From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding

Nicolas H. Thurin[1,*,†], Romin Pajouheshnia[2,†], Giuseppe Roberto[3,†], Caitlin Dodd[4], Giulia Hyeraci[3], Claudia Bartolini[3], Olga Paoletti[3], Hedvig Nordeng[4], Helle Wallach-Kildemoes[4], Vera Ehrenstein[5], Elena Dudukina[5], Thomas MacDonald[6], Giorgia De Paoli[6], Maria Loane[7], Christine Damase-Michel[8], Anna-Belle Beau[8], Cécile Droz-Perroteau[1], Régis Lassalle[1], Jorieke Bergman[9], Karin Swart[10], Tania Schink[11], Clara Cavero-Carbonell[12], Laia Barrachina-Bonet[12], Ainhoa Gomez-Lumbreras[13], Maria Giner-Soriano[13], María Aragón[13], Amanda J. Neville[14], Aurora Puccini[15], Anna Pierini[16], Valentina Ientile[17], Gianluca Trifirò[18], Anke Rissmann[19], Maarit K. Leinonen[20], Visa Martikainen[20], Sue Jordan[21], Daniel Thayer[21], Ieuan Scanlon[21], Mary E. Georgiou[22], Marianne Cunnington[22], Morris Swertz[9], Miriam Sturkenboom[23] and Rosa Gini[3]

In 2019, the Innovative Medicines Initiative (IMI) funded the ConcePTION project—Building an ecosystem for better monitoring and communicating safety of medicines use in pregnancy and breastfeeding: validated and regulatory endorsed workflows for fast, optimised evidence generation—with the vision that there is a societal obligation to rapidly reduce uncertainty about the safety of medication use in pregnancy and breastfeeding. The present paper introduces the set of concepts used to describe the European data sources involved in the ConcePTION project and illustrates the ConcePTION Common Data Model (CDM), which serves as the keystone of the federated ConcePTION network. Based on data availability and content analysis of 21 European data sources, the ConcePTION CDM has been structured with six tables designed to capture data from routine healthcare, three tables for data from public health surveillance activities, three curated tables for derived data on population (e.g., observation time and mother-child linkage), plus four metadata tables. By its first anniversary, the ConcePTION CDM has enabled 13 data sources to run common scripts to contribute to major European projects, demonstrating its capacity to facilitate effective and transparent deployment of distributed analytics, and its potential to address questions about utilization, effectiveness, and safety of medicines in special populations, including during pregnancy and breastfeeding, and, more broadly, in the general population.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ There is a societal obligation to rapidly reduce uncertainty about the safety of medication use in pregnancy and breastfeeding, using the framework of the heterogeneous European health data landscape.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ How can we preserve, leverage, and report on the heterogeneity in information provision across European data sources to support distributed pharmacoepidemiologic analyses of medicines during pregnancy and lactation?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ This study introduces and illustrates a set of concepts to represent heterogeneity across 21 data sources and the inception of the ConcePTION common data model.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ The tools created within the ConcePTION ecosystem will support the generation of evidence on the effectiveness and safety of medicines during pregnancy and breastfeeding.

[†]Contributed equally to the work.

Received July 1, 2021; accepted October 24, 2021. doi:10.1002/cpt.2476

[1]Bordeaux PharmacoEpi, INSERM CIC-P1401, Univ. Bordeaux, Bordeaux, France; [2]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands; [3]Agenzia regionale di sanità della Toscana, Florence, Italy; [4]PharmacoEpidemiology and Drug Safety Research Group, Department of Pharmacy, and PharmaTox Strategic Initiative, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway; [5]Department of Clinical Epidemiology, Aarhus University, Aarhus, Denmark; [6]MEMO Research, School of Medicine, University of Dundee, Dundee, UK; [7]Institute of Nursing and Health Research, Ulster University, Newtownabbey, UK; [8]INSERM, CERPOP: SPHERE, CIC 1436, Université de Toulouse, Toulouse, France; [9]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [10]PHARMO Institute for Drug Outcomes Research, Utrecht, The Netherlands; [11]Leibniz Institute for Prevention Research and Epidemiology–BIPS, Bremen, Germany; [12]Fundació per al Foment de la Investigació Sanitaria i Biomédica de la Comunitat Valenciana (FISABIO), Valencia, Spain; [13]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Universitat Autònoma de Barcelona, Bellaterra, Spain; [14]IMER Registry (Emila Romagna Registry of Birth Defects), Center of Epidemiology for Clinical Research, University of Ferrara, Ferrara, Italy; [15]Drug Policy Service, Emilia Romagna Region Health Authority, Bologna, Italy; [16]Epidemiology of Rare Diseases and Congenital Anomalies Unit, National Research Council-Institute of Clinical Physiology (CNR-IFC), Pisa, Italy; [17]Department of Biomedical and Dental Sciences and Morpho-functional Imaging, University of Messina, Messina, Italy; [18]Department of Diagnostics and Public Health, University of Verona, Verona, Italy; [19]Malformation Monitoring Centre Saxony-Anhalt, Medical Faculty, Otto-von-Guericke-University, Magdeburg, Germany; [20]Finnish Institute for Health and Welfare, Helsinki, Finland; [21]Faculty of Health and Life Science, Swansea University, Swansea, UK; [22]Glaxosmithkline, Stevenage, UK; [23]Department Datascience and Biostatistics, University Medical Center Utrecht, Utrecht, The Netherlands. *Correspondence: Nicolas H. Thurin (nicolas.thurin@u-bordeaux.fr)

More than 200 million women worldwide become pregnant every year,[1] and a majority take at least one medication during pregnancy.[2,3] However, merely 5% of available medications have been adequately monitored, tested, and labeled with safety information for use in pregnant or breastfeeding women.[4,5] Pregnant and breastfeeding women have long been overlooked in medical research, especially in randomized controlled clinical trials,[6] due to complex practical, ethical, and legal challenges. Since the 2000s, data generated in routine healthcare have been increasingly used and recognized as a valuable source for evidence generation, offering the possibility of filling evidence gaps left by randomized controlled clinical trials and rendering the lack of information on pregnancy and breastfeeding even more untenable from ethical, medical, and societal perspectives. In 2019, the Innovative Medicines Initiative (IMI) funded the ConcePTION project—Building an ecosystem for better monitoring and communicating safety of medicines use in pregnancy and breastfeeding: validated and regulatory endorsed workflows for fast, optimised evidence generation—with the vision that there is a societal obligation to rapidly reduce uncertainty about the safety of medication use in pregnancy and breastfeeding.[7] ConcePTION aims to contribute to filling the knowledge gap regarding the effects of medicines in pregnancy and lactation, by developing a system across European Data Access Providers (DAPs) that transforms existing and routinely collected healthcare data into evidence in a robust and transparent manner.

