uses. To obtain high-quality standardised patient data, those who record the data must be supported in encoding clinical concepts, and in documenting relations between them.

Part III) Reasoning and Ontologies for Semantic Interoperability

Our method to detect intra-axiom redundancies currently only applies to SNOMED CT. An interesting direction of future work would be to investigate *to what extent our method can be generalised*. It might also provide interesting insights to analyse *how redundancies evolved over the course of previous versions of SNOMED CT*.

In several chapters of this thesis, we have worked with OWL to represent SNOMED CT, *open*EHR archetypes and patient data. This is a non-standard approach whose implications should be further assessed. The most apparent advantage is that a logic-based representation language opens the doors for automated reasoning. Regarding patient data, reasoning can make implicit knowledge contained in EMRs explicit, and thereby machine-processable. Automated reasoning might also contribute to tackle the *boundary problem*, which refers to the problem that terminologies and information models can overlap, as they are developed independently from each other. For this reason, it is possible that differently modelled constructs are semantically equivalent, i.e. isosemantic. The question is *how such constructs can be detected to make information models and terminologies interoperable*.

10.6 OUTLOOK

This thesis answers the research question of *under which conditions healthcare quality indicators can be computed automatically by reusing data already collected during the clinical care process*. The presented results are a basis to support clinical practice and further areas of research where quality indicators are used to improve health outcomes of patients.

BIBLIOGRAPHY

[1] PricewaterhouseCoopers. *Transforming healthcare through secondary use of health data*, 2009.

[2] A. Geissbuhler, C. Safran, I. Buchan, R. Bellazzi, S. Labkoff, K. Eilenberg, A. Leese, C. Richardson, J. Mantas, P. Murray, and G. De Moor. *Trustworthy reuse of health data: A transnational perspective*. International Journal of Medical Informatics, 2012.

[3] Richard Lilford, Mohammed A. Mohammed, David Spiegelhalter, and Richard Thomson. *Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma*. Lancet, 363(9415):1147–54, April 2004.

[4] Martin Lawrence and Frede Olesen. *Indicators of Quality in Health Care*. European Journal of General Practice, 3(3):103–108, January 1997.

[5] Avedis Donabedian. *The Quality of Care: How Can It Be Assessed?* JAMA, 260:1743–1748, 1988.

[6] M. Beresford, R. Glynnejones, P. Richman, A. Makris, S. Mawdsley, D. Stott, M. Harrison, M. Osborne, R. Ashford, and J. Grainger. *The Reliability of Lymph-node Staging in Rectal Cancer After Preoperative Chemoradiotherapy*. Clinical Oncology, 17(6):448–455, September 2005.

[7] Scott Caplin, Jean-Philippe Cerottini, Fred T. Bosman, Michael T. Constanda, and Jean-Claude Givel. *For patients with Dukes' B (TNM Stage II) colorectal carcinoma, examination of six or fewer lymph nodes is related to poor prognosis*. Cancer, 83(4):666–72, August 1998.

[8] Fabio Cianchi, Annarita Palomba, Vieri Boddi, Luca Messerini, Filippo Pucciani, Giuliano Perigli, Paolo Bechi, and Camillo Cortesini. *Lymph node recovery from colorectal tumor specimens: recommendation for a minimum number of lymph nodes to be examined*. World Journal of Surgery, 26(3):384–9, March 2002.

[9] Gábor Cserni, Vincent Vinh-Hung, and Tomasz Burzykowski. *Is there a minimum number of lymph nodes that should be histologically assessed for a reliable nodal staging of T3NoMo colorectal carcinomas?* Journal of Surgical Oncology, 81(2):63–9, October 2002.

[10] Pedro Luna-Pérez, Saúl Rodríguez-Ramírez, Isabel Alvarado, Marcos Gutiérrez de la Barrera, and Sonia Labastida. *Prognostic significance of retrieved lymph nodes per specimen in resected rectal adenocarcinoma after preoperative chemoradiation therapy*. Archives of medical research, 34(4):281–6, 2003.

[11] Kazuhiko Yoshimatsu, Keiichiro Ishibashi, Arihiro Umehara, Hajime Yokomizo, Kiyohito Yoshida, Takashi Fujimoto, Kiyo Watanabe, and Kenji Ogawa. *How many lymph nodes should be examined in Dukes' B colorectal cancer? Determination on the basis of cumulative survival rate.* Hepato-gastroenterology, 52(66):1703–1706, 2004.

[12] *Landelijke richtlijn Coloncarcinoom, versie 2.0*.

[13] Helen A. Anema, Sabine N. van der Veer, Job Kievit, Elly Krol-Warmerdam, Claudia Fischer, Ewout Steyerberg, Dave A. Dongelmans, Auke C. Reidinga, Niek S. Klazinga, and Nicolet F. de Keizer. *Influences of definition ambiguity on hospital performance indicator scores: examples from The Netherlands.* European journal of public health, (Dmv):1–6, April 2013.

[14] Jeff Heflin and James Hendler. *Semantic interoperability on the web*. Technical report, University of Maryland, Department of Computer Science, 2000.

[15] Robert H. Brook, E izabeth A. McGlynn, and Paul G. Shekelle. *Defining and measuring quality of care: a perspective from US researchers.* International Journal for Quality in Health Care, 12(4):281–95, August 2000.

[16] A. E. Powell, H .T O. Davies, and R. G. Thomson. *Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls.* Quality & Safety in Health Care, 12:122–8, April 2003.

[17] J.B. Fowles, E.A. Kind, S. Awwad, J.P. Weiner, and K.S. Chan. *Performance measures using electronic health records: five case studies*, 2008.

[18] Carol P. Roth, Yee-Wei Lim, Joshua M. Pevnick, Steven M. Asch, and Elizabeth A. McGlynn. *The challenge of measuring quality of care from the electronic health record.* American Journal of Medical Quality, 24(5):385–394, 2009.

[19] Amy P. Abernethy, James E. Herndon II, Jane L. Wheeler, Krista Rowe, Jennifer Marcello, and Patwardhan Meenal. *Poor Documentation Prevents Adequate Assessment of Quality Metrics in Colorectal Cancer*. Journal of Oncology, 5(4):167–74, July 2009.

[20] Kitty S. Chan, Jinnet B. Fowles, and Jonathan P. Weiner. *Electronic health records and the reliability and validity of quality measures: a review of the literature.* Medical Care Research and Review, 67(5):503–27, October 2010.

[21] E. M. Burns, A. Bottle, P. Aylin, A. Darzi, R. J. Nicholls, and O. Faiz. *Variation in reoperation after colorectal surgery in England as an indicator of surgical performance: retrospective analysis of Hospital Episode Statistics*. BMJ, 343, August 2011.

[22] Thomas Beale. *Archetypes: Constraint-based domain models for future-proof information systems*. In OOPSLA 2002 workshop on behavioural semantics, pages 1–18, 2002.

[23] Gert Galster. *Why is clinical information not reused?* In Studies in Health Technology and Informatics, pages 624–628, 2012.

[24] Franz Baader, Sebastian Brandt, and Carsten Lutz. *Pushing the EL envelope*. IJCAI, 2005.

[25] Franz Baader, Sebastian Brandt, and Carsten Lutz. *Pushing the EL Envelope Further*. In 5th OWL Experiences and Directions Workshop, page 364, October 2008.

[26] Kent A. Spackman, Robert Dionne, Eric Mays, and Jason Weis. *Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED*. AMIA Symposium Proceedings, pages 712–6, January 2002.

[27] Ronald Cornet and Nicolette F. de Keizer. *Forty years of SNOMED: a literature review*. BMC Medical Informatics and Decision Making, 8 Suppl 1:S2, January 2008.

[28] Matthew Horridge and Sean Bechhofer. *The OWL API: A Java API for Working with OWL 2 Ontologies*. 6th OWL Experiences and Directions Workshop, 529:11–21, 2009.

[29] Dmitry Tsarkov and Ian Horrocks. *FaCT++ Description Logic Reasoner: System Description*. In Automated Reasoning, pages 292–297. Springer, 2006.

[30] Yevgeny Kazakov. *Consequence-Driven Reasoning for Horn SHIQ Ontologies*. In IJCAI, pages 2040–2045, 2009.

[31] Kathrin Dentler, Ronald Cornet, Annette ten Teije, and Nicolette F. de Keizer. *Comparison of reasoners for large ontologies in the OWL 2 EL profile*. Semantic Web Journal, 2:71–87, 2011.

[32] Atanas Kiryakov, Damyan Ognyanov, and Dimitar Manov. *OWLIM - A Pragmatic Semantic Repository for OWL*. In Web Information Systems Engineering Workshops, pages 182–192. Springer, 2005.

[33] Jeen Broekstra, Arjohn Kampman, and Frank Van Harmelen. *Sesame: A generic Architecture for Storing and Querying RDF and RDF Schema*. The Semantic Web - ISWC 2002, pages 54–68, 2002.

[34] Ruud Stegers, Annette ten Teije, and Frank Van Harmelen. *From Natural Language to Formal Proof Goal*. Managing Knowledge in a World of Networks, pages 51–58, 2006.

[35] Stephanie Medlock, Dedan Opondo, Saeid Eslami, Marjan Askari, Peter Wierenga, Sophia E de Rooij, and Ameen Abu-Hanna. *LERM (Logical Elements Rule Method): A method for assessing and formalizing clinical rules for decision support*. International Journal of Medical Informatics, 80(4):286–95, April 2011.

[36] Carl A. Williams, Angelia D. Mosley-Williams, and J. Marc Overhage. *Arthritis Quality Indicators for the Veterans Administration: Implications for Electronic Data Collection, Storage Format, Quality Assessment, and Clinical Decision Support*. In AMIA Symposium Proceedings, pages 806–10, January 2007.

[37] Matvey B. Palchuk, Anna A. Bogdanova, Tarang Jatkar, Jinlei Liu, Natalya Karmiy, Dan Housman, and Jonathan S. Einbinder. *Automating Quality Reporting with Health Quality Measures Format "eMeasures" and an Analytics Engine*. AMIA Symposium Proceedings, page 1205, 2010.

[38] Chunhua Weng, Samson W. Tu, Ida Sim, and Rachel Richesson. *Formal representation of eligibility criteria: A literature review*. Journal of Biomedical Informatics, 43(3):451–467, June 2010.

[39] Paolo Besana, Marc Cuggia, Oussama Zekri, Annabel Bourde, and Anita Burgun. *Using Semantic Web technologies for Clinical Trial Recruitment*. In The Semantic Web - ISWC 2010, pages 34–49, 2010.

[40] Chintan Patel, James Cimino, Julian Dolby, Achille Fokoue, Aditya Kalyanpur, Aaron Kershenbaum, Li Ma, Edith Schonberg, and Kavitha Srinivas. *Matching patient records to clinical trials using ontologies*. In The Semantic Web - ISWC 2007, pages 816–829. Springer-Verlag, 2007.

[41] Kathrin Dentler, Annette ten Teije, Ronald Cornet, and Nicolette F. de Keizer. *Towards the Automated Calculation of Clinical Quality Indicators*. Knowledge Representation for Health-Care, LNCS 6924:51–64, 2012.

