Vaccine Semantics

Automatic methods for recognizing, representing, and reasoning about vaccine-related information

Vaccin Semantiek

Geautomatiseerde methoden om vaccin-gerelateerde informatie te herkennen, te representeren, en erover te redeneren

Thesis

to obtain the degree of Doctor from the Erasmus University Rotterdam by command of the rector magnificus

Prof. dr. R.C.M.E. Engels

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on Tuesday 8 January 2019 at 15:30 hrs by

Benedikt Ferdinand Hellmut Becker

born in Benediktbeuern, Germany.

Erasmus University Rotterdam

Doctoral Committee

PROMOTER

Prof. dr. M.C.J.M. Sturkenboom

OTHER MEMBERS

Prof. dr. B.H.Ch. Stricker
Prof. dr. N.F. de Keizer
Dr. M.A.B. van der Sande

COPROMOTOR

Dr. ir. J.A. Kors

VACCINE SEMANTICS

Automatic methods for recognizing, representing, and reasoning about vaccine-related information

BENEDIKT BECKER

Chapter 3 of this thesis was partially developed in the context of the WHO project SPHQ13 – LOA 209. The research for chapters 4, 6 and 7 was funded by the Innovative Medicines Initiative Joint Undertaking under ADVANCE grant agreement 15557, with financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution.

Cover picture by Félix Becker Morales.

Typeset in LaTeX using the Palatino font
and based on the classicthesis package.

Digital version and online material available
at *http://hdl.handle.net/1765/111218*.

Für Maciel und Félix.

CONTENTS

1 INTRODUCTION 1

I LISTENING TO VACCINE SAFETY CONCERNS AND PUBLIC SENTIMENT

2 SOCIAL MEDIA FOR VACCINE SAFETY SURVEILLANCE 15
- 2.1 Introduction 16
- 2.2 Methods 17
- 2.3 Results 18
- 2.4 Discussion 24
- 2.5 Conclusions 29

3 SOCIAL MEDIA FOR FOLLOWING A VACCINE DEBATE 31
- 3.1 Introduction 32
- 3.2 Methods 33
- 3.3 Results 34
- 3.4 Discussion 38
- 3.5 Conclusion 40

II ACCESSING EXISTING EVIDENCE

4 VACCINE RECOGNITION AND CLASSIFICATION OF SCIENTIFIC ARTICLES 43
- 4.1 Background and significance 44
- 4.2 Materials and Methods 45
- 4.3 Results 50
- 4.4 Discussion 53

5 EXTRACTION OF CHEMICAL-INDUCED DISEASES 57
- 5.1 Introduction 57
- 5.2 Methods 58
- 5.3 Results 64
- 5.4 Discussion 68

III VERIFYING VACCINE B/R HYPOTHESES

6 SEMIAUTOMATIC CODING OF CASE DEFINITIONS 73
- 6.1 Introduction 74
- 6.2 Methods 75
- 6.3 Results 82
- 6.4 Discussion 84

7 ALIGNMENT OF VACCINE CODES USING THE VACCO ONTOLOGY 87

7.1 Background 88
7.2 Methods 90
7.3 Results 99
7.4 Discussion 103
7.5 Conclusion 104

8 GENERAL DISCUSSION 107
9 SUMMARY 117

BIBLIOGRAPHY 124

LIST OF FIGURES

Figure 1.1 Heterogeneous representation of medical outcomes 7
Figure 1.2 Approaches to harmonizing extraction queries 7
Figure 1.3 Aims and resources addressed in this thesis 11
Figure 2.1 Assertions of rosiglitazone/cardiovascular event-related posts 21
Figure 2.2 Assertions of HPV vaccine/infertility-related posts 25
Figure 3.1 Number of messages about the pentavalent vaccine 36
Figure 3.2 Authors of messages about the pentavalent vaccine 37
Figure 4.1 Example annotations from the reference corpus 46
Figure 4.2 Performance measures for the automatic indexing of vaccine literature 53
Figure 4.3 Heading-specific performance of models $VaccO_{VAC}$ and *CNN* 54
Figure 5.1 Workflow for CDR extraction 59
Figure 5.2 Example dependency parse tree 63
Figure 6.1 Key phases of CodeMapper and use of the UMLS 76
Figure 6.2 Screens of the CodeMapper application 79
Figure 6.3 Automatic evaluation of CodeMapper 81
Figure 6.4 Error categories of the CodeMapper evaluation 82
Figure 7.1 Structure of the VaccO ontology 91
Figure 7.2 Example for the compilation of vaccine code descriptors in VaccO 96
Figure 7.3 F-scores of the alignment algorithm with five similarity measures 101
Figure 8.1 Methods applied in this thesis 109
Figure 8.2 The proliferation of standards 112

LIST OF TABLES

Table 2.1 Overview of posts about rosiglitazone and cardiovascular adverse events 19
Table 2.2 Description of referenced web pages (rosiglitazone/cardiovascular events) 20
Table 2.3 Overview of posts about HPV vaccine and infertility 23
Table 2.4 Description of referenced web pages (HPV/infertility) 24
Table 3.1 Author countries and countries in content of social medial messages 37
Table 4.1 Annotations in the reference corpus of vaccine descriptions 51
Table 4.2 Performance measures of method $VaccO_{VDR}$ 51
Table 4.3 Error analysis of $VaccO_{VDR}$ 52
Table 5.1 Characteristics of the CDR corpus 59
Table 5.2 Performance of the Peregrine system 65
Table 5.3 Error analysis of RELigator 65
Table 5.4 Comparison of relation extraction systems 66
Table 5.5 Comparison of relation extraction systems on the CDR test data 67
Table 6.1 Case definitions and reference sets 80
Table 6.2 Performance measures of CodeMapper 83
Table 6.3 False-positive errors by CodeMapper 84
Table 6.4 False-negative errors by CodeMapper 84
Table 7.1 Property categories used to define groups 90
Table 7.2 Example inferences in VaccO using property chains 93
Table 7.3 Reference set for evaluating the code alignment algorithm 98
Table 7.4 Number of classes and terms in the VaccO ontology 99
Table 7.5 Error analysis of automatic code alignment 102

1

INTRODUCTION

Vaccines are among the most effective means for improving population health.[1–4] Smallpox, for example, which accounted for up to 500 million deaths in the twentieth century, has been eradicated thanks to a global vaccination programme.[5] The polio eradication initiative reduced the number of worldwide reported polio cases from 400,000 in 1988 to 22 in 2017.[6] And the number of measle cases worldwide decreased between 2000 and 2015 by 75% due to vaccinations, averting an estimated 20.3 million deaths.[7,8]

Besides its beneficial effects, a vaccination carries the risk of causing adverse events, as do other medical interventions. But because vaccines are administered to healthy adults and children, the weight of the benefits and risks (B/R) of a vaccine requires special consideration.[9] Prior to licensure, the vaccine B/R profile is assessed during preclinical and clinical development.[10,11] The B/R profile after marketing, however, is not entirely predictable from the preclinical and clinical assessment, where only selective populations are included, and follow-up and size of the investigated populations are limited. Therefore, the B/R profile of a vaccine may change after licensure, for example due to the emergence of rare and long-term side effects,[12] or due to differing effectiveness or risks in populations that were not covered by prelicensure clinical trials (e.g., pregnant women or children). Passive surveillance systems have been established to detect possible safety signals for medicines and vaccines after marketing based on individual case reports (e.g., EudraVigilance in Europe and VAERS in the US).[13,14]

A potential change in the B/R profile of a vaccine necessitates the reassessment of the profile based on available evidence regarding the coverage, benefits and risks of the vaccine and based on the empirical verification of cumulative life-cycle evidence in observational studies. Undetected changes in the B/R profile – but also the very success of a vaccination campaign by lowering the risk perception of the vaccine-preventable disease in the public – can give rise to public concerns towards vaccines, lowering the vaccine acceptance or even jeopardizing a vaccination programme.[15] Safety concerns about vaccines and changes in public sentiment have to be recognized early and acted on by communicating established knowledge to maintain public trust. In this context of post-licensure management of vaccines, the prompt and accurate extraction of vaccine-related information is fundamental with respect to three aims, which are addressed in this thesis: Listening to

vaccine safety issues and changes of sentiment in the public, accessing established evidence about a vaccine, and empirically verifying hypotheses about vaccine B/R in observational studies.

Vaccine-related information that is pertinent to these aims is available in various resources. Potential safety concerns and public sentiment are expressed in user-generated internet content (UGC) such as social media. Established knowledge about vaccines is recorded in scientific literature. And the empirical verification of hypotheses regarding the vaccine B/R is based on real-world evidence about vaccines in electronic health record (EHR) databases. The extraction of information from these resources would be straightforward if the information was represented homogeneously, i.e., by a unique symbol (code, word, or phrase) for each concept (vaccine or medical outcome). However, a common characteristic of the post-licensure information resources is their representational heterogeneity: the symbols used to represent equivalent information differ between – or even within – resources. In free-text resources such as UGC and scientific literature, differences in descriptions may be due to the use of different natural languages, terminologies, or levels of detail (e.g., vaccine products vs. pharmacological classes). In EHR databases, different coding systems are used to represent medical information. This representational heterogeneity encumbers the retrieval of vaccine-related information.

Automatic or semi-automatic methods promise to improve the extraction of vaccine-related information from resources where representational heterogeneity occurs. Such automatic methods instruct a computer to make sense of information independently of variance in its representation – a process that generally involves three steps: (1) recognizing the symbols that carry relevant information (e.g., specific words in free text or codes in databases), (2) representing the information independently from its symbols, and (3) interpreting the information against the background of domain-specific knowledge. We refer to this process as *vaccine semantics*. The pursued aim of post-licensure vaccine management determines the steps required for the information extraction, and the characteristics of the considered resources shape the automatic methods implementing each step. The following sections present the aims and relevant resources in more detail.

LISTENING TO SAFETY CONCERNS AND PUBLIC SENTIMENT

The World Wide Web has developed in the past decade into an unprecedented platform for forming interest communities and rapidly sharing information. Its large volume makes UGC a promising resource to monitor healthcare-related issues in the public. For example, the

tapping of social media messages covering personal experience with medical products has spawned much interest for using this information for the surveillance of diseases and drug safety.[16–23] Blogs and MySpace discussions have been evaluated for safety surveillance of vaccines,[24,25] and UGC and public news were proposed for monitoring public sentiment about vaccines and vaccinations programmes.[26–28] It is unknown, however, if or how an analysis of social media messages can contribute to the surveillance of vaccine safety or to the monitoring of public sentiment towards vaccines.

The use of social media messages for monitoring vaccine safety and sentiment is challenging because social media represent secondary data, i.e., data that are not originally intended for monitoring. Messages are authored in different languages and vaccine descriptions differ by idiom, terminology, and syntactic variations. To date, there are no standard methodologies for mining social media content for vaccine safety surveillance or monitoring of public sentiment. The field of natural-language processing (NLP), however, provides general methods for analysing natural language data and extracting information (see explanation box 1.1).

ACCESSING ESTABLISHED EVIDENCE ABOUT VACCINE B/R

Previously established evidence in the form of scientific literature constitutes an important component in the B/R assessment of vaccines. The amount of available scientific literature about human vaccines grows rapidly: more than 3,000 publications about human vaccines are made available on PubMed every year.[29] The sheer amount of articles hampers the manual screening of literature to retrieve study results about a specific vaccine or a class of vaccines. Automatic retrieval of specific vaccine information could help accelerating the post-marketing assessment of vaccines and communicating established knowledge to prevent or handle crises of public confidence in vaccination programs.

The retrieval of evidence about a vaccine from scientific literature involves three tasks: (1) identifying articles that are relevant for a given vaccine or vaccine class, (2) recognizing the vaccines in the text by their descriptions, and (3) extracting relational information about the vaccine, e.g., stated adverse events. However, the automatic retrieval of vaccine-related information from scientific literature is challenging due to the large syntactic variability in vaccine descriptions and their semantic relations that is common in research articles.

Literature databases provide a classification of published articles by indexing them with codes from a controlled vocabulary, such as the Medical Subject Headings (MeSH) for the PubMed literature

Explanation box 1.1: Natural-language processing

Natural-language processing (NLP) is a field of computer science and artificial intelligence, where automatic methods are developed for processing texts written in natural languages (such as English, in contrast to formal languages that target computers). Potential applications of NLP are information retrieval (the identification of relevant documents in a corpus) and information extraction (the extraction of structured information from a given document). NLP is challenging because the flexibility of natural language and the ubiquity of domain-specific and general background knowledge in human understanding opposes the deterministic execution of computer programs.
NLP methods typically process the input text in a pipeline involving several steps (and required steps depend on the task):

1. Sentence splitting: Identify the boundaries between sentences in the input text to limit each subsequent step to one sentence.
2. Tokenization: Split each sentence in a sequence of words or punctuation.
3. Lemmatization: Map words to their canonical form (lexeme) by removing morphological variations, e.g., 'goes' $\rightarrow$ 'go'.
4. Named-entity recognition (NER): Assign words or sequences of words to entity types, e.g., 'Aspirin' $\rightarrow$ *Drug*.
5. Parsing: Syntactically analyse a sentence to generate a parse tree over its words.
6. Normalization: Assign named entities to identifiers from a database, ontology, or knowledge graph to connect them with contextual information.
7. Semantic analysis: Extract semantic relations between normalized entities that are described in the text (e.g., 'Aspirin' treats 'pain').

database.[30,31] The classification constitutes an indispensable tool for quickly identifying relevant literature. However, indexing is a manual process and does not cover all recent publications.[32] Also, articles are indexed on a document level and the location of the text that led to the indexing is not retained. This text location is a critical starting point for the automatic extraction of relational information.

The identification of drugs in scientific literature has been the subject of extensive research, but little attention has been given so far to the identification of vaccines,[33] which differs from and is more difficult than drug identification due to the large syntactic variability in vaccine descriptions. Whereas a drug tends to be referred to by its (product or generic) name or by the name of its active ingredient, a vaccine is typically characterized by its properties, for example its immunization

targets and immunization strategy (e.g., 'monovalent conjugated vaccine against *Haemophilus influenzae* type B'). Available reference corpora for medicines contain only few vaccine mentions,[34] which prevents their use in training or evaluating automatic methods for vaccines.

Relational information about vaccines in scientific literature includes statements about the vaccine benefits (relating a vaccine with its vaccine-preventable disease) and risks (relating a vaccine with a potential adverse event). The manual relation extraction from scientific articles and their storage in structured databases is cumbersome and expensive, and existing inventories are fragmental.[35] Automation of the extraction promises to solve these problems. Previous research on relation extraction from scientific literature, however, largely focused on finding interactions between genes, proteins, and drugs.[36–44] General, automatic recognition of vaccine descriptions in scientific literature has not been attempted previously to the best of our knowledge. Attempts to extract relations between chemicals, including drugs and vaccines, and diseases have met with limited success, mostly due to the lack of a large-scale training corpus.[45]

EMPIRICAL VERIFICATION OF B/R HYPOTHESES IN OBSERVATIONAL STUDIES

After the approval of a vaccine, hypotheses about the vaccine B/R are tested by conducting observational studies in EHR databases. These studies generally aim at quantifying the effect of a vaccine exposure on a medical outcome, i.e., on the vaccine-preventable disease in effectiveness studies, or on an adverse event in safety studies. The study protocol describes the medical outcome in a case definition and the exposure by product names or pharmacological classes. To extract an exposure or a medical outcome from an EHR database, its description has to be translated into a database query comprised of pertinent codes from the database coding system (see explanation box 1.2).

To increase their statistic power, observational studies can be performed in a collaborative fashion by combining information from multiple EHR databases.[46] Medical outcomes, however, are recorded using different medical coding systems in European EHR databases (figure 1.1),[47] and vaccines are represented by product names and pharmacological classes using medical coding systems, drug coding systems, or database-specific custom coding systems.[48] The definition of the database-specific extraction queries and their harmonization, which is required to guarantee a consistent extraction of vaccinations and outcomes between databases, constitutes an important bottleneck in the conduction of collaborative observational studies in Europe.[46,49,50]

Explanation box 1.2: Coding of medical information

Medical information in electronic health record (EHR) databases is represented using codes from controlled medical coding systems (i.e., vocabularies). Different medical coding systems are used in European EHR databases. For example, the disease pneumonia is represented in by code R81 from the ICPC-2 coding system the Dutch IPCI database, by codes 480, 481, 482.2, 482.3, 482.9, 483, 485, 486, and 487.0 from ICD-9 CM in the Italian Lombardy Regional database, and by codes J12 up to J18 from ICD-10 CM in the Danish Aarhus database. The meaning of each code is defined by a short, textual description in the coding system. Standardized medical coding systems use a taxonomic hierarchy to subordinate more specific codes to more general codes. Some EHR databases use additionally free text and database-specific custom coding systems to record information that is not covered by the medical coding system of the database.

Two fundamental approaches to harmonization exist: broadening the definition given in the study protocol, or unifying the database codes (figure 1.2). A common ad hoc broadening approach is the manual mapping of the textual case definition to an individual extraction query for each database, based on an iterative process directed by earlier extraction results, results from the literature, and expert discussion. This mapping approach, however, requires great manual effort and does not reinforce consistency between queries. It was refined in the EU-ADR project[49,51] by using the Unified Medical Language System (UMLS),[52] a compendium of numerous medical coding systems including those commonly used to record medical information in EHR databases (explanation box 1.3). Diseases, symptoms, laboratory procedures, or tests were automatically identified in the case definition and represented by abstract concepts (i.e., concept unique identifiers (CUIs) from the UMLS). The list of concepts was manually revised in an iterative process. Lastly, the concepts were automatically projected into corresponding code sets from the targeted coding systems using the assignments between concepts and codes in the UMLS. Whereas the identification of concepts and their projection to codes was automatic, the overall workflow was not integrated and the development process was difficult to document, which hampered the reuse of the queries in subsequent studies.

In the unification approach, the codes used in the databases are mapped to one reference coding system that is used to define the outcome or exposure in the study protocol. Unification is suitable also for resolving heterogeneity in the representation of vaccines, which

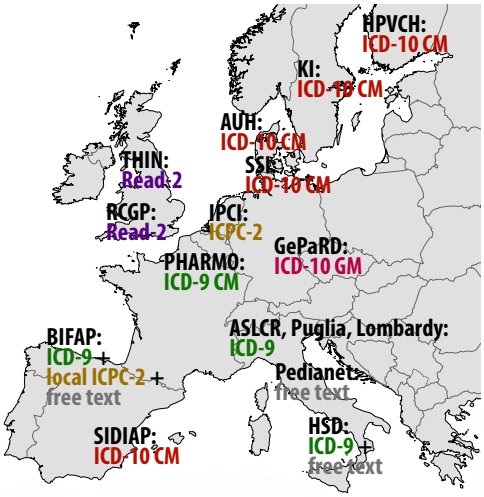


Figure 1.1: Examples for heterogeneous representation of medical outcomes in European EHR databases

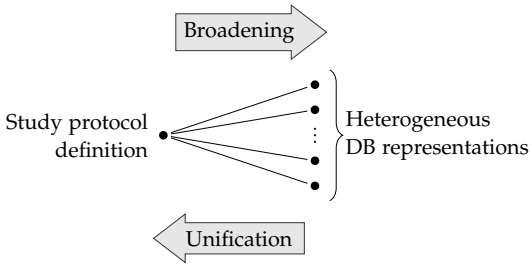
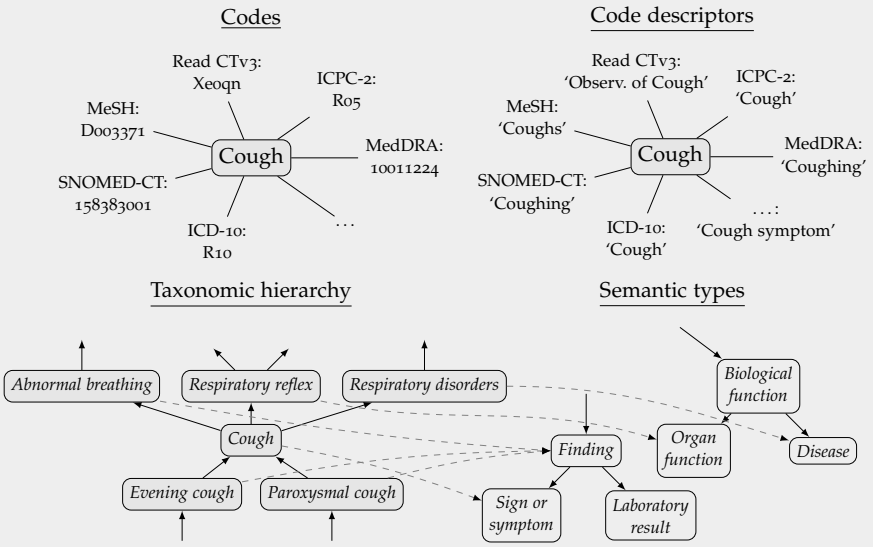


Figure 1.2: Approaches to harmonizing extraction queries for databases with representational heterogeneity

are often recorded using custom coding systems that lack mappings to other coding systems. The automatic unification of codes can be based on an analysis of the different components of the coding systems, such as the code descriptors, the taxonomic hierarchy, and information about the instances (e.g., the vaccine products that belong to the pharmacological class represented by a code).[54] Custom vaccine coding systems, however, usually lack taxonomic hierarchies and information about instances.[48] The only information about vaccine codes that is generally available are their descriptors, which use different languages, different terminologies, different levels of description (i.e., products and pharmacological classes), and different properties for describing equivalent vaccines (e.g., by vaccine-preventable diseases as 'tuberculosis vaccine' or by active ingredient as 'BCG'). Domain knowledge

Explanation box 1.3: The Unified Medical Language System

The Unified Medical Language System (UMLS) is a compendium of medical coding systems, which have been integrated by assigning codes from different coding systems but with a common meaning to one concept unique identifier (CUI). The UMLS contains more than 3.6 million CUIs that connect 14 million codes from 154 coding systems (in version 2018AA).[53] Code descriptors and hierarchies from the coding systems are preserved in the UMLS. Each CUI is further assigned to one or more of 127 semantic types, which define broad conceptual categories such *Disease* or *Substance*. The figure below illustrates the information in the UMLS related to the CUI C0010200, which represents the meaning of 'cough'.

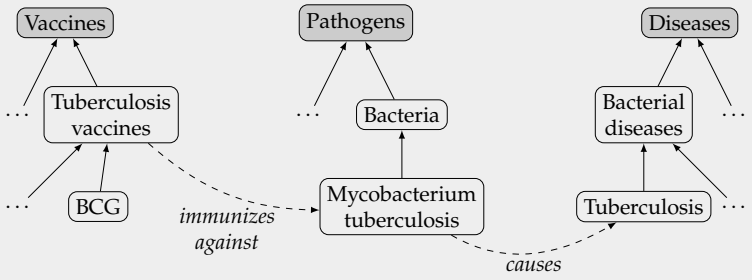


about vaccines is required to resolve these representational differences in the unification of the vaccine codes. A common way to make domain knowledge available to automatic processes is its formalization in an ontology (explanation box 1.4). However, existing vaccine ontologies focus on vaccine products and their immunological properties and are not suited to interpret vaccine descriptions.[55]

Explanation box 1.4: Ontologies

In computer science, an ontology is a formal definition of the entities in a domain including their relations ('the explicit specification of a conceptualization').[56,57] The entities in an ontology comprise (1) ground level objects (e.g., vaccine products or packages), (2) properties that describe a relation between two entities, and (3) classes that group other entities by defining common characteristics (e.g., the class for influenza vaccines contains any vaccine that immunizes against influenza) and that are arranged in a taxonomic hierarchy (e.g., all influenza vaccines are also viral vaccines). Different conceptualizations of a domain are valid and an ontology always represents a specific point-of-view on a domain. The de facto standard for describing ontologies is the Web Ontology Language (OWL2).

The figure below exemplifies an ontology of vaccines, pathogens, and diseases (solid arrows capture the taxonomic hierarchy, dashed arrows indicate relations with their properties). The class *BCG* is defined as a subclass of *Tuberculosis vaccine*s, which in turn is defined as the set of vaccines that immunize against *Mycobacterium tuberculosis*. The relation is specified by the property *immunizes against*. The ontology also formalizes the domain knowledge that *Mycobacterium tuberculosis* is the causal agent of *Tuberculosis*. This information can be applied to derive the fact that *BCG* is a vaccine that protects against *Tuberculosis*.



OUTLINE

This thesis explores automatic methods for solving representational heterogeneity of vaccine-related information to facilitate post-marketing benefit and risk assessment of vaccines (figure 1.3). Part I focuses on public social media. Their use for the surveillance of vaccine safety is evaluated in chapter 2, and for understanding the dynamics of the public sentiment towards a vaccine in chapter 3. Both chapters present basic methods for the identification and retrieval of relevant information (vaccines, medical outcomes, and locations) from public social media messages using different languages and terminologies. Part II covers the retrieval of established knowledge from scientific literature. Chapter 4 compares different methods for recognizing vaccine descriptions in scientific articles and for classifying articles by vaccines. Chapter 5 presents RELigator, a system that extracts causal relations between chemicals and diseases from scientific articles, which could eventually be specialized in the extraction of vaccine adverse events by combining it with the automatic recognition of vaccine descriptions. Part III deals with retrieval of vaccines and outcomes from electronic health records. Chapter 6 presents CodeMapper, a comprehensive web application that helps in broadening clinical definitions of medical outcomes to database queries. And chapter 7 closes with a novel approach to unify vaccine coding systems based on the VaccO ontology, which was created for the purpose of representing and reasoning about vaccine descriptions.

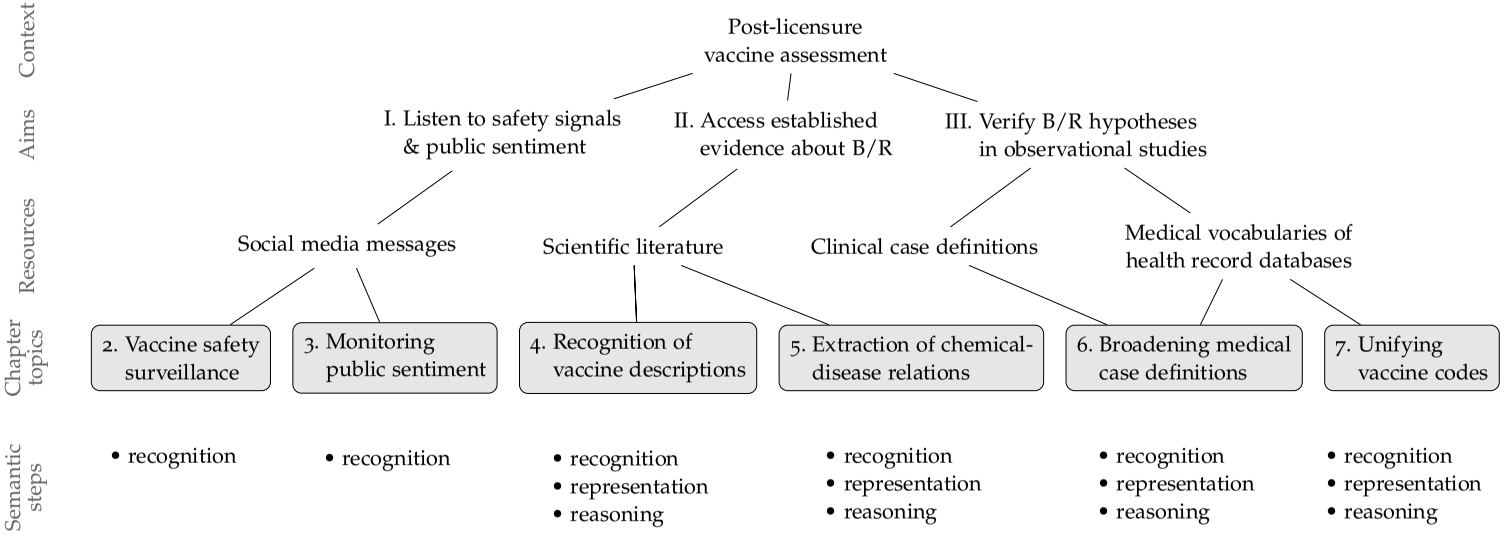


Figure 1.3: Aims and resources in the post-marketing management of vaccines, topics of the chapters in this thesis with required semantic steps

Part I

LISTENING TO VACCINE SAFETY CONCERNS AND PUBLIC SENTIMENT

2

SOCIAL MEDIA FOR VACCINE SAFETY SURVEILLANCE

ABSTRACT

OBJECTIVE To evaluate potential contribution of mining social media networks for medicines safety surveillance using the following associations as case studies: (1) rosiglitazone and cardiovascular events (i.e., stroke and myocardial infarction); and (2) human papillomavirus (HPV) vaccine and infertility.

METHODS We collected publicly accessible, English-language posts on Facebook, Google+, and Twitter until September 2014. Data were queried for co-occurrence of key words related to the drug/vaccine and event of interest within a post. Messages were analysed with respect to geographical distribution, context, linking to other web content, and author's assertion regarding the supposed association.

RESULTS A total of 2,537 posts related to rosiglitazone/cardiovascular events and 2,135 posts related to HPV vaccine/infertility were retrieved, with the majority of posts representing data from Twitter (98% and 87%, respectively) and originating from users in the US. Almost 25% of rosiglitazone-related posts and 75% of HPV vaccine-related posts referenced other web pages, mostly news items, law firms' websites, or blogs. Assertion analysis showed predominantly affirmation of the association rosiglitazone/cardiovascular events (72%, N=1,821) and of HPV vaccine/infertility (82%, N=1,753). There were only 10 posts describing personal accounts of rosiglitazone/cardiovascular adverse event experiences and 9 posts describing HPV vaccine problems related to infertility.

CONCLUSIONS Publicly available data from the considered social media networks were sparse and largely untrackable for the purpose of providing early clues of safety concerns regarding the prespecified case studies. Further research investigating other case studies and exploring other social media platforms are necessary to further characterize the usefulness of social media for safety surveillance.

Coloma PM, Becker BFH, Sturkenboom MCJM, van Mulligen EM, Kors JA. Evaluating Social Media Networks in Medicines Safety Surveillance: Two Case Studies. *Drug Saf* **38** (2015)

2.1 INTRODUCTION

The past decade has brought forth enormous growth and popularity of online communities and social networks, greatly expediting information exchange from one corner of the world to another. The concept of blogging has allowed virtually anybody with internet access to post his or her views and experiences on any topic at any time. Whilst the value of such online conversations has been exploited mostly by commercial enterprises to promote product improvement and innovation, healthcare has not been immune to this phenomenon of public engagement.[59–61] In the same spirit of eliciting greater patient participation, several investigators have begun to explore what social media can offer in terms of medicines safety surveillance.[19,22,62] Reporting of individual cases of suspected adverse drug reactions (ADRs) to regulatory authorities, mostly by physicians or other healthcare professionals, remains the cornerstone of pharmacovigilance. However, spontaneous reporting systems are hampered by various limitations, the most important of which is underreporting.[63,64]

Because social media represent secondary data, i.e., data that are not originally intended for surveillance, there are challenges to overcome with respect to terminology, traceability, and reproducibility. Apart from these technical challenges, practical policy guidelines are lacking on how potential safety signals from social media should be handled in the current regulatory framework. Although the US Food and Drug Administration (FDA) has released two guidance documents on the use of social media platforms for presenting benefit/risk information on prescription drugs and medical devices,[65] these documents are more concerned with product promotion than surveillance and 'do not establish legally enforceable rights or responsibilities'.[66] The European Medicines Agency (EMA)'s guideline on good pharmacovigilance practices (Module VI) provides provisions on how to deal with information on suspected adverse reactions from the internet or digital media and hold market authorization holder (MAH) responsible for reviewing web sites under their control for valid cases and reporting them accordingly, although there is no requirement to trawl internet sites not under the MAH's control.[67] To date there are no standard methodologies to mine user-generated data from social media for pharmacovigilance. In this study we sought to evaluate the potential contribution of mining social media networks for pharmacovigilance using examples of drug-event associations that have been flagged as potential signals: rosiglitazone and cardiovascular events (i.e., stroke and myocardial infarction); and human papillomavirus (HPV) vaccine and infertility.

2.2 METHODS

Postings were collected from three of the most widely used social media networking platforms (Facebook, Google+, and Twitter) using their respective search application programming interface (API). The search APIs return a set of public messages from the social network that match the query keywords. For each message the content is provided together with additional information about the message itself (date and content), about the status in a conversation (repost or reply to another message), and about author (user name and location). Messages were obtained from as far back as available until 25 September 2014. Only English-language posts were considered. Facebook provides only messages from the preceding month by their search API. The search API of Google+ obtains messages dating back to its establishment in 2011. The search API of Twitter is restricted to a time window of about one week. In order to supplement the Twitter data obtained via its search API, an additional search engine, Topsy was used.[68] Topsy is a real-time search engine for posts and shared content on social media, primarily on Twitter and Google+. As of this writing, Topsy had complete coverage of historical messages and has indexed every (public) tweet ever posted since 2006. For this particular study, only Twitter-related posts were retrieved via the free analytics service of Topsy.

2.2.1 *Case studies*

Usefulness of the above social media platforms for safety surveillance was evaluated using two examples of drug-adverse event associations that have previously been flagged as potential safety signals: (1) rosiglitazone and cardiovascular events (i.e., stroke and myocardial infarction); and (2) HPV vaccine and infertility. These two case studies were chosen because they represent associations that have triggered controversies and thus are likely to have been the subject of media attention as well as online discussions. Furthermore, the case studies involve different types of agents that are used by different subsets of the population under different circumstances, thus allowing investigation of diverse scenarios.

For each case study, data were queried for co-occurrence of the drug/vaccine of interest and the event of interest within the same post or tweet. Search queries were constructed using all possible drug-event keyword combinations. Event-related keywords consisted of clinical terms from the Unified Medical Language System (UMLS) as well as known abbreviations and layman's terms (search queries and event keywords are available as online supplementary material[69]). Drug-

related keywords consisted of international non-proprietary names and trade names.

2.2.2 *Assessment of suitability for use in safety surveillance*

Relevant posts were tallied (reposts/retweets excluded) and analysed with respect to geographical distribution, context, and linking to other web content. The country of origin of a message was automatically determined from the location information about the author. When the country was not available in a designated data field, it was automatically identified from the available location information by means of a list of names of countries, regions and cities. The frequency of message propagation (i.e., reposts or retweets) was calculated. The content of all posts were reviewed one by one to determine whether there was reference to a person's actual experience of having the (adverse) event of interest in relation to exposure to the drug (or vaccine) of interest. It was not the intention to assign or assess causality, but rather to describe the context of how the drug-event relationship is described. Posts were likewise analysed with respect to the author's assertion of the purported association between the drug (or vaccine) of interest and the event of interest. Somewhat analogous to sentiment analysis, assertion was judged as one of the following: (1) affirmative, if the post alluded to an affirmation of the association; (2) negating, if the post alluded to a negation of the association; or (3) neutral, if the post alluded to neither affirmation nor negation of the association. Manual review and annotation of the assertions was done by a physician/pharmacist (PMC). In addition, key dates during which important communication or regulatory actions occurred were marked and compared with the timeline of the posts.

