



www.advance-vaccines.eu

**Accelerated Development of VAccine beNefit-risk
Collaboration in Europe**

Grant Agreement n°115557

**D5.2 Initial fingerprinting of the
participating health care databases**

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

**V1.4
Final**

Lead beneficiary: EMC
Date: 28-11-2014
Nature: Report
Dissemination level: CO



 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 2/76


TABLE OF CONTENTS

Contents

TABLE OF CONTENTS	2
DOCUMENT INFORMATION	4
DOCUMENT HISTORY	6
DEFINITIONS	7
PARTICIPATING INSTITUTIONS IN WP 5	9
EXECUTIVE SUMMARY	10
CHAPTER 1. BACKGROUND	11
CHAPTER 2. OVERVIEW OF FINGERPRINTING PROGRAM	13
2.1 DESCRIPTIVE DATA	14
2.1.1 <i>Population fingerprinting</i>	14
2.1.1.1 Fingerprinting common input file - Population	14
2.1.1.2 Running population fingerprint	15
2.1.2 <i>Vaccine fingerprint</i>	17
2.1.2.1 Common input file	17
2.1.2.2 Vaccine ontologies	19
2.1.2.3 Combination of vaccine data in databases and ontologies	22
2.1.2.4 Analysis of vaccination fingerprint	22
2.1.3 <i>Event fingerprinting</i>	24
2.1.3.1 Common input file	24
2.1.3.2 Selection of events and mapping	24
2.1.3.3 Analysis	26
2.2 SUITABILITY FINGERPRINTING	26
CHAPTER 3. DATABASES IN THE CONSORTIUM	27
CHAPTER 4. RESULTS OF POPULATION FINGERPRINTING	31
4.1 PARTICIPATING DATABASES	31
4.2 DISTRIBUTION OF PERSON-TIME OVER CALENDAR-YEAR	32
4.3 POPULATION PYRAMIDS AND REPRESENTATIVENESS	34
4.3.1 <i>Denmark: Aarhus database (DK_AUH)</i>	34
4.3.2 <i>Denmark: The Danish Civil and Health Registration System (DK_SSI)</i>	37
4.3.3 <i>Sweden: population based-registers (SE-KI)</i>	40
4.3.4 <i>UK: THIN database (UK_THIN)</i>	43
4.3.5 <i>UK: RCGP Research and Surveillance Centre (UK_RCGP)</i>	46
4.3.6 <i>NL: Integrated Primary Care Information (NL_IPCI)</i>	48

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 3/76

4.3.7	<i>ES: Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria (ES_BIFAP)</i>	51
4.3.8	<i>IT: ASL della provincia di Cremona (IT_ASLCR)</i>	53
4.3.9	<i>IT: Pedianet (IT_PEDIANET)</i>	55
4.3.10	<i>FI: Finnish HPV cohort (FI_UTAHPVCHRT)</i>	58
CHAPTER 5. FURTHER STEPS IN FINGERPRINTING		60
CHAPTER 6. CONCLUSIONS		60
ANNEXES		61
ANNEX 1. DATABASE POPULATION FINGERPRINT INSTRUCTIONS		62
ANNEX 2: POSITIVE AND NEGATIVE VACCINE-OUTCOME ASSOCIATIONS		74

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 4/76


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	5.2	Title	Initial Fingerprinting of the participating health care databases
Work package	Number	5	Title	Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring

Delivery date	Contractual	Month 12	Actual	Month 14
Status	Draft version / V1.4		Draft Final <input checked="" type="checkbox"/>	
Nature	Report <input checked="" type="checkbox"/> Prototype <input type="checkbox"/> Other <input type="checkbox"/>			
Dissemination Level	Public <input type="checkbox"/> Confidential <input checked="" type="checkbox"/>			


Authors (Partner)	EMC, GSK		
Responsible Author	Miriam Sturkenboom (EMC) Peter Rijnbeek (EMC) Marius Gheorghe (EMC) Daniel Weibel (EMC) Benedikt Becker Germano Ferreira (GSK)	Email	m.sturkenboom@erasmusmc.nl p.rijnbeek@erasmusmc.nl m.gheorghe@erasmusmc.nl d.weibel@erasmusmc.nl b.becker@erasmusmc.nl germano.x.ferreira@gsk.com
	Partner	EMC, GSK	Phone
Description of the deliverable	This deliverable aims to describe the initial characterization of the databases that are available for the POC phase I studies, in the consortium		
Key words	Proof of concept, post-authorization studies, platform, distributed network. Collaboration, standard procedures		

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 6/76


DOCUMENT HISTORY

NAME	DATE	VERSION	DESCRIPTION
Miriam Sturkenboom	29-10-2014	1.0	First draft
Peter Rijnbeek	03-11-2014	1.0	Review
Marius Gheorghe	06-11-2014	1.0	Comments, data and graphics
Benedikt Becker	06-11-2014	1.0	Description of ontology
Miriam sturkenboom	07-11-2014	1.1	Review and incorporation of comments
Germano Ferreira	17-11-2014	1.1	Review
Miriam sturkenboom	18-11-2014	1.2	Incorporation of comments
Hanne Dorthé	23-11-2014		
Lisen Arnheim-Dahlstrom	21-11-2014		
Elisa Martin	21-11-2014		
Marco Villa	21-11-2014		
Anna Cantarutti	21-11-2014		
Lieke van der Aa, Patrick Mahy	21-11-2014		
Xavier Kurz	24-11-2014		
Miriam Sturkenboom, Marius Gheorghe, Benedikt Becker	24-11-2014	1.3	Incorporation of comments, recoding of database names, inclusion of references
Vincent Bauchau	26-11-2014		
Klara Berenci	27-11-2014		
Miriam Sturkenboom	28-11-2014	1.4	Incorporation of comments

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 7/76


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)
 - **EC.** European Commission
 - **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)
 - **CRX.** Crucell Holland BV (Netherlands)
 - **PFIZER.** Pfizer Limited (United Kingdom)
 - **TAKEDA.** Takeda Pharmaceuticals International GmbH (Switzerland)

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 8/76


- This preliminary glossary of terms used in the review will be further developed in ADVANCE accumulating the contributions for the various deliverables and work packages.
 - **Aggregated Data:** Summarized information
 - **Anonymised Data:** Data that cannot be traced back to the individual patient
 - **Data “controller”:** the “controller” shall mean the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data
 - **Data “processor”:** is anyone who processes personal data for a controller
 - **Data custodian:** An individual / organisation / or group responsible for providing the data to the ADVANCE platform. These can be data controllers or processors
 - **(Datasource) Fingerprinting:** characterization of the actual data characteristics of a database.
 - **ETL:** Extract, Transform, Load
 - **Harmonised Data:** The harmonised data follow a consensus and are formatted in the same way
 - **Identified or identifiable natural person:** means anyone who “can be identified, directly or indirectly, in particular by reference to an identification number or by one or more factors specific to his/her physical, physiological, mental, economic, cultural, or social identity.”
 - **Ontology:** An ontology is defined as a hierarchy, or specification of clinical concepts and their relationships within an given domain¹
 - **Original Data:** Data, as maintained by the Data Source or any organization which collects the data, before inclusion in the platform
 - **Personal Data:** any information relating to an identified or identifiable natural person (data subject); \
 - **Proof-of-concept studies (POC):** studies that will be conducted during ADVANCE project to evaluate and provide evidence of whether proposed methods, guidances and methods of integration and collaboration work
 - **Third party:** is anyone who processes data under “the direct authority” of a controller or processor
 - **Study:** de novo information generation following a specific question and protocol

¹ Gruber T. A translation approach to portable ontology specifications. Knowledge Acquisition 1993; 5 (2):199–220. doi:10.1006/knac.1993.1008

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 9/76

Participating institutions in WP 5

Participant No. / Short name	Names
Part.1/ EMC	Miriam Sturkenboom, Peter Rijnbeek, Johan van der Lei, Daniel Weibel, Caitlin Dodd, Kartini Gadroen, Benedikt Becker
Part.21/ GSK	Germano Ferreira
Part.2/ UNIBAS-UKBB	Jan Bonhoeffer, Yolanda Brauchli
Part.3/ EMA	Peter Arlett, Xavier Kurz
Part.5/ SURREY	Simon de Lusignan, Filipa Ferreira, Harshana Liyanage
Part.6/ P95	Thomas Verstraeten , Kaatje Bollaert
Part.10/ PEDIANET	Anna Cantarutti, Lara Tramontan, Luigi Comachio
Part.11/ KI	Lisen Arnheim Dahlström
Part.12/ ASLCR	Silvia Lucchi, Marco Villa, Salvatore Mannino
Part.13/ AEMPS-BIFAP	Elisa Martín, Consuelo Huerta, Miguel Gil , Ana Llorente
Part.14/ AUH	Klára Berencsi , Lars Pedersen
Part.15/ UTA	Matti Lehtinen, Simopekka Vanska
Part.16/ WIV-ISP	Patrick Mahy
Part.18/ SSI	Kåre Mølbak
Part.19/ RCGP	Douglas Fleming, Hayley Durnall
Part.20/ RIVM	Marianne Van der Sande, Susan Hahne, Jacco Wallinga Nicoline van der Maas , Hester de Melker, Wim van der Hoek , Fiona van der Klis, Nynke Rots , Willem Luytjes
Part.22/ SP	Caroline Legendre
Part.24/ SP MSD	Hélène Bricout, Susanne Hartwig, Géraldine Dominiak

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghie, Daniel Weibel, Germano Ferreira		Security: 10/76

EXECUTIVE SUMMARY

In the ADVANCE project, "fingerprinting" is used to describe the activities of producing a standard, automated description of observational healthcare database contents in order to understand data quality and appropriateness for studies on the benefits and risks of vaccines.

This ADVANCE project deliverable 5.2 describes the program and initial output for fingerprinting the databases available in the ADVANCE project. This deliverable is a first fingerprinting deliverable and it will be followed by a final one in project month 31 (D5.4).

Since databases are heterogeneous, we need to better understand the differences and the suitability of the databases for the various information requirements for benefit/risk analyses of vaccines.

In ADVANCE we will fingerprint


- 1) The population characteristics (size, time, follow-up, representativeness)
- 2) Vaccination coverage (selected vaccines)
- 3) Events of special interest: vaccine preventable disease, events of special interest for safety and events that define risk populations
- 4) Size of positive and negative vaccine-event associations

In this report we present the results of the population fingerprint of the databases that participate in ADVANCE and which can be included in the first phase of proof of concept studies.

The table below describes the data that are available at this point in time, throughout the ADVANCE project more data may become available.

Country	Database	Type of database	Years covered	Total Number of persons	Person years	Median follow-up (years)	25 th -75 th quartile follow-up (years)
UK	THIN	GP med. Record	1994 - 2013	8,326,238	51,333,610	4.91	1.8-10
	RCGP	GP med. Record	2003 - 2014	2,043,800	12,839,908	6.08	2.25-10.7
NL	IPCI	GP med. Record	1996 - 2014	1,786,405	5,468,684	2.83	1.5-4.3
	RIVM	Case surveillance					
ES	BIFAP	GP med. Record	1998 - 2014	4,800,538	25,680,609	4.91	2.3-8.2
IT	PEDIANET	GP med. Record	2004 - 2014	77,021	350,797	4.25	1.8-7.1
	ASLCR	Provincial record linkage	2002 - 2013	454,188	4,226,316	11.90	6.7-11.9
SE	KI	National record linkage	1998 - 2010	9,421,687	110,841,514	12.91	12.9-12.9
FI	HPV Cohorts	Trial/cohort	Prospective sinc 2007				
DK	AUH	Regional record linkage	2004 - 2013	1,741,051*	12,352,156*	10	3.7-10
	National (SSI)	National record linkage	1996 - 2014	7,512,032	103,492,835	18.66	8.3-18.7
Total				34,421,909	314,234,273		

*AUH data are also part of DK_SSI

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

Differences in stability of the population and representativeness differ. These should be considered in the POC studies, both for the design and the analyses.

Chapter 1. BACKGROUND

The ADVANCE project emerged from the need to build an integrated and sustainable framework for the continuous monitoring of the benefit/risk of vaccines.

The ADVANCE vision is to deliver “**Best evidence** at the **right time** to support **decision-making on vaccination** in Europe”, and the mission is to establish a prototype of a sustainable and compelling system that rapidly provides best available scientific evidence on vaccination benefits and risks post-licensure for well informed decisions. This will be achieved by developing and testing a code of conduct, rules of governance, technical infrastructures, data sources, methods, and workflows in a European network of stakeholders.


ADVANCE aims to build an integrated and collaborative framework. By *integrated* we currently mean a coordinated and harmonised approach by all stakeholders across the European member states with a stake on benefit and safety of vaccines and vaccination programs. European member states will continue having their national level approaches. ADVANCE will complement, not replace, activities in place at national level to monitor the benefits and risks of vaccines. It is envisioned that the following areas are those where the ADVANCE collaborative framework will bring added value:

- 1) To increase statistical power and allow for stratifications to subpopulations
- 2) To exploit variability in exposure to different types of vaccines and schedules across member states (e.g. adjuvanted versus non-adjuvanted)
- 3) To enlarge research capacity and expertise across the member states
- 4) To be able to estimate at and understand different results between member states
- 5) To help in interpretation of results across member states and support harmonised decision-making

ADVANCE Work Package (WP) 5 captures all activities that require actual use of data from the electronic healthcare databases, surveillance data, cohort data etc. and the technical infrastructure in order to generate information on benefits and risks of vaccines. Requests for the so-called Proof of Concept studies (POCs) will arrive from ADVANCE work packages 1, 3, and 4 and will be developed in close collaboration with these work packages, experts in the consortium and subsequently sent for consortium input.

Specific tasks in WP5 are:

- 1) Implementation of an information technology infrastructure that allows for collaborative studies
- 2) Fingerprinting of all databases in ADVANCE and those identified in WP 3 that may be willing to participate in POC studies later

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 12/7 6

- 3) Creation of ontologies, mapping of vaccines to a standard dictionary and outcome mapping for the respective proof of concept studies
- 4) Conduct of proof of concept studies proposed by WPs 1, 3, 4
- 5) Stakeholder feedback evaluation on the use of data and processes based on the proof-of-concept studies

In deliverable 5.1 we described task 1: the available infrastructure and the proposed way of collaboration and workflow for generating new evidence from secondary use of health care data. This workflow is the basis for the collaborative POC studies on vaccine benefit and risk which will start in project month 18.

The concept of bringing data together within and across countries with the purpose of addressing vaccine benefit/risk questions in a collaborative and integrated approach can be addressed in several ways with respect to:

- 1) Standardization of protocols to conduct studies on multiple data sources
- 2) Local data extraction
- 3) Transformation of the data into analytical datasets
- 4) Pooled analyses of data

In ADVANCE steps 1, 3 and 4 will be harmonized and centrally coordinated for specific studies upon appropriate protocol approval (see figure 1.1).

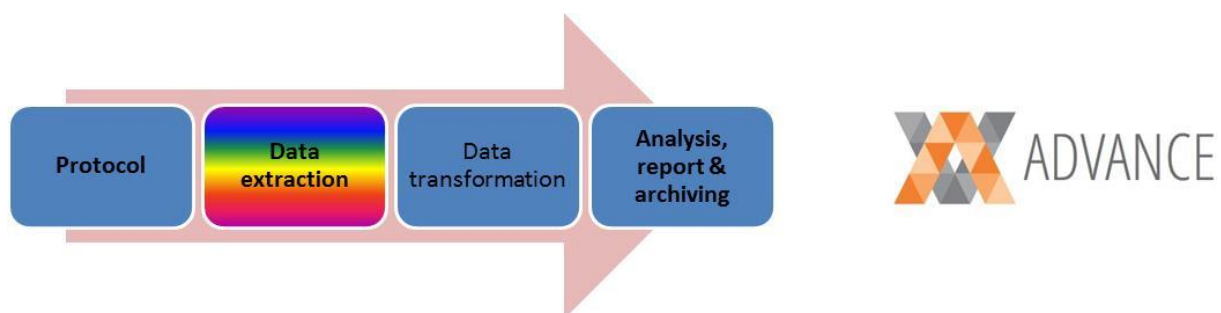



Figure 1.1: Distributed collaborative information generation workflow, with common protocol, standardized transformation and shared analyses while data extraction and original data remain local. This is the accepted ADVANCE approach.

We cannot fully harmonize step 2 for the following reasons:

- 1) Different structures of health care systems across EU member states
- 2) Different types of databases within a country and across EU member states (i.e. health care databases, claims databases, inpatient databases, surveillance networks, laboratory data, microbiology data, vaccination registries, medical record databases), if possible all these databases will be fingerprinted
- 3) Different content of similar types of databases across EU member states

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

- 4) Different coding/terminologies and language of similar information between databases in different EU member states

Since it is very unlikely that all these differences will (soon) be resolved, we have tried to find ways to understand the differences, to gain insight into the underlying determinants and mechanisms of data generation, and to address these differences in a consistent and informed way, such that we can actually use the data for the purpose of vaccine benefit/risk monitoring. This is translated in the following approach to data acquisition:


- 1) Use of local source data knowledge: Full involvement of the database custodian in data extraction processes and interpretation of the data to appreciate differences
- 2) Semantic harmonization: mapping of terminologies and variables for population, events (outcomes and covariates), vaccines and drugs & creation of ontologies and mappings of codes and terms to allow for specific data to be integrated into a common data model
- 3) Fingerprinting: (i.e characterizing) of what data is actually available in the databases by real data extraction (transparency)
 - a. Stepwise conversion of specific required study data into a simple common data model
 - b. Describing the data quantitatively using a common script and visualization
 - c. Iterative harmonization and verification of data extraction steps under item 2 across the databases
 - d. Benchmarking of data extracted against available external sources of information.
- 4) Knowledge & information management: Reporting of generated evidence and knowledge and making it available and accessible

This deliverable D5.2 will start with a description of the anticipated ADVANCE fingerprinting program (population, vaccine and events), followed by a meta description of the participating databases. Then we will present the first results of the population fingerprints, fingerprinting of vaccines and events will be presented in the next fingerprinting deliverable (D5.4) but will follow the program as described.

Chapter 2. Overview of fingerprinting program

In ADVANCE, we aim to characterize datasources on different levels. In WP3 the descriptive, content and ethical/privacy related data will be collected for each relevant database in the EU (Meta and meso level) (e.g. D3.2)

In WP 5 we will describe data on the micro level through fingerprinting. The entire fingerprinting program will describe the data in the databases and assess the suitability of the data for specific vaccine benefit and risk studies by data extraction from the databases.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

2.1 Descriptive data

In the fingerprinting task (Task 5.2), we will describe the databases based on the data that are locally extracted into the Jerboa common data model. The simple common input files (standardized files that will be created locally and which will allow launching of a standardized script to produce aggregated data) which were defined in the D5.1 and that we will use in ADVANCE comprise three files:

- 1) population
- 2) exposure (vaccines & drugs)
- 3) events (outcomes and covariates)

Database custodians will have to transform their local data into these common input files. These input files will be processed locally by a common tool called Jerboa Reloaded (see D5.1) that will generate aggregated fingerprinting data. Such input files are specific to the fingerprinting task, further input files will be defined on a study by study basis part of the ADVANCE research plan.

2.1.1 Population fingerprinting

2.1.1.1 Fingerprinting common input file - Population

The common input files for the population have been described in D5.1. It comprises the following variables both for the Jerboa Reloaded as well as SAS track.

PatientID	Patient ID
Gender	Can be either F or M, for Female or Male respectively
Birthdate	Date of birth
Startdate	Date when the patient is eligible to be included in the study. This is typically the date the patient is entered into the registration system (run in periods may be defined in Jerboa)
Enddate	Date at which the patient is no longer eligible for inclusion in the study (i.e. last data extraction, end of the study, death, transferring out from the practice/catchment area, last data update of data provider)

```

patientid,gender,birthdate,startdate,enddate
1,F,19590601,19950701,20050701
2,M,19960201,19960201,20050701
3,F,19850801,19950701,19971001
4,M,19881101,19950701,20050701
5,F,19830301,19950701,20050701
6,M,19550301,19950701,20050701
7,M,19590701,19950701,19971001
8,F,19840301,19950701,20040401
9,M,19880101,19950701,20050701
10,M,19601101,19950701,20050701
11,M,19651201,19950701,19970601
12,M,19431101,19960101,20050701
13,F,19920101,19950701,20050701
14,M,19460501,19950701,20050701
15,M,19560501,19950701,20050701

```


 IM - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 15/7 6

Figure 2.1: Example of local population input files

2.1.1.2 Running population fingerprint

All databases that agreed to participate in WP 5 were invited to run the population fingerprint. This could have been performed by a Jerboa (JAVA) track as well as a SAS track. Double coding of the Jerboa & SAS track was coordinated by Peter Rijnbeek (EMC). Marius Gheorge (EMC) was responsible for the JAVA code, Maria de Ridder(EMC) and Klara Berencsi (AUH) for the SAS code. The Jerboa code was developed for EMIF and is used in ADVANCE as EMC foreground.

Database Instructions for the Jerboa track are included in annex 1. A summary of the steps is provided below.

Data custodians could download programs and instructions from the OCTOPUS remote research environment (an server that allows remote access for collaborative analyses) using an FTP client (FileZilla). Instructions on how to use FileZilla and a video are available for all OCTOPUS users. If databases had questions, special teleconference meetings were organized.

The user is required to copy the Jerboa programs into a folder and adds the population input file. Jerboa then checks the input files and will report any errors found (e.g. wrong date format, empty fields). If no errors are found in the input data file(s) the application will proceed. An indication of the time left to finish the current step is given in the progress bar on the bottom of the screen.