[42] Lucila Ohno-Machado, John H. Gennari, Shawn N. Murphy, Nilesh L. Jain, Samson W. Tu, Diane E. Oliver, Edward Pattison-Gordon, Robert A. Greenes, Edward H. Shortliffe, and G. Octo Barnet. *The GuideLine Interchange Format - A Model for Representing Guidelines*. Journal of the American Medical Informatics Association, 5(4):357–372, 1998.

[43] Vimla L. Patel, Vanessa G. Allen, Jose F. Arocha, and Edward H. Shortliffe. *Representing Clinical Guidelines in GLIF: Individual and Collaborative Expertise*. Journal of the American Medical Informatics Association, 5(5):467–483, 1998.

[44] Kathrin Dentler, Ronald Cornet, Annette ten Teije, Kristien Tytgat, Jean Klinkenbijl, and Nicolette F. de Keizer. *The Reproducibility of CLIF, a Method for Clinical Quality Indicator Formalisation*. In Studies in Health Technology and Informatics, pages 113–117. IOS Press, 2012.

[45] Kathrin Dentler, Annette ten Teije, Ronald Cornet, and Nicolette F. de Keizer. *Semantic Integration of Patient Data and Quality Indicators based on openEHR Archetypes*. ProHealth 2012/KR4HC 2012, LNAI 7738:85–97, 2013.

[46] David Blumenthal and Marilyn Tavenner. *The Meaningful Use" Regulation for Electronic Health Records*. New England Journal of Medicine, pages 501–504, 2010.

[47] Diederick E. Grobbee, Arno W. Hoes, Theo J. M. Verheij, Augustinus J. P. Schrijvers, Erik J. C. Van. Ameijden, and Mattijs E. Numans. *The Utrecht Health Project: Optimization of routine healthcare data for research*. European Journal of Epidemiology, 20(3):285–290, January 2005.

[48] Jessica Ross, Samson W. Tu, Simona Carini, and Ida Sim. *Analysis of Eligibility Criteria Complexity in Clinical Trials*. In AMIA Summits on Translational Science Proceedings, page 46, 2010.

[49] Mike Conway, Richard L Berg, David Carrell, Joshua C Denny, Abel N Kho, Iftikhar J Kullo, James G Linneman, Jennifer a Pacheco, Peggy Peissig, Luke Rasmussen, Noah Weston, Christopher G Chute, and Jyotishman Pathak. *Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms.* AMIA Symposium, 2011:274–83, January 2011.

[50] Chunhua Weng, Xiaoying Wu, Zhihui Luo, Mary Regina Boland, Dimitri Theodoratos, and Stephen B Johnson. *EliXR: an approach to eligibility criteria extraction and representation.* Journal of the American Medical Informatics Association, 18 Suppl 1:i116–24, December 2011.

[51] Samson W. Tu, Mor Peleg, Simona Carini, Michael Bobak, Jessica Ross, Daniel Rubin, and Ida Sim. *A practical method for transforming free-text eligibility criteria into computable criteria.* Journal of Biomedical Informatics, pages 1–39, September 2010.

[52] Krystyna Milian, Anca Bucur, and Annette Ten Teije. *Formalization of clinical trial eligibility criteria: Evaluation of a pattern-based approach.* 2012 IEEE International Conference on Bioinformatics and Biomedicine, pages 1–4, October 2012.

[53] William K Thompson, Luke V Rasmussen, Jennifer a Pacheco, Peggy L Peissig, Joshua C Denny, Abel N Kho, Aaron Miller, and Jyotishman Pathak. *An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms.* AMIA Symposium, 2012:911–20, January 2012.

[54] V. Stroetman, Dipak Kalra, Pierre Lewalle, Alan Rector, J. Rodrigues, K. Stroetman, Gyorgy Surjan, Bedirhan Ustun, Martti Virtanen, and P. Zanstra. *Semantic Interoperability for Better Health and Safer Healthcare*. Technical report, The European Commission, 2009.

[55] K. Holzer and W. Gall. *Utilizing IHE-based Electronic Health Record systems for secondary use.* Methods of Information in Medicine, 50(4):319–25, January 2011.

[56] H.U. Prokosch and T. Ganslandt. *Perspectives for Medical Informatics*. Methods of Information in Medicine, pages 38–44, 2009.

[57] Jessica S. Ancker, Sarah Shih, Mytri P. Singh, and Andrew Snyder. *Root Causes Underlying Challenges to Secondary Use of Data*. In AMIA Symposium Proceedings, pages 57–62, 2011.

[58] James J. Cimino. *Collect Once, Use Many: Enabling the Reuse of Clinical Data through Controlled Terminologies*. Journal of AHIMA, 78(2):24–29, 2007.

[59] R. H. Brook, E. A. McGlynn, and P. D. Cleary. *Measuring quality of care*. New England Journal of Medicine, 335(13):966–970, 1996.

[60] S. M. Campbell, J. Braspenning, A. Hutchinson, and M. Marshall. *Research methods used in developing and applying quality indicators in primary care*. Quality and Safety in Health Care, 11(4):358–64, April 2002.

[61] Lisa M. Kern, Sameer Malhotra, Yolanda Barron, Jill Quaresimo, Rina Dhopeshwarkar, Michelle Pichardo, Alison M. Edwards, and Rainu Kaushal. *Accuracy of Electronically Reported "Meaningful Use" Clinical Quality Measures*. Annals of Internal Medicine, 158:77–83, 2013.

[62] Brian C. Drolet and Kevin B. Johnson. *Categorizing the world of registries.* Journal of Biomedical Informatics, 41(6):1009–20, December 2008.

[63] Kathrin Dentler, Annette ten Teije, Nicolette F. de Keizer, and Ronald Cornet. *Barriers to the Reuse of Routinely Recorded Clinical Data: A Field Report*. In Studies in Health Technology and Informatics, 2013.

[64] Danielle G. T. Arts, Nicolette F. de Keizer, and Gert-Jan Scheffer. *Defining and improving data quality in medical registries: a literature review, case study, and generic framework.* ...of the American Medical ..., pages 600–611, 2002.

[65] Zichtbare Zorg. *Kwaliteitsindicatoren*, 2012.

[66] W. van Gijn and C. J .H. van de Velde. *Improving quality of cancer care through surgical audit.* European Journal of Surgical Oncology (EJSO), 36 Suppl 1, September 2010.

[67] Dutch Institute for Clinical Auditing. *DICA Rapportages 2011*. 2011.

[68] Kathrin Dentler. *Formalised colorectal cancer surgery quality indicators.*, 2014.

[69] Patrick S. Romano, Hillary J. Mull, Peter E. Rivard, Shibei Zhao, William G. Henderson, Susan Loveland, Dennis Tsilimingras, Cindy L. Christiansen, and Amy K. Rosen. *Validity of selected AHRQ patient safety indicators based on VA National Surgical Quality Improvement Program data.* Health Services Research, 44(1):182–204, February 2009.

[70] J. M. Bland and B. K. Butland. *Comparing proportions in overlapping samples*. Technical report, 2011.

[71] Nicole Gray Weiskopf and Chunhua Weng. *Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.* Journal of the American Medical Informatics Association, 20(1):144–151, June 2012.

[72] S. D. Persell, J. M. Wright, J. A. Thompson, K. S. Kmetik, and D. W. Baker. *Assessing the validity of national quality measures for coronary artery disease using an electronic health record*. Archives of Internal Medicine, 166(20):2272, November 2006.

[73] David W. Baker, Stephen D. Persell, Jason A. Thompson, Neilesh Soman, Karen M. Burgner, David Liss, and Karen S. Kmetik. *Automated Review of Electronic Health Records to Assess Quality of Care for Outpatients with Heart Failure*. Annals of Internal Medicine, 146(4):270–277, 2006.

[74] J. P. Weiner, J. B. Fowles, and K. S. Chan. *New paradigms for measuring clinical performance using electronic health records*. International Journal for Quality in Health Care, 24(3):200–205, 2012.

[75] Paul C. Fu, Daniel Rosenthal, Joshua M. Pevnick, and Floyd Eisenberg. *The impact of emerging standards adoption on automated quality reporting.* Journal of Biomedical Informatics, 45(4):772–81, August 2012.

[76] Ferdinand T. Velasco, Floyd Eisenberg, and John Cooper. *Deriving Quality Measures from Electronic Health Record Systems*. In AMIA Symposium Proceedings, pages 1411–14, 2010.

[77] Eve A. Kerr, Dylan M. Smith, Mary M. Hogan, Sahah L. Krein, Leonard Pogach, Timothy P. Hofer, and Rodney A. Hayward. *Comparing clinical automated, medical record, and hybrid data sources for diabetes quality measures*. Journal on Quality Improvement, 28(10):555–565, 2002.

[78] Amanda Parsons, Colleen McCullough, Jason Wang, and Sarah Shih. *Validity of electronic health record-derived quality measurement for performance monitoring.* Journal of the American Medical Informatics Association, 19(4):604–609, 2012.

[79] Catherine H. MacLean, Rachel Louie, Paul G. Shekelle, Carol P. Roth, Debra Saliba, Takahiro Higashi, John Adams, John T. Chang, Caren J. Kamberg, David H. Solomon, Roy T. Young, and Neil S. Wenger. *Comparison of Administrative Data and Medical Records to Measure the Quality of Medical Care Provided to Vulnerable Older Patients*. Medical Care, 44(2):141–148, 2006.

[80] Paul C. Tang, Mary Ralston, Michelle Fernandez Arrigotti, Lubna Qureshi, and Justin Graham. *Comparison of Methodologies for Calculating Quality Measures Based on Administrative Data versus Clinical Data from an Electronic Health Record System: Implications for Performance Measures*. Journal of the American Medical Informatics Association, 14(1):10–15, 2007.

[81] A. A. Warsi, S. White, and P. McCulloch. *Completeness of data entry in three cancer surgery databases*. European Journal of Surgical Oncology (EJSO), 28(8):850–856, December 2002.

[82] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. *Secondary Use of EHR: Data Quality Issues and Informatics Opportunities.* AMIA Summits on Translational Science, pages 1–5, January 2010.

[83] George Hripcsak and David J. Albers. *Next-generation phenotyping of electronic health records.* Journal of the American Medical Informatics Association, pages 1–5, September 2012.

[84] Ibrahim Edhemovic, Walley J. Temple, Christopher J. de Gara, and Gavin C. E. Stuart. *The computer synoptic operative report - a leap forward in the science of surgery*. Annals of Surgical Oncology, 11(10):941–947, 2004.

[85] L. A. Mack, O. F. Bathe, M. A. Hebert, E. Tamano, W. D. Buie, T. Fields, and W. J. Temple. *Opening the black box of cancer surgery quality: WebSMR and the Alberta experience*. Journal of Surgical Oncology, 99(9):525–530, June 2009.

[86] Jason Park, Venu G. Pillarisetty, Murray F. Brennan, William R. Jarnagin, Michael I. D'Angelica, Ronald P. Dematteo, Daniel G. Coit, Maria Janakos, and Peter J. Allen. *Electronic synoptic operative reporting: assessing the reliability and completeness of synoptic reports for pancreatic resection.* Journal of the American College of Surgeons, 211(3):308–15, September 2010.