The creation of a federated network using distributed analytics comes with several barriers, such as interoperability, transparency, data protection regulations, and ease and speed in conducting analyses across multiple data sources.[8] The execution of distributed multidatabase studies using a common data model (CDM) and common analytics, through which analysis scripts are sent to the data rather than data being centralized for analysis, is one solution to these issues.[9,10] However, the choice of which CDM to adopt is a difficult and recurrent question.[8,11] CDMs, such as those of the US Food and Drug Administration (FDA) Sentinel System,[12] the National Patient-Centered Clinical Research Network (PCORnet),[13] or the Observational Medical Outcomes Partnership (OMOP)[14,15] have been used in the United States to answer public health questions and drug safety concerns. In Europe, several initiatives led to the creation of CDMs but mainly at a regional level,[16] for disease-specific studies (e.g., EUROlinkCAT),[17] or at study-specific level,[18] and no consensus on a single CDM has yet emerged. Over the last 3 years, the European Health Data & Evidence Network (EHDEN), funded by the IMI, has supported the adoption of the OMOP CDM across Europe to create a federated network,[19] and this initiative is ongoing. The extract, transform, and load (ETL) process of data to the OMOP CDM requires large time and financial resources, as conversion of the database structure (structural harmonization) and of the coding systems (semantic harmonization) are mandatory.[20] These challenges are compounded in the European context by heterogeneity of the data sources in terms of structure, provenance, and language. Fitting OMOP standards for all the DAPs involved in ConcePTION would have been impossible without jeopardizing the forthcoming stages of the project. In addition, technical choices in the processing of the data have profound consequences on the quality of the evidence generated. Driven by these perspectives, the ConcePTION project's strategy was to design a tailor-made CDM with the capacity to (i) preserve the granularity resulting from European data heterogeneity, (ii) be implemented within constrained timelines and budget, and (iii) conduct distributed analyses transparently and efficiently.

This paper aims to introduce the set of concepts used to describe the European data sources involved in the ConcePTION project and illustrates the ConcePTION CDM, which serves as the keystone of the federated ConcePTION network.

METHODS

Data collection and content analysis

In ConcePTION, the DAP refers to an organization with the ability and expertise to access, process, and analyze healthcare data to conduct pharmacoepidemiological research. The 20 DAPs involved in ConcePTION were selected based on their experience and interest in conducting studies relating to medicines safety in pregnancy and breastfeeding, as well as for their expertise in making use of their data source demonstrated by multiple high-quality scientific publications. In Europe, DAPs generally access data on a project or permit basis, as they are often not the data custodians.

A systematic characterization of the 21 data sources accessible by the 20 DAPs was conducted. Initially, the data sources were conceptualized as relational databases, or databases that could be unambiguously

broken down into distinct data tables. First, the data model and the corresponding data dictionary accessed by each DAP were requested, in original format. Second, pairs of researchers from the coordinating team reviewed the data dictionaries. Third, each DAP was asked to produce in written form a high-level description of the data source they use and the answers to a fixed set of questions for each data table to understand key characteristics:

- What triggers the creation of a record in the table?
- Is the table collected for all the population of your database, or only for a subpopulation?
- Can you comment on the completeness and quality of the table? If you do not have formal measurements, feel free to convey the assumptions you commonly make.
- What is the time span of the table, how often it is refreshed, and what is the lag time between data creation and the time when the data has the potential to be available to your organization?
- Include other comments you may want to share about this table.

Fourth, 120-minute structured interviews were conducted with 2 representatives from each DAP. Interviewers were instructed to review and clarify with the DAPs their answers to the above questions. Fifth, a content analysis was conducted on the interview answer documents.[21,22] Data were extracted in summary tables. DAPs revised the headers of these summary tables, their corresponding categories, and the populated content. The process was iterated until consensus was reached. This led to the identification of latent concepts that more accurately represented the data. Concepts were prioritized, which allowed for a concise yet exhaustive representation.

Design of the ConcePTION Common Data Model

A first draft of the ConcePTION CDM was created based on the basic requirements of pharmacoepidemiological studies in pregnancy: demographics, exposures, outcomes, healthcare encounters, procedures, death, and mother-child linkage. A formal comparison of this first CDM with the OMOP CDM version 6 was conducted, in order to align structures as much as possible. Wherever possible, the adoption of the OMOP conventions were considered to facilitate collaboration with DAPs already using the OMOP CDM. To verify completeness, a comparison with the FDA Sentinel System CDM was also conducted. The first version of the ConcePTION CDM was then updated to accommodate heterogeneity and specific linkages highlighted during DAP interviews (e.g., integration of EUROCAT tables).

RESULTS

Outcome of the content analysis

The 20 DAPs involved in ConcePTION were able to access and provide information on 21 data sources. The detailed information extracted from the interview answer documents are included in **Supplementary Material S1**.

The content analysis highlighted that a data table was too granular a unit of observation in which to deconstruct a data source. For instance, a pair of tables described by the Aarhus University contained, respectively, diagnoses and procedures from the same hospital stay, therefore all the answers to the interview questions were replicated across the two tables. The latent concepts identified after the content analysis are described in **Table 1**, hereafter referred to as the ConcePTION conceptual framework.

Table 1 ConcePTION conceptual framework

Concept			Definition
Data source	Data bank	Data bank	A collection of structured healthcare data, sustained by an organization, which mandates that a record is prompted whenever a specified class of events occurs to persons belonging to a specific population.
		Originator of a data bank	Organization which mandates and sustains the data bank.
		Prompt of a data bank	A specified class of events that prompt a record in the data bank.
		Underlying population of a data bank	Records are generated whenever the prompts happen to a specified population.
	Data source		A collection of data banks having the same underlying population, or overlapping populations, that can be linked to one another at an individual level.
	Data source population		An organization may have access to entire data banks, or a subset thereof. The corresponding population is referred to as the data source population.
Instance of a data source	Instance of a data source		Subset of a data source extracted for the purpose of conducting one or more studies.
	Instance population		The subpopulation of the data source population that is included in an instance of a data source.
	Instance of a common data model		An instance of a data source converted to a Common Data Model.

