Pharmacoepidemiology and Drug Safety

WILEY

ORIGINAL ARTICLE OPEN ACCESS

INSIGHT: A Tool for Fit-for-Purpose Evaluation and Quality Assessment of Standardized Observational Data Sources for Real World Evidence on Medicine and Vaccine Safety

Vjola Hoxhaj[1] | Constanza L. Andaur Navarro[1] | Judit Riera-Arnau[1,2] | Roel J. H. J. Elbers[1] | Ema Alsina[1] | Caitlin Dodd[1] | Miriam C. J. M. Sturkenboom[1]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands | [2]Clinical Pharmacology Service, Vall d'Hebron Hospital Universitari, Vall d'Hebron Barcelona Hospital Campus, Universitat Autònoma de Barcelona, Barcelona, Spain

Correspondence: Constanza L. Andaur Navarro (c.l.andaurnavarro@umcutrecht.nl)

Received: 21 May 2024 | **Revised:** 26 November 2024 | **Accepted:** 5 December 2024

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 Programme and EFPIA.

Keywords: common data model | data quality | drug safety | multi-database studies | pregnancy | quality checks | real-world data

ABSTRACT

Purpose: To describe the development of INSIGHT, a real-world data quality tool to assess completeness, consistency, and fitness-for-purpose of observational health data sources.

Methods: We designed a three-level pipeline with data quality assessments (DQAs) to be performed in ConcePTION Common Data Model (CDM) instances. The pipeline has been coded using R.

Results: INSIGHT is an open-source tool that identifies potential data quality issues in CDM-standardized instances through the systematic execution and summary of over 588 configurable DQAs. Level 1 focuses on conformance to the ConcePTION CDM specifications. Level 2 evaluates the temporal plausibility of events and uniqueness of records. Level 3 provides an overview of distributions, outliers, and trends over time to facilitate fit-for-purpose evaluation. Therefore, level 1 and 2 assure a proper data standardization, while level 3 provides information regarding the study population, and potential sub-populations. The DQAs are run locally and assessed centrally by a data quality revisor together with the data access provider's representatives.

Discussion: Data quality is the sum of several internal and external features of the data. While DQAs can provide reassurance about fitness-for-purpose for secondary-use data sources, improvements in data collection are essential to reduce errors and enhance overall data quality for Real World Evidence.

Conclusion: INSIGHT aims to support clinical and regulatory decision-making for medicines and vaccines by evaluating the quality of observational health data sources to support fit for purpose assessment. Assessing and improving data quality will enhance the reliability and quality of the generated evidence.

Study Registration: This research was registered in EU PAS registration with number EU50142.

1 | Introduction

Regulatory agencies and healthcare professionals acknowledge the relevance of Real-World Data (RWD) in informing clinical decisions and shaping public health policies, especially for populations underrepresented in clinical trials, such as pregnant women [1, 2]. Since the 1990s, pharmacoepidemiologists have used RWD and worked in distributed research networks to

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2025 The Author(s). *Pharmacoepidemiology and Drug Safety* published by John Wiley & Sons Ltd.

Summary

- Health data collected via electronic records (real-world data) have shown to be a valuable source of information to generate evidence on effectiveness and safety of medicines and vaccines.
- Due to the heterogenous and diverse nature of real-world data, ensuring data is appropriate for analyses impose challenges on the extension and complexity of data quality assessment.
- We developed INSIGHT, an open-source tool that assesses completeness, consistency, and facilitates fitness-for-purpose evaluation of real-world data sources standardized to ConcePTION common data model.
- INSIGHT integrates principles and data quality dimensions inspired by Sentinel, EUROCAT, and OHDSI.
- INSIGHT with its several checks, incorporates the dimensions outlined in the Data Quality Framework by the European Medical Agency.

quickly generate Real-World Evidence (RWE) for regulatory decisions. In these networks, multiple Data Access Providers (DAPs) collaborate, handling diverse data and employing state-of-the-art methods to produce RWE. This involves the use of a common protocol, a common data model and common analytics with standard tools, which have been improving over the past 15 years [3, 4].

Before conducting any analysis using RWD, it is imperative to ensure the quality of the data by conducting Data Quality Assessments (DQAs) to evaluate whether data sources are fit-for-use and fit-for-purpose prior to study initiation. ConcePTION along with other research networks such as US Food and Drug Administration (FDA) Sentinel, PCORnet, and the Observational Health Data Science and Informatics program (OHDSI) perform DQAs on the data standardized into common data models (CDM) used by these networks for distributed analyses, thus facilitating fitness-for-use and fitness-for-purpose evaluations [3–7].

As part of the IMI-ConcePTION project (www.imi-conception.eu), we have developed INSIGHT, a generic, open-source tool that performs three levels of DQAs on ConcePTION CDM-standardized data instances ensuring the generation of reliable evidence for pharmacoepidemiologic studies using RWD. In this article, we outline step-by-step its design and execution. Furthermore, we present an overview of the DQA workflow, including processes, visualizations, and responsible parties. We show examples of detected inconsistencies and errors, which serve as proof-of-concept, highlighting the tool's effectiveness in identifying potential data quality issues for the generation of reliable RWE.

2 | Methods

2.1 | ConcePTION Common Data Model

As part of a distributed analytics framework, Data Access Providers (DAPs) are required to convert their local data sources into the ConcePTION CDM. This CDM encompasses demographics, observation periods, medicines and vaccines exposure, events, procedures, mother–child linkage, visits, and clinical observations. Once the data sources are syntactically standardized, DQAs are conducted. The current version 2.2 includes 16 tables (Table 1). Additional details can be found elsewhere [7].

TABLE 1 | ConcePTION CDM tables.

Routine healthcare data	VISIT_OCCURRENCE
	EVENTS
	MEDICINES
	PROCEDURES
	VACCINES
	MEDICAL_OBSERVATIONS
Surveillance	EUROCAT
	SURVEY_ID
	SURVEY_OBSERVATIONS
Curated tables[a]	PERSONS
	OBSERVATION_PERIODS
	PERSON_RELATIONSHIPS
	PRODUCTS
	CDM_SOURCE
Metadata	METADATA
	CDM_SOURCE
	INSTANCE
	PRODUCTS

[a]Curated tables contain information that has been previously combine from several data banks to create a single value.

2.2 | Objectives of Data Quality Assessments

Various conceptual data quality frameworks have been published over the last years [8]. For the design of INSIGHT, we relied on the harmonized terminology for data quality dimensions put forth by Khan et al., and the quality indicators established by the US FDA Sentinel initiative [9, 10]. Notably, some quality indicators also overlapped with the ones developed by OHDSI, EUROCAT, and ADVANCE [11–13]. By incorporating these frameworks, we aimed to create a comprehensive and robust data quality verification pipeline for projects using the ConcePTION CDM. For additional details on the design, see protocol (EUPAS50142). In Figure 1, we show the hierarchy of DQAs in INSIGHT followed by a brief description of dimensions covered.

Level 1 provides insight on the completeness (i.e., data availability) and conformance of the ETL (Extract-Transform-Load) process to the ConcePTION CDM specifications, ensuring format, semantic, and structural (relational) coherence [14]. In addition, the plausibility of the data and the uniqueness of records is addressed, maintaining data integrity as well as identifying duplication errors.