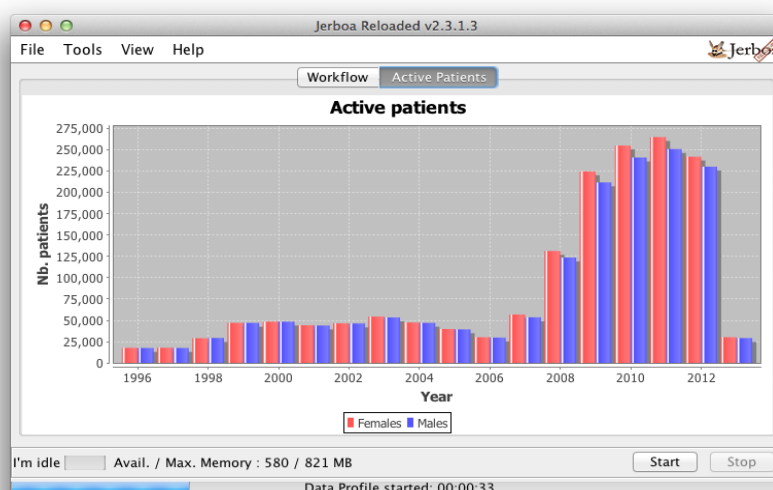



Figure 2.2: screenshot of Jerboa running and feedback locally

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

During the run user feedback is given in the form of a graph showing the active male and female patients in the database per year (See figure 2.2). When the fingerprinting is finished, fingerprinting graphs are presented to the user in a separate tab. In the results section of this deliverable, we included examples of these graphs. In the working folder an Adobe Acrobat pdf is created containing all the plots for local review, error finding and approval to release the data by the data custodian.

Jerboa generates the following characteristics for the population per gender, calendar year and age group:

1. Active subjects
The number of patients with at least one day of patient time
2. Age distribution at patient end
The age in months at the time the patient leaves the database (or at the last data update)
3. Age distribution at patient start
The age in months at the time the patient enters the database per year and age group
4. Age distribution by start of calendar year
The age of the patient at the first of January of each calendar year
5. Births in calendar year
The number of births in a calendar year
6. Amount of Observation time per subject
PatientEnd – PatientStart Period since the date patient enters the database to the date patient leaves the database
7. Observation time after start of a year
PatientEnd – first day of the year
8. Observation time before the start of a year
First day of the year – PatientStart
9. Observation time in a year
The amount of patient time available in year and/or per age group
10. Observation time in years
Count of the number of patients with <1,1,2,3,4, etc. years of patient time


The created result file contains the following columns for each characteristic:

1. The aggregation levels:

Name1	Name of the first aggregation variable, for example YEAR
Value1	Value of the first aggregation variable, for example 2001
Name2	Name of the second aggregation variable, for example AGE
Value2	Value of the second aggregation variable, for example 10-14
Gender	M, F, or T (total)

2. The statistics that are calculated for each aggregation level if appropriate:

Min, Max, Count/Sum, Mean, 25th percentile, Median, 75th percentile, SD

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

In the final step Jerboa will produce an encrypted file that contains the log file, the created pdf, the used script and the result file described above. If the database custodian is satisfied with the output and has obtained local clearance, then this file can then be uploaded to Octopus using the FileZilla procedure as described in the Octopus instructions.

In OCTOPUS the Jerboa results are decrypted and archived per database in the ADVANCE data repository in time stamped folders. All files generated by Jerboa automatically contain the version number of both Jerboa and the script and are time stamped as well. The log file contains the location of the input file on the machine of the data custodian. The script is added to the encrypted file for review. i.e. to make sure the data custodian has not made unknown changes in the script.

In the OCTOPUS environment R scripts are available that allow plotting of multiple database outputs in one graph. All results can be accessed by the partners using a remote desktop session (see D5.1 for more details)

2.1.2 Vaccine fingerprint

2.1.2.1 Common input file


The common input files for the vaccinations has been discussed in WP 5 after the delivery of the D5.1 From the databases the following variables will be requested

PatientID*:	Patient Identifier
Date*:	Different dates may be available, in this field we need the date of or nearest to administration (from WP 3 we will assess which date type this will correspond to)
ATC*:	Type of vaccine (ATC level) (mandatory): ATC should be provided in most precise level
Brand:	Brand name of the vaccine (as linkage to ontology)
Dose:	Recorded Sequence dose of vaccine (booster) (i.e 1, 2,3,4)
Lot:	Lot/ batch number of vaccine


The initial discussions with the databases showed that although additional information is often not available, most databases will have the ATC code or at least part of it. In the Anatomical Therapeutic Chemical (ATC)² classification system, the active substances are divided into different groups according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. Drugs are classified in groups at five different levels. The drugs are divided into fourteen main groups (1st level), with pharmacological/therapeutic subgroups (2nd level). The 3rd and 4th levels are chemical/pharmacological/therapeutic subgroups and the 5th level is the chemical substance³. Vaccines are coded in different Anatomical groups although the majority are part of the therapeutic subgroup J07. (see figure 2.3)

² http://www.whooc.no/atc_ddd_index/


³ http://www.whooc.no/atc/structure_and_principles/#principles

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 18/7 6

[Home](#)
[ATC/DDD application form](#)
[Order publications](#)
[WHO Centre](#)
[Contact us](#)
[Log in](#)



WHO Collaborating Centre for Drug Statistics Methodology



Norwegian Institute of Public Health

News

ATC/DDD Index

Updates included in the ATC/DDD Index

ATC/DDD methodology

ATC

DDD

ATC/DDD alterations, cumulative lists

ATC/DDD publications

Use of ATC/DDD

Courses

Meetings / open session

Deadlines

Links

Postal address:
WHO Collaborating Centre for Drug Statistics Methodology
Norwegian Institute of Public Health
P.O.Box 4404 Nydalen
0403 Oslo
Norway

Visiting/delivery address:
Marcus Thranes gate 6
0473 Oslo
Norway

Tel: +47 21 07 81 60
Fax: +47 21 07 81 46
E-mail: whocc@thi.no

[Copyright/Disclaimer](#)


[New search](#)

Found 39 entries containing 'vaccin'.

J07	VACCINES
J07A	BACTERIAL VACCINES
J07B	VIRAL VACCINES
J07C	BACTERIAL AND VIRAL VACCINES, COMBINED
J07X	OTHER VACCINES
J07AC	Anthrax vaccines
J07AD	Brucellosis vaccines
J07AE	Cholera vaccines
J07AF	Diphtheria vaccines
J07AG	Hemophilus influenzae B vaccines
J07AH	Meningococcal vaccines
J07AJ	Pertussis vaccines
J07AK	Plague vaccines
J07AL	Pneumococcal vaccines
J07AM	Tetanus vaccines
J07AN	Tuberculosis vaccines
J07AP	Typhoid vaccines
J07AR	Typhus (exanthematicus) vaccines
J07AX	Other bacterial vaccines
J07BA	Encephalitis vaccines
J07BB	Influenza vaccines
J07BC	Hepatitis vaccines
J07BD	Measles vaccines
J07BE	Mumps vaccines
J07BF	Polioomyelitis vaccines
J07BG	Rabies vaccines
J07BH	Rota virus diarrhea vaccines
J07BJ	Rubella vaccines
J07BK	Varicella zoster vaccines
J07BL	Yellow fever vaccines
J07BM	Papillomavirus vaccines
J07BX	Other viral vaccines
J07CA	Bacterial and viral vaccines, combined
L03AX03	BCG vaccine
J07AE51	cholera, combinations with typhoid vaccine, inactivated, whole cell
L03AX12	melanoma vaccine
J07AH09	meningococcus B, multicomponent vaccine
J07AH06	meningococcus B, outer membrane vesicle vaccine
J06BB07	vaccinia immunoglobulin

Figure 2.3: Entries for vaccination in ATC system (dated October 2014)

The divisions are made between bacterial (J07A), viral (J07B), bacterial & viral (J07C), other vaccines (J07D) and cancer vaccines (in L03). Bacterial vaccines are divided in 14 subgroups, viral vaccines are divided in 13 subgroups, grouping is based on the vaccine preventable disease. Within the groupings the most detailed level finishes with the type of antigen (see figure 2.4).

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 19/76

J07BB Influenza vaccines

ATC code	Name
J07BB01	<u>influenza, inactivated, whole virus</u>
J07BB02	<u>influenza, inactivated, split virus or surface antigen</u>
J07BB03	<u>influenza, live attenuated</u>

Figure 2.4: Most detailed ATC codes for vaccines

For full benefit risk analyses we will need additional information on the vaccines (e.g. valence, excipients). This information needs to be retrieved from other sources, therefore we will create a vaccine ontology that can provide additional information on the vaccines which may be useful for analysis as well as to enable the fingerprinting across multiple databases. The ontology will be part of the D5.5

2.1.2.2 Vaccine ontologies

Ontologies can be used to represent an ontological structure of a domain of discourse, contextual knowledge and lexical information in a single model.

The ontological structure is a *directed acyclic graph* (DAG) whose nodes are the concepts of a domain (individuals and classes). The connection of the DAG describe the *is-a* relation between concepts. That is, when the concepts *A* and *B* are in the *is-a* relation, then *A* shares the properties of *B* but is more specific. In the domain of vaccines, the set of nodes is comprised of all individual vaccine products (e.g. *FluMist*[®], *Pandemrix*[®]) and possible classes of vaccines (e.g. viral vaccines, flu vaccines). This allows the aggregation of vaccines by relevant classes for analyses. See figure 2.4 for an extract of a vaccine ontology.

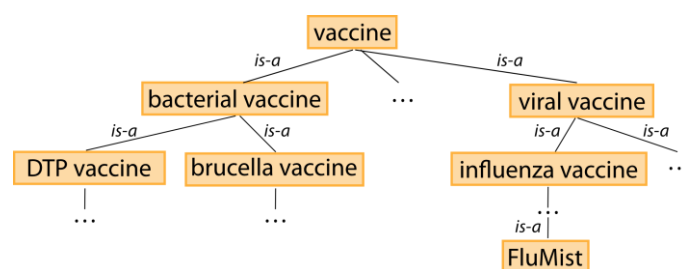



Figure 2.4 Extract of a vaccine ontology with *is-a* relations.

An ontology may further contain contextual information, such as concepts that are unrelated with respect to the *is-a* relation but related in different relations. For example, such a rich vaccine ontology may contain substances as concepts with relations between vaccines and substances to mark adjuvants or excipients of a vaccine. This allows the classification of individual vaccines and vaccination exposure by their components.

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

Finally, the lexical information of a concept is comprised of *terms* that are used in free-text *corpora* to refer to that concept. This information builds the basis of the automatic detection of concepts in text mining of for example literature or social media or databases.⁴

The terms ontology, controlled vocabulary, thesaurus are often used in similar ways – although they capture different characteristics of a domain. We will use the term ontology to refer to a model of ontological, contextual and lexical information.

Available vaccine ontologies

Several data sources are already available, that provide ontological, lexical and/or contextual information about vaccines. We will first describe some of their general characteristics and then assess their ontological and lexical information content.

The *Anatomical Therapeutic Chemical Classification System* (ATC) is provided and controlled by the WHO to classify drugs by anatomical, therapeutically, pharmacological and chemical characteristics⁵.

The *Medical Subject Headings* (MeSH) is a controlled vocabulary of medical terms used to index documents in MEDLINE/Pubmed. It is structured by *descriptors* that contain concepts and *is-a* relations⁶.

The *Unified Medical Language System* (UMLS) combines more than hundred medical source ontologies (including MeSH). UMLS unifies equivalent concepts from different sources and maintains an overall ontological structure.⁷

The *Vaccine Investigation and OnLine Information Network* (VIOLIN) provides the *Vaccine Ontology* (VO). It is a community-driven ontology of domain specific knowledge about vaccines such as documented adverse events, adjuvants, components and manufacturers⁸.

⁴ Feldman, Ronen, and James Sanger, eds. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, 2007. See chapter III: "Text mining preprocessing techniques" and chapter IV: "Categorization" for an general introduction to text mining. The following paper describes dictionary-based concept identification: Schuemie, Martijn J., Rob Jelier, and Jan A. Kors. "Peregrine: Lightweight gene name normalization by dictionary lookup." Proc of the Second BioCreative Challenge Evaluation Workshop. 2007.


⁵ ATC description at http://www.whocc.no/atc/structure_and_principles/ and index at http://www.whocc.no/atc_ddd_index/

⁶ Johnston D, Nelson SJ, Schulman J-LA, Savage AG, Powell TP. Redefining a Thesaurus: Term-Centric No More. Proceedings of the AMIA Symposium 1998:1025. MeSH browser at <http://www.nlm.nih.gov/mesh/MBrowser.html>

⁷ Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91. UMLS browser at <https://uts.nlm.nih.gov/metathesaurus.html> (login required).

⁸ He Y, Racz R, Sayers S, Lin Y, Todd T, Hur J, Li X, Patel M, Zhao B, Chung M, Ostrow J, Sylora A, Dungarani P, Ulysse G, Kochhar K, Vidri B, Strait K, Jourdain GW, Xiang Z. Updates on the web-based VIOLIN vaccine database and analysis system. Nucleic Acids Res. 2014 Jan.

Vaccine Ontology browser at <http://www.ontobee.org/browser/index.php?o=VO>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

Furthermore, two data sources provide information specifically about medical drugs authorized in the European Union. The *European Pharmacopoeia* (Ph. Eur.)⁹ provides recognised common standards for the quality of medicines and components. This might be incorporated into an ontology of vaccines as contextual knowledge. Since this information is only available at the level of vaccine classes, and provided only as monographs, further assessment of the viability is necessary.

The European Medical Agency (EMA) creates a database of information on medicinal products authorized in the EU as provided by article 57(2), of regulation (EC) 726/2004. This database will be available to the ADVANCE consortium in Q1/2015.

Preliminary assessment of the available sources

Ontological and lexical information about vaccines was extracted from ATC, MeSH, UMLS and VO. Several complexity measures have been proposed to determine fundamental characteristics of ontologies¹⁰. As a first evaluation of the extracted information, the following measures were applied:

- the *total size* of the entire ontology as the number of concepts in the ontology
- the *unique size* as the number of unique concepts in the ontology (disregarding shared branches of the ontology introduced by concepts with multiple parent concepts)
- the *maximal depth* of the ontology
- the *average branching factor* as the average number of child concepts
- the *number of terms* from all concepts (disregarding shared branches)
- the *average number of terms* per concept (disregarding shared branches)

The results of the evaluation are provided in table 2. The simple ontological structure of ATC and MeSH is apparent from the relatively low total sizes of the ontologies (111 and 101). However, MeSH provides 4.71 terms per concept whereas ATC provides only a single term of the concept (its name). UMLS contains MeSH as a (principal) source, and provides more ontological and lexical information. It contains 695 unique vaccine concepts with 5.51 terms at average.

VO provides very rich ontological information with 2144 unique concepts in a deeply structured manner (maximal depth of 11 with average branching factor of 4.87). But with only 1.07 terms per concept at average it provides very little lexical information apart from the concept's name.

Table 2: Different measures of the ontological and lexical information content on available data sources.


Ontology	total size	unique size	max depth	avg branching	num terms	avg num terms
ATC	111	111	4	3.66	111	1.00
MeSH	101	83	5	4.34	391	4.71
UMLS	776	694	7	4.96	3826	5.51
VO	2148	2144	11	4.87	2297	1.07

⁹ Council of Europe. *European Pharmacopoeia*, 8th Ed. Strasbourg: Council of Europe, 2013.

Online at <http://online6.edqm.eu/ep804/> (login required)

¹⁰ Yang, Zhe, Dalu Zhang, and Chuan Ye. Evaluation metrics for ontology complexity and evolution analysis. *e-Business Engineering*, 2006. ICEBE'06. IEEE International Conference on. IEEE, 2006.

Zhang, Hongyu, Yuan-Fang Li, and Hee Beng Kuan Tan. Measuring design complexity of semantic web ontologies. *Journal of Systems and Software* 83.5 (2010): 803-814.

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

Next steps

Two further steps need to be accomplished to complete this preliminary assessment of available sources. Firstly, ontological and lexical information about vaccines needs to be extracted from Ph. Eur., EMA's article 57 database and SNOMED CT and evaluated as above. Secondly, the sources need to be evaluated for their actual value to detect concepts in free-text corpora such as scientific literature, medical records and public or social media. This will be done by extracting concepts with the available lexical information from free-text corpora and comparing the results with manual annotations.

Furthermore, the creation of a mapping between ontological rich and lexical rich sources (such as UMLS and VO) will facilitate text mining in different types and levels of aggregation.

In D5.5 further options of how to present the ontology knowledge base will be explored and presented. Once the ontology can be used we can proceed with vaccine fingerprinting.

2.1.2.3 Combination of vaccine data in databases and ontologies

For vaccine benefit-risk analyses we may want to look at different levels of aggregation for vaccines. Analyses may be run on a class level (e.g. vaccine preventable disease level), which can be obtained from the ATC code, by type of antigen (ATC code), but if we want to look at vaccines with certain excipients we need to aggregate across vaccines. Information from the ontology will be mapped to the input files, based on the product name. This will be done in Jerboa/SAS, scripts will be provided. Thereto the input files will be enriched with additional variables.

2.1.2.4 Analysis of vaccination fingerprint

The vaccine fingerprint will be described by using vaccination coverage estimates as well as the number of doses per person.


To fingerprint the datasources in terms of vaccine coverage /uptake data we will:

- 1) Estimate coverage by age, gender, calendar year of the following vaccinations and compare these to the monitoring data from WHO¹¹, the ECDC funded VENICE II consortium¹² and available national statistics: Bacille Calmette-Guérin (BCG) vaccine, the third dose of diphtheria and tetanus toxoid and pertussis vaccine (DTP3), the third dose of polio vaccine – either oral polio vaccine or inactivated polio vaccine, the first dose diphtheria and tetanus toxoid and pertussis vaccine (DTP1) and the third dose of haemophilus influenza type b (Hib3), seasonal Influenza (compared to VENICE)¹³ and the first dose and third dose of human

¹¹ <http://www.who.int/immunization/en/>

¹² http://venice.cineca.org/the_project.html

¹³ VENICE II: Go on combining our efforts towards a European common vaccination policy! F D'Ancona on behalf of VENICE II group. Eurosurveillance, 2009, vol. 14 n.12: Seasonal influenza immunisation in Europe. Overview of recommendations

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

papillomavirus vaccinations (HPV)¹⁴. These vaccines are also reported to WHO and provide for benchmark

- 2) Assess timing of childhood immunizations in databases (by age) and compare these with local recommended schedule.
- 3) Calculate the number of doses per vaccinated individual.

Information from appraisal of methods in coverage estimations

The deliverable 4.1 describes that a search in the published literature and reports from the Vaccine European New Integrated Collaboration Effort (VENICE) consortium showed that there are currently no standardized method to estimate or report vaccine coverage in Europe. The most commonly used estimations are the number of vaccinated children at 12 months, at 24 months or at school entry. The D4.1 states that coverage estimates obtained using fixed months most likely will underestimate coverage compared with estimates obtained using birth cohorts. These two types of estimates might not be readily comparable. From the D4.1 discussion we also learn that childhood vaccines schedules differ a lot between European countries ¹⁵.

The D4.1 concludes that WHO receives yearly vaccine coverage estimates from all over the world, but they do not take into account that different vaccination coverage methods are used. Very few reports address limitations and uncertainties in the reported data, which may provide a false sense of certainty. Vaccine schedules do vary between countries and the different schedules can make coverage estimation difficult, however there are certain age groups where very few countries recommend vaccinations e.g. from 2-4 years and from 8-10 years. From a benefit-risk point of view, an additional coverage estimation should be performed for the age groups 2-4 years and 8-10 years to obtain more comparable estimates.

ADVANCE approach based on D4.1 recommendations


Within the fingerprint task on coverage, we will look at describing coverage at 12, 24, 48 and 120 months (BCG, DTP, polio, Hib), and a cumulative approach (Kaplan Meier) for birth cohorts. For HPV, we will assess coverage at age 16. For seasonal influenza vaccination, we will assess coverage by year of age. Benchmarking against other sources will be done, but have limitations as we learned from the D4.1.

Timing of vaccinations will be described by plots for age of vaccination by type and dose of vaccine and this will be compared with the information in the vaccine schedules.

and vaccination coverage for three seasons: pre-pandemic (2008/09), pandemic (2009/10) and post-pandemic (2010/11). J Mereckiene, S Cotter, A Nicoll, P Lopalco, T Noori, J T Weber, F D'Ancona, D Lévy-Bruhl, L Dematte, C Giambi, P Valentiner-Branth, I Stankiewicz, E Appelgren, D O'Flanagan, the VENICE project gatekeepers group. Eurosurveillance, 19 (16) 2014.

¹⁴ Health technology assessments on human papillomavirus vaccinations in Europe: a survey from Venice network, Frédérique Dorléans, Daniel Lévy-Bruhl, Cristina Giambi, Fortunato D'Ancona, Giuseppe La Torre, Suzanne Cotter, Jolita Mereckiene, Pawel Stefanoff, Eva Appelgren and the Vaccine European New Integrated Collaboration Effort (VENICE II) project gatekeepers, Italian Journal of Public Health, Volume 9, N. 1 (2012)

¹⁵ <http://vaccine-schedule.ecdc.europa.eu/Pages/Scheduler.aspx>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

Required steps

Prior to completion of the fingerprinting of vaccine coverage work under the Task 5.3 (M6-M36) and subsequent delivery of D5.4 (M31), we need to create the required vaccine ontology (D5.5 M31) and the Jerboa/SAS modules to process the data.

2.1.3 Event fingerprinting

2.1.3.1 Common input file

The common input files for the events have been described in D5.1. the file comprises the following variables both for the Jerboa as well as SAS track.


PatientID	Patient ID
Date	Date of the event
EventType	Type of event according to categories (any type of medical condition of interest): e.g. GBS, Narcolepsy..)
Code	Code of the event based on for example ICD-9 or free text, lab valueslaboratory values, etc.

The event type is a given name for project specific purposes. The code is the code that was used to extract the event, which may be in ICD-9/10, READ, ICPC, text, drugs, lab valueslaboratory values etc. The date of the event is what the database custodian needs to extract, it should be the earliest date that the event is recorded.