Table 2 Data bank families included in at least two ConcePTION data sources

Data bank (number of data sources including it)	Originator	Organizations which collect data	Prompts for records in the data bank family[b]	Typical content	Less common content
Hospital administrative records (16)	Healthcare payer[a]	Healthcare service providers: hospitals	Discharge from a hospitalization; a specialist encounter or emergency department visit may also prompt records	Diagnosis/signs/symptoms observed; main diagnosis that led to the hospitalization; administration of procedures; specialty of the ward; outcome	Socio-economic status, test results, hospital name, indicators of emergency hospitalization/overnight stay
Primary care medical records (5)	Network of primary care practices	Primary care practices	Contact between patients and their primary care practice: face-to-face/telephone/online. Records may be prompted when hospital discharge information is received by the practice (e.g., UK data banks)	Registration with the practice; diagnosis/signs/symptoms; prescription of diagnostic tests/medicines	Socio-economic status, test results, vaccination, dosing regimen or indication of prescribed medicines, hospital/specialist referrals
Pharmacy dispensings records (15)	Healthcare payer[a]	Healthcare service providers: pharmacies	Dispensing of a medicine for reimbursement by a healthcare payer. Prompted by community pharmacies, hospital pharmacies or both, mostly by dispensing for outpatient use and/or for outpatient administration, rarely by inpatient use	National code of the medicinal product, ATC code and amount dispensed	Link to the prescription, batch number, condition of pregnancy of the patient, batch number
Birth registry (12)	Public health/statistical authority,	Hospitals or midwives or children health services	Live births observed in hospital or at the first visit of the child. If the prompt is delivery in hospital, stillbirths prompt a record and are distinguished from spontaneous abortion by the gestational age at labor	Information on the mother, on the pregnancy, on the delivery, on the child(ren)	Information on the father
Induced terminations registry (4)	Public health/statistical authority, authority allowing terminations	Hospitals executing the procedure or practices authorizing the procedure	Request or execution of an induced termination. The record may be anonymous	Information on the circumstances of the termination, on the pregnancy and on the woman	Link to the corresponding hospital administrative record
Congenital anomaly registry (10)	Public health authority, research center	Hospitals, healthcare professionals involved in the delivery	Recording of a congenital anomaly, at birth or during a follow-up of several years; a fetal death with an anomaly	EUROCAT core variables	Other EUROCAT variables
Inhabitant registry (3)	Civic authority (national or regional)	Civic offices	Immigration/emigration in a country or region	Immigration/emigration date, birth date	Date of death, address
Registration with healthcare system (6)	Healthcare payer[a]	Healthcare service office	Registration with a healthcare system (primary care physician and/or health insurer)	Date of registration	Date of de-registration, physician/insurer name, address
Exemptions from copayment (7)	Healthcare payer[a]	Healthcare service office	Some healthcare payers admit exemptions from copayment due to some health conditions	Cause for exemption	
Death registry (6)	Public health authority	Public health authority	Recording of the cause of death	Principal cause of death	Secondary causes of death

ATC, Anatomic Therapeutic Chemical.

[a] Definition of health payers, from Donnelly *et al.*[23]: single payer refers to a health system that is financed by a single entity; in its common usage, that single entity is government. Examples: Italy, Denmark, Norway, Spain, and the United Kingdom. Multiple payer refers to a health system that is financed through more than a single entity, one of which may be governmental. Examples: Germany, the Netherlands, and France. Whether a healthcare system is single- or multiple-payer does not in and of itself define the system in terms of coverage. Universal coverage means simply that all people within a particular jurisdiction have access to healthcare, be it single- or multiple- payer. Examples: all countries in this study. [b] The prompts identified by the 20 DAPs are listed for each data bank family. However, not all prompts within a family result in creation of a record for each individual data bank within that family.

Representation of data sources and data banks

Data banks with similar record prompts and originators were grouped into 10 families (see **Table 2**). The most common data bank family was Hospital Administrative Records, which is included in 16 out of 21 data sources; followed by Pharmacy Dispensing Records, included in 15 data sources. Most data bank families were sustained by the payer of the healthcare system, others originated from public health or government statistical authorities. The exception was Primary Care Medical Records, included in five data sources: this type of data bank is created for clinical purposes and contains detailed clinical information. The most frequent record prompts resulted from: access to a medical facility (e.g., hospitalizations, visits to emergency departments, primary care practices, or specialist clinics), dispensing of healthcare services (e.g., a medicinal product, a vaccine, a diagnostic test), accessing medical facilities due to pregnancy (childbirth and pregnancy loss), registering with healthcare services' organization(s) (e.g., enlisting with a payer and immigration), or events related to population surveillance (e.g., death and emigration). Although record prompts were similar for data banks aggregated in the same family, local differences remained. For example, among data banks belonging to the family Pharmacy Dispensing Records, hospital pharmacies may prompt records only if they dispense medicinal products for outpatient use, or may also include outpatient administration (e.g., infusions), or, in rare cases, also inpatient use. Among data banks of the family Hospital Administrative Records, creation of a record may be prompted by one or more of the following: hospitalization, emergency department attendance, or specialist appointment. Similarly, in terms of content, all the data banks of a single family shared a core set of data items, but any individual data bank may collect additional content, such as socio-economic status, or link between prescribing a medicinal product and its subsequent dispensing.

Data source and data bank characteristics are summarized in **Supplementary Material S2**, including corresponding underlying populations, prompts, and content. Because most data banks are sustained by the healthcare payer, in most data sources, the underlying population comprises the lawful inhabitants of a geographic area. This is because national healthcare systems in Europe all have universal coverage, which is often supported by a single payer,[23] usually a national or regional government (e.g., the UK's National Health System (NHS)). In countries whose healthcare system has multiple payers (e.g., Germany), the underlying population of data banks sustained by a healthcare insurer is people registered with that healthcare insurer. Data sources, such as a Primary Care Medical Record data bank, have an underlying population defined by the overlap between the patients of the primary care practices generating the medical records and the population registered with the healthcare payers sustaining the other data banks representing services outside primary care. Some DAPs contributing to ConcePTION only access subpopulations of interest for research questions related to pregnancy, for example, pregnant women (e.g., the Centre Hospitalier Universitaire de Toulouse (CHUT)), or infants/fetuses with a congenital anomaly and their mothers (e.g., CNR-IFC).