[87] Ronald Cornet, Armand Van Eldik, and Nicolette De Keizer. *Inventory of tools for dutch clinical language processing*. Studies in health technology and informatics, 180:245, 2012.

[88] Jeremy Wyatt. *Acquisition and use of clinical data for audit and research*. Journal of Evaluation in Clinical Practice, 1(1):15–27, 1995.

[89] David Moner, José Alberto Maldonado, Diego Bosca, Jesualdo Tomas Fernandez-Breis, Carlos Angulo, Pere Crespo, Pedro J. Vivancos, and Montserrat Robles. *Archetype-based semantic integration and standardization of clinical data*. Engineering in Medicine and Biology Society, pages 5141–4, January 2006.

[90] Sebastian Garde, Petra Knaup, Thilo Schuler, and Evelyn Hovenga. *Can openEHR Archetypes Empower Multi-Centre Clinical Research?* Studies in Health Technology and Informatics, 116:971–6, January 2005.

[91] Sebastian Garde, Evelyn Hovenga, Jasmin Buck, and Petra Knaup. *Expressing clinical data sets with openEHR archetypes: a solid basis for ubiquitous computing*. International Journal of Medical Informatics, 76 Suppl 3:S334–41, December 2007.

[92] Christian D. Kohl, Sebastian Garde, and Petra Knaup. *Facilitating secondary use of medical data by using openEHR archetypes*. Studies in Health Technology and Informatics, 160(Pt 2):1117, 2010.

[93] Rong Chen, Patrik Georgii-Hemming, and Hans Å hlfeldt. *Representing a chemotherapy guideline using openEHR and rules*. Studies in Health Technology and Informatics, pages 653–657, 2009.

[94] Mar Marcos, José Alberto Maldonado, Begoña Martínez-Salvador, David Moner, Diego Boscá, and Montserrat Robles. *An archetype-based solution for the interoperability of computerised guidelines and electronic health records*. Artificial Intelligence in Medicine, pages 276–285, 2011.

[95] Leonardo Lezcano, Miguel-Angel Sicilia, and Carlos Rodríguez-Solano. *Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules*. Journal of Biomedical Informatics, 44(2):343–353, November 2011.

[96] Catalina Martínez-Costa, Marcos Menárguez-Tortosa, Jesualdo Tomás Fernández-Breis, and José Alberto Maldonado. *A model-driven approach for representing clinical archetypes for Semantic Web environments*. Journal of Biomedical Informatics, 42(1):150–64, February 2009.

[97] Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, and Ruslan Velkov. *OWLIM: A family of scalable semantic repositories*. Semantic Web Journal, 2(1):33–42, 2011.

[98] Alan L. Rector, Rahil Qamar, and Tom Marley. *Binding ontologies and coding systems to electronic health records and messages*. Applied Ontology, 4(1):51–69, 2009.

[99] Marcos Menárguez-Tortosa and Jesualdo Tomas Fernandez-Breis. *Validation of the openEHR Archetype Library by using OWL Reasoning*. Studies in Health Technology and Informatics, 169:789, January 2011.

[100] Stijn Heymans, Matthew McKennirey, and Joshua Phillips. *Semantic validation of the use of SNOMED CT in HL7 clinical documents*. Journal of Biomedical Semantics, 2(2), July 2011.

[101] Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. *The description logic handbook: theory, implementation, and applications*. Cambridge: Cambridge University Press, 2003.

[102] Franz Baader, Carsten Lutz, and Boontawee Suntisrivaraporn. *Is Tractable Reasoning in Extensions of the Description Logic EL Useful in Practice?* In Proceedings of the Methods for Modalities Workshop, page 5, 2005.

[103] Yuanbo Guo, Zhengxiang Pan, and Jeff Heflin. *LUBM: A Benchmark for OWL Knowledge Base Systems*. Web Semantics: Science, Services and Agents on the World Wide Web, 3(2-3):158–182, 2005.

[104] Li Ma, Yang Yang, Zhaoming Qiu, Guotong Xie, Yue Pan, and Shengping Liu. *Towards a Complete OWL Ontology Benchmark*. The Semantic Web: Research and Applications, pages 125–139, 2006.

[105] Tom Gardiner, Dmitry Tsarkov, and Ian Horrocks. *Framework for an Automated Comparison of Description Logic Reasoners*. In The Semantic Web - ISWC 2006, volume 4273, pages 654–667. Springer, 2006.

[106] Anni-Yasmin Turhan, Sean Bechhofer, Alissa Kaplunova, Thorsten Liebig, Marko Luther, Ralf Möller, Olaf Noppens, Peter Patel-Schneider, and Timo Suntisrivaraporn, Boontawee Weithöner. *DIG 2.0 - Towards a Flexible Interface for Description Logic Reasoners*. In 2nd OWL Experiences and Directions Workshop, volume 6. Citeseer, 2006.

[107] Boris Motik and R Studer. *KAON2-A scalable reasoning tool for the semantic web*. In The Semantic Web: Research and Applications, volume 17, 2005.

[108] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. *Pellet: A practical OWL-DL reasoner*. Web Semantics: Science, Services and Agents on the World Wide Web, 5(2):51–53, June 2007.

[109] Volker Haarslev and Ralf Müller. *RACER System Description*. Automated Reasoning, 2083:701–705, 2001.

[110] Jürgen Bock, Peter Haase, Qiu Ji, and Raphael Volz. *Benchmarking OWL reasoners*. In AReа2008 - Workshop on Advancing Reasoning on the Web: Scalability and Commonsense, 2008.

[111] Rob Shearer, Boris Motik, and Ian Horrocks. *HermiT: a Highly-Efficient OWL Reasoner*. In 5th OWL Experiences and Directions Workshop, 2008.

[112] Thorsten Liebig. *Reasoning with OWL - System Support and Insights*. Technical report, Universität Ulm, 2006.

[113] Alexandre Riazanov and Andrei Voronkov. *Vampire 1.1*. In Automated Reasoning, pages 376–380. Springer, 2001.

[114] Marko Luther, Thorsten Liebig, Sebastian Böhm, and Olaf Noppens. *Who the Heck Is the Father of Bob?* In The Semantic Web: Research and Applications, pages 66–80, 2009.

[115] Jeremy J Carroll, Ian Dickinson, Chris Dollin, Dave Reynolds, Andy Seaborne, and Kevin Wilkinson. *Jena: Implementing the Semantic Web Recommendations*. In Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, pages 74–83, 2004.

[116] Ravi Bhushan Mishra and Sandeep Kumar. *Semantic web reasoners and languages*. Artificial Intelligence Review, 35(4):339–368, December 2011.

[117] Boontawee Suntisrivaraporn. *Empirical evaluation of reasoning in lightweight DLs on life science ontologies*. In Proceedings of the 2nd Mahasarakham International Workshop on AI, 2008.

[118] Julian Mendez and Boontawee Suntisrivaraporn. *Reintroducing CEL as an OWL 2 EL Reasoner*. In Proceedings of the 2009 International Workshop on Description Logics, volume 477, 2009.

[119] Franz Baader, Carsten Lutz, and Boontawee Suntisrivaraporn. *CEL - A Polynomial-time Reasoner for Life Science Ontologies*. In Proceedings of the 3rd International Joint Conference on Automated Reasoning, volume 4130, pages 287–291. Springer, 2006.

[120] Michael J. Lawley and Cyril Bousquet. *Fast Classification in Protege: Snorocket as an OWL2 EL Reasoner*. In Australasian Ontology Workshop, pages 45–49, 2010.

[121] Ian Horrocks, Ulrike Sattler, and Stephan Tobies. *Practical Reasoning for Very Expressive Description Logics*. Logic Journal of IGPL, 8(3):239–264, 2000.

[122] Boris Motik, Rob Shearer, and Ian Horrocks. *Optimized Reasoning in Description Logics using Hypertableaux*. In Proceedings of the 21st International Conference on Automated Deduction, volume 4603, pages 67–83. Springer, 2007.

[123] Sebastian Rudolph, Tuvshintur Tserendorj, and Pascal Hitzler. *What Is Approximate Reasoning?* Web Reasoning and Rule Systems, pages 150–164, 2008.

[124] Ian R. Horrocks. *Optimising tableaux decision procedures for Description Logics*. PhD thesis, University of Manchester, 1997.

[125] Stephan Tobies. *Complexity Results and Practical Algorithms for Logics in Knowledge Representation*. PhD thesis, RWTH Aachen, 2001.

[126] Yevgeny Kazakov. *SRIQ and SROIQ are Harder than SHOIQ*. In Eleventh International Conference on Principles of Knowledge Representation and Reasoning, 2008.

[127] Sebastian Brandt. *Polynomial Time Reasoning in a Description Logic with Existential Restrictions, GCI Axioms, and-What Else?* In ECAI, volume 16, pages 298–302, 2004.

[128] Boontawee Suntisrivaraporn. *Polynomial-Time Reasoning Support for Design and Maintenance of Large-Scale Biomedical Ontologies*. PhD thesis, TU Dresden, 2009.

[129] Markus Krötzsch, Sebastian Rudolph, and Pascal Hitzler. *Description Logic Rules*. In ECAI, pages 80–84. IOS Press, 2008.

[130] Markus Krötzsch, Sebastian Rudolph, and Pascal Hitzler. *ELP: Tractable Rules for OWL 2*. In The Semantic Web - ISWC 2008, pages 649–664. Springer, 2008.

[131] Matthew Horridge, Bijan Parsia, and Ulrike Sattler. *Justification Oriented Proofs in OWL*. In The Semantic Web - ISWC 2010, volume 6496, pages 354–369. Springer, 2010.

[132] Francesco M. Donini, Maurizio Lenzerini, Daniele Nardi, and Andrea Schaerf. *Reasoning in Description Logics*. Principles of knowledge representation, pages 191–236, 1996.

[133] Thorsten Liebig, Marko Luther, Olaf Noppens, Mariano Rodriguez, Diego Calvanese, Michael Wessel, Matthew Horridge, Sean Bechhofer, Dmitry Tsarkov, and Evren Sirin. *OWLlink: DIG for OWL 2*. In 5th OWL Experiences and Directions Workshop, 2008.

[134] Franz Baader, Bernhard Ganter, Baris Sertkaya, and Ulrike Sattler. *Completing description logic knowledge bases using formal concept analysis*. IJCAI, 7:230–235, 2007.

[135] V. Haarslev, R. Möller, and S. Wandelt. *The revival of structural subsumption in tableau-based description logic reasoners*. In Proceedings of the 21st International Workshop on Description Logics, 2008.

[136] Edward Thomas, Jeff Z. Pan, and Yuan Ren. *TrOWL: Tractable OWL 2 Reasoning Infrastructure*. In The Semantic Web: Research and Applications, pages 431–435. Springer, 2010.

[137] Yuan Ren, Jeff Z. Pan, and Yuting Zhao. *Soundness Preserving Approximation for TBox Reasoning*. In Proceedings of the 25th AAAI Conference, pages 351–356, 2010.

[138] Samantha Bail, Bijan Parsia, and Ulrike Sattler. *JustBench: A Framework for OWL Benchmarking*. In The Semantic Web - ISWC 2010, 2010.

[139] Bijan Parsia, Christian Halaschek-Wiener, and Evren Sirin. *Towards Incremental Reasoning Through Updates in OWL-DL*. In Reasoning on the Web, 2006.