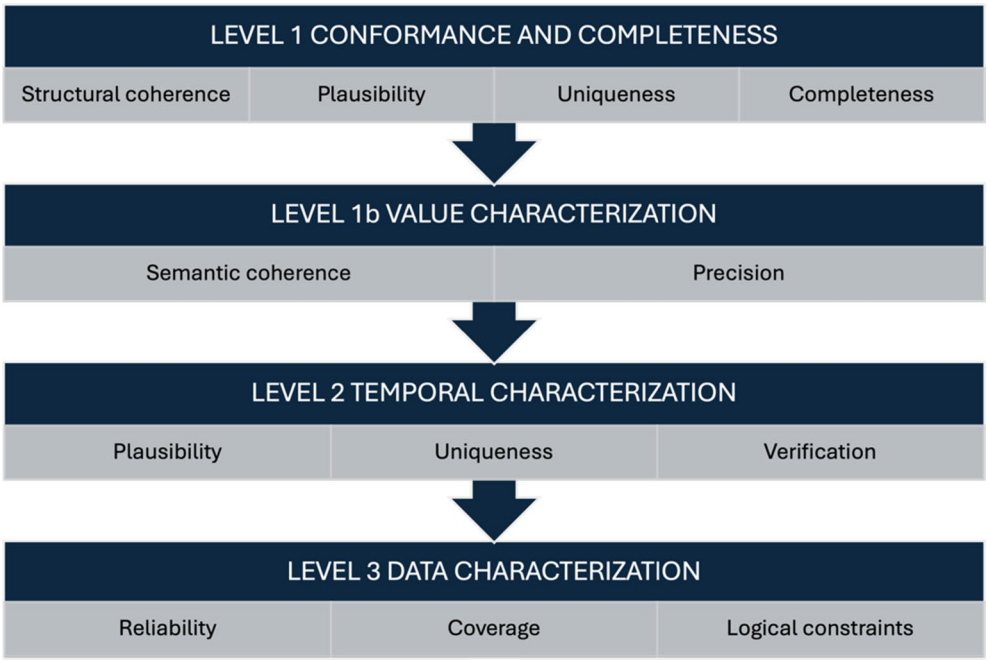


FIGURE 1 | Hierarchy and dimensions of data quality assessment in INSIGHT.

Level 1b provides a list of all extracted and loaded non-date variables across the ConcePTION CDM tables. The objective is to provide DAPs, principal investigators, and programmers, a comprehensive overview of the values of each non-date variable to allow for the construction of variables such as outcome, exposure and covariates. This level aims to enhance the understanding of the data content, facilitating semantic harmonization in the study scripts.

Level 2 provides an overview of the logical relationship and integrity of values within and between variables and tables. Its purpose is to verify the plausibility by examining records that occur outside the recorded person-time, identifying observations related to a person ID that is not present in the PERSONS table, among other similar inconsistencies.

Level 3 provides distributions of population, diagnoses, medicines, vaccines, lifestyle factors, pregnancy, and temporal trends over calendar time for each specific variable. The objective is to allow for inspection of temporal changes. These indicators can be compared between instances and between DAPs, but also against external benchmarks (e.g., national statistic reports, publications) to verify their fitness-for-purpose.

2.3 | Design of Data Quality Indicators

2.3.1 | Level 1: Conformance and Completeness

Level 1 comprises five major steps (Table 2). Step 1 verifies the presence and format of variables ensuring their inclusion and adherence to the designated format. Step 2 verifies missing records within the data instance and provides the extent of missingness overall (i.e., by CDM table and variable), by meaning and by calendar year, when feasible. Step 3 verifies whether date variables comply with the required format and validates day, month, and year values. Steps 1 to 3 are performed simultaneously on all ConcePTION CDM tables, except for the METADATA table. The results are consolidated into a single HTML report.

An intermediate step before Steps 4 to 5 generates a high-level characterization of the ConcePTION CDM instance by extracting information from the CDM_SOURCE and INSTANCE tables, highlighting the key attributes of the data instance. A verification of uniqueness of records is also performed as duplicates check. Step 4 verifies conformance of the ConcePTION CDM tables with the required conventions. These conventions are specific to each CDM table and refer to the table structure and vocabulary rules. Moreover, frequency tables are constructed for categorical variables, providing a summary of values. Step 5 produces the distribution of continuous variables and date variables. For continuous variables, descriptive statistics such as mean, median, interquartile ranges are calculated. Steps 4 and 5 are performed separately for each CDM table. An HTML report is generated for each table, providing a separate overview of the results. All issues identified need to be fixed by updating the ETL design, or when not possible these will be noted. In summary, Level 1 DQA aims to verify format coherence and completeness of each variable across the 15 CDM tables, encompassing a total of 356 individual indicators. The code is available at https://github.com/UMC-Utrecht-RWE/INSIGHT-Level1.

2.3.2 | Level 1b: Value Characterization

Level 1b relies on a successful Level 1 and consists of two steps (Table 3). In this level, various aspects related to the data's semantic coherence and precision are evaluated through 26 indicators (Table 3). Some examples include checking the pharmacological

TABLE 2 | Details on the quality checks involved in level 1: Step 0 to 5.

		HTML report
Step 0: METADATA	Checks the working directory and the consistent presence of all CDM tables and variables, irrespective of uppercase/lowercase and content	Yes
	Checks if all variables are written in lowercase.	
Step 1: Checks the table formatting	Checks if all rows of the .csv files in the working directory contain the correct number of fields	Yes
	Checks if all variables are present irrespective of their content	
	Checks if variables are written in lowercase	
	Checks for presence of mandatory variables as according to the CDM	
	Checks for presence of non-mandatory variables by comparing between the table of interest and the information recorded in the METADATA table	
	Checks presence of vocabularies for specific variables	
	Checks the format for all values and compare to a list of acceptable formats which have been filled out in the METADATA table	
Step 2: Missing data analysis	Tabulates missingness in all variables available in the instance (mandatory and non-mandatory), overall and by calendar year (in the tables that contain a date variable)	
	Provides missing data will be further stratified by data provenance[a] (in the tables that containing this variable)	
	Provides missing data overall will be displayed using bar charts for each CDM table and reported as counts and percentages	
	Provides missing data stratified by data provenance or calendar year will be displayed using line charts for each CDM table and reported as counts and percentages	
	Provides missing data stratified by data provenance and calendar year will be displayed using faceted bar charts for each CDM table and reported as counts and percentages	
Step 3: Date variables	Checks if dates are in the correct format (8 characters)	
	Checks if date variables contain allowable values: Year: 1995(study specific)-present; Month: 01–12; Day: 01–31	
Intermediate step	Provides high-level description of the CDM instance based on the information collected from CDM_SOURCES and INSTANCE tables	Yes
Step 4: Convention and counts of categorical variables	Checks if the table of interest contains any duplicate rows	Yes, per each CDM table
	Checks that all conventions for the table of interest have been adhered to	
	Constructs frequency tables of categorical variables, overall and by calendar year (when the table of interest contains a date variable).	
Step 5: Distribution of continuous variables and dates	For continuous variables mean, median, interquartile range, skewness, and kurtosis will be reported.	Yes, per each CDM table
	Distribution of date variables will be reported as counts of dates overall and by calendar year.	

[a]Data provenance is ascertained from the variable "meaning" present in each of the ConcePTION CDM tables.

level for ATC codes (e.g., fifth or seven) or the unit for treatment duration (i.e., days or weeks). This precision assessment allows to create dictionaries for semantic harmonization so researchers can interpret data uniformly across different DAPs. Level 1b does not generate an HTML report, but a set of csv files for each CDM table. The code is available at https://github.com/UMC-Utrecht-RWE/INSIGHT-Level1b.