2.3 RESULTS

2.3.1 *Rosiglitazone and cardiovascular events*

As shown in table 2.1, we retrieved a total of 2,537 posts related to rosiglitazone and cardiovascular events (i.e., stroke and myocardial infarction), with the overwhelming majority of posts (98%) representing data from Twitter. There were only two posts on Facebook, while there were 41 posts retrieved on Google+. About 10% of all posts were reposts or retweets. The country of origin (based on the holder of the social network account) could not be automatically identified in 59% of the posts; of the posts that could be identified, two-thirds was accounted for by the United States (US) while the remaining one-third was distributed

Table 2.1: Overview of posts about rosiglitazone and cardiovascular adverse events across social media networking platforms

Platform	Posts	Reposts	Links to other sites	Date range	Origin of post* (Count)
Facebook	2 (0.1%)	0	2 (100%)	07/2014–08/2014	US (1) Unknown (1)
Google+	41 (1.6%)	6 (15%)	41 (100%)	06/2012–08/2014	Unknown (31) US (9) Egypt (1)
Twitter	2,494 (98.3%)	250 (10%)	493 (20%)	05/2007–09/2014	Unknown (1,461) US (682) India (53) UK, Canada (50 each) Indonesia (31) Other (167)
Total	2,537	256 (10%)	536 (21%)		

* Based on account holder. Where applicable, only the top five countries are given.

Table 2.2: Description of web pages referenced by posts about rosiglitazone and cardiovascular events

Category of linked web pages	Facebook (N=2)	Google+ (N=41)	Twitter (N=493)	Total (N=536)
News	-	8 (20%)	188 (38%)	196 (37%)
Law firm's website or advertisement	1 (50%)	17 (41%)	139 (28%)	157 (29%)
Blog	-	13 (32%)	125 (25%)	138 (26%)
Health reference for professionals	-	2 (5%)	22 (5%)	24 (5%)
Patient community website	-	-	2 (1%)	4 (<1%)
Health education for patients	1 (50%)	-	-	1 (<1%)
Scientific journal	-	-	15 (3%)	15 (3%)
Video	-	1 (2%)	-	1 (<1%)

among 50 other countries or territories all over the world. Overall, 21% of posts (N=536) had links to other web pages (table 2.2). News items comprised more than one-third of the web pages referenced (N=196), followed by law firms' websites or advertisements (N=157) and blogs (N=138). There were 24 posts referring to health information websites intended for health professionals, 15 posts linking to scientific journals, four posts referring to a patient community website, one post linking to a hospital's patient education website and another to a YouTube video.

Assertion analysis done on all posts demonstrated predominantly affirmation of the association between rosiglitazone and cardiovascular events (72%, N=1,821), with the remainder more or less split between negating (13%) and neutral (15%). Most neutral posts were asking for further information or otherwise not directly related to the drug-adverse event association. There were posts by lawyers or reporters explicitly soliciting cases (N=12), but there were also posts (N=122) ridiculing lawyers' television commercials that asked patients who 'died while taking the drug' to call a particular number.

Figure 2.1 shows the trend of assertions over time in relation to events in the timeline of the association of interest. The highest peak of affirmative posts occurred in February 2010. In this particular month, the US Senate Finance Committee released a report based on a two-year inquiry of rosiglitazone, expressing concern that 'FDA has overlooked or overridden safety concerns cited by its own officials'.[70] The EMA's suspension of rosiglitazone's marketing authorization in the

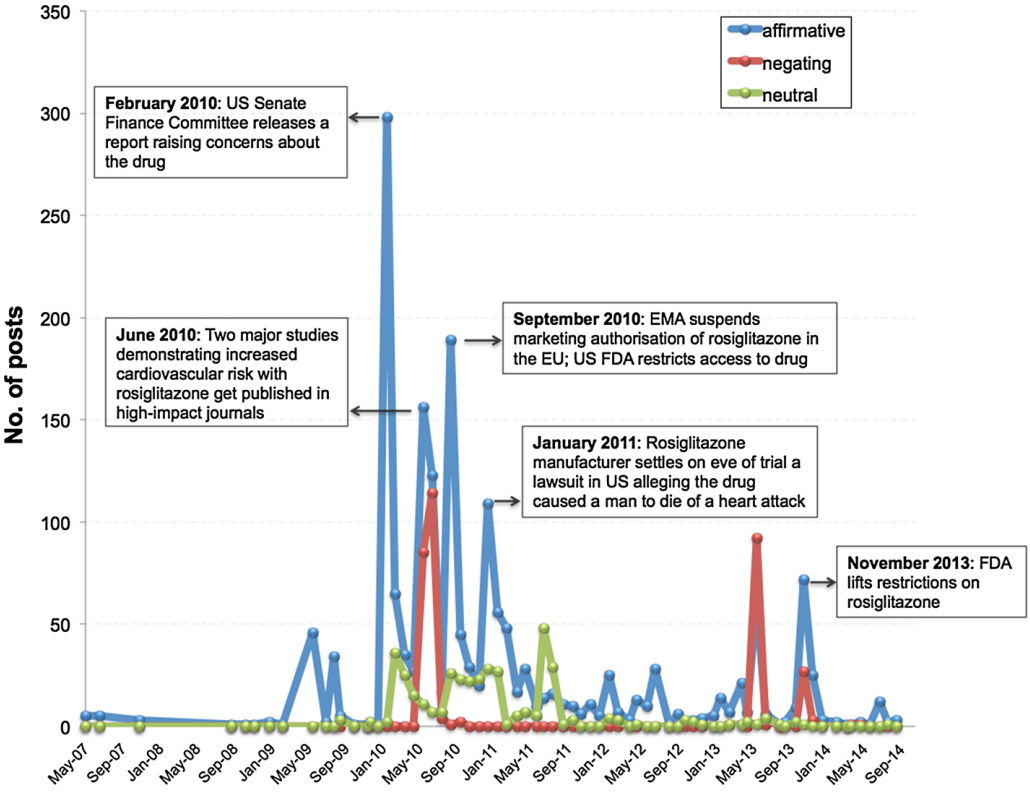


Figure 2.1: Trend of assertions of rosiglitazone/cardiovascular event-related posts over time.

European Union (EU) and the FDA restriction of access to the drug coincided with the second peak of affirmative posts in September 2010, while the simultaneous publication in high-impact journals of two studies demonstrating increased cardiovascular risk with use of rosiglitazone[71,72] coincided with the peak in June 2010. The peaks in negating assertions paralleled those of the affirmative, with the greatest peak in affirmations observed in June-July 2010 (and a smaller peak in November 2013), reflecting the active online debate that was happening regarding the issue. Figure 2.1 also shows that in June 2013 negating posts actually outnumbered the affirmative posts; the results of the FDA-mandated re-evaluation of the rosiglitazone (RECORD) trial became available online in June 2013.[73] The peak of neutral posts seen in July 2011 represented posts about news of rosiglitazone being potentially useful for neuropathic pain (although the pertinent study was already published online three months earlier[74]).

There were only 10 posts that appeared to be about experiences of the drug-adverse event association of interest. Four posts involved

the person posting the message himself or herself (one even claimed winning a legal case against the drug manufacturer); three involved somebody's brother-in-law; while there was one each for somebody's father, father-in-law, and grandmother. In addition, there were two posts referencing a patient community website that claimed 21,015 people reported to have a heart attack while taking rosiglitazone (representing '32% of all who reported side effects'). Interestingly, some posts (N=20) alleged other adverse events of rosiglitazone such as leg pain, abdominal pain and eye pain (all of which are symptoms suggestive of end-organ complications of diabetes, the primary indication for the drug), while others (N=67) alluded to a beneficial effect of the drug (prevention of neuropathic pain).

2.3.2 *HPV vaccine and infertility*

We retrieved a total of 2,135 posts related to HPV vaccine and infertility, again with the majority of posts (87%) representing data from Twitter (table 2.3). There were 23 posts on Facebook while there were 256 posts retrieved on Google+. Reposts or retweets comprised 22% of all posts. Similar to posts related to the previous case study on rosiglitazone, the country of origin was unknown for more than half of the HPV vaccine-related posts, with the US representing majority (N=519) of those posts that could be automatically identified. In contrast to the rosiglitazone-related posts, however, a large proportion of all posts (84%) referenced other web pages (table 2.4). Various blogs comprised almost half of the linked web pages referenced (N=847), followed by news items (N=650) and scientific journals (N=118). Most of the blogs commented on these same news items or journal articles. There were 109 posts referring to health information websites intended for health professionals, 49 posts linking to (mostly anti-vaccine) YouTube videos, while only a minority of posts were associated with lawyer's websites or advertisements (N=24).

The posts demonstrated predominantly affirmative assertion of the association between HPV vaccine and infertility (82%, N=1,753), with posts that negate the association accounting for 4% (N=81) and neutral posts accounting for the rest. Most neutral posts were asking for further information or were negative comments about the HPV vaccine in general but not directly related to infertility. Figure 2.2 shows the trend of assertions over time in relation to events in the timeline of the association of interest. The highest peak of affirmative posts occurred in November 2013 when two sisters, aged 20 and 19, alleged at a US federal court that Gardasil (trade name of the HPV vaccine) caused them to go into early menopause and become infertile. The build-up to

Table 2.3: Overview of posts about HPV vaccine and infertility across social media networking platforms

Platform	Posts	Reposts	Links to other sites	Date range	Origin of post* (Count)
Facebook	23 (1%)	6 (26%)	15 (65%)	04/2014–09/2014	Unknown (19) Bangladesh, India, The Philippines, United States (1 each)
Google+	256 (12%)	42 (16%)	249 (97%)	09/2011–09/2014	Unknown (178) United States (41) Australia, India (6 each) Canada (5) Spain, France (2 each) Other countries (16)
Twitter	1,856 (87%)	432 (23%)	1,538 (83%)	07/2008–09/2014	Unknown (1,039) United States (477) Canada (112) Australia (40) United Kingdom (37) Italy, Egypt (10 each) Other countries (131)
Total	2,135	480 (22%)	1,802 (84%)		

* Based on account holder. Where applicable, only the top five countries are given.

Table 2.4: Description of web pages referenced by posts about HPV vaccine and infertility

Category of linked web pages	Facebook (N=15)	Google+ (N=249)	Twitter (N=1,538)	Total (N=1,802)
News	4 (27%)	121 (49%)	525 (34%)	650 (36%)
Law firm's website or advertisement	-	3 (1%)	21 (1%)	24 (1%)
Blog	5 (33%)	100 (40%)	742 (48%)	847 (47%)
Health reference for professionals	-	8 (3%)	101 (7%)	109 (6%)
Scientific journal	-	1 (<1%)	117 (8%)	118 (6%)
Video	1 (7%)	16 (6%)	32 (2%)	49 (3%)
Multiple sites	5 (33%)	-	-	5 (<1%)

this peak appears to have been triggered by a study describing three young women who presented with secondary amenorrhea following HPV vaccination;[75] this study was first published online at the end of July 2013 (corresponding to the earlier, but smaller, peak in figure 2.2). Many of the posts within the period from August to October 2013 actually referred to an event that happened one year before: the publication of the first case report on the association of interest. This case report of a 16-year-old Australian girl who had premature ovarian failure after HPV vaccination was first published online in October 2012.[76]

There were nine posts that appeared to be accounts of HPV vaccine-adverse event experience. Six posts involved the person posting the message herself. One simply said she was '15 and infertile' because of the vaccine (the actual page appears to have been taken down after the initial data collection), while four other individuals claimed to have an ovarian cyst, delayed period (and negative pregnancy test), (vaginal) spotting, menopause and hot flashes because of the vaccine. One post was about somebody's friend who was '21 and infertile due to the HPV vaccine' and there were two posts from different mothers whose daughters had no (menstrual) periods after getting the vaccine.

2.4 DISCUSSION

In this study we aimed to characterize the data currently available from social media networking platforms and to determine if – and how – such data can be tapped for surveillance of two specific safety issues: rosiglitazone and cardiovascular events (i.e., stroke and myocardial infarction); and HPV vaccine and infertility. Rosiglitazone is a drug

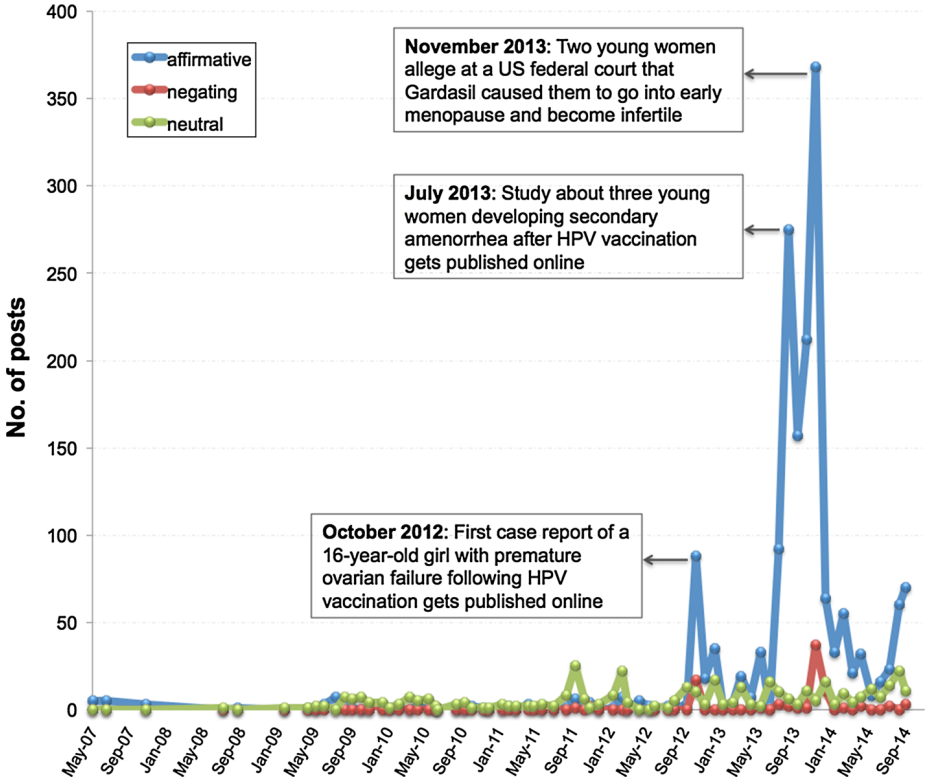


Figure 2.2: Trend of assertions of HPV vaccine/infertility-related posts over time.

indicated for a very prevalent disease, diabetes, and although such a disease is expected to occur in the middle-aged population – who comprise a relative minority of the population of Twitter users, it was precisely one of the aims of this study to illustrate that such a group and such condition of interest could be under-represented in social media networks, however huge these networks may be. The primary motivation for exploring social media as an additional resource for pharmacovigilance is to capture information that cannot be found in traditional sources. Among the three websites evaluated, Twitter provided the greatest number of (publicly available) posts potentially relevant to the two case studies but these represented mostly links to news items or, particularly for rosiglitazone and cardiovascular events, websites of personal injury lawyers rather than accounts of drug/vaccine-related adverse events. The ubiquity and instantaneous nature of the internet and social media networks supposedly provides a mechanism to find adverse drug (or vaccine, or medical device) experiences of laymen that are otherwise missed by ADR reporting

systems – and in real time. Thus, one of the more relevant questions to ask is whether data from social media networks can provide early signs of potential safety concerns. Despite the hype about social media representing 'big data,' the volume of relevant posts was sparse for the two case studies considered. Although Twitter has over 500 million users (more than half of whom are reportedly active), it was too 'young' a source to use, particularly for the case study on rosiglitazone. When FDA issued the safety alert on Avandia in May 2007 Twitter had only been in service for less than a year, was largely in its trial phase and thus still had few subscribers. The same argument can be said for Facebook, which became available in September 2006 and Google+, which was launched much later in September 2011. The problem that these social media sites did not have enough time to accumulate data should have been less of an issue for the HPV vaccine-infertility association, which is a more recent potential safety concern, and yet that does not seem to be the case.

Our findings corroborate what other researchers have shown regarding the geographic distribution of users of social media networks: a small number of countries, led by the US, account for a large share of the total user population and likewise make up the active and influential user population.[77,78] Although this is not totally unexpected, given that only English-language posts were obtained in this study, there can be implications on inferences drawn from research using data from social media networks.

There were (only) 10 and 9 accounts of adverse experiences related to rosiglitazone/cardiovascular events and HPV vaccine/infertility, respectively, but these experiences appeared to be more reactionary than anticipatory (meaning they were shared online after news about the safety issues broke out). Furthermore, verification of such allegations proved to be difficult considering the data privacy constraints (only publicly accessible data could be analysed) and in particular, establishing an identifiable patient and 'reporter' (required for valid safety reporting in traditional pharmacovigilance systems) is challenging, if not impossible. The scenario of unprincipled individuals spreading inaccurate – and even false - information is not unheard of and since social media is largely unregulated, cannot be avoided.[79] Interestingly, two posts identified in the current study referenced a health information and community website,[80] which claims to have studied (as of the time of writing this article) '65,460 people who had side effects while taking Avandia from FDA and social media,' and among them 21,015 had a 'heart attack'. In addition, there are 7,752 who had 'stroke'. The website provides statistics on when the heart attack/stroke was reported, age and gender of people who have heart attack/stroke when taking Avandia, 'time on Avandia when people have heart at-

tack/stroke,' 'severity of heart attack/stroke when taking Avandia,' 'top conditions involved for these people,' and 'top co-used drugs for these people.' All such information – if truthful – are relevant. However, nowhere is it stated which part of the information comes from social media and specifically from which social media (there are too many of them). More importantly, there is no description of how these reports were obtained, the actual configuration and content of the reports could not be traced, and the circumstances surrounding the alleged adverse events could not be verified. While the site does include a general disclaimer and a counsel to 'report adverse side effects to the FDA,' these sections are found at the end of the page and may be easily ignored.

White et al. utilized retrospective web search logs to make the case for internet users providing early clues about adverse drug events via their online information seeking.[18] Chary et al. have proposed tools for using data from social networks to characterize patterns of (recreational) drug abuse,[81] while Harpaz et al. have provided an extensive review on how state-of-the-art text mining for adverse drug events can leverage unstructured data sources, including social media.[82] Similar to the current study, Freifeld et al. used publicly available data from Twitter to obtain messages that resembled adverse event reports ('proto-AEs') related to 23 prespecified medical products.[19] Rather than focusing on a few specific events of interest, the Freifeld study collected all potential events (symptoms), thus resulting in more permutations of search terms, which explains why their study had a higher yield of relevant posts compared to our study. While our current study was more of a 'scoping' study across three social media networking platforms for two specific case studies, the study by Freifeld et al. had a different aim: to evaluate concordance between Twitter posts mentioning AE-like reactions and spontaneous reports received by the FDA Adverse Event Reporting System (FEARS). There is the implicit assumption of equivalent level of information between the two sources, which, among other things, necessitated the development of a dictionary to map internet vernacular to the standardized ontology Medical Dictionary for Regulatory Activities (MedDRA). Other researchers have explored the utility of more specific health-oriented websites and patient community forums to identify adverse drug events[83] and to better understand the impact of ADRs.[84] These types of social media sources are likely to provide more relevant content because their very nature allows for sharing of health-related concerns among patients with similar conditions ('like me') and would make verification easier since user registration is often mandatory and more exhaustive (the likelihood of faking an illness in this group is probably lower). Personal accounts of adverse events from such sources are often inaccessible to the public, although many of the prominent and moderated patient community websites

will allow access to further information under certain conditions of use (and sometimes for a fee). These more health-oriented social media platforms are certainly worth exploring, especially for surveillance of uncommon adverse events as well as those related to drugs indicated for rare conditions.

The potential value of mining data from social networks appears to be greatest for measuring awareness regarding potential safety concerns. Because this study focused only on English-language posts, there is the caveat that the findings are biased towards users from English-speaking countries, particularly the US, that comprise the majority of subscribers of these social networking sites. Both number of posts and assertion trend in the two case studies were predominantly driven by events that occurred in the US. Another caveat is that bad news is often more popular than good news. The case report of the 16-year-old girl from Australia who had premature ovarian failure after HPV vaccination fired up huge comments online while four studies (published earlier or around the same time)[85–88] that showed no evidence of increased risk for new adverse events, including those related to fertility, were practically ignored.

The other, perhaps even more relevant, question to ask is whether data from social media networks can be used to help corroborate, or refute, potential safety concerns by providing information where there is none. It is time to turn the impressionability of social media as an advantage and leverage it towards bringing balanced and evidence-based information to the internet and its multitude of users.

2.4.1 *Limitations*

Data were queried for co-occurrence of the drug/vaccine of interest and the event of interest within the same post or tweet, which may have limited the number of relevant posts obtained. Similarly, the use of publicly available data and English-language only posts may have contributed to sampling bias. The assertion analysis conducted may not always reflect the true opinion of the user, the very nature of social media promoting an open and unrestricted environment. A generalization cannot be made as to which among the social networking platforms provides the most valuable information since the amount and nature of commentaries generated and shared within each network is a function of its own culture and privacy restrictions. Moreover, the population of users of social networking sites comprises the relatively young (and healthy) and fairly educated who have access to internet.[89–91] The evaluation done was retrospective and the findings for these particular case studies considered may not necessarily reflect discussions about

safety concerns related to other drugs or other vaccines in the future. Because social media platforms are continually being re-engineered to improve the commercial service, there is the concern as to whether studies conducted on data collected from these platforms are reproducible, even one year later.[92] The phenomenon of 'blue team dynamics' has been described where the algorithm generating the data (and, consequently, user utilization) has been modified by service providers such as Google, Twitter, and Facebook in line with their business model.[92,93] Similarly, there is the so-called 'red team' dynamics, which occurs when social media platform users attempt to manipulate the data-generating process to support their own economic or political gain.[92,94]

2.5 CONCLUSIONS

Publicly available data from the considered social media networks were sparse and largely untrackable for the purpose of providing early clues of safety concerns regarding the prespecified case studies (rosiglitazone and stroke/myocardial infarction and HPV vaccine and infertility). The potential value of mining data from social networks appears to be greater for measuring awareness regarding emerging safety issues, with the caveat that this will be biased towards a younger and healthier population who comprise the majority of subscribers of these social networking sites. Further research investigating other case studies (including prospective investigations) and exploring other social media platforms are necessary to further characterize the usefulness of social media for post-marketing safety surveillance.

3

SOCIAL MEDIA FOR FOLLOWING A VACCINE DEBATE

ABSTRACT

BACKGROUND Public confidence in an immunization programme is a pivotal determinant of the programme's success. The mining of social media is increasingly employed to provide insight into the public's sentiment. This research further explores the value of monitoring social media to understand public sentiment about an international vaccination programme.

OBJECTIVE To gain insight into international public discussion on the paediatric pentavalent vaccine (DTP-HepB-Hib) programme by analysing Twitter messages.

METHODS Using a multilingual search, we retrospectively collected all public Twitter messages mentioning the DTP-HepB-Hib vaccine from July 2006 until May 2015. We analysed message characteristics by frequency of referencing other websites, type of websites, and geographic focus of the discussion. In addition, a sample of messages was manually annotated for positive or negative message tone.

RESULTS We retrieved 5771 messages. Only 3.1% of the messages were reactions to other messages, and 86.6% referred to websites, mostly news sites (70.7%), other social media (9.8%), and health-information sites (9.5%). Country mentions were identified in 70.4% of the messages, of which India (35.4%), Indonesia (18.3%), and Vietnam (13.9%) were the most prevalent. In the annotated sample, 63% of the messages showed a positive or neutral sentiment about DTP-HepB-Hib. Peaks in negative and positive messages could be related to country-specific programme events.

CONCLUSIONS Public messages about DTP-HepB-Hib were characterized by little interaction between tweeters, and by frequent referencing of websites and other information links. Twitter messages can indirectly reflect the public's opinion about major events in the debates about the DTP-HepB-Hib vaccine.

Becker BFH, Larson HJ, Bonhoeffer J, van Mulligen EM, Kors JA, Sturkenboom MCJM. Evaluation of a Multinational, Multilingual Vaccine Debate on Twitter. *Vaccine* **34** (2016)

3.1 INTRODUCTION

Vaccination programmes are among the most effective means for improving population health. But particularly at the time of programme introduction, they tend to be accompanied by public discussion.[27,96] This may increase public awareness of the vaccine and affect the programme beneficially.[97] However, public concern may lead to reduced uptake or even jeopardize the entire immunization programme.[98,99] Therefore, detecting changes in public sentiment early is important to understand its origin and dynamics and to inform appropriate measures to investigate concerns, guide public health decision making, or help identify issues with the vaccine or the vaccination programme.

Public attention and sentiment about vaccines have been evaluated previously by analysing different types of social-media messages and user-generated web content. Messages from the social-media platform MySpace were used for monitoring public sentiment about the human papillomavirus (HPV) vaccine.[24] Public news items about the HPV vaccine were shown to influence the public's awareness and opinion about HPV infection and vaccine in the United States (US).[25] Sentiments about an influenza vaccine shared through Twitter messages were found to correlate highly with US vaccination rates as reported by the US Centers for Disease Control and Prevention (CDC).[26] International debates about vaccines and the course and drivers of public confidence have also been studied through analysis of media sources such as news sites, blogs, and governmental reports.[27,28] Twitter and other social media have frequently been used for post-marketing surveillance of pharmaceutical safety issues.[19,21,22] Some studies have concluded that monitoring social media is more suitable for measuring public awareness of known safety issues than for providing clues about new safety signals (see chapter 2).[58]

Since 2001, a pentavalent paediatric vaccine against diphtheria, tetanus, pertussis, hepatitis B and *Haemophilus influenzae* type b (DTP-HepB-Hib) has been introduced into more than 70 low- and middle-income countries.[100] In a number of countries, the introduction of the vaccine was accompanied by a critical debate following a suspected association with the death of children, none of which have been deemed as causally related to the vaccine.[101] In India, a petition and a lawsuit was filed against the vaccine.[15,102] In Sri Lanka, Bhutan, and Vietnam, the market authorization for the vaccine was even temporarily suspended.[103]

In this study, we explore the value of public Twitter messages to gain insight into the multinational debate on the pentavalent vaccine.

3.2 METHODS

3.2.1 *Data collection*

The search query 'pentavalent OR pentavac OR quinvaxem' was used to retrieve messages about the pentavalent vaccine. The query terms were selected to retrieve messages from multiple national discussions about the vaccine, but not from all national or language-specific discussions (which would have required, amongst others, the inclusion of country-specific brand names and slang terms). The terms 'pentavac' and 'quinvaxem' are brand names of the pentavalent vaccine and specific to the vaccine as such. The term 'pentavalent' is also used in various other contexts (e.g., 'pentavalent' also occurs in chemistry and as user name on social media). To remove unrelated messages, a message retrieved by the term 'pentavalent' was only retained if it also contained the term 'child' or 'vaccine' (in the language of the message). The translations of 'child' and 'vaccine' in different languages were retrieved from OmegaWiki, a community-driven, multilingual dictionary.[104] OmegaWiki provided 94 terms for 'child' and 45 terms for 'vaccine'. The terms came from 67 different languages.

We used Twitter's advanced search web interface to collect messages retrospectively. The messages were collected on 1 May 2015. The advanced search interface provides the content and date of messages from the entire history of Twitter since 2006. We queried Twitter's web application programming interface (API) to retrieve additional data fields describing the language of the content, the identity of the author, the geographical location in his or her user-profile, and the interaction status of the message (original post, repost, or reply).

3.2.2 *Message analysis*

A random sample of 10% of the messages was selected for manual analysis. The message tone was manually analysed to gain insights into the sentiment about the pentavalent vaccine as reflected on Twitter. The two categories of message tone – *positive/neutral* and *negative* – and the criteria to assign the categories were the same as in a related study about public news.[28] A message was coded *negative* if it contained any indication of concern about the pentavalent vaccine or vaccination programme, e.g., information about an adverse event that occurred after immunization, vaccine suspension, or any other factor that might have a negative effect on the vaccine programme. A message was coded *positive/neutral* if it contained no indication of public concern about the vaccine or vaccination programme. Non-English messages were

translated using Google Translate while annotating.[105] Google Translate covered the languages of all messages in the sample, and the tone was apparent from the translations for all messages.

All authors of the messages in the random sample and the 50 authors creating most messages overall were characterized as *private person*, *news site*, *health information*, *health organization*, *government*, *vaccine-critical*, *manufacturer,* or *non-governmental organization (NGO)* based on their public Twitter profile.

To characterize the use of references (web links) in the collected messages, the most commonly referred (top-level) web domains were categorized as *news site*, *social media*, *health information*, *health organization*, and *other*. Additionally, all messages from the random sample that contained references, were manually assessed if the author added own content (i.e., if the message contained more than a link to or the title of the referred website).

We defined the geographical focus of a message by identifying the countries mentioned in the message or referred web pages. A dictionary of terms for geographical entities of countries (including cities and regions) was compiled from the GeoNames database to identify mentions of countries automatically.[106] To disambiguate terms that referred to entities in different countries, the country with the entity that had the largest population was selected. For example, 'Bali' is the name of a city in India and an island in Indonesia. Because the population of the Indonesian island is larger than that of the Indian city, mentions of 'Bali' were assigned to Indonesia. Messages that contributed to peaks in the message distribution over time were manually reviewed to identify the events that triggered the peaks.

The messages were analysed for occurrences of the standard format for reposts ('RT @user') to complement the information provided by the Twitter API. However, when evaluating public awareness and sentiment we did not distinguish between original posts and reposts, assuming that users primarily repost messages that reflect their own stance.

3.3 RESULTS

We retrieved 7,657 messages about the pentavalent vaccine from Twitter, of which 5,771 (75.3%) from 2,945 users remained after disambiguation. The number of messages grew over the years from 10 messages in 2008 to 2619 messages in 2013 (32 in 2009, 110 in 2010, 446 in 2011, and 1,033 in 2012). The numbers of messages should be seen against the background of a strong growth of Twitter messages until 2012, as well as the expanded introduction of the pentavalent vaccine and incidents of public resistance in some countries. After 2013 the number of messages

declined (1,091 in 2014 and 430 until May 2015). A histogram of all messages per month from 2012 until May 2015 is shown in figure 3.1 a).

In the manually annotated sample of 585 messages, 9 messages (1.5%) were false positives of the message retrieval and filtering and unrelated to the pentavalent vaccine. Among the 576 messages referring to the pentavalent vaccine, 37% had a negative tone and 63% of the messages had positive/neutral tone. The percentage of negative messages in the random sample reached its maximum in 2014 (2010: 9%, 2011: 24%, 2012: 19%, 2013: 43%, 2014: 51%, 2015: 36%) when reports of alleged cases of severe adverse events in Vietnam dominated the overall debate. The striped bars in figure 3.1 a) show the estimated number of messages with negative tone. No personal experience reports with the vaccine were found in the manually annotated messages.

Figure 3.2 shows the distribution of users from the random sample and of the top-50 tweeters over the user categories. In both sets most users were private persons or represented news sites. Health information sites, health organizations (including Global Alliance for Vaccines and Immunization (GAVI), World Health Organization (WHO), and CDC), governments, and vaccine-critical forums were overrepresented among the top-50 users. Many users (N=1,979; 67.2%) created only a single message, 920 users (31.1%) created 10 or less messages and the 50 users with the largest number of messages (1.7%) each created between 10 and 113 messages.

The dictionary of geographic entities contained 19,096 terms for 246 countries in 125 languages, with a median of 51 terms per language. In total 135 terms (0.7%) referred to entities in different countries and were disambiguated by population size. After a preliminary identification of countries, 78 terms were removed from the dictionary because the terms did not refer to geographical entities. Overall, 149 different countries were identified in 4,067 (70.4%) of the messages. The most frequently mentioned countries were India (2,047; 35.4%), Indonesia (1,056; 18.3%), Vietnam (803; 13.9%), and Pakistan (631; 10.9%). Most countries (104) were identified in less than 1% of the messages.

The most common languages of the messages were English (61.3%), Indonesian (16.1%), and Vietnamese (7.1%). English occupies a special role as the most common language on Twitter and as a common language for public communication in India. Countries were most frequently detected in Indian messages (79.2%), English messages (60.0%), French messages (44.4%), and Vietnamese messages (35.0%). Multiple countries were mentioned in 36.9% of the messages.

The country of origin could be identified by the information about the author for 3,067 (53.1%) messages. Most authors came from India (849; 27.6%), Indonesia (505; 16.4%), US (458; 14.9%), and Vietnam (267; 8.7%). The relationship between the country of the message author and

Figure 3.1: Number of messages about the pentavalent vaccine from January 2011 until April 2015 by month, with the estimated portion of messages with negative tone as striped bars. a) All messages, b) India, c) Indonesia, d) Vietnam, e) Pakistan. Messages between 2008 and 2010 (N=152; 2.6%) are not shown for clarity of the figure.

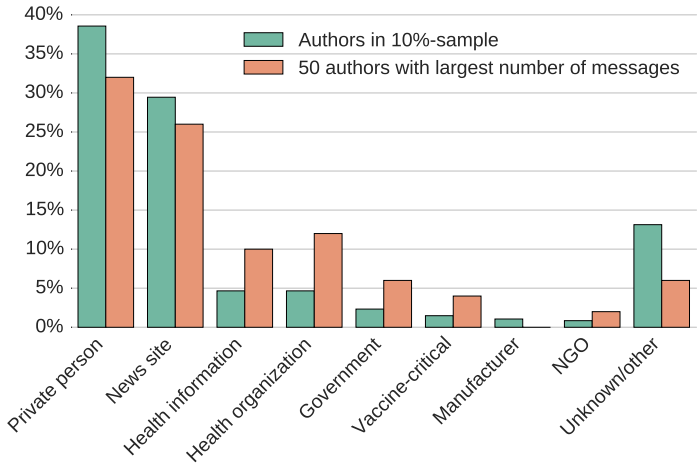


Figure 3.2: Distribution of users in the random sample and of the 50 users who created the largest number of messages, over different user categories.

Table 3.1: The relationship between the country of the message author and the country mentioned in the message content. Each cell contains the proportion of messages by authors from the country on the row, which mention the country in the column in their content (e.g., 74.3% of the messages from users in India are about India).

	Content									
Author	IN	ID	VN	PK	LK	US	BT	SR	JP	KR
IN	74.3	2.6	11.2	15.5	14.7	6.4	14.0	0.6	3.8	0.8
ID	2.6	83.2	0.8	0.2	0.2	8.3	0.2	34.9	10.3	0.0
US	41.5	6.3	17.9	13.1	13.5	18.1	12.7	0.4	5.7	8.5
VN	4.5	0.4	45.3	5.2	7.1	0.7	5.6	0.0	1.5	11.2

the country mentioned in the message content is shown in table 3.1. Each cell contains the proportion of messages by users from the country on the row, which mention the country in the column. The figure shows that authors largely focus on their own or neighbouring countries. Most messages from the US – which contributes the largest number of Twitter users but where the pentavalent vaccine was not a recently introduced vaccine – referred to events in India.

Only 158 messages (2.7%) were replies to other messages, and 180 messages (3.1%) were reposts. References to websites were very common, as 86.6% of the messages contained at least one reference. In the

manually annotated sample, the users provided original content in only 15.2% of the messages. The remaining messages only contained a link or copied content of the referred page. The most frequently referenced web domains were newspapers (70.7%), social media (9.8%), health information sites (9.5%), and health organizations (9.3%).

Most peaks of messages in figure 3.1 a) could be attributed to peaks of messages in individual countries, which in turn were in temporal relation to country-specific events as annotated in figures figure 3.1 b-e). Messages from India in December 2011 discussed the introduction of the pentavalent vaccine in the states Tamil Nadu and Kerala, and in May 2012 the introduction in five other states. The discussion about the vaccine gained momentum in India in 2013. Messages in January 2013 referenced news articles about child fatalities supposedly related to the pentavalent vaccine. The messages in April 2013 discussed the introduction of the pentavalent vaccine in eight more Indian states, but also criticized a supposed re-labelling of expired vaccines. The discussion between August and October 2013 included voices demanding the ban of the vaccine and continued discussions about the child fatalities. Numerous messages from February 2014 referred to articles alleging the association of the vaccine with child fatalities. Messages from May 2014 referred to the prequalification of a new brand of the pentavalent vaccine by the WHO. The messages from October and November 2014 discussed the introduction of the vaccine in the state of Rajasthan. Numerous news items in January, April, September and October 2013 primarily addressed the vaccination programme in India but also mentioned the vaccination programmes in Vietnam, Pakistan, and Sri Lanka, resulting in message peaks in the latter countries.

