# Finding a constrained number of predictor phenotypes for multiple outcome prediction

**Jenna M Reps[1], Jenna Wong[2], Egill A. Fridgeirsson[3], Chungsoo Kim[4], Luis H. John[3], Ross D. Williams[3], Patrick Ryan[1]**

**[1] Janssen Research and Development, Raritan, New Jersey, United States, [2] Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA., [3] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands; [4] Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea**

## Background

Patient-level prediction can help personalize medical treatment. Websites such as MDCalc (www.mdcalc.com) provide an easy way to implement various prediction models using online forms. Each model on the website is often a simple score-based model with 5-15 predictors. However, there is often little overlap in the predictors used across models except basic predictors such as age and sex. For example, CHA2DS2-VASc[1], HAS-BLED[2], ORBIT[3], Dual Antiplatelet Therapy (DAPT) score[4], Wells' criteria for pulmonary embolism [5] and APACHE II score [6] all contain less than 15 predictors individually, but to apply all six models would require 40 predictors. Three of these models are for a related prediction task. If hundreds of simple models were independently trained, it is likely that hundreds or thousands of predictors would be required if a patient wanted to know their risk for all outcomes. This prompts the question: Is it possible to accurately predict many outcomes using a constrained number of universally informative predictors? If so, a website could be created that asks users to fill out a single form specifying the patient's values to the constrained number of predictors and the website would return personalized predictions for hundreds or thousands of outcomes.

In this paper we propose a novel methodology for identifying a constrained number of predictors and then fit models using the constrained predictors to compare performance of these models to the worst-case scenario (models with age and sex only predictors) and best-case scenario (no constraint on the number of candidate predictors).

## Methods

*OMOP CDM Databases*

In this study we used eight databases mapped to the OMOP CDM, see Appendix A Table 1. Three databases were used to learn the constrained set of predictors and develop models using these predictors, three databases were used to just learn the predictors and two databases were used to just develop the models using the predictors. The model development only databases are a fair test of whether the predictors, learned on different data, do well in predicting the outcomes of interest.

*Finding a constrained set of candidate predictors*

We identified OMOP drug or condition concepts (concept_ids) that are associated with many incident outcomes across many target populations. We calculated the standardized mean difference (SMD) [7] of how frequently a medical concept was recorded in the year prior to index (target population start date) for patients in a target population who develop the outcome within 1-year vs patients in a target population who did not develop the outcome within 1-year. This was done across 65,664 combinations of 64 target cohorts for different new drug users, 171 outcomes and 6 databases (MDCR, MDCD, CCAE,

JMDC, Germany and Australia). We then aggregated the SMD values by counting how often the SMD was greater than 0.1 ("#-SMD-significant") and ordered the medical concepts by decreasing value of #-SMD-significant. A clinician then reviewed the list of medical concepts starting at the top, to identify concepts that corresponded to specific medical concepts and annotated the document stating the broad medical category. The top 1500 concepts were reviewed. We then created phenotypes for each medical topic identified from the top 1500 medical concepts.

*Validating the constrained set of candidate predictors*

To validate the constrained set of predictor phenotypes from the top 1500 concepts, we trained models to predict first occurrence of five outcomes: seizure, fracture, gastrointestinal (GI) bleed, diarrhea, and insomnia within 1 year of initial major depression disorder diagnosis for patients given antidepressant treatment within 30 days and observed in the data for at least 1 year prior to index.

We develop models for the five prediction tasks across five databases (MDCR, MDCD, CCAE, Optum EHR and Optum SES) using the PatientLevelPrediction framework [8] with four different model designs:

1. Constrained LR: Logistic regression with LASSO regularization using the constrained set of predictors plus age/sex.
2. Constrained GBM: Gradient boosting machine using the constrained set of predictors plus age/sex.
3. Worst-case LR: Logistic regression with LASSO regularization using age/sex only predictors.
4. Best-case LR: Logistic regression with LASSO regularization using all conditions/drugs concept ids recorded for at least one patient in the target population prior to index plus age/sex.

Models were developed using the standard PatientLevelPrediction process [8], with 75% of labelled data used to learn the model with 3-fold cross validation to pick the optimal hyper-parameter and 25% of the labelled data used to internally validate the models. In addition, we used cross-database validation to externally validate each model for the same task in the other four databases.

