Title: *Themis Part 2 – The ETL Data Quality Conversion Initiative*

PRESENTERS: Meghan Pettine
Melanie Philofsky

INTRO

- **There are many model specifications and conventions for the source data to OMOP CDM transformation**
- **Many of these rules are not currently tested**
- **This leads to questions regarding the adherence to CDM specifications and Themis conventions**

METHODS

- **Reviewed each model specification and convention for testability**
- **Wrote a verifiable SQL test for each specification and convention**
- **Classified it in Kahn's Data Quality Framework**

Kahn's Data Quality Framework

VERIFICATION	VALIDATION
CONFORMANCE	
VALUE CONFORMANCE	
RELATIONAL CONFORMANCE	
COMPUTATIONAL CONFORMANCE	
COMPLETENESS	
PLAUSIBILITY	
UNIQUENESS PLAUSIBILITY	
ATEMPORAL PLAUSIBILITY	
TEMPORAL PLAUSIBILITY	

Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw S-T, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. 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) [Internet]. 2016 Sep 11 [cited 2016 Sep 12];4(1).

ETL conventions need to be both human understandable and computer executable to establish data quality checks that can be applied consistently across the OHDSI community

Take a picture to download the full paper

RESULTS

- **Zip Code Checks, NPI Code Checks, Distribution of Lab Values (Conformance, Validation)**
- **The Visit during which an event occurred is recorded through reference to the Visit Occurrence Table (Relational Conformance, Verification)**
- **Valid concepts belong to the correct domain (Value Conformance, Verification)**
- **The drug_concept_id in the Drug Era table should only contain concepts with concept_class = 'Ingredient'. The ingredient is derived from the Drug Concepts in the Drug Exposure table that are aggregated into the Drug Era record (Computational Conformance, Verification)**
- **Distribution of records by domain (Completeness, Validation)**
- **Fields that are required, what percentage are populated versus mapped to 0 (Completeness, Verification)**
- **Percentage of IP, OP, and ER Visits, Distribution of Concept Prevalence (Plausibility, Validation)**
- **Start dates should be less than end dates (Temporal Plausibility, Verification)**
- **Visit end dates are mandatory, when not available in the source they will be derived (Atemporal Plausibility, Verification)**
- **Care Site is unique combination of location_id and place_of_service_source_value (Uniqueness Plausibility, Verification)**

OHDSI