2.3.3 | Level 2: Temporal Characterization

Level 2 consists of 8 steps (Table 4) with a total of 57 indicators and verifies temporal plausibility of date variables (Steps 2.1–2.3, 2.5–2.6) and consistency of encounter records across different tables (Steps 2.4, 2.7–2.8). For example, step 2.1 examines if health encounters occurred before an individual's date of birth,

TABLE 3 | Details on the quality checks involved in level 1b.

		HTML report
Step 1	Creates frequency tables for each variable within each CDM table (excluding METADATA, CDM_SOURCE, EUROCAT) by displaying counts for all values present.	No
Step 2	Creates contingency tables for categorical variables within each CDM table (excluding METADATA, CDM_SOURCE, EUROCAT) and shows the distribution of one variable across the categories of other variables.	No

TABLE 4 | Details on the quality checks involved in level 2: Step 1 to 8.

		HTML report
Step 1[a]	Checks dates of events before birth	Yes
Step 2[a]	Checks dates of events after death	Yes
Step 3[a]	Checks dates of events outside observation periods	Yes
Step 4[a]	Checks subjects observed in a CDM table of interest without a corresponding record in the PERSONS table	Yes
Step 5	Checks observations associated with a *visit_occurrence_id* which occur before the *visit_start_date* (optional, if the study cares of visits)	Yes
Step 6	Checks observations associated with a *visit_occurrence_id* which occur after the *visit_end_date* (optional, if the study uses variable visits)	Yes
Step 7	Checks observations associated with a *visit_occurrence_id* for which the associated *person_id* differs from that in the VISIT_OCCURRENCE table (optional)	Yes
Step 8	Checks for subjects indicated in PERSON_RELATIONSHIPS as the parent of a child whose birth date is less than 12 years prior to the recorded birth date of the linkage child (optional, if the study requires mother–child linkage)	Yes

[a]Mandatory steps by default in the pipeline. Other steps could become mandatory based on study requirements.

ensuring the chronological integrity of the data. Similarly, Step 2.4 verifies that person IDs with a record in any of the CDM tables are also listed in the PERSONS table. The code is available at https://github.com/UMC-Utrecht-RWE/INSIGHT-Level2.

2.3.4 | Level 3: Data Characterization

Level 3 creates tables and graphics for study-variables (Table 5) in 9 steps. First the population, its follow-up, and entry and exit patterns are described, allowing for verification of stability in the dynamic population. Subsequently study variables such as medicines, vaccines, diagnoses, pregnancy, lifestyle factors and EUROCAT indicators are created in the study population. Graphics allow to inspect temporal trends, including person-time, prevalence of medicine prescriptions, and incidence of events for the study population and subpopulations of interest. This facilitates the identification of any peaks, drops, or trends in the data and facilitates benchmarking between different instances, DAPs and external data. There are three versions of the output provided, one with no masking procedures applied, the two other versions allow for masking of cell counts. The first is masking all counts lower than five, and the second does not show absolute counts but only ranges. With over 249 indicators, Level 3 allows to deeply

TABLE 5 | Details on the quality checks involved in level 3: Step 1 to 9.

		HTML report
Step 0: Subpopulation availability[a]	Checks presence or absences of subpopulation[a] for iterative performance of level checks	No
Step 1: Study-source population	Provides counts, rates, and distributions of study population characteristics	Yes
Step 2: Dates	Creates the distribution of start-end, follow-up, and birth dates across time	Yes
Step 3: Visits and lifestyle	Provides counts and rates about lifestyle factors and use of healthcare facilities	Yes
Step 4: Medicines	Provides counts and rates across time of medicine use	Yes
Step 5: Vaccines	Provides counts and rates across time of vaccine exposure	Yes
Step 6: Diagnoses	Provides counts and rates across time of events of interest	Yes
Step 7: Pregnancy	Provides counts and rates of pregnant population stratified by status of pregnancy	Yes
Step 8: Population of interest	Provides counts and rates of events, medicine use and vaccine exposure distribution in a specific sub-group (e.g., pregnant women, subjects with a migraine diagnosis etc.)	Yes
Step 9: EUROCAT[b]	Provides Data Quality Indicators on congenital anomalies information in Europe	Yes

[a]Subpopulations refer to subsets of the source population and can be defined by the data provenance (i.e., hospital admission, general practitioner data, etc.).
[b]The number of quality checks and HTML reports will double per the number of subpopulations available on the data instance For DAPs with EUROCAT/EUROmediCAT data sources, steps 1 and 2 are mandatory.

assess whether data is fit for purpose. The code is available at https://github.com/UMC-Utrecht-RWE/INSIGHT-Level3.

2.4 | HTML Report

To ensure that data quality indicators can be inspected, HTML reports were developed containing summary tables and graphs that provide a visual representation of trends and patterns. Within Level 1, the first report lists the presence of the METADATA table, a mandatory table, in the working directory. The report on Step 1 to 3 provides information on the format coherence and completeness. Reports Step 4 to 5 are optional and based on study specific requirements of CDM tables (Table 2). Reports 1–4 in Level 2 are mandatory, whereas reports 5–8 are optional, depending on study requirements (Table 3). Level 3 produces nine reports of which report 1 and 2 are mandatory. The total number of reports depends on the study requirements and the number of subpopulations (Table 5).

2.5 | Responsible Parties

The R-scripts are available on the UMCU-RWE GitHub and maintained by UMC Utrecht. A public frozen version is available for this paper. DAPs are responsible for downloading and running the R scripts locally and addressing issues related to table formatting that results from Level 1 Step 1 to 3 and exceeding the a priori 5% inconsistencies threshold defined in Level 2. If any issues are reported, DAPs must revise the ETL of the data instance and rerun these steps until the required quality is attained or all constraints noted.

The outputs of the Level 1, 1b, 2, and 3 are uploaded to the UMCU Digital Research Environment (DRE) by the DAPs, for centralized review by the data quality revisor and the DAPs [15]. The data quality revisor collaborates with each DAP to address any issues flagged during the revision, as DAPs can provide background information on data collection, population characteristics, national healthcare system, availability of healthcare facilities, regulatory aspects, and expected disease incidence and prevalence. Observations are listed in the data quality approval form (Supporting Information S1). Figure 2 shows the workflow.

2.6 | Approval Form

An approval form must be completed centrally by the DAP and the data quality revisor, with one form per instance and per level check. This form serves as a quality assurance document that systematically and transparently consolidates relevant information from the DQAs. The form is an Excel document with four tabs, one per level check listing the relevant aspects to be assessed per output (Supporting Information S1). We expect to make this form also available via REDCap, an electronic data capture system (www.project-redcap.org).

3 | Illustrative Example

INSIGHT is available as a set of executable R scripts, currently supporting the use of ConcePTION CDM version 2.2 [16–19].

3.1 | Level 1

In Level 1, the key objective is to assess conformance of the ETL'ed. data to the ConcePTION CDM specifications. For instance, if alternative values such as "1" for males and "2" for females are used instead of the designated values (M, F, U, O) for the variable *sex_at_instance_creation* in the PERSONS table, it will be flagged as an error in Steps 1 to 3 report.