2.1.3.2 Selection of events and mapping

For the fingerprinting of events we have currently selected 78 events. It is important to start the fingerprinting of these events, as the ground work of defining, mapping, extraction and harmonization needs to be done in all databases in order to run any POC study or for that matter any study focusing on benefits/risk of vaccinations. The events will comprise:

- 1) **Vaccine preventable disease:** Diphtheria, Invasive Haemophilus influenzae disease, Invasive meningococcal disease, Invasive pneumococcal disease (IPD), Measles, Mumps, Pertussis, Polio, Rabies, Rubella, Tetanus.
- 2) **Events of special interest as defined by EMA** for the enhanced safety surveillance for seasonal influenza vaccines: Fever, including high grade fever, Vomiting and nausea, Malaise, Headache, Irritability (for under 5-year-old vaccinees), Crying (for under 5-year-old vaccinees), Decreased appetite, Injection site reactions (e.g. pain, erythema, swelling) including severity and persistence, Myalgia/arthritis, Hypersensitivity reactions, including ocular symptoms, Nasal congestion/rhinorrhoea, Oropharyngeal pain, Cough, Febrile convulsions, Epistaxis, Rash, Hypersensitivity reactions, including facial oedema, urticaria and very rare anaphylactic reactions, Wheezing (in young children).

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

- 3) **Serious safety events** (as defined by PRISM)¹⁶: Diabetes type 1, Idiopathic thrombocytopenic purpura, Henoch-Schönlein purpura, Tics, Guillain-Barre Syndrome, Bell's palsy, Transverse myelitis, Acute disseminated encephalomyelitis, Optic neuritis, Uveitis, Brachial neuritis, Narcolepsy, Myocarditis and pericarditis, Kawasaki disease, Bronchospasm, Intussusception, Spontaneous abortion, Still birth, Systemic Lupus erythematosus, Rheumatoid arthritis and juvenile rheumatoid arthritis, Birth defects, Anaphylactic shock (anaphylaxis) or acute systemic allergic reaction, Death
- 4) **Additional events of interest (as risk groups or other)**: Pregnancy, Acute myocardial infarction, Diabetes, Asthma, COPD, chronic bronchitis, HIV, Cancer, Chronic Kidney disease, Chronic Liver disease, Cardiomyopathies, Conduction disorders, Ventricular arrhythmia, atrial fibrillation, Cerebral ataxia, Hashimoto's disease/Graves disease, Glomerulonephritis (acute kidney disease), Autoimmune hemolytic anemia, Myasthenia Gravis, Urticaria

Events have been divided over the WP 5 members for definitions and guidance of the harmonization process.

One of the most difficult challenges in creating an integrated harmonised framework for information generation is the diversity in the content and coding of medical conditions and procedures in the electronic health care data sources. Different coding schemes for medical events (e.g. International Classification of Diseases (ICD9-CM and ICD-10), the International Classification of Primary Care (ICPC), and the READ Code (RCD) classification) and different sources of information (e.g., general practitioners' records, hospital discharge diagnoses, death registries, laboratory values, etc.) are available in various healthcare databases.

Mapping and harmonization is based on the following steps


1. **Clinical event definition.** We have agreed to define events based on ECDC definitions for vaccine preventable disease and Brighton Collaboration case definitions and standard textbooks and literature
2. **Re-use of mappings** of events that were defined in the past (e.g. in project listed in D5.1 VAESCO, SOS, ARITMO, SAFEGUARD, EU-ADR, GRIP as well as PRISM¹⁷)
3. **Use of MEDDRA** definitions for events of special interest from European Medicines Agency¹⁸
4. **Unmapped events.** For all unmapped events the following process will be used.

¹⁶ Yih WK, Lee GM, Lieu TA, Ball R, Kulldorff M, Rett M, Wahl PM, McMahill-Walraven CN, Platt R, Salmon DA. Surveillance for adverse events following receipt of pandemic 2009 H1N1 vaccine in the Post-Licensure Rapid Immunization Safety Monitoring (PRISM) System, 2009-2010. *Am J Epidemiol.* 2012 Jun 1;175(11):1120-8

¹⁷ Baker MA, Nguyen M, Cole DV, Lee GM, Lieu TA. Post-licensure rapid immunization safety monitoring program (PRISM) data characterization. *Vaccine* [Internet]. 2013 Dec

¹⁸ European Medicines Agency (EMA). Interim guidance on enhanced safety surveillance for seasonal influenza vaccines in the EU. London: EMA; 2014. Available from:

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/04/WC500165492.pdf

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

- a. Identification of Unified Medical Language System® (UMLS®) concepts for events that have not been previously mapped. This will be done by utilizing existing application tools
- b. Projection into different terminologies by using code mapper
- c. Review of codes by databases

5. Extraction of events from databases and harmonization process

Harmonization Process:

- a. Extraction of events by database custodians
- b. Review of code frequency distributions
- c. Calculation of age specific event incidence rates
- d. Discussion with and between databases about differences
- e. Benchmarking with rates from the literature

Iteration of steps *a* to *d* until we have reached understanding of why there may be differences.

5. **Information storage:** All these steps will be documented in an event specific report and stored for knowledge management, we will discuss with WP 3 how this should be made available and how best to deliver in D5.5

2.1.3.3 Analysis

To fingerprint the data sources in terms of events, we will run the following analyses:

- 1) Disease code/type of event distributions of the codes for each event (frequencies)
- 2) Age stratified incidence rates
- 3) Standardized rates (to WHO reference population)

This will be done using different types of definitions for the codes/lab values/text.


Requirement

Prior to starting the event fingerprinting, we need to progress on the mapping and definitions (Task 5.3) and optimize the available Jerboa /SAS modules for these steps. An ontology and knowledge management system needs to be created as part of task 5.5

2.2 Suitability fingerprinting

A second step in the fingerprinting exercise will be to assess the validity of the databases with respect to associations between vaccines and outcomes.

The suitability of databases to contribute to vaccine outcome studies will be tested by using selected 'true positive' and 'true negative' vaccine-event associations. The associations that will be run are

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	27/76

selected from the FP-7 funded GRIP project in which currently fourteen events are reviewed as being true-positively or negatively associated with any of the vaccines, agreement on the exact events and in the impact on safety reporting needs to be discussed with SC¹⁹. (see annex 2). Codes for events have also been mapped in the GRIP project.

Associations will be assessed preferably by using fully automated self-controlled case series approaches²⁰ which is one of the favourite designs proposed in D4.2. Additional designs for the fingerprinting may be added upon discussion with WP 4.

Requirements

Prior to starting this suitability fingerprint we need to progress both the vaccine fingerprinting and the event fingerprinting (Task 5.2). Programs need to be created and optimized in collaboration with WP 4.

Chapter 3. Databases in the consortium

Within the ADVANCE consortium we have included several partners that will make data under their custody available for the fingerprinting and proof of concept studies. Key characteristics of these databases as listed in the description of work are listed in table 1. The total estimated number of persons was 35 million.

Some modifications have occurred after project initiation, Associate partners have been added: FISABIO is participating with data from Valencia region and SIDIAP with data from Barcelona. Figure 3.1 shows the geographic distribution of partners with databases.

¹⁹ Yolanda Brauchli Pernu, Cassandra Nan, Thomas Verstraeten, Mariia Pedenko, Osemeke U. Osokogu, Daniel Weibel, Miriam Sturkenboom, Jan Bonhoeffer. Reference set for performance testing of paediatric vaccine safety signal detection methods and systems. *Submitted*

²⁰ Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* 1995; 51:228-235.


	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 28/7 6



Figure 3.1. Geographic distribution of partners with databases that will participate in fingerprinting and first phase of POC.




	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	29/7 6

Table 1: Characteristics of the databases of ADVANCE partners

Datasource	NL_IPCI	UK_THIN	UK_RCGP	ES_BIFAP	IT_ASLCR	IT_PEDIANET	DK-AUH	DK_SSI	SE_KI	FI_HPVCCHRT
Type of data	GP linked to RIVM vaccine registries	GP	GP	GP	Record linkage	Family pediatricians	Record linkage	Record linkage	Record linkage	RCT+ record linkage
Size	2 million	> 7 million	1 million	4 million	400,000	80 centers	1.8 million	5.5 million	9 million	5 million
Outpatient diagnoses	yes (ICPC)	yes (READ)	Yes (READ)	yes (ICPC+free text)	No	Yes	yes (ICD-10)	ICPC diagnosis for 70%	yes (ICD-10)	yes (ICD-10)
Inpatient diagnoses	yes (from letters/specialist)	yes (READ)	Yes (READ)	yes (partial)	yes (ICD9)	Yes	yes (ICD-10)	yes (ICD-10)	yes (ICD-10)	yes (ICD-10)
Signs/symptoms	Yes	yes (READ)	Yes (READ)	yes (ICPC+free text)	No	Yes	partial	Partial	No	no
Narratives	Yes	Yes	No	Yes	No	Yes	no	No	No	no
Blood tests	Yes	yes	No results	not systematically	No	Yes	yes	Yes	Yes	no
Microbiology	Yes	yes	Some results	not systematically	No	Yes	yes	Yes	Yes	no
Childhood vaccinations	Yes	yes	Yes (GP)	Yes	Yes	Yes	Sub pop.	Yes	Yes, whole pop. from 2013	yes
Influenza vaccinations	Yes	yes	Yes	Yes	No	No	Sub pop.	Yes	partial	yes
Travel vaccinations	No	not all	Yes	No	Yes	No	no	Partial	No	no
Elective vaccinations	No	not all	Yes	Not systematically	Yes	Yes	Partial	Partial	No	no
Professional vaccinations	No	not all	Yes (GP)	Not systematically	Yes	No	Partial	Partial	No	no
Prescribed/dispensed drugs	Yes	yes	Yes	yes	Yes	Yes	yes	Yes	Yes	yes

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	30/7 6

Datasource	NL_IPCI	UK_THIN	UK_RCGP	ES_BIFAP	IT_ASLCR	IT_PEDIANET	DK-AUH	DK_SSI	SE_KI	FI_HPVRTX
Ability to validate diagnoses against medical records	Yes	yes	Yes	yes	Yes	Yes	yes	Yes	Yes	yes
Ability to go back to patients	Yes	yes	No	yes	Yes	Yes	yes	Yes	yes	no
Lag time for updates	3 months	1 year	Weekly	1 year	4 months	Immediate	3 months	3 month currently but will be improved	1.5 years (county data faster)	annual
Approval requirements for access	GP approval board	ISAC	RCGP	by the AEMPS	only for clinical data	No	Notification to the Danish Data Protection Agency.	Notification to the Danish data protection agency.	Ethical approval	

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 31/76

Chapter 4. Results of population fingerprinting

4.1 Participating databases


Since May 2014, we have run three versions of the population fingerprint (to take out misunderstandings and an error in Jerboa). Table 3 tabulates the period of data, the number of patients and the cumulative amount of persontime per database

Table 3: Overview of population size in the databases (Nov 2014)

Country	Database	Type of database	Years covered	Total Number of persons	Person years of follow-up	Median follow-up (years)	25 th -75 th quartile follow-up (years)
UK	THIN	GP med. Record	1994 - 2013	8,326,238	51,333,610	4.91	1.8-10
	RCGP	GP med. Record	2003 – 2014*	2,043,800	12,839,908	6.08	2.25-10.7
NL	IPCI	GP med. Record	1996 – 2014*	1,786,405	5,468,684	2.83	1.5-4.3
	RIVM	Surveillance					
ES	BIFAP	GP med. Record	1998 – 2014*	4,800,538	25,680,609	4.91	2.3-8.2
IT	PEDIANET	GP med. Record	2004 – 2014*	77,021	350,797	4.25	1.8-7.1
	ASLCR	Provincial record linkage	2002 - 2013	454,188	4,226,316	11.90	6.7-11.9
SE	National (KI)	National record linkage	1998 – 2010	9,421,687	110,841,514	12.91	12.9-12.9
FI	Cohort	Trial/cohort					
DK	AUH	Regional record linkage	2004 – 2013	1,741,051\$	12,352,156\$	10	3.7-10
	National (SSI)	National record linkage	1996 – 2014*	7,512,032	103,492,835	18.66	8.3-18.7
Total				34,421,909	314,234,273		

*Data for 2014 not complete \$ are also part of DK_SSI

Table 3 shows that the databases that have currently supplied data capture a total of around 35 million subjects, with a total of more than 300 million person-years. Follow-up is very long in those systems that capture a regional or national population, whereas the GP databases have more dynamic populations, often because the number of participating practices change, or software is changing

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 32/76

4.2 Distribution of person-time over calendar-year

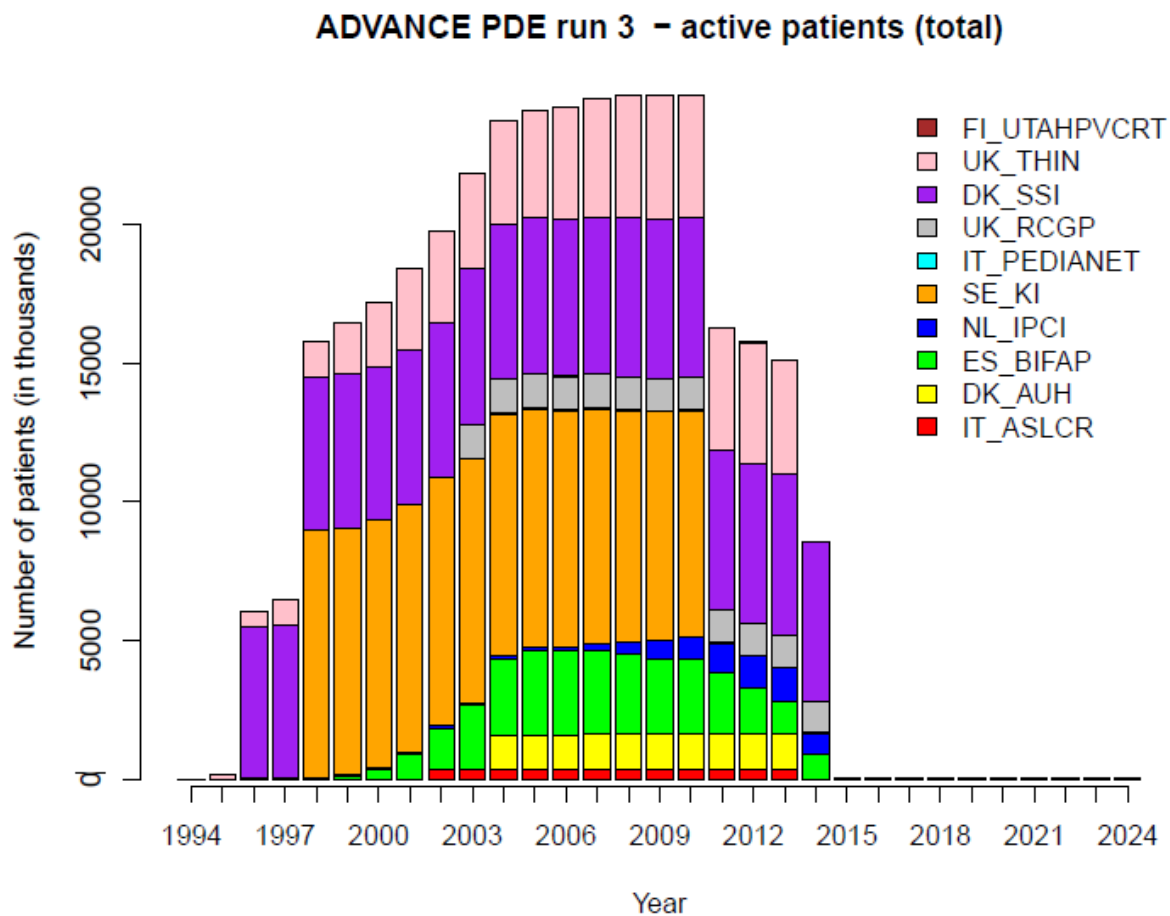



Figure 4.2.1.: Distribution of active patients over calendar time (active means: contributed at least one day of follow-up in that year)

This graph shows different patterns for the databases that submitted their fingerprinting data:

- 1) The SSI (DK) and THIN (UK) data cover the entire period from 1996-2013/14
- 2) The Swedish data (KI) are substantial but stop in 2010
- 3) GP databases: BIFAP(ES), IPCI(NL), RCGP (UK), and THIN (UK) seem to have shorter time between data delivery and last data supply (all contributing to data in 2013/2014)
- 4) GP databases tend to have more dynamic population time than the regional/national databases (i.e. SSI, KI, ASLCR, AUH)

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	
		Version: v1.4 – final
		Security: 33/76

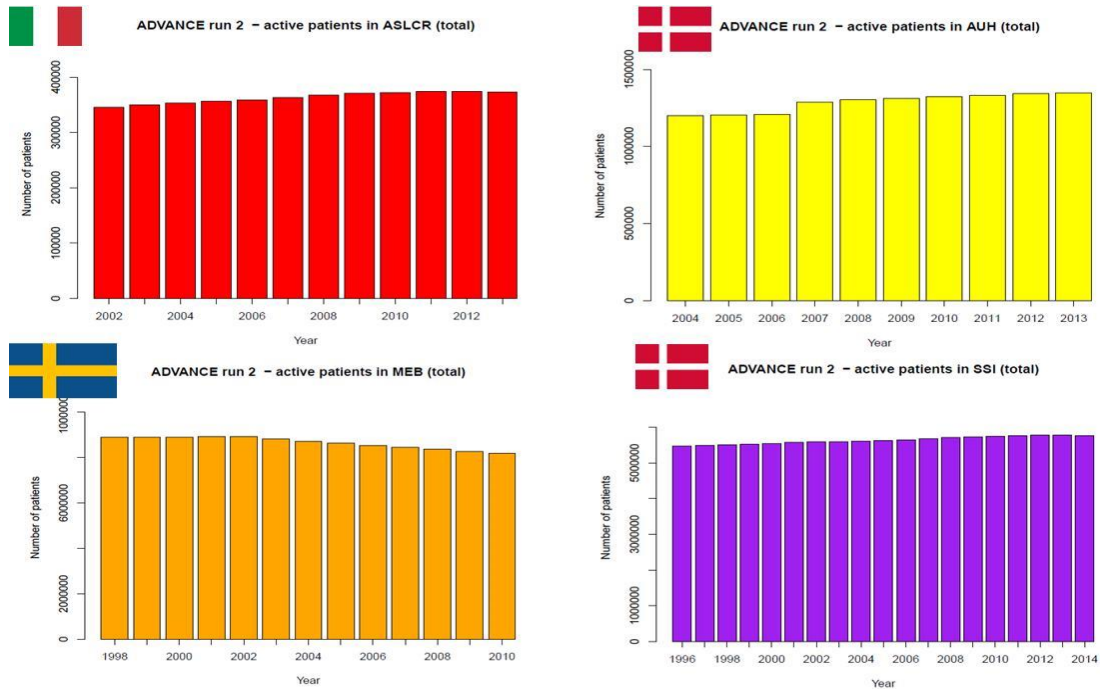


Figure 4.2.2. Distribution of active persons over calendar-years for regional/ national record linkage databases (note: Y-axes differ in scale)

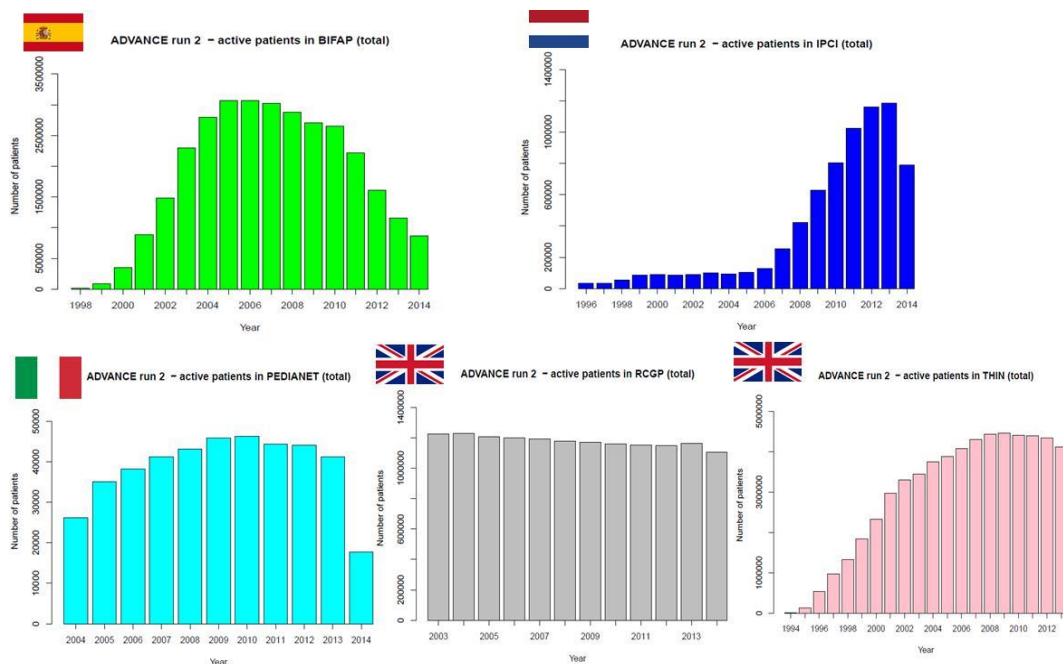



Figure 4.2.3. Distribution of persons over calendar-years for GP databases (note Y-axes differ in scale)

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 34/7 6

The HPV cohort from Finland is different, it started with a trial and is converted into a cohort that has a fixed follow-up until 2024.

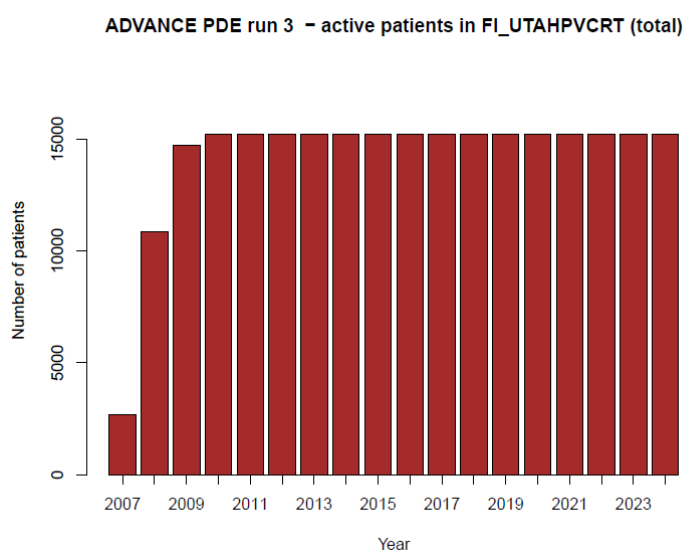


Figure 4.2.4. Distribution of persons over calendar-years for HPV cohort in Finland (note Y-axes differ in scale)

4.3 Population pyramids and representativeness


4.3.1 Denmark: Aarhus database (DK_AUH)

Description

The subset of The Primary Health Care Database available at AUH comprises data on the population of former North-Jutland, Aarhus, Ringkjøbing and Viborg counties, which since 2007 are called the Central Denmark Region and the North Denmark Region. This population covers a total of 1.8 million inhabitants and is representative of the population of Denmark.²¹

Data available on these subjects include information on services provided by general practitioners, dentists and other specialists. Influence of erroneous registrations, misclassifications and possible fraud is considered to be extremely small, and the registry is known to have a high internal validity. The registry contains further information on the patient (civil registration number, age, gender, region of residence, etc.), on the health care provider, on the health service (service code and service description, time of service, size of fee covered by the National Health Insurance) and on the dates (week numbers and the year of service). The categorization of services by service codes is extremely detailed, for example the different types of vaccinations have all their own unique service code.