The messages about Indonesia in March 2012, August 2013 and December 2014 were composed of references to a few news articles discussing the production and introduction of the pentavalent vaccine in Indonesia. The messages in May 2013 discussed the suspension of the vaccination programme in Vietnam. Messages about Vietnam in November 2013 referred mainly to news items alleging (severe) adverse effects of the vaccine. In Pakistan the pentavalent vaccine was introduced in late 2014 and the November messages reference news articles about the introduction. The messages from Pakistan in March 2015 discussed the suspension of a health official for having spoiled the national pentavalent vaccine supply due to inappropriate storage.

3.4 DISCUSSION

In this chapter, we conducted an analysis of Twitter messages to characterize multinational debates about the pentavalent vaccine and vac-

cination programmes. We combined an analysis of geographical focus of the messages and message tone over time.

The debates on Twitter were portrayed by peaks of messages covering events in country-specific vaccination programmes. The perceptions of events on Twitter were local: authors of messages were largely reacting to events in their own country or neighbouring countries, suggesting multiple national debates rather than a multinational debate. In contrast to a previous study that observed a broad variety of concerns about vaccines in news sites, blogs and governmental reports,[28] the dominant concern in our data was about the safety of the vaccine. Most messages were created by users representing private persons and news sites. Stakeholders in the vaccination programme were overrepresented among the 50 users who created the largest number of messages, which suggests that they adhered longer to the debate than private persons.

The Twitter messages had three salient properties that have also been observed in chapter 2: few interactions (replies, reposts) between users, virtual absence of personal reports (in our case about the vaccine), and frequent references to other websites, particularly news portals. Many messages were comprised of only a reference or the title of the referred website. This appears to indicate that the messages were mainly created by sharing content on social media rather than to communicate with other users in the social network, a pattern that was also observed in Twitter messages about a measles vaccine in the Netherlands.[107] These properties can at least partially be explained by the focus of this study on messages that were made publicly viewable. Sharing of personal vaccine experiences and user interaction may be more common in private messages but private messages were unavailable in our data set.

With the lack of personal reports about the vaccine, our data does not directly reflect the public's opinion about the vaccine issues or events. But the Twitter messages may, however, reflect the users' opinion indirectly: the fact that no additional content was added by the authors of the large majority of Twitter messages suggests that the authors generally concur with the referenced content. The question whether messages on Twitter can shape public opinion was not in the scope of this study, but other studies argue that clear and transparent communication about vaccines, e.g., through social media like Twitter, can improve uptake rates.[96,97]

The analysis of Twitter messages for evaluating public sentiment may constitute a bias towards a debate between relatively young people with internet access.[89] This age group corresponds precisely to young parents who decide whether or not the pentavalent vaccine should be administered to their children. The population bias in this study represents a focus on persons for whom the vaccination programme is

potentially relevant. While most Twitter users are from the US, many countries where the pentavalent vaccine has been introduced have significant numbers of Twitter users.[77]

Our study has some limitations. First, the pentavalent vaccine has been marketed with various other brand names that were not used as query terms in this study. The inclusion of further brand names into the search query could help to expand the study to more countries where there are national debates about the vaccine (e.g., the scope of the study could be expanded to the vaccination programme in Ukraine by including the local brand name Пентаксим). However, we do not expect the characteristics of Twitter messages as described above to differ significantly in other countries. Second, the disambiguation of terms for countries based on population sizes may result in misallocations and could be improved by taking the context of the mention in the message into account. Third, we automatically detected countries in messages but did not try to determine whether a country mention was vaccine-related. The analysis of geographical focus in the debate could be further refined by distinguishing between country mentions that are related to the vaccine and those that are not.

3.5 CONCLUSION

The continuous monitoring of public debates about vaccines can help to alert vaccination programmes to emerging issues that may cause public confidence to plummet. We showed the potential value of monitoring social media retrospectively based on manual analysis of messages. When applying automatic techniques for the analysis of tone and topic of messages, the presented approach could increase the capacity and speed to allow for real-time analysis of public vaccine debates.

Part II

ACCESSING EXISTING EVIDENCE

4

AUTOMATIC RECOGNITION OF VACCINE DESCRIPTIONS AND CLASSIFICATION OF ARTICLES IN SCIENTIFIC LITERATURE

ABSTRACT

OBJECTIVE Once licensed, the benefit-risk profile of a vaccine requires monitoring over its life cycle. For every assessment, prior published evidence should be considered, but the large amount of available literature hinders manual information extraction. We present automatic methods for two basic tasks in the extraction of vaccine information from scientific literature: recognition of vaccine descriptions (VDR) and classification of vaccine articles (VAC).

MATERIALS AND METHODS For VDR, we evaluated a method based on the VaccO ontology of vaccine descriptions. For VAC, we compared dictionary-based, ontology-based, and machine-learning methods to assign MeSH vaccine headings to scientific articles. The approaches were evaluated on a manually created reference corpus of vaccine descriptions (VDR) and on MeSH vaccine headings from articles indexed in PubMed (VAC).

RESULTS Our ontology-based method for VDR performed reasonably well (F-score 0.69) considering the inter-annotator agreement on the reference corpus (F-score 0.80). For VAC, a machine-learning method performed best (F-score 0.76) but was outperformed by the ontology-based method for infrequent MeSH vaccine headings.

DISCUSSION Vaccines are described considerably different than drugs, requiring the development of vaccine-specific tools for VDR. Our ontology-based approach could be improved by expanding the ontology and ameliorating the detection of vaccine-description contexts.

CONCLUSION Using the VaccO ontology is a novel and promising approach for VDR. Our reference corpus of vaccine descriptions is the first of its kind and publicly available. For VAC, machine-learning methods perform best for classifying common MeSH vaccine headings, but for infrequent headings an ontology-based approach worked better.

Becker BFH, He HY, Sturkenboom MCJM, Kors JA. Identifying and Normalizing Vaccine Descriptions in Scientific Literature (Submitted)

4.1 BACKGROUND AND SIGNIFICANCE

Once licensed, the benefit-risk profile of a vaccine requires monitoring over its life cycle.[109] For every new assessment, prior published evidence about the benefits and risks (B/R) should be considered. Retrieving evidence about a vaccine from scientific literature involves two tasks: identifying relevant publications, and extracting semantic information, such as relations (e.g., between vaccines and events) or information frames (e.g., describing study design, result measures, population, and statistical power).[110]

Available scientific literature about human vaccines is growing quickly (more than 3,000 articles were published every year since 2015), which hampers a purely manual handling of literature. Literature databases such as PubMed[29] index articles with terms from a controlled vocabulary (Medical Subject Headings (MeSH),[30,31] in the case of PubMed) and are an indispensable tool for quickly identifying and retrieving relevant literature. But indexing is a manual process and does not cover all recent publications.[32] Also, articles are indexed on a document level and the localization of vaccine-specific information in the text is not retained. Localization, however, is a critical starting point for automatically extracting semantic information.

Automatic recognition of medicines and extraction of semantic information from scientific literature, independent of indexing by Pubmed, has been the subject of extensive research.[33,45] Little attention has been given to vaccines, though, which are preventive drugs and are described differently than therapeutic drugs (medicines) in scientific literature. Whereas medicines tend to be referred to by their (product or generic) name or by the name of their active ingredients, vaccines are often specified by their properties, such as immunization strategy and immunization targets (e.g., 'conjugated vaccine against *Haemophilus influenzae* type B'). This difference may hamper the transfer of existing automatic tools for extracting information about medicines to the extraction of information about vaccines.

We will focus on two problems in the automatic extraction of vaccine information from scientific literature: vaccine-description recognition (VDR) and vaccine-article classification (VAC). VDR refers to the identification of the precise text that describes the vaccine under consideration. VDR is an essential first step in extracting semantic information about vaccines. VDR differs from medicine named-entity recognition (NER) in that vaccine descriptions show a much larger syntactic variability than medicine descriptions, and their recognition requires semantic analysis to exclude entities that can be part of a vaccine description when they occur outside a vaccine description (e.g., the sentence 'The influenza vaccine was administered to the patients'

refers to an influenza vaccine, whereas the sentence 'The vaccine was administered to influenza patients' does not describe the vaccine).

VAC assigns the whole article to one or more codes representing vaccines in a controlled vocabulary. VAC constitutes an automatic approach for literature indexing, and a reliable VAC method could support or replace the manual effort of indexing new vaccine articles for literature databases.

To our knowledge, no automatic methods exist for VDR. Available reference corpora for medicines contain only few vaccine mentions,[34] which prevents their use in training or evaluating automatic methods. Regarding VAC, an ontology-based method was proposed by Hur et al. using terms from the Vaccine Ontology (VO) in SciMiner, an indexing engine based on dictionaries and grammar rules.[111–113] They report good performance but focus on research articles about a single pathogen, *Brucellosis*, which leaves it uncertain how well their results generalize to the vaccine literature in general. Moreover, their evaluation task consists of separating vaccine-related from vaccine-unrelated *Brucellosis* literature. We found that a similar good performance as reported can be obtained by a simple keyword query for vaccines (available as supplementary material online[69]).

4.1.1 *Objective*

This chapter presents two methods for VDR, and evaluate them on a reference corpus of in-text annotations of vaccine descriptions. For VAC, we propose dictionary-based, rule-based and statistical approaches, and compare their performance with existing MeSH annotations of a large corpus of Medline articles.

4.2 MATERIALS AND METHODS

This chapter builds upon a background set of articles (titles and abstracts) representative for the vaccine research from the last 20 years that is relevant to the B/R assessment of vaccines after their licensing. The background set was created in three steps. First, we queried PubMed for all articles about vaccines in humans published between 1997 and 2016 (using the query 'Vaccines[MAJR] AND Humans[MH] AND english[LA] AND 1997:2017[DP] AND hasabstract[text]'). The query referred to the top-most MeSH heading *Vaccines*, which implies all 85 subordinated MeSH vaccine headings. We then obtained the MeSH headings and subheadings of each article (MeSH defines 78 subheadings that are used to narrow down the meaning of main headings). We only retained articles with MeSH subheadings that indicated relev-

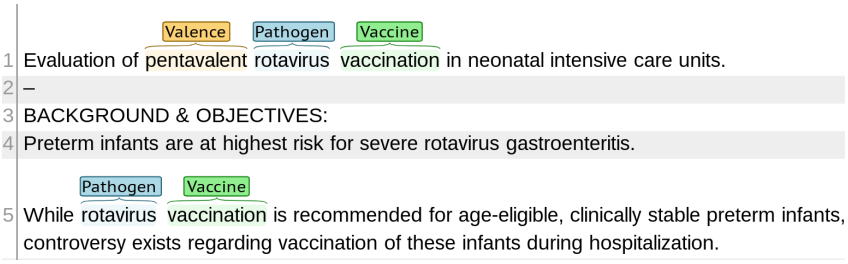


Figure 4.1: Example annotations from the reference corpus of in-text annotations of vaccine descriptions. Lines 1 and 5 contain vaccine descriptions with three and two entities. Line 4 does not contain a vaccine description.

ance for post-authorization B/R assessment (namely *Administration & dosage*, *Therapeutic use*, *Adverse effects*, *Economics*, *Supply & distribution*, and *History*) if they were a major subheading or a subheading of a major heading (headings or subheadings are usually marked as *major* if they are obtained from the title and/or statement of purpose).[32] The background set comprised 27,616 articles as of January 8, 2018.

4.2.1 *Reference set of vaccine descriptions*

To evaluate automatic methods for VDR, we manually created a reference corpus of in-text annotations of vaccine descriptions in scientific literature. Each vaccine description consists of a number of entities of types *Vaccine* (covering general terms such as 'vaccine' or 'immunization' as well as common abbreviations and products), *Immunization target* (including pathogens and diseases), *Vaccine strategy*, *Administration route*, *Ingredient*, *Valence*, or *Manufacturer*. These entity types correspond to the object properties defined in the VaccO Ontology of Vaccine Descriptions, which will be described in detail in chapter 7. Entities in vaccine descriptions that did not match these types were annotated with type *Other*. An entity was only annotated if it was part of a vaccine description in the same sentence. For example, in figure 4.1 the immunization target 'rotavirus' is part of a vaccine description in the lines 1 and 5, but not in line 4. Vaccines were not annotated when they were not characterized within the sentence (e.g., 'Finally, the vaccine authorization was suspended.').

Two annotators (BB and HY) independently annotated a random sample of 150 articles from the background set, and an arbiter (MS) resolved conflicting annotations afterwards to produce the reference corpus. Annotators and arbiter have at least four years of experience in vac-

cine research; HY and MS are pharmacists. The annotation guidelines and the reference corpus are publicly available.[114]

4.2.2 *Recognition of vaccine descriptions*

We compared two methods for VDR. The first method, $VaccO_{VDR}$, used the VaccO ontology, which includes 412 classes categorized into the entity types defined above (except *Other*). A class is described by one or more terms, and the terms of all classes in VaccO comprise its ontology dictionary. The VaccO dictionary contains 2,167 terms and was compiled from scientific literature and selected vocabularies of the Unified Medical Language System (UMLS).[52] The dictionary was stored in the Solr text search platform,[115] and the Solr TextTagger plugin was used to find terms from the dictionary in the input article.[116] If a term was found, $VaccO_{VDR}$ generated a candidate annotation, whose entity type was determined by the VaccO class described by the term. A candidate annotation was retained only if the sentence in which the term was found contained at least one annotation of type vaccine and one annotation of a different type.

The second approach used TaggerOne, a state-of-the-art drug identification system.[117] A specific model for identifying vaccines did not exist and could not be trained due to the lack of training data. Instead, we used a predefined model (`BC5CDRC`) from TaggerOne's software distribution, which has shown excellent performance in recognizing drugs and chemicals in scientific literature.[45]

4.2.3 *Vaccine article classification*

For VAC, we compared one dictionary-based, one ontology-based, two statistical (i.e., machine learning), and one hybrid method.

1. The *Dictionary* method was based on a dictionary of terms that describe the MeSH vaccine headings. The dictionary was generated in two steps using the UMLS. First, we selected all concept unique identifiers (CUIs) from the UMLS that correspond to the vaccine headings. Then, we aggregated all English terms that correspond to these CUIs in the following UMLS vocabularies: MeSH, MedDRA,[118] SNOMED-CT,[119] ICD-10 CM,[120] and CHV.[121] The dictionary was stored in the Solr text search platform and Solr TextTagger was used to find occurrences of the terms in an article. The MeSH vaccine headings that corresponded to the found terms were assigned to the article.

2. The *VaccO$_{VAC}$* method applied an algorithm for aligning vaccine coding systems that will be described in detail in chapter 7. In preparation, we extracted the property values from all MeSH vaccine headings. Property values are a flat, normalized representation of a vaccine description that combines information from the description and the VaccO ontology. For example, the description 'BCG vaccine' has the property values *[Immunization target: Tuberculosis; Strategy: Attenuated; Ingredient: BCG]*. To classify an article, we first extracted the property values from the article. All MeSH vaccine headings whose property values for immunization targets exactly matched those from the article, comprised the candidate set. Finally, we assigned all headings from the candidate set, whose property values had maximal similarity to the property values from the article (similarity was measured by Jaccard coefficient[122]).

3. The *RF-BoW* method was a random forest model using bags-of-words as features.[123] The bags-of-words were created by tokenizing the article text, lemmatizing the tokens and converting them to lowercase. *RF-BoW* was implemented in Python using the NLTK interface to WordNet[124] for lemmatization and using the Scikit-learn library with default parameters for creating the bag-of-words and training the random forest models.[125]

4. *The RF-VaccO* method was a random forest model using the VaccO property values identified in the text as its features. VaccO property values were extracted as described for method *VaccO$_{VAC}$*. The implementation of *RF-VaccO* used the random forest model from the Scikit-learn library with default parameters.

5. The *CNN* model was a convolutional neural network for multi-label classification of texts,[126] composed of one-dimensional convolutional layers with max pooling for 1-grams to 5-grams. The implementation is based on the Python library Magpie.[127] The input text was represented by word embeddings using word2vec[128] that were pretrained on all full-text articles from PubMed Central by Pyysalo et al.[129]

Methods *Dictionary* and *VaccO$_{VAC}$* were based on general information about vaccines and did not require task-specific training data. The statistical methods (*RF-BoW*, *RF-VaccO*, *CNN*) were trained on a random sample of 20,000 articles from the background set using the MeSH vaccine headings assigned in PubMed as targets.

4.2.4 *Evaluation*

To evaluate the methods for VDR, we applied them to the manually created reference corpus. The performance was measured by the strict and lenient precision, recall, and F-score, which are defined as follows.[130] An annotation is characterized by an article identifier, the entity type, and the start and end positions of the annotated term. The set of true-positive annotations (TP) comprises the generated annotations for which document identifier and entity type match those of the reference annotations, and start and end positions coincide (for strict performance measures) or overlap (for lenient performance measures). The set of false positives (FP) comprises all generated annotations excluding true positives, and the set of false negatives (FN) comprises all reference annotations excluding true positives. Precision is defined as $|\mathrm{TP}| / (|\mathrm{TP}| + |\mathrm{FP}|)$ and recall as $|\mathrm{TP}| / (|\mathrm{TP}| + |\mathrm{FN}|)$, where $|\mathrm{S}|$ denotes the cardinality of a set S. F-score is the harmonic mean of precision and recall and defined by $2 * \mathrm{precision} * \mathrm{recall} / (\mathrm{precision} + \mathrm{recall})$. We report lenient performance measures and describe differences between lenient and strict matches. Since TaggerOne did not annotate entity types, they were not considered in determining the performance of TaggerOne. To analyse the errors of $VaccO_{VDR}$, we randomly selected ten FP annotations and ten FN annotations of each entity type (if there were less than ten erroneous annotations, we took all) and categorized them by the reason for the error.

The inter-annotator agreement (IAA) was measured by the lenient F-score, and we describe the differences between lenient and strict matches. Note that the F-score is invariant to which of the two annotated sets serves as the reference when computing precision and recall.

To evaluate the automatic methods for VAC, we applied them to all 7,616 articles from the background set that were not used for training. The performance of a VAC method was measured by comparing the automatically assigned MeSH vaccine headings with those assigned on Pubmed as a reference, and computing the precision, recall, and F-score.

The dictionary-based and ontology-based methods assign a binary value to each MeSH vaccine heading. The statistical models generate a value between 0 and 1 for each MeSH vaccine heading, and a heading was considered present if the value was above a given threshold. To estimate the optimal threshold, we trained a model on 90% of the training data and chose the threshold that maximized the F-score of the model on the remaining training data.

4.3 RESULTS

4.3.1 *Reference corpus*

The annotators achieved an average IAA of 0.80 in creating the reference corpus. The IAA was high for entity types *Vaccine* (0.85), *Immunization target* (0.82), *Administration route* (0.87), and *Valence* (0.85). The IAA was 0.65 for *Strategy* and only 0.27 for *Ingredient*. The low agreement for ingredients was due to ambiguous terms (e.g., 'BCG' can refer to a vaccine or an ingredient) and annotation errors, where ingredients were not annotated on the lowest level of detail (e.g., annotating "tetanus toxoid" with one entity of type *Ingredient* instead of two entities of types *Immunization target* and *Strategy* according to the annotation guidelines). The IAA was undefined for entity types *Manufacturer* and *Other*, for which only one annotator created annotations. Of the overlapping annotations, the far majority (95.9%) were exact matches. Of the inexact matches, most comprised cases where one annotator included unessential information ('inactivated' versus 'inactivated whole-cell').

After adjudication of all annotation differences, the reference corpus of 150 articles contained 2,106 annotations of vaccines and vaccine properties (table 4.1). The median number of annotations in an article was 12, and the maximum was 49. Six articles were left unannotated because they did not contain a specific vaccine description but covered vaccines in general. Of the annotations of the *Vaccine* entity type, 661 (70.1%) consisted of 21 different terms that included 'vaccine', 'immunization', and variants thereof. The remaining vaccine annotations were accounted for by 9 other general vaccine terms (e.g., 'shot'), 38 unique common abbreviations, and 30 product names. Immunization targets were most frequently used to characterize vaccines. Only three manufacturers could be annotated, in two articles. Twenty-seven occurrences of 9 unique entities did not match the VaccO property categories and were annotated with type *Other* (e.g., 'patient-specific', 'parental').

4.3.2 *Identification of vaccine descriptions*

The average F-score of the $VaccO_{VDR}$ method for recognizing vaccine descriptions in the reference corpus was 0.69 (table 4.2). The F-score increased only slightly when entity types were ignored (0.71). The approach worked very well for recognizing terms of type *Valence* (F-score 1.00) and had reasonable performance for most other vaccine properties (*Immunization target*: 0.77, *Strategy*: 0.62, *Route*: 0.67). Terms of type *Ingredient* were identified with fair precision (0.70) but low recall (0.21). The far majority of overlapping annotations were strict

Table 4.1: Number of annotations, unique terms, and annotated articles in the reference corpus

Entity type	Annotations	Terms	Articles
Vaccine	942	98	143
Immunization target	815	171	135
Ingredient	109	49	29
Strategy	108	32	48
Route	67	21	21
Valence	35	11	16
Manufacturer	3	3	2
Other	27	9	13
Total	2,106	380	144

Table 4.2: Performance measures of method $VaccO_{VDR}$ for identification of vaccine properties stratified by entity type

Entity type	Precision	Recall	F-score
Vaccine	0.55	0.80	0.66
Immunization target	0.81	0.73	0.77
Ingredient	0.70	0.21	0.32
Strategy	0.54	0.72	0.62
Route	0.63	0.71	0.67
Valence	1.00	1.00	1.00
Manufacturer	0.33	0.33	0.33
Overall	0.64	0.74	0.69
Overall, ignoring type	0.66	0.76	0.71

matches (96.9%). TaggerOne using the `BC5CDRC` model was not suited for identifying vaccine descriptions (F-score 0.23 with precision 0.58 and recall 0.14).

The results of the error analysis are shown in table 4.3. Most FP annotations of $VaccO_{VDR}$ marked terms that can be part of a vaccine description but were used outside the context of vaccines, for example 'AIDS' in 'safety of the flu vaccine among AIDS patients'. Other less frequent reasons for FP annotations were ambiguous terms in the dictionary, which resulted in wrong entity types (e.g., 'BCG' can refer to a tuberculosis vaccine and to its active ingredient), and longer dictionary terms that comprised several terms and were favoured by Solr TextTagger over the individual terms. This error of $VaccO_{VDR}$ corresponded

Table 4.3: Reasons for false positive (FP) and false negative (FN) errors in VDR using $VaccO_{VDR}$

Error	Reason	Vaccine	Target	Ingr.	Strat.	Route	Manuf.	Total
FP	No vaccine context	10	10	1	5	9	2	37
	Ambiguous term	0	0	2	5	1	0	8
	Complex term	0	0	7	0	0	0	7
FN	Missing concept	2	3	5	5	6	1	22
	Missing term/abbrev.	5	5	2	1	3	1	17
	Missed vaccine context	0	2	1	4	1	0	8
	Contextual usage	3	0	2	0	0	0	5

to the annotation error where components were not annotated on the lowest level of detail (e.g., 'tetanus toxoid').

Most FN annotation errors were due to entities that were not represented in the VaccO ontology, for example, specific immunization targets, vaccine products and residual classes (e.g., 'non-adjuvanted'), which cannot be modelled semantically in VaccO due to limitations of its language, OWL2, to represent residual classes.[131] Other FN errors were due to missing terms and abbreviations for existing VaccO classes, missed detection of the vaccine context, and a use of terms that depends on the context (e.g., 'yellow fever coverage' refers to a yellow fever vaccine and not the immunization target).

4.3.3 *Vaccine article classification*

Figure 4.2 shows the performance of the five methods for automatic VAC. The *Dictionary* method had a low overall performance (precision: 0.43, recall: 0.47, F-score: 0.45), which indicates that the wordings of vaccine descriptions in scientific literature differ considerably from the definitions of vaccines in medical vocabularies. The $VaccO_{VAC}$ method, based on the VaccO ontology of vaccine properties, performed better than *Dictionary*, mainly because of higher precision (precision: 0.71, recall: 0.51, F-score: 0.60). The performance of *RF-BoW*, the random forest model using bags-of-words, was comparable to the performance of the $VaccO_{VAC}$ method (precision: 0.68, recall: 0.57, F-score: 0.62). The *RF-VaccO* method, combining VaccO property identification with statistical assignment of MeSH vaccine codes, outperformed $VaccO_{VAC}$ and *RF-BoW* (precision: 0.76, recall: 0.61, F-score: 0.68). The *CNN* model performed best with an F-score of 0.76 (precision: 0.82, recall: 0.71).

The heading-specific F-scores of *CNN* correlated strongly with the number of positive training articles per MeSH vaccine heading (Spear-

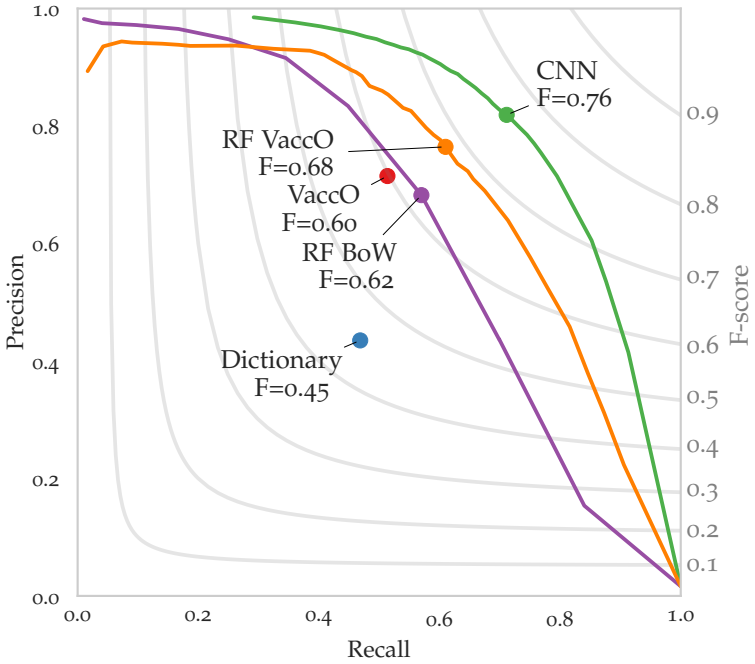


Figure 4.2: Performance measures of five methods for automatic indexing of vaccine literature. The dictionary-based and ontology-based approaches generate binary outcomes and their performances are indicated by point values. The statistical methods generate probability outcomes and their performances for different cut-off thresholds are shown as recall-precision curves. The F-scores for the estimated optimal threshold are indicated.

man coefficient $r = 0.753$; $p < 0.001$),[132] whereas such correlation was not observed for $VaccO_{VAC}$ ($r = 0.173$; $p < 0.124$). *CNN* performed better than $VaccO_{VAC}$ in 39 of the 40 most frequent MeSH vaccine headings, but the F-scores of $VaccO_{VAC}$ exceeded *CNN* for 34 of the 40 vaccine headings with the lowest number of training articles (figure 4.3).

4.4 DISCUSSION

In this chapter, we developed and evaluated automatic methods for the recognition of vaccine descriptions (VDR) and for the classification of vaccines in research articles (VAC), two basic tasks in the extraction of established evidence about vaccines from scientific literature. The manually created reference corpus of annotated vaccine descriptions is the first of its kind and publicly available to facilitate future research. Our proposed method for VDR, $VaccO_{VDR}$, performed reasonably (F-

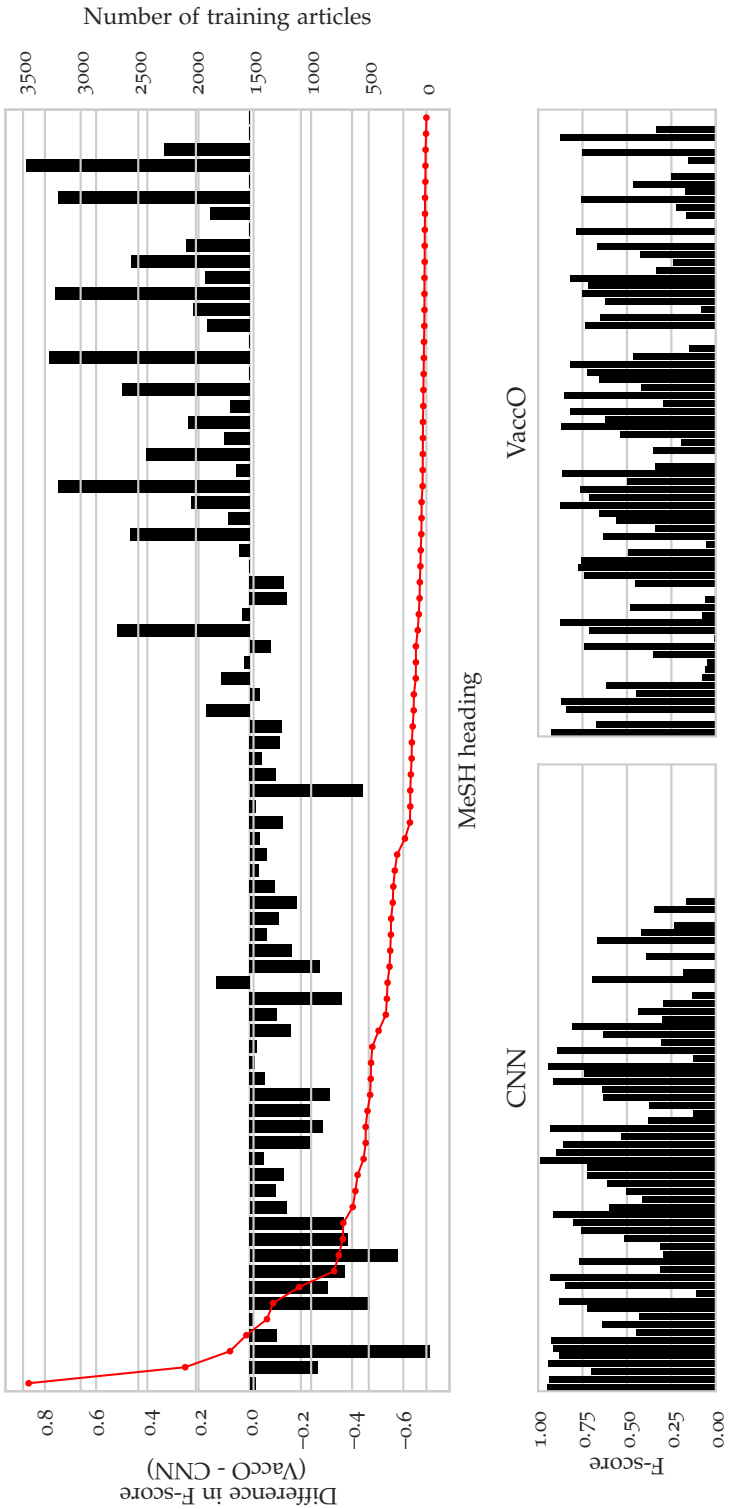


Figure 4.3: Top: Difference between the F-score of the $VaccO_{VAC}$ method and the F-score of the *CNN* method for individual MeSH vaccine heading (black bars, left axis), and number of articles in the training set indexed with the vaccine heading (red dots, right axis). Bottom: Heading-specific F-scores of methods *CNN* and $VaccO_{VAC}$.

score 0.69) in comparison with the IAA (0.80). For VAC, the *CNN* method performed best but had difficulty to correctly assign infrequent MeSH vaccine headings.

VDR may be considered a special case of drug recognition at first sight. However, vaccine descriptions differ considerably from descriptions of medicines, which prohibits the application of medicines-specific recognition methods for the recognition of vaccines. For example, the `BC5CDRC` model for TaggerOne, which excelled in recognizing drugs and chemicals, performed poorly in VDR. $VaccO_{VDR}$ did not require any training data but was only based on general domain knowledge from the VaccO ontology. The error analysis suggested two main changes to improve the performance of $VaccO_{VDR}$. First, the VaccO ontology, which was designed to represent descriptions of licensed vaccines in Europe, should be expanded with immunization targets and products from non-European vaccine research to improve the recall of $VaccO_{VDR}$. Second, the naïve detection of the vaccine context by testing sentence-co-occurrence between vaccine entities and property entities accounts for most FP and many FN errors. The test could be improved by automatically extracting syntactical patterns that commonly relate vaccines and their properties,[133,134] and ruling out vaccine properties that are not connected to a vaccine term by a common syntactic pattern.

The evaluation of our automatic methods for VAC illustrated the limitations and strengths of dictionary-based, ontology-based, statistical, and hybrid approaches. A simple dictionary of terms on the level of MeSH vaccine headings (method *Dictionary*) lacked the necessary flexibility for the large syntactic variation of vaccine descriptions in scientific literature. The $VaccO_{VAC}$ method used only general vaccine information from the VaccO ontology and was not tailored to the specific task of automatic literature classification. However, the underlying algorithm for code assignment in $VaccO_{VAC}$, which proved suitable for aligning descriptors between vaccine coding systems, had moderate performance for classifying vaccine literature. The better performance of the hybrid random forest model on the basis of VaccO vaccine property values (*RF-VaccO*) demonstrated the existence of a more robust approach for assigning vaccine headings based on property values. The convolutional neural network *CNN* constituted the best model for categorizing vaccine literature but required substantial training material to train classifiers that perform well. For most of the infrequent MeSH vaccine headings, *CNN* performed worse than $VaccO_{VAC}$. Vaccine headings, however, may necessarily be infrequent in the training data for novel or less-studied vaccines. Combining both methods dependent on the number of training articles may further improve performance.

4.4.1 *Conclusion*

VaccO$_{VDR}$ is a novel and promising approach for the automatic identification of vaccine descriptions in the rapidly increasing scientific literature. It could be used as a basic building block for the automatic extraction of relational information about vaccines, which is further developed in the next chapter. The continued research in automatic indexing of vaccine literature on the basis of ontologies is warranted because its performance is independent from the availability of training examples and may supplement machine-learning methods, which performed only strongly on frequent MeSH vaccine headings.