Figure 1 shows the links between each data source and the different families of data banks. All DAPs but one described a single data source; the CHUT described two. Three data sources were solely Congenital Anomalies Registries, and one was composed only of Primary Care Medical Records. All the other data sources included between 2 and 12 data banks, and no pair of data sources included the same combination of data banks. Ten data sources hosted one or several data bank(s) (category "Other") that were not represented elsewhere (e.g., transport claims, cancer registries, and specific outpatient claims).

ConcePTION Common data model

The ConcePTION CDM version 2.2 released in April 2021 is represented in **Figure 2**. The full sets of tables are described in **Supplementary Material S3**. The tables were divided into four sections:

- *Routine healthcare data*, accommodating data banks generated during routine healthcare, including Hospital Discharge Records, Primary Care Medical Records, and Pharmacy Dispensing Records.
- *Surveillance*, accommodating data banks generated from activities of public health surveillance (e.g., Birth Registry, Cancer Registry, and Death Registry), research (e.g., cohorts), and cross-sectional or longitudinal surveys (e.g., the Mother and Child Protection Centre database, accessed by the CHUT). The Congenital Anomaly Registry is included in this section in the standard EUROCAT format.[24]
- *Curated tables*, accommodate the periods of time when each person in the data source population is observed. They comprise summary information about each person (e.g., birth date), and the relationships between persons in the data source population (e.g., mother-child relationship). These tables are mostly created based on Inhabitant Registries, Registration with Healthcare Systems, and Birth and Death Registries.
- *Metadata*, accommodating information about the data source (e.g., type of coding system used) and its specific instance (e.g., date of last data update), as well as details on medicinal products marketed in the underlying population (e.g., product identifiers and number of pills in a box).

Two tables of the CDM, MEDICAL_OBSERVATIONS in Routine healthcare data and SURVEY_OBSERVATION in Surveillance, have an "Entity-Attribute-Value" (EAV) structure, illustrated in **Figure 3**. An EAV table can be conceptualized as a table with three columns: entity is the unit of observation (in our case, an identifier of the original record, e.g., person_id), attribute is an identifier of the observable event (in our case, the name of the columns in the original record, e.g., height, weight), and value is the observable event itself (in our case, the content of the original column e.g., "169" for height or "63" for weight).[25] In particular, the EAV format is suitable to load data banks of questionnaires, surveys, registries, or surveillance systems, or add information that does not fit in the structure of Routine healthcare data tables of the CDM, such as gestational age, delivery data, or birthweight, which are of interest for pregnancy and breastfeeding related studies.

Notably, each Routine healthcare and Surveillance table comes with origin and meaning columns, conveying the context that brings that record into existence: the former contains the

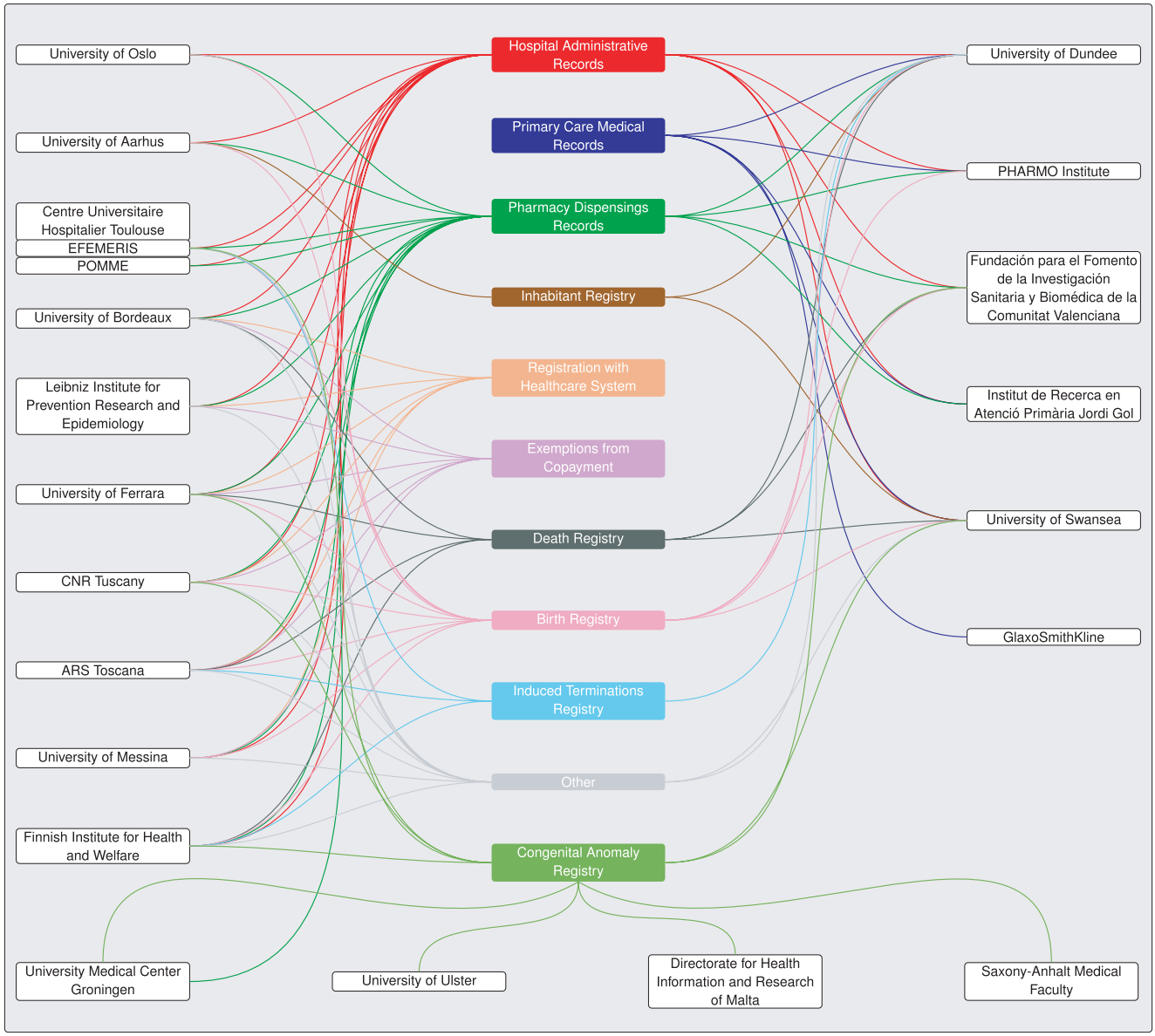

Figure 1 Data banks in each data source. Only data banks included in at least two data sources are represented, the others are summarized in "Other." The data banks are described in **Table 2**.

name of the data bank, the second a description of the prompt (e.g., discharge from hospital, or visit to the emergency department). An updated version of the comparisons between the OMOP CDM version 6 and the FDA Sentinel CDM with the ConcePTION version 2.2 is available on **Supplementary Material S4**.