Performance was estimated using area under the receiver operating characteristic curve (AUROC), which is a measure that determines how well the models rank patients in order of risk. A value of 1 means perfect ranking and a value of 0.5 means random ranking.

Note: when predicting seizure (or GI bleed), the constrained predictors contained history of seizure (or history of GI bleed). However, as we predicted first occurrence of the seizure (or GI bleed), it should result in the seizure (or GI bleed) predictor having a value of 0 for all patients in the target population for that task.

**Results**

*The constrained set of candidate predictors*

In total we calculated the #-SMD-significant for a total of 28,741 medical concepts. Ordering the concepts by descending value of #-SMD-significant, the top 1500 concepts were reviewed, and 52 medical topics were identified. Concepts that were too broad to have a clear medical condition, such as '4329041 Pain', were ignored. However, concepts such as "443784 Vascular disorder" were annotated as "Peripheral vascular disease/ coronary artery disease". We then created phenotypes for each of the 52 identified medical topics. Some topics, such as 'antibiotic use', were separated into multiple phenotypes

corresponding to the different antibiotic families. This resulted in our final set of 67 phenotype predictors plus age/sex, see Table 1 in Appendix B.

*Validating the constrained set of candidate predictors*

The internal and external AUROC results for the five prediction tasks per model design are displayed in figure 1. The constrained LR and GBM models performed similarly across the prediction tasks. For the outcomes: seizure, fracture and gastrointestinal bleed, the constrained models performed almost as well as the best-case LR that was able to select from thousands of candidate predictors. For diarrhea the best-case LR consistently achieved a higher AUROC than the constrained models but the constrained models' AUROCs were closer to the best-case LR AUROCs compared to the worst-case LR AUROCs. For insomnia, the constrained models performed midway between the base-case and worst-case LRs' AUROCs. However, insomnia was also the hardest outcome to predict as the best-case LRs' performances were between 0.6-0.7 AUROC. On average, the constrained LR/GBM models had a 4%/3.5% decrease in AUROC compared to the best-case LR across the validations and tasks investigated.

