**Title**: Preparing Hospital Electronic Medical Record Data for Research: Essential Data Validation

Schneider T1,2, Punjani T1, Silva JM1,3, Garcia MC1,4, Telford V1, Mbuagbaw L3,5, Gibson P6, Holbrook AM1,3,7

­1 Clinical Pharmacology Research, Research Institute of St Joes Hamilton, Hamilton, ON

2 eHealth Master’s Program, McMaster University, Hamilton, ON

3 Department of Health Research Methods, Evidence and Impact (HEI), McMaster University, Hamilton, ON

4 Temerty Faculty of Medicine, University of Toronto, Toronto, ON

5 Biostatistics Unit, Father Sean O'Sullivan Research Centre, St Joseph's Healthcare, Hamilton, ON

6 Division of Haematology/Oncology, McMaster Children's Hospital, Hamilton, ON

7 Division of Clinical Pharmacology and Toxicology, McMaster University, Hamilton, ON

Presented at the Research and Education Poster Session, St. Joseph’s Healthcare Hamilton, Hamilton, Canada, November 9, 2023.

**Background**

Electronic medical records (EMR) contain an abundance of data that can be extracted for

research, but the quality of the data is crucial for resulting statistical analysis and machine learning. There are data quality frameworks in the literature based on EMR data, however, there is a paucity of data validation methods to ensure high quality EMR data for research. We aim to validate major components of the Epic-Dovetale (EpicD) EMR to estimate data quality using the QT-prolonging medications (QTPMeds) and major adverse cardiac events (MACE) project as the prototype research.

**Approach**

An Entity Relationship Diagram was developed to understand the location and relationships of key tables. Computational data validation (comparing EpicD data extractions to Slicer Dicer and CIHI DAD counts) and manual data validation (comparing EpicD to manual chart review) were applied with adjustments frequently required. The validation statistics applied were plain agreement and those for diagnostic accuracy (positive predictive value (PPV), negative predictive value (NPV), sensitivity (Sn), specificity (Sp)).

**Results**

5 main universal data themes have been validated. 1) Demographics includes the number of eligible patients in the cohort, age, sex, gender, admission, discharge, and transfers. 70,079 patients were extracted from EpicD, compared to 70,170 from Slicer Dicer (99.9% agreement). 2) Exposures include the administration of known, possible, and conditional QTPMeds. Excluding operating room medications, 338 charts reviewed for known QTPMeds resulted in a PPV of 100%, NPV 86%, Sn 95%, and Sp 100%. 3) Outcomes are represented by MACE: a composite of ventricular arrhythmias, syncope, and death. Manual review of 338 charts resulted in a PPV of 93%, NPV 98%, Sn 99%, and Sp 81%. 4) Potential confounders include lab results, electrocardiogram and telemetry monitoring, cardiac comorbidities, and drug interactions. 736 lab results were validated for 100% agreement. 5) Timestamping included precise times for admission, medication administrations, and outcomes occurring after exposure. 56 charts reviewed for medication administration timing had 100% agreement.

**Conclusion**

EMR data validation is an essential, exacting process with its own methodologic quality principles. Once data extraction is perfected, EpicD data for key research themes shows a high degree of agreement and diagnostic accuracy.