ConcePTION Vocabulary

The analysis of the interview data suggested that the CDM should be able to handle rich and nonstandardized data vocabularies. Thus, the ConcePTION vocabulary was based on a combination of the data banks' dictionaries: each coded column of the ConcePTION CDM comes with a corresponding column indicating its dictionary. The dictionary may be a standardized coding system (e.g., International Classification of Disease 9th revision (ICD-9) or a specific coding system documented in the ConcePTION vocabulary either with a reference to a lookup table or with explicit decoding. The vocabularies of the meaning and origin columns are open: each DAP is free to include additional values, to capture the information that in their experience may be useful when processing the records for a study.

The detailed vocabulary of the ConcePTION CDM version 2.2 is available in **Supplementary Material S3**.

Extract, Transform, and Load from original data banks to the ConcePTION CDM

To support DAPs in the ETL process, an ETL design template was developed defining the link between source data (original data banks) to the target tables of ConcePTION CDM. In this

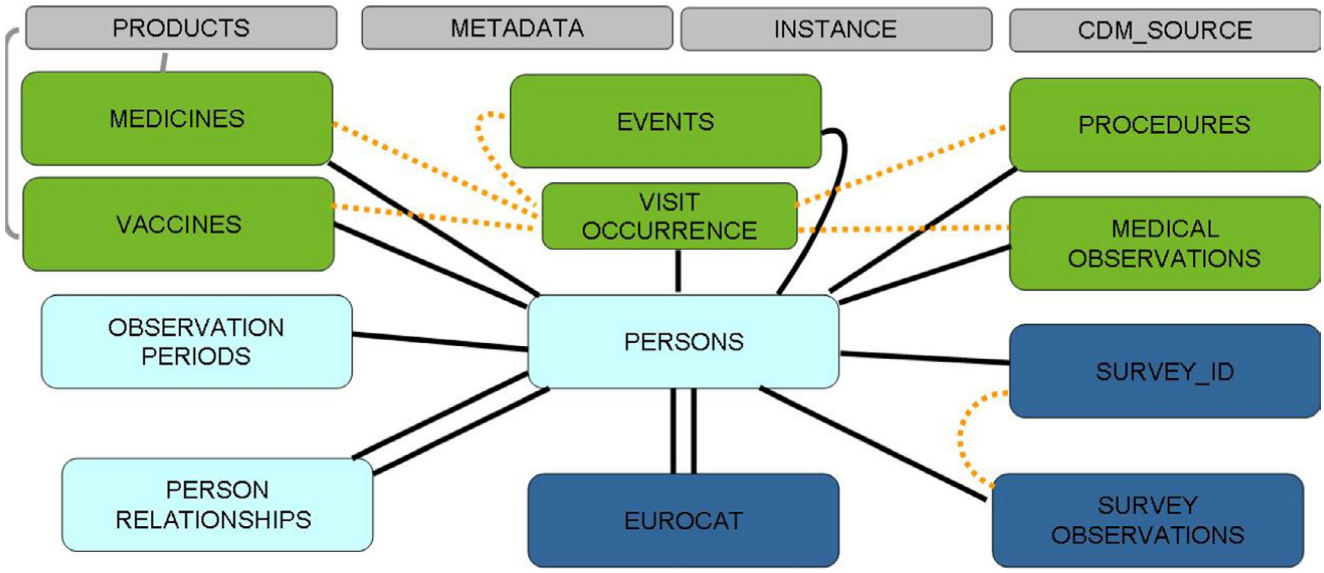

Figure 2 ConcePTION CDM version 2.2. Solid black lines refer to the linkage across records of the same person; dotted lines refer to linkage across items extracted from the same record; solid grey lines refer to linkage from items referring to a medicinal product to the product itself. Tables are color coded according to the section of the common data model they belong to: Routine healthcare data are represented in green, Surveillance data in dark blue, Curated data in light blue, and Metadata in grey. CDM, Common Data Model; ETL, Extract, Transform and Load.

person_id	HEIGHT	WEIGHT	GESTAGE_WEEKS
P1	169	63	37
P2	170	64	40

person_id	so_source_column	so_source_value
P1	HEIGHT	169
P1	WEIGHT	63
P1	GESTAGE_WEEKS	37
P2	HEIGHT	170
P2	WEIGHT	64
P2	GESTAGE_WEEKS	40

Figure 3 Example of Entity-Attribute-Value structure. The data contained in the upper table is represented in the lower table as an Entity Attribute Value fashion. Person_id is the Entity; HEIGHT, WEIGHT, and GESTAGE_WEEKS are the Attributes; and, for instance, 169 is the Value of Attribute HEIGHT for Entity P1.

document, for each table and column to be populated in the CDM, the following information was collected: (i) the rule that feeds each target column based on source columns and/or context information, and (ii) the rule that generates target record(s) from each source record. The ConcePTION ETL template is available in **Supplementary Material S5**.

There is a "many-to-many relationship" between the data banks of a data source and the tables of the ConcePTION CDM: every original data bank may feed multiple target tables, and every target table may be fed by multiple original data banks. The most common pairs of "original data bank family–target table" are represented in **Figure 4**.

DISCUSSION

The qualitative analysis of 21 data sources accessed by 20 European DAPs led to an understanding that the concepts of data tables and databases were insufficient to describe the landscape of European data, as the scope of those terms is restricted to information storage, whereas we need a description of how the information is generated, maintained, and accessed.[26] Therefore, we introduced a specific meaning for the terms data bank and data source: the former represents collections of data that are independent of a study and the latter represents an assimilation of data banks, where generation of observational evidence occurs. This conceptual framework allows to classify which prompts are included in each data source, which readily clarifies which data sources can provide information about spontaneous abortions or terminations, as well as ongoing pregnancies. This contributes in a transparent manner to the assessment of the suitability of data sources to address specific research questions. For instance, in the CONSIGN study,[27] this allowed the investigators to understand which data sources could retrieve timely information on women who were in the early stages of pregnancy when they contracted coronavirus disease 2019 (COVID-19). It was found that some data sources were more suitable for this assessment as they contained data banks with prompts that could be used to identify an ongoing pregnancy, which was not the case for all data sources. Together with experiences from prior projects,[18,28–33] these concepts served as support for the development of the

