 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1- Final
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	1/80



www.advance-vaccines.eu

Accelerated Development of VAccine beNefit-risk Collaboration in Europe

D5.5 Ontology for the Integration and Extraction of Vaccine-related Information in Europe: a proof of concept

**WP5 – Proof-of-concept studies of a framework
to perform vaccine benefit-risk monitoring**

v1


Draft date: September 29, 2017

Lead beneficiary: EMC/P95

Date: September 29, 2017

Nature: Report

Dissemination level: PU

 IMM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1- Final
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	2/80


Document information

Grant Agreement Number	115557	Acronym	ADVANCE
Full title	Accelerated Development of VAccine beNefit-risk Collaboration in Europe		
Project URL	http://www.advance-vaccines.eu		
IMI Project officer	Angela Wittelsberger (angela.wittelsberger@imi.europa.eu)		

Deliverable	Number	5.5	Title	An Ontology for the Integration and Extraction of Vaccine-related Information in Europe
Work package	Number	5	Title	Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring


Delivery date	Contractual	Month X	Actual	
Status	Current version / v1		Draft <input type="checkbox"/> Final <input checked="" type="checkbox"/>	
Nature	Report <input checked="" type="checkbox"/> Prototype <input type="checkbox"/> Other <input type="checkbox"/>			
Dissemination Level	Public <input checked="" type="checkbox"/> Confidential <input type="checkbox"/>			

Authors (Partner)	Name	Partner	Section
	Benedikt Becker	EMC	All
	Miriam Sturkenboom	P95	All
	Jan Kors	EMC	All
	Elisa Martin Merino	AEMPS	Vaccine inventory, Code alignment

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1- Final
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	3/80


	Giuseppe Roberto	ARS Toscana	Vaccine inventory
	Sofia Zastavnik	EMA	VaccO
	Klara Berencsi	AUH	Code alignment
	Talita Dualle Salles	SIDIAP	Code alignment
	Harshana Liyanage	SURREY	Code alignment, expert review
	Silvia Lucchi	ASLCR	Code alignment
	Giorgia Danieli	Pedianet	Code alignment
	Zubair Afzal	EMC	Research Proposal
	Jorgen Bauwens	UNIBAS	Research Proposal
	Jan Bonhoeffer	UNIBAS	Research Proposal
	Kartini Gadroewn	EMC	Research Proposal
	Harshana Liyanage	SURREY	Research Proposal
	Olivia Mahaux	GSK	Research Proposal, Eudravigilance application
	Vincent Bauchau	GSK	Research Proposal/Draft
Responsible Author	Benedikt Becker	Email	b.becker@erasmusmc.nl
	Partner EMC	Phone	

Description of the deliverable	
Key words	

 IMM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1- Final	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	4/80

Document history

NAME	DATE	VERSION	DESCRIPTION
Benedikt Becker (EMC), Zubair Afzal (EMC), Jorgen Bauwens (UNIBAS), Jan Bonhoeffer (UNIBAS), Jan Kors (EMC), Kartini Gadroen (EMC), Harshana Liyanage (SURREY), Olivia Mahaux (GSK), David Mullett, Sofia Zastavnik, Miriam Sturkenboom	May 16, 2016		Research plan
Benedikt Becker (EMC)	May 5, 2017	v0.1	Draft outline
Elisa Martin Merino (BIFAP), Giuseppe Roberto (ARS Toscana)	May 10, 2017		Review and expansion of vaccine inventory
Benedikt Becker (EMC)	May 26, 2017	v0.2	
Jan Kors (EMC)	May 29, 2017		Review
Benedikt Becker (EMC)	May 29, 2017	v0.3	
Miriam Sturkenboom	May 30, 2017		Review
Benedikt Becker (EMC)	May 30, 2017	v0.4	
Benedikt Becker (EMC)	June 21, 2017	v0.5	Added VaccO on Eudravigilance
Miriam Sturkenboom, Vincent Bauchau (GSK), Lina Titievsky (Pfizer)	June 23, 2017		Discussion
Benedikt Becker (EMC)	June 27, 2017	v0.6	Draft for project leads
Miriam Sturkenboom, Vincent Bauchau, Olivia Mahaux (GSK)	June 30, 2017		Review
Benedikt Becker (EMC)	July 12, 2017	v0.7	Draft for expert reviewers
Benedikt Becker (EMC)	Sep 13, 2017	v0.8	Update VaccO and code alignment
Harshana Liyanage (SURREY)	Sept 21, 2017		Review
Benedikt Becker (EMC)	Sep 28, 2017	v0.9	Address reviewer comments, describe web applications,
Natasha Yefimenko (Synapse)	Sep 29, 2017	v1	Final version


 IMM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1- Final	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	5/80

Definitions

- Participants of the ADVANCE Consortium are referred to herein according to the following codes:
 - **EMC.** Erasmus Universitair Medisch Centrum Rotterdam (Netherlands) - Coordinator
 - **UNIBAS.** Universitaet Basel (Switzerland) - **Managing entity of the IMI JU funding**
 - **EMA.** European Medicines Agency (United Kingdom)
 - **ECDC.** European Centre for Disease Prevention and Control (Sweden)
 - **SURREY.** The University of Surrey (United Kingdom)
 - **P95.** P95 (Belgium)
 - **SYNAPSE.** Synapse Research Management Partners, S.L. (Spain)
 - **OU.** The Open University (United Kingdom)
 - **LSHTM.** London School of Hygiene and Tropical Medicine (United Kingdom)
 - **PEDIANET.** Società Servizi Telematici SRL (Italy)
 - **KI.** Karolinska Institutet (Sweden)
 - **ASLCR.** Azienda Sanitaria Locale della Provincia di Cremona (Italy)
 - **AEMPS.** Agencia Española de Medicamentos y Productos Sanitarios (Spain)
 - **AUH.** Aarhus Universitetshospital (Denmark)
 - **UTA.** Tampereen Yliopisto (Finland)
 - **WIV-ISP.** Institut Scientifique de Santé Publique (Belgium)
 - **MHRA.** Medicines and Healthcare products Regulatory Agency (United Kingdom)
 - **SSI.** Statens Serum Institut (Denmark)
 - **RCGP.** Royal College of General Practitioners (United Kingdom)
 - **RIVM.** Rijksinstituut voor Volksgezondheid en Milieu * National Institute for Public Health and the Environment (Netherlands)
 - **GSK.** GlaxoSmithKline Biologicals, S.A. (Belgium) – EFPIA Coordinator
 - **SP.** Sanofi Pasteur (France)
 - **NOVARTIS.** Novartis Pharma AG (Switzerland)
 - **SP MSD.** Sanofi Pasteur MSD (France)
 - **JANSSEN.** Janssen Vaccines & Prevention B.V. (Netherlands)
 - **PFIZER.** Pfizer Limited (United Kingdom)
 - **TAKEDA.** Takeda Pharmaceuticals International GmbH (Switzerland)
 - **ARS Toscana** - Agenzia Regionale di Sanità, Toscana (Italy)
 - **IDIAP JORDI GOL.** Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (Spain)

Abbreviations

- **AIRR:** ADVANCE International Research Readiness
- **ASLCR:** Local Health Authority of Cremona database
- **ATC:** Anatomical Therapeutic Chemical Classification System
- **AUH:** Aarhus Universitetshospital databases

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1- Final	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	6/80

- **BIFAP:** Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria
- **CDM:** Common data model
- **DL:** Description logics
- **EHR:** Electronic health record
- **ECDC:** European Center for Disease Prevention and Control
- **MedDRA:** Medical dictionary for regulatory activities
- **OVEA:** Ontology of Vaccine Adverse Events
- **OWL:** Web ontology language
- **Pedianet:** Società Servizi Telematici SRL database
- **POC-I:** First proof-of-concept study in ADVANCE WP5
- **RCGP:** Royal College of General Practitioners database
- **SIDIAP:**
- **SSI:** Statens Serum Institut databases
- **SMQ:** Standardized MedDRA queries
- **THIN:**
- **VaccO:** Vaccine ontology for pharmacoepidemiology
- **VO:** VIOLIN Vaccine ontology
- **VIOLIN:** Vaccine Investigation and Online Information Network



 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	7/80

Table of contents


1. Introduction.....	12
2. The vaccine coding inventory	15
2.1. Introduction	15
2.2. Questionnaire	17
2.3. Conduct of interviews	20
2.4. Participating databases	20
2.5. Results	21
2.6. Discussion.....	27
3. VaccO: An Ontology of vaccine Properties	29
3.1. Ontology.....	29
3.2. Resources for VaccO	32
3.3. Inventory of criteria used for defining vaccine classes	34
3.4. The design of VaccO	35
3.5. Mappings.....	40
3.6. Identification of VaccO classes in free text	41
3.7. Representation of vaccines in VaccO	42
3.8. Property lists	43
3.9. VaccO analysis – an interactive illustration.....	43
3.10. Data distribution	45
4. VaccO application I: vaccine code selectiON.....	47
5. VaccO application II: Automatic alignment of vaccine codes based on multilingual code	

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	8/80

labels.....	50
5.1. Introduction	50
5.2. Methods	51
5.3. Results	56
5.4. Error analysis	60
5.5. User application for vaccine code alignment.....	61
5.6. Application in the ADVANCE framework	63
5.7. Discussion.....	64
6. VaccO application III: Vaccine product classification in a spontaneous reporting system	66
6.1. Introduction	66
6.2. Methods	67
6.3. Results	69
6.4. Discussion.....	70
7. References	73

Table of figures

Figure 1: Overview of the vaccine ontology project	14
Figure 2: Questions of the vaccine data inventory.	19
Figure 3: Visualization of the DL expression describing the class of attenuated Pertussis vaccines	31
Figure 4: Structure of the core VaccO ontology	38
Figure 5: Representation of descriptions of vaccine classes in VaccO	42
Figure 6: Input of the VaccO Analysis application.....	44
Figure 7: The VaccO Analysis application displays the internal representations of a vaccine	

 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	9/80

description	45
Figure 8: The input of the VaccO Selection application	48
Figure 9: VaccO Selection shows the property values in the supplied vaccine coding system, and the list of all codes.....	49
Figure 10: Example for the identification of target codes for a source code.	53
Figure 11: Overall average F-score using different thresholds for the selection of target codes.....	58
Figure 12: F-score for automatic vaccine code alignments.	58
Figure 13: The VaccO Alignment application takes two vaccine coding systems as input.....	62
Figure 14: Example of the output of the VaccO Alignment application for aligning ATC vaccine codes with ADVANCE Vactype.	63
Figure 15: Processing of Eudravigilance data and evaluation.	68

Table of tables

Figure 1: Overview of the vaccine ontology project	14
Figure 2: Questions of the vaccine data inventory.	19
Figure 3: Visualization of the DL expression describing the class of attenuated Pertussis vaccines	31
Figure 4: Structure of the core VaccO ontology	38
Figure 5: Representation of descriptions of vaccine classes in VaccO	42
Figure 6: Input of the VaccO Analysis application.....	44
Figure 7: The VaccO Analysis application displays the internal representations of a vaccine description	45
Figure 8: The input of the VaccO Selection application	48
Figure 9: VaccO Selection shows the property values in the supplied vaccine coding system, and the list of all codes.....	49
Figure 10: Example for the identification of target codes for a source code.	53
Figure 11: Overall average F-score using different thresholds for the selection of target codes.....	58
Figure 12: F-score for automatic vaccine code alignments.	58
Figure 13: The VaccO Alignment application takes two vaccine coding systems as input.....	62
Figure 14: Example of the output of the VaccO Alignment application for aligning ATC vaccine codes with ADVANCE Vactype.	63



	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	10/80

Figure 15: Processing of Eudravigilance data and evaluation.68

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	11/80

Executive summary


The implementation of benefit, coverage, and safety studies on vaccines requires access to a broad range of information, which is spread over various, isolated resources in Europe: For example, vaccinations are recorded in electronic health record databases of different types, the exact composition, documented adverse events, and contraindications are described in authorisation databases, summary of product characteristics (SPCs) and package leaflets. This scattering of vaccine-related information and the differences in representation of vaccine information in EHR databases impede the aggregation of information over resources, which are required to increase the readiness of vaccine benefits, coverage, and safety studies.

The objective of this deliverable is to create a shared representation and tools for identifying vaccine-related information in various resources.

We conducted an inventory on how vaccine-related information is represented in European electronic health record databases, and how the information was processed to match the common input files of the first ADVANCE proof-of-concept study.

We created an ontology – named VaccO – as a common representation of vaccine descriptions in different resources.

We developed three applications of VaccO in the conduction of studies about vaccine benefit/risk assessment: 1) For the analysis and selection of vaccine codes in a given vaccine coding system, 2) for the automatic mapping between vaccine coding systems, and 3) for the extrapolation of vaccine information from a spontaneous reporting system. Web applications were developed to accelerate the conduction of multi-database vaccine studies.


	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	12/80

1. INTRODUCTION

The implementation of benefit, coverage, and safety studies on vaccines requires access to a broad range of information, which is spread over various, isolated resources in Europe: Vaccinations are recorded in electronic health record (EHR) databases of different types (e.g., prescription/administration and primary care, vaccine registries) and using different coding systems for the recording of vaccinations (e.g., general coding systems like ATC [1], Read-v2 [2,3], and custom, database-specific coding systems). The authorizations of vaccines are recorded both in European (centralized procedure) and national medicine registries (national procedure). The exact composition, documented adverse events, and contraindications are described in the authorisation databases, summary of product characteristics (SPCs) and package leaflets. This scattering of vaccine-related information and the differences in representation of vaccine information in different data sources impede the aggregation of information over resources, which are required to increase the readiness of vaccine benefits, coverage, and safety studies.

ADVANCE aims to provide the rapid assessment of benefits and risks of vaccines including the rapid availability of information. With this deliverable we support this vision by developing an ontology of vaccine characteristics – named *VaccO* –, which are relevant in the conduction of vaccine benefit, coverage, and safety studies, and propose and test applications for improving availability of vaccine information using *VaccO*. Figure 1 gives an overview of the different projects. First, we describe the representation of vaccination information in the EHR databases that participated in the ADVANCE vaccine fingerprinting (section 2.). Subsequently we present the design of the *VaccO* ontology (section 3.). Finally, we show three applications of *VaccO*.

- a) A web application (“*VaccO* Selection”) has been developed for the analysis and selection of vaccine codes in a given vaccine coding system (section 4.).
- b) An algorithm based on *VaccO* has been developed and evaluated for the automatic

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	13/80

mapping between vaccine coding systems (section 5.). The accompanying web application (“VaccO Alignment”) can accelerate the conduction of multi-database vaccine studies, where each database has to map its vaccine codes to the coding system defined in the study protocol.


- c) An example of integrating vaccine information from the EUDRAVIGILANCE spontaneous reporting system with information in the VaccO ontology (section 6.).