5

EXTRACTION OF CHEMICAL-INDUCED DISEASES USING PRIOR KNOWLEDGE AND TEXTUAL INFORMATION

ABSTRACT

We describe our approach to the chemical-disease relation (CDR) task in the BioCreative V challenge. The CDR task consists of two subtasks: automatic disease named-entity recognition and normalization (DNER), and extraction of chemical-induced diseasess (CIDs) from Medline abstracts. For the DNER subtask, we used our concept recognition tool Peregrine, in combination with several optimization steps. For the CID subtask, our system, which we named RELigator, was trained on a rich feature set, comprising features derived from a graph database containing prior knowledge about chemicals and diseases, and linguistic and statistical features derived from the abstracts in the CDR training corpus. We describe the systems that were developed and present evaluation results for both subtasks on the CDR test set. For DNER, our Peregrine system reached an F-score of 0.757. For CID, the system achieved an F-score of 0.526, which ranked second among 18 participating teams. Several post-challenge modifications of the systems resulted in substantially improved F-scores (0.828 for DNER and 0.602 for CID). RELigator is available as a web service.[136]

5.1 INTRODUCTION

The extraction of chemicals, diseases, and their relationships from unstructured scientific publications is important for many areas of biomedical research, such as pharmacovigilance and drug repositioning.[82,137] Text-mining systems in combination with methods for literature-based discovery and network analysis hold promise for automatically generating new hypotheses and fresh insights.[138,139] The manual extraction of these entities and relations, and their storage in structured databases is cumbersome and expensive, and it is impossible for researchers or curators to keep pace with the ever-swelling number of papers that

Pons E + Becker BFH, Akhondi SA, Afzal Z, van Mulligen EM, Kors JA. Extraction of Chemical-Induced Diseases Using Prior Knowledge and Textual Information. *Database* **2016** (2016)

are being published. Automatic extraction of chemical-disease relations (CDRs) should solve these problems, but previous attempts have met with limited success. One of the difficulties that has to be addressed is the identification of relevant concepts, i.e., chemicals and diseases.[140,141] Concept identification goes beyond concept recognition in that not only the mention of a chemical or a disease has to be recognized, but that in addition a unique identifier has to be assigned, which links the concept to a source that contains further information about it.[142] Also the detection of relationships between the identified chemicals and diseases remains a challenging task,[143–145] partly because available annotated corpora to train and evaluate extraction algorithms are limited in size.[144,146]

In BioCreative V, one of the challenge tasks is the automatic extraction of CDRs from biomedical literature.[147] The CDR task comprises two subtasks. The first subtask involves automatic disease named-entity recognition and normalization (DNER) from a set of Medline documents, and can be considered as a first step in CDR extraction. The second subtask consists of extracting chemical-induced diseases (CID) and delivering the chemical-disease pairs per document.

Our team participated in both CDR subtasks. For the DNER subtask, we used our concept recognition tool Peregrine, in combination with several optimization steps.[148] For the CID subtask, we applied the optimized Peregrine system for disease concept recognition; for chemical concept recognition, we used tmChem, a chemical concept recognizer that was provided by the challenge organizers.[149] A relation extraction module was trained on a rich feature set, including features derived from a graph database containing prior knowledge about chemicals and diseases, and linguistic and statistical features derived from the training corpus documents.

In the following, we describe the systems that we developed for the BioCreative challenge, as well as several post-challenge improvements, and present evaluation results for both subtasks on the CDR training and test sets.

5.2 METHODS

Figure 5.1 shows the different steps in our workflow for CDR extraction from biomedical abstracts. The data, methods for entity recognition and normalization, and relation extraction methods are described below.

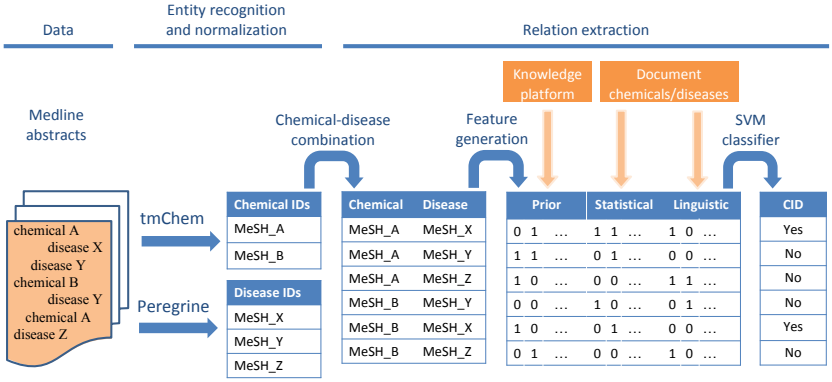


Figure 5.1: Workflow for CDR extraction. The chemical and disease entities in a Medline abstract are recognized and mapped to their corresponding MeSH identifiers by tmChem (for chemicals) and Peregrine (for diseases). For each possible combination of chemicals and diseases that are found in the document, features are generated based on prior knowledge from a knowledge platform, and based on statistical and linguistic information from the document. The features are fed to an SVM classifier to detect CIDs.

Table 5.1: Characteristics of the CDR corpus

Data	Training	Devel.	Test	Total
Abstracts	500	500	500	1,500
Chemical mentions	5,203	5,347	5,385	15,935
Unique chemical identifiers	1,467	1,507	1,435	4,409
Disease mentions	4,182	4,244	4,424	12,850
Unique disease identifiers	1,965	1,865	1,988	5,718
CDRs	1,038	1,012	1,066	3,116

5.2.1 *Data*

The CDR task data consist of a training, a development and a test set, each containing 500 Medline abstracts. Chemicals and diseases in the abstracts were manually annotated in the form of text offset, text span, and Medical Subject Headings (MeSH) identifier.[147] Chemical-disease interactions were annotated at the document level as MeSH-identifier pairs, but only if a mechanistic relationship between a chemical and a disease was explicitly mentioned in the abstract.[150] Therapeutic relationships between chemicals and diseases were not annotated. Table 5.1 shows the number of annotated (unique) identifiers of chemicals and diseases, and the number of annotated relationships.

5.2.2 *Entity Recognition and Normalization*

Chemical concept recognition was carried out using the tmChem chemical recognizer system.[149] The tmChem system was one of the best performing systems in the previous BioCreative IV chemical-named entity recognition (CHEMDNER) challenge.[45] It includes a dictionary look-up to map recognized chemicals to MeSH identifiers. tmChem is an ensemble system that combines two systems based on conditional random fields (CRFs), of which we only used the one that performed best in the CHEMDNER challenge. We trained this system on the 1,000 documents in the CDR training and development sets.

For the recognition and normalization of diseases, we employed our dictionary-based concept recognition system Peregrine.[148] Peregrine employs a user-supplied dictionary and splits the terms in the dictionary into sequences of tokens. When such a sequence of tokens is found in a document, the term and the concept associated with that term, is recognized in the document. Peregrine removes stopwords (we used the PubMed stopword list[151]) and tries to match the longest possible text phrase to a concept. It uses the Lexical Variant Generator tool (LVG) to reduce tokens to their stems before matching.[152]

We constructed a dictionary with concepts and corresponding terms taken from four biomedical vocabularies, as contained in the Unified Medical Language System (UMLS) 2015AA edition.[153] These are: MeSH; Medical Dictionary for Regulatory Activities (MedDRA); SNOMED Clinical Terms (SCT), and International Classification of Diseases Clinical Modifications version 10 (ICD-10 CM). The MetamorphoSys tool[153] was used to only include concepts that belong to the semantic group *Disorders*,[154] and to discard terms that are flagged as suppressible in the UMLS.

After a document was processed with Peregrine, several post-processing steps were executed. We extracted all abbreviations and their corresponding long forms,[155] and made sure that any combination of abbreviation and long form was tagged with the same concept. Adjacent term spans that were identified as the same concept were merged. For our challenge submission on the test set, we filtered out terms that Peregrine had tagged erroneously in the training and development data (false-positive terms).

After the challenge, we refined this approach by only removing false-positive terms if the ratio of true-positive to false-positive terms was lower than 0.3. This threshold was heuristically set based on the training data to prevent that an occasional false-positive detection would cause the removal of terms that were generally correctly recognized. Moreover, we performed a term-frequency analysis by indexing a random set of one million Medline abstracts and manually checking the

2,000 top-ranking terms found by Peregrine. Erroneously recognized terms were also removed. Finally, we added all terms that Peregrine had missed in the training set (false-negative terms) to the dictionary.

The UMLS identifiers of the concepts that resulted from the indexing and post-processing steps were mapped to MeSH identifiers with the IntraMap tool.[153] IntraMap contains a precompiled mapping table that links each UMLS concept to the semantically closest MeSH header.

5.2.3 *Relation Extraction*

We formulated the relation extraction task as a binary decision problem: for each possible pair of chemicals and diseases found in a document, determine whether there is a relationship. To train the relation extraction algorithm, we constructed training instances based on the perfect (gold-standard) entity annotations of the training data. Of the 10,693 possible pairs of annotated chemicals and diseases, 2,050 were labelled as positive instances because the pair had been annotated as a relationship by the reference. The other 8,643 pairs were labelled as negative instances. Co-occurrence pairs were allowed to cross the title-abstract border. For each instance, three sets of features were generated, based on prior knowledge and on statistical and linguistic information from the document.

Prior knowledge features

To generate features based on existing, prior knowledge, we used a graph database, the Euretos Knowledge Platform.[156] The Euretos Knowledge Platform is a commercial system and not freely available but life-science researchers can request free browsing access. We have obtained an academic license to use a local installation of the system. The graph database contains entities and relations from (curated) structured databases, such as UniProt, the Comparative Toxicogenomics Database and UMLS, and from scientific abstracts (semantic Medline).[157] Each connection between entities can have a set of named relations or predicates. Attached to each predicate is provenance information, including the different sources in which the relation was found and, per source, the number of records or abstracts with the relation. Euretos provides an application programming interface that was used to query the database for paths between two given entities. A path can be direct (i.e., the entities have a direct, one-directional (causal) or two-directional (non-causal), relationship) or indirect (the entities are connected through one intermediate entity; if the two relationships involved are one-directional, one relationship should point towards the intermediate entity and the other should point away from it). For

each path, a confidence score based on provenance information was computed that indicates how strongly the entities are related. If two entities were connected through both direct and indirect paths, the latter were ignored. If there were multiple paths of the same length, the total score and total provenance count were taken as the maximum of the path scores and path provenance counts, respectively. The provenance count of an indirect path was taken as the minimum of the provenance counts of the two predicates involved. We determined for each chemical-disease pair the path type (direct, indirect or no path), the confidence score, the number of paths, the set of predicates involved and the provenance count.

Statistical features

The statistical feature set contained, for each chemical-disease pair at the document level, the number of mentions of the chemical and of the disease and number of possible chemical-disease pairs in the document (i.e., number of chemical mentions times number of disease mentions). The ratios of these numbers to the numbers of all chemical mentions, all disease mentions and all possible chemical-disease pairs in the document were also taken as features. Additional features captured the minimal sentence and word distance between the mentions of the chemical and the disease. Binary features indicated whether the chemical, the disease or both were mentioned in the document title. The MeSH identifiers of the chemical and disease were included as nominal features.

Linguistic features

We used the Stanford CoreNLP parser in version 3.4.1 with the English PCFG parsing module to generate dependency trees of the sentences of each document, and determined *governing* verbs of chemicals and diseases, and *relating* words of chemical-disease pairs (figure 5.2). The governing verb of a word was defined as the first verb in the parse tree that was encountered when traversing the tree from the word towards the root. The relating word of a chemical-disease pair was defined as the first word in the parse tree that the chemical and disease had in common. If the chemical and disease mentions appeared in different sentences, the relating word was undefined.

Two sets of linguistic features were used. For the first set, only one pair of chemical and disease mentions in the document was considered. The pair was selected on the basis of the following heuristics. A pair with the chemical and disease mentions in the same sentence had precedence over a pair with mentions in different sentences, and a pair where no other chemical-disease pair could be found lower in the parse

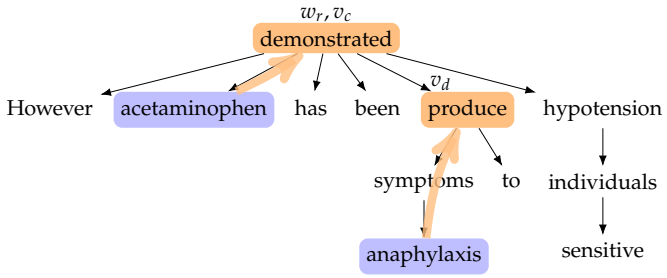


Figure 5.2: Example dependency parse tree for the sentence 'However, acetaminophen has been demonstrated to produce symptoms of anaphylaxis, including hypotension, in sensitive individuals.'. The governing verb v_d of the disease is 'produce'; the governing verb v_c of the chemical is 'demonstrated', which coincides with the relating word w_r.

tree had precedence over a pair for which this was not true. If there were only pairs with mentions in different sentences, the last pair with the chemical mention before the disease mention was selected. If no such pair existed, the first chemical and disease mentions in the document was selected. The following features were derived: governing verb of the chemical and of the disease, relating word, and governing verb of the relating word. Note that if the chemical and the disease occurred in different sentences, the governing verbs were taken from different parse trees and the relating word and its governing verb were undefined. Further features indicated whether the chemical was mentioned before the disease, and whether another chemical-disease pair could be found lower in the parse tree. After the challenge, we added four features that signified whether the relating word and the governing verb of the chemical, of the disease, and of the relating word, were negated. Negation was assessed by the presence of negation modifiers in the parse tree. Three more post-challenge features indicated whether the chemical was the same as the relating word, whether the governing verb of the disease was the same as the relating word, and whether both governing verbs were the same as the relating word.

For the second set of linguistic features, we aggregated information about the governing verbs and relating words from all possible pairs of chemical and disease mentions in the documents. This set contained one numeric feature for each governing verb or relating word encountered in the training set, indicating how many times that word was found as a governing verb or relation word for the chemical-disease pairs in the document.

Machine learning

Various machine-learning algorithms were explored, utilizing Weka machine learning libraries.[158] Performance was estimated by 10-fold cross-validation.

In a preliminary analysis in which we compared various classification algorithms, support vector machines (SVMs) proved to have superior performance. Therefore, we continued to optimize parameters for the SVM classification model. We used C-support vector classification with radial basis function kernel type, initially with default settings for cost (1.0) and gamma (0.0).

All numeric features were normalized to scale between zero and one. Because of the class imbalance the cost matrix of the SVM was set to 5:1, giving extra weight to the minority class. Utilizing the best performing feature set, we tuned the cost and gamma parameters by performing a grid search, again applying 10-fold cross-validation. During the grid search, we used a fixed decision threshold of 0.5 for the SVM. We subsequently varied the decision threshold to optimize the F-score of the SVM.

Evaluation

For each document, the disease concepts and the disease-chemical relationships found by our systems were compared with the gold-standard annotations, resulting in true-positive, false-positive and false-negative detections. Micro-averaged precision, recall and F-score were then computed over the whole document set. We implemented our final challenge systems as web services, which the CDR task organizers utilized for online system evaluation on the test set.

5.3 RESULTS

5.3.1 *DNER task*

Table 5.2 shows the performance of the Peregrine challenge system and the system with post-challenge modifications on the DNER test set. The challenge system obtained an F-score of 0.757. The modified system performed considerably better achieving an F-score of 0.828, well above the average F-score (0.760) of the 16 teams participating in the DNER task.[147]

To get insight in the cause of the remaining errors of the modified Peregrine system, we randomly selected and analysed 50 false-positive and 50 false-negative detections. Table 5.3 shows that almost half of the false-positives were due to incorrectly recognized terms, e.g., in

Table 5.2: Performance of the Peregrine challenge and post-challenge systems for disease normalization on the test set

System	Recall	Precision	F-score
Peregrine, challenge	0.772	0.737	0.757
Peregrine, post challenge	0.839	0.818	0.828

Table 5.3: Error analysis of 50 false-positive (FP) and 50 false-negative (FN) errors of the post-challenge Peregrine system

Error type	FP	FN
Term mapped to incorrect MeSH identifier	8	6
Term incorrectly on exclusion list	-	5
Term partially recognized	13	15
Term incorrectly recognized	23	-
Term not recognized	-	20
Annotation error	6	4

the form of an erroneous synonym ('patch' for 'plaque') or a term that is no disease ('glucose tolerance curve'). The largest group of false-negatives resulted from missing synonyms in the terminology. Interestingly, many of these synonyms were present in other vocabularies in the UMLS than the ones that we selected for building our terminology. A smaller number of terms were correctly recognized but then mapped to the wrong MeSH identifier, or were excluded because their true-/false-positive ratio was below the threshold of 0.3. Partial recognition of terms, e.g., 'carcinoma' in 'cervical carcinoma' or 'ethanol abuse' in 'cocaine and ethanol abuse', resulted in considerable numbers of false-positives as well as false-negatives. Finally, we encountered a number of gold-standard annotation errors. For example, in the term 'ST depression', an electrocardiographic concept, 'depression' had been annotated as a psychological disorder. As another example, a mention of the term 'death' had not been annotated, whereas the annotation guidelines explicitly state that this term should be annotated.

5.3.2 *CID Task*

Table 5.4 shows the results of different relation extraction systems on the CDR training and development data, using the gold-standard chemical and disease annotations to generate all possible chemical-disease pairs.

Table 5.4: Performance of different relation extraction systems on the CDR training and development data, given perfect entity annotations

System – features	Threshold*	Recall	Prec.	F-score
Sentence co-occurrence	n/a	0.725	0.313	0.437
Prior knowledge	n/a	0.664	0.405	0.503
SVM – all challenge	0.30	0.840	0.693	0.760
SVM – all post-challenge	0.34	0.854	0.753	0.801
– without prior knowledge	0.33	0.765	0.695	0.728
– without statistical	0.39	0.775	0.683	0.726
– without linguistic	0.38	0.842	0.701	0.765

* Probability threshold for the SVM to decide whether there is a relationship.

A baseline system based on sentence co-occurrence of entities gave an F-score of 0.437 with a recall of 0.725, indicating that more than a quarter of the relations spanned more than one sentence. The application of prior knowledge, assuming that a relation was present if a chemical and a disease were directly connected in the Euretos Knowledge Platform by a non-treatment predicate, resulted in an F-score of 0.503. When the SVM was trained with all the challenge features (i.e., without the negation and word correspondence features that we defined post-challenge), we achieved an F-score of 0.760. Including all our features further improved the F-score to 0.801. To assess the performance contribution of the different features sets, we retrained the system after removing each feature set in turn. Removal of the prior knowledge features or the statistical features resulted in a similar drop of performance (F-scores of 0.728 and 0.726, respectively). Leaving out the linguistic features reduced performance to some lesser extent (F-score 0.765).

Table 5.5 shows the performance results of the SVM classifier, using tmChem and Peregrine for entity normalization, on the CDR test set. For the CDR challenge, we submitted three runs using the SVM trained on the challenge features, in combination with tmChem and the Peregrine challenge system: one run used the decision threshold of 0.30 that resulted from our cross-validation experiments, the other two runs used thresholds of 0.20 and 0.40. The best F-score was 0.569, which was achieved for a threshold of 0.2. This result is higher than the F-score of 0.526 reported in the CDR challenge proceedings.[147,159] The reason is that the server showed occasional race-conditions during the challenge, which we only discovered and fixed after the challenge. Our system, which we named RELigator, ranked second among the systems of 18

Table 5.5: Performance of relation extraction systems on the CDR test data, for challenge and post-challenge SVM features and different entity annotations

SVM features	Chemical NER	DNER – features	Threshold*	Recall	Prec.	F-score
All challenge	tmChem	Peregrine – challange	0.20	0.601	0.540	0.569
All challenge	tmChem	Peregrine – challange	0.30	0.537	0.579	0.557
All challenge	tmChem	Peregrine – challange	0.40	0.467	0.605	0.527
All challenge	tmChem	Peregrine – post-challange	0.30	0.556	0.569	0.563
All post-challenge	tmChem	Peregrine – post-challange	0.34	0.570	0.637	0.602
All post-challenge	Gold standard	Gold standard	0.34	0.731	0.676	0.702

* Probability threshold for the SVM to decide whether there is a relationship.

participating teams in the CDR task (the best team achieved an F-score of 0.570).[147] Use of the improved, post-challenge Peregrine system only slightly improved performance (F-score 0.557 vs. 0.563 at a threshold of 0.3). However, the system trained with the additional post-challenge features yielded a considerably improved F-score of 0.602. For comparison, we also evaluated this SVM using the gold-standard entity annotations. This resulted in an F-score of 0.702.

5.4 DISCUSSION

We described our Peregrine-based system for disease normalization, and the RELigator system for CDR extraction. RELigator achieved an F-score of 0.526 for the CID challenge, which ranked second among 18 participating teams. Several post-challenge modifications of the systems resulted in a substantially improved F-score of 0.602 for CID, currently outperforming the best challenge submission. Evaluation of CID extraction using gold-standard entity annotations illustrates that the quality of entity recognition is still an important limitation.

Regarding the CDR extraction, our results indicate that knowledge-based features, statistical features and linguistic features each contribute to the final system performance, and thus contain at least partly complementary information.

Our original Peregrine system (F-score 0.757) was outperformed in the challenge by CRF-based disease recognition systems, with an F-score of 0.865 for the best performing system. The post-challenge modifications of Peregrine resulted in a substantial performance improvement (F-score 0.828). This result compares favourably with the F-score of 0.698 that we obtained in a previous study in which we also used Peregrine for disease concept recognition in a set of Medline abstracts.[140] The lower performance in that study may partly be explained by the more demanding task to recognize disease concepts from any vocabulary in the UMLS, not just from MeSH like in this study.

Our error analysis revealed that most disease recognition errors were terminology-related. Inclusion of other vocabularies from the UMLS to increase the coverage of synonyms in combination with filtering on semantic types and manual term curation, may further improve Peregrine's performance.

Remarkably, the gain in Peregrine performance before and after the challenge hardly increased the performance of the relation extraction pipeline (F-score rose from 0.557 to 0.563, using the challenge feature set and a decision threshold of 0.3 for the SVM classifier). There may be several reasons for this. First, relation extraction performance is

dependent on the performance of both the disease concept recognition and the chemical concept recognition. Improved disease recognition alone will therefore only be partially reflected in improved relation extraction. Second, disease recognition performance is based on the annotations of all unique disease mentions in the abstracts, whereas relation-extraction performance is based on disease annotations at the document level. The test contains 1,988 gold-standard annotations of unique disease mentions and 865 gold-standard disease annotations that are part of CDRs. Again, improved performance of the disease recognition step is likely to be only partially reflected in improved relation extraction.

Despite the noisy entity data for instance generation, we still performed second in the challenge for CID extraction. Because the performance of relation extraction is not evaluated independently of entity recognition, it is hard to put the CID results into perspective. The task, in-part inspired by the needs of CTD curators, did not distinguish between DNER and CID performance, while this seems essential to bring this task forward.

The inter-annotator agreement (IAA) for the CID corpus is not known. Wiegers et al. reported a surrogate IAA score of 77% for annotation of chemical-gene interactions.[160] This IAA averages agreement of each annotator against a gold standard, created by disagreement resolution, which presumably overestimates the true IAA. Our system has a micro-averaged F-score of 70% using gold-standard annotations, and may come within reach of the IAA. However, formal assessment of CID IAA needs to be performed.

Several improvements of the final model can be envisaged. The scope of syntactically connected chemical–disease pairs could be expanded through anaphora resolution. Governing and relating words could be encoded as word embeddings instead of nominal values, giving them a more compact and semantically rich representation. Simple token features in a window around chemical and disease could provide further context. Finally, the CDR annotations that we used to train our models were provided at the document level. We did not attempt to annotate the relation mentions in the document texts, which might have yielded stronger features.

Part III

VERIFYING VACCINE B/R HYPOTHESES

6

CODEMAPPER: SEMIAUTOMATIC CODING OF CASE DEFINITIONS

ABSTRACT

BACKGROUND Assessment of drug and vaccine effects by combining information from different healthcare databases in the European Union requires extensive efforts in the harmonization of codes as different vocabularies are being used across countries. In this chapter, we present a web application called CodeMapper, which assists in the mapping of case definitions to codes from different vocabularies, while keeping a transparent record of the complete mapping process.

METHODS CodeMapper builds upon coding vocabularies contained in the Metathesaurus of the Unified Medical Language System (UMLS). The mapping approach consists of three phases. First, medical concepts are automatically identified in a free-text case definition. Then, the user revises the set of medical concepts by adding or removing concepts, or expanding them to related concepts that are more general or more specific. Finally, the selected concepts are projected to codes from the targeted coding vocabularies. We evaluated the application by comparing codes that were automatically generated from case definitions by applying CodeMapper's concept identification and successive concept expansion steps using the revision operations, with reference codes that were manually created in a previous study.

RESULTS Automatic concept identification alone had a sensitivity of 0.246 and positive predictive value (PPV) of 0.420 for reproducing the reference codes. Three successive steps of concept expansion increased sensitivity to 0.953 and PPV to 0.616.

CONCLUSIONS Automatic concept identification in the case definition alone was insufficient to reproduce the reference codes, but CodeMapper's operations for concept expansion provide an effective, efficient and transparent way for reproducing the reference codes.

Becker BFH, Avillach P, Romio S, Mulligen EM, Weibel D, Sturkenboom MCJM, Kors JA. CodeMapper: Semiautomatic Coding of Case Definitions. A Contribution from the ADVANCE Project. *Pharmacoepidemiology and drug safety* **26** (2017)

6.1 INTRODUCTION

In order to increase the scale of pharmacoepidemiological studies, information from multiple electronic health record (EHR) databases should be combined in a distributed, collaborative fashion.[46] However, EHR databases use different coding vocabularies to record medical information,[162,163] such as the International Classification of Diseases Clinical Modifications version 9 (ICD-9 CM)[164] and version 10,[165] the International Classification of Primary Care version 2 (ICPC-2),[166] Read codes version 2 (Read-2)[167] and Read Clinical Terms version 3 (CTv3).[168] In multi-database studies, the extraction of an event typically requires several steps to achieve consistency between databases. A case definition that describes the event in the study protocol is translated into an operational definition, which is then mapped for each vocabulary into a set of codes that represents the event. The code sets are combined into queries for case identification and harmonized between databases by comparison with benchmarks from the literature and by feedback from the database custodians.

The creation of code sets for each vocabulary from the textual case definitions has been largely a manual process. Given the number and complexity of the targeted vocabularies, the mapping and harmonization process can pose an important bottleneck to the rapid implementation of collaborative epidemiological studies.[49,50] Furthermore, the rationale for including or excluding individual codes is not consistently documented, which hampers the possible reuse of code sets and queries in subsequent studies.

A previous attempt to accelerate the creation of code sets from multiple vocabularies was made in the EU-ADR project.[49,51,169] Medical concepts like diseases, symptoms, laboratory procedures, or tests were automatically identified in a case definition using the MetaMap program.[170] Code sets representing the concepts in the targeted vocabularies were then generated using the Unified Medical Language System (UMLS),[52] a biomedical terminology system that integrates many vocabularies including coding vocabularies commonly used in EHR databases. Whereas the identification of concepts and their projection to codes was automatic, the overall workflow was not integrated or recorded to facilitate the later reuse of the mapping. The approach was applied also in other European projects like GRIP,[171] VAESCO,[172] and EMIF.[173] Similar collaborative studies in the American, Asian and Pacific regions deal with less heterogeneous medical vocabularies (Mini-sentinel,[174] PRISM,[175] VSD,[176] and AsPEN[177]). Alternatively to adapting the event identification algorithm to the different databases, these databases can be mapped to a standardized coding system. A single event identification algorithm can then be used in different

databases. This approach is being pursued by OMOP and the OHDSI collaboration.[178,179]

We present a web application called CodeMapper, which has been developed in the ADVANCE project (Accelerated Development of VAccine beNefit-risk Collaboration in Europe).[180] It is based on the EU-ADR approach and assists in mapping case definitions to code sets from different vocabularies while keeping a record of the complete mapping process. We evaluate the application by comparing code sets that were automatically generated by CodeMapper with reference code sets that were manually created in a previous epidemiological study.

6.2 METHODS

CodeMapper's mapping approach consists of three phases (figure 6.1, top). First, medical concepts are automatically identified from a free-text case definition. The user can then revise the set of medical concepts by adding or removing concepts, or expanding a concept to more general or more specific concepts. For example, the concept *Coughing* can be expanded to more general concepts such as *Respiratory disorders* and *Abnormal breathing*. Expanding it to concepts that are more specific results in subtypes of coughing such as *Paroxysmal cough* and *Evening cough*. Finally, each concept is represented by (possibly several) codes in the targeted vocabularies, and the projection of the concepts to codes forms the result of the mapping process. In this section, we will describe the mapping approach, the CodeMapper application, and an evaluation of the approach.

6.2.1 *Mapping approach*

CodeMapper builds upon information from the Metathesaurus of the UMLS. The Metathesaurus is a compendium of many medical vocabularies, which have been integrated by assigning equivalent codes and terms from different source vocabularies to the same concepts. Each concept in the UMLS is identified by a CUI. For example, the concept *Coughing* (CUI: C0010200) is among others associated with the codes 786.2 (from ICD-9 CM), R05 (from ICD-10 CM) and XC07I (from CTv3). The Metathesaurus contains more than one million concepts connected to codes from 201 vocabularies. Each concept is assigned to one or more of 127 semantic types, which define broad conceptual categories like *Disease or syndrome*, *Finding*, or *Substance*. To provide even broader structure, semantic types are combined into 15 semantic groups.[154] We used version 2016AA of the UMLS in this evaluation.

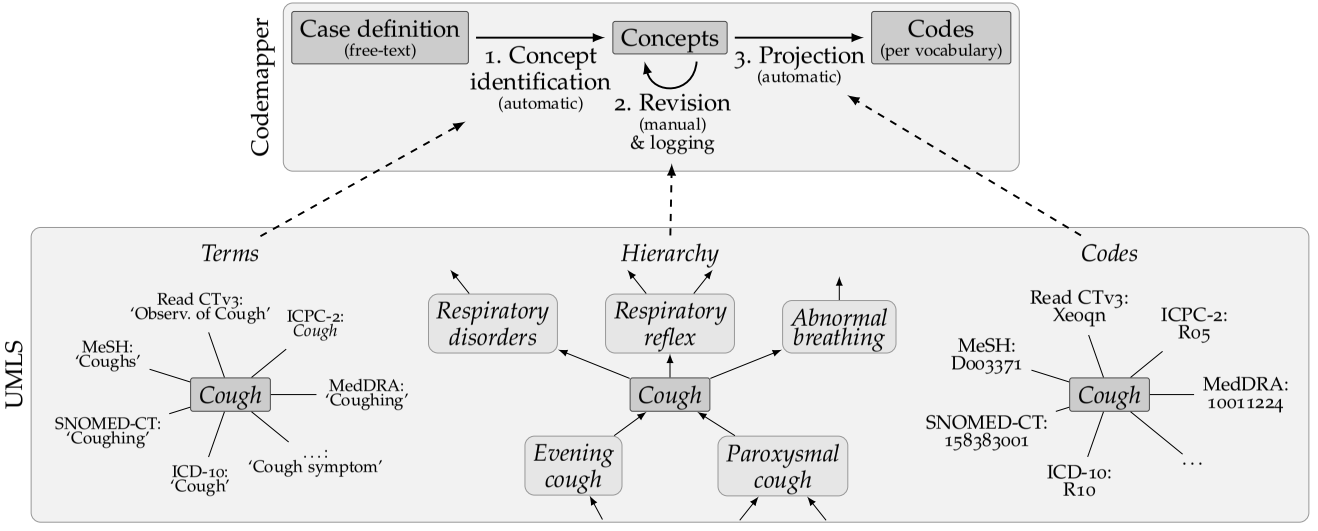


Figure 6.1: Key phases of CodeMapper (top) and the usage of information from the UMLS Metathesaurus, exemplified by the concept for *Cough* with concept unique identifier (CUI) C0010200 (bottom). Terms from the Metathesaurus drive the automatic identification of concepts in the free-text case definition. Hierarchical information about concepts in the Metathesaurus is used to retrieve related concepts during revision of the mapping. Information in the Metathesaurus is used to project the selected concepts to codes from the targeted vocabularies.

The automatic concept identification of CodeMapper is based on lexical information from the Metathesaurus (figure 6.1, left). The lexical information of a concept consists of terms that can be used in free-text to refer to that concept. We compiled a dictionary for the concepts in the semantic groups *Anatomy*, *Chemicals & drugs*, *Disorders*, *Genes & molecular Sequences*, *Living beings*, *Phenomena*, *Physiology*, and *Procedures* of non-suppressible, English terms from the following vocabularies: Medical Subject Headings (MeSH),[30] Medical Dictionary for Regulatory Activities (MedDRA),[118] SNOMED Clinical Terms (SCT),[119] ICD-9 CM, ICD-10 CM, ICPC-2, and CTv3. Our text-indexing engine Peregrine uses this dictionary to identify medical concepts in the case definition.[148]

CodeMapper provides two operations to improve the sensitivity of the mapping by expanding a concept to more general or more specific concepts, based on the hierarchical relationships defined in the Metathesaurus (figure 6.1, centre). Hierarchical relationships connect concepts that are more general or more specific in meaning. For example, the concept for *Coughing* is connected to the more general concept *Respiratory Disorders,* and to the more specific concept *Paroxysmal cough*. To expand a concept in CodeMapper, all concepts that have a more general (or more specific relationship) with it are identified and displayed in the application for selection by the user. Hierarchical relationships in the Metathesaurus are inherited from the source vocabularies or defined in the Metathesaurus.[181] Both types of hierarchical relationships are taken into account for concept expansion.

The projection of concepts to code sets from the targeted vocabularies follows the assignment of codes to concepts in the Metathesaurus (figure 6.1, right).

6.2.2 *Application*

The CodeMapper application is implemented as an open-source web application and freely available for non-commercial use.[182,183] CodeMapper has three screens. On the first screen, the user enters a clinical case definition of an event as free-text. Medical concepts are automatically identified in the text and highlighted inline. By default, only concepts that belong to the semantic group of *Disorders* are preselected for further processing in the application, but the user can select and deselect any identified concept depending on their relevance for the described event.

The second screen displays the mapping as a table with one row for each medical concept, and one column for each targeted vocabulary (figure 6.2 a). Each cell contains the names of the codes that are used to represent the medical concept of the row in the targeted vocabulary

of the column. The codes are displayed when the names are hovered over with the mouse. Several user operations are available for revising the mapping. The user can remove concepts from the mapping, query for and add new concepts, or automatically retrieve more general and more specific concepts. The retrieved concepts are shown in a list and can be selected by the user for inclusion in the mapping. The user can also add or remove vocabularies that should be targeted by the mapping (figure 6.2 b). After every operation, the codes are automatically updated and displayed in the table.

The third screen shows a list of all operations that have been made, for later traceability of the mapping process. When the user saves the mapping, he has to provide a summary of the modifications, which is incorporated into the mapping history. After saving, the mapping and history list are available to other users of the application. Comments can be attached to concepts to capture the discussion about the mapping. Concepts can be categorized by tags, which are inherited by the projected codes. Finally, the user can download the mapping as a spreadsheet file, for example to incorporate the codes into extraction queries. The spreadsheet file comprises the original free-text case definition, the concepts of the mapping, the codes for the targeted vocabulary, and the full history of the mapping process.

6.2.3 *Evaluation*

We evaluated the effectiveness of CodeMapper's approach for creating realistic code sets for a number of case definitions, by comparing code sets that were generated with CodeMapper with manually created reference code sets. We used case definitions and reference code sets from the FP-7-funded SAFEGUARD project, which was conducted in nine EHR databases in the EU and USA. (The full study protocol including the case definitions is available in the EU-PAS registry.[184]) This project was selected for the variety of mapped events and the range of targeted vocabularies. The manual mapping process consisted of deriving operational definition from the textual case definition, choosing codes from the targeted vocabularies without the use of the Metathesaurus, and refining the code set based on feedback from database custodians. The reference mappings also contained exclusion codes, which were not considered in the evaluation because they were generally not derived from the case definitions.