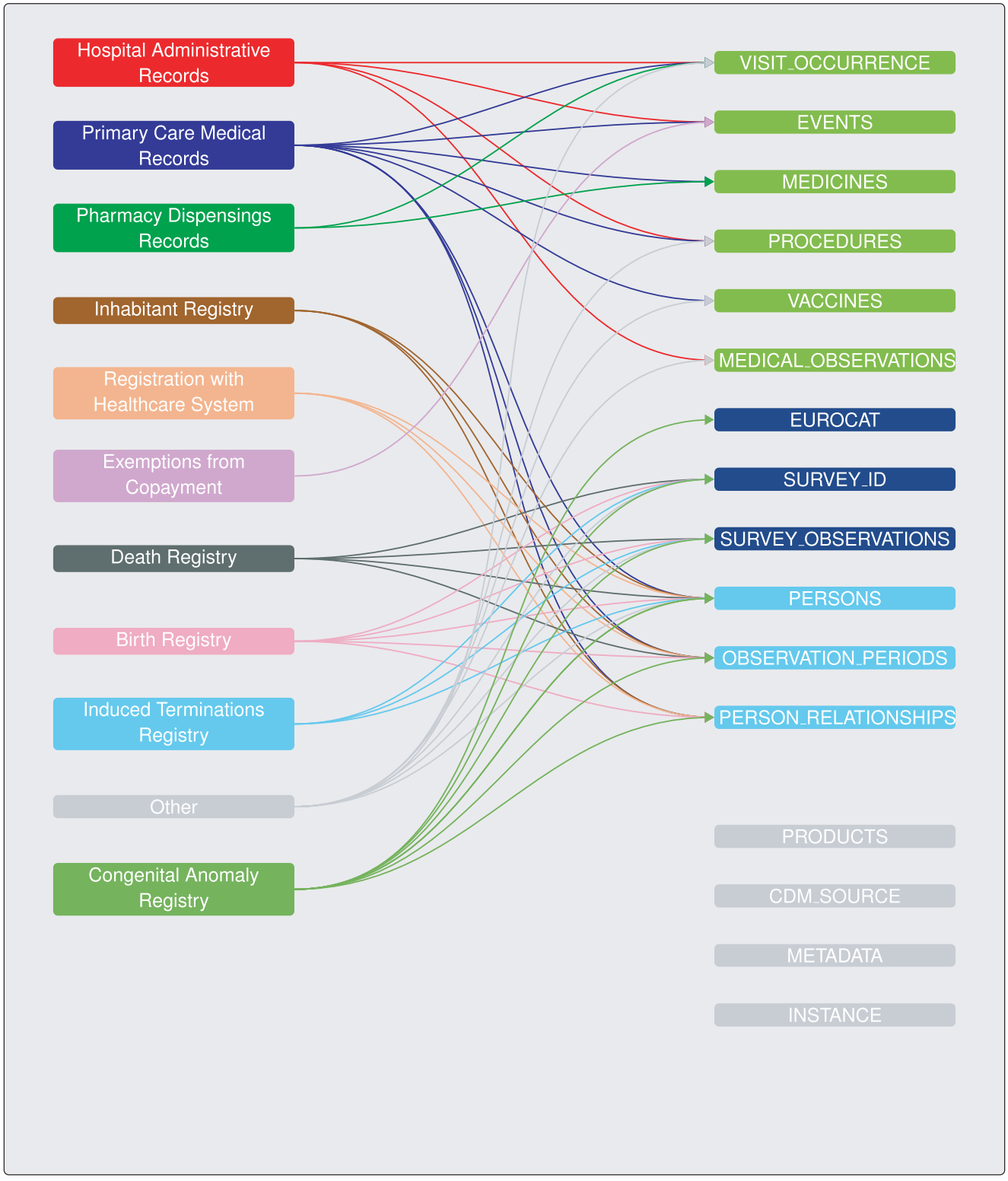

Figure 4 Common ETL between original families of data banks and tables of the ConcePTION CDM version 2.2. In the Figure, each arrow represents a pair formed by an original family of data bank and a ConcePTION CDM target table. CDM, Common Data Model.

ConcePTION CDM, a generic CDM addressing the needs of the ConcePTION ecosystem.

The ConcePTION CDM has been specifically designed to manage, within constrained timelines and budget, the heterogeneity inherent in the diverse data banks in Europe, which vary in both structure and content. This capacity to align very heterogeneous sources and to preserve their levels of detail allows leveraging of specific records, such as those relating to pregnancy, which are

challenging to explore in distributed approaches. In this context, technical choices may have profound consequences on the evidence generated. For example, information about the start of a pregnancy may be captured with different types of records across data banks—date of last menstrual period, or gestational age in weeks, or gestational age in days—and all of those slightly different data items have to be treated in a specific way to enable the estimation of an accurate date. At the same time, having this heterogeneity represented in a CDM allows source-specific data processing to be performed by a common program, which enhances transparency and reproducibility.

The absence of unique standard vocabulary is a key feature of the ConcePTION CDM. Absence of an *a priori* semantic harmonization made CDM deployment faster and more flexible across Europe, because neither translation nor alignment of terms are required. Between May 2020 and June 2021, 13 data sources were converted to the ConcePTION CDM, and successfully contributed to cross-national studies relying on it, including major European projects, such as ACCESS, where background incidence rates of 36 adverse events of special interest that may be used to monitor the benefit-risk profile of COVID-19 vaccines were calculated.[34] The preliminary results of this report contributed to regulatory decisions regarding the COVID-19 AstraZeneca vaccine.[35]

The design of the ConcePTION CDM implies that queries to be executed in distributed analyses must be adapted to the coding system used in participating data sources, which is not the case when dealing with fully harmonized CDMs, such as OMOP CDM.[15] When clinical phenotypes are defined across multiple data sources, relevant local concepts must be identified in close collaboration with DAPs to form *ad hoc* concept sets. The core script incorporating these concept sets can then be executed locally on instances of data sources mapped to the ConcePTION CDM. This flexible approach allows concepts to arise from different data banks, on a study-by-study basis: DAPs can indicate, based on their local knowledge and expertise, whether some concepts should be processed differently when retrieved from specific data banks. The higher the number of DAPs whose expertise must be included, the more time-consuming this process becomes when compared with querying common unique standard vocabularies. However, by supporting a high level of input from DAPs, local expertise may be integrated at each step of data processing, generating results while preserving the level of detail and the quality of the information.