These data can be linked to the national registry of patients that comprises information on admissions to Danish somatic hospitals, emergency rooms and outpatient clinics with diagnosis

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 35/76

codes and procedures registered. These databases have been used in numerous studies and are proven valid for pharmacoepidemiological research.

Population fingerprint

Aarhus provided data from 2004-2014 and has a very stable population, the cumulative amount was 1.7 million, the median follow-up is 10 years. The Danish population comprises around 5.6 million persons.

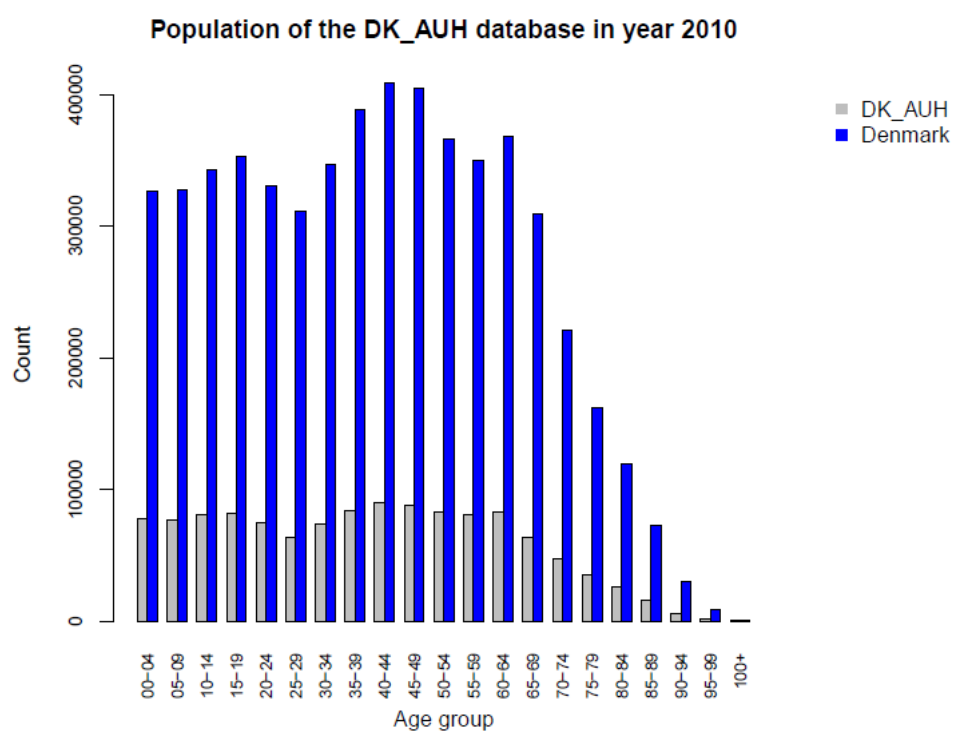



Figure 4.3.1.1. Distribution of Aarhus database (ADVANCE fingerprint run) and DK population by age (DK reference from United Nations)²¹ in absolute numbers

²¹ <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases			
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghie, Daniel Weibel, Germano Ferreira		Security:	36/76

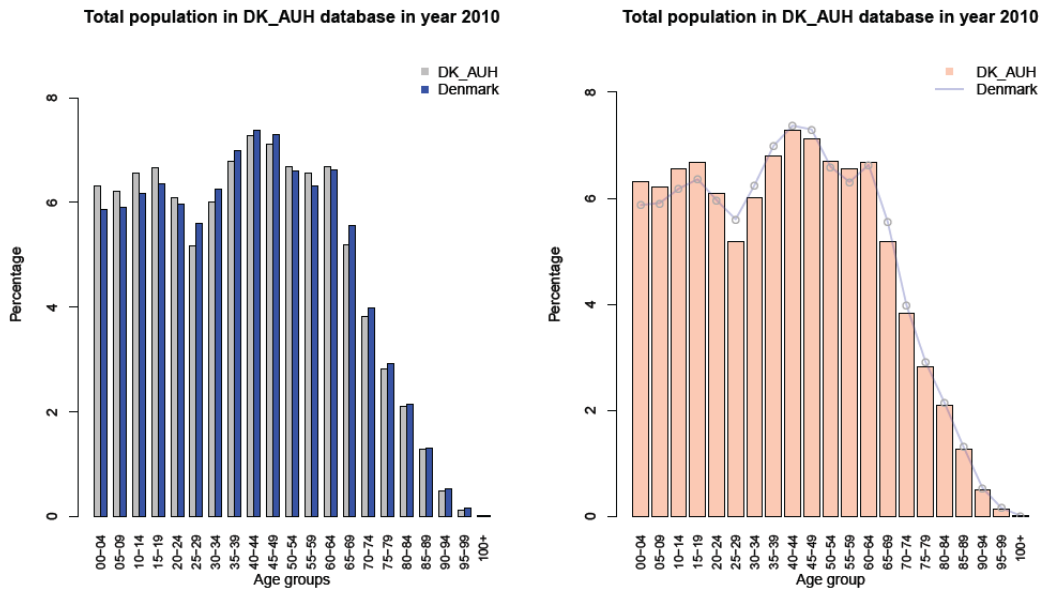


Figure 4.3.1.2. Relative distributions of the age of the population from Aarhus database (ADVANCE fingerprint run) and the Danish population (United Nations) in 2010

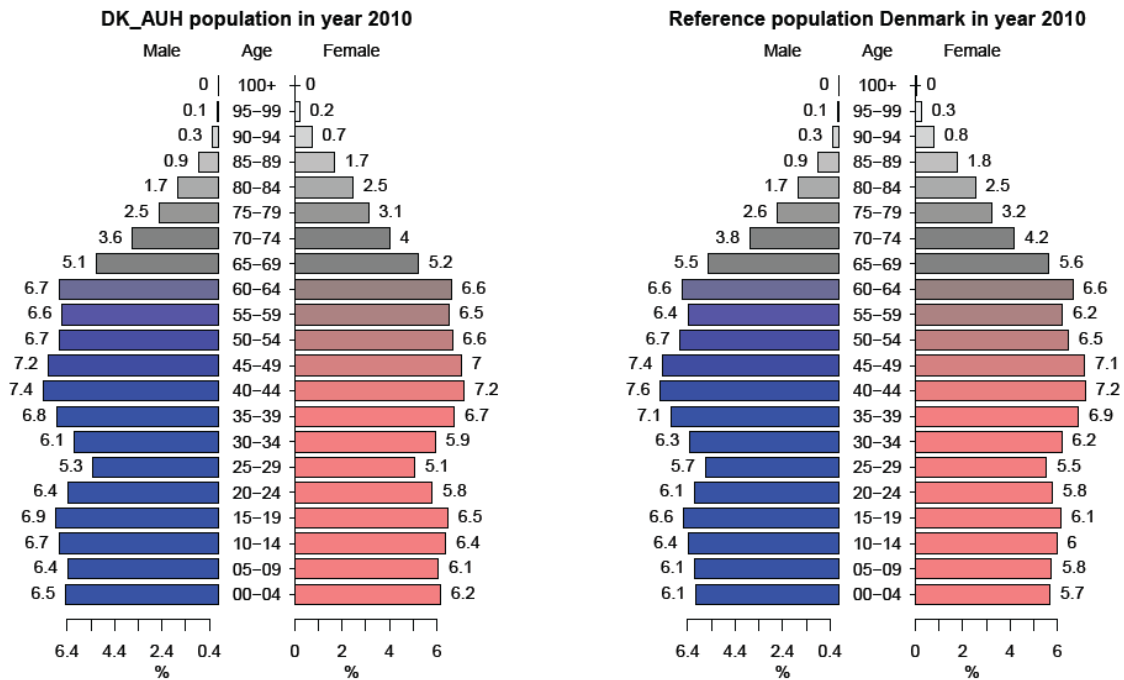



Figure 4.3.1.3. Distribution of Aarhus database (ADVANCE fingerprint run) and DK population by age and sex (DK reference from United Nations) in percentages

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	37/7 6

These data show

- 1) Aarhus database captures a little more than a quarter of the Danish population
- 2) The age and gender distribution are representative for Denmark
- 3) The long follow-up and stable population will be good for studies in birth cohorts and where there is need for long follow-up

4.3.2 Denmark: The Danish Civil and Health Registration System (DK_SSI)


Description

The Danish Civil and Health Registration System database is created ad hoc by linkage between the civil registration system, the vaccination registry, the patient registry plus other relevant databases (e.g., disease surveillance, medications, microbiology, pathology and so on).

The data are kept at Research Services at the SSI from where access can be granted, provided that the analysis has been reported and approved by the Danish Data Protection Agency. There is a fee for access. The vaccination registry includes personal ID for linkage. Also the database is not a single database, but that data are assembled and merged for specific studies according to the relevant Scientific questions.

Population fingerprint

SSI provided data from 1996-2014 and has a very stable population, the cumulative amount of persons was 7.5 million, the median follow-up is 18 years. The Danish population comprises around 5.6 million persons in a year.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 38/7 6

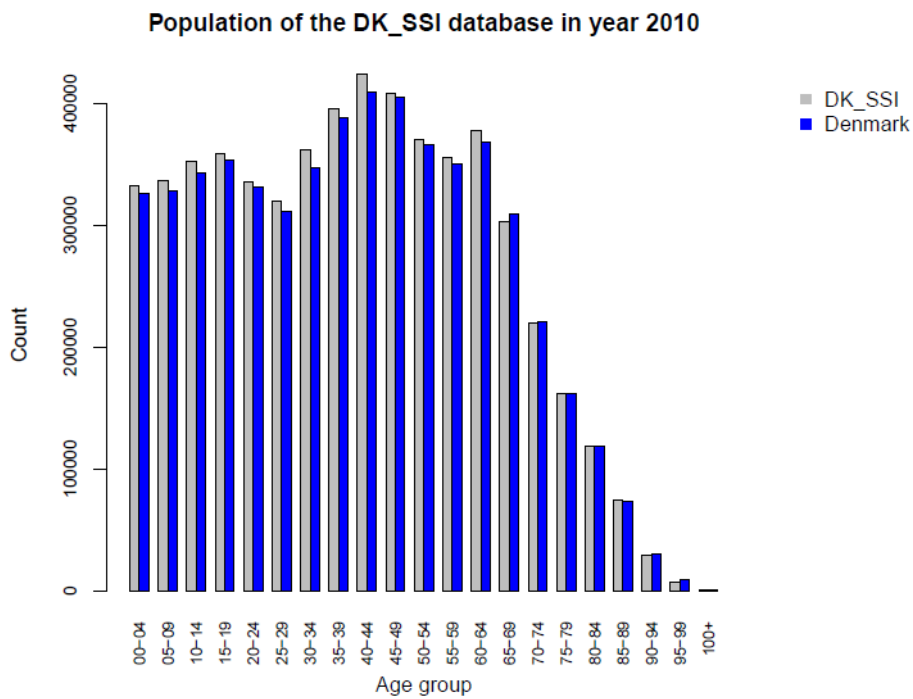



Figure 4.3.2.1. Distribution of Danish Civil and Health Registration System database (ADVANCE fingerprint run) and DK population by age (DK reference from United Nations)²² in absolute numbers.

²² <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	
	Version: v1.4 – final	
	Security:	39/76

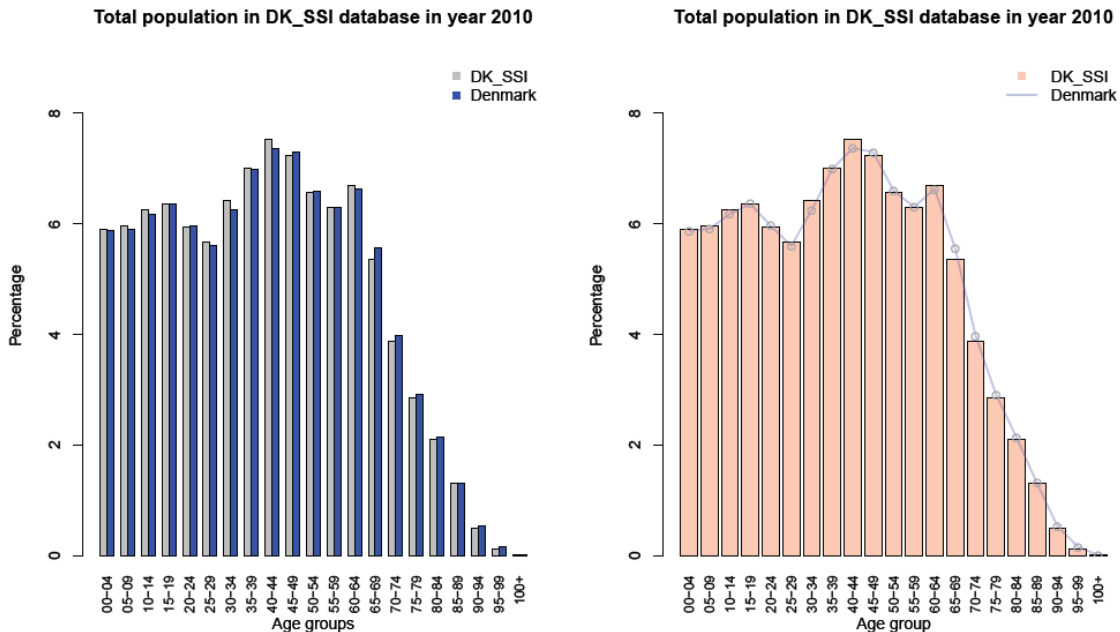


Figure 4.3.2.2. Relative distributions of the age of the population from Danish Civil and Health Registration System database (ADVANCE fingerprint run) and the Danish population (United Nations) in 2010.

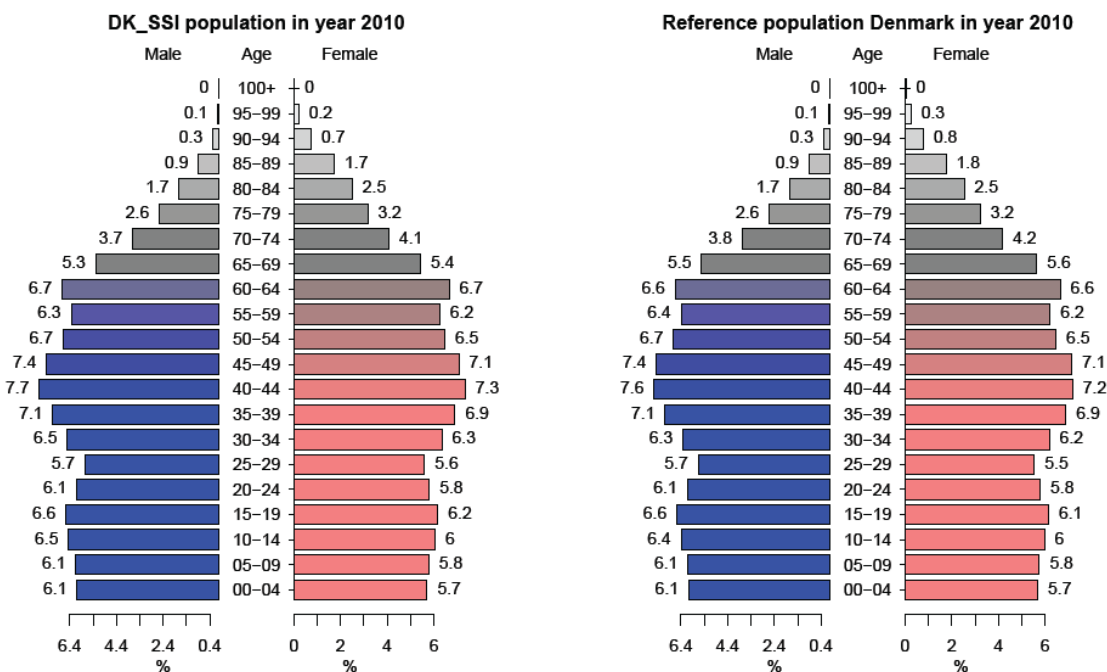



Figure 4.3.2.3. Distribution of Danish Civil and Health Registration database (ADVANCE fingerprint run) and DK population by age and sex (DK reference from United Nations) in percentages.

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	40/7 6

These data show:

- 1) The Danish Civil and Health Registration database captures the entire Danish population
- 2) Therefore the age and gender distribution are representative of Denmark
- 3) The long follow-up and stable population will be good for studies in birth cohorts and where there is need for long follow-up


4.3.3 Sweden: population based-registers (SE-KI)

Description

In Sweden there is not one single database that can be used for studies within ADVANCE. For each study the data will be requested and retrieved from relevant data sources i.e. linkage between the Total population register (Statistics Sweden), the vaccination registry (The public health agency), the patient registers (National Board of Health) plus other data sources that will be needed to answer specific research question (e.g., disease surveillance, drug prescription, microbiology and others). Each study will need an ethical approval. Each data owner charges a fee to handle the request and retrieve data. Linkage will be done using a unique personal identifier (PIN). Once linked the PINs will be substituted with unique study numbers instead. If asked for and motivated in the ethical application, the data holder that link all the data can keep a code key for approx. 3-5 years. This way follow-up studies can be done without starting the whole process from the beginning. The data requested will be stored at MEB, KI and can be used for other studies if ethical approval exist i.e. data can be re-used.

Population fingerprint

Karolinska Institutet provided data from 1998-2010 for this fingerprint (for POC studies this will be updated) and has a very stable population, the cumulative amount was 9.4 million, the median follow-up is 13 years. The Swedish population comprises around 9.7 million persons .

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 41/76

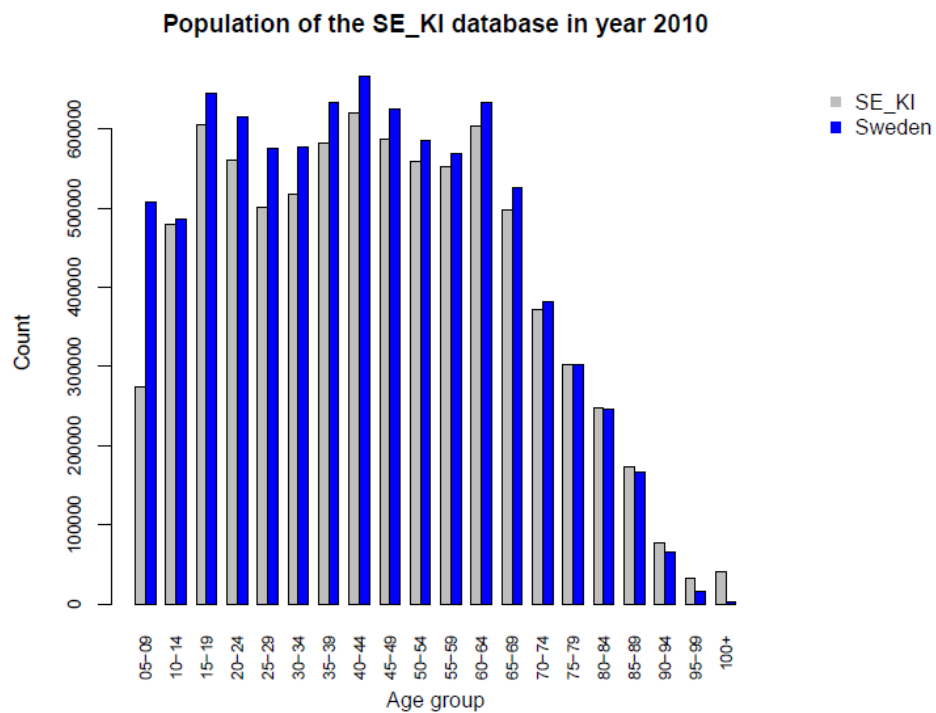



Figure 4.3.3.1. Distribution of Swedish database (ADVANCE fingerprint run) and SE population by age (SE reference from United Nations)²³ in absolute numbers.

²³ <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases			
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghie, Daniel Weibel, Germano Ferreira		Security:	42/7 6

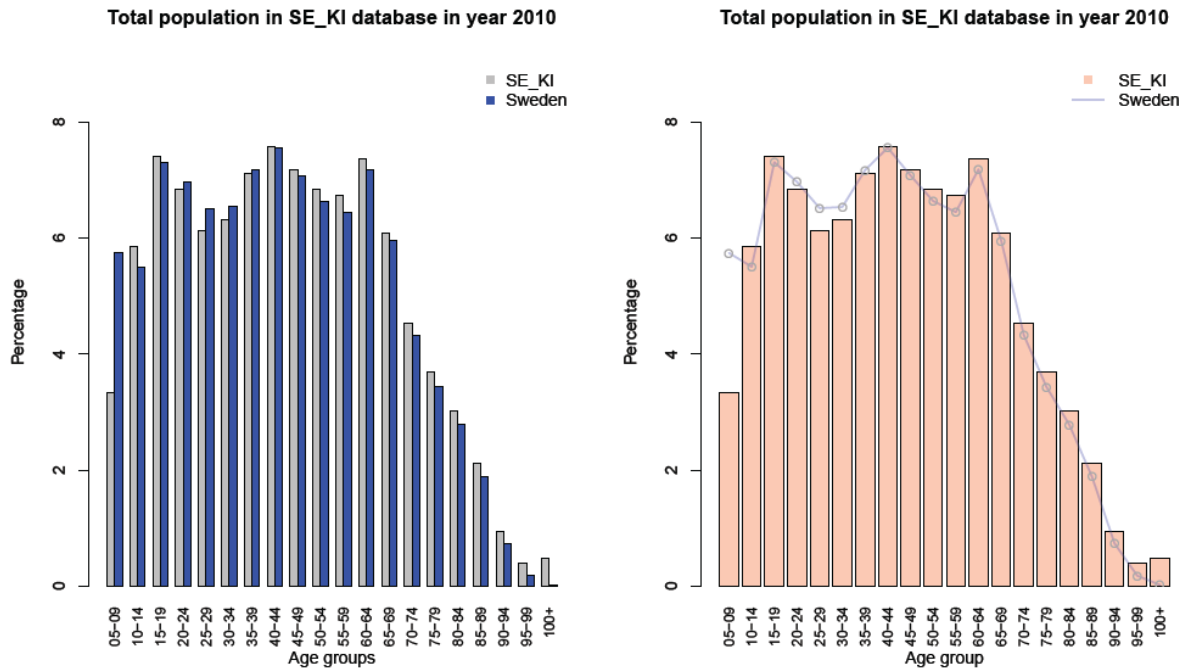


Figure 4.3.3.2. Relative distributions of the age of the population from Swedish database (KI)(ADVANCE fingerprint run) and the Swedish population (United Nations) in 2010.