[140] Xinxin Zhu, Jung-Wei Fan, David M. Baorto, Chunhua Weng, and James J. Cimino. *A review of auditing methods applied to the content of controlled biomedical terminologies*. Journal of Biomedical Informatics, 42(3):413–25, June 2009.

[141] Kent A. Spackman. *Normal forms for description logic expressions of clinical concepts in SNOMED RT*. AMIA Symposium Proceedings, pages 627–31, January 2001.

[142] Stefan Schlobach and Ronald Cornet. *Logical Support for Terminological Modeling*. Studies in Health Technology and Informatics, 107:439–43, January 2004.

[143] Stephan Grimm and Jens Wissmann. *Elimination of redundancy in ontologies*. The Semantic Web: Research and Applications, pages 260–274, 2011.

[144] Ronald Cornet and Stefan Schulz. *Relationship groups in SNOMED CT*. Studies in Health Technology and Informatics, 150:223–227, 2009.

[145] Yevgeny Kazakov, Markus Krötzsch, and František Simančík. *Concurrent Classification of EL Ontologies*. The Semantic Web - ISWC 2011, pages 305–320, 2011.

[146] Ronald Cornet and Ameen Abu-Hanna. *Auditing description-logic-based medical terminological systems by detecting equivalent concept definitions*. International Journal of Medical Informatics, 77(5):336–45, May 2008.

[147] Ronald Cornet and Ameen Abu-Hanna. *Two DL-based methods for auditing medical terminological systems*. In AMIA Symposium Proceedings, pages 166–70, January 2005.

[148] James J. Cimino. *Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS metathesaurus*. In AMIA Symposium Proceedings, pages 120–4, January 2001.

[149] Ronald Denaux, Dhaval Thakker, Vania Dimitrova, and Anthony G. Cohn. *Entendre: Interactive Semantic Feedback for Ontology Authoring*. In The Semantic Web - ISWC 2011 - Demo Track, 2011.

[150] Yi Peng, Michael H. Halper, Yehoshua Perl, and James Geller. *Auditing the UMLS for redundant classifications*. In AMIA Symposium Proceedings, pages 612–6, January 2002.

[151] Rafael S. Gonçalves, Bijan Parsia, and Ulrike Sattler. *Ecco: A Hybrid Diff Tool for OWL 2 ontologies*. In 9th OWL Experiences and Directions Workshop, 2012.

[152] Zhisheng Huang, Frank Van Harmelen, A ten Teije, and Kathrin Dentler. *Knowledge-based Patient Data Generation*. In Knowledge Representation for Health-Care, 2013.

[153] H. R. Rubin, P. Pronovost, and G. B. Diette. *From a process of care to a measure: the development and testing of a quality indicator.* International Journal for Quality in Health Care, 13(6):489–96, December 2001.

[154] Jan Mainz. *Defining and classifying clinical indicators for quality improvement*. International Journal for Quality in Health Care, 15(6):523–30, December 2003.

[155] Jan Mainz. *Developing evidence-based clinical indicators : a state of the art*. International Journal for Quality in Health Care, 15:5–12, 2003.

[156] Louise C. Walter, Natalie P. Davidowitz, Paul A. Heineken, and Kenneth E. Covinsky. *Pitfalls of Converting Practice Guidelines*. JAMA, 291(20), 2004.

[157] A. R. Aronson. *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.* AMIA Symposium Proceedings, pages 17–21, January 2001.

[158] Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, and Mark A. Musen. *BioPortal: ontologies and integrated data resources at the click of a mouse.* Nucleic acids research, 37(Web Server issue):W170–3, July 2009.

A SUMMARY

Today, hospitals and general practitioners are requested to compute dramatically increasing numbers of healthcare quality indicators to monitor and to improve the quality of their delivered care, and to be compared with each other. This computation is mostly performed manually, which is time-consuming, expensive and error-prone because indicators are typically released in inherently ambiguous natural language. A concurrent development is the rapid adoption of Electronic Medical Records, resulting in increased volumes of routinely recorded healthcare data, and opening the door to the automated computation of quality indicators. Therefore, the main research question tackled in this thesis is:

> *Under which conditions can healthcare quality indicators be computed automatically by reusing data already collected during the clinical care process?*

We demonstrated that the automated computation of healthcare quality indicators by reusing data already collected during the clinical care process is feasible. However, several conditions must be met:

I) Indicators and their formalisation

Indicators need to be formalised to be automatically computable. Based on a literature study and a requirements analysis, we developed CLIF, a method to formalise quality indicators into unambiguous queries that can be run against patient data. We created a web-based tool that implements the formalisation method to lead users through the formalisation process, and thereafter examined the method's reproducibility in a case study, and its generalisability by formalising the entire set of Dutch indicators for general practitioners. Our studies confirmed that CLIF can lead to maximally reproducible results and that it is generalisable to a broad set of Dutch quality indicators, but that unambiguous indicators and the cooperation of trained experts are required. Both the tool and the sets of formalised indicators have been made available online.[1]

II) Patient data and its (re)usability

1 https://github.com/kathrinrin/clif, http://figshare.com/authors/Kathrin_Dentler/452665

Regarding the secondary use of patient data, we conducted our research in the clinical setting of the GIOCA, the Gastro-Intestinal Oncology Centre Amsterdam. To automatically compute quality indicators, patient data needs to be available. We attempted to gather all raw source data that is required to compute the set of indicators relevant for the GIOCA. We identified barriers that impede the secondary use of patient data and provided recommendations on how to prevent them. Patient data needs to be of adequate quality to be reused. We assessed the data quality inside our hospital by comparison to data submitted to the DSCA, the Dutch Surgical Colorectal Audit, and its influence on quality indicator results by a statistical analysis. We demonstrated that data quality can have a significant impact on quality indicator results. High quality implies the use of well-established healthcare standards, so that the meaning of data becomes machine-processable. To integrate data from various heterogeneous sources, and also to bridge the semantic gap between indicators and patient data, standard information models and large, lightweight, logics-based terminologies such as SNOMED CT play an important role. We represented both the data and the indicators based on the standard information model *open*EHR archetypes, and proved the concept by automatically computing the formalised indicators.

III) Semantic interoperability

Automated reasoners provide support to meaningfully use logics-based terminologies, for example by selecting not only certain concepts, but also their sub-concepts. Based on a literature study, we defined characteristics of reasoning engines, and categorised eight reasoners along these characteristics. Our results show that reasoners can vary substantially, and that their characteristics should be taken into account when choosing a reasoner for a specific application scenario. Finally, the quality of medical terminologies must be ensured by automated auditing methods, as these terminologies are typically too large to be audited manually. A relevant quality factor is non-reduncancy. We extended and operationalised an already existing definition to detect intra-axiom redundancies in SNOMED CT and showed that 12% of concepts in the employed SNOMED CT version contained redundant elements.

The presented results are a basis to support clinical practice and further areas of research where quality indicators are used to improve health outcomes of patients.

B SAMENVATTING

Ziekenhuizen en huisartsen moeten tegenwoordig een groot en toenemend aantal van kwaliteitsindicatoren berekenen en aanleveren om de kwaliteit van hun zorg te monitoren en te verbeteren, maar ook om verantwoording af te leggen en met elkaar vergeleken te worden. De berekening van indicatoren gebeurt meestal handmatig, maar het proces is tijdintensief, duur en foutgevoelig omdat indicatoren normaliter in ambigue natuurlijke taal geschreven zijn. Gelijktijdig maken zorgverleners toenemend gebruik van het elektronisch patiëntendossier, waardoor steeds meer routinematig verzamelde zorggegevens beschikbaar worden. Daarom is de hoofdonderzoeksvraag van dit proefschrift, getiteld "Automatisch berekenen van kwaliteitsindicatoren in de zorg - Hergebruik van patiëntdata en semantische interoperabiliteit":

> *Onder welke voorwaarden kunnen kwaliteitsindicatoren in de zorg automatisch worden berekend door hergebruik van reeds tijdens de klinische zorgproces verzamelde gegevens?*

We lieten zien dat de geautomatiseerde berekening van kwaliteitsindicatoren door hergebruik van routinematig verzamelde zorggegevens haalbaar is. Echter er moet aan een aantal voorwaarden voldaan zijn:

I) Indicatoren en hun formalisering

Indicatoren moeten worden geformaliseerd om ze automatisch te kunnen berekenen. Op basis van een literatuurstudie en een analyse van eisen ontwikkelden we CLIF, een methode om kwaliteitsindicatoren te formaliseren in eenduidige queries die op basis van patiëntgegevens kunnen worden uitgevoerd. We implementeerden een web-based tool om gebruikers stap voor stap door het proces van formalisering te leiden. Daarna hebben we de reproduceerbaarheid van onze methode in een case study onderzocht en de generaliseerbaarheid getoetst door de gehele set van Nederlandse indicatoren voor huisartsen te formaliseren. Deze studies toonden aan dat CLIF tot reproduceerbare resultaten lijdt, en dat de methode generaliseerbaar is naar een brede set van Nederlandse kwaliteitsindicatoren, maar tevens dat eenduidig beschreven indicatoren en samenwerking tussen getrainde experts noodzakelijke voorwaarden zijn. Zowel onze tool als ook de sets van geformaliseerde indicatoren zijn online beschikbaar.[1]

1 https://github.com/kathrinrin/clif, http://figshare.com/authors/Kathrin_Dentler/452665

II) (Her-)bruikbaarheid van patiëntgegevens

We onderzochten het secundair gebruik van patiëntgegevens in de klinische setting van de GIOCA, het Gastro-Intestinaal Oncologisch Centrum Amsterdam. Om kwaliteitsindicatoren automatisch te kunnen berekenen moeten patiëntgegevens beschikbaar zijn. Door onze inspanningen om alle brongegevens te verzamelen die nodig zijn om de voor de GIOCA relevante set van indicatoren te berekenen, identificeerden we barrières die het secundaire gebruik van patiëntgegevens hinderen en definieerden we aanbevelingen om deze te slechten. Patiëntgegevens moeten van voldoende kwaliteit zijn om hergebruikt te kunnen worden. Wij hebben de kwaliteit van de gegevens in één ziekenhuis vergeleken met gegevens uit de DSCA, de Dutch Surgical Colorectal Audit, en we hebben aangetoond dat de kwaliteit van de gegevens de indicatorresultaten aanzienlijk kan beïnvloeden. Om goede datakwaliteit te realiseren is het gebruik van standaarden en normen uit de gezondheidszorg essentieel omdat hierdoor de betekenis van data voor computers verwerkbaar wordt. Om gegevens uit diverse heterogene bronnen te integreren, en ook om de semantische kloof tussen indicatoren en patiëntgegevens te overbruggen, spelen informatiemodellen en grote, logica-gebaseerde terminologieën zoals SNOMED CT een belangrijke rol. Wij representeerden zowel de patiëntgegevens en als de indicatoren op basis van openEHR archetypen en lieten zien dat de geformaliseerde indicatoren hierdoor automatisch te berekenen zijn.