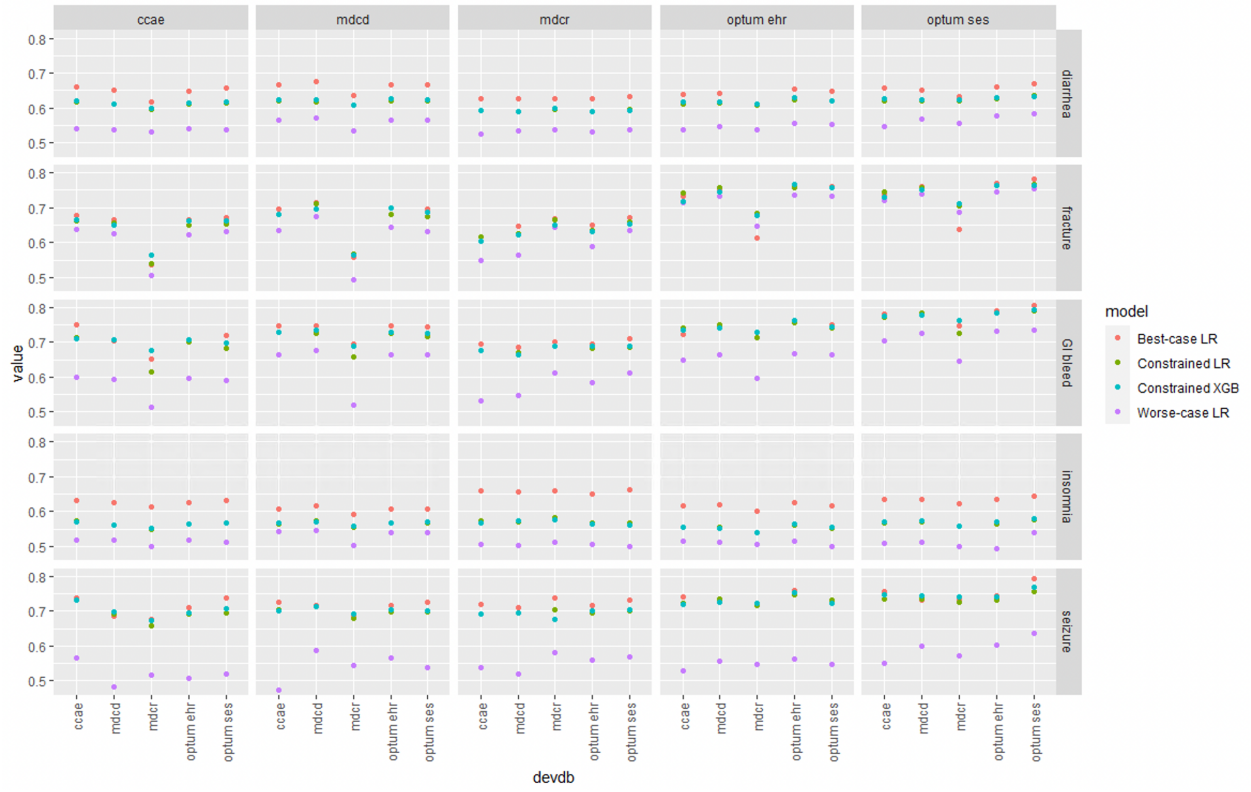


**Figure 1. Internal and external AUROC performance across five prediction tasks for the constrained LR, constrained GBM, best-case LR and worst-case LR. The x-axis is the development database, the facet-grid columns are the external validation databases, and the facet-grid rows are the prediction tasks.**

## Conclusion

We propose a novel method to identify a constrained set of predictors that can be used to predict multiple outcomes accurately. Our results show that models developed using a constrained number of predictors (~67 + age/sex) shared across prediction tasks can result in models that, in most cases, perform similarly

to models trained using thousands of candidate predictors. The results are promising and suggest it may be possible to develop a website where users can answer ~67 questions and see their future risks for hundreds or thousands of outcomes.

In this study we selected an arbitrary number of predictors in the constrained predictor set and in future work it would be interesting to determine whether there is an optimal number of predictors to use. In addition, there are alternative methods that could be implemented to find the constrained set of predictors and future work could investigate alternative methods. Finally, we only validated the predictor set across five outcomes and a single target population. In future work it would be informative to see whether the results hold when more prediction tasks and databases are included.

## References


1. Lip GY, Nieuwlaat R, Pisters R, et al. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-72.
2. Zhu W, He W, Guo L, et al. The HAS-BLED Score for Predicting Major Bleeding Risk in Anticoagulated Patients With Atrial Fibrillation: A Systematic Review and Meta-analysis. Clin Cardiol. 2015;38(9):555-61.
3. O'Brien EC, et al. The ORBIT bleeding score: a simple bedside score to assess bleeding risk in atrial fibrillation. *European Heart Journal*. 2015;36(46):3258–3264.
4. Mihatov N, Secemsky EA, Kereiakes DJ, et al. Utility of the dual antiplatelet therapy score to guide antiplatelet therapy: A systematic review and meta-analysis. Catheter Cardiovasc Interv. 202;97(4):569-578.
5. Zhang NJ, Rameau P, Julemis M, et al. Automated Pulmonary Embolism Risk Assessment Using the Wells Criteria: Validation Study. JMIR Form Res. 2022;6(2):e32230.
6. Akavipat P, Thinkhamrop J, Thinkhamrop B, Sriraj W. ACUTE PHYSIOLOGY AND CHRONIC HEALTH EVALUATION (APACHE) II SCORE - THE CLINICAL PREDICTOR IN NEUROSURGICAL INTENSIVE CARE UNIT. Acta Clin Croat. 2019;58(1):50-56.
7. Zhang XHD. A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics*. 2007;89(4):552–61.
8. Reps, JM., et al., Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *JAMIA*. 2018;25(8):969–975.