The scope of this deliverable is the development of an ontology for the conduction of vaccine benefit/risk studies and of algorithms based on the VaccO ontology to make vaccine information more accessible and retrievable.

What is already available?

Several resources about vaccine information are already available. The Vaccine Investigation and Online Network (VIOLIN) published several ontologies for the standardization and integration of vaccine information, including the VIOLIN vaccine ontology (VO) [4] and the Ontology of Vaccine Adverse Events (OVAE) [5]. VIOLIN provides several tools for accessing information about vaccines in a knowledge database including vaccine components, mechanisms, vaccine design and literature mining [6,7]. VO focuses on vaccine products licensed in the US and Canada. However, VaccO focuses on vaccine characteristics and classes that are relevant in the context of pharmacoepidemiologic studies, products that are licensed in the European Union, and text mining applications. The formal and technological background of the VIOLIN vaccine ontology and VaccO are compatible, and integration between the two ontologies is pursued but outside of scope of this deliverable.

Drug databases provide information about vaccine products. The Article 57 database (Art57 DB) [8] implements the article 57(2) of EC regulation No 726/2004, and is issued by the European Medicines Agency (EMA). The Art57 DB provides information about medical products authorised in the European Union, including their scientific composition,

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	14/80

indications, and authorization details. RxNorm [9,10] has a comparable scope of information for clinical drugs authorised in the US, and is issued by the US National Library of Medicine. To supersede country-specific drug vocabularies and Art57 DB, five ISO standards have been developed for the international identification of medicinal products (IDMP) [11,12]. IDMP became a regulation for the recording of drug information for EMA and FDA. The initiative covers the standardized recording of information about medical substances, dose forms, units of measurement, medical products, and pharmaceutical products. Notably, IDMP contains classifications of medical products to a number of medical coding systems, and provides a drug dictionary, a term browser, and a text encoder.

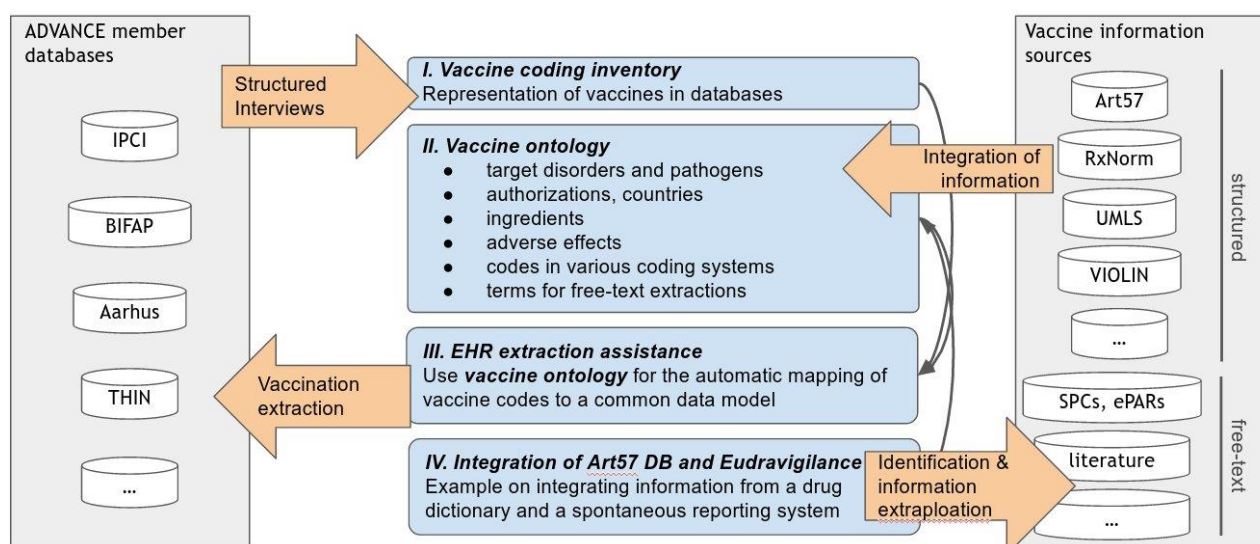



Figure 1: Overview of the vaccine ontology project

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	15/80


2. THE VACCINE CODING INVENTORY

2.1. Introduction

The recording of vaccinations and the availability of different aspects about vaccinations in electronic health care databases (EHR) depend on the characteristics of the specific EHR database. The characteristics include, for instance, the healthcare setting where data is recorded, the purpose of data collection, and the healthcare system organization from which information is captured. For example, information on vaccine administrations can be recorded in primary care databases regardless of healthcare service reimbursement, but with the intention to complete the patient clinical history for health assistance purposes. Conversely, information on vaccinations that are administered in different healthcare settings (such as public health agencies) are unlikely to be consistently available in a primary care database, rather these are recorded in vaccination registries for the monitoring of specific immunization campaigns. Moreover, some details about the vaccination might or might not be recorded in a specific database. For instance, injection site and dose number are usually available in a primary care database while they might not need to be recorded in an administrative/reimbursement database, for which it is only important what the costs are. Data entry strategies vary even between database types: Information may be entered by administrative personnel, doctors, nurses, or even by patients themselves, which may influence the availability and reliability of information.

The characteristics of EHR databases in Europe and beyond have been examined in the AIRR study as part of ADVANCE WP3 deliverable 3.4 (“Catalogue and meta-profiles of data sources for vaccine benefit-risk monitoring”). Reasons for exposure misclassification have been analysed in project 3 of ADVANCE WP4 deliverable 4.4 (“Impact of disease and exposure misclassification on estimation of vaccine effectiveness”).


Beyond the content, the representation of vaccine information recording can differ between databases. Different medical vocabularies are used in European EHR databases for the

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	16/80

recording of vaccines and vaccinations. Representations include medical vocabularies like dedicated drug classification systems (e.g., ATC), general medical coding systems (e.g., Read codes), the target disorders for the vaccine (i.e., varicella immunization), and custom, database-specific vocabularies, as well as free-text information. Differences in the representation of vaccine information between databases hamper the uniform execution of multi-database benefit/risk studies, because available vaccination information needs to be transformed into a common representation, to aggregate information or compare between databases. The transformation of local data to a common representation currently requires the development and application of database-specific algorithms.

Vaccination information was converted by each database participating in the ADVANCE POC-I study into common input files, an anonymized and homogeneous representation that built the basis for the processing and fingerprinting of vaccination information in the POC-I study. The vaccine common input files required a characterisation of the vaccine by immunization targets and discrimination between whole-cell and acellular vaccines in the case of Pertussis, which was represented by a vocabulary called *ADVANCE Vactype*. The vocabulary contained 28 codes for individual immunization targets (e.g., code *INF* for Influenza vaccines, and *aPE* for acellular Pertussis vaccines) that could be combined for representing combination vaccines. The common input files further contained the ATC code, product name, and recorded dose if available, and a derived dose if the dose was not available (see the ADVANCE WP5 deliverables D5.2 and D5.4 for further details about the common input files).

We developed a questionnaire to examine how vaccine-related information is represented in EHR databases, and how the information was processed to match the common input files of the POC-I. Structured interviews were conducted with the custodians of all EHR databases (N=8) that participated in the vaccine fingerprinting for the POC-I study of WP5 by October 2016. The resulting inventory of vaccine-related information helped the development of tools and information sources for future extraction of vaccination


	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	17/80

information from the concerned databases.

2.2. Questionnaire

A questionnaire was developed to examine how vaccine-related information is represented in EHR databases, and how the information was transferred into the vaccine data file that was used in ADVANCE.

The questionnaire contains thirteen questions which are shown in Figure 2. Questions Q1 and Q2 investigate the general layout of the EHR databases and the tables that contain information about vaccinations. The questions of Q3 investigate the settings of the data recording. Questions Q4-Q6 investigate availability, location and representation of vaccination information in the database. Questions Q7-Q9 investigate the extraction and transformation of vaccination information to match the ADVANCE common data model. Question 9 investigates the resolution of ambiguity in vaccination information. Questions Q10-Q13 investigate the local vaccination programme. The full questionnaire is available on SharePoint under [WP5 Documents – WP5.5 Vaccine ontology – Vaccine inventory – Vaccine data inventory Questions](#).

 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	18/80

Storage of Vaccination Data in the ADVANCE Member Databases

Questionnaire for the vaccination data inventory

- Q1. Please provide a list of tables in your database and their relationships.
- Q2. Which are the tables of your database that collect information on vaccinations?
- Q3a. What is the purpose of the recording?
- Q3b. Which events trigger the creation of records in the tables?
- Q3c. Is the registration of the triggering event compulsory? Are all fields of the table compulsory?
- Q3d. Does the event recording take place in a specific healthcare setting?
- Q3e. Who enters the information?
- Q3f. Which coding terminology is used?
- Q3g. Were there changes in data recording over time? When?
- Q3h. Does the table contain free-text field?
- Q3i. Which language is used in the free-text?
- Q3j. How do you extract structured data from free-text content?
- Q4. Which vaccine types (VPD) are captured? You can proceed to the next question if you already provided this information in the AIRR survey.
- Q5. Please describe which information is available in your database (Date, brand name, vaccine target, ATC code, dose, manufacturer, lot/batch number, injection site, vaccinating facility), in which table it is stored and how it is represented. All vaccine-specific exceptions of availability and representation of information should be indicated.
- Q6. Which other vaccination-related information are captured in the database?
- Q7a. How did you extract the vaccine type for the vaccine fingerprinting (variable Vactype in the CDM)? Which fields from your database did you use, and how did you assign the vaccine types?
- Q7b. If ATC classifications are unavailable in your database, how did you assign ATC codes to the vaccinations? Which problems arose in the process or prevented you from providing ATC codes?
- Q8. How exactly did you extract vaccination doses for the CDM?
- Q9. If one vaccination triggers several entries in distinct data domains/tables of the database (e.g., prescription and administration), how do you use the different components of the information on vaccine exposure in vaccine studies?
- Q10. Where are children vaccinated in your country for officially scheduled vaccines?
- Q11. Where do people receive their travel vaccines? Do they end up in your database?
- Q12. Who is vaccinated against seasonal influenza (eligible criteria)?
- Q13. How is influenza vaccination delivered in your country, and do you capture the data?



	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	19/80

Figure 2: Questions of the vaccine data inventory.

 IMi - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	20/80

2.3. Conduct of interviews


The questionnaire was first sent to the database custodian and filled by him or her, allowing the database custodian to consult other members in their team when necessary. The answers were then discussed in in-person meeting or on telephone where the questionnaire was edited collaboratively by the interviewer and the database custodian using Google Documents.

2.4. Participating databases

The questionnaire was filled for each database that participated in the vaccine fingerprinting of POC-I by October 2016. Two databases are each located in Spain, Denmark, and United Kingdom, and three databases were located in Italy. They provide health records from primary health care, inpatient care and outpatient care, reimbursement claims, and vaccine recordings (see Table 1).

Table 1: Databases in vaccine data inventory. Database types were captured in the AIRR study.

Database/partner	Country	Type
ES-SIDIAP	Spain	Primary health care, Vaccination registry, Pharmacy dispensing records
ES-BIFAP	Spain	Primary health care
DK-SSI	Denmark	Outpatient, Inpatient EHRs, Reimbursement, Disease surveillance, Vaccination registry, Population data, Vital records, Pharmacy dispensing records, Specialised care
DK-AUH	Denmark	Reimbursements, Quality assurance, Administration of health services
UK-THIN	United Kingdom	Primary care, Outpatient EHRs, Prevention records
UK-RCGP	United Kingdom	Primary care, Disease surveillance
IT-ASLCR	Italy	Reimbursement, Vaccination registry, Pharmacy dispensing records
IT-Pedianet & Venetia regional database	Italy	Primary care, Outpatient EHRs, Vaccination registry, Vital records, Pharm. dispensing records, Specialized care, Adverse event reporting, Population health surveys, Health care costs

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	21/80

2.5. Results

All completed questionnaires and a Excel spreadsheet summarizing the results are available on the ADVANCE SharePoint under [WP5 Documents – WP5.5 Vaccine ontology –Vaccine inventory](#). Relevant vaccination information was generally dispersed over several tables, and a table with patient information was used to link information from different tables. Most databases contained one or more dedicated tables to record vaccine administration (SIDIAP, SSI, AUH, ASLCR, Pedianet, Venetia DB), and vaccinations were recorded in tables together with other information in THIN and RCGP (see Table 2).

The purpose of data recording was reported as the creation of personal health records (SIDIAP, BIFAP, SSI, THIN, RCGP), drug reimbursements (SSI, AUH), reimbursements of health services (AUH), vaccine monitoring (SSI, ASLCR, Pedianet + Regional DB), (commercial) research (THIN), recording of prescriptions (RCGP) (see


	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	22/80

Table 3). Vaccine administrations are usually recorded for all non-travelling vaccines (see


	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	23/80

Table 4). Only Pedianet records vaccination of only one immunization target (Influenza). Vaccinations are described by varying characteristics in the databases. Neither vaccine brand names, immunization targets, ATC codes, manufacturers, lot numbers, injection sites, nor vaccination facilities are available in all databases. Most information about vaccines such as brand names, immunization targets, doses, lot numbers, and vaccination facilities is recorded in database-specific custom codes (see


	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	24/80

Table 5). Free text was considered in BIFAP (Spanish free text), ASLCR (Italian), and Pedianet and regional databases (Italian), and processed using keyword queries (see


 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	25/80


Table 6).

The immunization targets were extracted for conversion to the ADVANCE Vactype coding systems from brand name (ASLCR, PEDIANET), market authorization number (ASLCR), ATC codes (SSI), custom codes (SIDIAP, PEDIANET), and using existing lookup tables (ATC codes from a vaccine list by EMA in THIN). Algorithms working on the labels of a custom coding system were developed and applied in BIFAP after validating all potential sites in the database with recordings of vaccination information.

Dose information had been validated already during data recording by rejecting duplet doses in SSI. In other databases where dose information was not available it was derived from the age and dose counts based on national vaccination schedules (ASLCR, SIDIAP, BIFAP, AUH, Pedianet).

Table 2: Databases tables with vaccination information. Table names in italic link with other tables, table names set in bold contain information about vaccinations.