Additionally, Level 1 identifies unallowable values in date variables. In Figure 3, we show an example from the EVENTS table, the indicator *error_year* is displayed in blue, highlighting records with dates before 1995 or in the future. Most DAPs have data banks starting after 1995, that is why this cut off was

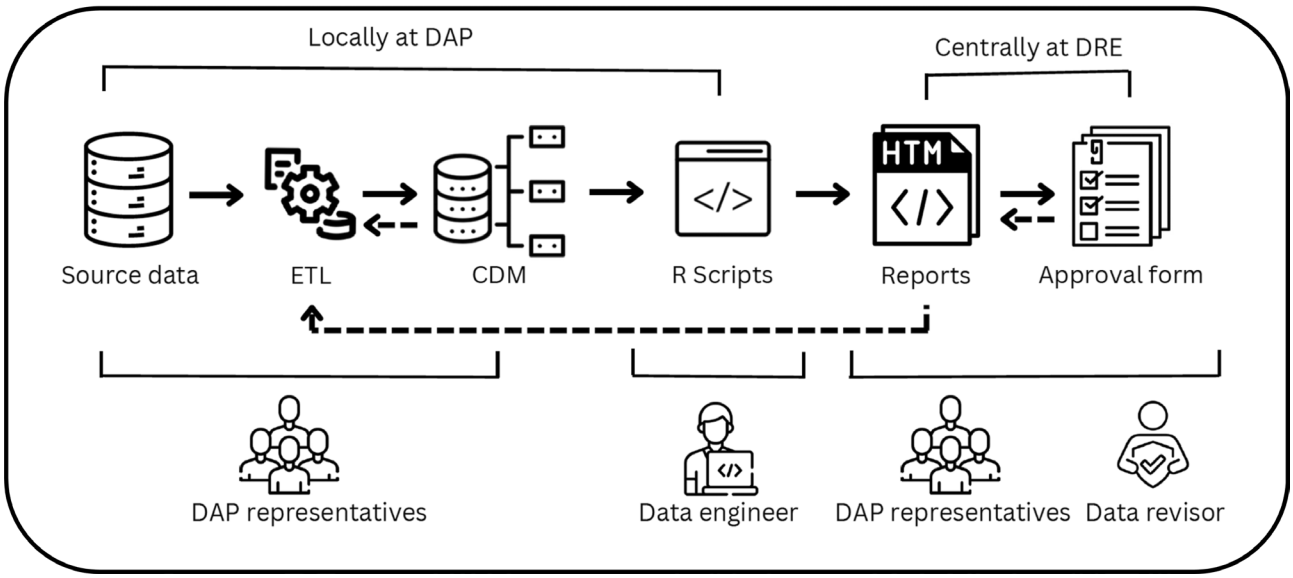


FIGURE 2 | Workflow of INSIGHT.

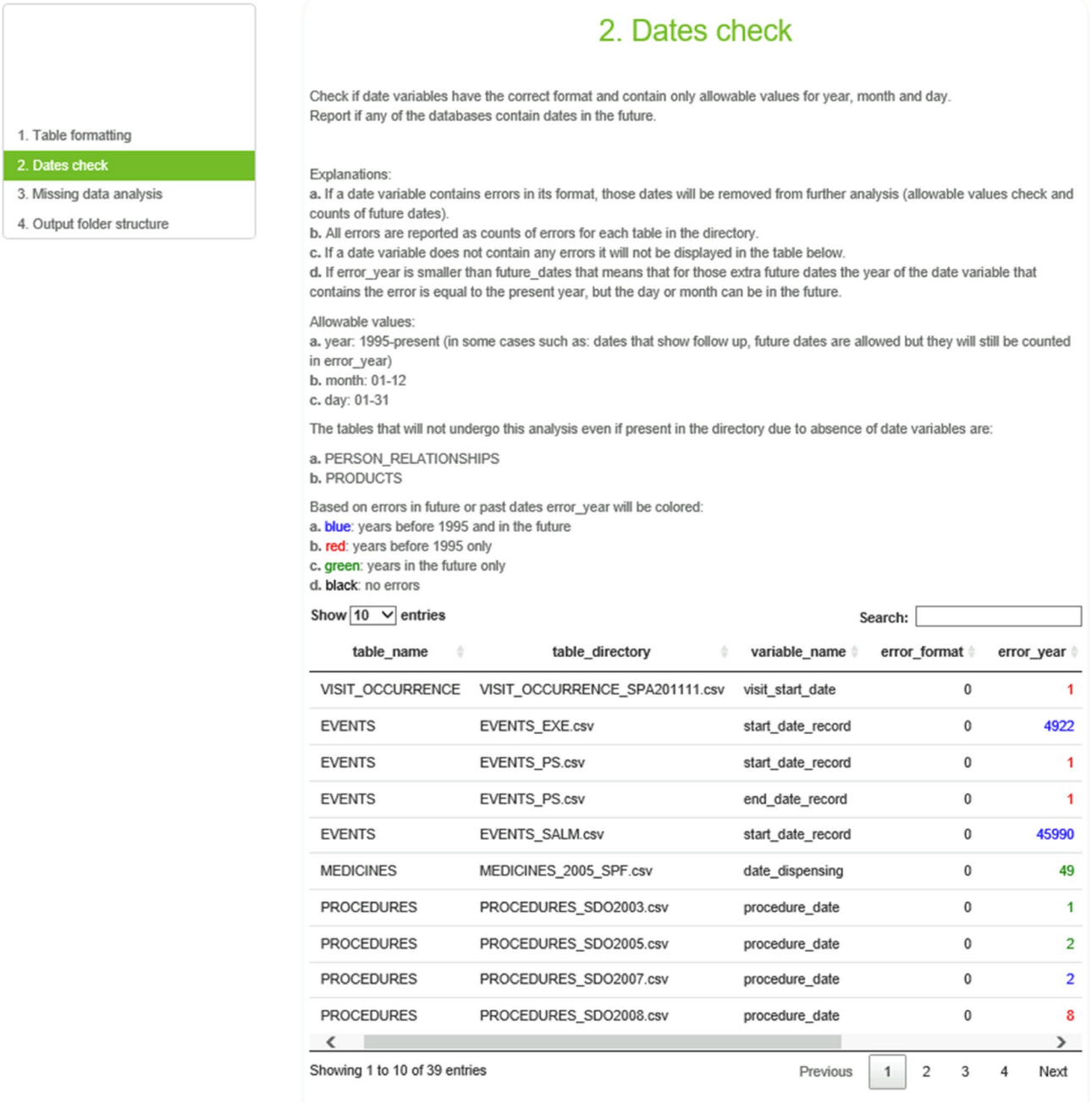


FIGURE 3 | Example of level 1 check.

chosen. The indicator *future_dates* show records with dates in the future in comparison to the script running date.