SAFEGUARD studied nine events: acute pancreatitis, bladder cancer, haemorrhagic stroke, heart failure, ischemic stroke, acute myocardial infarction, pancreatic cancer, sudden cardiac death, and ventricular arrhythmia. One event (sudden cardiac death) was excluded from the

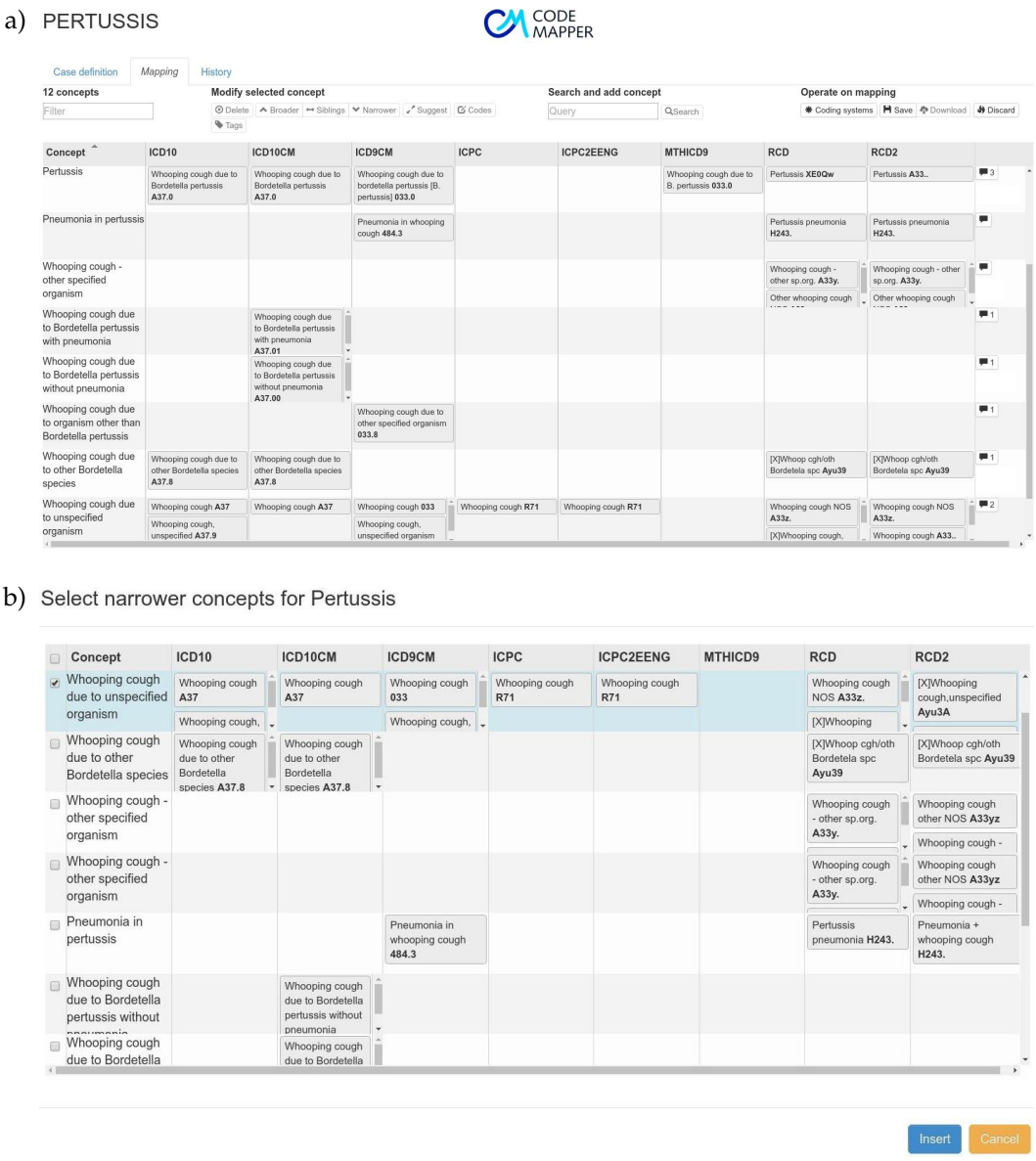


Figure 6.2: a) The second screen of the CodeMapper application shows the mapping as table and provides operations to revise the concepts of a mapping. The cells show the code names from the vocabulary stated in the column that correspond to the concept of the row. Individual codes are shown when hovering the terms. The balloon symbols in the last column indicate the number of comments attached to a concept. b) Example of the operation for concept expansion: The list of concepts that are more specific than *Pertussis* are displayed for the selection and insertion in the mapping.

Table 6.1: Number of words in case definitions and number of codes in the reference set. The numbers of exclusion codes are given in brackets.

Event	Case definition (word count)	Codes ICD-9	ICD-10	ICPC-2[a)]	READ-2
Acute pancreatitis	49	1 (0)	6 (0)	1 (0)	7 (0)
Bladder cancer	87	12 (0)	12 (0)	1 (3)	91 (0)
Hemorrhagic stroke	48	3 (2)	22 (2)	1 (2)	36 (0)
Ischemic stroke	53	10 (0)	11 (0)	2 (1)	20 (0)
Acute myocardial infarction	39	11 (1)	7 (0)	1 (6)	- [b)]
Pancreatic cancer	19	8 (0)	9 (0)	1 (1)	109 (0)
Ventricular arrhythmia	234	5 (0)	5 (0)	1 (1)	27 (0)
Sum	529	50 (3)	72 (2)	8 (14)	290 (0)
Average	75.5	7.1	10.2	1.1	48.3

a) Additional text-based queries for IPCI database

b) Text-based query only for GePaRD database

evaluation because of several missing code sets, and another (heart failure) because the case definition contained only a short symptomatic description of the event, unrelated to the codes representing the event. The events were mapped for nine EHR databases with four vocabularies: Medicare, PHARMO, HSD and regional EHR databases from Lombardy and Puglia (all these databases use ICD-9 CM), GePaRD (ICD-10, German modifications), IPCI and BIFAP (both ICPC-2 and keywords), and CPRD (Read-2). For ICD-9 CM we selected the code sets for Medicare as the reference since it contained less database-specific additions than the other code sets using ICD-9 CM. The codes for GePaRD are contained by the ICD-10 and ICD-10 CM vocabularies in the UMLS, so we combined the codes generated by CodeMapper for these vocabularies. To generate codes for Read-2, a translation table between Read-2 and CTv3 was integrated into CodeMapper, because the Metathesaurus covers only CTv3 and not Read-2. (The mapping table is available at the Health & Social Care Information Centre.) Codes from the IPCI mapping were trimmed to three digits to adjust for the database-specific codes in IPCI.

Overall, the reference code sets contained 529 codes (table 6.1). The size of the reference code sets vary widely between vocabularies: on average, the code sets for Read-2 contain 48.3 codes, whereas the code sets for ICPC-2 contain 1.1 codes. This discrepancy is firstly due to the differences of granularity of the vocabularies (Read-2 has 77,290 codes in the Metathesaurus, ICPC-2 only 1,397). Secondly, the queries to the IPCI database (to which the ICPC-2 code sets were targeted) are

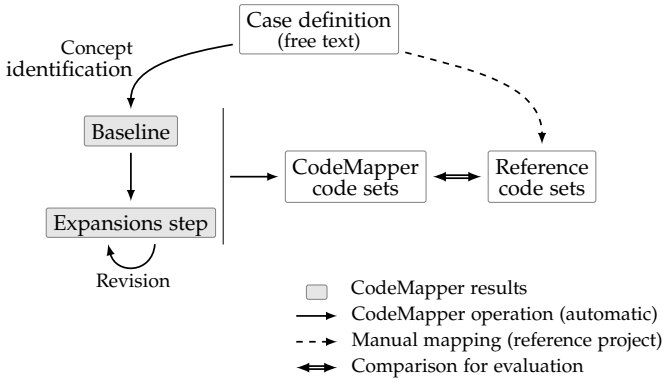


Figure 6.3: Automatic evaluation of CodeMapper. Reference code sets were created manually for each targeted vocabulary from the free-text case definition of an event. The baseline mappings and expansion steps were generated automatically from the same case definition using the operations available in CodeMapper.

supported by keyword searches on the free-text portion of the IPCI medical records and additional exclusion criteria.

Different code sets were generated fully automatically by CodeMapper for the events of the reference project based on the same case definitions. The baseline code sets resulted from the concepts identified in the case definition (figure 6.3). We then simulated the actions of an 'informed user' who seeks to improve the sensitivity of the mapping. We assumed that this user would expand the concepts and, from all possible concepts that are more general or more specific, would only retain those that are relevant to the event. Based on the reference set we were able to automatically simulate the 'informed user's' actions. The resultant set of concepts defined a new code set, which always contained all codes from the preceding code set. We simulated four of these expansion steps on successive concept sets.

For each target vocabulary and event, the generated code sets were compared with the reference code sets. We determined the number of true-positive codes (TP), false-positive codes (FP), and false-negative codes (FN), and computed sensitivity as $TP/(TP+FN)$ and positive predictive value (PPV) as $TP/(TP+FP)$. We report for each vocabulary the sensitivity and PPV averaged over all events in the reference set.

6.2.4 *Error analysis*

We then carried out an automatic error analysis of the false-positive and false-negative codes after the third expansion step (figure 6.4). The definitions of the error categories were based on the notion of sibling

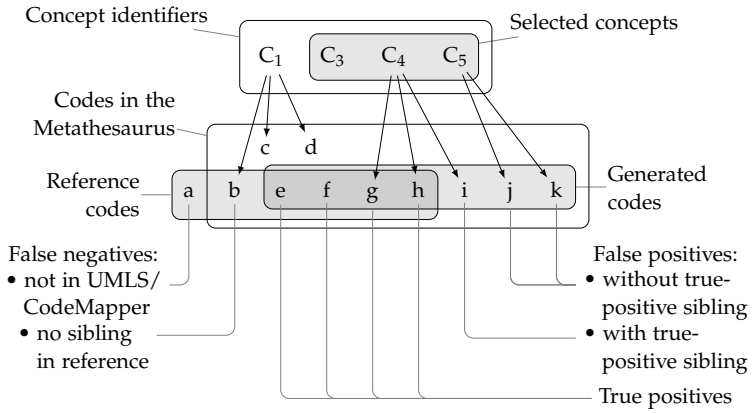


Figure 6.4: Categories of false negatives and false positives in the error analysis. Two codes are siblings if they are associated with the same concept.

codes: two codes are siblings if they are linked to the same concept in the Metathesaurus. For false negatives, we distinguished between codes that are not contained in the Metathesaurus and codes whose siblings are not in the reference sets. False positive codes were categorized as having or not having a true-positive sibling code.

6.3 RESULTS

6.3.1 *Baseline*

The baseline mapping created by CodeMapper comprised 46 concepts and had an average sensitivity of 0.246 for reproducing the reference code sets (table 6.2). The average PPV of the baseline mapping was 0.420. Without filtering by the semantic group of *Disorders*, the number of concepts would increase from 46 to 77 without affecting the sensitivity of the codes sets.

6.3.2 *Concept expansion*

The average sensitivity of the baseline mapping greatly improved in the first expansion step, to 0.818. Sensitivity further increased in the second (0.940) and third (0.953) expansion steps. All ICPC-2 codes were produced after the first expansion step and all ICD-10 CM codes were produced after the second step. The sensitivity increased incrementally for Read-2 and ICD-9 CM. The average PPV improved after one expansion step (0.633) and decreased slightly after two (0.621) and three (0.616) expansion steps. The performance did not improve

Table 6.2: Number of concepts and performance measures of the baseline mapping and after the expansion steps. Numbers per vocabularies are macro-averages over all events.

Revision (concepts)		ICD-9	ICD-10	ICPC-2	READ-2	Avg.
Baseline (46)	Sens.	0.300	0.195	0.357	0.131	0.246
	PPV	0.387	0.380	0.500	0.411	0.420
Step 1 (183)	Sens.	0.858	0.848	1.000	0.568	0.818
	PPV	0.483	0.558	0.762	0.729	0.633
Step 2 (297)	Sens.	0.914	1.000	1.000	0.846	0.940
	PPV	0.463	0.509	0.762	0.749	0.621
Step 3 (335)	Sens.	0.929	1.000	1.000	0.882	0.953
	PPV	0.462	0.498	0.762	0.742	0.616

further in a fourth expansion step. The sensitivity was lower after three expansion steps when using only hierarchical relationships that were either inherited from the source vocabularies (0.928) or defined in the Metathesaurus (0.879).

6.3.3 *Error analysis*

False-positive codes were generated in all vocabularies after the third expansion step (N=234, table 6.3). Most false-positive codes had true-positive siblings (N=164; 70.1%). False-positive codes without true-positive siblings (N=70; 29.9%) resulted from the initial concept identification step because the concept expansion steps that simulated the informed user added only concepts with true-positive codes.

False-negative codes occurred only for the vocabularies Read-2 and ICD-9 CM (table 6.4). Most false negative codes did not have any siblings in the reference set (N=24; 68.6%), suggesting that these codes were added to the reference set due to database specific needs. Other false-negative Read-2 codes were not contained in the conversion table from CTv3 to Read-2 codes, or the CTv3 codes corresponding with the Read-2 codes were not in the Metathesaurus (N=11; 31.4%).

A mapping constructed to maximize sensitivity by selecting concepts to generate all available codes from the reference sets had a sensitivity of 0.991 and PPV of 0.733.

Table 6.3: Number of false-positive codes after three expansion steps by vocabulary and error category, and their percentage of all false-positive codes.

Vocabulary	FP category	Count	Percentage
ICD-9 CM	With TP sibling	52	22.2
	No TP sibling	22	9.4
ICD-10	With TP sibling	66	28.2
	No TP sibling	30	12.8
ICPC-2	With TP sibling	3	1.3
	No TP sibling	1	0.4
READ-2	With TP sibling	43	18.4
	No TP sibling	17	7.3
Overall	With TP sibling	164	70.1
	No TP sibling	70	29.9

Table 6.4: Number of false-negative codes after three expansion steps by vocabulary and error category, and their percentage of all false-negative codes.

Vocabulary	FN category	Count	Percentage
READ-2	No sibling in reference	19	54.3
	Not in UMLS	11	31.4
ICD-9 CM	No sibling in reference	5	14.3
Overall	No sibling in reference	24	68.6
	Not in UMLS	11	31.4

6.4 DISCUSSION

In this chapter, we presented CodeMapper, a web application that assists in the mapping of textual case definitions to code sets from multiple vocabularies, which is often a bottleneck in the implementation of epidemiological multi-database studies. We showed the effectiveness of CodeMapper's approach by simulating an informed usage of the application.

Creating a mapping only by the automatic identification of medical concepts in the case definition was insufficient for reproducing the reference code sets (sensitivity 0.246). The mapping process cannot be replaced by a simple indexing step. However, the goal of CodeMapper

is to support an informed user in creating such mappings, and CodeMapper's operations for concept expansion provide an effective and efficient way for this. The reference code sets were regenerated with a sensitivity of 0.953 and PPV of 0.616 after only three expansion steps. The reference codes for ICPC-2 were even completely regenerated after the first expansion step and the reference codes for ICD-10 CM after the second expansion step. The sensitivity always increased between expansion steps because codes were always retained in subsequent steps. The increase of sensitivity came at the costs of a slight decrease in PPV, which is a consequence of the introduction of false-positive codes that are siblings of newly added true-positive codes. The sensitivity was lower for more granular vocabularies, where more expansion steps were required.

The performance of the mapping that simulated maximal sensitivity (0.991 with associated PPV of 0.733) forms an upper bound of CodeMapper's performance for the presented evaluation. The imperfect sensitivity is due to reference codes that are missing in the UMLS or in the mapping between Read-2 and CTv3. The moderate PPV may be due to inconsistencies in the reference code sets or in the Metathesaurus. The reference code sets may be inconsistent between vocabularies for two reasons. First, the inclusion of one code in the reference mapping did not always imply the inclusion of all sibling codes in the targeted vocabularies, which is reflected by the high percentage of false positives that have true-positive siblings. Second, different code sets were created for databases with the same vocabularies, which can be necessary to compensate for characteristics of the databases. For example, when an event is only available as inpatient diagnosis in one database, a drug that is usually prescribed in case of the event in outpatient setting can be included in the query as a proxy. Such database-specific additions also explain some false-negative codes without siblings in the reference set. Inconsistencies in the Metathesaurus such as missing identification of equivalent codes and incomplete coverage of vocabularies have been discussed before.[185–188]

The SAFEGUARD reference set contained only codes for diagnoses but no codes for laboratory, imaging, or ECG results. CodeMapper's operations for concept expansion would not be suitable for generating such codes for diagnostic tests because the corresponding concepts are not hierarchically related to the concepts for the diagnoses given in the case definition. However, these result codes can be generated using CodeMapper's approach if the concepts for diagnostic tests are mentioned in the case definition.

When exclusion criteria are indicated in the case definition, CodeMapper's approach can be applied to map them to codes, but they must then manually be tagged for exclusion to inform the data extraction

process. Automatic negation extraction could be used to automate the identification of exclusion criteria in the case definition.[189]

The use of the Metathesaurus in CodeMapper's approach brings practical limitations with it. Differences in granularity between vocabularies can affect the consistency of the generated code sets. This problem could be identified by incorporating information about code usage in EHR databases into CodeMapper. Vocabularies that are not part of the UMLS can only be targeted with CodeMapper by integrating additional mapping tables. And database-specific code sets cannot be maintained easily because code sets are generated per vocabulary.

CodeMapper has been applied to the mapping of 45 events in the ADVANCE project so far. The automatic concept identification and revision operations allowed a quick drafting and interactive exploration of the code sets, without requiring extensive knowledge of each targeted vocabulary. Feedback from medical experts and database custodians, and harmonization between databases were crucial to identify missing codes and concepts, and was collected in CodeMapper for subsequent revision of the mappings. Together with the detailed history of all steps that resulted in the mapping, CodeMapper facilitated an integrated and transparent management of the overall mapping process.

In conclusion, the CodeMapper web application constitutes a single entry point for the different phases of the terminology mapping process for multi-database studies. The expansion operations provide a more efficient and systematic way to add relevant related codes to the mapping than browsing the source vocabularies. The integration of the mapping process into a single application, and the recording of user operations make the mapping process traceable, and the mappings more suitable for reuse in subsequent studies.

7

ALIGNMENT OF VACCINE CODES USING THE VACCO ONTOLOGY OF VACCINE DESCRIPTIONS

ABSTRACT

BACKGROUND Vaccine information is represented in European electronic health record (EHR) databases using various clinical and database-specific coding systems and drug vocabularies. The lack of harmonization constitutes a challenge in reusing EHR data in collaborative benefit-risk studies about vaccines.

METHODS We designed an ontology of properties commonly used in vaccine descriptions, VaccO, with a dictionary for the analysis of multilingual vaccine descriptions. Based on the VaccO ontology, we implemented different algorithms for the alignment of vaccine coding systems, i.e., the identification of corresponding codes from different coding systems. The algorithms were evaluated by comparing their results with manually created alignments in two reference sets including clinical and database-specific coding systems with multilingual code descriptors.

RESULTS The best-performing algorithm represented vaccine descriptions as logical statements about entities in VaccO and used an ontology reasoner to identify common properties between corresponding vaccine codes. The evaluation demonstrated excellent performance of the approach (F-scores 0.91 and 0.96).

CONCLUSION The VaccO ontology allows the identification, representation, and comparison of heterogeneous descriptions of vaccine and vaccine groups. The automatic alignment of vaccine coding systems accelerates the readiness of EHR databases in collaborative vaccine studies.

AVAILABILITY The VaccO ontology and three web applications implementing the analysis and alignment of vaccine codes are publicly available.

Becker BFH, Kors JA, Mulligen EM, Sturkenboom MCJM. Alignment of Vaccine Codes Using the VaccO Ontology of Vaccine Properties (Submitted)

7.1 BACKGROUND

The ADVANCE project (Accelerated Development of VAccine beNefit-risk Collaboration in Europe) is building systems to provide best evidence to support decision-making on vaccination in Europe based on the reuse of electronic health record (EHR) data.[109] An important aspect is the extraction of vaccine exposure data from healthcare records across Europe. One of the challenges in reusing EHRs is the lack of harmonization of vaccine information in EHR databases.

A vaccine can be described on different levels: stating the product, its pharmacologic group, or its characteristics in an ontology. The level of description determines which additional information is available about the recorded vaccine. First, a vaccine can be indicated on the level of individual products using its commercial or generic name or using a code from a normalized drug or vaccine terminology. Drug terminologies unify different names of vaccines and provide many product properties, e.g., ingredients and authorizations. Several such drug terminologies exist. The Article 57 database (Art57 DB), issued by the European Medicines Agency, provides information about medical products authorized in Europe, including their composition, indications, and authorization details.[191] RxNorm, issued by the US National Library of Medicine, and the National Drug Codes from the US Food and Drug Administration (FDA) have a comparable scope of information for therapeutic drugs and vaccines authorized in the United States.[192,193]

Second, a vaccine can be recorded more generally by its pharmacologic group, usually indicated by a code from a medical coding system. A code is defined by a descriptor (a short textual phrase) that refers to the vaccine properties that are shared between group members, e.g., the disease or pathogen that a vaccine seeks to prevent ('Influenza vaccines' or 'H1N1 vaccines'), or the vaccine strategy ('attenuated vaccines' or 'inactivated vaccines'; we use the same property names as Plotkin where applicable[1]). Some coding systems possess a taxonomic hierarchy that subordinates codes representing more specific vaccine groups to codes representing more general groups. Only the information stated in the code descriptor and implied by the hierarchy is available about a recorded vaccine. Vaccine codes are defined in several medical coding systems including diagnosis coding systems (e.g., SNOMED Clinical Terms (SCT),[119] Read-2 codes,[167,194] or Medical Subject Headings (MeSH)[195]), drug classification systems (e.g., Anatomical Therapeutic Chemical Classification System (ATC)[196]), and custom coding systems that may be specific for a particular EHR database and possibly using non-English descriptors.[48,197] Some coding systems comprise codes in a taxonomic hierarchy and codes for individual vaccines

(e.g., National Drug File Reference Terminology (NDF-RT)[198,199] and British National Formulary (BNF)[200]).

Third, vaccines can be represented by referring to information from an ontology. An ontology is an unambiguous definition of the entities and relations in a domain ('the explicit specification of a conceptualization').[57,201] The entities in the domain of vaccines may include vaccines (individual products and vaccine groups), immunization targets, ingredients, manufacturers, and market authorizations. Vaccine properties can be inferred from the information available in the ontology. The Vaccine Investigation and Online Network (VIOLIN) maintains the Vaccine Ontology (VO), to date the most comprehensive ontology of immunological information about vaccines, with the objectives of standardizing data and enabling computer-assisted reasoning about vaccines in the United States and Canada.[55] VIOLIN provides several tools for accessing information about vaccines, including vaccine components, mechanisms, vaccine design, and literature.[202,203]

Currently, vaccine benefit-risk studies that utilize vaccine information from EHR databases with different coding systems have to go through a tedious manual semantic harmonization process to align the codes.[48,204] The automatic alignment of vaccine coding systems would accelerate the readiness to obtain information from the EHR databases for vaccine benefit-risk studies.

Various approaches were previously proposed for aligning ontologies in general,[54] medical coding systems,[185,205–209] and drug coding systems.[210,211] These approaches commonly use lexical, instance-based, or hierarchical information about concepts. However, not all approaches are applicable to the alignment of the vaccine coding systems used in EHR databases. Lexical techniques create alignments based on lexical comparison of code descriptors, which is unsuitable for coding systems with descriptors in different languages. For instance-based techniques, the similarity of two classes is asserted by comparing the instances that belong to each class, but coding systems usually do not contain information about the membership of individual products to vaccine codes. Hierarchical techniques employ the taxonomic hierarchy of the ontology, but vaccine coding systems used in EHR databases are often not hierarchically structured.

Codes in general drug coding systems are commonly defined by chemical structure, therapeutic intent, physiologic effect, mechanism of action, and pharmacokinetics.[212,213] The predominant property category for defining vaccine classes is the immunization target (corresponding to the therapeutic intent), but vaccine types (corresponding to the production method) and administration routes, which are used in the definitions of vaccine codes, are uncommon in general drug coding systems. These differences between descriptors in general drug coding

Table 7.1: Categories of properties used to define vaccine groups. A check mark (✓) indicates that a property category (row) is used for defining vaccine codes in a coding system (column).

Prop. category	SCT	Read-2	MeSH	ATC	BNF	AHD
Pathogen	✓	✓	✓	✓	✓	✓
Disease	✓	✓	✓	✓	✓	✓
Strategy	✓	✓	✓	✓		✓
Ingredient		✓	✓		✓	✓
Route		✓	✓	✓		✓
Valence		✓	✓	✓		

systems and descriptors in vaccine coding systems further hamper the transfer of algorithms for aligning drug coding systems to vaccine coding systems.

In this chapter, we describe and evaluate an automatic approach for aligning vaccine coding systems with multilingual code descriptors. For this purpose we developed the VaccO Ontology of Vaccine Descriptions that models properties used in the definition of vaccine codes, which contrast to the immunological properties of vaccines modelled in existing ontologies. Our alignment approach uses VaccO to identify and represent vaccine properties in their descriptions, and an ontology reasoner to identify corresponding descriptions.

7.2 METHODS

7.2.1 *Construction of the VaccO ontology*

A vaccine code in a medical coding system stands for an individual vaccine product or for a pharmacologic group of vaccines. To prepare the creation of the VaccO ontology, we analysed the categories of properties used to define the vaccine groups in a number of general, drug-specific, and custom, database-specific coding systems: SCT, Read-2, MeSH, ATC, BNF, and Additional Health Data (AHD) from the database of the The Health Improvement Network (THIN).

Immunization targets (i.e., vaccine-preventable diseases and their pathogens) were used in all coding systems for the definition of vaccine codes (table 7.1). Vaccine-preventable diseases and pathogens may be used interchangeably to describe equivalent vaccine groups (e.g., 'Vaccine against cervical cancer' and 'Human papillomavirus vaccine'). Vaccine codes were further defined based on vaccine strategies, ingredients (including adjuvants and active ingredients), routes of administration,

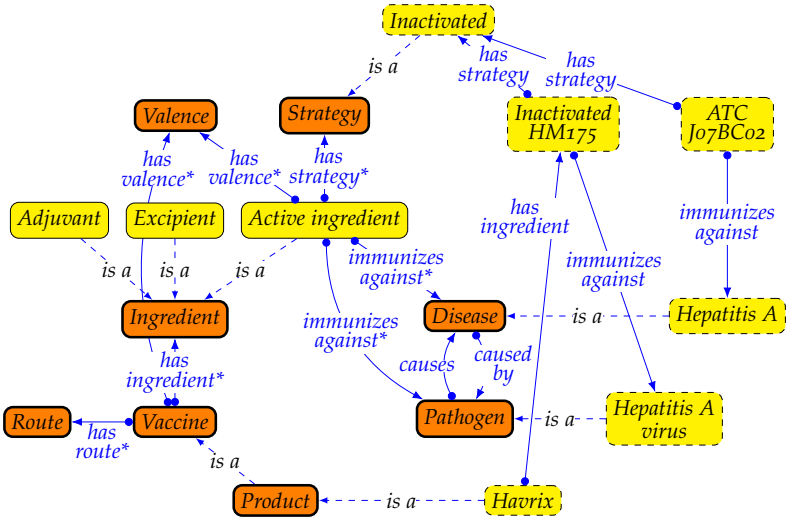


Figure 7.1: Structure of the core VaccO ontology. Fundamental classes are indicated by bold frames. Properties marked with an asterisk are propagated along subclass relations (*is a*) and containment relations (*has ingredient*). Their domains are expanded along the same relations. Also shown are examples for the representations of a vaccine product from the Article 57 database ('Havrix'), and of a vaccine group defined by ATC code J07BC02 ('Hepatitis A, inactivated') with dashed frames. The visualization follows the Graffoo specification.[215]

and valences (which denote either the number of pathogen strains targeted by a vaccine or the number of components in combination vaccines).

The VaccO ontology is specified using the Web Ontology Language (OWL2).[131] Classes in OWL2 are hierarchically structured by the subclass relation (*is-a*) and specified by expressions of description logic (DL) that describe the class properties.[214] For example, a class for influenza vaccines can be defined by the DL expression *Vaccine* that *immunizes-against Influenza*, where *Vaccine* and *Influenza* refer to other classes and *immunizes-against* refers to a property. A class definition can further contain one or more terms that describe the meaning of the class in free text.

To improve readability of the DL expressions, we omit the existential operator (some) in the notation because only existential and no universal object property restrictions are used in VaccO. We use that as a synonym of and in the context of property restrictions. Class names are capitalized, DL keywords are underlined, and names of object properties have lowercase names.

The vaccine property categories, vaccines, and vaccine products comprised the fundamental classes of the VaccO ontology: *Vaccine*, *Valence*, *Route*, *Ingredient*, *Strategy*, *Disease*, and *Pathogen* (see figure 7.1 above). Classes for pharmacological groups and vaccine products were defined as subclasses of *Vaccine*. Subclasses of fundamental classes and their English terms were compiled from the following resources:

- Vaccine products and their ingredients were extracted from the Art57 DB.
- Common pharmacological vaccine groups and their abbreviations (e.g., 'DTaP') were identified in vaccine literature[1,2,4,216,217] and a monograph from the US Centers for Disease Control and Prevention.[218]
- Vaccine strategies and terms were extracted from descriptions in literature, classes in the VO ontology, and vaccine codes in MeSH.
- A general, formalized resource for indication of drugs and vaccines is not publicly available to our knowledge. Classes for pathogens and diseases, and causal relationships between them were instead manually extracted from the descriptions of MeSH headings ('scope notes'). Terms were compiled from the codes that the Unified Medical Language System (UMLS)[52] links to the MeSH headings of pathogens and diseases in the following coding systems: Consumer Health Vocabulary (CHV),[219] International Classification of Diseases Clinical Modifications version 10,[165] Medical Dictionary for Regulatory Activities,[118] MeSH, the taxonomy of the National Center for Biotechnology Information,[220] and SCT.
- Administration routes were identified in the Art57 DB and the VO ontology, and terms (including common abbreviations) were compiled from literature and a monograph of the FDA.[221]
- Classes and terms for valences ('1-valent' up to '30-valent') were created automatically, and common terms for valence 1-10 were added manually (e.g., 'pentavalent').

Relations between classes are expressed in OWL2 using (existential) object properties. An object property is defined by its domain and by its range. For example, the domain of the object property *has-ingredient* is the class *Vaccine* and its range is the class *Ingredient*. Other object properties in VaccO are *immunizes-against* (relating *Vaccine* and *Active ingredient* with *Pathogen* and *Disease*), *has-strategy* (relating *Vaccine* and *Active-ingredient* with *Strategy*), *has-valence* (relating *Vaccine* with *Valence*), and *has-route* (relating *Vaccine* with *Route*, *causes* (relating

Table 7.2: Example inferences in VaccO using property chains

Available information	Property chain	Inferred information
• *v* is-a *Vaccine* that *has-ingredient I* • *I* is-a *Active-ingredient* that *immunizes-against Flu*	• *has-ingredient* ∘ *immunizes-against* ⇒ *immunizes-against*	• *v* is-a *Vaccine* that *immunizes-against Flu*
• *v* is-a *Vaccine* that *has-ingredient I* • *I* is-a *Active-ingredient* that *has-strategy Inactivated*	• *has-ingredient* ∘ *has-strategy* ⇒ *has-strategy*	• *v* is-a *Vaccine* that *has-strategy Inactivated*
• *v* is-a *Vaccine* that *immunizes-against Hib* • *Hib* is-a *Pathogen* that *causes Cervical-cancer*	• *immunizes-against* ∘ *causes* ⇒ *immunizes-against*	• *v* is-a *Vaccine* that *immunizes-against Cervical cancer*

Pathogen with *Disease*), and *caused-by* (relating *Disease* with *Pathogen*). Property chains were defined to propagate properties of ingredients to the containing vaccine, and to unify pathogens and diseases as immunization targets when they are in a causal relation (table 7.2). For example, the property chain *has-ingredient* ∘ *immunizes-against* ⇒ *immunizes-against* states that if a vaccine has an ingredient that immunizes against a specific target (left-hand side), the vaccine immunizes also against the target (right-hand side).

7.2.2 *Representation of vaccine descriptions in VaccO*

The representation of vaccine descriptions in VaccO involves three steps: The identification of vaccine properties in the free-text description, the compilation of the vaccine properties into logical expressions in the ontology, and the normalization of the comprised information as property values.

Identification of vaccine properties in free text

The set of all terms assigned to the classes in an ontology is called the ontology dictionary. The VaccO ontology dictionary constitutes the basis for identifying references to its classes in free text. Each occurrence of a term from the dictionary in an input text is considered a reference to the associated class. We refer to the set of

classes identified in an input text t as $C(t)$. For example, the input text $t =$ 'Live/attenuated influenza vaccine' contains references to the classes in $C(t) = \{Influenza, Attenuated\}$.

We prepared the dictionary of VaccO for multilingual input by automatically translating all English terms using GoogleTranslate to Spanish, Italian, and Catalan (the languages of the vaccine code descriptors in the ADVANCE data sources).[105] The multilingual dictionary is stored in the Apache Solr text search platform, and a Solr plugin for dictionary-based concept identification, Solr TextTagger, is used to identify occurrences of terms from the ontology dictionary in free text.[115,116]

Compilation of vaccine properties into the VaccO class

The representation of vaccine descriptions in VaccO is based on the compilation of a VaccO class c identified in the descriptor to a DL expression describing a vaccine, $[\![c]\!]$. The compilation depends on the category of c and corresponds to c itself if it is a vaccine (a class being a DL expression), or to the class of vaccines with a specific property if c is a vaccine property:

$$[\![c]\!] := \begin{cases} c & \text{if } c \text{ is a } \textit{Vaccine} \\ \textit{Vaccine } \underline{\text{that}} \textit{ has-strategy } c & \text{if } c \text{ is a } \textit{Strategy} \\ \textit{Vaccine } \underline{\text{that}} \textit{ immunizes-against } c & \text{if } c \text{ is a } \textit{Strategy} \text{ or } \textit{Disease} \\ \textit{Vaccine } \underline{\text{that}} \textit{ has-ingredient } c & \text{if } c \text{ is a } \textit{Ingredient} \\ \textit{Vaccine } \underline{\text{that}} \textit{ has-valence } c & \text{if } c \text{ is a } \textit{Valence} \\ \textit{Vaccine } \underline{\text{that}} \textit{ has-route } c & \text{if } c \text{ is a } \textit{Route} \end{cases}$$

For example, the disease class *Tuberculosis* is compiled to the DL expression *Vaccine* that *immunizes-against Tuberculosis*. A set of classes is compiled into the conjunction of the compiled individual classes, $[\![\{c_1, \ldots, c_n\}]\!] := [\![c_1]\!]$ and $\ldots$ $[\![c_n]\!]$.

A description t of a vaccine is represented by the compiled vaccine class, $V(t)$, defined by the result of compiling the classes identified in the description, $[\![C(t)]\!]$. For example, the vaccine class for the descriptor 'Live/attenuated influenza vaccine' is defined by the DL expression *Vaccine* that *immunizes-against Influenza* and *has-strategy Attenuated*.

Normalization to property values

The property values $P(t)$ of a vaccine description t are an assignment of each object property in VaccO (*immunizes-against*, *has-route*, etc.) to all subclasses of the property range that conform to the vaccine description and the information available in VaccO. Formally, the property values

$P(t)$ contain for each property p each subclass c of the range of p, where $VaccO \models [\![C(t)]\!] \sqsubseteq$ *Vaccine* that p c (using the notation by Baader[214]). For example, the property values for the descriptor 'DTwP' are [*immunizes-against*: *Diphtheria*, *Tetanus*, *Pertussis*; *has-strategy*: *Inactivated*].

The compiled vaccine class links information from the vaccine description with information in the VaccO ontology. An ontology reasoner is required to access information implied by the ontology, and the comparison of two compiled vaccine classes can only assess specification, generalization, or equivalence. However, the property values are an explicit representation of all information about a vaccine description implied by the ontology, and they can be compared with each other more flexibly using set-similarity measures. Furthermore, equivalent vaccine descriptions based on pathogens ('Influenza virus vaccine'), disease ('Flu vaccine'), abbreviations ('IIV3'), or products ('Influvac') are normalized to the same property value [*immunizes-against*: *Influenza*].

