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. Achieves [2] 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 [2]. 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.

Methods

5. Data Quality Dashboard

Table 6b. Number of passed, failed, and total DQO checks per category. For DQO < 2, the checks with value NA are not included.

<table>
<thead>
<tr>
<th>Category</th>
<th>Pass</th>
<th>Fail</th>
<th>Total</th>
<th>% Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flexibility</td>
<td>141</td>
<td>8</td>
<td>1,500</td>
<td>90.6%</td>
</tr>
<tr>
<td>Performance</td>
<td>75</td>
<td>17</td>
<td>1,072</td>
<td>70.8%</td>
</tr>
<tr>
<td>Transparancy</td>
<td>275</td>
<td>7</td>
<td>1,857</td>
<td>79.1%</td>
</tr>
<tr>
<td>Completeness</td>
<td>70</td>
<td>8</td>
<td>78</td>
<td>91.5%</td>
</tr>
<tr>
<td>Total</td>
<td>2,570</td>
<td>66</td>
<td>2,636</td>
<td>96.4%</td>
</tr>
</tbody>
</table>

Drug Exposure Diagnostics

The Drug Exposure Diagnostics is an R package [4], developed by DARWIN EU® CC and is used to summarize ingredients-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 presented in the report as a summary (Table 3) and used upon onboarding for data quality checks and during study feasibility for more in-depth analysis. Indicative quality checks performed:

- Whether the route corresponds to the expectation. For example, Albuterol is expected to be inhaled.
- Distribution of amounts gives an idea of the strengths prescribed. For example, Albuterol is expected to be prescribed mainly as 500mg. As an example, we always expect a high number (puffs) or a low number (inhalers), but this should be consistent.

Table 3. Example results for Drug Exposure Diagnostics run during CdmOnboarding, for two of the eleven ingredients. Distribution of exposure days. An important quality check here is whether a duration is available, or expected to be prescribed mainly as 500mg.

Table 3. Example results for Drug Exposure Diagnostics run during CdmOnboarding, for two of the eleven ingredients. Distribution of exposure days. An important quality check here is whether a duration is available, or expected to be prescribed mainly as 500mg.

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/OHDSI/Achilles
5. https://github.com/OHDSI/EHDENCDMInspection

Anne van Winzum, Maxim Moinat, Sofia Bazakou