III) Semantische interoperabiliteit

Geautomatiseerde reasoners (redeneertools) bieden ondersteuning om logica-gebaseerde terminologieën zinvol te gebruiken, bijvoorbeeld door het selecteren van bepaalde concepten, maar ook van hun subconcepten. Op basis van een literatuurstudie identificeerden we kenmerken van reasoners, en categoriseerden acht reasoners langs deze kenmerken. Onze resultaten tonen aan dat de reasoners aanzienlijk kunnen variëren. Daarom moet bij het kiezen van een reasoner voor een specifieke toepassing rekening gehouden worden met hun kenmerken. Ten slotte moet de kwaliteit van medische terminologieën gewaarborgd worden door geautomatiseerde controle, aangezien deze terminologieën meestal te groot zijn voor handmatige controle. Een relevante kwaliteitskenmerk is het vermijden van redundantie. We hebben een reeds bestaande definitie voor intra-axioma redundantie uitgebreid en geoperationaliseerd, en konden aantonen dat 12% van de concepten in de gebruikte SNOMED CT versie overbodige elementen bevatte.

De gepresenteerde resultaten zijn een basis om de klinische praktijk en verdere gebieden van onderzoek te ondersteunen waar kwaliteitsindicatoren gebruikt worden om gezondheidsresultaten van patiënten te verbeteren.

SIKS Dissertatiereeks

====
1998
====

1998-1 Johan van den Akker (CWI)
DEGAS - An Active, Temporal Database of Autonomous Objects

1998-2 Floris Wiesman (UM)
Information Retrieval by Graphically Browsing Meta-Information

1998-3 Ans Steuten (TUD)
A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective

1998-4 Dennis Breuker (UM)
Memory versus Search in Games

1998-5 E.W.Oskamp (RUL)
Computerondersteuning bij Straftoemeting

====
1999
====

1999-1 Mark Sloof (VU)
Physiology of Quality Change Modelling;
Automated modelling of Quality Change of Agricultural Products

1999-2 Rob Potharst (EUR)
Classification using decision trees and neural nets

1999-3 Don Beal (UM)
The Nature of Minimax Search

1999-4 Jacques Penders (UM)
The practical Art of Moving Physical Objects

1999-5 Aldo de Moor (KUB)
Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems

1999-6 Niek J.E. Wijngaards (VU)
Re-design of compositional systems

1999-7 David Spelt (UT)
Verification support for object database design

1999-8 Jacques H.J. Lenting (UM)
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.

====
2000
====

2000-1 Frank Niessink (VU)

Perspectives on Improving Software Maintenance

2000-2 Koen Holtman (TUE)
Prototyping of CMS Storage Management

2000-3 Carolien M.T. Metselaar (UVA)
Sociaal-organisatorische gevolgen van kennistechnologie;
een procesbenadering en actorperspectief.

2000-4 Geert de Haan (VU)
ETAG, A Formal Model of Competence Knowledge for User Interface Design

2000-5 Ruud van der Pol (UM)
Knowledge-based Query Formulation in Information Retrieval.

2000-6 Rogier van Eijk (UU)
Programming Languages for Agent Communication

2000-7 Niels Peek (UU)
Decision-theoretic Planning of Clinical Patient Management

2000-8 Veerle Coup (EUR)
Sensitivity Analyis of Decision-Theoretic Networks

2000-9 Florian Waas (CWI)
Principles of Probabilistic Query Optimization

2000-10 Niels Nes (CWI)
Image Database Management System Design Considerations,
Algorithms and Architecture

2000-11 Jonas Karlsson (CWI)
Scalable Distributed Data Structures for Database Management

====
2001
====

2001-1 Silja Renooij (UU)
Qualitative Approaches to Quantifying Probabilistic Networks

2001-2 Koen Hindriks (UU)
Agent Programming Languages: Programming with Mental Models

2001-3 Maarten van Someren (UvA)
Learning as problem solving

2001-4 Evgueni Smirnov (UM)
Conjunctive and Disjunctive Version Spaces with
Instance-Based Boundary Sets

2001-5 Jacco van Ossenbruggen (VU)
Processing Structured Hypermedia: A Matter of Style

2001-6 Martijn van Welie (VU)
Task-based User Interface Design

2001-7 Bastiaan Schonhage (VU)
Diva: Architectural Perspectives on Information Visualization

2001-8 Pascal van Eck (VU)
A Compositional Semantic Structure for Multi-Agent Systems Dynamics.

2001-9 Pieter Jan 't Hoen (RUL)
Towards Distributed Development of Large Object-Oriented Models,
Views of Packages as Classes

2001-10 Maarten Sierhuis (UvA)
Modeling and Simulating Work Practice
BRAHMS: a multiagent modeling and simulation language
for work practice analysis and design

2001-11 Tom M. van Engers (VUA)
Knowledge Management:
The Role of Mental Models in Business Systems Design

====
2002
====

2002-01 Nico Lassing (VU)
Architecture-Level Modifiability Analysis

2002-02 Roelof van Zwol (UT)
Modelling and searching web-based document collections

2002-03 Henk Ernst Blok (UT)
Database Optimization Aspects for Information Retrieval

2002-04 Juan Roberto Castelo Valdueza (UU)
The Discrete Acyclic Digraph Markov Model in Data Mining

2002-05 Radu Serban (VU)
The Private Cyberspace Modeling Electronic Environments
inhabited by Privacy-concerned Agents

2002-06 Laurens Mommers (UL)
Applied legal epistemology;
Building a knowledge-based ontology of the legal domain

2002-07 Peter Boncz (CWI)
Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications

2002-08 Jaap Gordijn (VU)
Value Based Requirements Engineering: Exploring Innovative
E-Commerce Ideas

2002-09 Willem-Jan van den Heuvel (KUB)
Integrating Modern Business Applications with Objectified Legacy Systems

2002-10 Brian Sheppard (UM)
Towards Perfect Play of Scrabble

2002-11 Wouter C.A. Wijngaards (VU)
Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12 Albrecht Schmidt (Uva)
Processing XML in Database Systems

2002-13 Hongjing Wu (TUE)
A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU)
Agent Interaction: Abstract Approaches to Modelling, Programming and

Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT)
Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16 Pieter van Langen (VU)
The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UVA)
Understanding, Modeling, and Improving Main-Memory Database Performance

====
2003
====

2003-01 Heiner Stuckenschmidt (VU)
Ontology-Based Information Sharing in Weakly Structured Environments

2003-02 Jan Broersen (VU)
Modal Action Logics for Reasoning About Reactive Systems

2003-03 Martijn Schuemie (TUD)
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04 Milan Petkovic (UT)
Content-Based Video Retrieval Supported by Database Technology

2003-05 Jos Lehmann (UVA)
Causation in Artificial Intelligence and Law - A modelling approach

2003-06 Boris van Schooten (UT)
Development and specification of virtual environments

2003-07 Machiel Jansen (UvA)
Formal Explorations of Knowledge Intensive Tasks

2003-08 Yongping Ran (UM)
Repair Based Scheduling

2003-09 Rens Kortmann (UM)
The resolution of visually guided behaviour

2003-10 Andreas Lincke (UvT)
Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture

2003-11 Simon Keizer (UT)
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT)
Dutch speech recognition in multimedia information retrieval

2003-13 Jeroen Donkers (UM)
Nosce Hostem - Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN)
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD)
Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI)

Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17 David Jansen (UT)
Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM)
Learning Search Decisions

====
2004
====

2004-01 Virginia Dignum (UU)
A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02 Lai Xu (UvT)
Monitoring Multi-party Contracts for E-business

2004-03 Perry Groot (VU)
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04 Chris van Aart (UVA)
Organizational Principles for Multi-Agent Architectures

2004-05 Viara Popova (EUR)
Knowledge discovery and monotonicity

2004-06 Bart-Jan Hommes (TUD)
The Evaluation of Business Process Modeling Techniques

2004-07 Elise Boltjes (UM)
Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes

2004-08 Joop Verbeek (UM)
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise

2004-09 Martin Caminada (VU)
For the Sake of the Argument; explorations into argument-based reasoning

2004-10 Suzanne Kabel (UVA)
Knowledge-rich indexing of learning-objects

2004-11 Michel Klein (VU)
Change Management for Distributed Ontologies

2004-12 The Duy Bui (UT)
Creating emotions and facial expressions for embodied agents

2004-13 Wojciech Jamroga (UT)
Using Multiple Models of Reality: On Agents who Know how to Play

2004-14 Paul Harrenstein (UU)
Logic in Conflict. Logical Explorations in Strategic Equilibrium

2004-15 Arno Knobbe (UU)
Multi-Relational Data Mining

2004-16 Federico Divina (VU)
Hybrid Genetic Relational Search for Inductive Learning

2004-17 Mark Winands (UM)
Informed Search in Complex Games

2004-18 Vania Bessa Machado (UvA)
Supporting the Construction of Qualitative Knowledge Models

2004-19 Thijs Westerveld (UT)
Using generative probabilistic models for multimedia retrieval

2004-20 Madelon Evers (Nyenrode)
Learning from Design: facilitating multidisciplinary design teams

====
2005
====

2005-01 Floor Verdenius (UVA)
Methodological Aspects of Designing Induction-Based Applications

2005-02 Erik van der Werf (UM)
AI techniques for the game of Go

2005-03 Franc Grootjen (RUN)
A Pragmatic Approach to the Conceptualisation of Language

2005-04 Nirvana Meratnia (UT)
Towards Database Support for Moving Object data

2005-05 Gabriel Infante-Lopez (UVA)
Two-Level Probabilistic Grammars for Natural Language Parsing

2005-06 Pieter Spronck (UM)
Adaptive Game AI

2005-07 Flavius Frasincar (TUE)
Hypermedia Presentation Generation for Semantic Web Information Systems

2005-08 Richard Vdovjak (TUE)
A Model-driven Approach for Building Distributed Ontology-based Web Applications

2005-09 Jeen Broekstra (VU)
Storage, Querying and Inferencing for Semantic Web Languages

2005-10 Anders Bouwer (UVA)
Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments

2005-11 Elth Ogston (VU)
Agent Based Matchmaking and Clustering - A Decentralized Approach to Search

2005-12 Csaba Boer (EUR)
Distributed Simulation in Industry

2005-13 Fred Hamburg (UL)
Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen

2005-14 Borys Omelayenko (VU)
Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics

2005-15 Tibor Bosse (VU)
Analysis of the Dynamics of Cognitive Processes

2005-16 Joris Graaumans (UU)
Usability of XML Query Languages

2005-17 Boris Shishkov (TUD)
Software Specification Based on Re-usable Business Components

2005-18 Danielle Sent (UU)
Test-selection strategies for probabilistic networks

2005-19 Michel van Dartel (UM)
Situated Representation

2005-20 Cristina Coteanu (UL)
Cyber Consumer Law, State of the Art and Perspectives

2005-21 Wijnand Derks (UT)
Improving Concurrency and Recovery in Database Systems by
Exploiting Application Semantics

====
2006
====

2006-01 Samuil Angelov (TUE)
Foundations of B2B Electronic Contracting

2006-02 Cristina Chisalita (VU)
Contextual issues in the design and use of information technology in organizations

2006-03 Noor Christoph (UVA)
The role of metacognitive skills in learning to solve problems

2006-04 Marta Sabou (VU)
Building Web Service Ontologies

2006-05 Cees Pierik (UU)
Validation Techniques for Object-Oriented Proof Outlines