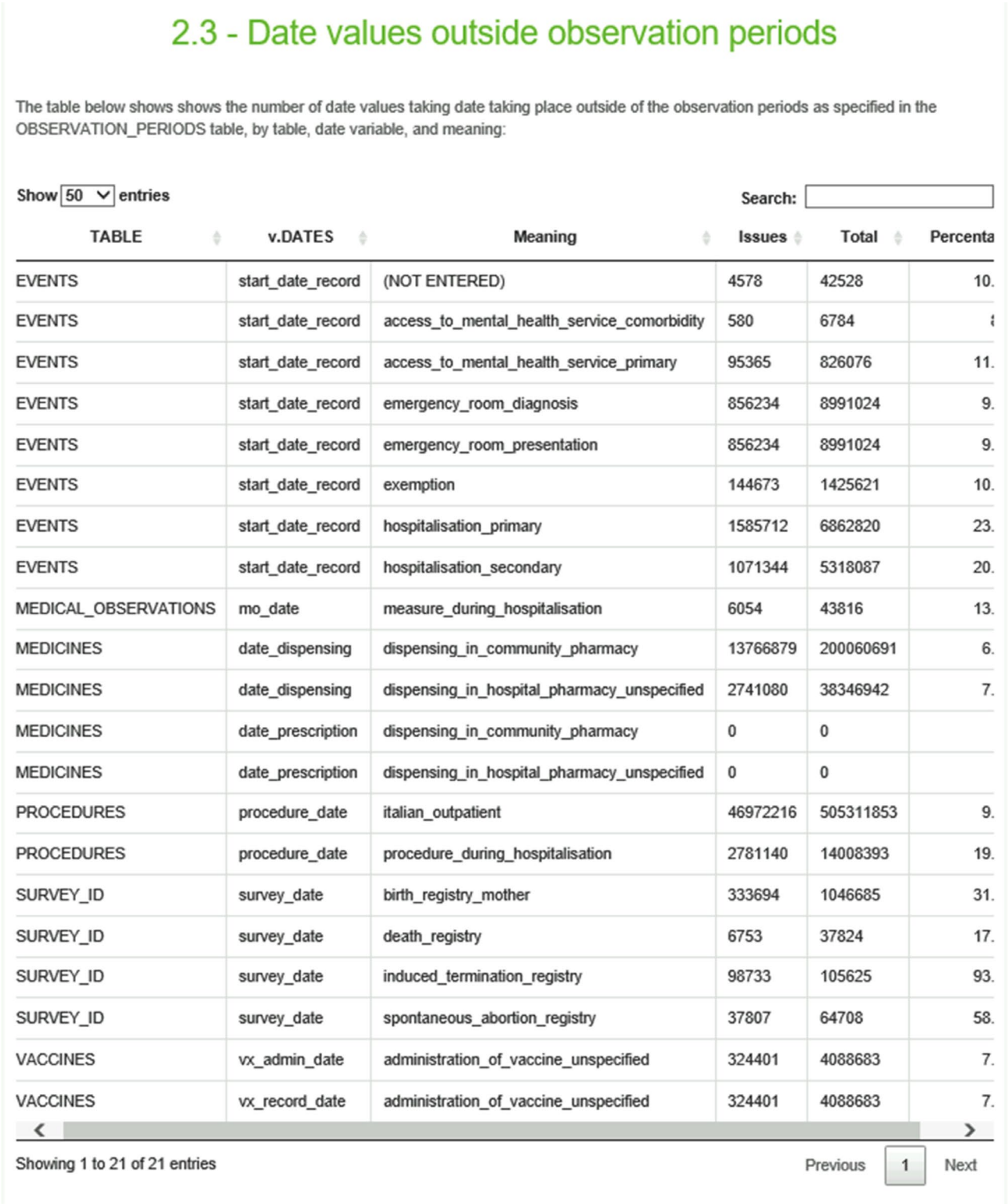
DAPs are expected to resolve these issues by updating their ETL specifications to adhere to the specified rules and requirements, if possible.

3.2 | Level 1b

Level 1b allows users to visualize various aspects of the data, such as the composition of codes in the PROCEDURES table (e.g., number, number with letters, symbols like "$" or "?") and provides insights into the original value definitions for study variables in the MEDICAL_OBSERVATIONS or SURVEY_OBSERVATIONS tables. Additionally, by assessing the content of the ConcePTION CDM tables the user can confirm that each table was fed appropriately.

3.3 | Level 2

Step 2.3 highlights records where dates fall outside the observation period (Figure 4). The construction of this period is a crucial step. Start dates of observation can be obtained from various sources, usually inhabitant registries. However, accessibility to these sources may vary and alternative resources such as first visit to a general practitioner are used. In Figure 4, the high



2.3 - Date values outside observation periods

The table below shows shows the number of date values taking date taking place outside of the observation periods as specified in the OBSERVATION_PERIODS table, by table, date variable, and meaning:

Show 50 entries Search:

TABLE	v.DATES	Meaning	Issues	Total	Percenta
EVENTS	start_date_record	(NOT ENTERED)	4578	42528	10.
EVENTS	start_date_record	access_to_mental_health_service_comorbidity	580	6784	
EVENTS	start_date_record	access_to_mental_health_service_primary	95365	826076	11.
EVENTS	start_date_record	emergency_room_diagnosis	856234	8991024	9.
EVENTS	start_date_record	emergency_room_presentation	856234	8991024	9.
EVENTS	start_date_record	exemption	144673	1425621	10.
EVENTS	start_date_record	hospitalisation_primary	1585712	6862820	23.
EVENTS	start_date_record	hospitalisation_secondary	1071344	5318087	20.
MEDICAL_OBSERVATIONS	mo_date	measure_during_hospitalisation	6054	43816	13.
MEDICINES	date_dispensing	dispensing_in_community_pharmacy	13766879	200060691	6.
MEDICINES	date_dispensing	dispensing_in_hospital_pharmacy_unspecified	2741080	38346942	7.
MEDICINES	date_prescription	dispensing_in_community_pharmacy	0	0	
MEDICINES	date_prescription	dispensing_in_hospital_pharmacy_unspecified	0	0	
PROCEDURES	procedure_date	italian_outpatient	46972216	505311853	9.
PROCEDURES	procedure_date	procedure_during_hospitalisation	2781140	14008393	19.
SURVEY_ID	survey_date	birth_registry_mother	333694	1046685	31.
SURVEY_ID	survey_date	death_registry	6753	37824	17.
SURVEY_ID	survey_date	induced_termination_registry	98733	105625	93.
SURVEY_ID	survey_date	spontaneous_abortion_registry	37807	64708	58.
VACCINES	vx_admin_date	administration_of_vaccine_unspecified	324401	4088683	7.
VACCINES	vx_record_date	administration_of_vaccine_unspecified	324401	4088683	7.

Showing 1 to 21 of 21 entries Previous 1 Next

FIGURE 4 | Example of level 2 check.

percentages of dates outside the observation period indicate that the current observation period fails to capture all relevant events.

Moreover, Level 2 assesses the presence of subjects across the CDM tables. Ideally, all *person_id* in the CDM tables should be in the PERSONS table. An error in step 4 indicates a need to re-evaluate the design of the PERSONS table.