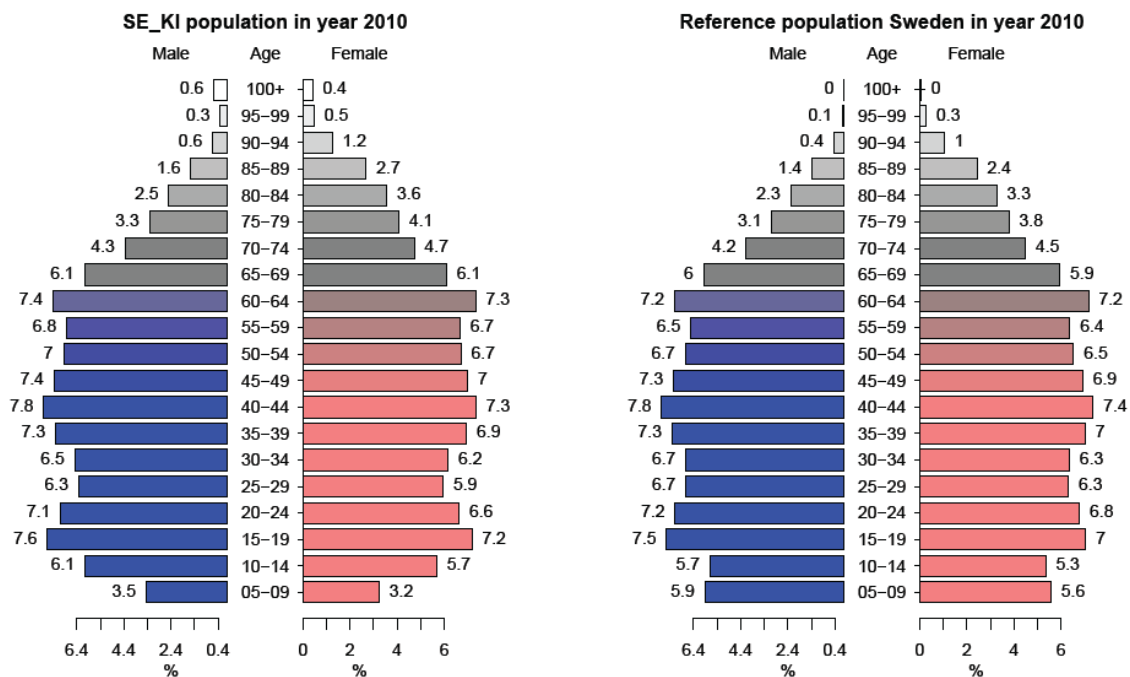



Figure 4.3.3.3: Distribution of Swedish (KI) database (ADVANCE fingerprint run) and SE population by age and sex (SE reference from United Nations) in percentages

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

These data show

- 1) The data on Sweden as submitted by KI capture almost the entire Swedish population
- 2) Therefore the age and gender distribution are representative of Sweden
- 3) The long follow-up and stable population will be good for studies in birth cohorts and where there is need for long follow-up

4.3.4 UK: THIN database (UK_THIN)

Description

The Health Improvement Network (THIN) is a database of primary care medical records from the UK. General practitioners are trained to complete their medical records using the Vision general practice computer system (InPractice Systems, London, UK). This electronic record serves as the primary medical records for the practice.


Data recorded in THIN include demographics, details from GPs' visits such as medical diagnoses and prescriptions written by the GPs, diagnoses from specialist referrals and hospital admissions, some results of laboratory tests, some lifestyle characteristics and other measurements as taken in the practice. Within the database, diagnoses are recorded using READ codes. Prospective data collection for THIN began in September 2002, with electronic medical records that date back to 1985. In addition, practices may retrospectively enter significant medical events into the electronic medical record. The database has around 2.7 million active patients registered. Recently a validation study was conducted by Lewis et al (2007) which concluded that "THIN data that are collected outside of the Clinical Practice Research Datalink (CPRD) appear as valid as the data collected as part of the CPRD"²⁴.

As the primary aim of the collection of data in the THIN database is patient management, data will reflect only those events that are deemed to be relevant to patient's care. In addition, use of THIN data is not appropriate in studies where individual ethnicity, occupation, employment, and/or socio-economic status are important variables. As for all prescription databases, over-the-counter (OTC) drug use and non-compliance to medication prescriptions might be an issue. As the average follow-up within the THIN database is 5 years, the THIN database is not suitable to conduct long-term follow-up studies. Approval needs to be obtained for each study from the THIN governance board.

Population fingerprint

THIN data were provided by EMC under the Erasmus academic license. It captures data from 1994-2013 (including the non-up to standard time) and has a growing population, the cumulative amount was 8.3 million persons, the median follow-up is 5 years. The UK population comprises around 63 million persons and THIN only represents a small proportion.

²⁴ Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf.* 2007;16(4):393-401.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 44/7 6

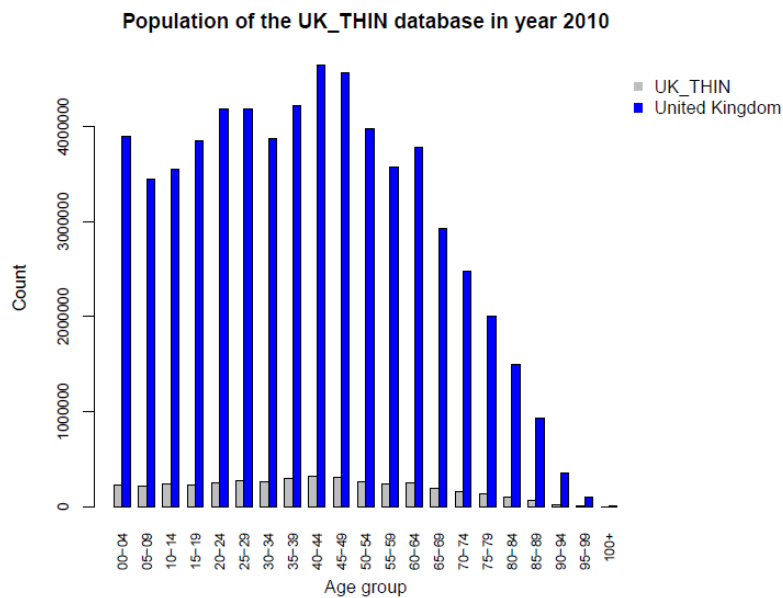



Figure 4.3.4.1. Distribution of the THIN database (ADVANCE fingerprint run) and UK population by age (UK reference from United Nations)²⁵ in absolute numbers (2010).

²⁵ <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghie, Daniel Weibel, Germano Ferreira	
	Version: v1.4 – final	Security: 45/76

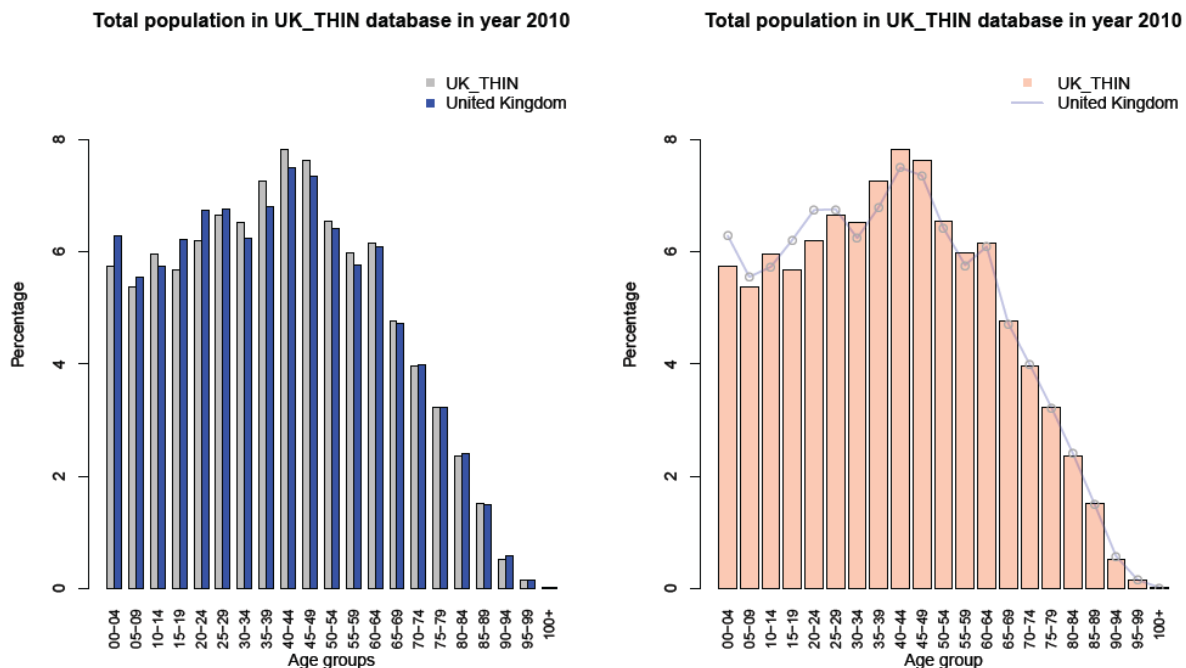


Figure 4.3.4.2. Relative distributions of the age of the population from THIN database (ADVANCE fingerprint run) and the UK population (United Nations) in 2010.

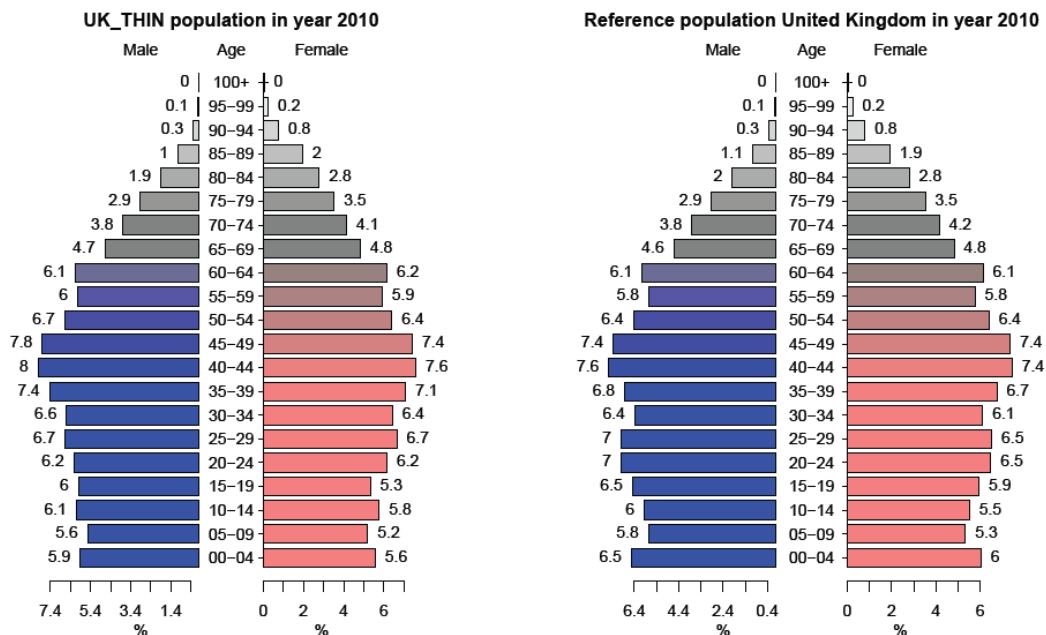



Figure 4.3.4.3. Distribution of the THIN database (ADVANCE fingerprint run) and UK population by age and sex (UK reference from United Nations) in percentages.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 46/76

These data show

- 1) The THIN database captures only a small proportion of the UK population (around 5%)
- 2) The age and gender distribution are representative for UK, in the younger age categories a slight underrepresentation
- 3) The short follow-up will be difficult for studies in birth cohorts and where there is need for long follow-up

4.3.5 UK: RCGP Research and Surveillance Centre (UK_RCGP)

Description

The RCGP Research and Surveillance Centre’s (RSC) extensive network of spotter practices currently extracts data from over 100 practices throughout England and Wales. The aim of the surveillance scheme is to provide a timely picture of consultations by diagnosis with sentinel GPs in England and Wales. Pseudonymised patient-linked data extending back to 2003. The data custodian is Professor Simon de Lusignan who is contracted to the RCGP via University of Surrey.

Population fingerprint

RCGP submitted data from 2003-2014. The population size is very stable over time as we saw above. The cumulative amount was 2 million persons, and the median follow-up is 6 years. The UK population comprises around 63 million persons and RCGP only represents a small proportion.

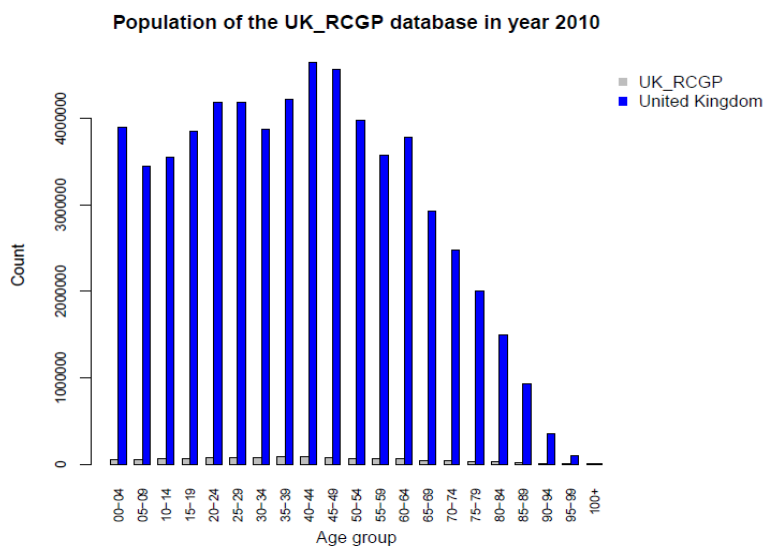



Figure 4.3.5.1. Distribution of the RCGP database (ADVANCE fingerprint run) and UK population by age (UK reference from United Nations)²⁶ in absolute numbers (2010)

²⁶ <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	
	Version: v1.4 – final	
	Security:	47/76

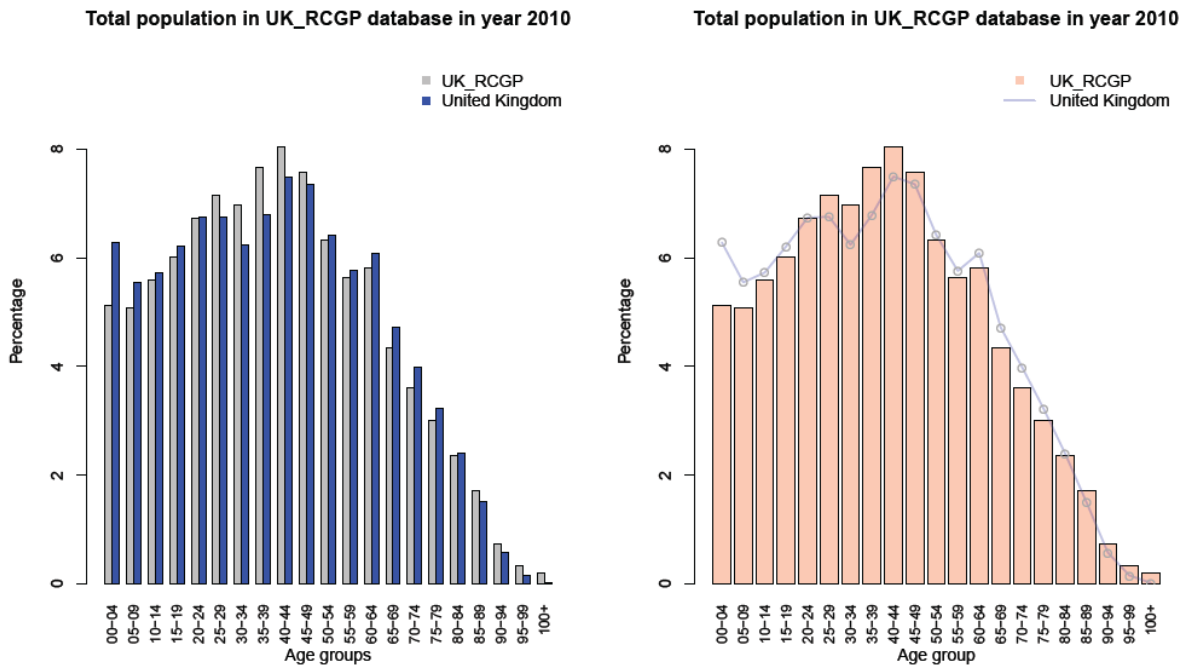


Figure 4.3.5.2. Relative distributions of the age of the population from RCGP database (ADVANCE fingerprint run) and the UK population (United Nations) in 2010.

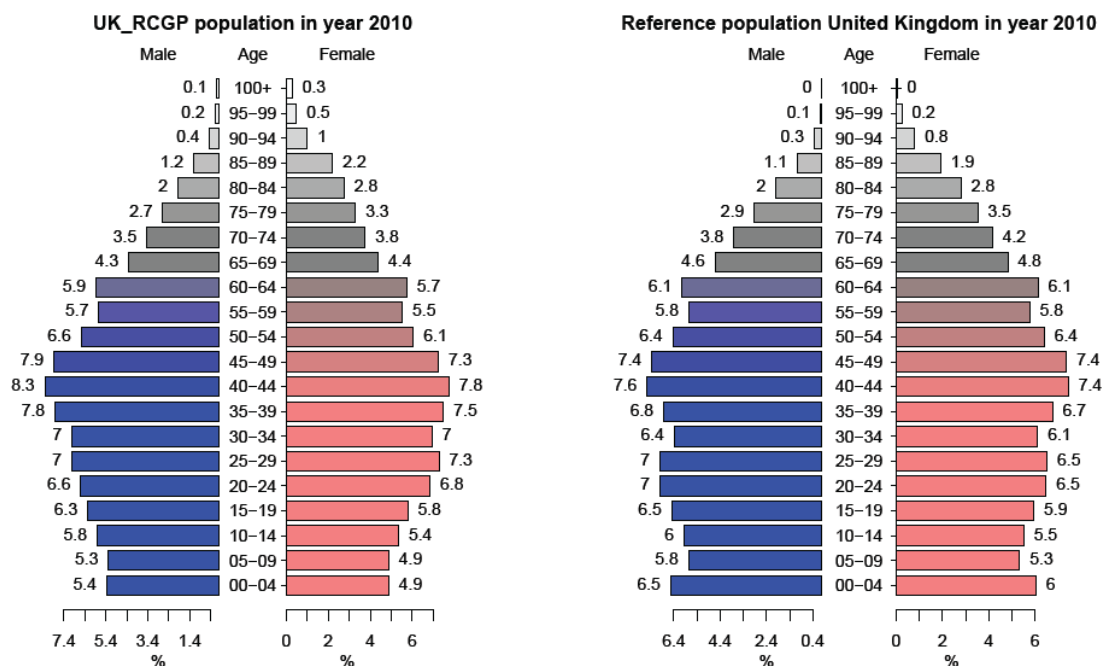



Figure 4.3.5.3: Distribution of the RCGP database (ADVANCE fingerprint run) and UK population by age and sex (UK reference from United Nations) in percentages

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	48/76

These data show:

- 1) The RCGP database captures only a small proportion of the UK population (around 3.5%)
- 2) The age and gender distribution are representative for UK, although in the younger age categories there is an underrepresentation
- 3) The short follow-up will be difficult for studies in birth cohorts and where there is need for long follow-up

4.3.6 NL: Integrated Primary Care Information (NL_IPCI)

Description


In 1992 the Integrated Primary Care Information Project (IPCI) was started by the Department of Medical Informatics of the Erasmus University Medical Centre. IPCI is a longitudinal observational database that contains data from computer-based patient records of a selected group of general practitioners (GPs) throughout The Netherlands, who voluntarily chose to supply data to the database. Collaborating practices are located throughout The Netherlands and the collaborating GPs are comparable to other GPs in the country according to age and gender. In the Netherlands, all citizens are registered with a GP practice, which forms the point of care and acts as a gatekeeper in a two-way exchange of information with secondary care. The medical records of each patient can therefore be assumed to contain all relevant medical information including medical findings and diagnosis from secondary care.

The International Classification of Primary Care (ICPC) is the coding system for patient complaints and diagnoses, but diagnoses and complaints can also be entered as free text. Prescription data such as product name, quantity prescribed, dosage regimens, strength and indication are entered into the computer²⁷. The National Database of Drugs, maintained by the Royal Dutch Association for the Advancement of Pharmacy, enables the coding of prescriptions, according to the ATC classification scheme recommended by the World Health Organization. Approval needs to be obtained for each study from the IPCI governance board.

Population fingerprint

IPCI submitted data from 1996-2014. The population size is very dynamic over time as we saw above, from 2007 onwards there was a substantial growth in size since data from other information systems was added. The cumulative amount was 1.8 million persons, but the median follow-up is only 2.8 years. The NL population comprises around 16 million persons and IPCI only represents a small proportion.