2006-06 Ziv Baida (VU)
Software-aided Service Bundling - Intelligent Methods & Tools
for Graphical Service Modeling

2006-07 Marko Smiljanic (UT)
XML schema matching – balancing efficiency and effectiveness by means of clustering

2006-08 Eelco Herder (UT)
Forward, Back and Home Again - Analyzing User Behavior on the Web

2006-09 Mohamed Wahdan (UM)
Automatic Formulation of the Auditor's Opinion

2006-10 Ronny Siebes (VU)
Semantic Routing in Peer-to-Peer Systems

2006-11 Joeri van Ruth (UT)
Flattening Queries over Nested Data Types

2006-12 Bert Bongers (VU)
Interactivation - Towards an e-cology of people, our technological environment, and the arts

2006-13 Henk-Jan Lebbink (UU)
Dialogue and Decision Games for Information Exchanging Agents

2006-14 Johan Hoorn (VU)
Software Requirements:
Update, Upgrade, Redesign - towards a Theory of Requirements Change

2006-15 Rainer Malik (UU)
CONAN: Text Mining in the Biomedical Domain

2006-16 Carsten Riggelsen (UU)
Approximation Methods for Efficient Learning of Bayesian Networks

2006-17 Stacey Nagata (UU)
User Assistance for Multitasking with Interruptions on a Mobile Device

2006-18 Valentin Zhizhkun (UVA)
Graph transformation for Natural Language Processing

2006-19 Birna van Riemsdijk (UU)
Cognitive Agent Programming: A Semantic Approach

2006-20 Marina Velikova (UvT)
Monotone models for prediction in data mining

2006-21 Bas van Gils (RUN)
Aptness on the Web

2006-22 Paul de Vrieze (RUN)
Fundaments of Adaptive Personalisation

2006-23 Ion Juvina (UU)
Development of Cognitive Model for Navigating on the Web

2006-24 Laura Hollink (VU)
Semantic Annotation for Retrieval of Visual Resources

2006-25 Madalina Drugan (UU)
Conditional log-likelihood MDL and Evolutionary MCMC

2006-26 Vojkan Mihajlovic (UT)
Score Region Algebra: A Flexible Framework for Structured Information Retrieval

2006-27 Stefano Bocconi (CWI)
Vox Populi: generating video documentaries from semantically annotated media repositories

2006-28 Borkur Sigurbjornsson (UVA)
Focused Information Access using XML Element Retrieval

====
2007
====

2007-01 Kees Leune (UvT)
Access Control and Service-Oriented Architectures

2007-02 Wouter Teepe (RUG)
Reconciling Information Exchange and Confidentiality: A Formal Approach

2007-03 Peter Mika (VU)
Social Networks and the Semantic Web

2007-04 Jurriaan van Diggelen (UU)
Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach

2007-05 Bart Schermer (UL)
Software Agents, Surveillance, and the Right to Privacy:
a Legislative Framework for Agent-enabled Surveillance

2007-06 Gilad Mishne (UVA)
Applied Text Analytics for Blogs

2007-07 Natasa Jovanovic (UT)
To Whom It May Concern - Addressee Identification in Face-to-Face Meetings

2007-08 Mark Hoogendoorn (VU)
Modeling of Change in Multi-Agent Organizations

2007-09 David Mobach (VU)
Agent-Based Mediated Service Negotiation

2007-10 Huib Aldewereld (UU)
Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols

2007-11 Natalia Stash (TUE)
Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System

2007-12 Marcel van Gerven (RUN)
Bayesian Networks for Clinical Decision Support:
A Rational Approach to Dynamic Decision-Making under Uncertainty

2007-13 Rutger Rienks (UT)
Meetings in Smart Environments; Implications of Progressing Technology

2007-14 Niek Bergboer (UM)
Context-Based Image Analysis

2007-15 Joyca Lacroix (UM)
NIM: a Situated Computational Memory Model

2007-16 Davide Grossi (UU)
Designing Invisible Handcuffs.
Formal investigations in Institutions and Organizations for Multi-agent Systems

2007-17 Theodore Charitos (UU)
Reasoning with Dynamic Networks in Practice

2007-18 Bart Orriens (UvT)
On the development an management of adaptive business collaborations

2007-19 David Levy (UM)
Intimate relationships with artificial partners

2007-20 Slinger Jansen (UU)
Customer Configuration Updating in a Software Supply Network

2007-21 Karianne Vermaas (UU)
Fast diffusion and broadening use:
A research on residential adoption and usage of broadband internet
in the Netherlands between 2001 and 2005

2007-22 Zlatko Zlatev (UT)
Goal-oriented design of value and process models from patterns

2007-23 Peter Barna (TUE)
Specification of Application Logic in Web Information Systems

2007-24 Georgina Ramírez Camps (CWI)
Structural Features in XML Retrieval

2007-25 Joost Schalken (VU)
Empirical Investigations in Software Process Improvement

====
2008
====

2008-01 Katalin Boer-Sorbán (EUR)
Agent-Based Simulation of Financial Markets: A modular,continuous-time approach

2008-02 Alexei Sharpanskykh (VU)
On Computer-Aided Methods for Modeling and Analysis of Organizations

2008-03 Vera Hollink (UVA)
Optimizing hierarchical menus: a usage-based approach

2008-04 Ander de Keijzer (UT)
Management of Uncertain Data - towards unattended integration

2008-05 Bela Mutschler (UT)
Modeling and simulating causal dependencies
on process-aware information systems from a cost perspective

2008-06 Arjen Hommersom (RUN)
On the Application of Formal Methods to Clinical Guidelines,
an Artificial Intelligence Perspective

2008-07 Peter van Rosmalen (OU)
Supporting the tutor in the design and support of adaptive e-learning

2008-08 Janneke Bolt (UU)
Bayesian Networks: Aspects of Approximate Inference

2008-09 Christof van Nimwegen (UU)
The paradox of the guided user: assistance can be counter-effective

2008-10 Wauter Bosma (UT)
Discourse oriented summarization

2008-11 Vera Kartseva (VU)
Designing Controls for Network Organizations: A Value-Based Approach

2008-12 Jozsef Farkas (RUN)
A Semiotically Oriented Cognitive Model of Knowledge Representation

2008-13 Caterina Carraciolo (UVA)
Topic Driven Access to Scientific Handbooks
2008-14 Arthur van Bunningen (UT)
Context-Aware Querying; Better Answers with Less Effort
2008-15 Martijn van Otterlo (UT)
The Logic of Adaptive Behavior: Knowledge Representation and Algorithms
for the Markov Decision Process Framework in First-Order Domains.

2008-16 Henriette van Vugt (VU)
Embodied agents from a user's perspective

2008-17 Martin Op 't Land (TUD)
Applying Architecture and Ontology to the Splitting and Allying of Enterprises

2008-18 Guido de Croon (UM)
Adaptive Active Vision

2008-19 Henning Rode (UT)
From Document to Entity Retrieval:
Improving Precision and Performance of Focused Text Search

2008-20 Rex Arendsen (UVA)
Geen bericht, goed bericht.
Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer
met de overheid op de administratieve lasten van bedrijven

2008-21 Krisztian Balog (UVA)
People Search in the Enterprise

2008-22 Henk Koning (UU)
Communication of IT-Architecture

2008-23 Stefan Visscher (UU)
Bayesian network models for the management of ventilator-associated pneumonia

2008-24 Zharko Aleksovski (VU)
Using background knowledge in ontology matching

2008-25 Geert Jonker (UU)
Efficient and Equitable Exchange in Air Traffic Management Plan Repair
using Spender-signed Currency

2008-26 Marijn Huijbregts (UT)
Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled

2008-27 Hubert Vogten (OU)
Design and Implementation Strategies for IMS Learning Design

2008-28 Ildiko Flesch (RUN)
On the Use of Independence Relations in Bayesian Networks

2008-29 Dennis Reidsma (UT)
Annotations and Subjective Machines -
Of Annotators, Embodied Agents, Users, and Other Humans

2008-30 Wouter van Atteveldt (VU)
Semantic Network Analysis:
Techniques for Extracting, Representing and Querying Media Content

2008-31 Loes Braun (UM)
Pro-Active Medical Information Retrieval

2008-32 Trung H. Bui (UT)
Toward Affective Dialogue Management
using Partially Observable Markov Decision Processes

2008-33 Frank Terpstra (UVA)
Scientific Workflow Design; theoretical and practical issues

2008-34 Jeroen de Knijf (UU)
Studies in Frequent Tree Mining

2008-35 Ben Torben Nielsen (UvT)
Dendritic morphologies: function shapes structure

====
2009
====

2009-01 Rasa Jurgelenaite (RUN)
Symmetric Causal Independence Models

2009-02 Willem Robert van Hage (VU)
Evaluating Ontology-Alignment Techniques

2009-03 Hans Stol (UvT)
A Framework for Evidence-based Policy Making Using IT

2009-04 Josephine Nabukenya (RUN)
Improving the Quality of Organisational Policy Making using Collaboration Engineering

2009-05 Sietse Overbeek (RUN)
Bridging Supply and Demand for Knowledge Intensive Tasks
- Based on Knowledge, Cognition, and Quality

2009-06 Muhammad Subianto (UU)
Understanding Classification

2009-07 Ronald Poppe (UT)
Discriminative Vision-Based Recovery and Recognition of Human Motion

2009-08 Volker Nannen (VU)
Evolutionary Agent-Based Policy Analysis in Dynamic Environments

2009-09 Benjamin Kanagwa (RUN)
Design, Discovery and Construction of Service-oriented Systems

2009-10 Jan Wielemaker (UVA)
Logic programming for knowledge-intensive interactive applications

2009-11 Alexander Boer (UVA)
Legal Theory, Sources of Law & the Semantic Web

2009-12 Peter Massuthe (TUE, Humboldt-Universität zu Berlin)
Operating Guidelines for Services

2009-13 Steven de Jong (UM)
Fairness in Multi-Agent Systems

2009-14 Maksym Korotkiy (VU)
From ontology-enabled services to service-enabled ontologies
(making ontologies work in e-science with ONTO-SOA)

2009-15 Rinke Hoekstra (UVA)
Ontology Representation - Design Patterns and Ontologies that Make Sense

2009-16 Fritz Reul (UvT)
New Architectures in Computer Chess

2009-17 Laurens van der Maaten (UvT)
Feature Extraction from Visual Data

2009-18 Fabian Groffen (CWI)
Armada, An Evolving Database System

2009-19 Valentin Robu (CWI)
Modeling Preferences, Strategic Reasoning and Collaboration
in Agent-Mediated Electronic Markets

2009-20 Bob van der Vecht (UU)
Adjustable Autonomy: Controling Influences on Decision Making

2009-21 Stijn Vanderlooy (UM)
Ranking and Reliable Classification

2009-22 Pavel Serdyukov (UT)
Search For Expertise: Going beyond direct evidence

2009-23 Peter Hofgesang (VU)
Modelling Web Usage in a Changing Environment

2009-24 Annerieke Heuvelink (VUA)
Cognitive Models for Training Simulations

2009-25 Alex van Ballegooij (CWI)
"RAM: Array Database Management through Relational Mapping"

2009-26 Fernando Koch (UU)
An Agent-Based Model for the Development of Intelligent Mobile Services

2009-27 Christian Glahn (OU)
Contextual Support of social Engagement and Reflection on the Web

2009-28 Sander Evers (UT)
Sensor Data Management with Probabilistic Models

2009-29 Stanislav Pokraev (UT)
Model-Driven Semantic Integration of Service-Oriented Applications

2009-30 Marcin Zukowski (CWI)
Balancing vectorized query execution with bandwidth-optimized storage

2009-31 Sofiya Katrenko (UVA)
A Closer Look at Learning Relations from Text

2009-32 Rik Farenhorst (VU) and Remco de Boer (VU)
Architectural Knowledge Management: Supporting Architects and Auditors

2009-33 Khiet Truong (UT)
How Does Real Affect Affect Affect Recognition In Speech?