3.4 | Level 3

In Level 3, DQAs become more extensive, as their goal is to facilitate the evaluation of fitness-for-purpose. Figure 5 illustrates a demographic tree indicating a similar distribution between females and males. If the study focuses on females and young males, it is possible that fathers have been included in the study

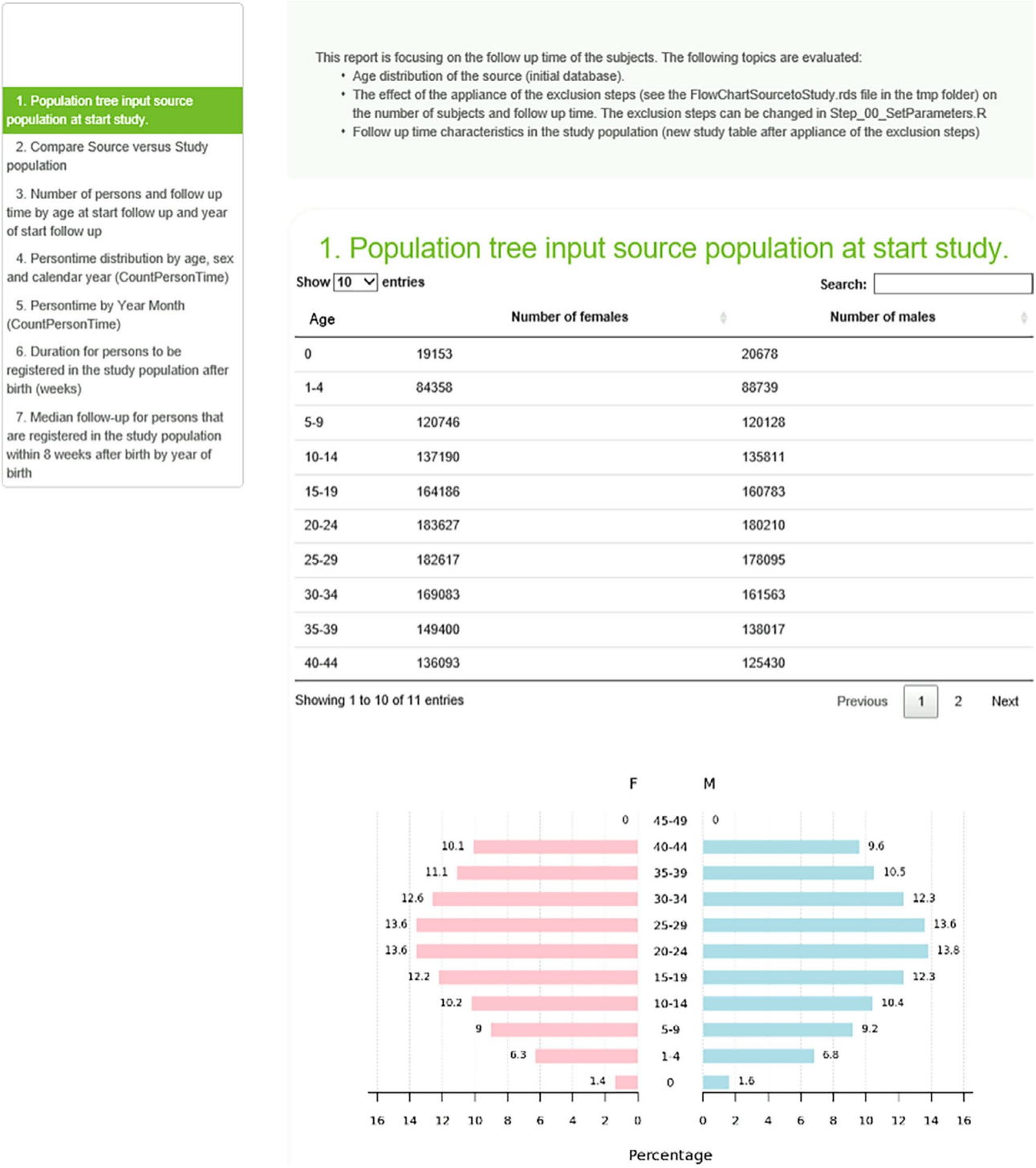


FIGURE 5 | Example of level 3 check.

population, which does not align with the study requirements but is not necessarily an error. Another example is gestational diabetes diagnostic codes in males. This could be explained by recording diagnoses under male children. However, this percentage should be low to reflect the reality of the condition, otherwise it may indicate an error.

In level 3, one of the checks is distribution analysis of start and end dates of follow-up, by year and month. In Figure 6 all records start in January and July of 2004. This pattern indicates a potential ETL specification to filter out records outside the study period of interest. The peak in July suggests imputation due to missing data in date of birth, which has been used as start of observation.

Level 3 plays a crucial role in detecting and reflecting the consequences of technical errors in previous levels. By identifying and addressing these issues, the goal is to improve the overall quality

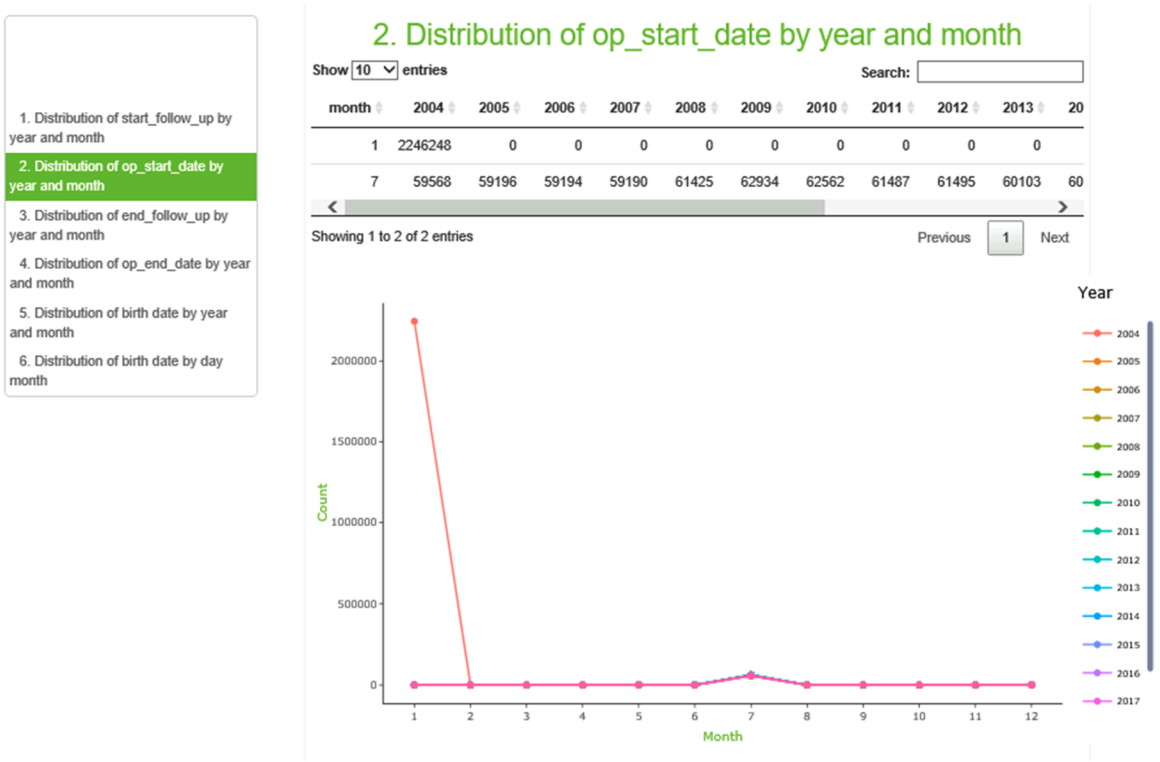


FIGURE 6 | Example of level 3 check. *op_start_date* refers to when individuals start being follow up in a particular databank.

of the data source. INSIGHT is an iterative process of running and updating, until the DAPs and researchers have confidence in the data source's quality.