²⁷ Vlug AE, van der Lei J, Mosseveld BM, van Wijk MA, van der Linden PD, Sturkenboom MC, et al. Postmarketing surveillance based on electronic patient records: the IPCI project. *Methods of information in medicine.* 1999;38(4-5):339-44.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 49/7 6

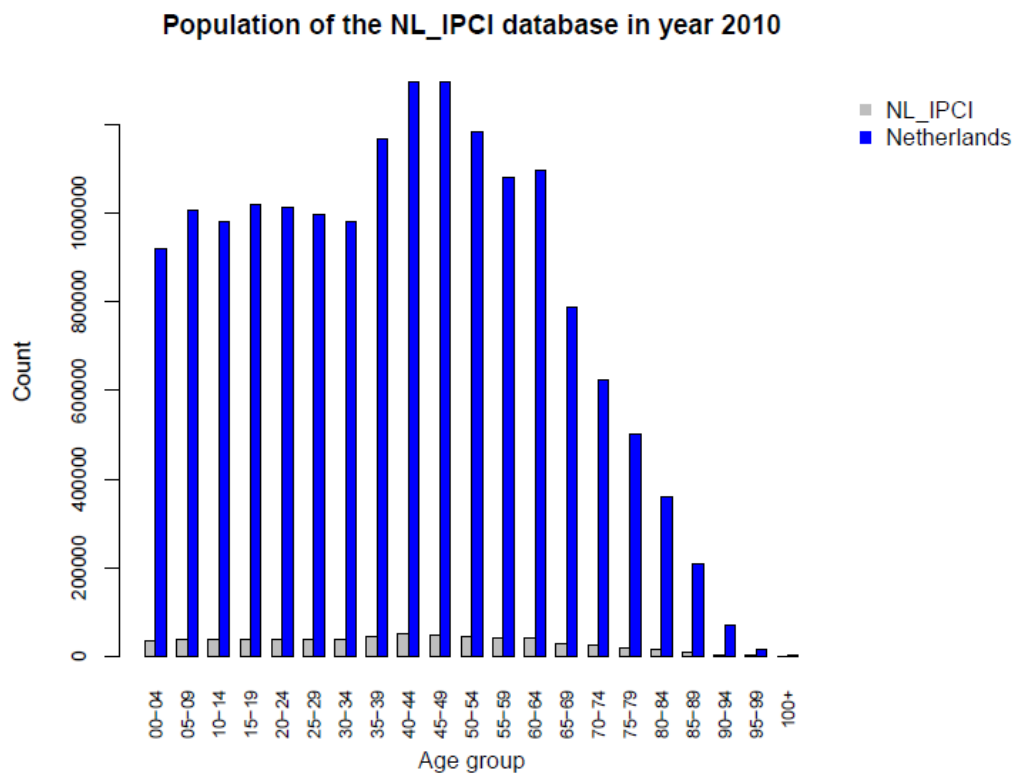



Figure 4.3.6.1. Distribution of the IPCI database (ADVANCE fingerprint run) and NL population by age (NL reference from United Nations)²⁸ in absolute numbers (2010).

²⁸ <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases			
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghie, Daniel Weibel, Germano Ferreira		Security:	50/76

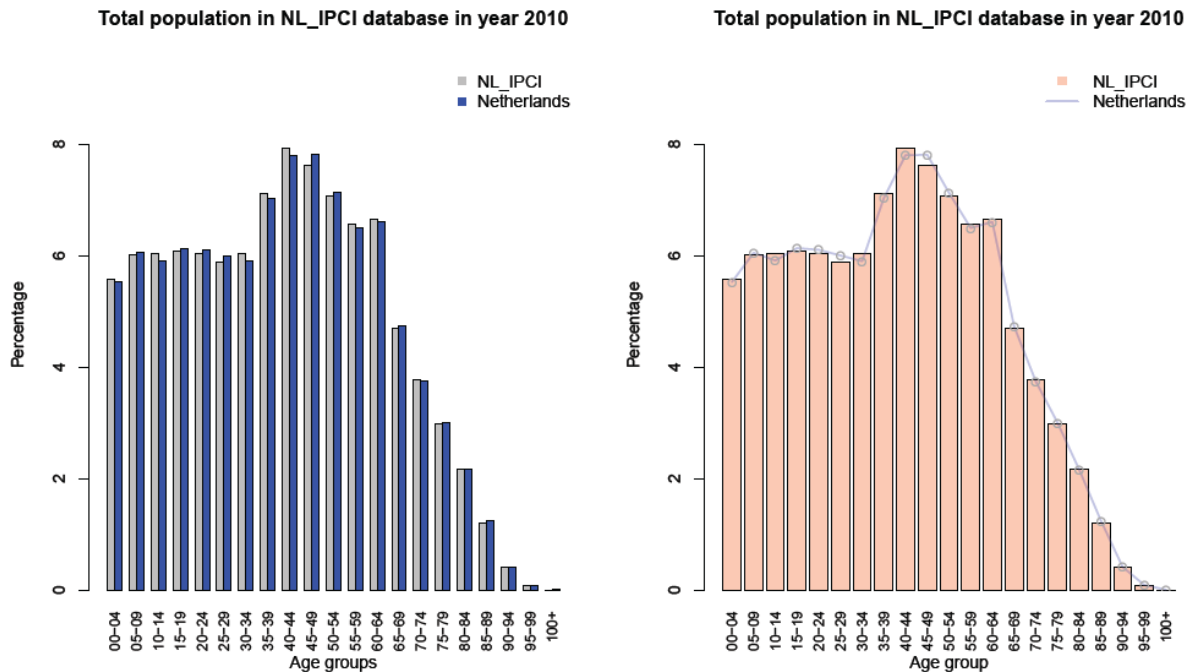


Figure 4.3.6.2. Relative distributions of the age of the population from IPCI database (ADVANCE fingerprint run) and the NL population (United Nations) in 2010.

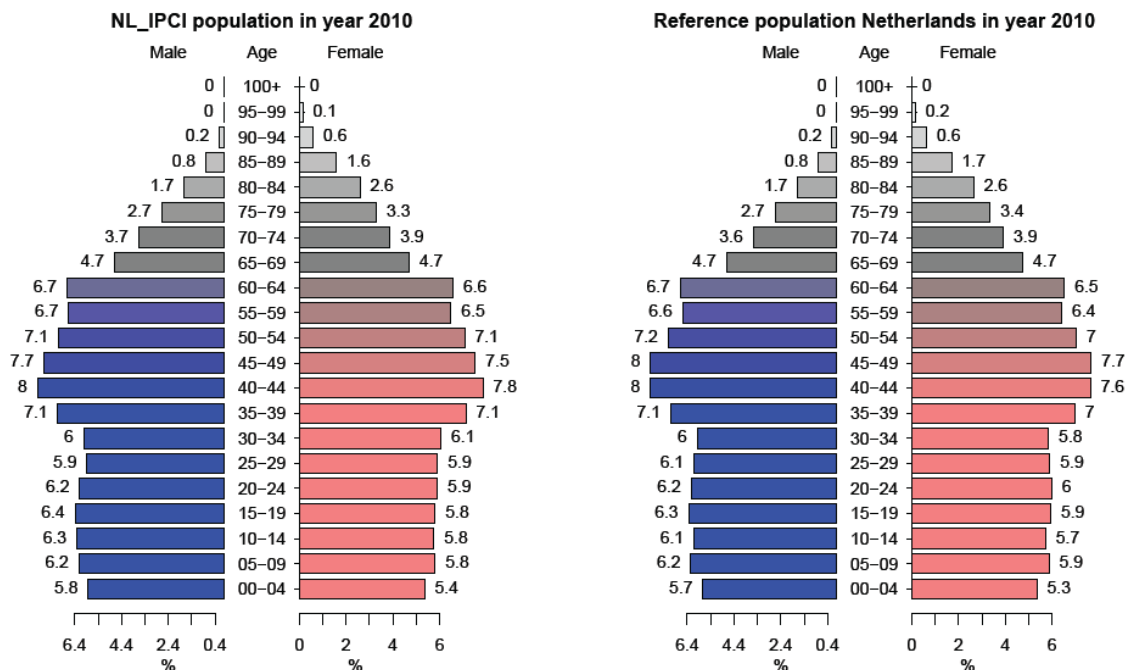



Figure 4.3.6.3. Distribution of the IPCI database (ADVANCE fingerprint run) and NL population by age and sex (NL reference from United Nations) in percentages.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 51/76

These data show:

- 1) The IPCI database captures only a small proportion of the NL population (around 3%)
- 2) The age and gender distribution are representative for NL
- 3) The short follow-up will be difficult for studies in birth cohorts and where there is need for long follow-up

4.3.7 ES: Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria (ES_BIFAP)

Description

BIFAP (Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria) (<http://www.bifap.org/>) is a computerised database of medical records of Primary Care operated by the Spanish Medicines Agency (AEMPS) as a non-profit research project. Currently, the database includes 2,239 physicians (including general practitioners and paediatricians) from 9 different regions in Spain. These regions collaborate with BIFAP and send their data to BIFAP every year. BIFAP database includes clinical and prescription data from around 4,5 million patients covering around 8.6% of the Spanish population. The database is restricted to research for regulatory purposes and to independent research by healthcare professionals of the local health authorities and the national health system.

Population fingerprint

BIFAP submitted data from 1998-2014. The population size is dynamic over time as we saw above, from 2004 onwards there was a substantial growth in size. The cumulative amount was 4.8 million persons, but the median follow-up is 5 years. The Spanish population comprises around 47 million persons and BIFAP only represents a small proportion.

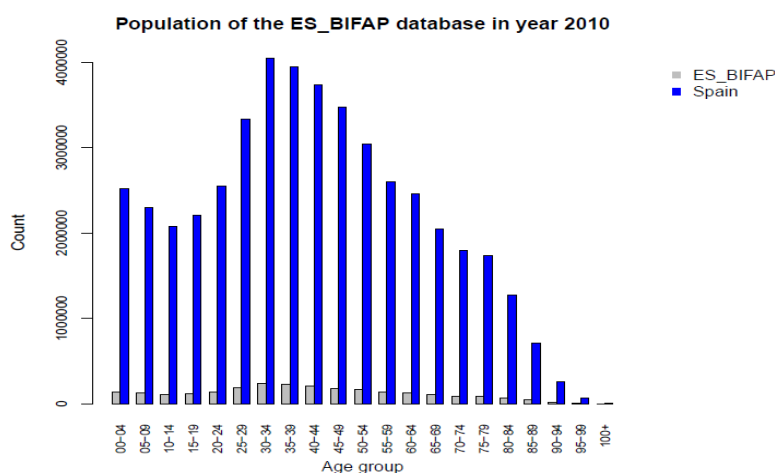



Figure 4.3.7.1. Distribution of the BIFAP database (ADVANCE fingerprint run) and ES population by age (ES reference from United Nations)²⁹ in absolute numbers (2010).

²⁹ <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	
	Version: v1.4 – final	
	Security:	52/76

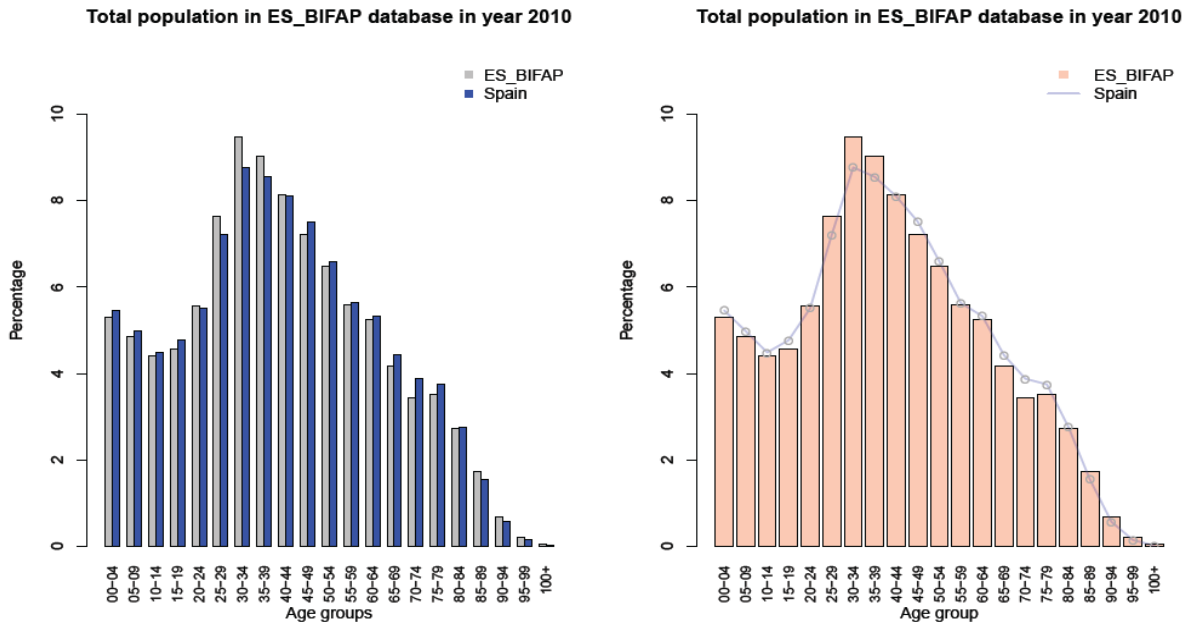


Figure 4.3.7.2. Relative distributions of the age of the population from BIFAP database (ADVANCE fingerprint run) and the ES population (United Nations) in 2010.

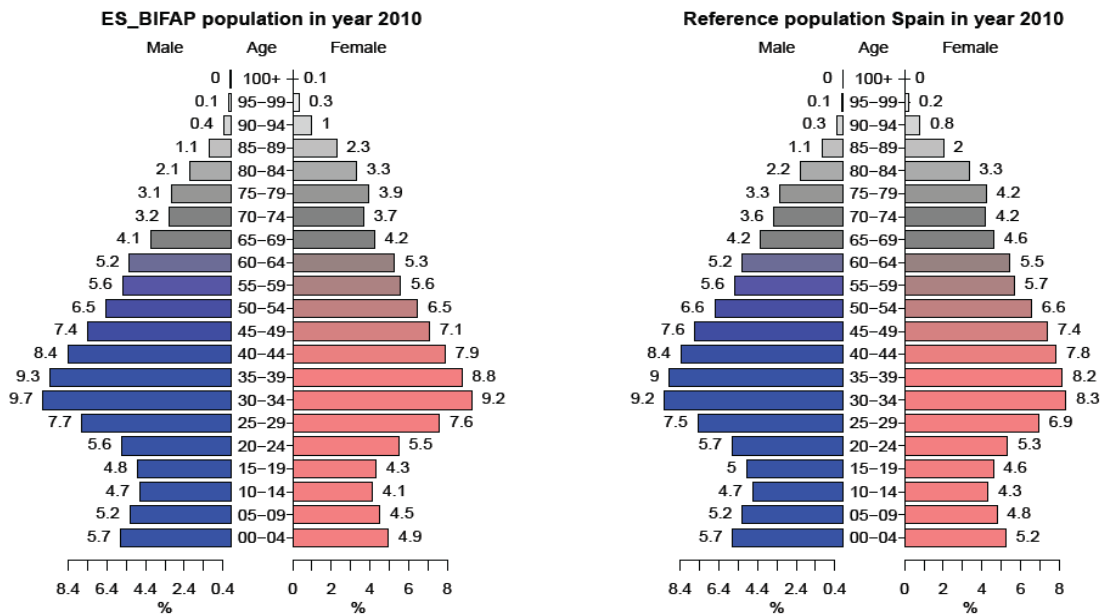



Figure 4.3.7.3. Distribution of the BIFAP database (ADVANCE fingerprint run) and ES population by age and sex (ES reference from United Nations) in percentages.

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	53/7 6

These data show

- 1) The BIFAP database captures only a small proportion of the Spanish population (around 8%)
- 2) The age and gender distribution are representative for Spain, except for the fourth decade of life where there is slight overrepresentation
- 3) The short follow-up will be difficult for studies in birth cohorts and where there is need for long follow-up

4.3.8 IT: ASL della provincia di Cremona (IT_ASLCR)


Description

ASLCR is a record linkage database. It contains all the mortality data (with cause of death), hospitalizations (with diagnosis), outpatient visits, drug prescriptions of the citizens. Moreover, it contains the registry of all the vaccinations administered by (or notified to) the Local Health Authority and the registry of infectious diseases to be notified by law.

The Local Health Authority (ASL) of Cremona is the institution in charge of the health of the citizens living in the province of Cremona. The ASL is responsible for the provision of all health-related services (prevention, treatment, residential care, etc.).

Population fingerprint

ASLCR submitted data from 2002-2013. The population size is very stable over time as we saw above. The cumulative amount was 450 thousand persons, but the median follow-up is 12 years. The IT population comprises around 61 million persons and ASLCR only represents a small proportion from the North of Italy.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 54/7 6

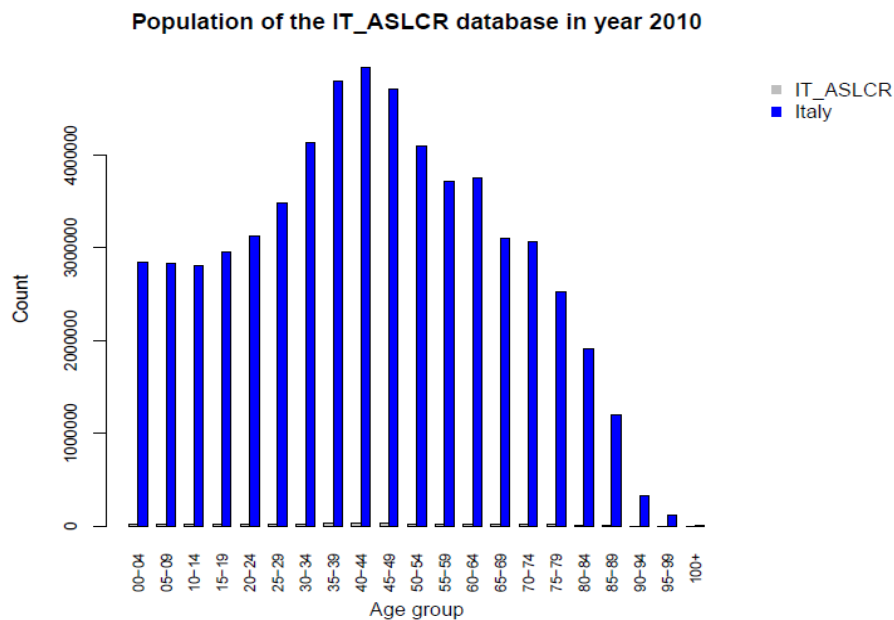


Figure 4.3.8.1. Distribution of the ASLCR database (ADVANCE fingerprint run) and IT population by age (IT reference from United Nations)³⁰ in absolute numbers (2010).

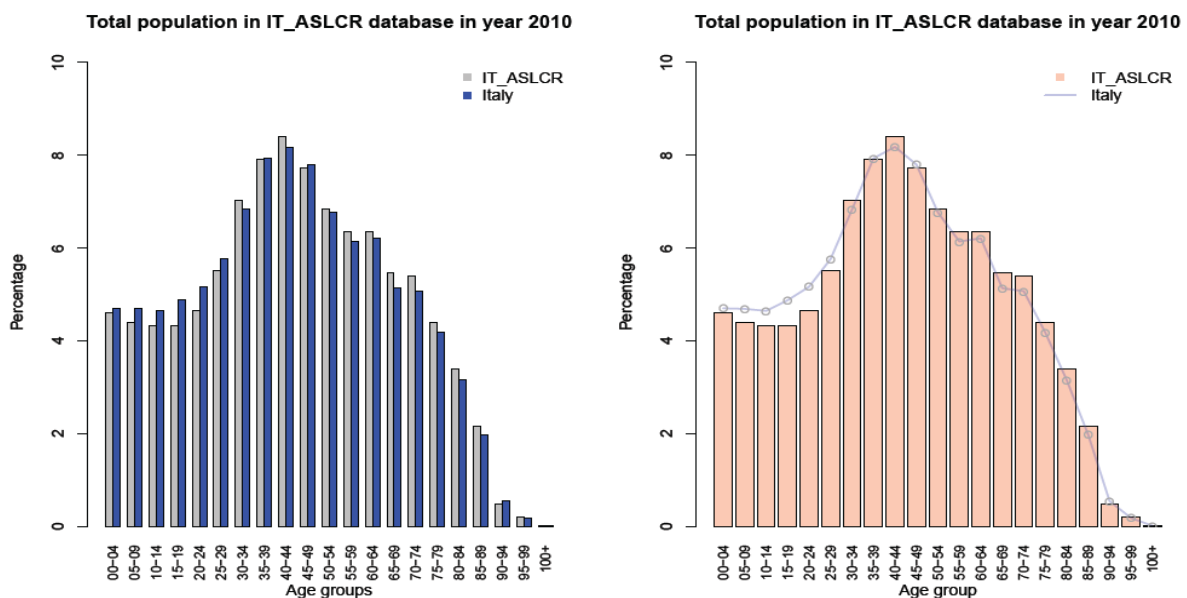



Figure 4.3.8.2. Relative distributions of the age of the population from ASLCR database (ADVANCE fingerprint run) and the IT population (United Nations) in 2010.