Database	Tables
SIDIAP	<i>Patient</i> , Vaccine , Diagnosis, Hospitalization, Lab tests, Prescriptions, Dispensations, + 13 other
BIFAP	<i>Patient</i> , Vaccines , Prescriptions , Dispensations , General data , Comments , Visits, Clinical episodes, Referrals, Radiologies, Side effects, Laboratory results, Prevalent conditions, Prevalent antecedents
SSI	<i>Patient</i> , Danish vaccination registry (DDV) , Notifications infectious diseases, hospitalizations & diagnosis, microbiological test results, prescriptions/dispensations
AUH	<i>Patient</i> , DB of reimbursed prescriptions , registry of patients , service registry , civil registration system
THIN	<i>Patient</i> , Additional Health Data (Codes, comments, AHD codes), Therapy (Pack, Drug, Dosage), Medical (Codes, Comments)
RCGP	<i>Patient</i> , Prescription , Event, NHS organisation
ASL Cremona	<i>Patient</i> , Vaccine administration
Pedianet	<i>Patient</i> , Influenza vaccine administration , Death, Procedures, Measurements, Drugs, ER access, Test results, Hospitalizations, Visit to pediatrician

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	26/80

Venetia DB	<i>Patient</i> (linked to Pedianet), Vaccination administration
------------	--


 IMi - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	27/80

Table 3: Tables with vaccination information and their recording purposes

Database	Table	Purpose of recording
SIDIAP	Vaccine	Register of health professional in the electronic health record of the patient
BIFAP	Vaccines, Prescriptions, Dispensations, General data, Comments	Creation of individual electronic clinical history
SSI	Danish Vaccination Register (DDV)	Personal health record, reimbursement, vaccine monitoring (coverage, safety, effectiveness)
AUH	DB of Reimbursed Prescriptions (DNDRP), Registry of Patients (DNRP), Health Service Register (DNHSR)	Drug reimbursement (DNDRP), Administration of health services (DNRP), Reimbursement of health services (DNHSR)
THIN	Additional Health Data	(Commercial) research, monitoring, GP recording
RCGP	Prescription	Recording of prescriptions
ASL Cremona	Vaccine administration	Vaccine monitoring
Pedianet + Regional DB	Influenza vaccination administration (Pedianet) + Vaccine administration (Regional DB)	Clinical practice and care, vaccine monitoring


 IMl - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	28/80

Table 4: Available data fields about vaccinations (Y=mandatory, O=optional, -=not available)

Database	Vaccines	Brand name	Imm. targets	ATC	Dose	Manufacturer	Lot / batch	Inj. site	Vacc. facility
SIDIAP	All	-	Y	-	Y	-	-	Y	Y
BIFAP	All	Y (not always)	Y	-	Y (not always)	Y (not always)	Y (not always)	O	Y (not always)
SSI	All (since 2015)	Y	Y	Y	Y (pediat.)	-	Y (since 2015)	-	Y
AUH	All (since 2015)	Y (DNRP)	Y	Y (DNRP)	-	O (DNDRP)	-	-	O (DNRP, DNHRs)
THIN	All	-	Y	-	O	-	-	-	-
RCGP	BCG, HiB, HPV, Pert, Polio, Infl	Y	-	-	O	-	O	O	O
ASL Cremona	All but BCG, Mening, Infl, Herpes Zoster	Y	O	O	-	Y	O	-	-
Pedinet	Infl	O	Y	-	O	-	O	Y	Y
Regional DB	All but HepA, Herpes zoster	O	Y	-	Y	-	O	Y	Y


 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	29/80

Table 5: Representation of vaccinations, and possible conversion to ADVANCE common input files

Database	Brand name	Imm. targets	ATC	Dose	Manuf.	Lot/ batch	Inj. site	Vacc. facility
<i>SIDIAP</i>	-	?	-	?	-	-	?	?
<i>BIFAP</i>	Text	Custom codes	From custom codes	Text	Text	Text	Text	
<i>SSI</i>	Custom codes	Local danish codes	From vacc type, brand	Reimbursement codes	-	?	-	National register codes
<i>AUH</i>	Custom codes		ATC (DNDRP), Custom codes (DNRP, DNHSR)		Custom codes			Custom codes
<i>THIN</i>	-	From AHD	-	Custom code	-	-	-	-
<i>RCGP</i>	EMIS	-	-	?	-	Custom code	Read v2	Custom code (practice ID)
<i>ASL Cremona</i>	Text	Text	ATC		Text	Text	-	-
<i>Pedianet</i>	Text	Custom codes	-	-	-	Text	-	Text
<i>Venetian DB</i>	Custom codes	Custom codes	-	Integer	-	Text	Custom code	Custom code



	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	30/80

Table 6: Use of free-text in populating the ADVANCE common input files

Database	Free text	Extraction
SIDIAP	-	
BIFAP	Spanish	Keyword queries
SSI	-	
THIN	- (available on demand, not used)	
RCGP	-	
ASL Cremona	Italian	Keyword queries
Pedianet + regional DBs	Italian	Keyword queries + manual review

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	31/80


2.6. Discussion

Most databases that participated in the vaccine coding inventory are public, primary care databases. Primary care is the official setting for systematic, paediatric vaccinations in most countries, which means that participating databases potentially contain all vaccinations on children in the covered areas. The Danish SSI database has been set up to record all vaccinations in the country.


Significant differences in the representation of vaccines and vaccinations were observed between databases. Vaccines were characterised by brand names, immunization targets, and ATC codes, but no vaccine properties were recorded in all databases. Most information was represented in database-specific vocabularies. Mappings to general drug vocabularies or coding systems were missing for the custom vocabularies, and had to be created for the conduction of the ADVANCE POC-I study. Free-text EHRs were not used in any database for providing information about vaccinations in the ADVANCE common input files. Keyword queries were used on custom codes for preparing the creation of mappings from the custom coding systems to the ADVANCE common input files.

Databases where dose information was unavailable developed algorithms for deriving dose information. The algorithms were developed individually by the databases but were similar in being based on the age of the vaccinees, the order of vaccine administrations, and the national vaccination schedule. The vaccination schedules of European countries are made available by the European Centre for Disease Prevention and Control (ECDC) (<http://vaccine-schedule.ecdc.europa.eu>).

The heterogeneous representation of vaccines hampers the rapid implementation of vaccine benefit, coverage, and risk studies, because the implementation requires the mapping of custom vaccine coding systems to the common input files. The integration of a new database requires the creation of a mapping of another custom vaccine vocabulary to the common input files. On top of that, the representation of vaccines in the common input files must be chosen as a smallest common denominator of information available in the

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	32/80

participating databases. We propose an algorithm for the automatic mapping of custom coding systems to suitable vocabularies in common input files in section 5.

 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	33/80

3. VACCO: AN ONTOLOGY OF VACCINE PROPERTIES


We created a vaccine ontology – named *VaccO* – to assist the development of study protocols and the extraction of information about vaccinations from electronic health records, spontaneous reporting systems, and scientific literature. The ontology serves as a common representation of vaccine descriptions in different resources, and helps to integrate information across resources. The taxonomic hierarchy of the ontology allows for aggregating information about vaccines on different levels, combining information and evidence from specific vaccines when needed. The background knowledge encoded in the ontology and information from linked ontologies can be provided in vaccine-related workflows.

In subsequent sections we describe the construction of *VaccO* and show example applications of *VaccO* targeting the following tasks:

1. Provide a website for simple selection of ATC vaccine codes;
2. Facilitate extraction of vaccine information from EHR databases by creating automatic alignments between vaccine coding systems (i.e., between database-specific vocabularies and a common data model);
3. Complete coding of vaccines and information on vaccines in a spontaneous reporting database.

3.1. Ontology


An ontology is a formal and computer-readable description of the entities in a thematic domain including their properties, interrelations, and taxonomic hierarchy [13,14]. An ontology of the vaccine domain comprises for example *vaccines*, *vaccine components*, *ingredients*, *manufacturers*, and *immunization targets*. The domain entities are categorised by classes that are defined by common characteristics of its entities, for example the class of *viral vaccines* comprised of all vaccines targeting disorders that are caused by viruses.

 IMM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	34/80

Interrelations between classes and entities are described by (object) properties. For example, a property *contains* links a vaccine and its ingredients. Classes are arranged in a taxonomic hierarchy, which allows for aggregation of information across elements of those classes. For example, specifying that the class *influenza* is as a *subclass* of class *virus vaccine* states that each influenza vaccine is also a viral vaccine. Classes can be linked to third-party ontologies, which facilitates the integration and exchange of information between ontologies. Numerous biomedical ontologies are available in the NCBO BioPortal [15] and the OBO Foundry [16].

The de facto standard for semantic web and linked data technologies is the Web Ontology Language (OWL) [17]. Classes in OWL are linked by assertions of equivalences and subclass relations. Classes are defined using expressions of description logics (DL) [18,19]. A DL-expression can have the following forms, and is interpreted as sets of individuals (classes are denoted by capitalized, italic letters or words, properties by uncapitalized letters or words, and keywords of DL expressions are underlined words):

- The name of a class (*C*), whose interpretation is the set of members of class *C*.
- An intersection of one or more DL-expressions (*C and D*), whose interpretation is the set of individuals that are member of the interpretations of the given DL-expressions. The keywords that can be used synonymously: (*C and D* is equivalent to *C that D*)
- A disjunction of one or more DL-expressions (*C or D*), whose interpretation is the set of individuals that are member of the interpretations of at least one of the DL-expressions
- A negation of another DL-expression (*not C*), whose interpretation is the set of all individuals that are not in the interpretation of the given DL-expression.
- An existential restriction (*p some C*) describes the property values of a class, and is interpreted as the individuals where some filler of property *p* is in the interpretation

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	35/80

of C.

- f) A universal value restriction ($p \text{ value } C$), whose interpretation are the individuals where all fillers of property p are in the interpretation of C .

For example, the set of attenuated Pertussis vaccines is defined by the DL-expression “Vaccine that has-component (Vaccine component that immunizes-against some Pertussis and has-type some attenuated”. The DL-expression is visualized in Figure 3 using the Graffoo specification [20].

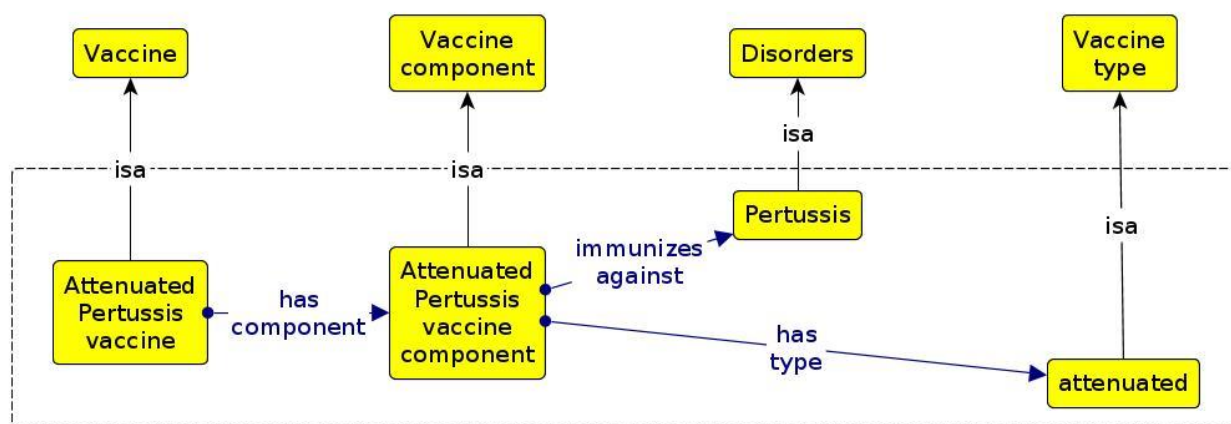



Figure 3: Visualization of the DL expression describing the class of attenuated Pertussis vaccines.

The definition of a class in VaccO contains one or more terms (words and phrases) that describe the class in free text. The set of all terms in an ontology is called its “dictionary”. The dictionary forms the basis of the automatic identification of classes in free text. Classes in VaccO that are created for improving the structure can also be marked for being ignored in the concept identification process.

Ontologies are created to formalize the knowledge within a given domain as a categorization of the domain entities into classes, their relations, and the taxonomic hierarchy. However, the categorization and taxonomization is not intrinsic to the domain


	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	36/80

but depends on the desired perspective, which in turn affects the conclusions that can be drawn from the ontology. For example, an ontology of birds that is based on appearance and behaviour may imply the new world vulture is closely related to the old world vulture. But in an ontology based on genetics, the new world vulture is instead closer related to storks! The design of an ontology represents always a specific perspective on the domain. Because of the ADVANCE mission and vision VaccO has been designed to capture information and concepts about vaccines that are relevant in the conduction of pharmacoepidemiologic studies.

3.2. Resources for VaccO

Information about vaccines and vaccine class criteria is distributed about various existing resources (see Table 9). Detailed information about vaccine products that are authorized in the European Economic Area is provided by the “Article 57 database” (Art57 DB). The Art57 DB is issued and maintained by the European Medicines Agency (EMA) and implements the requirement of marketing-authorization holders for medicinal products for human use to submit and maintain medicinal product information to the EMA (Article 57(2) of Regulation (EC) No 726/2004). The Art57 DB contains information about all nationally authorized products in Europe and products that were centrally authorized for all European countries. Information is provided on the level of marketed authorizations of vaccines and includes amongst others the authorization country, authorization date and possibly the date of withdrawal, different product names (long product name, short product name, and generic name, if any), the scientific composition of the vaccine (active ingredients, excipient, and adjuvant), the pharmaceutical form, the administration route, an ATC code, and the indications (MedDRA terms).

The Vaccine Investigation and Online Information Network maintains a vaccine ontology (VO) [4] that captures vaccines, documented adverse events [5,21], adjuvants, components, and manufacturers of vaccines. VO defines a comprehensive taxonomy of

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	37/80

vaccine products and related categories, and is formalized in the standardized format OWL which facilitates exchange with other information ontologies. However, VO provides information only about vaccines authorized in the United States and Canada.

The Medical Subject Headings (MeSH) provide a hierarchically organized terminology for indexing and cataloguing of biomedical information. MeSH is created and updated by the United States National Library of Medicine (NLM) and is used to catalogue scientific articles in the PubMed database. MeSH contains over 25,000 subject headings that are arranged in a (multi-)hierarchy, together with a short description of each heading. MeSH descriptions of vaccine subject headings contain prevented disorders or immunized pathogens and the MeSH descriptions of disorders often refer to causative agents. These interrelations will be used in VaccO for specifying the immunization targets.

The Unified Medical Language System (UMLS) [22] is a metathesaurus of more than 150 clinical terminologies, including classification systems for medications like RxNorm and ATC. The UMLS is published by the National Institute of Health. Equivalent codes in different terminologies are identified and assigned to concept unique identifiers (CUIs). The UMLS provides terms (words and phrases) for CUIs by aggregating the terms from the source terminologies. Multilingual source terminologies provide translations of terms in different languages. The UMLS defines a taxonomic hierarchy on top of its CUIs, and CUIs are categorized by their semantic types (e.g., organism, organ, substance, or disease). The UMLS contributes existing mappings between medical vocabularies and terms for creating an ontology dictionary.

Additional vocabularies were used as the basis of the classes and terms covering administration routes [23] and for the definition of classes for common combination vaccines and their abbreviations, for example *MMR* for Measles, Mumps, Rubella vaccines [24].


 IMi - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	38/80

Table 7: Information resources for VaccO and their content. Content categories in bold letters are integrated into VaccO.