In addition, by capturing the specific nature and origin of routine healthcare data through the meaning and origin variables, the ConcePTION CDM offers the possibility of stratifying analyses by type of data bank or prompt, allowing investigators to quantify the differences between data sources.[31,32] For example, a condition captured through hospital administrative records may differ in terms of clinical meaning from the same condition captured through primary care medical records: a discharge primary diagnosis referring to diabetes in hospital settings foreshadows a patient potentially more complicated to manage, or a more severe condition, than a diabetes diagnosis coded in primary care. On the other hand, if a type of data bank or prompt is missing in a data source, a fixed concept set will be less sensitive, as fewer cases are detectable. For example, consider a data source A, where you have a primary care medical record data bank and a hospital administrative record data bank, and data source B with hospital administrative records only, and consider a fixed concept set "diagnosis of diabetes." Most persons with diabetes will be visiting their general practitioner due to their condition, which will prompt primary care medical records. Therefore, in data source A, this concept set will have high sensitivity. Fewer persons with diabetes are admitted to the hospital due to their condition, therefore fewer records will be prompted in B, hence the same concept set will have lower sensitivity. The impact on clinical results of the type of data included in analyses has been clearly observed in the ACCESS study.[34] During background rate incidence calculation, some data sources were analyzed twice, using different combinations of data banks: Primary Care Medical Record used alone or in combination with Hospital Administrative data. In all data sources, the combination of both data banks provided consistently higher estimates of incidence rates for many acute adverse events of special interest representing acute outcomes (e.g., acute kidney injury), highlighting the importance of the knowledge and the careful selection of the suitable type of data bank and prompt to generate accurate results.

There are limitations to this analysis that need to be highlighted. Interviews to characterize the data sources included in this study were conducted with one DAP per data source. Other DAPs may conceptualize the data sources slightly differently based on their experiences and expertise. Nevertheless, this only underlines the value of utilizing the concepts presented here to help facilitate the thorough and transparent description of real-world data sources for use in studies of medicines. The application of the ConcePTION framework to additional data sources, especially outside of Europe, could further confirm the robustness of the concepts and capability of the CDM to adequately capture the necessary information to answer drug safety and effectiveness questions. However, it is worth mentioning that evidence of the generalizability of this approach has already been brought in the frame of the project Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies (MINERVA)[36]—an activity funded by the European Medicines Agency to define and collect a set of metadata to describe real-world data sources—where it was applied to additional data sources, including a treatment registry and a biobank. The ConcePTION CDM may need continual modifications arising from changes to, or the addition of new information collected by the data banks, reflecting changes in the clinical world. However, our approach allows us to adapt it as needed to respond to future emerging health concerns and to generate real-world evidence in relation to these. Furthermore, the usefulness of the ConcePTION CDM to address concerns outside the context of medicine safety in pregnancy and lactation has already been demonstrated through the completion of the ACCESS study.[34]

This research, conducted in the ConcePTION project, demonstrates the unique features of European data sources and allows their content to be represented in a unified conceptual framework. Based on this, the ConcePTION CDM was developed to facilitate effective and transparent deployment of distributed analytics, taking into

account the European context, and allowing the agile generation of answers to questions about utilization, effectiveness, and safety of medicines in special populations, including during pregnancy and breastfeeding, and, more broadly, in the general population.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

The authors acknowledge John Logie, Miriam Gatt, Janet Sultana, Annarita Armaroli, and Ursula Kirchmayer for their contribution to this paper, and Alice Leblond for her support in the submission process.

FUNDING

The ConcePTION project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 821520. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA.

CONFLICT OF INTEREST

N.H.T., C.D.-P., and R.L. are researchers at Bordeaux PharmacoEpi, an independent research platform of the Bordeaux University and its subsidiary the ADERA SAS, which performs financially supported studies for public and private partners. G.R., G.H., C.B., O.P., and R.G. are employed by ARS, a public health agency that conducts or participates in pharmacoepidemiology studies compliant with the ENCePP Code of Conduct. The budget of ARS is partially sustained by such studies. R.P. received funding for two projects under EMA contract EMA/2017/09/PE/04, which makes use of the ConcePTION common data model and other materials described in this manuscript. V.E. and E.D. are salaried employees by the Department of Clinical Epidemiology, Aarhus University, which is involved in studies with institutional funding from regulators and from various pharmaceutical companies, as research grants to and administered by Aarhus University. None of these studies is related to the current study. T.M.D. has received funding from UK NIHR and previously from Menarini, as well as consultancy fees from Astra Zeneca. K.S. is an employee of the PHARMO Institute for Drug Outcomes Research. This independent research institute performs financially supported studies for government and related healthcare authorities and several pharmaceutical companies. T.S. is an employee of the Leibniz Institute for Prevention Research and Epidemiology – BIPS, an independent, non-profit research institute, which performs among others financially supported studies for government and related healthcare authorities and pharmaceutical companies. A.G.-L., M.G.-S., and M.A. are employees of IDIAPJGol. They are working on other projects funded by pharmaceutical companies in the institution, which are not related to this study and with no personal profit. G.T. has been involved in advisory boards and has received research grants from public and private partners, which are not related to the topic of this paper. M.E.G. and M.C. are employees of, and hold shares in GSK. M.S. has been the principal investigator on an EMA requested post-authorization safety study (Novartis) not related to the topic, and has received research grant and pending grants. All other co-authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

N.H.T., R.P., G.R., C.D., R.G., and Mi.S. wrote the manuscript. R.G. and Mi.S. designed the research. All authors performed the research. All authors analyzed the data.

