Incorporating measurement values into patient-level prediction with missing entries: 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

Background

The OHDSI PatientLevelPrediction framework, tools and OMOP common data model enable researchers to develop patient-level prediction models for various prediction tasks using large observational healthcare data [1]. Numerous models have been developed [2-3] using the OHDSI tools and standardizations. The PatientLevelPrediction framework uses standardized features that are engineered using one-hot encoding based on whether a patient had a certain medical code recorded in the prior x-days relative to index. Models developed using these features have often performed well, but performance may be improved by including suitable measurements.

Measurements are often difficult to include into prediction models developed using large observational healthcare data. The main issues limiting the inclusion of measurements are: 1) the measurements are not standardized in the OMOP common data model (measurements can be recorded with different units and sometimes with a unit unknown) and 2) it is common to have sparsely recorded measurements due to the data being observational (causing missing data issues).

Standardizing certain measurements may be possible if a researcher manually defines how to convert different units into a standard unit; this needs to be done on a per-measurement basis. In addition, many current regression tools implemented in the OHDSI tool-stack are unable to directly include missing data and imputation methods are often unsuitable at health observational data scale. Fortunately, Bayesian inference [4] coherently admits simultaneous modeling of missing values. In this paper, we perform an initial feasibility study into including measurements into models developed using large observational healthcare data.

Methods

The initial stage of this study is to provide information as to the feasibility for using measurements from data conforming to the OMOP CDM, independent of how such imputed measures may impact models.

We will determine what measurements are feasible for the prediction tasks of interest by finding measurements that are recorded in the year prior to index for 5% or more patients in the target population (patients with pharmaceutically treated depression indexed at start of treatment). This will be done across five databases mapped to the OMOP common data model: MDCD, MDCR, CCAE, Optum SES and Optum EHR.

As the measurements can be recorded with different units, for each measurement occurring for >= 5% of the target population, we will also investigate the unit type distribution and how often no unit is recorded. This will provide information about how feasible it is to standardize the measurements.

Results & Discussion

Are there common measurements that are recorded across databases?
Table 1 shows the per database counts of measurements which are included for different percentages of the total target population of patients. Optum EHR has the best coverage of measurements for patients with five measurements being recorded for 95% or more of patients. However, coverage in the claims databases is not high overall. Due to few measurements being recorded for >= 50% of patients in the claims data, there is an insufficient number of measurements that are recorded sufficiently across more than one of the databases investigated. However, 38 measurements were recorded in at least 5% of the target population in all 5 data sources such as blood glucose, lipase, and iron measurements [1].

How complex is standardizing the measurement units?
Figure 1 shows the units used to record body weight in Optum EHR and illustrates issues with measurement units being non-standard in the OMOP CDM. The majority of body weight measurements in Optum EHR are valid (kg, pound or ounce), but ~5% had no unit. For body weight this means we may lose 5% of the measurements due to the unit being unclear. Researchers using body weight in Optum EHR must first standardize weight by converting pounds/ounces to kg. However, for the top 21 recorded measurements in Optum EHR, Table 2 shows the measurements are often mostly standardized.

Can we identify a set of measurements to include in prediction models?
Table 2 provides additional information on the twenty-one measurements that were recorded for >=75% of the study population in Optum EHR. The recommended units are shown in column 'Units'. It may be possible to include these 21 measurements, standardized to the recommended unit, into models developed using Optum EHR.

Conclusion
In this paper we performed a preliminary investigation into the feasibility of incorporating measurements into prediction models developed using large observational healthcare databases.

Across the five datasets investigated we found that claims data have few measurements recorded for >=50% of the target population investigated. This limits the number of measurements that are recorded sufficiently across multiple datasets. Consequently, if measurements are included into prediction models, it may be difficult to perform external validation. We also observed that measurements are often recorded with different units and for some measurements, the unit is unknown. Therefore, inclusion of measurements into prediction models requires manual standardization of units and measurements with missing units or infeasible values may be excluded. This will further increase missingness.

Our feasibility study highlighted Optum EHR as being the most suitable dataset investigated to use to develop prediction models using measurements. In future work it may be possible to include the 21 measurement that occurred for >=75% of the target population.

---

[1] Please see the published study site at https://github.com/ohdsi-studies/PlpMeasurmentFeasability

In future work we will further evaluate the feasibility of imputing such missing values will impact the performance of prediction models. This work will initially focus on the use of Optum EHR as all other databases lack significant coverage for more than a small percentage of subjects in the example target population.

Data Source	5%	10%	25%	50%	75%	95%	100%
Optum EHR	265	163	88	49	21	5	0
Optum SES	193	112	50	4	0	0	0
CCAE	101	29	9	3	0	0	0
MDCD	90	43	13	3	0	0	0
MDCR	109	45	8	2	0	0	0

Table 1. Number of databases with measurements taken for at least x% of patients in the target population (patients treated for newly diagnosed depression).