Resource	Available information
Art57 DB	Marked authorisation, Manufacturer, Names, Strength, Form, Active ingredient, Adjuvant, Pharm form, Route, ATC, Indication
VO	Pathogen, Disease (1), Licensure, Host, Manufacturer, Distributor, Route, Part (has part), Antigen type (has quality), Role, Subunit antigen, Allergen (obsolete), Contraindication, Form (specified input), Route, Licensure
MeSH	Vaccine targets, associated disorders and microbes
FDA, CDC vocabularies	Routes, combination vaccine abbreviations
UMLS	Existing code mappings, terms for concept identification

3.3. Inventory of criteria used for defining vaccine classes

A vaccine code in a medical coding system stands for an individual vaccine product, or defines a group of vaccines based on some common properties. To prepare the creation of the VaccO ontology, we analysed vaccine code descriptors in the following general, drug-specific, and custom database-specific coding systems to identify the categories of properties that are used for defining vaccine codes: Additional Health Data (AHD) used in the THIN database, ATC, BNF, MeSH, Read-2, and SNOMED-CT. The most frequent categories were the immunization targets, i.e. pathogens and disorders (see Table 7). Pathogen strains and broader groups of pathogens or disorders (e.g. bacteria and viruses) were used in most coding systems. Other categories are vaccine types, abbreviations of combination vaccines (e.g., “DTwP” for Diphtheria-Tetanus-wholecell-Pertussis), ingredients (including adjuvants and active ingredients), administration routes, vaccine components (of combination vaccines), and valences, which denote either the number of pathogen strains targeted by a vaccine or the number of vaccine components.


 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto		Security: PU 39/80


Table 8: Criteria used to define vaccine classes in medical vocabularies. A cross (X) indicates that the criterion in the column is used for defining vaccine codes in the vocabulary on the row. The digit 1 indicates that only one concept of the category is used.

Coding system	Pathogens	Pathogen categories	Strains	Disorders	Disorder types	Active ingredients	Valences	Vaccine types	Adjuvants	Routes	Vaccine abbreviations	Populations	Other
ATC	X	X	X	X			X	X		X			
MeSH	X	X		X		1	X	X	1	X	X		
SNOMED-CT	X	X	X	X	X			X					
THIN drugcodes	X		X	X		1				X	X	X	Strength
AHD	X		X	X		1		X		X	X	X	Antigen amount (low/high), Vaccine type (combined)
BNF	X			X		1					X		
Read2	X		X	X		1	X	X		X	X	X	Brand, Dose, Antigen amount, Administrator
Pedinet				1	1			1					Brand
Venetian DB	X			X								X	
SIDIAP				X	X							X	Age group
ADVANCE Vactype	X			X				1					

3.4. The design of VaccO


The aim of VaccO is to provide a formal, computer-readable description of the vaccine properties that are relevant in the context of vaccine coverage, benefit and risk studies, rather than to model all aspects of immunology and the vaccine manufacturing process. As such, VaccO defines only classes for criteria that are used in the definition of vaccine codes as identified in the inventory of criteria for defining vaccine classes. The fundamental classes of VaccO are not vaccine products but their characteristics.

The fundamental classes of VaccO are the property categories identified in the previous section (see Figure 4). More specific classes, subordinated to the fundamental classes,

 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	40/80

and their terms were compiled from the following resources to allow the representation of definitions of vaccine class in VaccO:

- Disorders, pathogens and their relationships were extracted from the descriptors of MeSH concepts. Terms for disorders and pathogens were compiled by identifying equivalent concept unique identifiers CUIs in NLM's Unified Medical Language System (UMLS) [22], and selecting terms for these concepts from the vocabularies Consumer Health Vocabulary (CHV, [25]), ICD-10 CM [26], MedDRA, MeSH, NCBI taxonomy, and SNOMED-CT.
- Vaccine types in VaccO were extracted from descriptions in literature [27–29], classes in the VIOLIN ontology, and MeSH codes. Terms for vaccine types were manually compiled from the same resources.
- Vaccine products and their manufacturers and ingredients were integrated from the Art57 DB. In Art57 DB, a product that is authorized in different European countries has one identifier for each authorization and possibly differing product names. Similarly, a manufacturer has a different identifier in each country where it received a marked authorization. Products and manufacturers from the Art57 DB were grouped to give each manufacturer with authorizations in different countries and each product that is authorized in different countries a shared identity. The 3655 country-dependent vaccine products from the Art57 DB were grouped into 178 product groups. The 114 manufacturers from Art57 DB were grouped into 33 manufacturer groups.
- Administration routes were identified in the Art57 DB, VIOLIN ontology, and terms were compiled from literature and an FDA monograph [23].
- Vaccine combinations and abbreviations and their definitions were compiled from vaccine literature and a CDC monograph [24].

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	41/80

- Classes for valences (monovalent up to “30-valent”) were created with automatically generated terms (e.g., “5-valent”), and manually defined terms for valence 1-10 (e.g., “pentavalent”).

Object properties are used in VaccO to define the relations between pathogens and disorders (*causes*, *caused-by*), between vaccines and immunization targets (*has-target*), types (*has-type*), routes (*has-route*), components (*has-component*), ingredients (*has-ingredient*), and valences (*has-valence*).


Property chains were defined to propagate the properties along other properties. 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). Other property chains in VaccO state that a vaccine type and an immunization target are propagated over ingredients and components:

- *has ingredient* ◦ *has type* ⇒ *has type*
- *has component* ◦ *has type* ⇒ *has type*
- *has ingredient* ◦ *immunizes against* ⇒ *immunizes against*
- *has component* ◦ *immunizes against* ⇒ *immunizes against*

Vaccines that immunize against a given microbe immunize also against the disorders that are specified as caused by that microbe, and vice versa:

- *immunizes against* ◦ *causes* ⇒ *immunizes against*
- *immunizes against* ◦ *is caused by* ⇒ *immunizes against*

VaccO has not been aligned with an upper level ontology such as the Basic Formal Ontology. The VaccO ontology has been created as an application ontology in the context of multi-database vaccine studies, specifically for identifying and representing descriptions of vaccines and vaccine groups. The VIOLIN ontologies (VO, OVAE) provide an

 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	42/80

exhaustive domain ontology of vaccines. But lacking a dictionary and classes for vaccine properties, they are not applicable for identifying and representing vaccine descriptions. An alignment between VaccO and the VIOLIN ontologies and an upper level ontology could be created if the semantic interoperability gained by the alignments is beneficial for the applications.

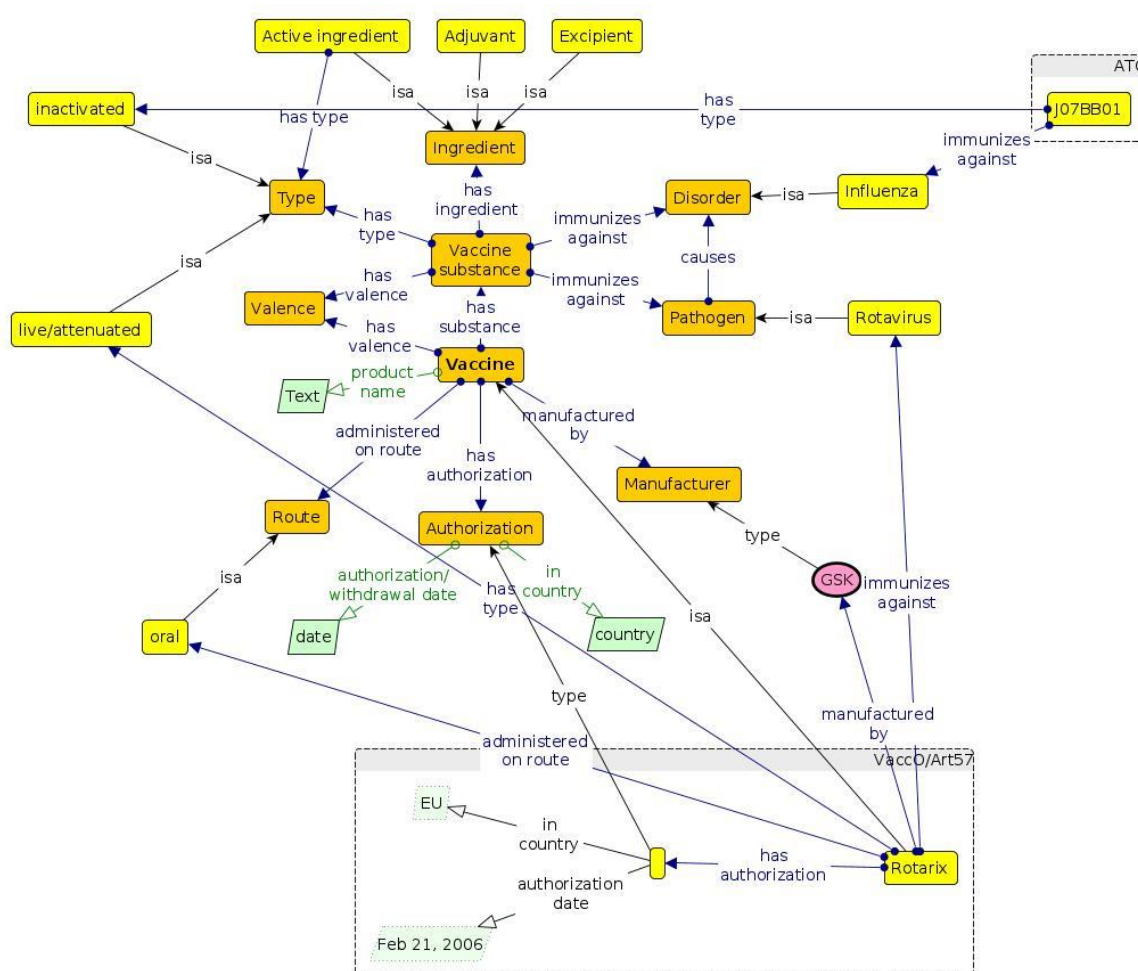


Figure 4: Structure of the core VaccO ontology, with fundamental classes representing property categories in orange and concrete subclasses in yellow. Also shown are examples for the representations of a vaccine product from the Article 57 database (Rotarix), and of a vaccine class defined by ATC code J07BB01 (“Influenza, inactivated, whole virus”). Only concrete classes that have a relation with are only shown in yellow. Visualization follows the Graffoo specification.


 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	43/80

Figure 4 also contains examples on how vaccines are represented in VaccO. The ATC code *J07BB01* defines the class of inactivated influenza vaccines (“influenza, inactivated, whole virus”). It is represented by the DL expression “*Vaccine that immunizes against some Influenza and has type some inactivated*”. The product (GSK’s *Rotarix*) is a vaccine product and thus shares all properties of a vaccine. It is specified to be immunizing against the Rotavirus, has a live/attenuated vaccine type, is manufactured by GSK and has a European market authorization since February 2006.

The VaccO ontology contains 320 vaccine classes with 731 terms (see Table 9): 206 vaccine products, 36 classes of common vaccine abbreviations, and ad-hoc classes for each vaccine component (e.g., *Pertussis vaccines*), administration route (e.g., *Oral vaccines*), and vaccine type (e.g. *Attenuated vaccines*) in VaccO. The product classes are related to classes for ingredients, which are categorized as active ingredients (310 classes), excipients (170 classes), and adjuvants (21 classes, some ingredients have multiple roles in Art57 DB). Classes for nine vaccine types were created with 34 terms: attenuated, conjugate, subunit, inactivated, polysaccharide, recombinant, synthetic, and toxoid. The 109 pathogen classes in VaccO contain 857 English terms. Pathogens were categorized by their biological domain: bacteria (65 classes), eukaryotes (6), and viruses (42 classes), including 42 classes for pathogen strains. The 49 classes for disorders contain 755 terms. Nine classes for administration routes were defined with 23 terms. VaccO contains 67 ad-hoc classes for vaccine components based on their pathogen categories, immunization targets, which unify pathogens and disorders when they are in causal relation (e.g., *Cholera vaccine component* is defined as equivalent to *Vaccine component that immunizes-against Cholera disorder and immunizes-against Vibrio cholerae pathogen*).



 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	44/80

Table 9: Number of direct subclasses and subclasses in VaccO and number of terms in the associated dictionary.

Fundamental class	Classes	Terms
Vaccine	321	706
Ingredient	497	505
Vaccine type	9	35
Pathogen	67	733
Disorder	49	755
Administration route	9	23
Vaccine component	67	68
Valence	30	71

3.5. Mappings

To facilitate the interchange of information with and between other information sources, VaccO integrates references and equivalence annotations to other information sources. Disorders, pathogens and strains refer to codes in MeSH, and UMLS CUIs where applicable. Administration routes refer to FDA codes as defined in the FDA data standards manual. VaccO classes for administration routes, vaccine targets and vaccine types are identified with their equivalences classes from VO [30]. Products and ingredients refer to their corresponding entries in the Art57 DB.

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	45/80


Other drug vocabularies like IDMP and Rx-Norm are currently not part of VaccO but could be integrated with VaccO by specifying equivalences between classes in VaccO and classes in IDMP and Rx-Norm.

3.6. Identification of VaccO classes in free text

The set of all terms assigned to the classes of an ontology is called the ontology dictionary. The ontology dictionary constitutes the basis for identifying VaccO classes in free text. Each occurrence of a term from the dictionary in an input text is considered a reference to the associated class. The classes identified in an input text t will be referred to as $C(t)$. For example, the input text “Oral, attenuated polio vaccine, bivalent” contains references to the the classes of the poliomyelitis disorder, the vaccine type attenuated, and bivalence, and the oral administration route class (references are marked in Figure 5, top left).

The identification of classes in VaccO in free text is based on the ontology dictionary, i.e. the set of terms associated with all classes in the ontology. Each occurrence of a term from the dictionary in an input text is considered a reference to the associated class. The classes identified in an input text t will be referred to as $C(t)$, and each identified class belongs to a category of vaccines, pathogens, disorders, routes, or vaccine types. For example, the input text “Pertussis, inactivated, whole cell, combinations with toxoids” contains references to the disorder *Pertussis*, the vaccine type *Inactivated*, and the vaccine combination class *DT* of common toxoid vaccines (see Figure 5, top left).

The identification of classes from VaccO is based on the Apache Solr text search platform [31]. A document was created in the Solr database for each term in the VaccO dictionary. This document contained the term itself and the class identifier. A Solr plugin for concept identification [32] is used to identify occurrences of terms from the dictionary in the input text.

 IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	46/80

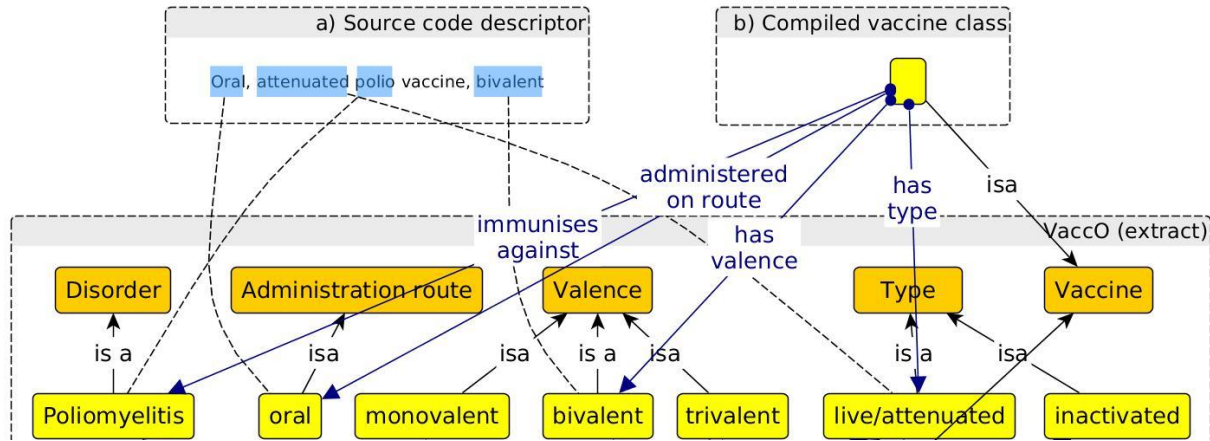


Figure 5: Representation of descriptions of vaccine classes in VaccO. a) Occurrences of VaccO classes are identified in the description. b) The identified VaccO classes are compiled into a vaccine class in VaccO.

3.7. Representation of vaccines in VaccO

A class is defined in an OWL ontology by an DL expression characterising its relations to other classes. To represent the free-text description of a vaccine in VaccO, we first compile each class identified in the description into a DL expression using an algorithm denoted by $\llbracket \cdot \rrbracket$, and defined as follows:


$$\llbracket \text{Vaccine } v \rrbracket \rightarrow v$$

$$\llbracket \text{Pathogen } p \rrbracket \rightarrow \text{Vaccine } \underline{\text{that}} \text{ immunizes-against } \underline{\text{some}} \text{ } p$$

$$\llbracket \text{Disorder } d \rrbracket \rightarrow \text{Vaccine } \underline{\text{that}} \text{ immunizes-against } \underline{\text{some}} \text{ } d$$

$$\llbracket \text{Router } r \rrbracket \rightarrow \text{Vaccine } \underline{\text{that}} \text{ is-administered-on-route } \underline{\text{some}} \text{ } r$$

$$\llbracket \text{Type } t \rrbracket \rightarrow \text{Vaccine } \underline{\text{that}} \text{ has-type } \underline{\text{some}} \text{ } t$$

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	47/80

For example, the vaccine class *Pediarix* is compiled to itself (a class being a DL expression), and the disorder class *Tuberculosis* is compiled to the DL expression *Vaccine that immunizes-against Tuberculosis*.

A set of VaccO classes is then compiled into the conjunction of compiled VaccO classes:

$$\llbracket \{c_1, c_2, \dots\} \rrbracket := \llbracket c_1 \rrbracket \text{and} \llbracket c_2 \rrbracket \text{and} \dots$$

Finally, the representation of a vaccine description t in VaccO is the class defined as equivalent to the DL expression $\llbracket \mathcal{C}(t) \rrbracket$. For example, the vaccine class for the descriptor “oral, attenuated polio vaccine, bivalent” is defined as equivalent to the DL expression “*Vaccine that has-route oral and has-type attenuated and immunizes-against Poliomyelitis and has-valence 2-valent*” (see Figure 5). Representation of vaccines in VaccO using properties


3.8. Property lists

A vaccine class can be converted into a property list. A property list maps each fundamental class in VaccO to the identifiers of its classes that are in relation with the vaccine class, unifying pathogens and disorders as immunization targets. For example, the property list for the descriptor “DTwP” is [target: diphtheria, tetanus, pertussis; type: inactivated]. The property lists of a vaccine description based on pathogens (“Influenzavirus vaccine”), disorders (“Flu vaccine”), abbreviations (“IIV3”), or products (“Fluzone”) share the property list [target: influenza]. A subclass c_1 of the fundamental class c_2 is included in the property list of the vaccine class c_3 if and only if c_3 implies $\llbracket c_1 \rrbracket$.

3.9. VaccO analysis – an interactive illustration

We created an application as an interactive illustration of the VaccO analysis process, called “VaccO Analysis”.

VaccO Analysis takes the textual description of a vaccine or vaccine group and the

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	48/80

language of the description as input (see Figure 6). Upon clicking the button “Analyse”, VaccO analyses the vaccine description, and displays three internal representations of a vaccine description (see Figure 7):

1. The VaccO classes that were mentioned in the description under the heading “Tags”,
2. the compiled DL expression under the heading “Expression”, and
3. the representation of the vaccine description as properties lists under the heading “Properties”.

VaccO Analysis is available on <https://euadr.erasmusmc.nl/VaccO/#!/analysis>.

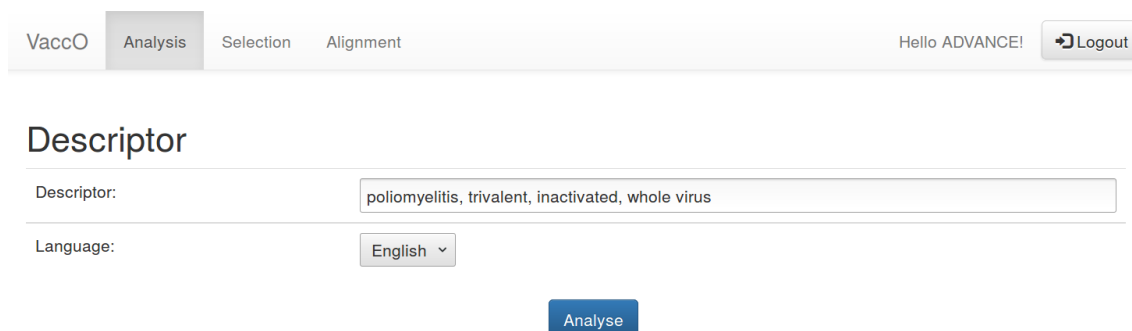



Figure 6: Input of the VaccO Analysis application: The descriptor of a vaccine or vaccine group, and its language.

 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	49/80

Analysis

Tags

String	Start	End	Classes
poliomyelitis	0	13	Poliovirus pathogen Poliomyelitis
trivalen	14	23	3-valent
, inactivat	24	35	inactivated
d, wh	36	41	inactivated

Expression

- immunizes against Poliovirus pathogen
- has valence 3-valent
- immunizes against Poliomyelitis
- has type inactivated
- Vaccine


Properties

Property	Values
Strain	
Target	poliomyelitis viral
AdministrationRoute	
Valence	valence_03 polyvalent
Manufacturer	
VaccineType	inactivated
Population	


Figure 7: The VaccO Analysis application displays the internal representations of a vaccine description: The VaccO classes that were mentioned in the description (Tags), the compiled DL expression (Expression), and the properties list (Properties).

3.10. Data distribution

VaccO is distributed in the format of the web ontology language [17] (OWL). The format was chosen to facilitate interoperability with other knowledge sources and to allow domain-specific inference about vaccines. VaccO was designed using Protégé (<http://protege.stanford.edu/>), a leading ontology-engineering application developed at Stanford University School of Medicine [33].

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	50/80

The VaccO ontology is available on the ADVANCE SharePoint under [WP5 Documents – WP5.5 Vaccine ontology – Vaccine ontology – OWL ontologies](#).

 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	51/80


4. VACCO APPLICATION I: VACCINE CODE SELECTION

We created an application for the analysis and selection of vaccine codes in a given vaccine coding system, called “VaccO Selection”. The application creates the property lists of the codes in a vaccine coding system, and allows the identification of vaccine codes based on their property values.

VaccO Selection takes a vaccine coding system and its language as input. The vaccine coding system is supplied as a text where each line is composed of 1) the code, 2) the pipe symbol “|”, and 3) the descriptor of the vaccine or vaccine group, for example “J07BJ01 | rubella, live attenuated” for code “J07BJ01” with English descriptor “rubella, live attenuated”. Several predefined vaccine coding systems can be loaded using the button “Use existing”. Each code must describe a vaccine or vaccine group.

After clicking the button “Select”, all codes in the vaccine coding system are analysed by VaccO, property lists are created for each code, and the output is shown. The output (see Figure 9) is organized in columns. One column is shown for every property category (e.g., “Vaccination target”, “Vaccine type”, or “Administration route”) that has been identified in the descriptors of any supplied vaccine code. Each column contains the list of all values that were identified for the property category in any descriptor of the supplied vaccine coding system (for example “Rubella”, “Poliomyelitis”, and “Influenza” the column for property category “Vaccination target”). Each property value can be selected using the checkbox next to it.

The final column contains a list of codes from the vaccine coding system. On selecting one or more property values in the columns to the left, the list is narrowed to codes, where the associated property lists contain the selected property values. For example, when “Pertussis” is selected in the column for property category “Target”, only codes of Pertussis vaccines are shown in the final column. When the checkbox “Include combinations” is deselected, the list of vaccine codes is filtered to the codes, where the

 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	52/80

associated property lists that contain *only* the selected property values, but no other property values. For example, when “Pertussis” is selected in the column “Target”, only codes are shown in the final column that represent vaccine that target only Pertussis but no combination vaccines. The button “Clear” resets the selection of property values.

VaccO Selection is available at <https://euadr.erasmusmc.nl/VaccO/#!/selection>.

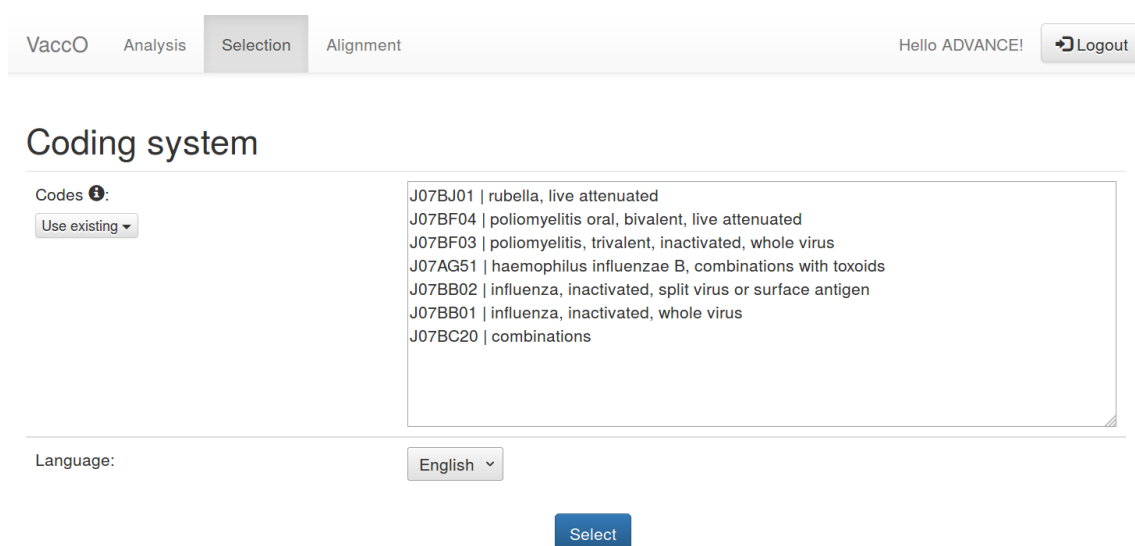



Figure 8: The input of the VaccO Selection application: A vaccine coding system and the language that is used in the descriptors.

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	53/80


Selection

Clear

☒ Include combinations

Target	Type	Route	Valence	Codes												
<input type="checkbox"/> Rubella <input type="checkbox"/> Viral <input type="checkbox"/> Poliomylitis <input type="checkbox"/> Influenza <input type="checkbox"/> Tetanus <input type="checkbox"/> Bacterial <input type="checkbox"/> Haemophilus Influenzae Type B <input type="checkbox"/> Diphtheria	<input type="checkbox"/> Unattenuated <input type="checkbox"/> Attenuated <input type="checkbox"/> Live <input type="checkbox"/> Inactivated <input type="checkbox"/> Recombinant	<input type="checkbox"/> Oral	<input type="checkbox"/> Valence 02 <input type="checkbox"/> Polyvalent <input type="checkbox"/> Valence 03	<table> <thead> <tr> <th>Code</th> <th>Label</th> </tr> </thead> <tbody> <tr> <td>J07AG51</td> <td>haemophilus influenzae B, combinations with toxoids</td> </tr> <tr> <td>J07BB01</td> <td>influenza, inactivated, whole virus</td> </tr> <tr> <td>J07BB02</td> <td>influenza, inactivated, split virus or surface antigen</td> </tr> <tr> <td>J07BC20</td> <td>combinations</td> </tr> <tr> <td>J07BF03</td> <td>poliomylitis, trivalent, inactivated, whole virus</td> </tr> </tbody> </table>	Code	Label	J07AG51	haemophilus influenzae B, combinations with toxoids	J07BB01	influenza, inactivated, whole virus	J07BB02	influenza, inactivated, split virus or surface antigen	J07BC20	combinations	J07BF03	poliomylitis, trivalent, inactivated, whole virus
Code	Label															
J07AG51	haemophilus influenzae B, combinations with toxoids															
J07BB01	influenza, inactivated, whole virus															
J07BB02	influenza, inactivated, split virus or surface antigen															
J07BC20	combinations															
J07BF03	poliomylitis, trivalent, inactivated, whole virus															

Figure 9: VaccO Selection shows the property values in the supplied vaccine coding system, and the list of all codes. Selecting one or more values restricts the list to the codes where the associated property list contains the selected values.


 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	54/80

5. VACCO APPLICATION II: AUTOMATIC ALIGNMENT OF VACCINE CODES BASED ON MULTILINGUAL CODE LABELS

5.1. Introduction

The vaccine coding inventory of section 2. demonstrated that vaccine information in European electronic health record databases is represented in various terminologies, including evolving database-specific terminologies, and codes are described using various, non-English languages. The implementation of a multi-database epidemiological study requires the harmonization of vaccine information from different databases in a common data model. Mappings are often missing between database vocabularies and the vocabulary used in the vaccine common input files. The process of mapping database-specific vaccine codes to the common data model is time-consuming and required for each database that participates in the study.

Various techniques have been proposed for aligning general ontologies [34], medical coding systems [35–40], and drug coding systems [41,42]. The most common approaches for aligning drug coding systems use lexical, instance-based, or hierarchical information about classes. However, their application to vaccine coding systems used in EHR databases is generally precluded by the characteristics of the coding systems: Lexical techniques create alignments based on comparison between the code descriptors, which is not directly applicable between coding systems using different languages. For instance-based techniques, the similarity of two classes is asserted based on the individual vaccines to the classes, but such information is usually unavailable in vaccine coding systems. Hierarchical techniques employ the structure of the taxonomic hierarchy of the ontology, however, vaccine coding systems used in EHR records are often not hierarchically structured.