2009-34 Inge van de Weerd (UU)
Advancing in Software Product Management: An Incremental Method Engineering Approach

2009-35 Wouter Koelewijn (UL)
Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling

2009-36 Marco Kalz (OUN)
Placement Support for Learners in Learning Networks

2009-37 Hendrik Drachsler (OUN)
Navigation Support for Learners in Informal Learning Networks

2009-38 Riina Vuorikari (OU)
Tags and self-organisation: a metadata ecology for learning resources in a multilingual context

2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin)
Service Substitution – A Behavioral Approach Based on Petri Nets

2009-40 Stephan Raaijmakers (UvT)
Multinomial Language Learning: Investigations into the Geometry of Language

2009-41 Igor Berezhnyy (UvT)
Digital Analysis of Paintings

2009-42 Toine Bogers
Recommender Systems for Social Bookmarking

2009-43 Virginia Nunes Leal Franqueira (UT)

Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients

2009-44 Roberto Santana Tapia (UT)
Assessing Business-IT Alignment in Networked Organizations

2009-45 Jilles Vreeken (UU)
Making Pattern Mining Useful

2009-46 Loredana Afanasiev (UvA)
Querying XML: Benchmarks and Recursion

====
2010
====

2010-01 Matthijs van Leeuwen (UU)
Patterns that Matter

2010-02 Ingo Wassink (UT)
Work flows in Life Science

2010-03 Joost Geurts (CWI)
A Document Engineering Model and Processing Framework for Multimedia documents

2010-04 Olga Kulyk (UT)
Do You Know What I Know?
Situational Awareness of Co-located Teams in Multidisplay Environments

2010-05 Claudia Hauff (UT)
Predicting the Effectiveness of Queries and Retrieval Systems

2010-06 Sander Bakkes (UvT)
Rapid Adaptation of Video Game AI

2010-07 Wim Fikkert (UT)
Gesture interaction at a Distance

2010-08 Krzysztof Siewicz (UL)
Towards an Improved Regulatory Framework of Free Software.
Protecting user freedoms in a world of software communities and eGovernments

2010-09 Hugo Kielman (UL)
A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging

2010-10 Rebecca Ong (UL)
Mobile Communication and Protection of Children

2010-11 Adriaan Ter Mors (TUD)
The world according to MARP: Multi-Agent Route Planning

2010-12 Susan van den Braak (UU)
Sensemaking software for crime analysis

2010-13 Gianluigi Folino (RUN)
High Performance Data Mining using Bio-inspired techniques

2010-14 Sander van Splunter (VU)
Automated Web Service Reconfiguration

2010-15 Lianne Bodenstaff (UT)
Managing Dependency Relations in Inter-Organizational Models

2010-16 Sicco Verwer (TUD)
Efficient Identification of Timed Automata, theory and practice

2010-17 Spyros Kotoulas (VU)
Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications

2010-18 Charlotte Gerritsen (VU)
Caught in the Act: Investigating Crime by Agent-Based Simulation

2010-19 Henriette Cramer (UvA)
People's Responses to Autonomous and Adaptive Systems

2010-20 Ivo Swartjes (UT)
Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative

2010-21 Harold van Heerde (UT)
Privacy-aware data management by means of data degradation

2010-22 Michiel Hildebrand (CWI)
End-user Support for Access to Heterogeneous Linked Data

2010-23 Bas Steunebrink (UU)
The Logical Structure of Emotions

2010-24 Dmytro Tykhonov
Designing Generic and Efficient Negotiation Strategies

2010-25 Zulfiqar Ali Memon (VU)
Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective

2010-26 Ying Zhang (CWI)
XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines

2010-27 Marten Voulon (UL)
Automatisch contracteren

2010-28 Arne Koopman (UU)
Characteristic Relational Patterns

2010-29 Stratos Idreos (CWI)
Database Cracking: Towards Auto-tuning Database Kernels

2010-30 Marieke van Erp (UvT)
Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval

2010-31 Victor de Boer (UVA)
Ontology Enrichment from Heterogeneous Sources on the Web

2010-32 Marcel Hiel (UvT)
An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems

2010-33 Robin Aly (UT)
Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval

2010-34 Teduh Dirgahayu (UT)
Interaction Design in Service Compositions

2010-35 Dolf Trieschnigg (UT)
Proof of Concept: Concept-based Biomedical Information Retrieval

2010-36 Jose Janssen (OU)
Paving the Way for Lifelong Learning;
Facilitating competence development through a learning path specification

2010-37 Niels Lohmann (TUE)

Correctness of services and their composition

2010-38 Dirk Fahland (TUE)
From Scenarios to components

2010-39 Ghazanfar Farooq Siddiqui (VU)
Integrative modeling of emotions in virtual agents

2010-40 Mark van Assem (VU)
Converting and Integrating Vocabularies for the Semantic Web

2010-41 Guillaume Chaslot (UM)
Monte-Carlo Tree Search

2010-42 Sybren de Kinderen (VU)
Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach

2010-43 Peter van Kranenburg (UU)
A Computational Approach to Content-Based Retrieval of Folk Song Melodies

2010-44 Pieter Bellekens (TUE)
An Approach towards Context-sensitive and User-adapted Access
to Heterogeneous Data Sources, Illustrated in the Television Domain

2010-45 Vasilios Andrikopoulos (UvT)
A theory and model for the evolution of software services

2010-46 Vincent Pijpers (VU)
e3alignment: Exploring Inter-Organizational Business-ICT Alignment

2010-47 Chen Li (UT)
Mining Process Model Variants: Challenges, Techniques, Examples

2010-48 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets

2010-49 Jahn-Takeshi Saito (UM)
Solving difficult game positions

2010-50 Bouke Huurnink (UVA)
Search in Audiovisual Broadcast Archives

2010-51 Alia Khairia Amin (CWI)
Understanding and supporting information seeking tasks in multiple sources

2010-52 Peter-Paul van Maanen (VU)
Adaptive Support for Human-Computer Teams:
Exploring the Use of Cognitive Models of Trust and Attention

2010-53 Edgar Meij (UVA)
Combining Concepts and Language Models for Information Access

====
2011
====

2011-01 Botond Cseke (RUN)
Variational Algorithms for Bayesian Inference in Latent Gaussian Models

2011-02 Nick Tinnemeier (UU)
Organizing Agent Organizations.
Syntax and Operational Semantics of an Organization-Oriented Programming Language

2011-03 Jan Martijn van der Werf (TUE)
Compositional Design and Verification of Component-Based Information Systems

2011-04 Hado van Hasselt (UU)
Insights in Reinforcement Learning;
Formal analysis and empirical evaluation of temporal-difference learning algorithms

2011-05 Base van der Raadt (VU)
Enterprise Architecture Coming of Age -
Increasing the Performance of an Emerging Discipline.

2011-06 Yiwen Wang (TUE)
Semantically-Enhanced Recommendations in Cultural Heritage

2011-07 Yujia Cao (UT)
Multimodal Information Presentation for High Load Human Computer Interaction

2011-08 Nieske Vergunst (UU)
BDI-based Generation of Robust Task-Oriented Dialogues

2011-09 Tim de Jong (OU)
Contextualised Mobile Media for Learning

2011-10 Bart Bogaert (UvT)
Cloud Content Contention

2011-11 Dhaval Vyas (UT)
Designing for Awareness: An Experience-focused HCI Perspective

2011-12 Carmen Bratosin (TUE)
Grid Architecture for Distributed Process Mining

2011-13 Xiaoyu Mao (UvT)
Airport under Control. Multiagent Scheduling for Airport Ground Handling

2011-14 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets

2011-15 Marijn Koolen (UvA)
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

2011-16 Maarten Schadd (UM)
Selective Search in Games of Different Complexity

2011-17 Jiyin He (UVA)
Exploring Topic Structure: Coherence, Diversity and Relatedness

2011-18 Mark Ponsen (UM)
Strategic Decision-Making in complex games

2011-19 Ellen Rusman (OU)
The Mind's Eye on Personal Profiles

2011-20 Qing Gu (VU)
Guiding service-oriented software engineering - A view-based approach

2011-21 Linda Terlouw (TUD)
Modularization and Specification of Service-Oriented Systems

2011-22 Junte Zhang (UVA)
System Evaluation of Archival Description and Access

2011-23 Wouter Weerkamp (UVA)
Finding People and their Utterances in Social Media

2011-24 Herwin van Welbergen (UT)
Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

2011-25 Syed Waqar ul Qounain Jaffry (VU)
Analysis and Validation of Models for Trust Dynamics

2011-26 Matthijs Aart Pontier (VU)
Virtual Agents for Human Communication- Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

2011-27 Aniel Bhulai (VU)
Dynamic website optimization through autonomous management of design patterns

2011-28 Rianne Kaptein (UVA)
Effective Focused Retrieval by Exploiting Query Context and Document Structure

2011-29 Faisal Kamiran (TUE)
Discrimination-aware Classification

2011-30 Egon van den Broek (UT)
Affective Signal Processing (ASP): Unraveling the mystery of emotions

2011-31 Ludo Waltman (EUR)
Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

2011-32 Nees-Jan van Eck (EUR)
Methodological Advances in Bibliometric Mapping of Science

2011-33 Tom van der Weide (UU)
Arguing to Motivate Decisions

2011-34 Paolo Turrini (UU)
Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

2011-35 Maaike Harbers (UU)
Explaining Agent Behavior in Virtual Training

2011-36 Erik van der Spek (UU)
Experiments in serious game design: a cognitive approach

2011-37 Adriana Burlutiu (RUN)
Machine Learning for Pairwise Data,
Applications for Preference Learning and Supervised Network Inference

2011-38 Nyree Lemmens (UM)
Bee-inspired Distributed Optimization

2011-39 Joost Westra (UU)
Organizing Adaptation using Agents in Serious Games

2011-40 Viktor Clerc (VU)
Architectural Knowledge Management in Global Software Development

2011-41 Luan Ibraimi (UT)
Cryptographically Enforced Distributed Data Access Control

2011-42 Michal Sindlar (UU)
Explaining Behavior through Mental State Attribution

2011-43 Henk van der Schuur (UU)
Process Improvement through Software Operation Knowledge

2011-44 Boris Reuderink (UT)

Robust Brain-Computer Interfaces

2011-45 Herman Stehouwer (UvT)
Statistical Language Models for Alternative Sequence Selection

2011-46 Beibei Hu (TUD)
Towards Contextualized Information Delivery:
A Rule-based Architecture for the Domain of Mobile Police Work