³⁰ <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 55/76

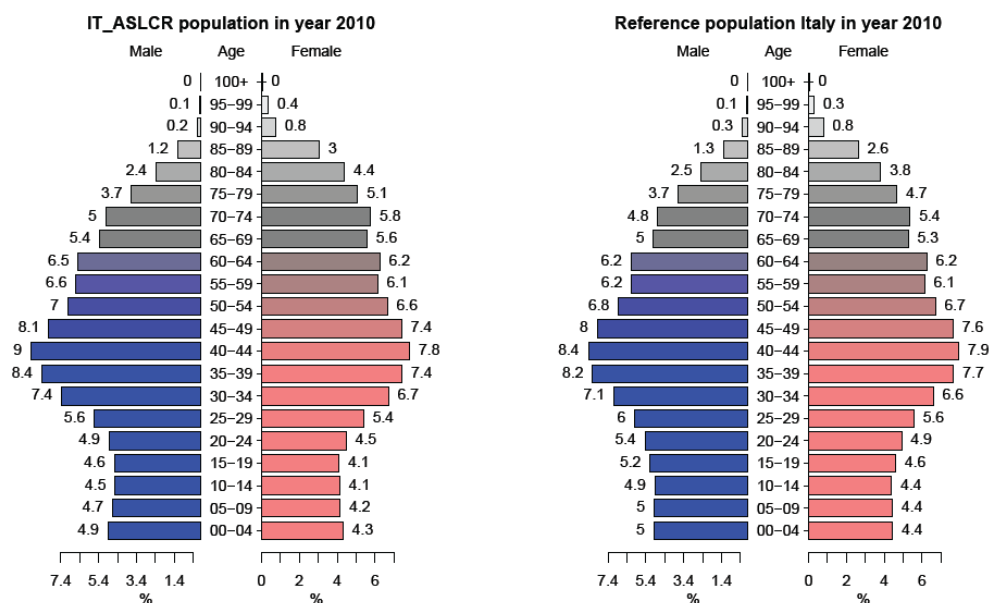


Figure 4.3.8.3. Distribution of the ASLCR database (ADVANCE fingerprint run) and IT population by age and sex (IT reference from United Nations) in percentages.


These data show

- 1) The ASLCR database captures only a small proportion of the IT population (less than 1%)
- 2) The age and gender distribution are roughly representative for IT, the younger ages are underrepresented, which reflects the situation in the North of Italy
- 3) The follow-up is relatively long which will be important for studies in birth cohorts and where there is need for long follow-up

4.3.9 IT: Pedianet (IT_PEDIANET)

Description

Pedianet is a paediatric research database including data collected by primary care paediatricians (FPs - Family Practitioners) during their routine daily practice since 2003. This system is based on the transmission of specific data (determined by individual studies) from computerised clinical files, which the paediatricians in the network fill out during their daily professional activities. Informed consent is required from the parents. Such data are collected anonymously by a central server in Padua, where the data are validated and elaborated. Pedianet allows linkage capabilities to other database via patient code. Each patient is identified by the system from a ID-patient and ID paediatric. Thanks to this system, all the information are anonymous, but at the same time, it is possible to link them with other database through these

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 56/76

code. For the ADVANCE project, Pedianet will participate with the Veneto data as only this database contains vaccine information.

Population fingerprint

Pedianet submitted data from 2004-2014. The population size is stable over time as we saw above. The cumulative amount was 77 thousand children, and the median follow-up is only 4.2 years, this is the subset that can be linked to vaccinations. The IT population comprises around 61 million persons and this subset of PEDIANET only represents a small proportion of all Italian children.

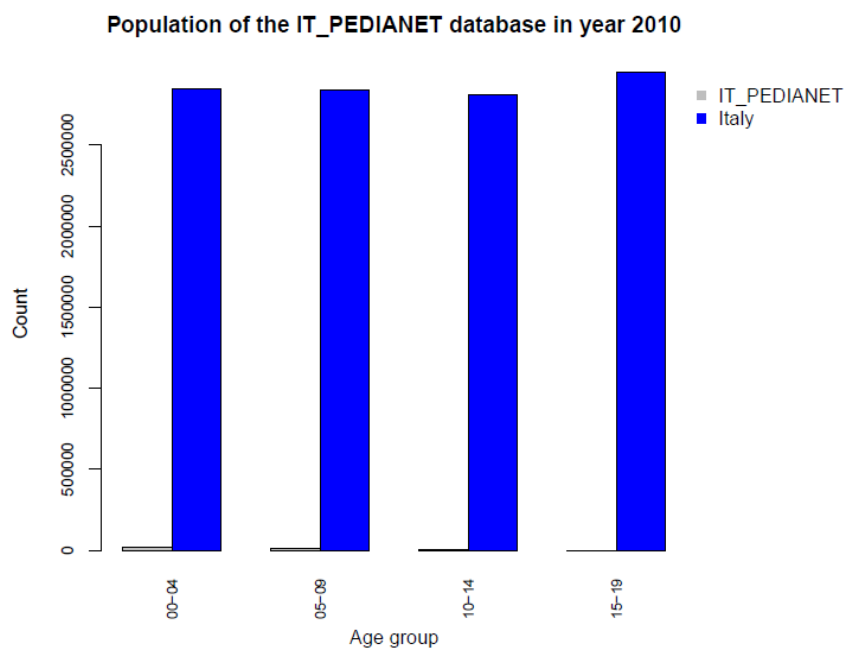



Figure 4.3.9.1. Distribution of the Pedianet database (ADVANCE fingerprint run) and IT pediatric population by age (IT reference from United Nations)³¹ in absolute numbers (2010).

³¹ <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 57/7 6

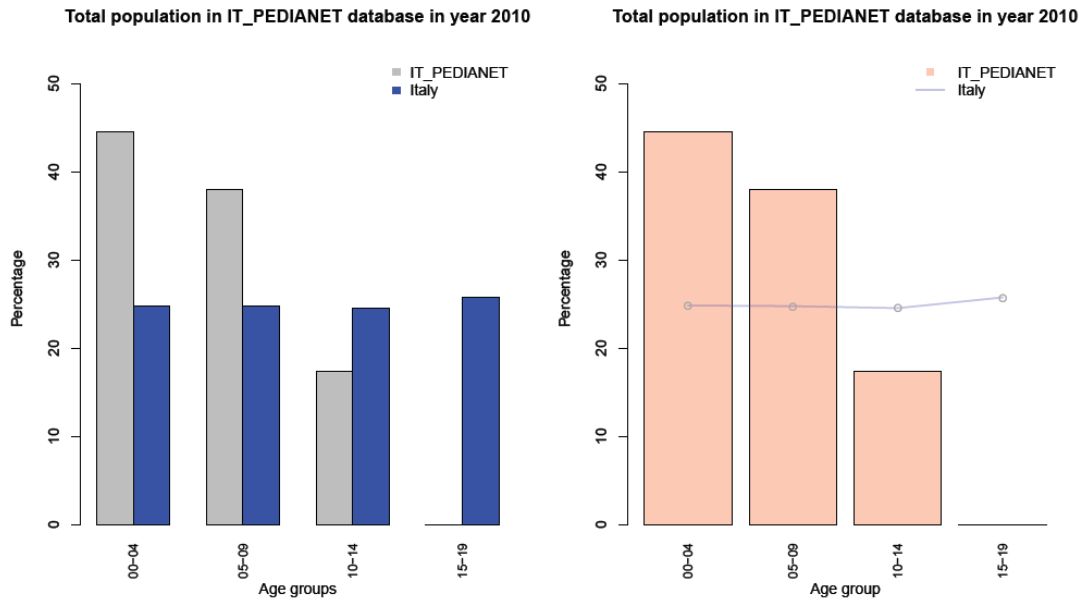


Figure 4.3.9.2. Relative distributions of the age of the population from Pedianet database (ADVANCE fingerprint run) and the IT pediatric population (United Nations) in 2010.

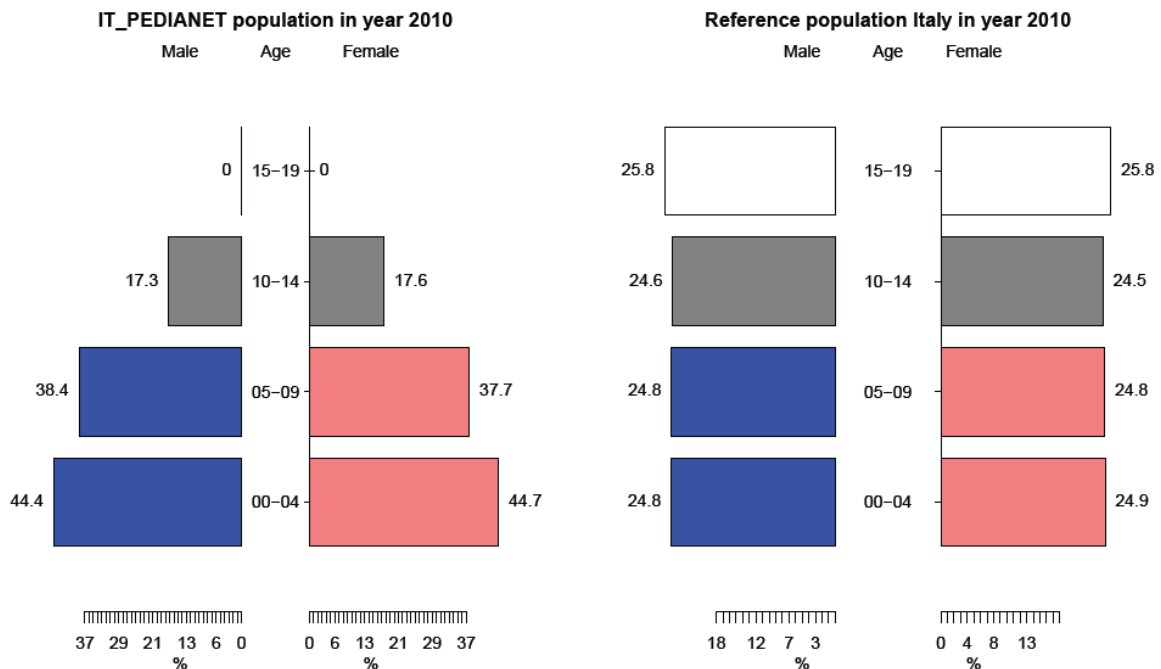



Figure 4.3.9.3. Distribution of the Pedianet database (ADVANCE fingerprint run) and IT population by age and sex (IT reference from United Nations) in percentages.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 58/76

These data show

- 1) The PEDIANET database captures only a small proportion of the IT pediatric population (less than 1%)
- 2) The age and gender distribution are not representative for IT pediatric population, the younger ages are overrepresented
- 3) The follow-up is short which will be important for studies in birth cohorts and where there is need for long follow-up

4.3.10 FI: Finnish HPV cohort (FI_UTAHPVCHRT)

Description

HPV-CRT Cohort: participants of a community randomized trial vaccinated as early adolescents with HPV 16/18 vaccine, HBV vaccine or unvaccinated.

Data custodian: University of Tampere/HES/STD-research group

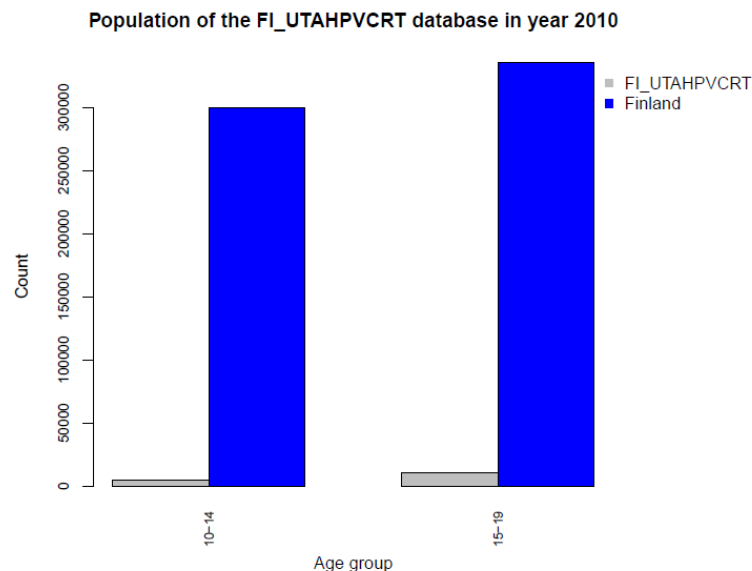



Figure 4.3.10.1. Distribution of the UTA HPV cohort (ADVANCE fingerprint run) and FI pediatric/adolescent population by age (FI reference from United Nations)³² in absolute numbers (2010).

³² <http://esa.un.org/unpd/wpp/Excel-Data/population.htm>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 59/76

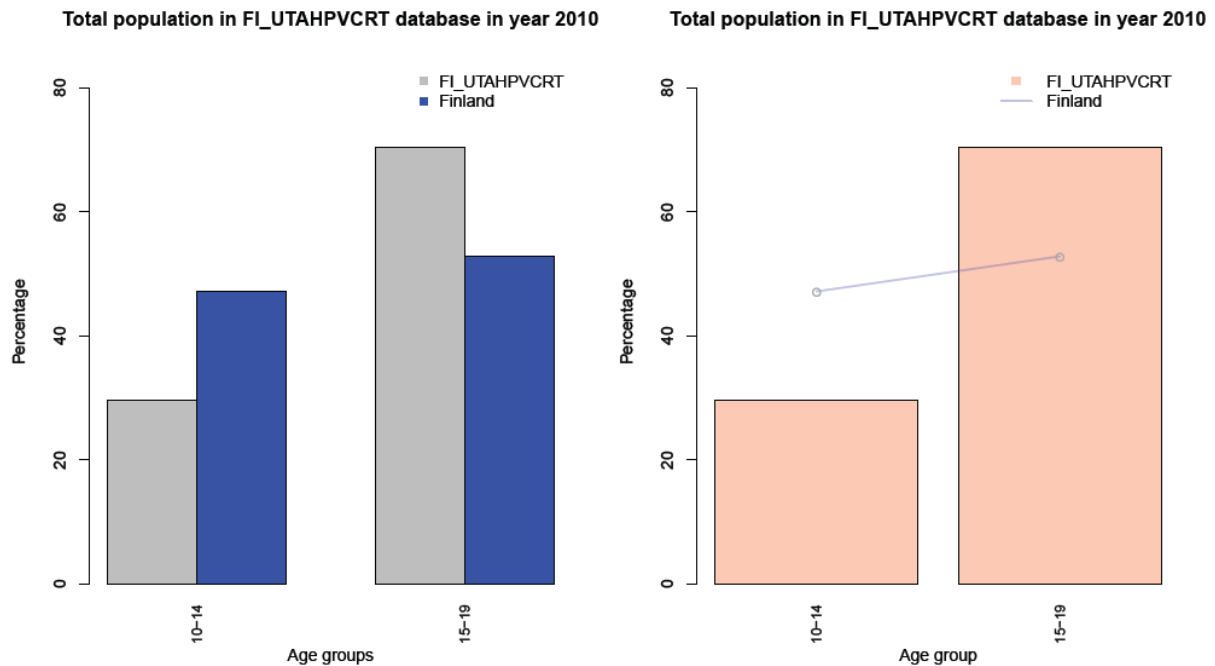


Figure 4.3.10.2. Relative distributions of the age of the population from UTA HPV cohort (ADVANCE fingerprint run) and the FI pediatric population (United Nations) in 2010.

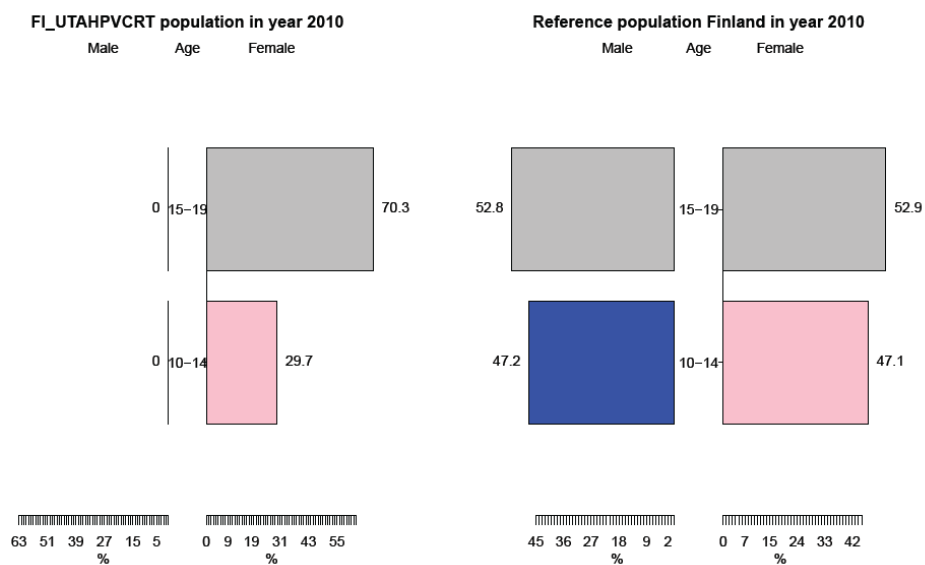



Figure 4.3.10.3. Distribution of the UTA HPV cohort (ADVANCE fingerprint run) and FI population by age and sex (FI reference from United Nations) in percentages.

	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	60/76

These data show

- 1) The UTA HPV cohort captures only a small proportion of the FI pediatric/adolescent population (less than 1%)
- 2) The age and gender distribution are not representative for FI pediatric/adolescent population, which is logical given the fact that this population represents a trial


Chapter 5. Further steps in fingerprinting

Next steps in the fingerprinting will be events followed by vaccinations and suitability, as outlined in chapter 2. The timelines will be event fingerprinting Dec 2014-March 2015, Vaccines from April 2015-June 2015, suitability will be done afterwards.


In addition, we will fingerprint the associate partner databases and databases that reside with vaccine manufacturers (e.g. FISABIO, IDIAP, CPRD).

Chapter 6. Conclusions

- The ADVANCE fingerprinting process showed to be well accepted, feasible and highly informative both to the ADVANCE as to the Local data custodian. Some technical issues were encountered but all could be resolved together with the IT team.
- This deliverable describes what was achieved in the first year and proposed what will be done in the ADVANCE data fingerprinting program. The fingerprinting aims at describing the type, amount and suitability of data in the databases and their suitability to implement timely benefit/risk monitoring and assessment of vaccines.
- We have started with population fingerprinting and have shown the results.
- Substantial differences exist between databases that participate in the consortium, this can be utilized in POC studies.
- Medical record databases generally have short follow-up and a very dynamic population, since patients may change general practitioner/family pediatrician, or because practices start or stop participation. On the contrary the regional/provincial/national record linkage databases have a very stable population with long follow-up.
- Most databases are representative of the national population, except PEDIANET and the UTA HPV cohort. This is not a problem as long as internal validity is regarded but should be considered when extrapolating.
- This first wave of fingerprinting results will be essential in planning of ADVANCE research plan and selection of the Phase I POC and studies data requirements.
- Further planned steps are to implement the vaccine exposure and event fingerprinting upon progress in the development of ontologies and mapping processes and methods this should be finished by June 2015.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

ANNEXES


	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 62/76

Annex 1. Database population fingerprint instructions



database fingerprinting run

USER INSTRUCTIONS

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 63/76

Introduction

In the ADVANCE project we federate the databases using a distributive network approach. We use the software package called Jerboa Reloaded to elaborate all the databases locally and produce common aggregated and anonymized output datasets (See Figure 1). The first version of Jerboa was developed within the EU-ADR project (FP7-ICT-2007-215847) and has since then been used successfully in several projects.




Figure 1. JERBOA model for distributed computing on databases

Primary data extraction run

To get a first database fingerprint we will run the Jerboa module called Primary Data Extraction on population level, e.g., the amount of patient time in the database before a certain year, the number of active patients in a year etc. The software will generate graphs to show you the extracted primary data. For this run only one simple input files needs to be created.

Database owners are asked to prepare the input file for Jerboa (see below) and to run the Jerboa program. The output of the program will be automatically encrypted and needs to be uploaded to the Remote Research Environment called OCTOPUS at Erasmus MC for further analysis as shown in figure 2. If you do not have access yet we will contact you. Instructions on how to obtain and run Jerboa can be found in Appendix 1.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 64/76

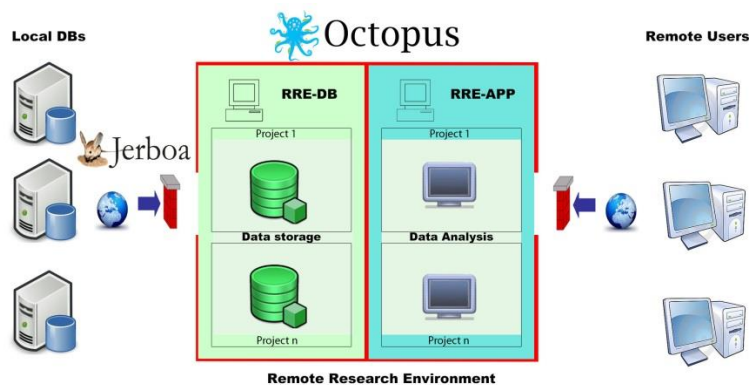


Figure 2. The OCTOPUS Remote Research Environment

If you have questions or doubts about Jerboa Reloaded or OCTOPUS please send an email to rre@erasmusmc.nl

Jerboa Data preparation

For this current primary data extraction run we only need a patient file that contains the following fields :

PatientID Patient Identifier.
Birthdate Date of birth.
Gender Gender of the patient.
Startdate Date from which the patient is eligible to be included in the study. This is typically the date the patient is entered into the registration system (date of registration with insurance/region, date GP started to collaborate).
Enddate Date after which the patient is no longer eligible for inclusion in the study (e.g. end of registration with GP, insurance, moving out, death, last data draw down (whichever is earliest)

NOTE: if patients occur multiple times in the database with different entry and exit codes, please include only the first time the patient is entered. This is relevant only for claims databases.


Example of patients input file:

```
patientid,gender,birthdate,startdate,enddate
1,F,19590601,19950802,20050701
2,M,19830301,19960912,20060903
```

Patient IDs

A patient ID is an alphanumeric string of characters that uniquely identifies a patient. Patient IDs can be numbers (1, 2, 3, etc.) or combination of numbers and letters (a01, a02, b01, etc.). There is no restriction in the length of the Patient ID.

Important: No duplicate patient IDs are allowed.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

Date formatting

You are not allowed to mix data formats in one file. All dates should be formatted as:

YYYYMMDD or YYYY/MM/DD or YYYY-MM-DD
 DDMMYYYY or DD/MM/YYYY or DD-MM-YYYY

For example: the 28th of March, 2008, is formatted as:

20080328 or 2008/03/28 or 2008-03-28
 28032008 or 28/03/2008 or 28-03-2008

Gender

The gender of a patient can have one of the following values:

FEMALE	Female	female	F	f
MALE	Male	male	M	m
UNKNOWN	Unknown	unknown	U	u

File format


Jerboa Reloaded allows multiple file formats. The input table should be either in CSV (Comma Separated Values) format or in TXT format with tabs delimited values or semicolon-delimited values. The first row should contain the column headers (the column header names provided above are mandatory). The order of the columns and rows is not important.

