# A comprehensive *report* that provides *insights* into the *completeness*, *transparency*, and *quality* of the Extract Transform Load (ETL) process.

*CDM Onboarding R package for data quality assessment.*

Data quality assessment of an observational health data set is an important aspect when deciding whether the data is suitable to answer a selected research question. Accordingly, the OHDSI (Observational Health Data Sciences and Informatics) community has developed a variety of tools to enable this need. Achilles [5] and the Data Quality Dashboard [3] are incorporated into the CDM Onboarding report.

The CDM Onboarding is an R package, developed by the DARWIN EU® Coordination Centre (CC) based on the EHDEN CDM Inspection report. The outcome is a comprehensive report that provides insights into the completeness, transparency, and quality of the Extraction, Transform, and Load (ETL) process step of the OHDSI journey. Additionally, at the DARWIN EU® onboarding process it helps with the evaluation of the new data partner's preparedness to join the DARWIN EU® data network and actively participate in research studies [1]. Building upon the CDM Inspection report foundation from the EHDEN consortium [2], the CDM Onboarding package integrates supplementary checks, enhancing the provision of valuable information. Noteworthy examples of these additional checks are highlighted below.

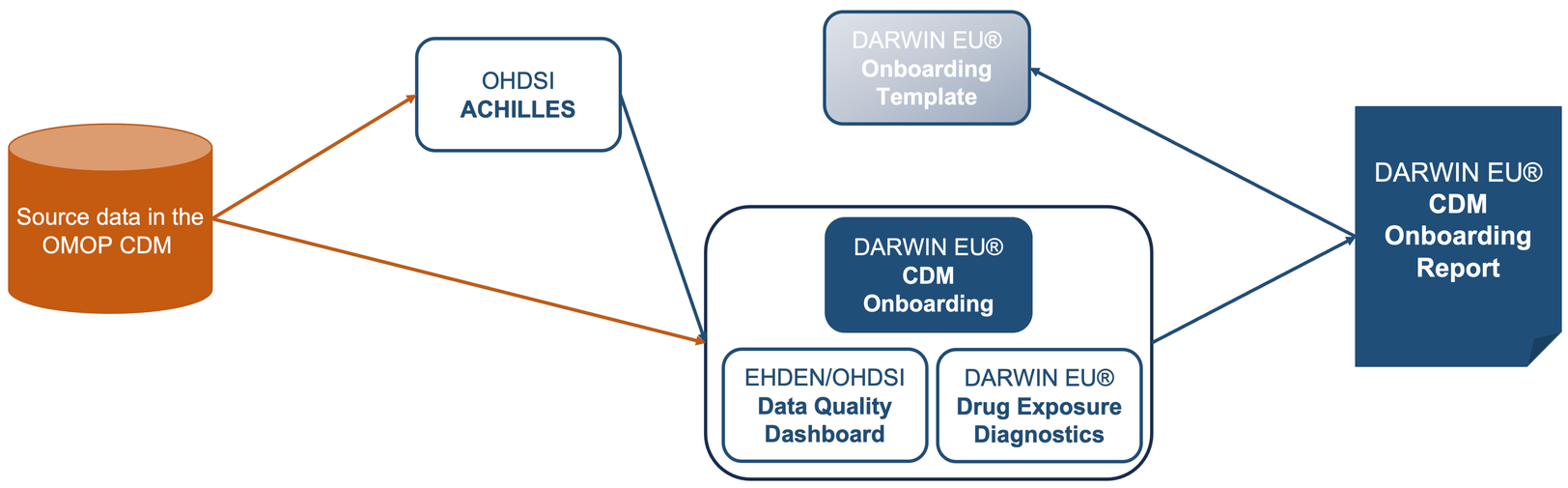



Figure 1: A schematic representation of the input for the CDM Onboarding report, a crucial part of the onboarding and refresh process for the DARWIN EU® Network Operations pillar.

## Methods

### 1 Data Quality Dashboard

The results file extracted by the Data Quality Dashboard [3] on the OMOP CDM instance can now be provided to give an overview of the number of passed and failed checks and will be displayed in the report.

**5. Data Quality Dashboard**

DataQualityDashboard Version: 2.0.0
DataQualityDashboard executed at 2024-05-05 11:52:57 in 8 mins.

**Table 40.** Number of passed, failed and total DQD checks per category. For DQD v2, the checks with status 'NA' are not included.

| Category | Pass | Fail | Total | %Pass |
|---|---|---|---|---|
| Plausibility | 141 | 8 | 2,500 | 5.6% |
| Conformance | 791 | 3 | 1,072 | 73.8% |
| Completeness | 202 | 55 | 464 | 43.5% |
| Total | 3,970 | 66 | 4,036 | 98.4% |

### 2 Drug Exposure Diagnostics

The Drug Exposure Diagnostics is an R package [4], developed by DARWIN EU® CC and it is used to summarize ingredient-specific drug exposure data in the OMOP CDM. As an example, we have selected a set of eleven common ingredients with different ways of administration to run Drug Exposure Diagnostics on. The results are presented in the report as a summary (Table 1) and used upon onboarding for data quality checks and during study feasibility for more in-depth analysis. Indicative quality checks performed:

- Whether the routes correspond to the expectation. For example, Acetaminophen is expected to be given orally, whilst Albuterol is expected to be inhaled.
- Distribution of amount gives an idea of the strengths prescribed. For example, Acetaminophen is expected to be prescribed mainly as 500mg.
- Distribution of quantities, i.e. how many tablets are prescribed at a time. For example, for Albuterol we either expect a high number (puffs) or a low number (inhalers), but this should be consistent.
- Distribution of exposure days. An important quality check here is whether a duration is available, or always 1, indicating the exposure end date always equals the start date.