4 | Discussion

The use of RWD in research and decision-making has gained significant attention from regulatory agencies such as FDA, PMDA, and EMA. They have emphasized the importance of valid and trustworthy RWD for generation of RWE [1, 2]. In this article, we present INSIGHT, a tool to assess fit-for-use (research independent) and fit-for-purpose (research specific) of RWD.

Multi-database studies are becoming increasingly popular [20]. With the abundance of healthcare data available, DAPs can leverage various data sources including routine healthcare data, surveillance systems, inhabitant registries, and more. However, the utility of these granular and diverse data sources relies on their quality. Data quality encompasses several features, including representativeness, precision, and accessibility, among other dimensions. To evaluate quality, EMA and Heads of Medicines Agency have proposed the Data Quality Framework (DQF) for EU medicines regulation which includes five quality dimensions: reliability, extensiveness, coherence, timeliness, and relevance [14]. We have mapped INSIGHT's data quality dimensions to Kahn's framework (right), and the EMA DQF (left) (Supporting Information S2).

It is essential to recognize that data quality is influenced by numerous factors, mostly the health care system, the purpose of data recording and the type of data recorder. INSIGHT DQAs provide a generic and transparent overview of data quality indicators, aiming to identify inconsistencies in the data, which can be categorized into three main types: extraction or convention errors, script errors, and patterns for acknowledgment and documentation. The outputs support the study team to understand the data and facilitate the semantic harmonization. The purpose of DQAs varies ranges from rectifying extraction errors to fine-tuning scripts to suit project requirements. For example, level 3-step 0 provides the opportunity to create sub-population of interest to perform DQAs in parallel to the overall study population, facilitating more detailed fit-for-purpose evaluations. Furthermore, documenting data patterns ensures accurate interpretation and understanding of data nuances and limitations.

4.1 | Previous Data Quality Assessment Tools

Various research networks have developed their own DQA processes with varying levels and indicators. For example, the US FDA Sentinel uses a four-level approach focusing on completeness and validity, accuracy and integrity, consistency of trends, and plausibility [5, 21]. The EUROCAT network, has a comprehensive DQA process, that includes indicators such as case ascertainment, accuracy of diagnoses, data completeness, timeliness, and availability of denominator data [13, 22]. Researchers using the OMOP CDM have access to tools like the White Rabbit and the Data Quality Dashboard (DQD) for assessing data quality at source data and CDM-standardized database [4, 23–25].

A comparison of DQAs across six research networks revealed variations in the number of quality checks, ranging from 875 up

to 3234 [26]. The DQAs processes also varied, including differences in centralization, distributed coordination of DQAs, programing languages, and staff involvement. These differences can be attributed to specific network requirements, analytical focus, as well as the maturity of their DQA models. INSIGHT provides a set of over 588 configurable quality indicators, whose number can increase if study requirements account for vaccines or even duplicate, if assessment for sub-population is necessary, and uses a centralized approach wherein reports containing the quality indicators are shared in DRE for review.

4.2 | Strengths and Limitations

INSIGHT was built during the ConcePTION project but has been made generic and is now adopted by other research networks using the ConcePTION CDM (ConcePTION project, VAC4EU, EU PE & PV network, SIGMA consortium). Our approval system is based on a flexible approval form and relies on study requirements, over predetermined conditions. However, there are limitations to be acknowledged. First, script running time depends on hardware, data size and complexity. Components like the central processing unit, random access memory, and computer data storage help determine process efficiency. Second, the pipeline requires iterative reruns after data issues are fixed until desired quality is achieved or at least, explained. Steps must be executed sequentially to avoid downstream issues. Last, fit-for-purpose evaluation is time and resource-intensive. While summary reports and dashboards aid assessment, automated detection of trends and verification against external benchmarks need to be facilitated further. Automating outlier and pattern detection is an area for future development given the increase complexity and limited skilled human resources. Overall, the DQA workflow is subject to continuous improvements.

The INSIGHT R scripts are openly available through GitHub, along with a training course by VAC4EU and educational videos to encourage other stakeholders to adopt this DQA pipeline and adhere to principles of open science. This tool has been already successfully implemented in EMA tendered risk minimization studies (EUPAS31095, EUPAS21001), COVID vaccines effectiveness study (EUPAS40404, EUPAS42504), Post-Authorization Safety Studies (EUPAS43593, EUPAS44424, EUPAS41725, EUPAS45461), and CONSIGN (EUPAS39226, EUPAS39438, EUPAS40317) [27]. INSIGHT is compliant with the proposed DQF for EU medicines regulation launched in October 2022 [14].

4.3 | Future Research

Interoperability between CDMs and their specific DQAs processes remains a challenge, as each CDM must be used in conjunction with its customized tools. This implies that applying external DQAs tools to an already CDM-standardized data source would require an additional ETL process tailored to that CDM. To address this, we propose to establish standards and guidelines for achieving interoperability across different CDMs. This would facilitate the exchange of DQAs tools and methodologies between research networks and regulatory bodies, streamlining the quality assessment process.

In INSIGHT, we aim to distinguish between ‘errors’ that is, concrete violations to predefined rules and ‘issues’ that is, potential inconsistencies without rule violation. Further enhancement of our tool should include a summary indicator regarding the performance of the ETL process and data instances to allow for continuous improvement on the standardization processes.

5 | Conclusion

In this article, we introduced INSIGHT, a comprehensive pipeline for conducting DQAs of ConcePTION CDM-standardized datasets. The tool aims to facilitate the evaluation of the fitness-for-use and fitness-for-purpose of RWD. By implementing this tool, researchers can assess the adequacy and reliability of the data, ensuring that the evidence generated is robust and reliable. INSIGHT is a valuable tool for enhancing the credibility of RWD in clinical and regulatory decision-making.

5.1 | Plain Language Summary

Research networks use health data collected via electronic records (often called real-world data) to carry on research on the safety of medicines and vaccines once available for prescription. Real-world data is often obtained from diverse data sources, such as hospital records or even administrative registries. Checks on the data are needed to ensure the health information stored is suitable and reliable for research. INSIGHT is an open-source tool that helps (1) ensuring all necessary patient details and medical data for the study are present, (2) making sure the data is consistent across different parts of the record and over time, (3) ensuring the data is formatted and categorized in a standardized way, making it easier to use in research, and (4) facilitating assessment on whether the data accurately reflects the patient's health and treatments. INSIGHT consist of several scripts containing each of these quality checks written in R language (a statistical software) to systematically and rapidly visualize issues or error with the data by using computational power. By using INSIGHT, researchers can have greater confidence that the RWD they use are reliable and useful for conducting studies.

Author Contributions

V.H., C.D., and M.C.J.M.S. started with the initial design; V.H., R.J.H.J.E., E.A., and RvdB programmed the R scripts; V.H. and C.L.A.N. wrote the initial draft of the manuscript. V.H., C.L.A.N., J.R.A., R.J.H.J.E., E.A., and M.C.J.M.S., provided critical feedback and edited the manuscript. The research leading to these results was conducted as part of the ConcePTION consortium. This paper only reflects the personal views of the stated authors.

Acknowledgments

We would like to thank Rutger van den Bor for supporting the programming of INSIGHT Level 2, Claudia Bartolini for testing the

scripts, and Rosa Gini for her expertise and collaboration throughout this project.

Ethics Statement

The authors state that no ethical approval was needed.