1. Sedgh, G., Singh, S. & Hussain, R. Intended and unintended pregnancies worldwide in 2012 and recent trends. *Stud. Fam. Plann.* **45**, 301–314 (2014).
2. Lupattelli, A. *et al*. Medication use in pregnancy: a cross-sectional, multinational web-based study. *BMJ Open* **4**, e004365 (2014).
3. Saha, M.R., Ryan, K. & Amir, L.H. Postpartum women's use of medicines and breastfeeding practices: a systematic review. *Int. Breastfeed. J.* **10**, 28 (2015).
4. Mazer-Amirshahi, M., Samiee-Zafarghandy, S., Gray, G. & van den Anker, J.N. Trends in pregnancy labeling and data quality for US-approved pharmaceuticals. *Am. J. Obstet. Gynecol.* **211**, e1–690. e11 (2014).
5. Byrne, J.J., Saucedo, A.M. & Spong, C.Y. Evaluation of drug labels following the 2015 pregnancy and lactation labeling rule. *JAMA Netw. Open* **3**, e2015094 (2020).
6. Kaye, D.K. The moral imperative to approve pregnant women's participation in randomized clinical trials for pregnancy and newborn complications. *Philos. Ethics Humanit. Med.* **14**, 11 (2019).
7. IMI Innovative Medicines Initiative | ConcePTION | Building an ecosystem for better monitoring and communicating of medication safety in pregnancy and breastfeeding: validated and regulatory endorsed workflows for fast, optimised evidence generation. *IMI Innovative Medicines Initiative* at <http://www.imi.europa.eu/projects-results/project-factsheets/conception>.
8. Schneeweiss, S., Brown, J.S., Bate, A., Trifirò, G. & Bartels, D.B. Choosing among common data models for real-world data analyses fit for making decisions about the effectiveness of medical products. *Clin. Pharmacol. Ther.* **107**, 827–833 (2020).
9. Gini, R. *et al*. Different strategies to execute multi-database studies for medicines surveillance in real-world setting: a reflection on the European model. *Clin. Pharmacol. Ther.* **108**, 228–235 (2020).
10. Kent, S. *et al*. Common problems, common data model solutions: evidence generation for health technology assessment. *Pharmacoeconomics* **39**, 275–285 (2021).
11. European MedIcines Agency. A common data model for Europe? - Why? Which? How? - workshop report. 35 (London, United Kingdom, 2018) at <https://www.ema.europa.eu/en/documents/report/common-data-model-europe-why-which-how-workshop-report_en.pdf>.
12. Platt, R. *et al*. The new sentinel network — improving the evidence of medical-product safety. *N. Engl. J. Med.* **361**, 645–647 (2009).
13. Fleurence, R.L. *et al*. Launching PCORnet, a national patient-centered clinical research network. *J. Am. Med. Inform. Assoc.* **21**, 578–582 (2014).
14. Overhage, J.M., Ryan, P.B., Reich, C.G., Hartzema, A.G. & Stang, P.E. Validation of a common data model for active safety surveillance research. *J. Am. Med. Inform. Assoc.* **19**, 54–60 (2012).
15. Hripcsak, G. *et al*. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inform.* **216**, 574–578 (2015).
16. But, A. *et al*. Cancer risk among insulin users: comparing analogues with human insulin in the CARING five-country cohort study. *Diabetologia* **60**, 1691–1703 (2017).
17. Morris, J.K. *et al*. EUROlinkCAT protocol for a European population-based data linkage study investigating the survival, morbidity and education of children with congenital anomalies. *BMJ Open* **11**, e047859 (2021).
18. Valkhoff, V.E. *et al*. Population-based analysis of non-steroidal anti-inflammatory drug use among children in four European countries in the SOS project: what size of data platforms and which study designs do we need to assess safety issues? *BMC Pediatr.* **13**, 192 (2013).
19. IMI Innovative Medicines Initiative | EHDEN | European Health Data and Evidence Network. *IMI Innovative Medicines Initiative* at <http://www.imi.europa.eu/projects-results/project-factsheets/ehden>.

20. OHDSI Uniform Data Representation. *The Book of OHDSI: Observational Health Data Sciences and Informatics* (OHDSI, New York, NY, 2019).
21. Hsieh, H.-F. & Shannon, S.E. Three approaches to qualitative content analysis. *Qual. Health Res.* **15**, 1277–1288 (2005).
22. Elo, S. & Kyngäs, H. The qualitative content analysis process. *J. Adv. Nurs.* **62**, 107–115 (2008).
23. Donnelly, P.D., Erwin, P.C., Fox, D.M. & Grogan, C. Single-payer, multiple-payer, and state-based financing of health care: introduction to the special section. *Am. J. Public Health* **109**, 1482–1483 (2019).
24. European Commission EUROCAT Network at <https://eu-rd-platform.jrc.ec.europa.eu>.
25. Brandt, C.A. *et al*. Metadata-driven creation of data marts from an EAV-modeled clinical research database. *Int. J. Med. Informatics* **65**, 225–241 (2002).
26. Clinical Data Interchange Standards Consortium Glossary at <https://www.cdisc.org/standards/glossary>.
27. CONSIGN study: COVID-19 infection and medicines in pregnancy - a multinational registry based study. *EU PAS register* at <http://www.encepp.eu/encepp/viewResource.htm?id=39439>.
28. Coloma, P.M. *et al*. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol. Drug Saf.* **20**, 1–11 (2011).
29. Trifirò, G. *et al*. Use of azithromycin and risk of ventricular arrhythmia. *Can. Med. Assoc. J.* **189**, E560 (2017).
30. Dieleman, J. *et al*. Guillain-Barré syndrome and adjuvanted pandemic influenza A (H1N1) 2009 vaccine: multinational case-control study in Europe. *BMJ* **343**, d3908 (2011).
31. Gini, R. *et al*. Quantifying outcome misclassification in multi-database studies: the case study of pertussis in the ADVANCE project. *Vaccine* **38**, B56–B64 (2020).
32. Roberto, G. *et al*. Identifying cases of type 2 diabetes in heterogeneous data sources: strategy from the EMIF project. *PLoS One* **11**, e0160648 (2016).
33. Becker, B.F.H. *et al*. CodeMapper: semiautomatic coding of case definitions. A contribution from the ADVANCE project. *Pharmacoepidemiol. Drug Saf.* **26**, 998 (2017).
34. Willame, C. *et al*. *Background Rates of Adverse Events of Special Interest for Monitoring COVID-19 Vaccines* (Zenodo, Geneva, Switzerland, 2021) https://doi.org.https://doi.org/10.5281/zenodo.5255870.
35. European Medicines Agency Signal assessment report on embolic and thrombotic events (SMQ) with COVID-19 Vaccine (ChAdOx1-S [recombinant]) – Vaxzevria (previously COVID-19 Vaccine AstraZeneca) (Other viral vaccines). (2021) at <https://www.ema.europa.eu/en/documents/prac-recommendation/signal-assessment-report-embolic-thrombotic-events-smq-covid-19-vaccine-chadox1-s-recombinant_en.pdf>.
36. Strengthening Use of Real-World Data in Medicines Development: Metadata for Data Discoverability and Study Replicability. *EU PAS Register* at <http://www.encepp.eu/encepp/viewResource.htm?id=39323>.