 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	55/80

Automatic alignment of vaccine coding systems would harmonize and accelerate the integration of EHR databases in pharmacoepidemiologic studies about vaccines. Here we describe an automatic approach for aligning vaccine coding systems that possibly use multiple languages in their descriptors, by representing vaccine codes as classes in VaccO. We evaluate the approach by comparing the automatically generated alignments with manually created alignments.

5.2. Methods


5.2.1. Automatic mapping of vaccine codes

The objective of the code alignment is to assign each code of a source coding system to the closest corresponding code in a target coding system. The correspondence is determined by a similarity measure between source and target codes, where 1 indicates maximal similarity and 0 indicates no similarity. A source code is assigned to the target code that has maximal similarity with the source code above a threshold. The threshold is varied between 0 and 1 in 10 steps. Multiple target codes may be assigned when more than one target code has maximal similarity, and only the most general target codes with maximal similarity are assigned when the target coding system has a taxonomic hierarchy.


Assignment of target codes

We evaluated the alignment algorithm using two baseline similarity measures, and three similarity measures using VaccO.

- Method *Words* implements a simple lexical technique. Each descriptor is split into individual words, and similarity between two codes is measured by the Jaccard coefficient of the sets of words. The Jaccard coefficient of two sets s and t is defined as $|s \cap t| / |s \cup t|$.

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	56/80

- Method *MetaMap* abstracts over word inflections and synonyms by using MetaMap [43] to identify UMLS CUIs in the descriptors of all source and target codes. Similarity is measured by the Jaccard coefficient of the sets of CUIs.
- Method *Classes* represents codes by VaccO classes that were identified in the code descriptors (see section 3.6.). Similarity is defined by the Jaccard coefficient of sets of VaccO classes.
- Method *Equivalence* compiles all codes into vaccine classes in VaccO. Similarity between two codes is 1 if their vaccine classes are equivalent and 0 otherwise. Equivalence between vaccine classes is tested using an ontology reasoner.
- Method *Properties* represents codes as property lists. The alignment is a two-stage process. First, the set of candidate target codes is identified based on matching immunization target properties. We then compute the overlap between the properties in the source code and the properties the candidate target code. The overlap is defined as the Jaccard coefficient between the values in the property lists. See Figure 10, bottom for an illustration of the process

 IM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	57/80

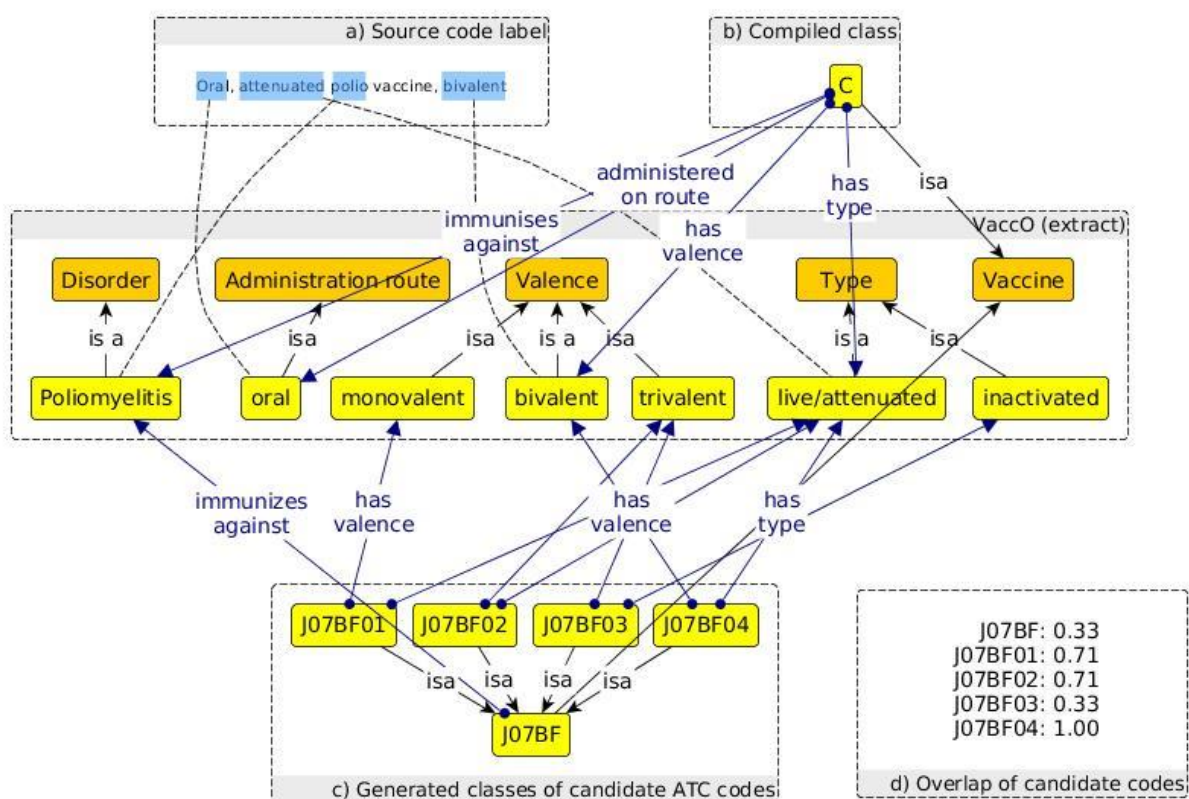



Figure 10: Example for the identification of target codes for a source code. a) Occurrences of VaccO classes are identified in the source code description. b) The identified VaccO classes are compiled into a VaccO sub class of Vaccines. c) Candidate target codes are identified by matching immunisation targets in their compiled class representation. d) The overlap between the compiled class and the candidate target classes is computed. The target codes that maximise the overlap are the result of the alignment (J07BF04).

5.2.2. Identification of VaccO classes in multilingual free text

We prepared the dictionary of VaccO for multilingual input by automatically translating it using GoogleTranslate [44] to Spanish, Italian, and Catalan. The multilingual dictionary is stored in the Apache Solr text search platform, and occurrences of the terms in free text are identified using the Solr TextTagger plugin. Tokenization, stemming algorithms, stop word lists, and elision filters have been configured for each language.

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	58/80

5.2.3. Evaluation

Reference mappings

We used two sets of reference mappings in our evaluation (see Table 10). The first set uses the ADVANCE Vactype coding system as target. Vactype uses English descriptors and comprises 28 codes for individual immunization targets that can be combined for combination vaccines (15 combinations are used in the reference set). The reference set has five custom vaccine coding systems of European primary-care EHR databases as sources: SIDIAP with Catalan descriptors, BIFAP with Spanish descriptors, the Italian paediatric database Pedianet with both English and Italian descriptors, and regional primary care database of Venetia with Italian descriptors. The mappings of the Vactype reference set were manually created and validated by the database custodians in the context of the ADVANCE project.

The second set comprises existing mappings in the UMLS with target coding system ATC. We included the coding systems in UMLS with the largest amount of vaccine codes mapped to vaccine ATC codes: Veterans Affairs National Drug File (VANDF), MeSH, Consumer Health Vocabulary (CHV), Vaccine Administered (CVX), National Drug File Reference Terminology (NDF-RT).

Reflexive alignments in which Vactype or ATC were both the source coding system and the target coding system were included in the evaluation to assess the completeness of the interim representation in the alignment process.

We corrected nine incorrectly assigned codes in the Vactype reference set (four in BIFAP, five in SIDIAP), and 17 codes in the ATC reference set that were not assigned to the most specific ATC code (three in CHV, two in CVX, two in MeSH, six in NDF-RT, and four in VANDF).



 IMi - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	59/80

Table 10: Target coding systems, languages and number of source codes in the reference alignments in the evaluation.

Target	Source	Source language	Source codes
Vactype	Vactype	English	43
	BIFAP	Spanish	761
	Venetia	English	21
	Pedianet ^{en}	English	9
	Pedianet ^{it}	Italian	9
	SIDIAP	Catalan	98
ATC	ATC	English	114
	BIFAP	Spanish	528
	VANDF	English	18
	MeSH	English	23
	CHV	English	26
	CVX	English	18
	NDFRT	English	40

Performance measure

The performance of the alignment methods was measured for each pair of source and target coding system by the precision ($TP/(TP + FP)$), recall ($TP/(TP + FN)$), and F-score ($((2 * precision * recall)/(precision + recall))$). TP stands for the number of true positive code assignments, FP for the number of false positive code assignments, and FN for the number of false negative code assignments. We also report for both reference sets the average performance measures over all source coding systems (excluding reflexive alignments).


 IMi - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	60/80

5.3. Results

A threshold of 0.1 maximized for all alignment methods the average F-score (see Table 11/Figure 11), and achieves a balance between precision (F-score = precision with 95%-CI [0.002; 0.02]) and recall (F-score = recall with 95%-CI [-0.06; 0.02]). The threshold is kept constant at this value in the following analyses.

The method *Tokens* was insufficient for the automatic alignment in both reference sets (see Table 11 with more details in Table 12) because it does not deal with multilingual input. Also the lower performance of both baseline methods *Tokens* and *Metamap* in the Vactype reference set than in the ATC reference was due to the use of descriptors with non-matching languages in the Vactype reference set. *Classes* was the only method that performed worse in the ATC reference set than in the Vactype reference set, because code descriptors in the ATC reference set used broader property categories, which are not identified as references to the same vaccine group. The *Properties* method performed best in both reference sets, dealing well with multilingual input and differing descriptions. Method *Equivalence* performed lower than method *Properties*: Testing equivalence between compiled codes was too strong as criterion for identifying corresponding target codes, especially when the source coding system is more granular than the target coding systems (BIFAP and MeSH).

All methods performed perfectly on Pedianet with English descriptors, which contain only codes for influenza vaccines. Only the baseline methods, which are not adapted for multilingual descriptors, performed lower using the Italian descriptors than the English descriptors of Pedianet. The performance on the Venetia source coding system was worse for all methods but *Tokens* because the reference mappings were created using contextual knowledge (all Pertussis codes refer to acellular Pertussis vaccines because all authorised Pertussis vaccines are acellular).

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	61/80


Both methods *Metamap* and *Classes* use concept sets as intermediate representation (UMLS CUIs and VaccO classes, respectively). The performance differences between the two methods are largest in the Vactype reference set with non-English source coding systems (Pedianet with Italian descriptors, SIDIAP, and BIFAP).

The F-scores of all reflexive alignments were higher than 0.95, which indicates that the intermediate representation of the method is capable of representing the information in the target coding systems. Some ATC codes could not be distinguished using sets of CUIs, VaccO classes or property lists (e.g., J07B and J07BX with descriptors “VIRAL VACCINES” and “Other viral vaccines”), which lowered the precision for the reflexive mapping in the ATC reference set.

By ignoring the threshold, the average recall improved to 0.93 in the Vactype reference set and to 0.97 in the ATC reference set. The average precision decreased to 0.65 in the Vactype reference set and to 0.97 in the ATC reference set.

Table 11: Average F-score using different thresholds for the selection of target codes over both reference sets. Method Equivalence does not use a threshold and is not shown.

Threshold	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Tokens	0.161	0.230	0.214	0.177	0.078	0.060	0.044	0.031	0.031	0.031	0.031
Stems	0.615	0.743	0.723	0.693	0.549	0.507	0.438	0.369	0.365	0.365	0.365
Metamap	0.504	0.638	0.492	0.453	0.441	0.432	0.377	0.367	0.370	0.370	0.370
Classes	0.643	0.756	0.755	0.744	0.719	0.715	0.734	0.722	0.728	0.717	0.713
Properties	0.843	0.929	0.929	0.928	0.916	0.916	0.874	0.870	0.870	0.857	0.857

 IMM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	62/80

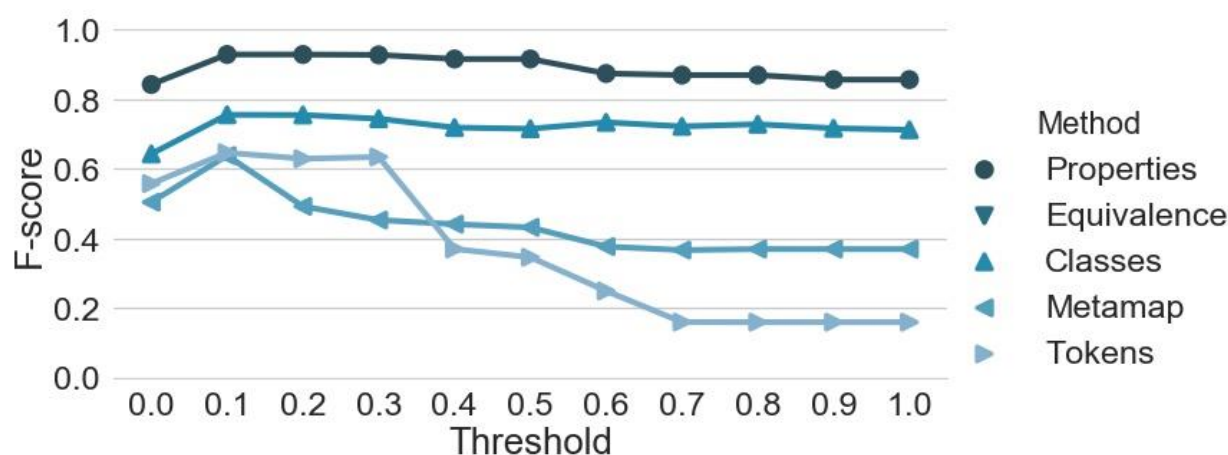


Figure 11: Overall average F-score using different thresholds for the selection of target codes.

