Scalable, Shareable, and Computable Clinical Knowledge for AI-Based Processing of Hospital-Based Nursing Data (SC2K)

Motivation: Data recorded by nurses are the most voluminous of hospital data yet are poorly understood by non-nurses, leading to misinterpretations of key data such as the timing of care interventions a patient actually received. Misinterpretations contribute to data quality issues when developing AI tools that use nursing data.

Project Alignment with Assistant Secretary for Technology Policy (ASTP) Leading Edge Acceleration Projects (LEAP) in Health IT 2024 Special Emphasis Notice (SEN): Area 1: Develop innovative ways to improve healthcare-data quality to support responsible development of AI tools in healthcare.

- Nurses are considered the most trusted profession. The SC2K study will enable efficiencies for AI developers in healthcare (industry and academic research) to use the vast amount of available - but poorly understood - nursing data in their AI models to promote health and wellbeing at scale.
- Study products will increase efficiencies in building AI tools by gathering, validating, and openly sharing key nursing facts that are necessary to know when using and interpreting nursing data in AI models.
- These facts will become knowledge graphs integrated directly into AI models to improve efficiency for AI developers and to minimize data quality issues and the need for additional costly requirements gathering with nurses.
- In alignment with the Make American Healthy Again (MAHA) report and focus on diabetes, the first clinical scenario addressed by this study is glucose management to inform more accurate and efficient AI tools.

Accomplishments to Date: We developed a computational model to differentiate missing data from missed care and defined, with expert nurses', the clinical scenario of glucose management in acute and critical care to validate this model. The model includes a key variable of minimally acceptable safe, quality nursing care, which is used for computational validations. We completed 10 sessions with 19 nurses; iterative analyses are underway for knowledge graph generation.

Significant Findings to Report: Our team defined heuristics for working with nursing EHR data which stipulate *Do Not Assume that*: 1) missing data equals missed care as one field in the EHR is not an accurate representation of the totality of care delivered, 2) protocols should be followed exactly as the logic states as many patients have multiple problems and protocol exceptions do occur, 3) all data capture is equal as some data are captured because policy or requirements mandate them while others are recorded voluntarily because the nurse identified them as clinically relevant and important, 4) you understand the process that caused the data capture nor the temporal sequence of how data capture related to actual care processes as a nursing intervention may include an action or the absence of an action and a nursing assessment may be an original assessment or a reassessment post-intervention, which may all occur in the same hour, 5) you know the clinical procedures and protocols that apply to your population as protocols evolve over time and may differ by unit, clinician, and institution, 6) the same EHR used across institutions has the same configurations and settings as configurations can vary widely and are highly dynamic, 7) all values are reliable as measures can differ in accuracy, precision, and calibration and required fields that include "hard stops" may have variable data quality, 8) all structured values make sense clinically as structured fields are not the same as clinical concepts mapped to standardized clinical terminology, 9) a SQL query retrieved all the values needed for your modeling task as nursing flowsheets are tricky - their complexity and volume make it easy to miss fields, 10) the value of narrative data based on its length nor assume that low frequency fields are of low value as nurses use short narrative comments to highlight significance and convey clinical context, 11) that information will be reliably captured in the same fields over time as the EHR is a living, breathing entity, and 12) that nursing data are statistically consistent over time, across different units, or across different institutions, nor assume that these data were extracted completely, correctly, or consistently, rather data patterns change as care evolves. Our knowledge graphs include actual nursing practice patterns such as identifying nurses' clinical judgements that appropriately require variation from protocols to ensure safe care in specific and common circumstances. For example, a nurse may wait to administer insulin until after a meal if a patient is not fully alert or has a low appetite in order to prevent a critical hypoglycemic event. This variability influences and can be detected in AI models by analyzing temporal patterns across different groups.

Dissemination to Date: Varkhedi V, Cato K, Albers D, Tiase V, Joshi S, Thate J, Connell K, Hull W, Finnegan A, Rossetti S. Translating Nursing Data into Computational Metrics: An Evaluation Guideline for Inpatient Intravenous and Subcutaneous Insulin Management. Paper Presentation at AMIA Annual Symposium, Atlanta, November 15-19 2025.

Upcoming Activities: We continue to validate our computational model with additional clinical scenarios and use cases across our clinical data sets and build knowledge graphs for integration into AI models.

This project is supported by the Assistant Secretary for Technology Policy (ASTP) of the U.S. Department of Health and Human Services (HHS) under 90AX0042/01-02, Scalable, Shareable, and Computable Clinical Knowledge for AI-Based Processing of Hospital-Based Nursing Data, $998,903. This information or content and conclusions are those of the author and should not be construed as the official position or policy of, nor should any endorsements be inferred by ASTP, HHS, or the U.S. Government. *PI Contact: Sarah Rossetti, RN, PhD,* sac2125@cumc.columbia.edu