The representation of vaccine classes and the conversion to property values was implemented in Java using the the OWL2 application programming interface and the JFact ontology reasoner.[222,223]

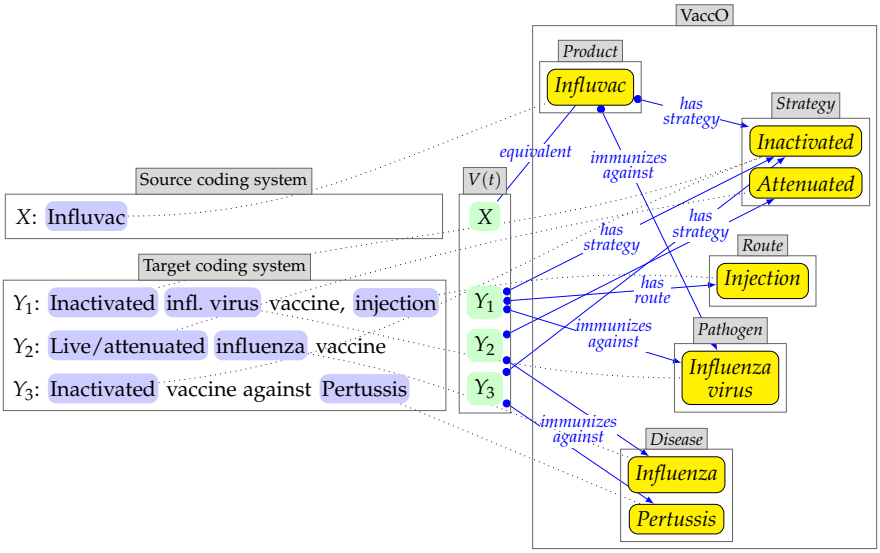
7.2.3 *Automatic code alignment and evaluation*

An alignment between a source coding system and a target coding system assigns each source code to its closest corresponding target code. Our algorithm for creating an alignment first determines for each source code a similarity score with every target code (where 1 indicates maximal similarity and 0 indicates no similarity). The target code with the highest similarity score is assigned to the source code, provided that the score was larger than a preset similarity threshold. If the maximum score does not reach the threshold, no target code is assigned. If multiple target codes have the same maximum similarity score larger than the threshold, all target codes are assigned unless the target coding system has a taxonomic hierarchy. In that case, only the most general target codes with maximum similarity are assigned.

Alignment methods

We evaluated our alignment algorithm using two baseline similarity methods and three similarity methods involving the representation of vaccine descriptions in VaccO as described in section 7.2.2 above. Example alignments for the VaccO-based methods are shown in figure 7.2.

- Method *Tokens* implemented a simple lexical technique. Each code descriptor was tokenized, and the similarity between two



Representation		X	Y_1	Y_2	Y_3
VaccO classes $C(t)$	Product:	*Influvac*			
	Pathogen:		*Infl. virus*		
	Disease:			*Influenza*	*Pertussis*
	Strategy:		*Inactivated*	*Attenuated*	*Inactivated*
	Route:		*Injection*		
Compilation $[\![C(t)]\!]$		*Influvac*	*Vaccine that imm.-ag. Infl. virus and has-strat. Inactivated and has-route Injection*	*Vaccine that imm.-ag. Influenza and has.-strat. Attenuated*	*Vaccine that imm.-ag. Pertussis and has-strat. Inactivated*
Property values $P(t)$	imm.-against:	*Influenza*	*Influenza*	*Influenza*	*Pertussis*
	has-strategy:	*Inactivated*	*Inactivated*	*Attenuated*	*Inactivated*
	has-route:		*Injection*		

Figure 7.2: Example for the compilation of vaccine code descriptors t into classes in VaccO. Above: VaccO classes are identified in the code descriptors (blue boxes in the source and target code descriptors on the left) and compiled into vaccine classes (green boxes X, Y_1, Y_2, and Y_3 in the centre, $V(t)$). Below: Representation of the vaccine descriptors in the VaccO similarity methods. The classes identified in the descriptor of code X do not overlap with those in the descriptors of codes Y_1, Y_2, or Y_3, and the DL-expressions are not equivalent, resulting in a similarity of 0 for similarity methods *Classes* and *Equivalence*. However, property values of code X and the target codes overlap, and X is assigned to code Y_1, which has maximal similarity with X ($Y_1 : 0.5, Y_2 : 0.3, Y_3 : 0$).

codes was measured by the Jaccard coefficient of the two sets of tokens. The Jaccard coefficient of two sets s and t is defined as $|s \cap t| \,/\, |s \cup t|$.

- Method *Metamap* used the MetaMap program to identify UMLS concept unique identifiers (CUIs) for each code descriptor, abstracting over word inflections and synonyms.[170] MetaMap used a dictionary of English terms, and thus can only find concepts in English text. Similarity was defined by the Jaccard coefficient of the two sets of CUIs.

- Method *Classes* represented a code with descriptor t as the set of classes identified in the code descriptor, $C(t)$. Similarity was defined by the Jaccard coefficient of the classes of the source code and the classes of the target code.

- Method *Equivalence* represented a code with descriptor t by the compiled vaccine class, $V(t)$. Similarity between two codes was 1 if their compiled vaccine classes are equivalent and 0 otherwise. Assessing equivalence involved information implied from the VaccO ontology and is checked using the ontology reasoner.

- Method *Properties* represented a code with descriptor t by its property values, $P(t)$. The similarity between a source code and target code was defined as 0 if the values of property *immunizes-against* differed, and by the overlap between the property values otherwise. The overlap was defined as the Jaccard coefficient between the property values.

Reference mappings

To evaluate our code alignment algorithm, we used two reference sets with manually curated alignments (table 7.3). The first reference set used the Vactype coding system as a target. Vactype was developed as a pragmatic solution to harmonize the vaccine descriptors in the databases that participated in an early vaccine studies of the ADVANCE project.[204] It used English descriptors, and currently comprises 43 codes (for 28 single immunization targets with strategies, and 15 combinations). The *Vactype* reference set used five custom vaccine coding systems with multilingual descriptors from European EHR databases as source coding systems: the Catalonian Information System for Research in Primary Care (SIDIAP) with Catalan descriptors,[224] the Spanish Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria (BIFAP) with Spanish descriptors,[225] the Italian paediatric database Pedianet with both English and Italian descriptors,[226] and the regional primary care database of Venetia with Italian descriptors.

Table 7.3: Vaccine coding systems, languages, and number of source codes in the reference sets

Target	Source	Language	Codes
Vactype	Vactype	English	43
	BIFAP	Spanish	761
	SIDIAP	Catalan	98
	Venetia	English	21
	Pedianet-en	English	9
	Pedianet-it	Italian	9
ATC	ATC	English	114
	NDF-RT	English	40
	CHV	English	26
	MeSH	English	23
	VANDF	English	18
	CVX	English	18

The alignments in the *Vactype* reference set were manually created and validated by the database custodians in a project proof-of-concept study of the ADVANCE.[204]

The second reference set comprised alignment from the UMLS to the ATC target coding system. As of 2017, the ATC system contained 114 vaccine codes (with prefix J07). The coding systems with the largest number of mappings to vaccine ATC codes in the UMLS were used as source coding systems in the ATC reference set: Veterans Affair National Drug File (VANDF), MeSH, CHV, Vaccine Administered (CVX), and NDF-RT. We corrected 17 code assignments where the source codes were not assigned to the most specific, corresponding ATC code in the UMLS (the corrections are available in table S1 of the online supplementary material[69]).

Reflexive alignments in which either Vactype or ATC was both the source coding system and the target coding system were included in the evaluation to assess the completeness of the intermediate representation used by the different similarity methods.

Performance measures

The comparison of an automatically generated alignment with a reference alignment is based on the number of correctly generated assignments (true positives, TP), the number of incorrectly generated assignments (false positives, FP), and the number of reference assignments that were not generated (false negatives, FN). The performance of a

Table 7.4: Number of classes and terms in the VaccO ontology

Fundamental class	Classes	Terms
Ingredient	497	505
Vaccine	321	706
Pathogen	104	863
Disease	49	759
Valence	30	71
Strategy	9	35
Route	9	23
Total	1,019	2,962

generated alignment was assessed by its precision (TP/ (TP + FP)), recall (TP/ (TP + FN)), and F-score (2 ∗ precision ∗ recall / (precision + recall)). We also report the average performance measures over all source coding systems in each reference set (excluding reflexive alignments).

7.3 RESULTS

The VaccO ontology contained 321 vaccine classes with 706 terms (table 7.4) including vaccine products (206 classes), common pharmacological groups (36), and auxiliary classes corresponding to immunization targets (e.g., *Pertussis vaccines*), administration route (e.g., *Oral vaccines*), and vaccine strategy (e.g., *Attenuated vaccines*). The 497 classes for ingredients were categorized as active ingredients (310 classes), excipients (170), and adjuvants (21; some ingredients serve multiple roles in the underlying Art57 DB). Classes for nine vaccine strategies with 34 terms were created: *Live/attenuated*, *Conjugated*, *Subunit*, *Inactivated*, *Polysaccharide*, *Recombinant*, *Synthetic*, *DNA*, and *Toxoid*. The 104 classes for pathogens contain 863 English terms. Pathogens were categorized by their biological domain: *Bacteria* (56 classes), *Viruses* (42), and *Protozoa* (6), including 42 classes for pathogen strains. VaccO defines 49 classes for diseases with 759 terms, 30 valence classes with 71 terms, and 9 classes for administration routes with 23 terms.

7.3.1 *Automatic code alignment*

Figure 7.3 shows the performance results of our alignment algorithm with different similarity methods in the two reference sets. These results were generated with a similarity threshold of 0.1. This threshold proved

to yield the highest average F-score over all alignments when we varied the threshold between 0 and 1 in steps of 0.1 (performance measures for all thresholds are available in table S2 of the supplementary online material[69]).

The F-scores of the reflexive alignments were higher than 0.99 on the *Vactype* reference set and higher than 0.93 on the *ATC* reference set. The reason for the slightly lower performance on the *ATC* reference set is that codes for residual classes cannot be represented in OWL2 (e.g., J07BX with descriptors 'Other viral vaccines') and some ATC codes are defined without reference to specific vaccine properties (e.g., J07 for 'VACCINES', J07BC20 for 'Combinations'). Overall, the reflexive mapping results indicated that the intermediate representations are capable of representing the descriptors of the target coding systems.

The baseline methods *Tokens* and *Metamap* performed poorly in the *Vactype* reference set with non-English descriptors because they were not designed to deal with multilingual input. On the *ATC* reference set, with only English code descriptors, their performance was higher. The other three methods, which used the multilingual VaccO dictionary, performed better on the *Vactype* reference set, with method *Properties* performing best for each source coding system (average F-score 0.91).

The performance was generally higher on the *ATC* reference set than on the *Vactype* reference set. Only method *Classes* performed worse on the *ATC* reference set than on the Vactype reference set, because a large variety of properties was used in the code descriptors for the same vaccine groups in the *ATC* reference set (e.g., 'Flu vaccine' vs. 'Influenza virus vaccine'). These different descriptors were represented by different sets of VaccO classes, resulting in little similarity. The performance of methods *Equivalence* and *Properties* was less vulnerable to the variety of descriptions. Overall, method *Properties* performed best (average F-score 0.96) in the *ATC* reference set.

Changing the similarity threshold resulted in lower F-scores, but increased precision or recall. With a threshold of 0.1, the F-score of method *Properties* averaged over all alignments in both reference sets was 0.93, with a precision of 0.94 and a recall of 0.92 (table S3 of the online supplementary material[69]). A threshold of 0.0 decreased precision to 0.81 and increased recall to 0.95 (F-score 0.85). A threshold of 1.0 increased precision to 0.97 and decreased recall to 0.78 (F-score 0.86).

7.3.2 *Error analysis*

We analysed the errors made by method *Properties* (with a similarity threshold of 0.1) to identify remaining problems. For each pair of

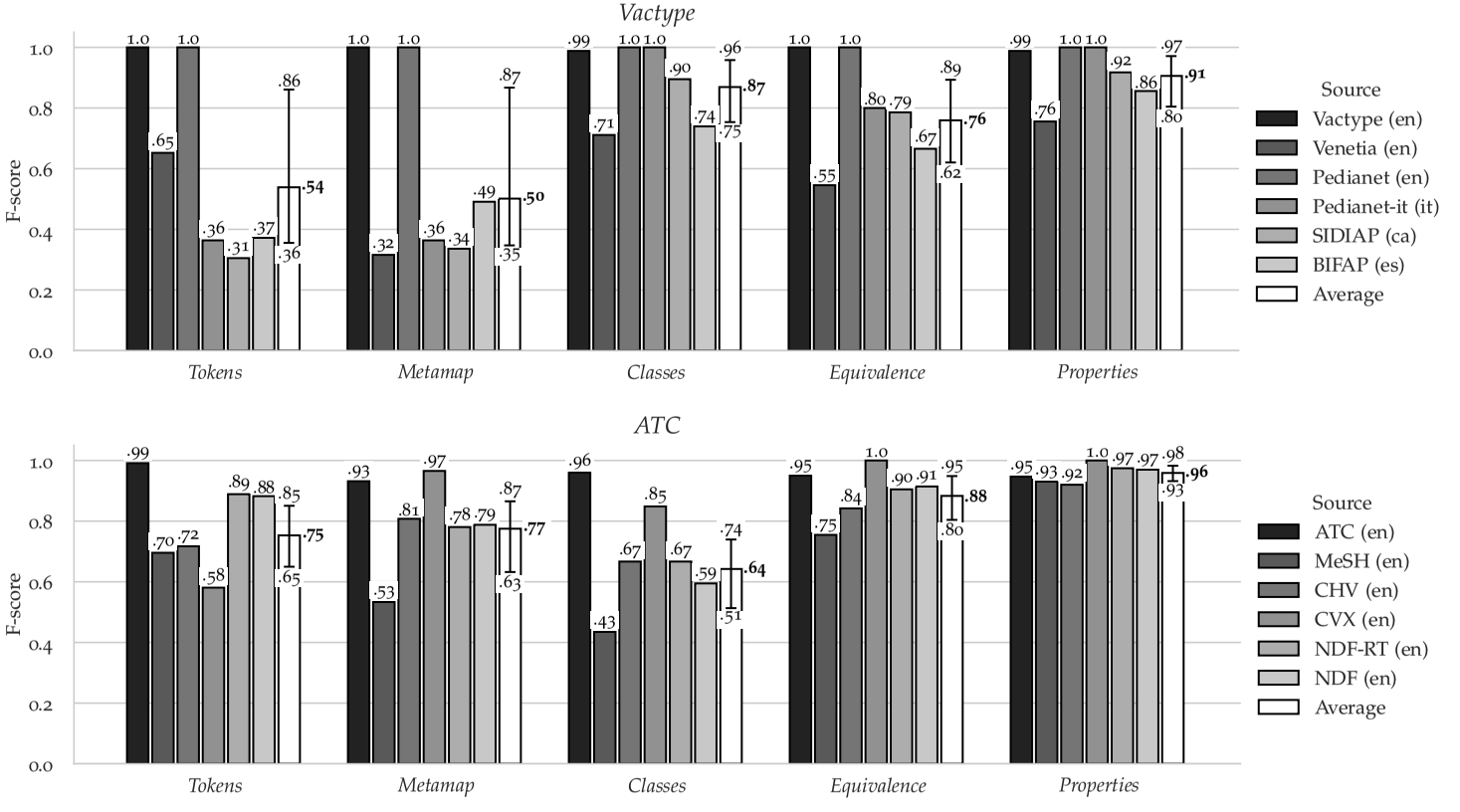


Figure 7.3: F-scores of our alignment algorithm using five similarity measures in the *Vactype* reference set (above) and in the *ATC* reference set (below). Black columns show the performance of the reflexive alignments and white columns show the average performance of the non-reflexive alignments with 95% bootstrap confidence intervals.

Table 7.5: Error analysis of automatic code alignment using the *Properties* method with a threshold of 0.1.

	Vactype		*ATC*			
	FN	FP	FN	FP	Total	%
Incorrect class identification	14	6	4	1	25	39.0
Lack of contextual knowledge	7	14	0	0	21	32.8
Incomplete representation	1	0	12	5	18	28.1

source and target coding systems, we took all alignment errors. If there were more than 10 false positive (FP) or false negative (FN) errors we sampled 10 FP errors and 10 FN errors. A total 64 errors were considered and categorized in different sources of error.

The largest error source was the incorrect identification of classes in the code descriptors, mostly in the multilingual *Vactype* reference set (table 7.5). These errors were caused by missing or ambiguous terms in the ontology dictionary. A second source of error in the *Vactype* reference set, was the lack of contextual knowledge in VaccO about the availability of vaccines. This knowledge had been used in creating the Vactype reference alignments, e.g., knowledge that only acellular vaccines are authorized was used to assign the source code of 'Pertussis vaccine' to the target code 'Acellular pertussis vaccines'. The lack of contextual knowledge gave rise to the low performance of all methods in the Venetia source coding system. Thirdly, incomplete representation in the similarity method was a large error source in the *ATC* reference set. This includes errors where two target codes are semantically identical (e.g., ATC codes J07B for 'Viral vaccines' and J07BX for 'Other viral vaccines'), where properties in the code descriptor do not correspond to classes in VaccO (J07AH06 for 'meningococcus B, outer membrane vesicle vaccine'), or where codes are not defined based on specific vaccine properties (J07 for 'VACCINES', J07BC20 for 'combinations').

7.3.3 *Web applications*

Three web applications accompany the VaccO ontology. The application *Analyse* allows the user to enter a vaccine description and displays the identified classes, compiled DL-expression, and property values (similar to figure 7.2). The application *Selector* analyses a user-provided vaccine coding system, and enables the user to select codes based on their VaccO vaccine properties. The application *Alignment* allows the user to upload two arbitrary vaccine coding systems and generates and

displays an alignment between them using the algorithm described above.

The VaccO ontology and three web applications implementing the analysis and alignment of vaccine codes are publicly available and open source.[114,227]

7.4 DISCUSSION

In this chapter we described VaccO, an application ontology for representing vaccine descriptions, and an algorithm for the automatic alignment of vaccine codes between general clinical and database-specific vaccine coding systems using multilingual code descriptors.

The alignment of vaccine coding systems presents three major difficulties: multilingual code descriptors, the use of different properties to describe the equivalent vaccine classes (e.g., by disease as in 'Flu vaccine' or by pathogen as in 'Influenza virus vaccine'), and differing granularities of the source and target coding system. Our reference sets presented these difficulties by comprising code descriptors in English, Spanish, Italian, and Catalan, and contained general medical coding systems, drug coding systems, and custom database coding systems. The balance between precision and recall of the *Properties* method can be shifted by changing the similarity threshold. Use of a lower threshold to maximize the recall could be advantageous in the use case where the automatically generated alignments are manually validated, as the removal of false-positive alignments generally requires less effort than the manual detection of missing, false-negative alignments and the selection of the correct codes.

VaccO is agnostic of any specific vaccine coding system and designed to represent the descriptors of any vaccine coding system. For this reason, the fundamental vaccine class subsumes no vaccine codes but only auxiliary classes, classes representing common vaccine abbreviations, and vaccine products. Vaccine products are included in VaccO to derive their properties when comparing code descriptors based of products with descriptors of pharmacological groups. VaccO focuses on European vaccines with its integration of the Art57 DB. Integration of other vaccine vocabularies could be used to change the geographical focus (e.g., RxNorm[228] for the United States or databases implementing ISO standard for the Identification of Medicinal Products (IDMP)[229]).

The presented VaccO ontology and the VO ontology[55] model the domain of vaccines. The two ontologies, however, are designed from different points of view: VO models vaccine products and their immunological properties, whereas VaccO models properties to describe vaccines in coding systems. Classes in VaccO and VO coincide where

vaccine descriptions correspond to immunological properties of vaccine products, e.g., with respect to pathogens and ingredients. Differences between VaccO and VO result from the following deviations of properties in vaccine descriptions from the immunological properties of vaccine products:

- Vaccine descriptions can be based on derived properties, which are not represented in VO (e.g., diseases and vaccine strategies derived from pathogens and ingredients, respectively).
- A vaccine immunizes against a pathogen, whereas a vaccine descriptions use pathogens and their corresponding vaccine-preventable diseases interchangeably. This ambiguity conflicts with the definition of the property *'vaccine immunization against microbe'* in VO and is modelled in VaccO by incorporating diseases and their causative relation with pathogens and by permitting pathogens and diseases in the range of property *immunizes-against*.
- Vaccine descriptions imprecisely attach properties to vaccines, e.g., a vaccine can be described by a strategy, whereas the strategy is actually a property of one of its active ingredient. VaccO models equivalences between such imprecise descriptions using property chain rules.

The therapeutic role of a drug is usually treated as an intent in biomedical ontologies. However, the differentiation between the intended and factual therapeutic role is unessential for representing descriptors of vaccine coding systems (e.g., all 2018 flu vaccines fall under the description 'influenza vaccines' even if not all instances immunize against the disease) and not modelled by the property *immunizes-against*.

VaccO was designed as an application-ontology for our code alignment algorithm. The algorithm did not require the integration of VaccO with other ontologies such as VO or an upper-level ontology. But VaccO is based on the OWL2 standard, which facilitates a technically simple integration with other ontologies when required.

7.5 CONCLUSION

The proposed method for aligning vaccine coding systems performed excellently on a wide range of vaccine coding systems using different languages, which suggests broad applicability of the approach. The automatic alignment of vaccine coding systems can accelerate the readiness of EHR databases in collaborative vaccine studies. The alignment method demonstrated the use of an application ontology to

identify and represent vaccine descriptions, and the use of an ontology reasoner to comparing them. VaccO could likewise be applied to the extraction of vaccine-related information from other free-text resources, e.g., literature, spontaneous reports, public news, or scientific literature (see chapter 4).

8 GENERAL DISCUSSION

Post-marketing management and decision-making about vaccines builds on the early detection of safety concerns and changes in public sentiment, the accurate access to established evidence, and the ability to promptly quantify effects and verify hypotheses about the vaccine benefits and risks (B/R). A variety of resources provide relevant information but they use different representations, which makes rapid evidence generation and extraction challenging. This thesis presented automatic methods for interpreting heterogeneously represented vaccine information (figure 8.1). Part I used automatic methods for the first step in vaccine semantics, the recognition of information, whereas the other steps were accomplished manually. Parts II and III developed and evaluated automatic methods and resources for the recognition, representation, and reasoning about vaccine-related information.

Chapter 2 was motivated by the question whether public social media capture information that is relevant for the post-marketing management of vaccines. The study provided negative evidence regarding the use of social media for monitoring vaccine safety, but chapter 3 suggested that changes in public attention and sentiment about vaccines and vaccination programmes could be monitored in public social media.

Part II introduced methods that underlie the automatic extraction of established evidence about vaccines from scientific literature. Chapter 4 addressed the recognition of vaccine descriptions and the classification of articles by vaccines, two basic tasks in extracting vaccine-related information. Our automatic method to recognize vaccine descriptions based on the VaccO ontology performed reasonably well in comparison with the inter-annotator agreement in a manually created reference corpus of vaccine descriptions. A convolutional neural network (CNN) performed best for categorizing vaccine literature but required substantial training material for creating a well-performing classifier, which may not be available for novel or less-studied vaccines. Chapter 5 presented a model for the automatic extraction of causal relations between chemicals and diseases from scientific literature.

Part III contrasted two automatic approaches to harmonizing extraction queries for vaccines and outcomes from EHR databases that use different medical coding systems, which is often a bottleneck in the implementation of collaborative observational studies. Chapter 6 introduced the web application CodeMapper that implements a semi-automatic approach to mapping textual case definitions to database-

specific extraction codes. CodeMapper was evaluated by simulating an informed usage, which showed the efficacy of its user operations. Chapter 7 presented an approach for solving heterogeneous representation of vaccine information in EHR databases by automatically creating alignments between database coding systems. The algorithm for code alignment is based on the newly created ontology of vaccine descriptions, VaccO. The automatically generated alignments between vaccine coding systems were compared with manually created alignments, which demonstrated excellent performance of the approach.

IMPROVING INFORMATION AVAILABILITY FOR POST-MARKETING MANAGEMENT OF VACCINES

Most of the presented work has been developed in the context of the ADVANCE project, which has been launched in 2013 to bring together the different steak-holders of post-marketing management of vaccines.[180] The thesis at hand offers a number of insights and applications.

We observed that public social media is not very helpful for identifying vaccine safety signals. Public social media about vaccines mostly witnessed the perception of other internet content rather than that they expressed private experience with vaccines. Spontaneous reporting systems may be better suited for providing early cues to vaccine safety issues,[230] but they suffer from other drawbacks such as under-reporting and unverified reports.[63,64,231] Public social media messages, however, proved valuable for monitoring the public sentiment about vaccines. Furthermore, social media are a powerful tool in shaping public sentiment, as recent political campaigns in the UK and US suggested. The next question to ask is whether social media can be harnessed to manage a public sentiment crisis about a vaccine by targeted dissemination of available evidence about vaccines.[232]

The use of social media for vaccine safety surveillance or public sentiment monitoring has several limitations. The analysis of social media constitutes a bias towards relatively young people with internet access.[89] Our first study had an additional bias towards an English-speaking communities due to the query keywords that were used to retrieve relevant messages. And because social media platforms are continually being re-engineered to improve the commercial service, there is the concern as to whether studies conducted on data collected from these platforms are reproducible.[92]

Automatic extraction of vaccine-related information from scientific literature provides rapid access to available evidence. The accurate recognition of vaccine descriptions is fundamental to the extraction of any

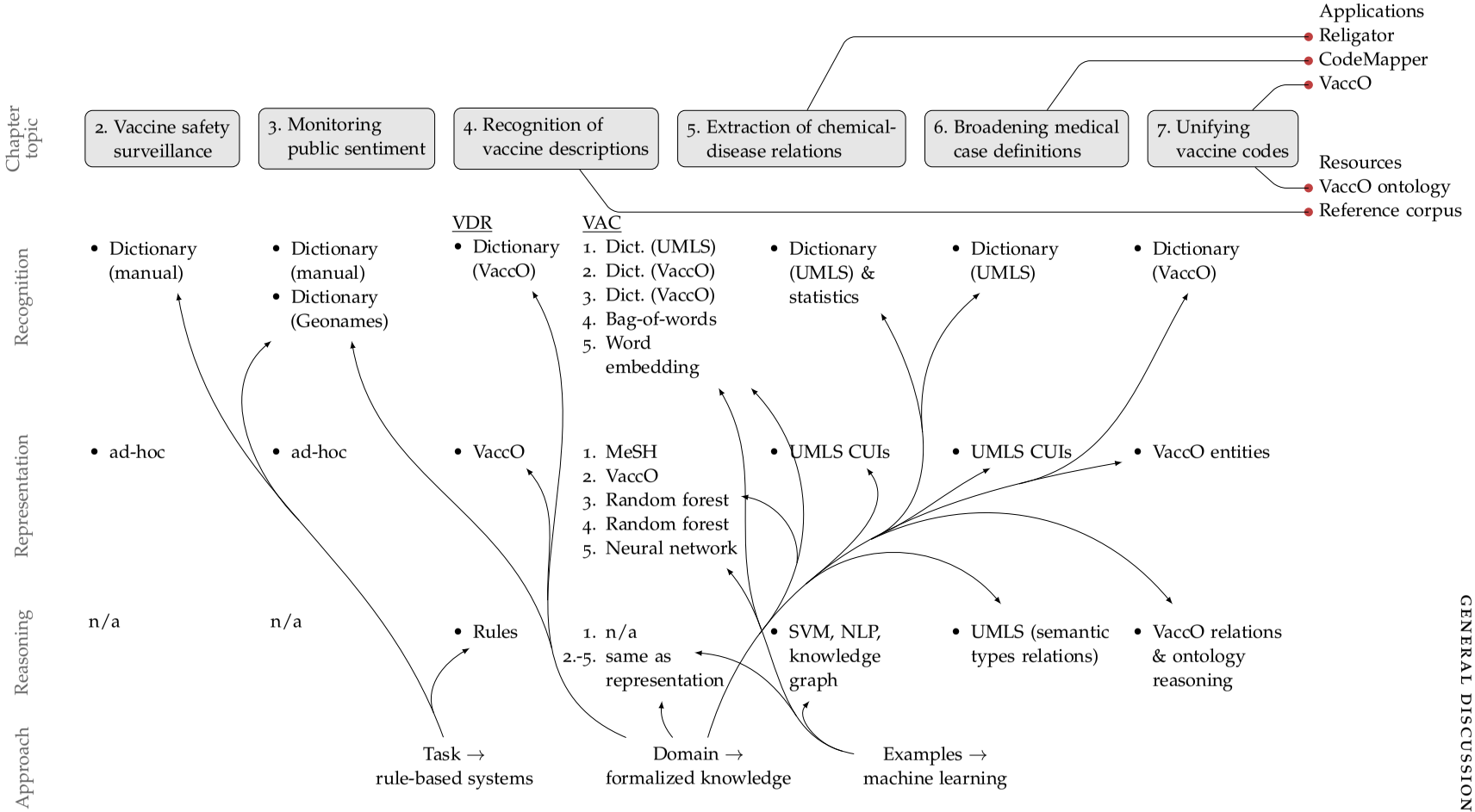


Figure 8.1: Methods applied in this thesis for the different steps of vaccine semantics (and continuation of figure 1.3)

relational information about vaccines. The RELigator model for extracting chemically-induced diseases could be specialized in the extraction of vaccine adverse events by narrowing its scope of causal agents from chemicals to vaccines using the recognition of vaccine descriptions. Automatic extraction of vaccine-related information from scientific literature is applicable in two settings. On the one hand, available evidence about a vaccine can be compiled on request, for example to rapidly assess a vaccine safety signal or to communicate evidence in order to raise public confidence into a vaccination program. On the other hand, automatic extraction could be used for creating and maintaining a comprehensive inventory of established evidence about vaccines in general. Partial and manually created attempts exist,[35] but a complete and up-to-date inventory would be useful for researchers for fast information access, or for health professionals and the public to inform individual vaccination decisions. The transfer of the proposed methods for information extraction from scientific literature to other free-text resources of established vaccine evidence remains to be assessed (e.g., clinical trial reports or summary of product characteristics).

Several computer applications resulted from this thesis and can assist the implementation of collaborative observational studies about vaccine B/R by solving the representational heterogeneity of medical outcomes and vaccines in EHR databases. The CodeMapper application has been evaluated or used in a number of research projects and industry (ADVANCE, European Medical Information Framework (EMIF),[173] National Vaccine Program Office (NVPO),[233] and Entresto®) since its inception. It has, for example, been applied in the ADVANCE project for the mapping of 45 outcomes. The automatic concept identification and revision operations allowed a quick drafting and interactive exploration of the code sets, without requiring extensive knowledge of the targeted vocabularies. Feedback from medical experts and database custodians, and harmonization between databases were crucial to identify missing codes and concepts, and was collected in CodeMapper for subsequent revision of the mappings. Together with the detailed history of all steps that resulted in the mapping, CodeMapper facilitated an integrated and transparent management of the overall mapping process.

In addition to knowledge of database codes, terminology mapping requires clinical knowledge about the event, epidemiological knowledge including an understanding of the study goals, and knowledge about the representation of medical events in the databases, which are introduced during the manual revision in CodeMapper's approach. According to the experience in the ADVANCE project, terminology mapping of the events can be performed most efficiently by smaller teams where single members contribute expertise in several of these topics. Each team should be led by the principal investigator of the

study who should have expertise in extraction of events from healthcare databases, has the objective of the study in view, and takes final decisions about the exclusion or inclusion of medical concepts based on the input of the event team.[234]

CodeMapper has several limitations, most of them owing to the basic design decision that the manual revision operates only on medical concepts, and code sets are automatically generated based on the assignments of codes to concepts in the UMLS. This design decision was taken to ensure equivalence between the code sets generated for different coding systems, assuming consistent assignments in the UMLS. As a result, CodeMapper cannot create extraction queries for databases that use custom coding systems not included in the UMLS or free text to represent medical information. And inconsistencies in the UMLS may crop up as inconsistencies between the code sets generated by CodeMapper.[185–188]

CodeMapper's approach could be developed further into two directions. First, mapping a medical outcome using CodeMapper starts with an analysis of a textual case definition. Alternatively, a mapping could likewise be built upon an existing code set that has been used before to extract the outcome from a single database. CodeMapper's engine for identifying medical concepts also recognizes codes, which would permit an experimental evaluation if the proposed workflow could be improved by using an existing code set instead of a textual case definition as a starting point. Second, a discrepancy between medical case definitions and EHRs impeded the creation of database extraction queries using CodeMapper's approach in practice. Whereas the purpose of medical case definitions is to establish a diagnosis based on signs and symptoms under full availability of information (i.e., by examining a patient), EHRs contain only information about a medical event that is relevant for the purpose of the database (e.g., final diagnoses in primary care databases, or medications in reimbursement databases). Consequently, extraction queries have to take the specific purposes of the databases into account. This could be realized by defining for each medical event several operational case definitions that describe the event with respect to the different database purposes. CodeMapper's approach can then be used to broaden the operational definitions to code sets used in the EHR database of the same types. Gini et al., for example, created separate mappings with CodeMapper to generated different code sets for EHR databases of different types.[235]

The VaccO ontology provides the basis for recognizing, representing, and reasoning about vaccine descriptions, and can be applied to the extraction of vaccinations from EHR databases by means of three web applications.[114] The application 'Analyse' allows the user to enter a vaccine description (e.g., a code descriptor) and presents the properties

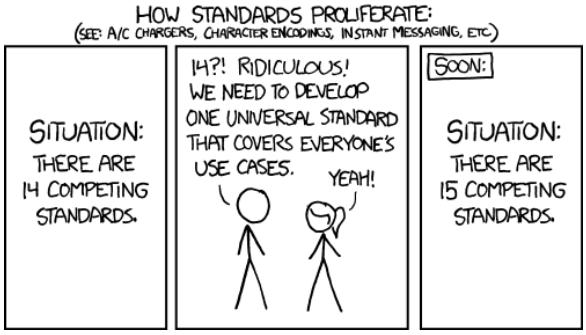


Figure 8.2: An XKCD comic hits the nail on the head once again (source: *https://xkcd.com*)

in the description derived using the domain-specific knowledge from the VaccO ontology. The application 'Selector' analyses all codes of a user-provided vaccine coding system and allows for filtering the vaccine codes by their derived properties. The application 'Alignment' lets the user upload two vaccine coding systems and generates an alignment between them based on the code descriptors. This can be used to create code mappings between the coding system used to define a vaccine exposure in the study protocol and the coding systems used to represent vaccinations in the EHR databases.

Methods for broadening event definitions or for unifying database codes will remain useful to the conduction of collaborative observational studies as long as the representational heterogeneity between databases persists. Different initiatives have been put forwards to eliminate the heterogeneity altogether by establishing a common data model (CDM) that specifies the database layout and coding systems to which databases have to conform to enable the extraction of medical outcomes based on a single database query. Several standardization initiatives exist, for example the OMOP common data model from the OHDSI project for EHR databases, and the ISO IDMP standard for recording medical products including vaccines.[179,229] However, due to the 'glacial speed' of changes in EHR databases, a potentially slow adaption can temporarily result in an even more heterogeneous situation (figure 8.2). Moreover, the migration of a database to a CDM still requires the mapping of custom, database-specific coding systems to the coding system imposed by the CDM.

TYPES OF REPRESENTATIONAL HETEROGENEITY AND THEIR RESOLUTION

This thesis tapped a number of resources for accessing vaccine-related information: public social media messages for vaccine safety concerns and public sentiment vaccines, scientific literature for established evidence about vaccines, and EHR databases for vaccinations and outcomes. We observed three types of representational heterogeneity.