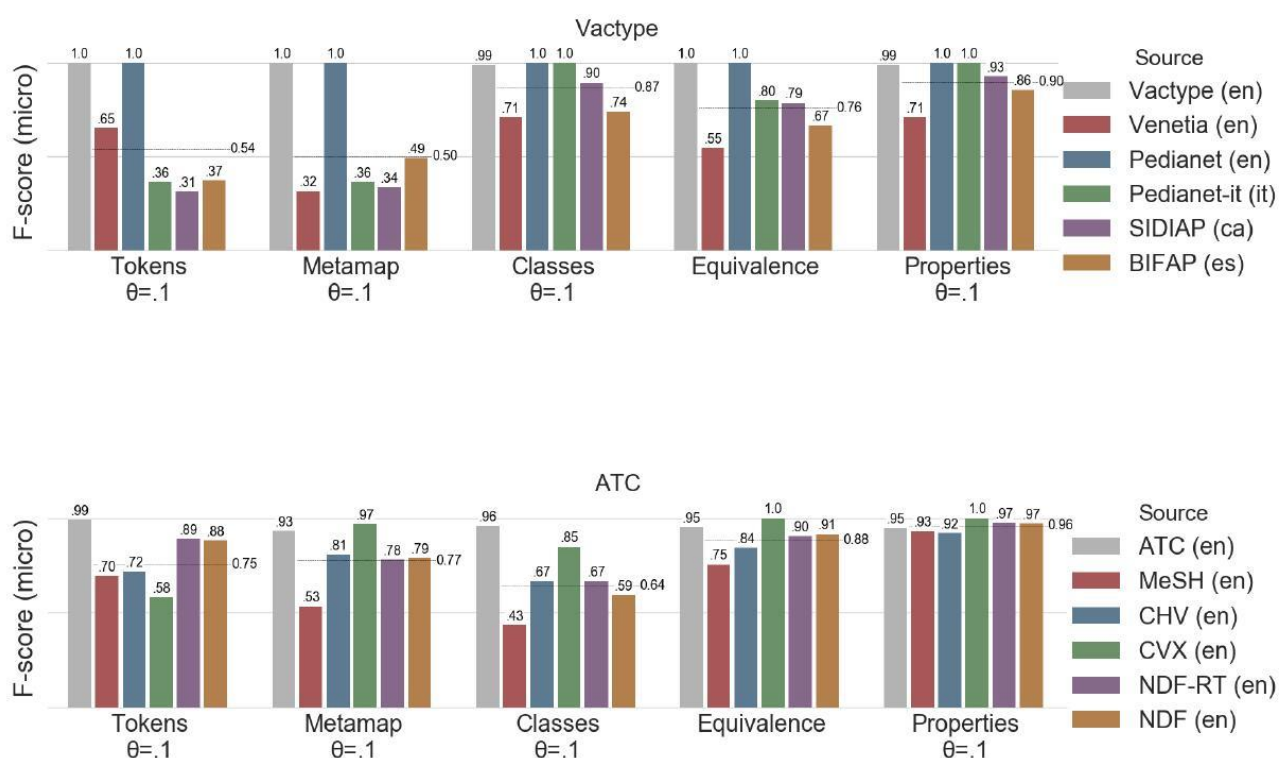


Figure 12: F-score for automatic vaccine code alignments. Performance measures of reflexive alignments (same source and target vocabulary) are shown in grey. The dashed line indicates the mean F-score of the method excluding reflexive alignments.


 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe				
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring			Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto			Security: PU	63/80


Table 12: Detailed performance measures of F-score (F), precision (P), and recall (R) for the vaccine code alignment using a threshold of 0.1 in a) the ATC reference set, and b) the Vactype reference set. Averages do not take the measures of reflexive alignment into account.

a)

Source	Vactype			Venetia			Pedianet			Pedianet-it			SIDIAP			BIFAP			Average		
Measure	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Tokens	1.000	1.000	1.000	0.600	0.714	0.652	1.000	1.000	1.000	1.000	0.222	0.364	0.760	0.198	0.314	0.735	0.248	0.371	0.819	0.477	0.540
Metamap	1.000	1.000	1.000	0.353	0.286	0.316	1.000	1.000	1.000	1.000	0.222	0.364	0.870	0.208	0.336	0.787	0.359	0.493	0.802	0.415	0.502
Classes	1.000	0.977	0.988	0.667	0.762	0.711	1.000	1.000	1.000	1.000	1.000	1.000	0.953	0.844	0.895	0.790	0.695	0.739	0.882	0.860	0.869
Equivalence	1.000	1.000	1.000	0.750	0.429	0.545	1.000	1.000	1.000	1.000	0.667	0.800	0.917	0.688	0.786	0.885	0.534	0.666	0.910	0.663	0.759
Properties	1.000	0.977	0.988	0.667	0.762	0.711	1.000	1.000	1.000	1.000	1.000	1.000	0.918	0.938	0.928	0.861	0.852	0.856	0.889	0.910	0.899

b)

Source	ATC			MeSH			CHV			CVX			NDFRT			VANDF			Average		
Measure	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Tokens	0.991	0.991	0.991	0.696	0.696	0.696	0.704	0.731	0.717	0.563	0.600	0.581	0.878	0.900	0.889	0.882	0.882	0.882	0.744	0.762	0.753
Metamap	0.908	0.956	0.932	0.545	0.522	0.533	0.808	0.808	0.808	1.000	0.933	0.966	0.762	0.800	0.780	0.813	0.765	0.788	0.786	0.765	0.775
Classes	0.973	0.947	0.960	0.435	0.435	0.435	0.643	0.692	0.667	0.778	0.933	0.848	0.636	0.700	0.667	0.550	0.647	0.595	0.608	0.681	0.642
Equivalence	0.905	1.000	0.950	0.667	0.870	0.755	0.774	0.923	0.842	1.000	1.000	1.000	0.864	0.950	0.905	0.889	0.941	0.914	0.839	0.937	0.883
Properties	0.955	0.939	0.947	1.000	0.870	0.930	0.958	0.885	0.920	1.000	1.000	1.000	1.000	0.950	0.974	1.000	0.941	0.970	0.992	0.929	0.959


 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	64/80

5.4. Error analysis

We characterize the remaining problems of the *Properties* method in the alignment of vaccine coding systems by an error analysis. We randomly selected from each source coding system at most 20 source codes that were assigned to false positive codes, false negative codes, or both, and assigned error categories to them.

Most errors in the sample were due to errors in the identification of VaccO classes in the descriptors (see Table 13). False positives were caused by ambiguous terms (e.g., “pentavalent” can refer to DTP-HIB-HepB and DTP-HIB-Polio), and false negatives were caused by spelling variations, typographic errors, and unknown abbreviations in the code descriptors. Several codes in the ATC reference set could not be represented (e.g. the ATC code J07 for “Vaccines” has an empty property list), or could not be distinguished (e.g., both ATC codes J07A for “Bacterial vaccines” and J07AX for “Other bacterial vaccines” have the same properties list [target: bacteria]). Other alignment errors were due to the use of contextual knowledge while creating the references in the Vactype reference set, which is not used or available in the automatic alignment, and by using products in the descriptors that are unavailable in VaccO.

Table 13: Error categories in the automatic alignment using the Properties method for the Vactype and ATC reference sets. Each error can result in a false negative (FN) alignment, false positive (FP) alignments, or both.

 IMM - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	65/80


Error category	Resulting error type in the alignment	Vactype		ATC	
		Count	Percentage	Count	Percentage
FN in identification of VaccO classes	FN	16	43.2%	2	22.2%
FP in identification of VaccO classes	FP	6	16.2%	1	11.1%
Not represented in properties	FN	-	-	5	55.6%
Contextual knowledge required	FP/FN	13	35.1%	-	-
Unknown product	FN	2	5.4%	1	11.1%

5.5. User application for vaccine code alignment

We have created a user application for the alignment of vaccine coding systems, called “VaccO Alignment”.

VaccO Alignment takes two vaccine coding system as input, one as source and one as target (see Figure 13). The vaccine coding system are supplied as a text where each line is composed of 1) the code, 2) the pipe symbol “|”, and 3) the descriptor of the vaccine or vaccine group, for example “J07BJ01 | rubella, live attenuated” for code “J07BJ01” with descriptor “rubella, live attenuated”. Each code must describe a vaccine or vaccine group. Several predefined vaccine coding systems can be loaded using the button “Use existing”.

On clicking the button “Align”, VaccO analyses the codes in the source and target vaccine coding systems (as described in section 3.) and creates an alignment between the source and target coding system is generated using the algorithm described above. The resulting alignment is displayed as a table (see Figure 14), where each row contains a source code and its descriptor together with the target code and its descriptor that

 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	66/80

was assigned. The table contains multiple rows for one source code if the alignment is not unique for the source code. The table that displays the alignment can be selected, copied and pasted into a Spreadsheet document such as Excel for further processing.

VaccO Alignment is available at <https://euadr.erasmusmc.nl/VaccO/#!/alignment>.


VaccO
Analysis
Selection
Alignment
Hello ADVANCE!
Logout

Source
Codes ⓘ:
Use existing ▼
J07BJ01 | rubella, live attenuated
J07BF04 | poliomyelitis oral, bivalent, live attenuated
J07BF03 | poliomyelitis, trivalent, inactivated, whole virus
J07AG51 | haemophilus influenzae B, combinations with toxoids
J07BB02 | influenza, inactivated, split virus or surface antigen
J07BB01 | influenza, inactivated, whole
Language: English ▼

Target
Codes ⓘ:
Use existing ▼
INF | Influenza
TET | Tetanus
PNE | Pneumococcal disease
VAR | Varicella
CHO | Cholera
DIP-HEB-TET-aPE | Diphtheria, Hepatitis B, Tetanus, acellular Pertussis
DIP-HIB-POL-TET-aPE | Diphtheria, Haemophilus influenzae type b, Poliomyelitis, Tetanus, acellular Pertussis
Language: English ▼

Align

Figure 13: The VaccO Alignment application takes two vaccine coding systems as input.

 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	67/80

Result


Source	Source label	Target	Target label
J07BJ01	rubella, live attenuated	RUB	Rubella
J07BF04	poliomyelitis oral, bivalent, live attenuated	POL	Poliomyelitis
J07BF03	poliomyelitis, trivalent, inactivated, whole virus	POL	Poliomyelitis
J07AG51	haemophilus influenzae B, combinations with toxoids	DIP-HIB-PER-TET	Diphtheria, Haemophilus influenzae type b, Pertussis, Tetanus
J07BB02	influenza, inactivated, split virus or surface antigen	INF	Influenza

Figure 14: Example of the output of the VaccO Alignment application for aligning ATC vaccine codes with ADVANCE Vactype.

5.6. Application in the ADVANCE framework

The VaccO Alignment application can be used to accelerate the conduction of multi-database vaccine studies. For the pooled analysis of vaccinations from multiple EHR databases, each database has to map its vaccine codes to a common coding system that was defined in the study protocol. Mappings from the vaccine coding systems of the EHR databases to the common coding system do usually not exist and have to be created manually, which is a time-consuming process.

A mapping from the vaccine coding system used in an EHR database to the common coding system can be automatically generated using the VaccO Alignment application. The vaccine coding system used by the EHR database is supplied as source coding system, and the common coding system is supplied as the target coding system. The automatically generated mapping should be regarded as a draft and validated manually. But the remaining manual work is reduced the removal of false positive target codes (11% of the generated codes according to an average precision of 0.89 in the Vactype reference set), and to the identification source codes where the correct target code has


 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	68/80

not been generated (in 9% of source codes according to an average recall of 0.91 in the Vactype reference set).


The balance between precision and recall of the Properties method can be shifted by changing its threshold. Using a lower threshold for maximizing the recall can be advantageous in the use case where the automatically generated alignment is used as an input for manual validation and correction, and the identification and correction of a false negative assignments requires more effort (identifying the error and selecting the correct code from the complete target coding system) than the identification of false positives (removing the codes).

5.7. Discussion

Alignment of vaccine coding systems has three main difficulties: Differing languages in descriptors, differing properties used for describing the same vaccine classes (e.g., by an active ingredient as “BCG” or by target disorder as “Tuberculosis vaccine”), and the identification of most relevant codes across different levels of granularity. The Properties method deals with all three difficulties as seen in its high performance in reference set using multiple languages (Vactype), and the reference set with larger variation of the use of property categories (ATC). The ability of the Properties method lies in the representation of the information from the code descriptor as VaccO vaccine classes, and its normalized representation of vaccine classes as property lists that allows their inexact comparison. Our reference sets used code descriptors in English, Spanish, Italian, and Catalan, and contained general medical coding systems, drug coding systems, and custom database coding systems. This wide range of vaccine coding systems indicates the generalizability of our evaluation.

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	69/80

Using VaccO for the alignment of vaccine coding systems demonstrates the power of a domain-specific ontology for extracting and representing information from free text. The approach could likewise be applied in other free-text resources, e.g. scientific literature, spontaneous reports, and public news.

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	70/80


6. VACCO APPLICATION III: VACCINE PRODUCT CLASSIFICATION IN A SPONTANEOUS REPORTING SYSTEM

6.1. Introduction

Spontaneous reporting systems like Eudravigilance are important information sources about the safety of vaccines after their market approval. Eudravigilance is maintained by the European Medicines Agency (EMA) and contains 6.2 million spontaneous reports of possible adverse reactions [45]. Eudravigilance reports issued to research organisations [46] contain only a limited set of data elements and exclude any data elements that allow the identification of personal data. Particularly, most drugs are only characterised by their active substances. We assess which properties of vaccines can be identified by the active substances using VaccO and how vaccine properties can be induced from knowledge about existing vaccine products.

Available data elements in Eudravigilance

For research organisations, access to Eudravigilance is restricted to data elements that comply with the European personal data protection legislation to research organisations. This includes information about the active ingredients of the drug, the adverse reaction coded in MedDRA, and the age of the reporter if available. The country or date of the report or any other identifier that provide hints about the country or any personal data of the reporter are restrained. Notably, product names are only available for centrally authorised products. The Eudravigilance access policy [46] describes the availability of

 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	71/80

data elements in detail.

6.2. Methods


Processing of Eudravigilance data

The VaccO ontology is used to identify properties of vaccines in the labels of the active substances (see Figure 15). Additional properties are then extrapolated from matching vaccine products available in VaccO. For example, if all vaccines that target the microbes and disorders identified in the active substances are from the same manufacturer, this manufacturer is included as extrapolated property of the vaccine.

We used case reports about paediatric vaccinations that were extracted from Eudravigilance for the GRiP [47] project in our evaluation. The GRiP dataset contained adverse drug reactions reported in the paediatric population (younger than 18 years) between 2002 and 2016. We used in our evaluation reports from the GRiP dataset that reported at least one vaccine which had been identified by a query on vaccine related keywords in the reported active substances.

Each report in Eudravigilance contains a number of drugs, and each drug refers to one or more active substances, and possibly to a product name and ATC code. We evaluated our approach for enriching Eudravigilance information with the VaccO ontology in a random sample of 500 uniquely reported vaccines (two vaccines were considered different if product name, ATC code, administration route or the descriptions of active substances differed). Vaccines were discarded from further analysis if they included an empty description for one of its active substances.

We identified for all remaining vaccines the products in VaccO that contain the given

 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	72/80

active substances identified by their descriptions (“Products from active substances” in Figure 15). We then identified the VaccO classes in the description of each active ingredient, and compiled the classes into a DL expression (as described in section 3.6. and 3.7.). The intersection of the DL expressions was used to represent the vaccine reported to Eudravigilance in VaccO (see section 3.7.). We computed the properties of the intersection (*Direct properties* in Figure 15), and the set of possible products that conform to the intersection (*Possible products*). The *extrapolated properties* were computed from the intersection of the DL expression representing the possible products. For all drugs that were reported with a product name we computed the vaccine properties from the class representing the product in VaccO (*Product properties*).

