# Real world evidence using measurement values for patient -level prediction models: a feasibility study

**Xiaoyu Wang[1,2], Jenna Reps[1,3], Anthony Sena[1,3], James P. Gilbert[1], Marc A Suchard[4,5]**

**[1]Janssen Research and Development, Titusville, NJ**
**[2] Statistical Science Department, Duke University, Durham, NC**
**[3] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands**
**[4] VA Informatics and Computing Infrastructure, US Department of Veterans Affairs, Salt Lake City, UT**
**[5] Department of Biostatistics, University of California, Los Angeles, CA**

## Abstract

The integration of measurements into patient-level prediction models using observational healthcare data has the potential to enhance real world evidence (RWE) for patients and providers. The Observational Health Data Sciences and Informatics (OHDSI) patient-level prediction (PLP) framework and observational medical outcomes partnership (OMOP) common data model (CDM) enables researchers to develop patient-level prediction models for various prediction tasks using large observational healthcare data. This research investigates the feasibility of this integrating measurement values into patient-level prediction models to specifically forecast stroke instances 1 to 365 days following atrial fibrillation, despite obstacles such as non-standardization and data sparseness. Benchmarks were established using LASSO and GBM models that account for variables such as age groups, sex, drugs, and conditions. These models were then compared with alternative models that incorporate 21 standardized measurements with mean value and Bayesian models imputations. The study's methodology delves into the distribution of unit types and the frequency of unrecorded units for each measurement appearing in at least 5% of the target population. This examination highlights the limitations of traditional regression tools in the OHDSI tool-stack and conventional imputation methods at the health observational data scale, suggesting Bayesian inference as a potential alternative.

Several large insurance claims and electronic healthcare record (EHR) data sets were evaluated, and findings point to Optum EHR as the most promising dataset for constructing prediction models, identifying 21 measurements recorded in at least 75% of the Optum EHR target population. Measurements are also recorded with different units and for some measurements, the unit is unknown. These results imply that while challenges such as manual standardization and missing data management exist, the inclusion of these elements in prediction models is, indeed a feasible endeavor. The research has broader implications for improving patient-level prediction models using observational healthcare data and paves the way for future measurement feasibility research. Further studies would examine the impact of imputing missing values on model performance, especially within the OMOP CDM dataset, given its superior measurement coverage.