▼Table 1. Example results for Drug Exposure Diagnostics runduring CdmOnboarding, for two of the eleven ingredients. Distributions are reported as median (q05-q95).

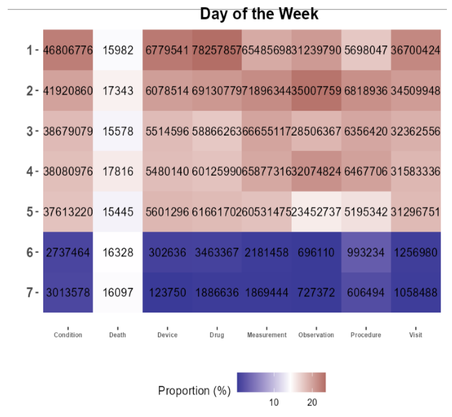
| Ingredient | Number of records | Route (n,%) | Amount distrib. [null or missing] | Quantity distrib. [null or missing] | Exposure days distrib. [null or missing] | Neg. Days n (%) |
|---|---|---|---|---|---|---|
| acetaminophen | 68266 | Oral (65604, 96%);<br>Rectal (2662, 3.9%);<br>Intra-arterial (38, 0%);<br>No matching concept (172, 0.1%) | 500 (0-1000) [7659, 7.7%] | 50 (14-250) [2252, 3.3%] | 36 (1-372) [0, 0%] | 5 (0%) |
| albuterol | 67308 | Respiratory tract (66308, 99.9%);<br>Oral (50, 0.1%);<br>Intra-arterial (11, 0%);<br>No matching concept (479, 0%) | 0.4 (0.2-1.8) [66354, 97.2%] | 1 (1-9) [341, 0.5%] | 1 (1-1) [0, 0%] | 2 (0%) |

### 3 Technical infrastructure

Additional assessments were incorporated to evaluate the technical infrastructure. Extracted information includes the indexes applied on the OMOP CDM instances, as well as those that are missing, which can aid in addressing performance issues. Furthermore, the report now lists all versions of the HADES and DARWIN EU packages, with missing packages highlighted.

### 4 Temporal details

Additional sections with temporal information were added; giving insights into overlapping observation periods, distribution of records over days of the week/month and time ranges of each domain. Figure 2 shows an example of the distribution of records over time. In general, less clinical events happen during the weekend, Death being an exception, and less events happen on the 31st of the month.



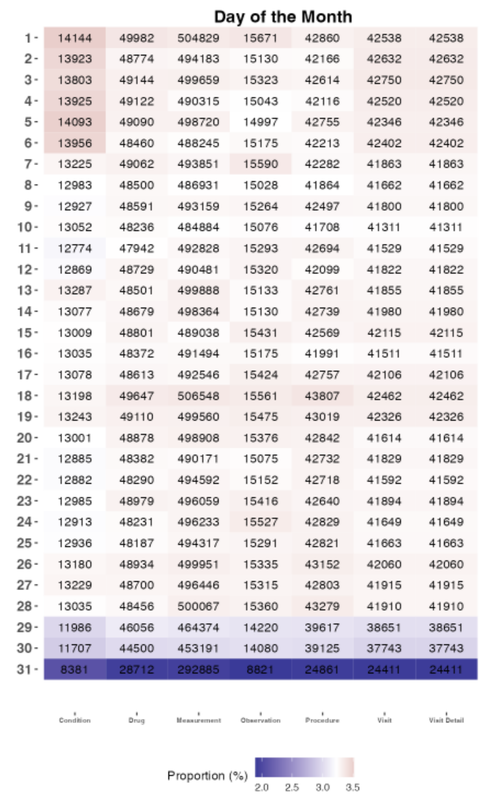




**Figure 2: Example results of the distribution of records over time**

**With these new developments, the DARWIN EU® CC gets further insights into the quality of data converted to the OMOP CDM. This is already showing value in the onboarding of new DARWIN EU® data partners. In the future the plan is to make further developments in the CDM Onboarding package and integrate more R packages like the CDM Connector benchmark.**


**References**
1.https://github.com/darwin-eu/CdmOnboarding
2.https://github.com/EHDEN/cdminspection
3.https://github.com/OHDSI/DataQualityDashboard
4.https://github.com/darwin-eu/DrugExposureDiagnostics
5.https://github.com/OHDSI/Achilles



Anne van Winzum, Maxim Moinat, Sofia Bazakou