Background and aim
• Data quality (DQ) assessment considers a mix of intrinsic and contextual factors to assess fitness for research and quality improvement purposes.
• Three intrinsic DQ categories - Conformance, Completeness and Plausibility [1].
• Two DQ assessment contexts - Verification with organizational data and Validation against an accepted gold standard.
• The electronic Practice Based Research Network (ePBRN) data repository stores longitudinal clinician and patient health information, uploaded from information systems of health services in South Western Sydney [2]. The ePBRN source databases are two GP EHRs, each with a different data model. The ePBRN dataset is mapped to the OMOP Common Data Model.

AIM: to evaluate if and how the quality of the ePBRN data was altered by the Extraction-Transform-Load (ETL) process.

Methods
• Cohort: patients prescribed with at least one opioid.
• Create pre-ETL (ePBRN) “drug_exposure” and “person” tables.
• ePBRN tables mapped, using the ETL process, to an OMOP CDM table.
• DQ was measured pre- and post-ETL, using one indicator each from the three categories of the DQ assessment framework.
• Completeness = “Absence of data values at a single moment in time”;
• Conformance = “Data values that identify a single object and not duplicated”;
• Plausibility (Uniqueness) = “Number of unique records identified by a set of attributes, in this case gender, birth_datetime, race, ethnicity, location_id”.

Conclusions
• Data quality is affected because conventions in the source and targets differ.
• Two-thirds (62.52%) of “drug_concept_id” in the CDM contain the value “0”.
• For “Quantity”, the NULL values in pre-ETL (reflecting the completeness 71.23%) are converted into “0” post-ETL (reflecting the conformance 70.69%). The lack of consensus on mapping to “0” or “null” needs to be addressed.
• The same reasons apply to the low “Drug_ID” conformance post-ETL.

Results
• The cohort included 28,457 patients and 2,329,081 drug prescriptions in the pre-ETL ePBRN dataset; and 28,154 patients and 2,201,030 prescriptions in the OMOP CDM dataset.
• The person “uniqueness” was 27990 (98.40 %) pre-ETL and 27704 (98.36%) post-ETL.
• The ETL process duplicated records in drug_exposure table (60% more records post-ETL).

Table 1. Indicators of Conformance and Completeness of “drug_exposure” pre and post ETL

<table>
<thead>
<tr>
<th>Variables</th>
<th>Pre-ETL (n=2329081)</th>
<th>Post-ETL (n=2201030)</th>
<th>Pre-ETL (n=2329081)</th>
<th>Post-ETL (n=2201030)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>person_id</td>
<td>28457 (100.00%)</td>
<td>28154 (100.00%)</td>
<td>28457 (100.00%)</td>
<td>28154 (100.00%)</td>
</tr>
<tr>
<td>gender_concept_id</td>
<td>28435 (99.92%)</td>
<td>28130 (99.99%)</td>
<td>28435 (99.92%)</td>
<td>28130 (99.99%)</td>
</tr>
<tr>
<td>year_of_birth</td>
<td>28140 (99.84%)</td>
<td>28154 (100.00%)</td>
<td>28140 (99.84%)</td>
<td>28154 (100.00%)</td>
</tr>
</tbody>
</table>

Table 2. Indicators of Conformance and Completeness of “persons” pre and post ETL

<table>
<thead>
<tr>
<th>Variables</th>
<th>Pre-ETL (n=2329081)</th>
<th>Post-ETL (n=2201030)</th>
<th>Pre-ETL (n=2329081)</th>
<th>Post-ETL (n=2201030)</th>
</tr>
</thead>
<tbody>
<tr>
<td>race_concept_id</td>
<td>28457 (100.00%)</td>
<td>28154 (100.00%)</td>
<td>28457 (100.00%)</td>
<td>28154 (100.00%)</td>
</tr>
<tr>
<td>ethnicity_concept_id</td>
<td>28457 (100.00%)</td>
<td>28154 (100.00%)</td>
<td>28457 (100.00%)</td>
<td>28154 (100.00%)</td>
</tr>
</tbody>
</table>

References:
1. Kahn M, et al. eGEMs (Generating Evidence & Methods to improve patient outcomes). Vol. 4: Iss. 1, Article 18. DOI: http://dx.doi.org/10.13063/2327-9214.1244

Contact: Professor Teng Liaw slaiaw@unsw.edu.au