Data for impact: Does ETL affect their quality?

Siaw-Teng Liaw PhD FACMI, **Aldir Borelli** BSc, **Guan-Nam Guo** BSc, **Jitendra Jonnagaddala** PhD

WHO Collaborating Centre on eHealth,
School of Public Health & Community Medicine, UNSW Sydney, Australia

WHO Collaborating Centre on eHealth (AUS-135)

Never Stand Still | Medicine

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.
- The low post-ETL completeness for “race_country_id” is due to a lack of CDM concepts for "Australian" or “Aboriginal” or “Torres Strait Islander”.
- Mapping errors can occur due to differing semantics in the two different source databases for the ePBRN data repository e.g. race and ethnicity.

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).

	Conformance		Completeness	
Variables	**Pre-ETL (n=2329081)**	**Post-ETL (n=2201030)**	**Pre-ETL (n=2329081)**	**Post-ETL (n=2201030)**
Patient_UUID	2329081 (100 %)	2201030 (100 %)	1369851 (100 %)	2201030 (100 %)
PROVIDER_ID	2329081 (100 %)	2172497 (98.7 %)	1659190 (71.24 %)	2201030 (100 %)
SCRIPTID	2328665 (99.98 %)	2201030 (100 %)	2329081 (100 %)	2201030 (100 %)
Drug_ID	2328179 (99.96 %)	824845 (37.48 %)	2329081 (100 %)	2201030 (100 %)
Ingredient	2329081 (100 %)	2201030 (100 %)	2328179 (99.96 %)	2201030 (100 %)
Strength	2328392 (99.96 %)	2201030 (100 %)	1640304 (70.43 %)	2201030 (100 %)
Dose	2282736 (97.58 %)	*Sig: 4185039 (100 %)*	1918601 (82.38 %)	*Sig: 4185039 (100 %)*
Frequency	2304587 (98.86 %)		2139811 (91.87 %)	
Quantity	2329024 (99.99 %)	1555826 (70.69 %)	1659101 (71.23 %)	2201030 (100 %)
Date	2328938 (99.99 %)	2201030 (100 %)	2329081 (100 %)	2201030 (100 %)

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

	Conformance		Completeness	
Variables	**pre-ETL (n=28457)**	**post-ETL (n=28154)**	**pre-ETL (n=28457)**	**post-ETL (n=28154)**
person_id	28457 (100.00%)	28154 (100.00%)	28457 (100.00%)	28154 (100.00%)
gender_concept_id	28435 (99.92%)	28130 (99.99%)	28435 (99.99%)	28130 (99.91%)
year_of_birth	28140 (99.84%)	28154 (100.00%)	28457 (100.00%)	28154 (100.00%)
month_of_birth	28457 (100.00%)	28154 (100.00%)	28457 (100.00%)	28154 (100.00%)
day_of_birth	28457 (100.00%)	28154 (100.00%)	28457 (100.00%)	28154 (100.00%)
birth_datetime	28457 (100.00%)	28154 (100.00%)	28457 (100.00%)	28154 (100.00%)
location_id	27719 (99.64%)	27538 (99.70%)	28424 (99.98%)	28117 (99.87%)
race_concept_id	28457 (100.00%)	28154 (100.00%)	20333 (71.45%)	186 (0.66%)
ethnicity_concept_id	28457 (100.00%)	28154 (100.00%)	20333 (71.45%)	20109 (71.43%)

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

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
2. Liaw ST, et al. *Proc Am Med Inform Assoc Ann Symp* 2011, pp 785-94. Washington DC

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