Conflicts of Interest

V.H., C.L.A.N., J.R.A., and M.C.J.M.S. are currently salaried employees at University Medical Center Utrecht, which receives institutional research funding from pharmaceutical companies and regulatory agencies and is administered by University Medical Center Utrecht. R.J.H.J.E., C.D., and E.A. were salaried employees of University Medical Center Utrecht at the time this project was performed.

Data Availability Statement

The quality check pipeline for ConcePTION CDM v2.2 instances is available as an open-source R scripts in the following link https://github.com/UMC-Utrecht-RWE. Protocol and Statistical Analysis Plan (SAP) can also be found there. Furthermore, a training course for DAPs and researchers is available upon request via VAC4EU. Researchers can register and complete this course online. Upon successfully passing the final test, they will be awarded certification.

References

1. U.S. Food and Drug Administration, "Framework for FDA's Real-World Evidence Program," 2023 accessed February 1, 2023, https://www.fda.gov/media/120060/download.

2. Heads of Medicines Agency, "European Medicines Agency HMA-EMA Joint Big Data Taskforce Summary Report," accessed February 1, 2023, https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf.

3. L. G. Qualls, T. A. Phillips, B. G. Hammill, et al., "Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®)," *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)* 6, no. 1 (2018): 3, https://doi.org/10.5334/egems.199.

4. C. Blacketer, F. J. Defalco, P. B. Ryan, and P. R. Rijnbeek, "Increasing Trust in Real-World Evidence Through Evaluation of Observational Data Quality," *Journal of the American Medical Informatics Association* 28, no. 10 (2021): 2251–2257, https://doi.org/10.1093/jamia/ocab132.

5. R. Ball, M. Robb, S. Anderson, and G. Dal Pan, "The FDA's Sentinel Initiative-A Comprehensive Approach to Medical Product Surveillance," *Clinical Pharmacology and Therapeutics* 99, no. 3 (2016): 265–268, https://doi.org/10.1002/cpt.320.

6. S. Schneeweiss, J. S. Brown, A. Bate, G. Trifirò, and D. B. Bartels, "Choosing Among Common Data Models for Real-World Data Analyses Fit for Making Decisions About the Effectiveness of Medical Products," *Clinical Pharmacology and Therapeutics* 107, no. 4 (2020): 827–833, https://doi.org/10.1002/cpt.1577.

7. N. H. Thurin, R. Pajouheshnia, G. Roberto, et al., "From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding," *Clinical Pharmacology and Therapeutics* 111, no. 1 (2022): 321–331, https://doi.org/10.1002/cpt.2476.

8. J. Bian, T. Lyu, A. Loiacono, et al., "Assessing the Practice of Data Quality Evaluation in a National Clinical Data Research Network Through a Systematic Scoping Review in the Era of Real-World Data," *Journal of the American Medical Informatics Association* 27, no. 12 (2020): 1999–2010, https://doi.org/10.1093/jamia/ocaa245.

9. M. G. Kahn, T. J. Callahan, J. Barnard, et al., "A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data," *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)* 4, no. 1 (2016): 18, https://doi.org/10.13063/2327-9214.1244.

10. Sentinel Operations Center, "Sentinel Data Quality Assurance Practices," 2017, https://www.sentinelinitiative.org/sites/default/files/data/distributed-database/Sentinel_DataQAPractices_Memo.pdf.

11. M. Sturkenboom, T. Braeye, L. van der Aa, et al., "ADVANCE Database Characterisation and Fit for Purpose Assessment for Multi-Country Studies on the Coverage, Benefits and Risks of Pertussis Vaccinations," *Vaccine* 38 (2020): B8–B21, https://doi.org/10.1016/j.vaccine.2020.01.100.

12. C. Blacketer, E. A. Voss, F. Defalco, et al., "Using the Data Quality Dashboard to Improve the Ehden Network," *Applied Sciences (Switzerland)* 11, no. 24 (2021): 11920, https://doi.org/10.3390/app112411920.

13. M. Loane, H. Dolk, A. Kelly, C. Teljeur, R. Greenlees, and J. Densem, "Paper 4: EUROCAT Statistical Monitoring: Identification and Investigation of Ten Year Trends of Congenital Anomalies in Europe," *Birth Defects Research. Part A, Clinical and Molecular Teratology* 91, no. SUPPL. 1 (2011): S31–S43, https://doi.org/10.1002/bdra.20778.

14. EMA and Heads of Medicines Agency, "Data Quality Framework for EU medicines regulation," accessed February 1, 2023, https://www.ema.europa.eu/en/news/high-quality-data-empower-data-driven-medicines-regulation-european-union.

15. anDREa Consortium, "anDREa," accessed February 1, 2023, www.andrea-cloud.eu.

16. V. Hoxhaj, C. Andaur Navarro, C. Dodd, and M. Sturkenboom, "ConcePTION-INSIGHT Level 1 Checks," accessed February 1, 2023, https://github.com/UMC-Utrecht-RWE/ConcePTION-Level1.

17. R. J. Elbers, A. Cid Royo, V. Hoxhaj, C. Andaur Navarro, and M. Sturkenboom, "ConcePTION-INSIGHT Level 1b Checks," accessed February 1, 2023, https://github.com/UMC-Utrecht-RWE/ConcePTION-Level1b.

18. V. Hoxhaj, R. van den Bor, C. Dodd, C. Andaur Navarro, and M. Sturkenboom, "ConcePTION-INSIGHT Level 2 Checks," accessed February 1, 2023, https://github.com/UMC-Utrecht-RWE/ConcePTION-Level2.

19. V. Hoxhaj, R. J. Elbers, E. Alsina, C. Andaur Navarro, C. Dodd, and M. Sturkenboom, "ConcePTION-INSIGHT Level 3 Checks," accessed February 1, 2023, https://github.com/UMC-Utrecht-RWE/ConcePTION-Level3.

20. A. Pacurariu, K. Plueschke, P. McGettigan, et al., "Electronic Healthcare Databases in Europe: Descriptive Analysis of Characteristics and Potential for Use in Medicines Regulation," *BMJ Open* 8, no. 9 (2018): e023090, https://doi.org/10.1136/bmjopen-2018-023090.

21. R. E. Behrman, J. S. Benner, J. S. Brown, M. McClellan, J. Woodcock, and R. Platt, "Developing the Sentinel System—A National Resource for Evidence Development," *New England Journal of Medicine* 364, no. 6 (2011): 498–499, https://doi.org/10.1056/NEJMp1014427.

22. European Comission, "Data Quality Indicators in EUROCAT," accessed February 1, 2023, https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/DQI-List-of-Data-Quality-Indicators-since-2012.pdf.

23. OHDSI, "Data Quality Dashboard (DQD)," accessed February 1, 2023, https://ohdsi.github.io/DataQualityDashboard/.

24. OHDSI, "CDM OHDSI Common Data Model," accessed February 1, 2023, https://github.com/OHDSI/CommonDataModel/releases.

25. OHDSI, "WhiteRabbit," accessed February 1, 2023, https://github.com/OHDSI/WhiteRabbit.

26. T. J. Callahan, A. E. Bauck, D. Bertoch, et al., "A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks," *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)* 5, no. 1 (2017): 8, https://doi.org/10.13063/2327-9214.1287.

27. E. Maisonneuve, O. de Bruin, E. Hurley, et al., "CONSIG-International Meta-Analysis Protocol & Statistical Analysis Plan (1.1)," Zenodo, https://doi.org/10.5281/zenodo.7112746.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.