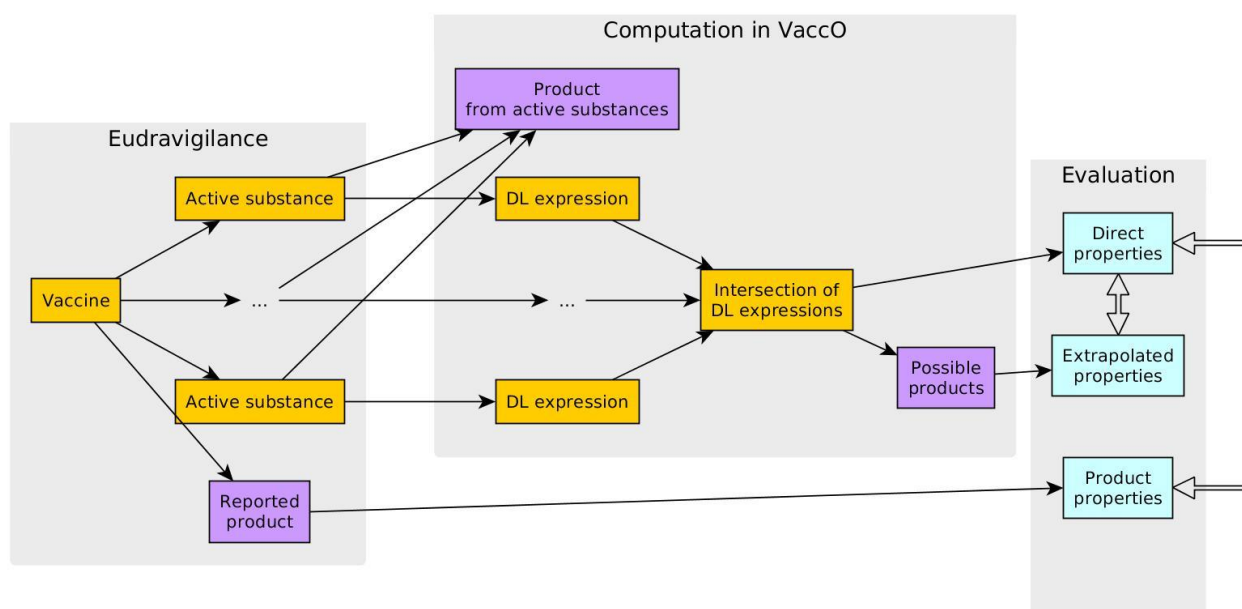



Figure 15: Processing of Eudravigilance data and evaluation. The comparison between direct properties and product properties assesses the extraction of information from the active substances. The comparison between extrapolated properties and product properties assesses

 IM I - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	73/80

the value of using VaccO for extrapolating information in Eudravigilance.


Evaluation

To evaluate the derivation of direct properties from the reported active substances, we measured the performance to match the product properties with the direct properties among all drugs that are reported with a product name. The performance was computed for each drug and property as the recall (i.e., sensitivity) and precision (i.e., positive-predictive value) between the values in the property lists. For example, the recall of the direct properties [target=Pertussis, Hepatitis A] with respect to the product properties [target=Pertussis, Diphtheria, Tetanus] is 0.33 and its precision is 0.5. We report the performance measures separately for each property averaged over all vaccine.

To evaluate the value of VaccO for extrapolating information in Eudravigilance, we measured the performance to match the product properties with the extrapolated property values.

6.3. Results

ATC classification and exact product information was only available in 18% (N=90) of the 500 vaccines in our evaluation sample. All reported products were represented in VaccO/Art57 DB. Most vaccines (93.6%; N=468) contained descriptions of all active substances, and were retained for further analysis. In the resulting set, it was only possible for 10.0% of the vaccines to identify products in Art57 DB using the verbatim descriptions of active substances (*Products from active substances* in Figure 15). In 38.0% of the vaccines, a product could be identified based on the DL-expression representing the active substances (*Possible products* in Figure 15).

 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	74/80

The performance to identify vaccine properties in the descriptions of active substances was evaluated in the 328 vaccines from the evaluation set where the product name was available under academic license and represented in VaccO, and where the descriptions of all active ingredients were available. The results are shown in Table 14 for the identification of properties based on reported active ingredients only, and for identification of extrapolated properties based on known matching vaccine products.


Table 14: Recall (sensitivity) and precision (positive-predictive value) for regenerating vaccine properties from active substances, and their extrapolation using known products that match the DL-expression of the vaccine.

Property	Method	Recall	Precision
Vaccine type	Direct	0.53	0.79
	Extrapolated	0.61	0.79
Vaccine target	Direct	0.96	0.99
	Extrapolated	0.96	0.99
Administration route	Direct	0.35	0.90
	Extrapolated	0.64	0.91

6.4. Discussion

Exact product information and ATC code were unavailable under the academic licence for most vaccines in Eudravigilance. Only vaccination reports concerning the 18% of vaccines that contained product information and ATC code could be used directly in academic, pharmacoepidemiologic studies that analyse reports about specific vaccines or that stratify by vaccine products or properties.

The literal descriptions of active substances did not constitute a reliable base for


 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	75/80

identifying vaccine products. Vaccine products could only be identified for 10% of the vaccines based on the verbatim description of active substances. This may be due either to variations in the descriptions between Art57 DB, which is the information source about vaccine ingredients in VaccO, and descriptions of active substances in Eudragilance, or due to vaccines reported in Eudragilance that are not included in Art57 DB (because they lost marked authorisation before creation of Art57 DB). Corresponding vaccine products could be identified for 38% of the vaccines the evaluation set based on the DL-expressions representing the active substances in VaccO.


Immunization targets are the fundamental property of vaccines and could be reliably identified from the active substances (recall 0.96, precision 0.99). The identification of vaccine types in the description of active substances performed lower (recall 0.53, precision 0.61). Possible reasons for this lower performance are false positives or false negatives in the identification of VaccO classes in the active substances, or missing information in the descriptions of the active substances. Additional error analysis is required to identify exact reasons. Administration routes could be identified from the reported information with a recall of 0.35 and a precision of 0.90.

Extrapolation of vaccine properties based on products in VaccO/Art57 DB that match the DL-expression compiled from the active substances improved the recall for identifying the vaccine types to 0.61. Extrapolation did change neither the precision for identifying vaccine types nor the performance for identifying immunization targets. The performance of identifying administration routes improved to a recall 0.61 and a precision 0.91 using extrapolation.

Extrapolation of vaccine properties using VaccO improves the identification of vaccine


	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	76/80

properties but its utility follows the predictivity of the set of possible products. This is a trade-off: When many vaccine properties are already determined by the active substances, the set of possible products is also homogeneous and is more likely to contain common properties. But when information about the vaccine properties is not derivable from the active substances, the set of possible products is large, and lacks common properties. Further experiments are required to assess if the utility of the extrapolation approach can be improved by including extrapolated properties when they are shared by a given percentage of possible products.


 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	77/80

7. REFERENCES


- [1] World Health Organization. WHOCC - ATC/DDD Index n.d. https://www.whooc.no/atc_ddd_index/ (accessed July 4, 2017).
- [2] Read JD, Sanderson HF, Sutton YM. Terming, Encoding, Grouping, The Language of Health. Proc. Int. Med. Inf. Assoc. 8th World Congr. Med. Inform. Vanc., 1995.
- [3] Schulz EB, Barrett JW, Brown PJB, Price C. The Read Codes: evolving a clinical vocabulary to support the electronic patient record. Conf. Proc. Electron. Health Rec. Eur. Newton CAEHR, 1996, p. 131–40.
- [4] He Y, Cowell L, Diehl AD, Mobley HL, Peters B, Ruttenberg A, et al. VO: vaccine ontology. Proc. 1st Int. Conf. Biomed. Ontol., vol. 2009, Buffalo, NY, USA; 2009.
- [5] Marcos E, Zhao B, He Y. The Ontology of Vaccine Adverse Events (OVAE) and its usage in representing and analyzing adverse events associated with US-licensed human vaccines. J Biomed Semant 2013;4:40.
- [6] Xiang Z, Zheng W, He Y. BBP: Brucella genome annotation with literature mining and curation. BMC Bioinformatics 2006;7:347.
- [7] Özgür A, Xiang Z, Radev DR, He Y. Mining of vaccine-associated IFN- γ gene interaction networks using the Vaccine Ontology. J Biomed Semant 2011;2:S8.
- [8] European Medicines Agency. Reporting requirements for authorised medicines - Guidance documents n.d. http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000336.jsp (accessed July 4, 2017).
- [9] Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Prof 2005;7:17–23.
- [10] Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc 2011;18:441–448.
- [11] International Organization for Standardization. ISO 11616:2012 - Health informatics -- Identification of medicinal products -- Data elements and structures for the unique identification and exchange of regulated pharmaceutical product information n.d. <https://www.iso.org/standard/55035.html> (accessed July 4, 2017).
- [12] IDMP Standards - Identification of Medicinal Products n.d. <https://www.idmp1.com/> (accessed July 4, 2017).
- [13] Gruber TR. Toward principles for the design of ontologies used for knowledge sharing? Int J Hum-Comput Stud 1995;43:907–928.
- [14] Gruber T. What is an Ontology. WWW Site [Httpwww-ksl Stanf Edukstwhatis--Ontol Html](http://www-ksl.stanford.edu/kstwhatis--Ontol.html) Accessed 07-09-2004 1993.

 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	78/80

- [15] Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story M-A, et al. The national center for biomedical ontology. J Am Med Inform Assoc 2011;19:190–195.
- [16] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25:1251–1255.
- [17] Word Wide Web Consortium. OWL 2 Web Ontology Language Document Overview (Second Edition) n.d. <https://www.w3.org/TR/owl2-overview/> (accessed July 4, 2017).
- [18] Baader F. The description logic handbook: Theory, implementation and applications. Cambridge university press; 2003.
- [19] Baader F, Horrocks I, Sattler U. Description logics. Found Artif Intell 2008;3:135–179.
- [20] Peroni S. Graffoo Specification. Graffoo Specif 2013. <http://www.essepuntato.it/graffoo/specification/current.html> (accessed June 26, 2017).
- [21] He Y, Sarntivijai S, Lin Y, Xiang Z, Guo A, Zhang S, et al. OAE: The Ontology of Adverse Events. J Biomed Semant 2014;5:29. doi:10.1186/2041-1480-5-29.
- [22] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32:281–291.
- [23] Center for Drug Evaluation and Research. Data Standards Manual (monographs) - Route of Administration n.d. <https://www.fda.gov/drugs/developmentapprovalprocess/formssubmissionrequirements/electronic submissions/datastandardsmanualmonographs/ucm071667.htm> (accessed May 23, 2017).
- [24] Center for Disease Control and Prevention. U.S. Vaccine Names n.d. <https://www.cdc.gov/vaccines/terms/usvaccines.html> (accessed July 2, 2017).
- [25] Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA, others. Characteristics of consumer terminology for health information retrieval. Methods Inf Med-Method Inf Med 2002;41:289–298.
- [26] Pavillon G, Maguin M. The 10th revision of the International Classification of Diseases. Rev Epidemiol Sante Publique 1992;41:253–5.
- [27] Plotkin S, Orenstein O, Offit P. Vaccines. Expert Consult Basic Books; 2013.
- [28] Hamborsky J, Kroger A, Wolfe C. Epidemiology and Prevention of Vaccine-preventable Diseases: The Pink Book: Course Textbook. Public Health Foundation; 2015.
- [29] US Department of Health Services. Understanding vaccines; what they are; how they work. Dostupno Na Httpwww Niaid Nih Govtopicsvaccinesdocumentsundvacc Pdf Datum Posled Pristupa 2015;31.
- [30] Lin Y, He Y. Ontology representation and analysis of vaccine formulation and administration and their effects on vaccine immune responses. J Biomed Semant 2012;3:17.
- [31] The Apache Software Foundation. Apache Solr - n.d. <http://lucene.apache.org/solr/> (accessed July 25, 2017).

 ADVANCE IMI - 115557	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1 – Draft
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	79/80

- [32] SolrTextTagger: A text tagger based on Lucene / Solr, using FST technology. OpenSextant; 2017.
- [33] Musen MA. The Protégé project: A look back and a look forward. *AI Matters* 2015;1:4–12.
- [34] Euzenat J, Shvaiko P. *Ontology matching*. Springer Science & Business Media; 2013.
- [35] Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proc. AMIA Symp.*, American Medical Informatics Association; 1998, p. 815.
- [36] Soualmia LF, Golbreich C, Darmoni SJ. Representing the MeSH in OWL: Towards a semi-automatic migration. *KR-MED*, vol. 102, 2004, p. 81–87.
- [37] Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. *AMIA. Annu. Symp. Proc.*, vol. 2005, American Medical Informatics Association; 2005, p. 266.
- [38] Assem M van, Malaisé V, Miles A, Schreiber G. A Method to Convert Thesauri to SKOS. *Semantic Web Res. Appl.*, Springer, Berlin, Heidelberg; 2006, p. 95–109. doi:10.1007/11762256_10.
- [39] 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 Inf* 2007;76 Suppl 3:S353-361. doi:10.1016/j.ijmedinf.2007.03.004.
- [40] Merabti T, Grosjean J, Soualmia LF, Joubert M, Darmoni SJ. Aligning biomedical terminologies in French: towards semantic interoperability in medical applications. INTECH Open Access Publisher; 2012.
- [41] Winnenburg R, Rodriguez L, Callaghan FM, Sorbello A, Szarfman A, Bodenreider O. Aligning Pharmacologic Classes Between MeSH and ATC. *VDOS ICBO*, 2013.
- [42] Winnenburg R, Bodenreider O. A framework for assessing the consistency of drug classes across sources. *J Biomed Semant* 2014;5:30. doi:10.1186/2041-1480-5-30.
- [43] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, American Medical Informatics Association; 2001, p. 17.
- [44] Google Inc. Google Translate n.d. <https://translate.google.com/> (accessed July 25, 2017).
- [45] EMA. 2015 Annual Report on EudraVigilance for the European Parliament, the Council and the Commission. 2016.
- [46] European Medicines Agency. EudraVigilance access policy for medicines for human use 2011.
- [47] GRiP. Global Research in Paediatrics (GRiP) - paediatric clinical pharmacology studies n.d. <http://www.grip-network.org/index.php/cms/en/home> (accessed July 2, 2017).

	D5.5 An Ontology for the Integration and Extraction of Vaccine-related Information in Europe		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1 – Draft	
	Author(s): Benedikt Becker (EMC), Miriam Sturkenboom, Jan Kors, Elisa Martin Merino, Giuseppe Roberto	Security: PU	80/80

8. NOTES

The files on the ADVANCE SharePoint that are referred to in this deliverable are also available on the Synapse SharePoint at the following address:

https://synapsemanagers-my.sharepoint.com/personal/nyefimenko_synapse-managers_com/Documents/ADVANCE%20D5.5%20Vaccine%20Ontology