Note that missing values are not allowed! If a record in the input table is found with at least one missing value, the entire record is considered inconsistent and placed in a list of erroneous records. Jerboa Reloaded will automatically detect the patient file based on the header so the name of the file is irrelevant (we suggest you use for example 2013-05-10-Patient.txt to keep track of multiple versions).

Important: The input file is **always** checked for integrity before processing. If inconsistencies are found, an error log is produced for each input file and the user is asked to correct all present errors in order to continue. For instructions on the error log see appendix 1.

Appendix 1. Running Jerboa

Jerboa requires the latest Java version in order to run (version 7+). You can download it from here: <http://www.java.com/en/download/manual.jsp/>

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security:

[Following this link, choose the appropriate Java version for your operating system \(e.g., Windows, Mac OS X, Linux\) and its type \(e.g., 32 bits or 64 bits\). To find out on what type of operating system you are running, do the following:](#)

- [On Windows : right click on My Computer → Properties → see System Type](#)
- [On Linux: open a terminal \(Ctrl + Alt + T\) and type getconf LONG_BIT](#)

Instructions on how to install Java can be found here as well, but if you need help please let us know. Possibly, you need the help of your local technical staff with administrator’s rights to install new software on your machine.

Downloading the latest version of Jerboa and the script from Octopus

The database owners need to have access to Octopus to be able to download and upload files. If you do not have access yet, please ask for an application form by sending an email to re@erasmusmc.nl.

The [latest version](#) of Jerboa and the script for the current run can always be found in Octopus using FileZilla. Instructions on how to use FileZilla can be found in the documentation and video sent to all OCTOPUS users.

When you login using FileZilla you will see a folder named Jerboa-EU-ADR. In this folder you can find Jerboa, the script (.jsf) and documentation in a zip file.

Download and unzip the zip file into a folder and copy the script file into the folder containing your input files.

Running Jerboa

Double-click on the JerboaReloaded.jar file to start Jerboa. After accepting the license, ou will see the screen in Figure 1.

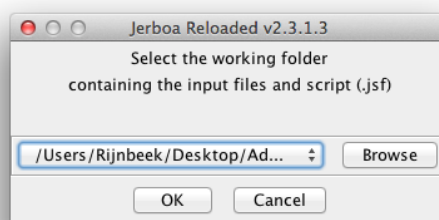



Figure 1. Opening a working folder

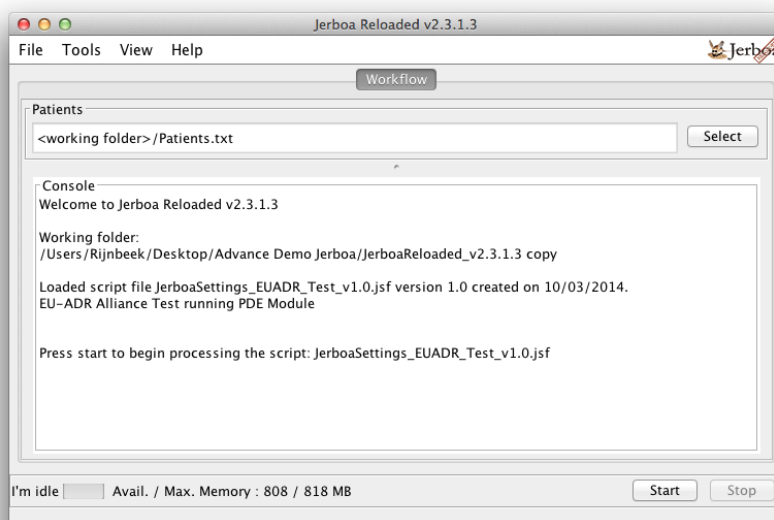
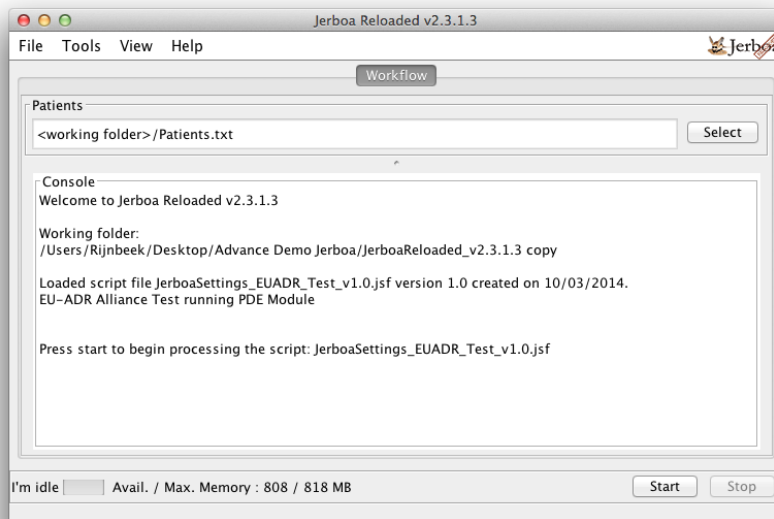
1. You can choose your working folder containing the input data file by clicking the browse button. In the first run of the Jerboa software, the folder where the JerboaReloaded.jar file is located is selected as default. If this folder corresponds to the location of your input file(s), just press OK.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 67/76

Previously used workspaces are remembered and available to open by clicking the dropdown list at the left of the browse button.

Important: Make sure that the provided **script file** (e.g., script.jsf) is in your chosen working folder.

2. Once a working folder is selected press OK to continue. The screen in Figure 2 will appear. As long as the patient file contains all mandatory columns, it will automatically be loaded and recognized, as shown in the Patients file panel on the upper side of the screen. If no patient file is found this will be indicated. **Note** that multiple patient files in the same folder are not allowed for this run.




 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 68/76

Figure 2. The application has successfully loaded the patients file

- Now click the start button and select your database. If your database is not listed you can add it using the add button

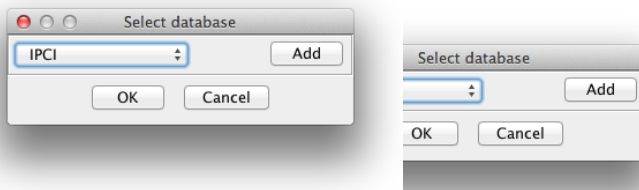


Figure 3. Selecting your database

- The application will check the input file and will report any errors found

4.a if errors are found in the input file(s), the user is informed as shown in Figure 4.

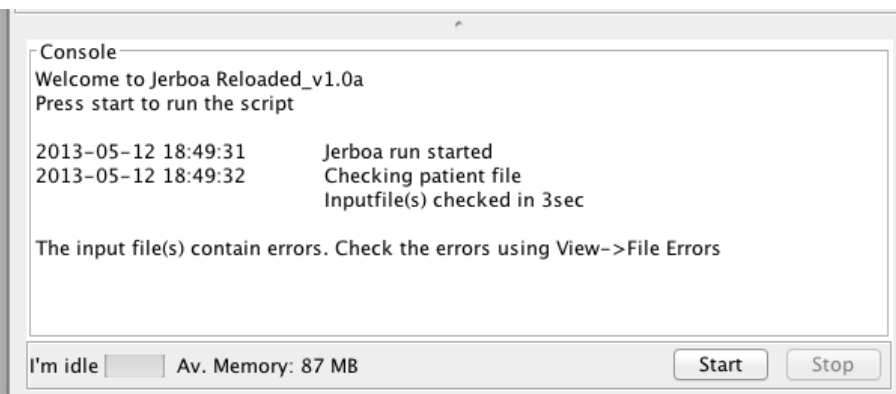

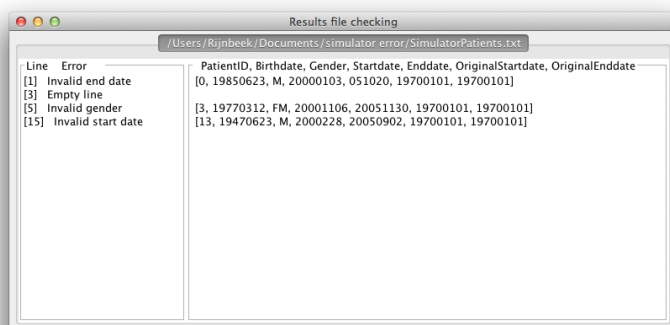


Figure 4. Errors in the input file(s)

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 69/76

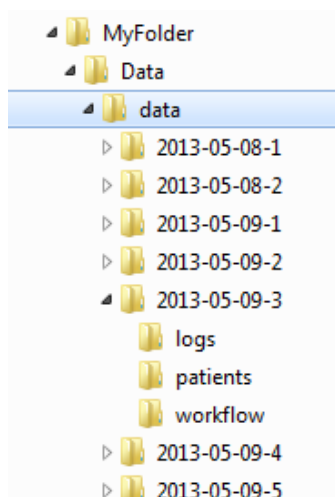
The user can check the errors by clicking the « View » menu and selecting « File Errors ». The following screen shown in Figure 5 will appear showing on the left side the error message and on the right side the actual content of the record in the input file. Alternatively, an error log file is generated for each input file. This files can be found in the « logs » folder of the current Jerboa run.


Important: In the working folder, a folder called « jerboa » is created. This folder contains all the files generated during each run of the Jerboa software. For each run, an individual folder is created inside the « jerboa » folder. The folder name is formed by the date of the run and the run number. This will allow you to keep a log of previous runs; Figure 6 shows an example of the folder structure created after multiple runs.



Figure

5. Results file checking



 IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 70/7 6

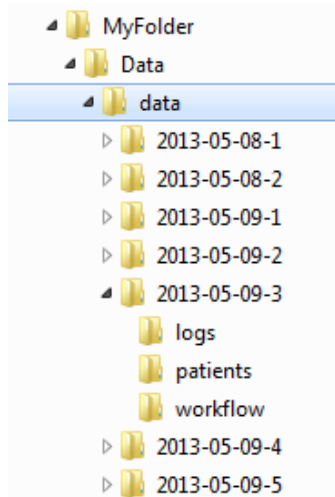


Figure 6: Folder structure created during Jerboa runs

4.b. If no errors are found in the input data file(s) the application will proceed and you can just sit back and relax. An indication of the time left to finish the current step is given in the progress bar on the bottom of the screen.

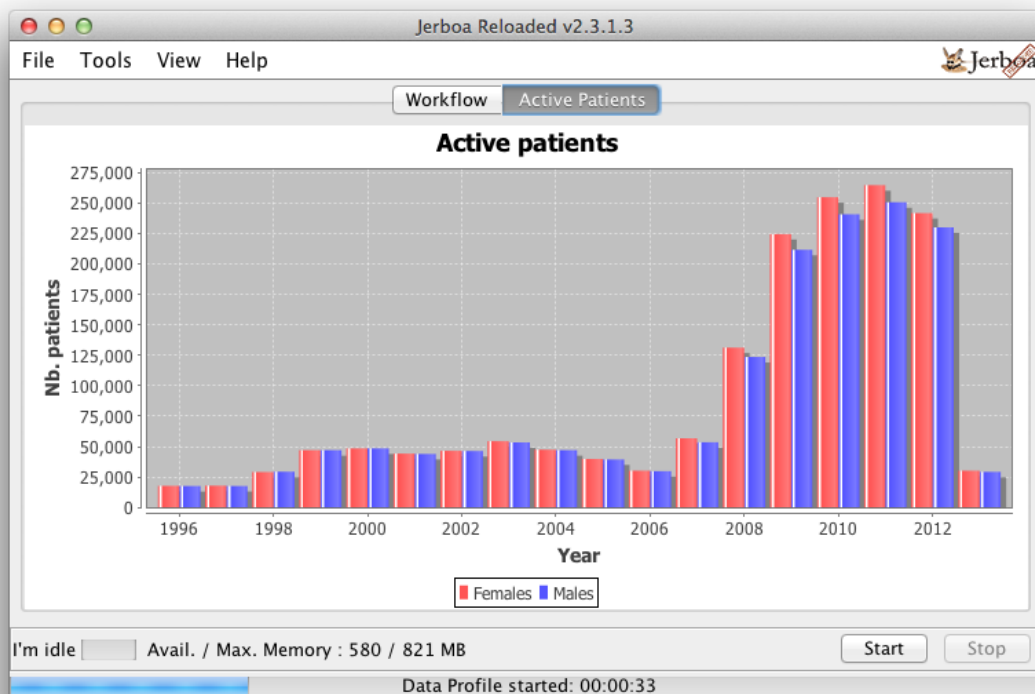

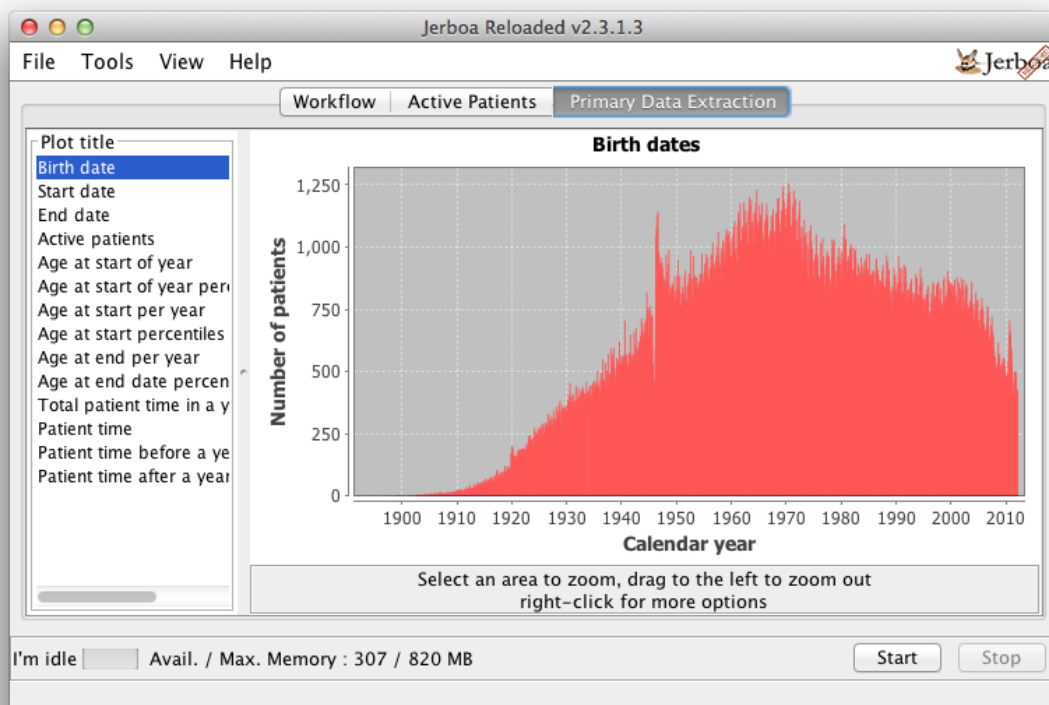



Figure 7. Input data checking was successful and processing the data following the script

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 71/7 6

During the run feedback is given in the form of graph showing the active male and female patients in your database per year. For each newly generated graph a tab on the top is created. When the data profiling has finished graphs are presented to the user in the PDE tab. In Figure 8 some examples are shown.



 IMM - 115557	D5.2 Fingerprinting of the participating health care databases	
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security: 72/76

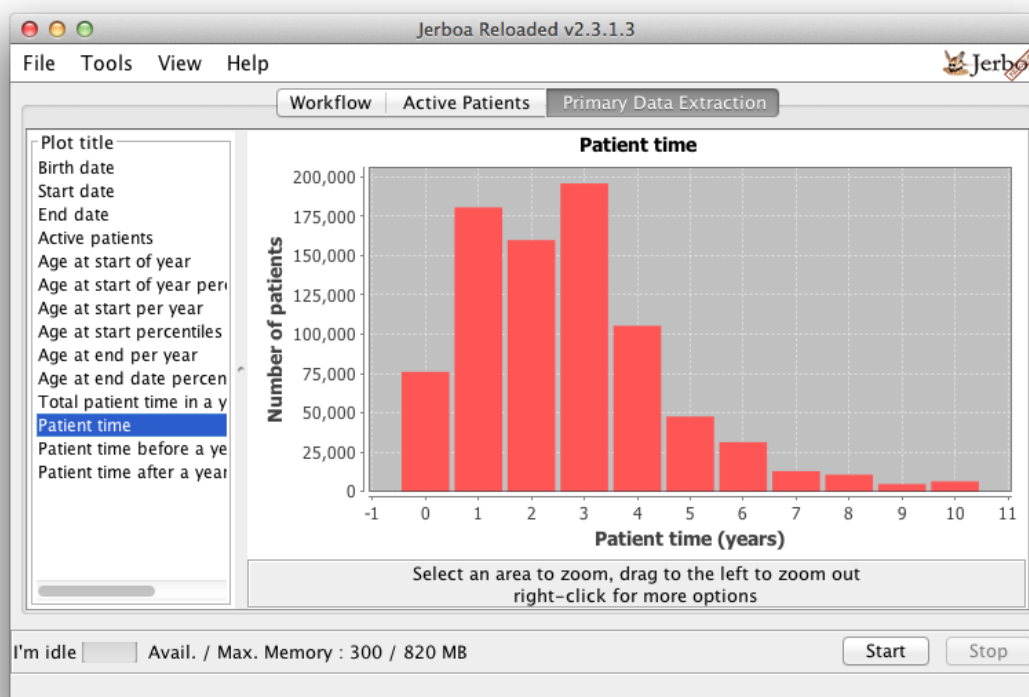


Figure 8. Examples on simulated data


For some of the graphs we generate males and females separate as well (use Next and Previous buttons). It is possible to zoom in by drawing a zoom window in the Graph. If you drag to the left the graph will zoom out to its original view. Right-click for more zooming options like zooming only one axis or print the graph.

In the working folder a pdf is created containing all the plots

- In the final step Jerboa will produce an .enc file. This is an encrypted file containing the output files. The file is to be found in the folder of the current run (e.g., MyFolder/Data/jerboa/2013-05-09-03/). The location is also shown in the console. This file should be sent to EMC following the procedure described below.


Sending of data to EMC using the Remote Research Environment Octopus

The .enc file should be uploaded to Octopus using the FileZilla procedure as described in the Octopus instructions. Please create a folder with the name of the project in your upload folder and upload the JERBOA output there.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	73/7 6

Send an email to rre@erasmusmc.nl with subject “[RRE FTP] <Project> Jerboa data upload from database #####”.


For any questions regarding Jerboa or Octopus, please use the same email address.

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	74/7 6

Annex 2: positive and negative vaccine-outcome associations

Modified (restricted list) from table from the reference set as defined in the GRIP project by Brauchli et al.

Vaccines	Anaphylaxis	Thrombocytopenia	Convulsions	Encephalitis	Arthritis	GBS	Wheezing / Reactive Airway disease	IDDM	Bell's Palsy
BCG	UC (1)	NC (1)	NC (1)	NC (1)	NC (1)	UC (2)	NC (2)	NC (2)	NC (2)
DTaP	IOM (1)	PC	IOM (1)	IOM (1), VIT	IOM (1)	IOM (1)	NC (2)	IOM (1)	IOM (1)
DTPw	VIT	UC (1)	UC (1)	VIT	IOM (1)	NC (2)	NC (2)	NC (2)	NC (1)
HAV	IOM (1)	UC (3)	NC (1)	NC (1)	NC (2)	IOM (1)	UC (2)	NC (1)	IOM (1)
HBV	IOM (1)	PC	IOM (1)	IOM (1)	IOM (1)	IOM (1)	UC (2)	NC (2)	UC (2)
PV	UC (2)	UC (3)	UC (2)	NC (1)	NC (2)	NC (1)	WHO	NC (1)	NC (1)
Influenza (any)	IOM (1)	UC (1)	IOM (1)	IOM (1)	IOM (1)	IOM (2)	IOM (1)**	NC (1)	IOM (1)
MV	IOM (1)	NC (2)	UC (2)	IOM (1)	NC (2)	IOM (1)	NC (2)	NC (1)	NC (1)
MMR	IOM (1)	VIT	IOM (1)*	IOM (1)	IOM (1)	IOM (1)	UC (1)	IOM (1)	NC (1)
VZV	IOM (1)	IOM (1)	IOM (1)	IOM (1)	IOM (1)	IOM (1)	UC (2)	NC (2)	UC (2)

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring		Version: v1.4 – final
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira		Security: 75/76

OPV	NC (1)	NC (1)	NC (1)	NC (1)	NC (1)	UC (1)	NC (2)	NC (2)	NC (1)
RV	NC (1)	NC (1)	NC (2)	NC (2)	NC (1)	NC (1)	NC (2)	NC (1)	NC (1)
Hib	UC (2)	NC (1)	UC (2)	NC (1)	NC (1)	UC (2)	NC (2)	NC (2)	NC (1)

Detailed information about basis for classification; association classified by:

1) Literature review:

- PC = positive control
- NC (1) = negative control - absence of evidence, NC (2) = negative control - evidence against
- UC (1) conflicting evidence, UC (2) absence of evidence or evidence of absence and ≥ 3 independent case reports/case series or proven pathomechanism or in SPC, UC (3) some evidence but not enough for positive control
- NM = pathomechanism not possible
- MG = review: Gold MS. Hypotonic-hyporesponsive episodes following pertussis vaccination: a cause for concern? Drug Saf. 2002;25(2):85-90. Review.²⁰

2) Official report:

IOM (1) = Reports of the Institute of Medicine, Adverse Effects of Vaccines: Evidence and Causality (2011)⁹

IOM (2) = Reports of the Institute of Medicine, Influenza Vaccines and neurological complications (2004)¹⁰


VIT = Vaccine Injury Table (July 2011)¹²

WHO = WHO, World Health Organisation, vaccine reaction rates information sheets¹¹

* = for febrile seizure

** = unclassifiable for children <5 years

	Positive control
	Negative control
	Unclassifiable

 ADVANCE IMI - 115557	D5.2 Fingerprinting of the participating health care databases		
	WP5. Proof-of-concept studies of a framework to perform vaccine benefit-risk monitoring	Version: v1.4 – final	
	Author(s): Miriam Sturkenboom, Peter Rijnbeek, Benedikt Becker, Jan Kors, Marius Gheorghe, Daniel Weibel, Germano Ferreira	Security:	76/7 6