2011-47 Azizi Bin Ab Aziz (VU)
Exploring Computational Models for Intelligent Support of Persons with Depression

2011-48 Mark Ter Maat (UT)
Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

2011-49 Andreea Niculescu (UT)
Conversational interfaces for task-oriented spoken dialogues:
design aspects influencing interaction quality

====
2012
====

2012-01 Terry Kakeeto (UvT)
Relationship Marketing for SMEs in Uganda

2012-02 Muhammad Umair (VU)
Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

2012-03 Adam Vanya (VU)
Supporting Architecture Evolution by Mining Software Repositories

2012-04 Jurriaan Souer (UU)
Development of Content Management System-based Web Applications

2012-05 Marijn Plomp (UU)
Maturing Interorganisational Information Systems

2012-06 Wolfgang Reinhardt (OU)
Awareness Support for Knowledge Workers in Research Networks

2012-07 Rianne van Lambalgen (VU)
When the Going Gets Tough:
Exploring Agent-based Models of Human Performance under Demanding Conditions

2012-08 Gerben de Vries (UVA)
Kernel Methods for Vessel Trajectories

2012-09 Ricardo Neisse (UT)
Trust and Privacy Management Support for Context-Aware Service Platforms

2012-10 David Smits (TUE)
Towards a Generic Distributed Adaptive Hypermedia Environment

2012-11 J.C.B. Rantham Prabhakara (TUE)
Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

2012-12 Kees van der Sluijs (TUE)
Model Driven Design and Data Integration in Semantic Web Information Systems

2012-13 Suleman Shahid (UvT)
Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

2012-14 Evgeny Knutov (TUE)
Generic Adaptation Framework for Unifying Adaptive Web-based Systems

2012-15 Natalie van der Wal (VU)
Social Agents. Agent-Based Modelling of Integrated Internal
and Social Dynamics of Cognitive and Affective Processes

2012-16 Fiemke Both (VU)
Helping people by understanding them -
Ambient Agents supporting task execution and depression treatment

2012-17 Amal Elgammal (UvT)
Towards a Comprehensive Framework for Business Process Compliance

2012-18 Eltjo Poort (VU)
Improving Solution Architecting Practices

2012-19 Helen Schonenberg (TUE)
What's Next? Operational Support for Business Process Execution

2012-20 Ali Bahramisharif (RUN)
Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing

2012-21 Roberto Cornacchia (TUD)
Querying Sparse Matrices for Information Retrieval

2012-22 Thijs Vis (UvT)
Intelligence, politie en veiligheidsdienst: verenigbare grootheden?

2012-23 Christian Muehl (UT)
Toward Affective Brain-Computer Interfaces:
Exploring the Neurophysiology of Affect during Human Media Interaction

2012-24 Laurens van der Werff (UT)
Evaluation of Noisy Transcripts for Spoken Document Retrieval

2012-25 Silja Eckartz (UT)
Managing the Business Case Development in Inter-Organizational IT Projects:
A Methodology and its Application

2012-26 Emile de Maat (UVA)
Making Sense of Legal Text

2012-27 Hayrettin Gurkok (UT)
Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games

2012-28 Nancy Pascall (UvT)
Engendering Technology Empowering Women

2012-29 Almer Tigelaar (UT)
Peer-to-Peer Information Retrieval

2012-30 Alina Pommeranz (TUD)
Designing Human-Centered Systems for Reflective Decision Making

2012-31 Emily Bagarukayo (RUN)
A Learning by Construction Approach for Higher Order Cognitive Skills Improvement,
Building Capacity and Infrastructure

2012-32 Wietske Visser (TUD)
Qualitative multi-criteria preference representation and reasoning

2012-33 Rory Sie (OUN)
Coalitions in Cooperation Networks (COCOON)

2012-34 Pavol Jancura (RUN)
Evolutionary analysis in PPI networks and applications

2012-35 Evert Haasdijk (VU)
Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics

2012-36 Denis Ssebugwawo (RUN)
Analysis and Evaluation of Collaborative Modeling Processes

2012-37 Agnes Nakakawa (RUN)
A Collaboration Process for Enterprise Architecture Creation

2012-38 Selmar Smit (VU)
Parameter Tuning and Scientific Testing in Evolutionary Algorithms

2012-39 Hassan Fatemi (UT)
Risk-aware design of value and coordination networks

2012-40 Agus Gunawan (UvT)
Information Access for SMEs in Indonesia

2012-41 Sebastian Kelle (OU)
Game Design Patterns for Learning

2012-42 Dominique Verpoorten (OU)
Reflection Amplifiers in self-regulated Learning

2012-43 Withdrawn

2012-44 Anna Tordai (VU)
On Combining Alignment Techniques

2012-45 Benedikt Kratz (UvT)
A Model and Language for Business-aware Transactions

2012-46 Simon Carter (UVA)
Exploration and Exploitation of Multilingual Data for Statistical Machine Translation

2012-47 Manos Tsagkias (UVA)
Mining Social Media: Tracking Content and Predicting Behavior

2012-48 Jorn Bakker (TUE)
Handling Abrupt Changes in Evolving Time-series Data

2012-49 Michael Kaisers (UM)
Learning against Learning -
Evolutionary dynamics of reinforcement learning algorithms in strategic interactions

2012-50 Steven van Kervel (TUD)
Ontologogy driven Enterprise Information Systems Engineering

2012-51 Jeroen de Jong (TUD)
Heuristics in Dynamic Sceduling;
a practical framework with a case study in elevator dispatching

====
2013
====

2013-01 Viorel Milea (EUR)
News Analytics for Financial Decision Support

2013-02 Erietta Liarou (CWI)
MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing

2013-03 Szymon Klarman (VU)
Reasoning with Contexts in Description Logics

2013-04 Chetan Yadati (TUD)
Coordinating autonomous planning and scheduling

2013-05 Dulce Pumareja (UT)
Groupware Requirements Evolutions Patterns

2013-06 Romulo Goncalves (CWI)
The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience

2013-07 Giel van Lankveld (UvT)
Quantifying Individual Player Differences

2013-08 Robbert-Jan Merk (VU)
Making enemies: cognitive modeling for opponent agents in fighter pilot simulators

2013-09 Fabio Gori (RUN)
Metagenomic Data Analysis: Computational Methods and Applications

2013-10 Jeewanie Jayasinghe Arachchige (UvT)
A Unified Modeling Framework for Service Design.

2013-11 Evangelos Pournaras (TUD)
Multi-level Reconfigurable Self-organization in Overlay Services

2013-12 Marian Razavian (VU)
Knowledge-driven Migration to Services

2013-13 Mohammad Safiri (UT)
Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly

2013-14 Jafar Tanha (UVA)
Ensemble Approaches to Semi-Supervised Learning Learning

2013-15 Daniel Hennes (UM)
Multiagent Learning - Dynamic Games and Applications

2013-16 Eric Kok (UU)
Exploring the practical benefits of argumentation in multi-agent deliberation

2013-17 Koen Kok (VU)
The PowerMatcher: Smart Coordination for the Smart Electricity Grid

2013-18 Jeroen Janssens (UvT)
Outlier Selection and One-Class Classification

2013-19 Renze Steenhuizen (TUD)
Coordinated Multi-Agent Planning and Scheduling

2013-20 Katja Hofmann (UvA)
Fast and Reliable Online Learning to Rank for Information Retrieval

2013-21 Sander Wubben (UvT)
Text-to-text generation by monolingual machine translation

2013-22 Tom Claassen (RUN)
Causal Discovery and Logic

2013-23 Patricio de Alencar Silva (UvT)
Value Activity Monitoring

2013-24 Haitham Bou Ammar (UM)
Automated Transfer in Reinforcement Learning

2013-25 Agnieszka Anna Latoszek-Berendsen (UM)
Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System

2013-26 Alireza Zarghami (UT)
Architectural Support for Dynamic Homecare Service Provisioning

2013-27 Mohammad Huq (UT)
Inference-based Framework Managing Data Provenance

2013-28 Frans van der Sluis (UT)
When Complexity becomes Interesting: An Inquiry into the Information eXperience

2013-29 Iwan de Kok (UT)
Listening Heads

2013-30 Joyce Nakatumba (TUE)
Resource-Aware Business Process Management: Analysis and Support

2013-31 Dinh Khoa Nguyen (UvT)
Blueprint Model and Language for Engineering Cloud Applications

2013-32 Kamakshi Rajagopal (OUN)
Networking For Learning;
The role of Networking in a Lifelong Learner's Professional Development

2013-33 Qi Gao (TUD)
User Modeling and Personalization in the Microblogging Sphere

2013-34 Kien Tjin-Kam-Jet (UT)
Distributed Deep Web Search

2013-35 Abdallah El Ali (UvA)
Minimal Mobile Human Computer Interaction

2013-36 Than Lam Hoang (TUe)
Pattern Mining in Data Streams

2013-37 Dirk Börner (OUN)
Ambient Learning Displays

2013-38 Eelco den Heijer (VU)
Autonomous Evolutionary Art

2013-39 Joop de Jong (TUD)
A Method for Enterprise Ontology based Design of Enterprise Information Systems

2013-40 Pim Nijssen (UM)
Monte-Carlo Tree Search for Multi-Player Games

2013-41 Jochem Liem (UVA)
Supporting the Conceptual Modelling of Dynamic Systems:
A Knowledge Engineering Perspective on Qualitative Reasoning

2013-42 Léon Planken (TUD)
Algorithms for Simple Temporal Reasoning

2013-43 Marc Bron (UVA)

Exploration and Contextualization through Interaction and Concepts

====
2014
====

2014-01 Nicola Barile (UU)
Studies in Learning Monotone Models from Data

2014-02 Fiona Tuliyano (RUN)
Combining System Dynamics with a Domain Modeling Method

2014-03 Sergio Raul Duarte Torres (UT)
Information Retrieval for Children: Search Behavior and Solutions

2014-04 Hanna Jochmann-Mannak (UT)
Websites for children: search strategies and interface design -
Three studies on children's search performance and evaluation

2014-05 Jurriaan van Reijsen (UU)
Knowledge Perspectives on Advancing Dynamic Capability

2014-06 Damian Tamburri (VU)
Supporting Networked Software Development

2014-07 Arya Adriansyah (TUE)
Aligning Observed and Modeled Behavior

2014-08 Samur Araujo (TUD)
Data Integration over Distributed and Heterogeneous Data Endpoints

2014-09 Philip Jackson (UvT)
Toward Human-Level Artificial Intelligence:
Representation and Computation of Meaning in Natural Language

2014-10 Ivan Razo-Zapata (VU)
Service Value Networks

2014-11 Janneke van der Zwaan (TUD)
An Empathic Virtual Buddy for Social Support

2014-12 Willem van Willigen (VU)
Look Ma, No Hands: Aspects of Autonomous Vehicle Control

2014-13 Arlette van Wissen (VU)
Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains

2014-14 Yangyang Shi (TUD)
Language Models with Meta-information

2014-15 Nataliya Mogles (VU)
Agent-Based Analysis and Support of Human Functioning in Complex
Socio-Technical Systems: Applications in Safety and Healthcare

2014-16 Krystyna Milian (VU)
Supporting trial recruitment and design by automatically interpreting eligibility criteria