**APPENDIX A**

| Country | Time Period | Use in Study | Name (Abbreviation) | Description |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| USA | 2000-2022 | Predictors and Model development | The IBM® MarketScan® Commercial Database (CCAE) | Includes health insurance claims across the continuum of care (e.g., inpatient, outpatient, outpatient pharmacy, carve-out behavioral healthcare) as well as enrollment data from large employers and health plans across the United States who provide private healthcare coverage for employees, their spouses, and dependents. |
| USA | 2000-2022 | Predictors and Model development | The IBM® MarketScan® Medicare Supplemental Database (MDCR) | Represents the health services of retirees in the United States with Medicare supplemental coverage through employer-sponsored plans. |
| USA | 2006-2022 | Predictors and Model development | The IBM® MarketScan® Multi-State Medicaid Database (MDCD) | Reflects the healthcare service use of individuals covered by Medicaid programs in numerous geographically dispersed states. The database contains the pooled healthcare experience of Medicaid enrollees, covered under fee-for-service and managed care plans. |
| USA | 2000-2022 | Model development | Optum's Clinformatics® Data Mart (Optum CDM) | Derived from a database of administrative health claims for members of large commercial and Medicare Advantage health plans. The database includes data over a 14-year period (1/2007 through 12/2021). |
| USA | 2007-2022 | Model development | Optum's longitudinal EHR repository (Optum EHR) | Derived from dozens of healthcare provider organizations in the United States, that include more 57 contributing sources and 111K sites of care. |
| Japan | 2005-2022 | Predictors | JMDC | Consists of data from more than 250 Health Insurance Associations covering workers aged less than 75 and their dependents. JMDC data includes data on membership status of the insured people and claims data provided by insurers under contract. Claims data are derived from monthly claims issued by clinics, hospitals and community pharmacies. The size of JMDC population is about 10% of people in the whole of Japan. |

| | | | | |
|---|---|---|---|---|
| Germany | 2012-2022 | Predictors | IQVIA®Disease Analyzer Germany (Germany) | A longitudinal patient database providing anonymized information from continuing physician and patient interaction on consultations, diagnoses and treatment within Primary Care. It contains a data from approximately 2,500 office-based doctors in Germany. |
| Australia | 2017-2022 | Predictors | IQVIA®LPD in Australia (Australia) | A longitudinal patient database providing anonymized information from continuing physician and patient interaction on consultations, diagnoses and treatment within Primary Care. Data are delivered by 900 office-based doctors in Australia. |

**APPENDIX B**

*Small set of predictors*

*Table 1- The table containing the constrained set of predictors*

| **Predictor** |
|---|
| Acetaminophen prescription |
| Alcoholism |
| Anemia |
| Angina |
| Antibiotic use (separated by family) |
| Antiepileptics (pain) |
| Anxiety |
| Osteoarthritis |
| Aspirin |
| Asthma |
| Atrial fibrillation |
| Hormonal contraceptives |
| Cancer |
| Acute kidney injury |
| Chronic kidney disease |
| Congestive heart failure |
| Chronic obstructive pulmonary disorder (COPD) |

| |
|---|
| Coronary artery disease |
| Depression |
| Diabetes type 1 |
| Diabetes type 2 |
| Deep vein thrombosis (DVT) |
| Dyspnea |
| Edema |
| Gastroesophageal reflux disease (GERD) |
| Gastrointestinal (GI) bleed |
| Heart valve disorder |
| Hepatitis |
| Hyperlipidemia |
| Hypertension |
| Hypothyroidism |
| Inflammatory bowel disorder (IBD) |
| Inpatient visit |
| Low back pain |
| Neuropathy |
| Obesity |
| Opioids |
| Osteoporosis |
| Peripheral vascular disease |
| Pneumonia |
| Psychotic disorder |
| Respiratory failure |
| Rheumatoid arthritis |
| Seizure |
| Sepsis |
| Skin ulcer |
| Sleep apnea |
| Smoking |
| Steroids |
| Hemorrhagic stroke |

| |
|---|
| Non-hemorrhagic stroke |
| Urinary tract infection (UTI) |

*Existing Models and predictors*

CHA2DS2-VASc uses 7 predictors (age, sex, CHF, Hypertension, Stroke/TIA/throcomembolism, Vascilar disease)

HAS-BLED uses 9 (Hypertension, renal disese, liver disease, stroke, prior major bleed, labile INR, age, medications that cause bleeds, alcohol useage)

ORBIT uses 5 (sex, age, bleeding history, GFR, treatment with antiplatelet agents)

Dual Antiplatelet Therapy (DAPT) score uses 9: age, smoking, diabetes, current MI, prior MI, PACLitaxel-eluting stend, stend diagmeter, CHF and Vein graft stent)

Wells' citeria for pulmonary embolism uses 6: DVT, Likely PE diagnosis, Heart rate, immobilization, pevious PE, hemoptysis, malignancy.

APACHE II score uses 15: history of severe organ failure, age, temperature, arterial pressure, pH, heart rate, respiratory rate, sodium, potassium, creatinine, renal failure, hematocrit, white blood count, Glasgow coma scale, FiO2