First, representational heterogeneity in free text may be due to the use of different languages and terminologies, of synonyms (e.g., 'immunization' or 'vaccination'), and of syntactic term variation. Free text was pervasive in this thesis and the primary content of social media messages, case definitions, and scientific literature, and secondary content as descriptors in vaccine coding systems. Linguistic heterogeneity was addressed in the step of recognition of relevant information. The presented methods were mostly based on dictionaries, which aggregated various terms describing a concept. Dictionaries proved a simple and reliable tool for the recognition of information. Their use in RELigator showed that the performance of dictionary-based approaches is comparable to the performance of machine-learning approaches but they require less or no training data. Dictionary-based approaches can easily be adapted to multilingual input: To recognize countries in social media messages we compiled a multilingual dictionary from the Geonames database, which includes geographical terms in various languages and terminologies.[106] And the dictionary of the VaccO ontology was initially compiled from different resources of English terms and subsequently translated to a number of European languages using an automatic translation service.[105] Comprehensive multilingual resources are generally rare or unbalanced between languages (e.g., the UMLS), but automatic translation services cover by now hundreds of languages and achieve high performance.[236,237]

Second, medical information is represented in European EHR database using different medical coding systems, which was addressed in part III. Any collaborative observational study that covers databases with different coding systems requires mappings between the coding systems. For CodeMapper, we used existing mappings between the coding systems to harmonize the different extraction queries for a medical event. But the mappings between coding systems are not always available and with VaccO we developed an method to automatically create mappings between vaccine coding systems.

Third, corresponding information can be represented on different taxonomic levels (e.g., products vs. pharmacological classes) and the description of equivalent information can be based on different properties (e.g., defining a vaccine by pathogen as in 'HPV vaccine' or by disease

as in 'cervical cancer vaccine'). Such conceptual heterogeneity occurs independently of the type of content (free text or code descriptors) and its resolution requires automatic reasoning using domain-specific knowledge. The best-performing method for the alignment of vaccine codes represented vaccine descriptions in the VaccO ontology and applied an ontology reasoner to resolve conceptual differences.

Automatic methods for resolving representational heterogeneity can be based on an analysis of the task, the domain, or exemplary answers, depending on the specific task and the availability of information resources, and with consequences for the interpretability and sustainability of the resulting models.[54] An analysis of the task (including the resource) underlies the creation of rule-based methods. For example, we manually compiled dictionaries that captured how specific vaccines and outcomes are described in social media. Task-based methods can easily be created in an iterative process to solve small-scale problems, but they are not well suited for large-scale tasks due to their inflexibility. For example, the heuristics used to annotate vaccine descriptions based on sentence co-occurrence between vaccines and properties turned out as major error source in our model for vaccine-description recognition.

Automatic methods for vaccine semantics can be implemented based on formalized domain-knowledge. The formalization of domain-specific knowledge requires initially great manual effort, but such resource can subsequently be used to solve different tasks in a domain. For example, information from the UMLS was applied in all three steps of vaccine semantics throughout this thesis: the recognition of concepts was based on dictionaries compiled from the UMLS (chapters 4 to 7), concepts were represented by UMLS CUIs in RELigator and CodeMapper, and the reasoning underlying the semantic operations in CodeMapper were based on the semantic relations from the UMLS. The VaccO ontology was created for the alignment of vaccine coding systems and applied to the recognition of vaccine descriptions. Further examples of domain-specific knowledge in this thesis were the Euretos knowledge graph for creating feature sets for RELigator and the Geonames database for recognizing countries in multilingual social media messages.

Finally, methods for vaccine semantics can be automatically derived from a training material of exemplary answers to the task by supervised machine learning. Machine-learning methods are independent from a manual analysis of the task or domain and capable of developing internal representations and generalizations for a large variety of tasks.[238] They offer high flexibility for automatic information extraction and categorization. In the context of this thesis, training material were only available for the classification of vaccine literature and the relation extraction of chemical-induced diseases (from Pubmed and the Biocreative V challenge, respectively). We compared different machine-learning

methods and feature sets for the classification of scientific articles, and a CNN using word-embeddings performed best. The RELigator system used a support vector machine to classify chemical-induced diseases.

However, the training of machine learning models requires substantial annotated, task-specific material, which are unavailable for most information extraction tasks in the B/R monitoring of vaccines. Additional experiments for chapter 4 showed that typically 50-60 articles were required for one vaccine code to train a CNN model that performed better than the ontology-based model. In the absence of task-specific training material, formalized, domain-specific knowledge is fundamental for implementing automatic methods for solving representational heterogeneity. Machine learning further results in black-box models that make the analysis and understanding of specific behaviour challenging. The behaviour of applications of domain-knowledge is usually interpretable, which allows, for example, for tracking down the reasons of incorrect behaviour to errors in the formalized domain knowledge in the application, where they can be corrected. The *RF-VaccO* model and RELigator showed the value of combining domain-specific knowledge with machine learning. Integrating deep learning models (e.g., CNN) with domain-specific knowledge may be challenging, but promises to further improve performance in classification tasks.[238]

The sustainability of automatic methods depends on their currency, i.e., on keeping them up-to-date with respect to changes in the task and the context. Most common changes in the context of vaccines come from new vaccine developments (e.g., immunization targets and vaccine strategies) and the authorization of new vaccine products. Bringing rule-based and example-based methods up-to-date requires considerable effort and specialized personnel: for rule-based methods, a programmer has to update the program code with potentially growing complexity to reflect the changes, and updating example-based methods requires the manual creation of training material that reflects the changes, and the retraining of the model.[239] Resources of formalized domain-knowledge can be created and updated in two ways. When created manually, dedicated software is commonly used and can also be applied for integrating changes (Protégé in the case of the VaccO ontology[240]). When compiled from a broader resource by a computer program (from EMA's article 57 database in the case of the VaccO ontology[241]), the creation process can be repeated when the underlying resource has been updated. In either case, sustainability of automatic methods to vaccine semantics requires comprehensive instructions how the relevant data was created, to enable a consistent and timely update.

9

SUMMARY

Post-marketing management and decision-making about vaccines builds upon the timely detection of vaccine safety signals and changes in public sentiment, established evidence about a vaccine, and the quantification of benefits and risks and the verification of hypotheses about the benefits and risks (B/R) of vaccines in observational studies. Prompt and accurate access to such vaccine-related information is fundamental to ensure the safety of vaccination programmes and to maintain public confidence. Several resources provide relevant insights about vaccines but the representation of information differs between and within resources, which impedes information extraction. This thesis proposed automatic methods to retrieve vaccine-related information from resources in view of such representational heterogeneity.

Part I was motivated by the question if public social media messages capture information that is relevant for the post-marketing management of vaccines. Chapter 2 explored the use of social media messages to monitor vaccine safety concerns by analysing English messages mentioning an HPV vaccine and infertility, a supposed adverse event. Publicly available data from the considered social media networks were sparse and largely untrackable for the purpose of providing early clues of safety concerns. Chapter 3 targeted the monitoring of public sentiment towards a vaccine in a multinational debate by analysing international, multilingual messages about a pentavalent vaccine, which had previously been introduced in numberous lower and middle-income countries. The messages that were analysed in both studies possessed three salient properties: frequent references to other websites (particularly news pages), few interactions between users (i.e., few replies or reposts), and virtual absence of personal reports about the vaccines. This may, however, come as little surprise given that only public messages were analysed, whereas personal experiences may rather be expressed in private messages, which were unavailable for the purpose of the studies.

The debate as manifested in social media was portrayed by peaks of messages following events in country-specific vaccination programmes. The perception of events was local: users reacted largely to events in their own country or adjacent countries, suggesting multiple, national debates rather than a multinational debate. The messages did not directly reflect the public's sentiment about the vaccine issues or events, because most messages by private persons were composed only by

references to news pages that covered major events in the implementation of vaccination programmes. The messages, however, may reflect the users' sentiment indirectly exactly because the lack of additional, personal content suggests that the authors generally concur with the referenced content. The dominant concern in referred websites was about the safety of the vaccine.

Part II proposed automatic methods underlying the extraction of established evidence about vaccines from scientific literature. Chapter 4 focused on the recognition of vaccine descriptions and on the classification of research articles by vaccines. The proposed method for the recognition of vaccine descriptions used the VaccO ontology of vaccine descriptions to recognize vaccines and their properties, and a simple heuristic to annotate vaccine descriptions based on the co-occurrence of vaccines and properties within sentences. The method performed reasonably well (F-score 0.69) in comparison to the inter-annotator agreement of the manually created reference corpus of vaccine descriptions (F-score 0.80). The error analysis revealed two possible ways for improving the approach. First, the VaccO ontology should be expanded with non-European vaccines, vaccines under development, and their properties. Second, the rule-based heuristic to annotate descriptions based on sentence co-occurrence was a major error source and should take into account the textual context of recognized vaccines and properties. The reference corpus of vaccine descriptions that we developed is the first of its kind and publicly available for training and testing of future methods.

For the classification of vaccine articles, we compared dictionary-based, ontology-based, statistical, and hybrid methods to assign vaccine codes from the MeSH vocabulary by an analysis of the article title and abstract. A simple dictionary of vaccine descriptions lacked the necessary flexibility in light of the large syntactic variation of vaccine descriptions in scientific literature (F-score 0.45). The ontology-based method derived a list of properties of the articles from the vacco ontology and applied the alignment algorithm (chapter 7) for the assignment of relevant MeSH vaccine codes. The algorithm, which proved suitable for aligning descriptors between vaccine coding systems, had only moderate performance for categorizing vaccine literature (F-score 0.60). The better performance of a hybrid random forest model based on the property list derived from VaccO appeared a more robust approach for assigning vaccine headings (F-score 0.68). A CNN constituted the best model for categorizing vaccine literature but required substantial training material to train classifiers that perform well (F-score 0.76). The performance of the CNN with respect to a single code correlated strongly with the number of articles with that code in the training set. The CNN performed better than the ontology-based method for almost

all frequent vaccine codes but was outperformed by it for most infrequent codes. Vaccine codes, however, may necessarily be infrequent in the training data for novel or less-studied vaccines. To further improve performance, the neural network-based method and the ontology-based methods could be combined dependening on the number of training examples.

Chapter 5 covered the extraction of causal relations between chemicals and diseases rather than vaccines, but this work would support equally well vaccine-related extractions. The extraction system, RELigator, was implemented by a support vector machine, using feature sets based on a linguistic analysis of the article, information extracted from the Euretos knowledge graph, and statistics about words in the training set. RELigator achieved an F-score of 0.60, and each of the features sets contributed to the final system performance.

Part III presented methods to retrieve vaccines and outcomes from EHR databases that use different coding systems. In chapter 6, we presented CodeMapper, a web application that assists in the mapping of clinical case definitions into code sets from multiple coding systems, which is often a bottleneck in the implementation of collaborative observational studies. CodeMapper constitutes a single entry point for all phases of creating database-specific code sets for a medical outcome: the identification of relevant medical concepts from the case definition, the manual revision of medical concepts, and the projection to database-specific code sets. A number of semantic operations in CodeMapper help in retrieving related medical concepts during the manual revision. The overall mapping process is automatically recorded and manual descriptions of the process are regularly requested, which makes the mapping process traceable and the mappings more suitable to subsequent studies. CodeMapper and its source code are publicly available. We evaluated the effectiveness of CodeMapper's approach by simulating an informed usage. Creating a mapping without revising the concepts was insufficient for reproducing the reference code sets, indicating that the mapping process cannot be replaced by a simple indexing step. However, the goal of CodeMapper is to support an informed user in creating such mappings, and our evaluation showed that CodeMapper's semantic operations provided an effective and efficient way (sensitivity 0.95 and precision 0.62).

Chapter 7 presented the VaccO ontology of vaccine descriptions, and evaluated different automatic methods for the alignment of corresponding vaccine codes between coding systems. The VaccO ontology defines classes for vaccine products and vaccine properties, namely immunization targets (including pathogens and diseases), immunization strategies, administration routes, valences, and ingredients. VaccO is agnostic of any specific vaccine coding systems: the basic class of

vaccines does not subsume any codes but only classes that represent pharmacological vaccine groups, common vaccine abbreviations (e.g., 'DTaP') and vaccine products, which are defined by their relations to other entities in the ontology. Each class includes a list of terms, which can be used to identify references to the class in free text. The VaccO ontology is publicly available.

The automatic methods for the alignment of corresponding vaccine codes between different coding systems were based on multilingual code descriptors, the only information about codes that is available in all vaccine coding systems. Two baseline methods were defined based on the overlap of words in the descriptors, and on the overlap of UMLS concepts identified in the descriptors. The baseline methods were compared with three methods built on the VaccO ontology, defined by the overlap of VaccO classes recognized in the descriptors, the equivalence between the logical expressions that represent the descriptors in VaccO, and the overlap between properties inferred about the descriptors from the ontology. The baseline methods lacked domain-specific knowledge about vaccines and vaccine descriptions, and were generally outperformed by the methods built on the VaccO ontology. The method based on the overlap of recognized VaccO classes outperformed the baseline methods in the reference set with multilingual descriptors, due to the multilingual dictionary in the VaccO ontology. The method based on the equivalence of logical expressions correctly aligned codes of equivalent vaccine groups described by different properties (e.g., 'HPV vaccine' and 'cervical cancer vaccine'). The method based on a flat representation of inferred information additionally identified matching codes across taxonomic levels, and demonstrated excellent performance (F-scores 0.91 and 0.96 in two reference sets).

Accurate, automatic methods for extracting information from resources with representational heterogeneity for the post-licensure vaccine management can be based on task-specific rules, formalized domain-knowledge, or example solutions using supervised machine learning. Formalized domain-knowledge provides the highest flexibility for specific vaccine-related tasks, where training corpora are unavailable.

SAMENVATTING

Vaccins worden door veel gezonde mensen gebruikt en zijn zeer effectief gebleken in het voorkomen van infectieziekten. Nadat vaccins op de markt komen is het belangrijk dat de positieve effecten en bijwerking goed worden gevolgd zodat het publiek vertrouwen houdt in de programma's. Hiervoor is snelle en goede toegang tot relevante informatie van essentieel belang. Er zijn vele verschillende bronnen met relevante informatie over vaccins, maar de representatie van de informatie verschilt tussen en binnen deze bronnen, wat het extraheren van informatie bemoeilijkt. Dit proefschrift beschrijft verschillende geautomatiseerde methoden om informatie over vaccins uit bronnen met representationele heterogeniteit, te halen.

Deel I richt zich op publieke sociale media als informatie bron. In hoofdstuk 2 onderzochten we of social media berichten gebruikt konden worden om signalen op te pikken omtrent HPV-vaccin en onvruchtbaarheid, een veronderstelde bijwerking. Hoofdstuk 3 keek of social media analyse een bijdrage zou kunnen hebben bij het monitoren van de multinationale publieke opinie over het pentavalente vaccin. De berichten die in deze studies werden geanalyseerd, hadden drie opvallende eigenschappen: frequente verwijzingen naar andere websites (met name nieuwspagina's), weinig interacties tussen gebruikers (d.w.z. weinig antwoorden of reposts), en het vrijwel ontbreken van persoonlijke rapporten over de vaccins. Het debat zoals dat zich in de sociale media manifesteerde, werd gekarakteriseerd door pieken in de berichten naar aanleiding van gebeurtenissen in landenspecifieke vaccinatieprogramma's. De perceptie van de gebeurtenissen was lokaal: de gebruikers reageerden grotendeels op gebeurtenissen in hun eigen land of aangrenzende landen, wat meerdere nationale debatten in plaats van een multinationaal debat suggereert. De berichten gaven niet direct het sentiment van het publiek over de vaccinatiekwesties of -gebeurtenissen weer, omdat de meeste berichten van particulieren alleen waren samengesteld uit verwijzingen naar nieuwspagina's over belangrijke gebeurtenissen bij de uitvoering van vaccinatieprogramma's. De berichten kunnen echter indirect het sentiment van de gebruikers weerspiegelen, juist omdat het gebrek aan aanvullende, persoonlijke inhoud suggereert dat de auteurs het in het algemeen eens zijn met de inhoud waarnaar verwezen wordt. De belangrijkste zorg op de websites waarnaar wordt verwezen, was de veiligheid van het vaccin.

Deel II richtte zich op de de wetenschappelijke literatuur als bron voor informatie. Hoofdstuk 4 richtte zich op de herkenning van vaccinbeschrijvingen en de classificatie van wetenschappelijke artikelen. Om vaccins te herkennen werd de VaccO-ontologie gebruikt. De methode presteerde redelijk goed (F-score 0,69) in vergelijking met de interannotatorovereenkomst van het handmatig gemaakte referentiecorpus van vaccinbeschrijvingen (F-score 0,80). De foutanalyse bracht twee mogelijke manieren aan het licht om de aanpak te verbeteren. Ten eerste moet de VaccO-ontologie worden uitgebreid met niet-Europese vaccins en vaccins in ontwikkeling, samen met hun eigenschappen. Ten tweede was de gebruikte heuristiek een belangrijke bron van fouten en moet in plaats daarvan rekening worden gehouden met de tekstuele context van herkende vaccins en eigenschappen. Het referentiecorpus van vaccinbeschrijvingen dat werd ontwikkeld, is het eerste in zijn soort en publiek beschikbaar voor training en testen van toekomstige methoden.

Voor de classificatie van vaccinartikelen vergeleken we woordenboekgebaseerde, ontologie-gebaseerde, statistische en hybride methoden om vaccincodes toe te wijzen uit de MeSH-vocabulaire. Een eenvoudig woordenboek van vaccinbeschrijvingen miste de nodige flexibiliteit in het licht van de grote syntactische variatie van vaccinbeschrijvingen in de wetenschappelijke literatuur (F-score 0,45). De ontologie-gebaseerde methode berekende de vlakke representatie van informatie uit de VaccO-ontologie voor de artikelen en paste het aligneringsalgoritme (hoofdstuk 7) toe voor de toewijzing van relevante MeSH-vaccincodes. Het algoritme, dat geschikt bleek voor het aligneren van descriptoren uit verschillende vaccincoderingssystemen, had een matige prestatie voor het categoriseren van vaccinliteratuur (F-score 0,60). De betere prestaties van een hybride random forest model gebaseerd op de vlakke representatie van VaccO-informatie bleek een robuustere aanpak voor het toewijzen van vaccinkoppen (F-score 0,68). Een convolutioneel neuraal netwerk (CNN) vormde het beste model voor het categoriseren van vaccinliteratuur, maar vereiste een substantiële hoeveelheid trainingsmateriaal om goed presterende classificatoren te trainen (F-score 0,76). De prestaties van de CNN voor een specifieke code correleerden sterk met het aantal artikelen met die code in de trainingsset. De CNN presteerde beter dan de ontologie-gebaseerde methode voor bijna alle frequent voorkomende vaccincodes, maar presteerde minder goed voor de meeste weinig voorkomende codes van nieuwe of minder bestudeerde vaccins. De neurale netwerkgebaseerde methode en de ontologiegebaseerde methoden zouden kunnen worden gecombineerd, afhankelijk van het aantal beschikbare trainingsvoorbeelden, om de prestaties verder te verbeteren.

Hoofdstuk 5 had betrekking op de extractie van causale verbanden tussen chemische stoffen en ziekten, maar dit werk zou evengoed vaccingerelateerde extracties kunnen ondersteunen. Het extractiesysteem, RELigator, werd geïmplementeerd als een support vector machine, met featuresets gebaseerd op een taalkundige analyse van het artikel, informatie uit de Euretos-kennisgraaf en statistieken over woorden in de trainingsset. RELigator behaalde een F-score van 0,60 en elk van de featuresets droeg bij aan de uiteindelijke systeemprestatie.

Deel III presenteerde methoden om vaccins en uitkomsten te halen uit internationale elektronische zorg databases- die gebruik maken van verschillende coderingssystemen. In hoofdstuk 6 presenteerden we een webapplicatie – CodeMapper – die helpt bij het vertalen van klinische case definities naar codesets van verschillende coderingssystemen, wat vaak een knelpunt is bij de implementatie van collaboratieve observationele studies. CodeMapper vormt één enkel toegangspunt voor alle fasen in het creëren van database-specifieke codesets voor een medische uitkomst: de identificatie van relevante medische concepten in de casusdefinitie, de handmatige herziening van medische concepten, en de projectie naar database-specifieke codesets. Verschillende semantische operaties in CodeMapper helpen bij het vinden van gerelateerde medische concepten tijdens de handmatige revisie. Het hele mappingproces wordt automatisch geregistreerd en er worden regelmatig manuele beschrijvingen van het proces gevraagd, waardoor het mappingproces traceerbaar is en de mappings beter geschikt zijn voor latere studies. CodeMapper en zijn broncode zijn publiek beschikbaar. We hebben de effectiviteit van CodeMapper geëvalueerd door het simuleren van een geïnformeerd gebruik bij de mapping van een aantal casusdefinities. Het maken van een mapping zonder revisie van de concepten was onvoldoende om de referentiecodesets te reproduceren, wat aangeeft dat het mapping proces niet kan worden vervangen door een eenvoudige indexeringsstap. Het doel van CodeMapper is echter om een geïnformeerde gebruiker te ondersteunen bij het maken van dergelijke mappings en uit onze evaluatie bleek dat de semantische operaties van CodeMapper daarvoor een effectieve en efficiënte manier zijn (sensitiviteit 0,95 en precisie 0,62).

Hoofdstuk 7 presenteerde VaccO, een applicatie-ontologie voor vaccinbeschrijvingen, en evalueerde verschillende geautomatiseerde methoden voor de alignering van overeenkomstige vaccincodes in verschillende coderingssystemen. De VaccO-ontologie definieert klassen voor vaccinproducten en voor vaccineigenschappen, namelijk immunisatiedoelen (pathogenen en ziekten), immunisatiestrategieën, toedieningsroutes, valenties en ingrediënten. VaccO is agnostisch voor specifieke vaccincoderingssystemen: de basisklasse van vaccins omvat geen codes, maar alleen klassen die farmacologische vaccingroepen,

algemene vaccinafkortingen (bv. 'DTaP') en vaccinproducten vertegenwoordigen, en die worden gedefinieerd door hun relaties met andere entiteiten in de ontologie. Elke klasse bevat een lijst van descriptoren, die gebruikt kunnen worden om verwijzingen naar de klasse in vrije tekst te identificeren. De VaccO-ontologie is openbaar.

De geautomatiseerde methoden voor de alignering van overeenkomstige vaccincodes tussen verschillende coderingssystemen waren gebaseerd op meertalige codebeschrijvingen, de enige informatie over codes die algemeen beschikbaar is in vaccincodeersystemen. Twee basismethoden werden gedefinieerd: woordoverlap in de descriptoren en overlap van UMLS-concepten in de descriptoren. De basismethoden werden vergeleken met drie methoden gebaseerd op de VaccO-ontologie: overlap van de VaccO-klassen die in de descriptoren worden herkend, equivalentie tussen de logische uitdrukkingen die de descriptoren in VaccO vertegenwoordigen, en overlap tussen de vlakke representaties van de informatie die werd afgeleid over de descriptoren uit de ontologie. De basismethoden ontbrak het aan domeinspecifieke kennis over vaccins en vaccinbeschrijvingen, en werden in het algemeen overtroffen door de methoden gebaseerd op de VaccO-ontologie. De methode gebaseerd op de overlap van herkende VaccO-klassen presteerde beter dan de basismethoden in de referentieset met meertalige descriptoren, als gevolg van het meertalige woordenboek in de VaccO-ontologie. De methode gebaseerd op equivalentie van logische uitdrukkingen, aligneerde de codes van corresponderende vaccingroepen die werden beschreven door verschillende eigenschappen (bv. 'HPV vaccin' en 'cervicaal kankervaccin'). De methode gebaseerd op een vlakke representatie van afgeleide informatie identificeerde ook overeenkomstige codes op verschillende taxonomische niveaus, en gaf uitstekende prestaties (F-scores 0,91 en 0,96 in twee referentiesets).

Nauwkeurige, geautomatiseerde methoden voor het extraheren van informatie uit bronnen met representatieve heterogeniteit voor het post-licentiebeheer van vaccins, kunnen worden gebaseerd op taakspecifieke regels, geformaliseerde domeinkennis of voorbeeldoplossingen in combinatie met machinaal leren. Geformaliseerde domeinkennis biedt de hoogste flexibiliteit voor specifieke vaccingerelateerde taken, waar trainingscorpora niet beschikbaar zijn.

BIBLIOGRAPHY

1. Plotkin SA. Vaccines: Past, Present and Future. *Nature Medicine* **11** (2005).
2. Plotkin SA, Orenstein O, Offit P. *Vaccines* (Elsevier, 2013).
3. Offit PA, DeStefano F. *Vaccine Safety* in *Vaccines* (eds Plotkin SA, Orenstein WA, Offit PA) (Elsevier, 2013).
4. Hamborsky J, Kroger A, Wolfe C. *Epidemiology and Prevention of Vaccine-Preventable Diseases* (Public Health Foundation, Washington, US, 2015).
5. Koplow DA. *Smallpox: The Fight to Eradicate a Global Scourge* (Univ of California Press, 2004).
6. World Health Organization. *South-East Asia Region Certified Polio-Free* ⟨*http://www.searo.who.int/mediacentre/releases/2014/pr1569/en/*⟩ (visited on 20/07/2018).
7. US Centers for Disease Control and Prevention. *The Global Impact of Vaccines* ⟨*https://www.cdc.gov/globalhealth/infographics/immunization/global_impact_of_vaccines.htm*⟩ (visited on 20/07/2018).
8. Patel MK. Progress Toward Regional Measles Elimination – Worldwide, 2000–2015. *MMWR Morb Mortal Wkly Rep* **65** (2016).
9. Caplan AL, Schwartz JL. *Ethics* in *Vaccines* (eds Plotkin SA, Orenstein WA, Offit PA) (Elsevier, 2013).
10. Baylor NW, Marshall VB. *Regulation and Testing of Vaccines* in *Vaccines* (eds Plotkin SA, Orenstein WA, Offit PA) (Elsevier, 2013).
11. Ehmann F, Kurz X, Cavaleri M, Arlett P. *Regulation of Vaccines in Europe* in *Plotkin's Vaccines (Seventh Edition)* (Elsevier, 2013).
12. Giese M. *Introduction to Molecular Vaccinology* 383 pp. (Springer, 2016).
13. European Medicines Agency. *EudraVigilance Access Policy for Medicines for Human Use* 2011.
14. US Centers for Disease Control and Prevention. *Vaccine Safety Monitoring Monitoring* ⟨*https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/index.html*⟩ (visited on 22/07/2018).
15. Nair H, Hazarika I, Patwari A. A Roller-Coaster Ride: Introduction of Pentavalent Vaccine in India. *J Glob Health* **1** (2011).
16. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS et al. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature* **457** (2009).
17. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ. Predicting Consumer Behavior with Web Search. *PNAS* **107** (2010).
18. White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-Scale Pharmacovigilance: Listening to Signals from the Crowd. *J Am Med Inform Assoc* **20** (2013).
19. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R et al. Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products in Twitter. *Drug safety* **37** (2014).
20. Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J et al. *Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks* in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2010).

21. O'Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith KL et al. *Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions* in *AMIA Annual Symposium Proceedings* (American Medical Informatics Association, 2014).

22. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K et al. Utilizing Social Media Data for Pharmacovigilance: A Review. *Journal of biomedical informatics* **54** (2015).

23. Harpaz R, DuMouchel W, Schuemie M, Bodenreider O, Friedman C et al. Toward Multimodal Signal Detection of Adverse Drug Reactions. *Journal of biomedical informatics* **76** (2017).

24. Keelan J, Pavri V, Balakrishnan R, Wilson K. An Analysis of the Human Papilloma Virus Vaccine Debate on MySpace Blogs. *Vaccine* **28** (2010).

25. Kelly BJ, Leader AE, Mittermaier DJ, Hornik RC, Cappella JN. The HPV Vaccine and the Media: How Has the Topic Been Covered and What Are the Effects on Knowledge about the Virus and Cervical Cancer? *Patient education and counseling* **77** (2009).

26. Salathé M, Khandelwal S. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *PLoS Comput Biol* **7** (2011).

27. Larson HJ, Cooper LZ, Eskola J, Katz SL, Ratzan S. Addressing the Vaccine Confidence Gap. *The Lancet* **378** (2011).

28. Larson HJ, Smith DM, Paterson P, Cumming M, Eckersberger E et al. Measuring Vaccine Confidence: Analysis of Data Obtained by a Media Surveillance System Used to Analyse Public Concerns about Vaccines. *The Lancet infectious diseases* **13** (2013).

29. US National Libary of Medicines. *PubMed* ⟨*https://www.ncbi.nlm.nih.gov/pubmed/*⟩ (visited on 08/03/2018).

30. Rogers F. Medical Subject Headings. *Bulletin of the Medical Library Association* **51** (1963).

31. National Libary of Medicines. *Medical Subject Headings* ⟨*https://www.nlm.nih.gov/mesh/*⟩ (visited on 08/03/2018).

32. US National Libary of Medicines. *Principles of MEDLINE Subject Indexing* ⟨*https://www.nlm.nih.gov/bsd/disted/meshtutorial/principlesofmedlinesubjectindexing/*⟩ (visited on 02/04/2018).

33. Liu S, Tang B, Chen Q, Wang X. Drug Name Recognition: Approaches and Resources. *Information* **6** (2015).

34. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep Learning with Word Embeddings Improves Biomedical Named Entity Recognition. *Bioinformatics* **33** (2017).

35. Pernus YB, Nan C, Verstraeten T, Pedenko M, Osokogu OU et al. Reference Set for Performance Testing of Pediatric Vaccine Safety Signal Detection Methods and Systems. *Vaccine* **34** (2016).

36. Giuliano C, Lavelli A, Romano L. *Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature* in *11th Conference of the European Chapter of the Association for Computational Linguistics* (2006).

37. Li J, Zhang Z, Li X, Chen H. Kernel-Based Learning for Biomedical Relation Extraction. *Journal of the Association for Information Science and Technology* **59** (2008).

38. Fundel K, Küffner R, Zimmer R. RelEx—Relation Extraction Using Dependency Parse Trees. *Bioinformatics* **23** (2007).

39. Tari L, Anwar S, Liang S, Cai J, Baral C. Discovering Drug-Drug Interactions: A Text-Mining and Reasoning Approach Based on Properties of Drug Metabolism. *Bioinformatics* **26** (2010).

40. Björne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T. Complex Event Extraction at PubMed Scale. *Bioinformatics* **26** (2010).

41. Simpson MS, Demner-Fushman D. *Biomedical Text Mining: A Survey of Recent Progress* in *Mining Text Data* (Springer, Boston, MA, 2012).

42. Quan C, Wang M, Ren F. An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature. *PLOS ONE* **9** (2014).

43. Li G, Ross KE, Arighi CN, Peng Y, Wu CH et al. miRTex: A Text Mining System for miRNA-Gene Relation Extraction. *PLOS Computational Biology* **11** (2015).

44. Song M, Kim WC, Lee D, Heo GE, Kang KY. PKDE4J: Entity and Relation Extraction for Public Knowledge Discovery. *Journal of Biomedical Informatics* **57** (2015).

45. Krallinger M, Leitner F, Rabal O, Vazquez M, Oyarzabal J et al. CHEMDNER: The Drugs and Chemical Names Extraction Challenge. *J Cheminform* **7** (2015).

46. Trifirò G, Coloma P, Rijnbeek P, Romio S, Mosseveld B et al. Combining Multiple Healthcare Databases for Postmarketing Drug and Vaccine Safety Surveillance: Why and How? *Journal of internal medicine* **275** (2014).

47. De Lusignan S, Liaw ST, Michalakidis G, Jones S. Defining Datasets and Creating Data Dictionaries for Quality Improvement and Research in Chronic Disease Using Routinely Collected Data: An Ontology-Driven Approach. *Inform Prim Care* **19** (2011).

48. ADVANCE consortium. *Ontology for the Integration and Extraction of Vaccine-Related Information in Europe: A Proof of Concept.* Deliverable 5.5 (IMI, 2017).

49. Avillach P, Mougin F, Joubert M, Thiessard F, Pariente A et al. A Semantic Approach for the Homogeneous Identification of Events in Eight Patient Databases: A Contribution to the European EU-ADR Project. *Stud Health Technol Inform* **150** (2009).

50. Gini R, Schuemie M, Brown J, Ryan P, Vacchi E et al. Data Extraction and Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* **4** (2016).

51. Avillach P, Coloma PM, Gini R, Schuemie M, Mougin F et al. Harmonization Process for the Identification of Medical Events in Eight European Healthcare Databases: The Experience from the EU-ADR Project. *Journal of the American Medical Informatics Association* **20** (2013).

52. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods of information in medicine* **32** (1993).

53. National Libary of Medicines. *UMLS Statistics - 2018AA Release* ⟨*https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html*⟩ (visited on 05/07/2018).

54. Euzenat J, Shvaiko P. *Ontology Matching* 2nd ed. (Springer, Berlin, Heidelberg, 2013).

55. He Y, Cowell L, Diehl AD, Mobley HL, Peters B et al. *VO: Vaccine Ontology* in *Proceedings of the 1st International Conference on Biomedical Ontology* 1st International Conference on Biomedical Ontology (Buffalo, NY, USA, 2009).

56. Gruber TR. Toward Principles for the Design of Ontologies Used for Knowledge Sharing? *International journal of human-computer studies* **43** (1995).

57. Gruber TR. *Ontology* in *Encyclopedia of Database Systems* (2009).

58. Coloma PM, Becker BFH, Sturkenboom MCJM, van Mulligen EM, Kors JA. Evaluating Social Media Networks in Medicines Safety Surveillance: Two Case Studies. *Drug Saf* **38** (2015).

59. Adams SA. Sourcing the Crowd for Health Services Improvement: The Reflexive Patient and "Share-Your-Experience" Websites. *Soc Sci Med* **72** (2011).

60. McKee M, Cole K, Hurst L, Aldridge RW, Horton R. The Other Twitter Revolution: How Social Media Are Helping to Monitor the NHS Reforms. *BMJ* **342** (2011).

61. Greene JA, Choudhry NK, Kilabuk E, Shrank WH. Online Social Networking by Patients with Diabetes: A Qualitative Evaluation of Communication with Facebook. *J Gen Intern Med* **26** (2011).

62. Knezevic MZ, Bivolarevic IC, Peric TS, Jankovic SM. Using Facebook to Increase Spontaneous Reporting of Adverse Drug Reactions. *Drug safety* **34** (2011).

63. Goldman SA. Limitations and Strengths of Spontaneous Reports Data. *Clin Ther* **20 Suppl C** (1998).

64. Gonzalez-Gonzalez C, Lopez-Gonzalez E, Herdeiro MT, Figueiras A. Strategies to Improve Adverse Drug Reaction Reporting: A Critical and Systematic Review. *Drug Saf* **36** (2013).

65. US Food and Drug Administration. *About the Center for Drug Evaluation and Research - For Industry: Using Social Media* ⟨*https://www.fda.gov/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/ucm397791.htm*⟩ (visited on 02/08/2018).

66. US Food and Drug Administration. *Guidance for Industry. Internet/Social Media Platforms with Character Space Limitations – Presenting Risk and Benefit Information for Prescription Drugs and Medical Devices* ⟨*http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm401087.pdf*⟩ (visited on 30/04/2015).

67. European Medicines Agency. *Guideline on Good Pharmacovigilance Practices (GVP). Module VI – Management and Reporting of Adverse Reactions to Medicinal Products* ⟨*http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/09/WC500172402.pdf*⟩ (visited on 30/04/2015).