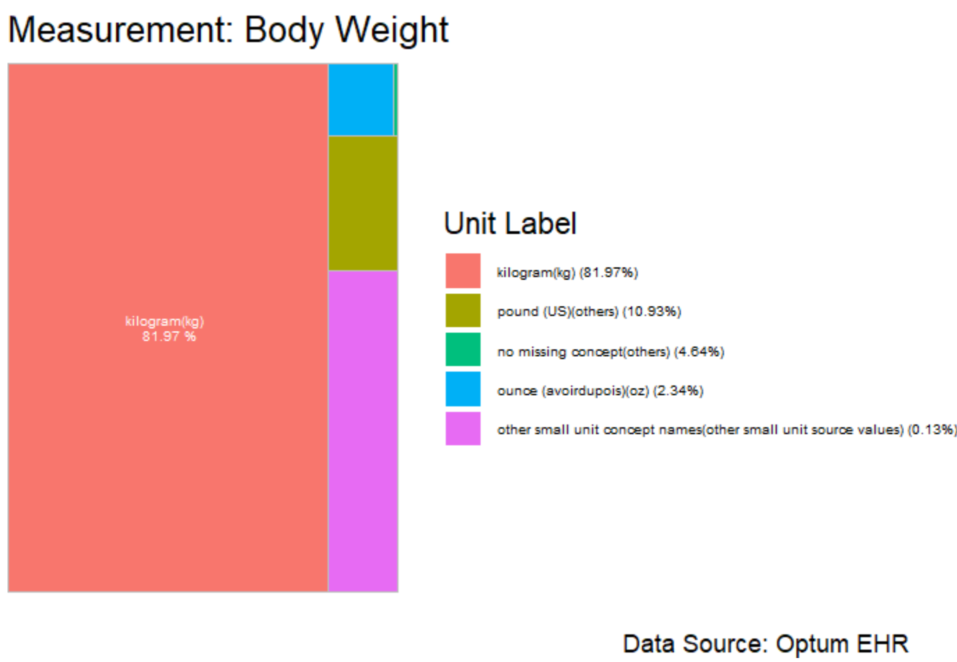

Figure 1. Frequency of units for measurement of body weight in Optum EHR. Units that don’t map to standard measure concepts were merged into “No matching concept”.

Measurement	Units	Unit Source Values	Percent Coverage
Blood urea nitrogen measurement	milligram per deciliter	mg/dl	67.95%
Body height	centimeter, inch (US)	cm, in	90.97%
Body mass index (BMI) [Ratio]	kilogram per square meter	kg/m2	13.76%
Body temperature	degree Celsius	deg c	99.98%
Body weight	kilogram, pound (US)	kg, lb	92.78%
Calcium.ionized/Calcium.total corrected for albumin in Blood	milligram per deciliter	mg/dl	99.31%
Chloride [Moles/volume] in Saliva (oral fluid)	millimole per liter	mmol/l	99.52%
Cotinine/Creatinine [Mass Ratio] in Urine	milligram per deciliter	mg/dl	99.53%
Diastolic blood pressure	millimeter mercury column	mm Hg	84.59%
Erythrocytes [#/volume] in Blood	million per microliter	x10^6/ul	84.26%
Glucose [Mass/volume] in Serum or Plasma	milligram per deciliter	mg/dl	81.38%
Hematocrit [Volume Fraction] of Blood	percent	%	73.37%
Hemoglobin [Mass/volume] in Blood	gram per deciliter	g/dl	59.81%
Leukocytes [#/volume] in Blood	thousand per microliter	x10^3/ul	83.19%
MCV [Entitic volume]	femtoliter	fl	76.9%
Oxygen [Partial pressure] in Blood	percent	%	99.09%
Penicillin G potassium [Mass] of Dose	millimole per liter	mmol/l	99.42%
Pulse intensity of Unspecified artery palpation	counts per minute	bpm	100%
Respiratory rate	counts per minute, counts per minute	breaths/min, bpm	94.77%
Sodium [Moles/volume] in Saliva (oral fluid)	millimole per liter	mmol/l	99.53%
Systolic blood pressure	millimeter mercury column	mm Hg	100%

Table 2. Measurement concepts and dominant units found with for at least 75% of patients for target population in Optum EHR. The standard unit percentage refers to the total coverage that map to vocabulary concepts (and can therefore be mapped to a common unit).

References

1. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association. 2018 Aug;25(8):969-75.
2. Johnston SS, Morton JM, Kalsekar I, Ammann EM, Hsiao CW, Reps J. Using machine learning applied to real-world healthcare data for predictive analytics: an applied example in bariatric surgery. Value in Health. 2019 May 1;22(5):580-6.
3. Wang Q, Reps JM, Kostka KF, Ryan PB, Zou Y, Voss EA, Rijnbeek PR, Chen R, Rao GA, Morgan Stewart H, Williams AE. Development and validation of a prognostic model predicting symptomatic hemorrhagic transformation in acute ischemic stroke at scale in the OHDSI network. PLoS One. 2020 Jan 7;15(1):e0226718.
4. Nishimura A, Suchard MA. Prior-Preconditioned Conjugate Gradient Method for Accelerated Gibbs Sampling in “Large n, Large p” Bayesian Sparse Regression. Journal of the American Statistical Association. 2022 May 6:1-4.