68. Topsy Inc. *Topsy.Com* (defunct in December 2015). ⟨*https://topsy.com*⟩ (visited on 01/02/2014).

69. Becker BFH. *Vaccine Semantics – Online Material* ⟨*http://hdl.handle.net/1765/111218*⟩.

70. US Senate Committee on Finance. *Grassley, Baucus Release Committee Report on Avandia* ⟨*https://www.finance.senate.gov/release/grassley-baucus-release-committee-report-on-avandia*⟩ (visited on 02/08/2018).

71. Graham DJ, Ouellet-Hellstrom R, MaCurdy TE, Ali F, Sholley C et al. Risk of Acute Myocardial Infarction, Stroke, Heart Failure, and Death in Elderly Medicare Patients Treated with Rosiglitazone or Pioglitazone. *JAMA* **304** (2010).

72. Nissen SE, Wolski K. Rosiglitazone Revisited: An Updated Meta-Analysis of Risk for Myocardial Infarction and Cardiovascular Mortality. *Arch Intern Med* **170** (2010).

73. Mahaffey KW, Hafley G, Dickerson S, Burns S, Tourt-Uhlig S et al. Results of a Reevaluation of Cardiovascular Outcomes in the RECORD Trial. *Am Heart J* **166** (2013).

74. Takahashi Y, Hasegawa-Moriyama M, Sakurai T, Inada E. The Macrophage-Mediated Effects of the Peroxisome Proliferator-Activated Receptor-Gamma Agonist Rosiglitazone Attenuate Tactile Allodynia in the Early Phase of Neuropathic Pain Development. *Anesth Analg* **113** (2011).

75. Colafrancesco S, Perricone C, Tomljenovic L, Shoenfeld Y. Human Papilloma Virus Vaccine and Primary Ovarian Failure: Another Facet of the Autoimmune/Inflammatory Syndrome Induced by Adjuvants. *Am J Reprod Immunol* **70** (2013).

76. Little DT, Ward HRG. Premature Ovarian Failure 3 Years after Menarche in a 16-Year-Old Girl Following Human Papillomavirus Vaccination. *BMJ case reports* **2012** (2012).

77. Kulshrestha J, Kooti F, Nikravesh A, Gummadi PK. *Geographic Dissection of the Twitter Network.* in *ICWSM* (2012).

78. Sullivan SJ, Schneiders AG, Cheang CW, Kitto E, Lee H et al. 'What's Happening?' A Content Analysis of Concussion-Related Traffic on Twitter. *Br J Sports Med* **46** (2012).

79. Edwards IR, Lindquist M. Social Media and Networks in Pharmacovigilance: Boon or Bane? *Drug Saf* **34** (2011).

80. eHealthMe. *eHealthMe - Personalized Medication Management* ⟨*https://www.ehealthme.com/*⟩ (visited on 30/08/2018).

81. Chary M, Genes N, McKenzie A, Manini AF. Leveraging Social Networks for Toxicovigilance. *J Med Toxicol* **9** (2013).

82. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D et al. Text Mining for Adverse Drug Events: The Promise, Challenges, and State of the Art. *Drug Saf* **37** (2014).

83. Chee BW, Berlin R, Schatz B. Predicting Adverse Drug Events from Personal Health Messages. *AMIA Annu Symp Proc* **2011** (2011).

84. Butt TF, Cox AR, Oyebode JR, Ferner RE. Internet Accounts of Serious Adverse Drug Reactions: A Study of Experiences of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis. *Drug Saf* **35** (2012).

85. Slade BA, Leidel L, Vellozzi C, Woo EJ, Hua W et al. Postlicensure Safety Surveillance for Quadrivalent Human Papillomavirus Recombinant Vaccine. *JAMA* **302** (2009).

86. Agorastos T, Chatzigeorgiou K, Brotherton JML, Garland SM. Safety of Human Papillomavirus (HPV) Vaccines: A Review of the International Experience so Far. *Vaccine* **27** (2009).

87. Gee J, Naleway A, Shui I, Baggs J, Yin R et al. Monitoring the Safety of Quadrivalent Human Papillomavirus Vaccine: Findings from the Vaccine Safety Datalink. *Vaccine* **29** (2011).

88. Klein NP, Hansen J, Chao C, Velicer C, Emery M et al. Safety of Quadrivalent Human Papillomavirus Vaccine Administered Routinely to Females. *Arch Pediatr Adolesc Med* **166** (2012).

89. Chou WyS, Hunt YM, Beckjord EB, Moser RP, Hesse BW. Social Media Use in the United States: Implications for Health Communication. *Journal of Medical Internet Research* **11** (2009).

90. Nef T, Ganea RL, Müri RM, Mosimann UP. Social Networking Sites and Older Users - a Systematic Review. *Int Psychogeriatr* **25** (2013).

91. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D et al. eHealth Literacy and Web 2.0 Health Information Seeking Behaviors among Baby Boomers and Older Adults. *J Med Internet Res* **17** (2015).

92. Lazer D, Kennedy R, King G, Vespignani A. Big Data. The Parable of Google Flu: Traps in Big Data Analysis. *Science* **343** (2014).

93. Butler D. When Google Got Flu Wrong. *Nature* **494** (2013).

94. Mustafaraj E, Metaxas P. From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search. *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line* (2010).

95. Becker BFH, Larson HJ, Bonhoeffer J, van Mulligen EM, Kors JA et al. Evaluation of a Multinational, Multilingual Vaccine Debate on Twitter. *Vaccine* **34** (2016).

96. Wolfe RM, Sharp LK. Anti-Vaccinationists Past and Present. *BMJ* **325** (2002).

97. Poland GA, Jacobson RM. *Vaccine Safety: Injecting a Dose of Common Sense* in *Mayo Clinic Proceedings* **75** (Elsevier, 2000).

98. Yahya M. Polio Vaccines – "No Thank You!" Barriers to Polio Eradication in Northern Nigeria. *Afr Aff (Lond)* **106** (2007).

99. HPA COVER. *Summary of Trends in Vaccination Coverage in the UK. Annual COVER Report: 2005/06.* (Health Protection Agency, 2006).

100. Gavi Vaccine Alliance. *Pentavalent Vaccine Support* ⟨*http://www.gavi.org/support/nvs/pentavalent/*⟩ (visited on 22/08/2016).

101. World Health Organization. Pentavalent Vaccine in Asian Countries. *Weekly epidemiological record* **88** (2013).

102. Mudur G. Antivaccine Lobby Resists Introduction of Hib Vaccine in India. *BMJ* **340** (2010).

103. Benny P. Pentavalent Vaccine - Critisised in Asian Countries. *International Journal of Preventive and Therapeutic Medicine* **2** (2014).

104. OmegaWiki contributors. *OmegaWiki - A Dictionary in All Languages* ⟨*http://www.omegawiki.org/Meta:Main_Page*⟩ (visited on 27/08/2018).

105. Google Inc. *Google Translate* ⟨*https://translate.google.com/*⟩ (visited on 25/07/2017).

106. GeoNames contributors. *GeoNames Geographical Database* ⟨*http://www.geonames.org/*⟩ (visited on 01/08/2018).

107. Mollema L, Harmsen IA, Broekhuizen E, Clijnk R, De Melker H et al. Disease Detection or Public Opinion Reflection? Content Analysis of Tweets, Other Social Media, and Online Newspapers during the Measles Outbreak in The Netherlands in 2013. *Journal of medical Internet research* **17** (2015).

108. Becker BFH, He HY, Sturkenboom MCJM, Kors JA. Identifying and Normalizing Vaccine Descriptions in Scientific Literature (Submitted).

109. Sturkenboom MCJM. Advancing Collaborative Vaccine Benefits and Safety Research in Europe via the ADVANCE Code of Conduct. *Vaccine* **36** (2018).

110. Fillmore CJ, Baker CF. *Frame Semantics for Text Understanding* in *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL* (2001).

111. Hur J, Xiang Z, Feldman EL, He Y. Ontology-Based Brucella Vaccine Literature Indexing and Systematic Analysis of Gene-Vaccine Association Network. *BMC immunology* **12** (2011).

112. He Y, Racz R, Sayers S, Lin Y, Todd T et al. Updates on the Web-Based VIOLIN Vaccine Database and Analysis System. *Nucleic acids research* (2013).

113. Hur J, Schuyler AD, States DJ, Feldman EL. SciMiner: Web-Based Literature Mining Tool for Target Identification and Functional Enrichment Analysis. *Bioinformatics* **25** (2009).

114. Biosemantics working group. *The VaccO Ontology of Vaccine Descriptions* ⟨*https://biosemantics.org/software/vacco*⟩ (visited on 23/08/2018).

115. The Apache Software Foundation. *Apache Solr* ⟨*http://lucene.apache.org/solr/*⟩ (visited on 25/07/2017).

116. SolrTextTagger community. *SolrTextTagger: A Text Tagger Based on Lucene/Solr Using FST Technology* ⟨https://github.com/OpenSextant/SolrTextTagger⟩ (visited on 06/07/2017).

117. Leaman R, Lu Z. TaggerOne: Joint Named Entity Recognition and Normalization with Semi-Markov Models. *Bioinformatics* **32** (2016).

118. Brown EG, Wood L, Wood S. The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Safety* **20** (1999).

119. Donnelly K. SNOMED-CT: The Advanced Terminology and Coding System for eHealth. *Studies in health technology and informatics* **121** (2006).

120. World Health Organization. *The International Statistical Classification of Diseases and Health Related Problems ICD-10* (World Health Organization, 2004).

121. Zeng QT, Tse T. Exploring and Developing Consumer Health Vocabularies. *J Am Med Inform Assoc* **13** (2006).

122. Jaccard P. Étude Comparative de La Distribution Florale Dans Une Portion Des Alpes et Des Jura. *Bull Soc Vaudoise Sci Nat* **37** (1901).

123. Breiman L. Random Forests. *Machine learning* **45** (2001).

124. Bird S, Klein E, Loper E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* NLTK (O'Reilly Media, Inc., 2009).

125. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B et al. Scikit-Learn: Machine Learning in Python. *Journal of machine learning research* **12** (2011).

126. Kim Y. Convolutional Neural Networks for Sentence Classification. arXiv: *1408.5882* (2014).

127. Stypka J. *Magpie: Deep Neural Network Framework for Multi-Label Text Classification* 2018.

128. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. *Distributed Representations of Words and Phrases and Their Compositionality* in *Advances in Neural Information Processing Systems 26* (eds Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ) (Curran Associates, Inc., 2013).

129. Pyysalo S, Ginter F, Moen H, Salakoski, T., Ananiadou, S. Distributional Semantics Resources for Biomedical Text Processing. *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan* (2013).

130. Botsis T, Buttolph T, Nguyen MD, Winiecki S, Woo EJ et al. Vaccine Adverse Event Text Mining System for Extracting Features from Vaccine Safety Reports. *J Am Med Inform Assoc* **19** (2012).

131. Word Wide Web Consortium. *OWL 2 Web Ontology Language Document Overview (Second Edition)* ⟨https://www.w3.org/TR/owl2-overview/⟩ (visited on 04/07/2017).

132. Kokoska S, Zwillinger D. *CRC Standard Probability and Statistics Tables and Formulae* (Crc Press, 1999).

133. Hearst MA. *Automatic Acquisition of Hyponyms from Large Text Corpora* in *Proceedings of the 14th Conference on Computational Linguistics-Volume 2* (Association for Computational Linguistics, 1992).

134. Gupta S, Manning C. *Improved Pattern Learning for Bootstrapped Entity Extraction* in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (2014).

135. Pons E + Becker BFH, Akhondi SA, Afzal Z, van Mulligen EM, Kors JA. Extraction of Chemical-Induced Diseases Using Prior Knowledge and Textual Information. *Database* **2016** (2016).

136. Biosemantics working group. *RELigator* ⟨*https://biosemantics.org/software/religator*⟩ (visited on 23/08/2018).

137. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P et al. Computational Drug Repositioning: From Data to Therapeutics. *Clinical Pharmacology & Therapeutics* **93** (2013).

138. Preiss J, Stevenson M, Gaizauskas R. Exploring Relation Types for Literature-Based Discovery. *J Am Med Inform Assoc* **22** (2015).

139. Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A. Literature Mining, Ontologies and Information Visualization for Drug Repurposing. *Brief Bioinform* **12** (2011).

140. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using Rule-Based Natural Language Processing to Improve Disease Normalization in Biomedical Text. *J Am Med Inform Assoc* **20** (2013).

141. Leaman R, Islamaj Doğan R, Lu Z. DNorm: Disease Name Normalization with Pairwise Learning to Rank. *Bioinformatics* **29** (2013).

142. Krauthammer M, Nenadic G. Term Identification in the Biomedical Literature. *Journal of Biomedical Informatics. Named Entity Recognition in Biomedicine* **37** (2004).

143. Buyko E, Beisswanger E, Hahn U. *The Extraction of Pharmacogenetic and Pharmacogenomic Relations–a Case Study Using PharmGKB* in *Biocomputing 2012* (World Scientific, 2012).

144. Gurulingappa H, Mateen-Rajpu A, Toldo L. Extraction of Potential Adverse Drug Events from Medical Case Reports. *Journal of Biomedical Semantics* **3** (2012).

145. Kang N, Singh B, Bui C, Afzal Z, van Mulligen EM et al. Knowledge-Based Extraction of Adverse Drug Events from Biomedical Text. *BMC bioinformatics* **15** (2014).

146. Van Mulligen EM, Fourrier-Reglat A, Gurwitz D, Molokhia M, Nieto A et al. The EU-ADR Corpus: Annotated Drugs, Diseases, Targets, and Their Relationships. *Journal of Biomedical Informatics. Text Mining and Natural Language Processing in Pharmacogenomics* **45** (2012).

147. Wei CH, Kao HY, Lu Z. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *BioMed research international* **2015** (2015).

148. Schuemie MJ, Jelier R, Kors JA. *Peregrine: Lightweight Gene Name Normalization by Dictionary Lookup* in *Proceedings of the Biocreative 2 Workshop* **140** (2007).

149. Leaman R, Wei CH, Lu Z. tmChem: A High Performance Approach for Chemical Named Entity Recognition and Normalization. *Journal of cheminformatics* **7** (2015).

150. Davis AP, Wiegers TC, Roberts PM, King BL, Lay JM et al. A CTD–Pfizer Collaboration: Manual Curation of 88 000 Scientific Articles Text Mined for Drug–Disease and Drug–Phenotype Interactions. *Database* **2013** (2013).

151. US National Center for Biotechnology Information. *Stopwords* ⟨*https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/*⟩ (visited on 30/08/2018).

152. Divita G, Browne AC, Rindflesch TC. *Evaluating Lexical Variant Generation to Improve Information Retrieval.* in *Proceedings of the AMIA Symposium* (American Medical Informatics Association, 1998).

153. Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Res* **32** (suppl_1 2004).

154. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. *Studies in health technology and informatics* **84** (o 1 2001).

155. Schwartz AS, Hearst MA. *A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text* in *Biocomputing 2003* (World Scientific, 2002).

156. Euretos B.V. *Euretos - In Silico Target / Biomarker Discovery & Validation* ⟨*https://www.euretos.com/*⟩ (visited on 01/08/2018).

157. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: A PubMed-Scale Repository of Biomedical Semantic Predications. *Bioinformatics* **28** (2012).

158. Frank E, Hall MA, Witten IH. The WEKA Workbench. *Data mining: Practical machine learning tools and techniques* **4** (2016).

159. Pons E + Becker BFH, Akhondi SA, Afzal Z, van Mulligen EM, Kors JA. *RELigator: Chemical-Disease Relation Extraction Using Prior Knowledge and Textual Information* in *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop* (2015).

160. Wiegers TC, Davis AP, Cohen KB, Hirschman L, Mattingly CJ. Text Mining and Manual Curation of Chemical-Gene-Disease Networks for the Comparative Toxicogenomics Database (CTD). *BMC Bioinformatics* **10** (2009).

161. Becker BFH, Avillach P, Romio S, Mulligen EM, Weibel D et al. CodeMapper: Semiautomatic Coding of Case Definitions. A Contribution from the ADVANCE Project. *Pharmacoepidemiology and drug safety* **26** (2017).

162. De Lusignan S, Minmagh C, Kennedy J, Zeimet M, Bommezijn H et al. A Survey to Identify the Clinical Coding and Classification Systems Currently in Use across Europe. *Studies in health technology and informatics* (2001).

163. De Lusignan S. Codes, Classifications, Terminologies and Nomenclatures: Definition, Development and Application in Practice. *Journal of Innovation in Health Informatics* **13** (2005).

164. US Department of Health and Human Services and others. *ICD 9 CM. The International Classification of Diseases. 9. Rev: Clinical Modification.; Vol. 1: Diseases: Tabular List. ; Vol. 2: Diseases: Alphabetic Index. ; Vol. 3: Procedures: Tabular List and Alphabetic Index.* (US Government Printing Office, 1980).

165. Pavillon G, Maguin M. The 10th revision of the International Classification of Diseases. *Rev Epidemiol Sante Publique* **41** (1993).

166. Lamberts H, Okkes IM, World Organization of National Colleges and Academies of Family Physicians. ICPC-2, International Classification of Primary Care (1998).

167. Read JD, Sanderson HF, Sutton YM. *Terming, Encoding, Grouping, The Language of Health* in *Proceedings International Medical Information Association's 8th World Congress on Medical Informatics. Vancouver* (1995).

168. O'Neil M, Payne C, Read J. Read Codes Version 3: A User Led Terminology. *Methods of information in medicine* **34** (1995).

169. Coloma PM, Schuemie MJ, Trifirò G, Gini R, Herings R et al. Combining Electronic Healthcare Databases in Europe to Allow for Large-Scale Drug Safety Monitoring: The EU-ADR Project. *Pharmacoepidem Drug Safe* **20** (2011).

170. Aronson AR. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proc AMIA Symp* (2001).

171. GRiP. *Global Research in Paediatrics (GRiP) - Paediatric Clinical Pharmacology Studies* ⟨*http://www.grip-network.org/index.php/cms/en/home*⟩ (visited on 02/07/2017).

172. Eurosurveillance Editorial Team. ECDC in Collaboration with the VAESCO Consortium to Develop a Complementary Tool for Vaccine Safety Monitoring in Europe. *Eurosurveillance* **14** (2009).

173. EMIF Consortium. *European Medical Information Framework* ⟨*https://www.emif.eu*⟩ (visited on 03/08/2018).

174. Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW et al. Design Considerations, Architecture, and Use of the Mini-Sentinel Distributed Data System. *Pharmacoepidemiology and drug safety* **21** (2012).

175. Baker MA, Nguyen M, Cole DV, Lee GM, Lieu TA. Post-Licensure Rapid Immunization Safety Monitoring Program (PRISM) Data Characterization. *Vaccine* **31** (2013).

176. Chen RT, Glasser JW, Rhodes PH, Davis RL, Barlow WE et al. Vaccine Safety Datalink Project: A New Tool for Improving Vaccine Safety Monitoring in the United States. *Pediatrics* **99** (1997).

177. AsPEN collaborators, Andersen M, Bergman U, Choi NK, Gerhard T et al. The Asian Pharmacoepidemiology Network (AsPEN): Promoting Multi-National Collaboration for Pharmacoepidemiologic Research in Asia. *Pharmacoepidemiol Drug Saf* **22** (2013).

178. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a Common Data Model for Active Safety Surveillance Research. *Journal of the American Medical Informatics Association* **19** (2012).

179. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* **216** (2015).

180. ADVANCE consortium. *Accelerated Development of Vaccine Benefit-Risk Collaboration in Europe* ⟨*http://www.advance-vaccines.eu/*⟩ (visited on 28/07/2017).

181. Bean A, Green R. *Relationships in the Organization of Knowledge* 239 pp. (Springer Science & Business Media, 2013).

182. Becker BFH. *CodeMapper – Source Code* ⟨*https://archive.softwareheritage.org/swh:1:dir:bdc4a344d2b8bc53bd546bc5fcba4e82b1fbb4bf*⟩.

183. Biosemantics working group. *CodeMapper: Semi-Automatic Coding of Case Definitions* ⟨*https://biosemantics.org/software/codemapper*⟩ (visited on 23/08/2018).

184. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. *Safety Evaluation of Adverse Reactions in Diabetes - Comparative Studies* ⟨*http://www.encepp.eu/encepp/viewResource.htm?id=8323*⟩ (visited on 03/08/2018).

185. Bodenreider O, Burgun A, Botti G, Fieschi M, Le Beux P et al. Evaluation of the Unified Medical Language System as a Medical Knowledge Source. *Journal of the American Medical Informatics Association* **5** (1998).

186. Cimino JJ, Min H, Perl Y. Consistency across the Hierarchies of the UMLS Semantic Network and Metathesaurus. *Journal of biomedical informatics* **36** (2003).

187. Erdogan H, Erdem E, Bodenreider O. Exploiting UMLS Semantics for Checking Semantic Consistency among UMLS Concepts. *Studies in health technology and informatics* **160** (0 1 2010).

188. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of Alternative Standardized Terminologies for Medical Conditions within a Network of Observational Healthcare Databases. *Journal of biomedical informatics* **45** (2012).

189. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *Journal of biomedical informatics* **42** (2009).

190. Becker BFH, Kors JA, Mulligen EM, Sturkenboom MCJM. Alignment of Vaccine Codes Using the VaccO Ontology of Vaccine Properties (Submitted).

191. European Medicines Agency. *Guidance Documents for the Article 57 Database* ⟨*http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000336.jsp*⟩ (visited on 27/11/2015).

192. Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: Prescription for Electronic Drug Information Exchange. *IT professional* **7** (2005).

193. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized Names for Clinical Drugs: RxNorm at 6 Years. *Journal of the American Medical Informatics Association* **18** (2011).

194. Schulz EB, Barrett JW, Brown PJB, Price C. *The Read Codes: Evolving a Clinical Vocabulary to Support the Electronic Patient Record* in *Conference Proceedings: Toward an Electronic Health Record Europe. Newton: CAEHR* (1996).

195. Lowe HJ, Barnett GO. Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches. *Jama* **271** (1994).

196. World Health Organization. *WHOCC - ATC/DDD Index* ⟨*https://www.whocc.no/atc_ddd_index/*⟩ (visited on 04/07/2017).

197. ADVANCE consortium. *Initial Fingerprinting of the Participating Health Care Databases.* Deliverable 5.2 (IMI, 2016).

198. Brown SH, Elkin PL, Rosenbloom ST, Husser CS, Bauer BA et al. VA National Drug File Reference Terminology: A Cross-Institutional Content Coverage Study. *Medinfo* **11** (Pt 1 2004).

199. Carter JS, Brown SH, Erlbaum MS, Gregg W, Elkin PL et al. *Initializing the VA Medication Reference Terminology Using UMLS Metathesaurus Co-Occurrences.* in *Proceedings of the AMIA Symposium* (American Medical Informatics Association, Maryland, US, 2002).

200. Mehta D. *British National Formulary* (Pharmaceutical Press, London, UK, 2005).

201. Gruber TR. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* **5** (1993).

202. Xiang Z, Zheng W, He Y. BBP: Brucella Genome Annotation with Literature Mining and Curation. *BMC bioinformatics* **7** (2006).

203. Özgür A, Xiang Z, Radev DR, He Y. Mining of Vaccine-Associated IFN-γ Gene Interaction Networks Using the Vaccine Ontology. *Journal of biomedical semantics* **2** (2011).

204. ADVANCE consortium. *Results of POC-Phase 1 Studies.* Deliverable 5.6 (IMI, 2017).

205. Soualmia LF, Golbreich C, Darmoni SJ. *Representing the MeSH in OWL: Towards a Semi-Automatic Migration.* in *KR-MED* **102** (2004).

206. Fung KW, Bodenreider O. *Utilizing the UMLS for Semantic Mapping between Terminologies* in *AMIA Annual Symposium Proceedings* **2005** (American Medical Informatics Association, Bethesda, US, 2005).

207. Van Assem M, Malaisé V, Miles A, Schreiber G. *A Method to Convert Thesauri to SKOS* in *The Semantic Web: Research and Applications* European Semantic Web Conference (Springer, Berlin, Heidelberg, Germany, 2006).

208. Marquet G, Mosser J, Burgun A. A Method Exploiting Syntactic Patterns and the UMLS Semantics for Aligning Biomedical Ontologies: The Case of OBO Disease Ontologies. *Int J Med Inform* **76 Suppl 3** (2007).

209. Merabti T, Grosjean J, Soualmia LF, Joubert M, Darmoni SJ. *Aligning Biomedical Terminologies in French: Towards Semantic Interoperability in Medical Applications* (InTech, Rijeka, Croatia, 2012).

210. Winnenburg R, Rodriguez L, Callaghan FM, Sorbello A, Szarfman A et al. *Aligning Pharmacologic Classes Between MeSH and ATC.* in *VDOS+ DO@ ICBO* (2013).

211. Winnenburg R, Bodenreider O. A Framework for Assessing the Consistency of Drug Classes across Sources. *Journal of Biomedical Semantics* **5** (2014).

212. Atkinson AJ, Huang SM, Lertora JJ, Markey SP. *Principles of Clinical Pharmacology* (Academic Press, Cambridge, US, 2012).

213. Pathak J, Chute CG. Analyzing Categorical Information in Two Publicly Available Drug Terminologies: RxNorm and NDF-RT. *J Am Med Inform Assoc* **17** (2010).

214. Baader F. *The Description Logic Handbook: Theory, Implementation and Applications* (Cambridge university press, Cambridge, US, 2003).

215. Peroni S. *Graffoo Specification* ⟨*http://www.essepuntato.it/graffoo/specification/current.html*⟩ (visited on 26/06/2017).

216. National Institute of Allergy and Infectious Diseases. *Understanding Vaccines; What They Are; How They Work* (NIH Publication, Washington, US, 2003).

217. Baxter D. Active and Passive Immunity, Vaccine Types, Excipients and Licensing. *Occup Med* **57** (2007).

218. Center for Disease Control and Prevention. *U.S. Vaccine Names* ⟨*https://www.cdc.gov/vaccines/terms/usvaccines.html*⟩ (visited on 02/07/2017).

219. Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA et al. Characteristics of Consumer Terminology for Health Information Retrieval. *Methods of Information in Medicine-Methodik der Information in der Medizin* **41** (2002).

220. Federhen S. The NCBI Taxonomy Database. *Nucleic Acids Res* **40** (2012).

221. US Food and Drug Administration. *Data Standards Manual (Monographs) - Route of Administration* ⟨*https://www.fda.gov/drugs/developmentapprovalprocess/formssubmissionrequirements/electronicsubmissions/datastandardsmanualmonographs/ucm071667.htm*⟩ (visited on 23/05/2017).

222. Horridge M, Bechhofer S. The Owl Api: A Java API for Owl Ontologies. *Semantic Web* **2** (2011).

223. JFact community. *JFact DL Reasoner* ⟨*http://jfact.sourceforge.net/*⟩ (visited on 24/10/2017).

224. Information System for Research in Primary Care. *SIDIAP - General Details* ⟨*http://www.sidiap.org/index.php/database/general-details*⟩ (visited on 31/08/2017).

225. Agencia Española de Medicamentos y Productos Sanitarios. *Base de Datos Para La Investigación Farmacoepidemiológica En Atención Primaria* ⟨*http://www.bifap.org*⟩ (visited on 31/08/2017).

226. Progetto Pedianet. *Pedianet a Unique Opportunity, for Research in Pediatric Primary Care - Pedianet Project* ⟨*http://www.pedianet.it/en/*⟩ (visited on 31/08/2017).

227. Becker BFH. *VaccO Ontology – Source Code* ⟨*https://archive.softwareheritage.org/swh:1:dir:015d1ebe57a83657169437797dc0a0ea4b3b7323f*⟩.

228. Bennett CC. Utilizing RxNorm to Support Practical Computing Applications: Capturing Medication History in Live Electronic Health Records. *Journal of Biomedical Informatics. Translating Standards into Practice: Experiences and Lessons Learned in Biomedicine and Health Care* **45** (2012).

229. European Medicines Agency. *European Medicines Agency - Data on Medicines (ISO IDMP Standards) - Substance, Product, Organisation and Referential (SPOR) Master Data* ⟨*http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_001849.jsp*⟩ (visited on 12/04/2018).

230. Dodd C, Pacurariu A, Osokogu OU, Weibel D, Ferrajolo C et al. Masking by Vaccines in Pediatric Drug Safety Signal Detection in the EudraVigilance Database. *Pharmacoepidemiology and Drug Safety* (2018).

231. Shimabukuro TT, Nguyen M, Martin D, DeStefano F. Safety Monitoring in the Vaccine Adverse Event Reporting System (VAERS). *Vaccine* **33** (2015).

232. Bahri P, Melero MC. Listen to the Public and Fulfil Their Information Interests – Translating Vaccine Communication Research Findings into Guidance for Regulators. *British Journal of Clinical Pharmacology* **84** (2018).

233. US Department of Health & Human Services. *National Vaccine Program Office* ⟨*https://www.hhs.gov/nvpo/index.html*⟩ (visited on 20/08/2018).

234. ADVANCE consortium. *Report on Tested Methods for Accelerated Assessment of Vaccination Coverage, Vaccine Benefits, Risks and Benefit-Risk.* Deliverable 4.4 (IMI, 2017).

235. Gini R. *Tackling Heterogeneity in Disease Misclassification in Multi-Database Studies: The Case Study of Pertussis in the ADVANCE Project.* 2018.

236. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv: *1409.0473* (2014).

237. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv: *1609.08144* (2016).

238. Marcus G. Deep Learning: A Critical Appraisal. arXiv: *1801.00631* (2018).

239. Sculley D, Holt G, Golovin D, Davydov E, Phillips T et al. *Machine Learning: The High Interest Credit Card of Technical Debt* in *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)* (2014).

240. Musen MA. The Protégé Project: A Look Back and a Look Forward. *AI matters* **1** (2015).

241. European Medicines Agency. *Reporting Requirements for Authorised Medicines - Guidance Documents* ⟨*http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000336.jsp*⟩ (visited on 04/07/2017).

242. Akhondi SA, Pons E, Afzal Z, van Haagen H, Becker BFH et al. Chemical Entity Recognition in Patents by Combining Dictionary-Based and Statistical Approaches. *Database* **2016** (2016).

243. Akhondi SA, Pons E, Afzal Z, van Haagen H, Becker BFH et al. *Patent Mining: Combining Dictionary-Based and Machine-Learning Approaches* in *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop* (2015).

ABOUT THE AUTHOR

Benedikt Becker was born in Germany, 1983. He studied computer science and cognitive science at Ludwig-Maximilian University Munich and Albert-Ludwig University Freiburg. In his student project, he developed a constraint-based type checker for a subset of the OCaml language. He obtained his Diploma in 2011 with a research thesis in cognitive science about Bayesian rationality and spatial reasoning.

Between 2007 and 2011 he developed web applications and research tools in the context of medical inference and cognitive science as a freelance software engineer in Germany. From 2012 to 2013 he worked as a scientific programmer at IRILL, France, on the Ocsigen/Eliom web application framework.

From 2013 to 2017 he was a scientific researcher and PhD candidate in the Biosemantics group of the Erasmus Medical Center, Netherlands. His work focused on applications of natural-language processing and machine learning methods for monitoring the benefits and risks of vaccines, and has been supervised by Prof. Dr. Miriam Sturkenboom and Dr. Jan Kors.

As of 2018, he is working at INRIA Saclay, France, on combining formal proofs and symbolic execution for the verification of Shell scripts.

PHD PORTFOLIO

PhD training

- Epidemiology (Erasmus MC Summer School, 2014)
- History of Epidemiologic Ideas (Erasmus MC Summer School, 2014)
- Biomedical and Scientific English Writing and Communication (Erasmusc MC, 2014)
- Scientific Presentations (Erasmusc MC, 2013)
- Integrity in Science (Erasmusc MC, 2013)

Poster presentation

- CodeMapper: Semi-automatic mapping of case definitions (Ireland, ISPE 2016)

Oral presentations

- VaccO: An ontology of vaccine properties and one application (ADVANCE GAM 9, 2017)
- CodeMapper: Theory & practice (ADVANCE GAM 9, 2017)
- The European vaccine inventory (ADVANCE GAM 7, 2016)
- (Towards) automated methods for the analysis of attention and sentiment in the media (ADVANCE WP1 workshop, 2015)
- Text mining for post-licensure vaccine assessment (ADVANCE WP4 workshop, 2014)

Software

- CodeMapper: Semi-automatic coding of case definitions (*https://biosemantics.org/software/codemapper*)
- The VaccO ontology of vaccine descriptions, with three applications (*https://biosemantics.org/software/vacco*)

Project reports

- ADVANCE deliverable D5.5: Vaccine Ontology Report
- ADVANCE deliverable D4.9: Methods for vaccine benefit-risk monitoring, including vaccine coverage, safety and effectiveness

– ADVANCE deliverable D4.4: Report on tested methods for accelerated assessment of vaccination coverage, vaccine benefits, risks and benefit-risk
– WHO report SPHQ13 - LOA 209: Pentavalent Vaccine and Infant mortality – Analysis of public confidence in India as an integral part of comprehensive safety monitoring

Award

– Second place: CDR task in the BioCreative IV chemical-named entity recognition challenge (CHEMDNER, 2015)

PUBLICATIONS

- ★ Becker BFH, Kors JA, Mulligen EM, Sturkenboom MCJM. Alignment of Vaccine Codes Using the VaccO Ontology of Vaccine Properties (Submitted)
- ★ Becker BFH, He HY, Sturkenboom MCJM, Kors JA. Identifying and Normalizing Vaccine Descriptions in Scientific Literature (Submitted)
- Dodd C, Pacurariu A, Osokogu OU, Weibel D, Ferrajolo C, Vo DH, Becker B, Kors JA, Sturkenboom M. Masking by Vaccines in Pediatric Drug Safety Signal Detection in the EudraVigilance Database. *Pharmacoepidemiology and Drug Safety* (2018)
- ★ Becker BFH, Avillach P, Romio S, Mulligen EM, Weibel D, Sturkenboom MCJM, Kors JA. CodeMapper: Semiautomatic Coding of Case Definitions. A Contribution from the ADVANCE Project. *Pharmacoepidemiology and drug safety* **26** (2017)
- ★ Becker BFH, Larson HJ, Bonhoeffer J, van Mulligen EM, Kors JA, Sturkenboom MCJM. Evaluation of a Multinational, Multilingual Vaccine Debate on Twitter. *Vaccine* **34** (2016)
- ★ Pons E + Becker BFH, Akhondi SA, Afzal Z, van Mulligen EM, Kors JA. Extraction of Chemical-Induced Diseases Using Prior Knowledge and Textual Information. *Database* **2016** (2016)
- Akhondi SA, Pons E, Afzal Z, van Haagen H, Becker BFH, Hettne KM, van Mulligen EM, Kors JA. Chemical Entity Recognition in Patents by Combining Dictionary-Based and Statistical Approaches. *Database* **2016** (2016)
- Coloma PM, Becker BFH, Sturkenboom MCJM, van Mulligen EM, Kors JA. Evaluating Social Media Networks in Medicines Safety Surveillance: Two Case Studies. *Drug Saf* **38** (2015)
- ★ Pons E + Becker BFH, Akhondi SA, Afzal Z, van Mulligen EM, Kors JA. *RELigator: Chemical-Disease Relation Extraction Using Prior Knowledge and Textual Information* in *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop* (2015)
- Akhondi SA, Pons E, Afzal Z, van Haagen H, Becker BFH, Hettne KM, Van Mulligen EM, Kors JA. *Patent Mining: Combining Dictionary-Based and Machine-Learning Approaches* in *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop* (2015), 102–109