

Rethinking Model Complexity in Clinical Prediction Models

Fleur Vereijken¹, Jenna Reys^{1,2}, Peter Rijnbeek¹, Ross Williams¹

¹Erasmus University Medical Centre, Rotterdam, The Netherlands

²Johnson & Johnson, Titusville, NJ, USA

Background

Clinical prediction models are increasingly developed to guide decision making. It is often assumed that a well-defined prediction model provides a single estimate for a given patient, task and dataset. This assumption often does not hold. For example, multiple models predicting mortality after acute myocardial infarction, with similar validity, produced considerably different predictions for individual patients [1]. As such attention is shifting toward the issue of model uncertainty in terms of individual prediction [2]. Consideration of model uncertainty is essential within healthcare applications, where they are being used to support clinical decision making that can directly affect clinical outcomes. For example, a patient predicted to have a 5% risk by model A and a 15% risk by model B may or may not receive intervention if a decision threshold of 10% is applied.

Several factors influence both population-level predictive performance and instability in individual risk estimates, such as model development size [3, 4], modelling choices [5, 6], model complexity. One commonly recommended approach to improve population-level discrimination is to include all available baseline covariates and let the data determine which have the most predictive power over the outcome of interest [7]. It is possible that increasing model complexity will lead to more instability, so we want to highlight the trade-off between model complexity, model predictive performance, and model stability.

Therefore, we empirically study the interaction between model complexity, varied by the number of predictors included and its effect on population-level performance and individual prediction instability across increasing model development sizes.

Methods

A. Data sources

We analyzed one longitudinal US Claims databases and one Dutch primary care database:

(i) Merative MarketScan® MultiState Medicaid Database (MDCD).

(ii) The Integrated Primary Care Information (IPCI) database [8].

All datasets were converted to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [9].

B. Data sets

For our analysis, we used a total of 4 data sets (2 prediction tasks per database) (Table 1).

Table 1 Overview of prediction tasks used.

Task	Target	Outcome	Time at risk (years)
1	Individuals aged 55-84 (between January 1, 2014, and December 31, 2014)	Dementia	5

2	Patients (45-65 year) with an outpatient visit in 2017 (no prior cancer or screening for suspected cancer)	Lung cancer	3
---	--	-------------	---

C. Model choice

For this analysis, we used logistic regression with L1 regularization, also referred to as the least absolute shrinkage and selection operator (LASSO) [10] using the Cyclops Package [11]. Additionally we included the algorithm Iterative Hard Thresholding (IHT) [12].

D. Candidate predictors

Candidate predictors included demographics (gender, age in years, race, ethnicity, and index month), condition diagnoses, drug exposures, procedures performed, measurement occurrences, observations, and device exposures. We evaluated nine candidate predictor sets:

1. LASSO using age + sex only
2. LASSO with Univariate Feature Selection (UFS), $k = 15$
3. LASSO with UFS, $k = 25$
4. LASSO with UFS, $k = 50$
5. LASSO with UFS, $k = 100$
6. LASSO with constrained predictor set [13] ($k = 51$)
7. IHT using a Bayesian Information Criterion (BIC) penalty, $k = 100$
8. IHT using a Akaike Information Criterion (AIC) penalty, $k = 100$
9. LASSO with no restriction (full model)

E. Model development

Data were split into training and test sets, with a single held-out test set used to evaluate all models within each prediction task and database. For datasets with fewer than 6,250 outcome events, 25% of the full dataset was reserved for testing. For larger datasets, a random subset was sampled such that, after a 75%/25% train–test split, the training set contained 5,000 outcome events.

Learning curves were generated by sampling training subsets with increasing numbers of outcome events (350 to 5,000, where available) using event-stratified sampling to preserve the event rate. For each training subset, models were developed using 200 bootstrap resamples to estimate individual-level uncertainty.

F. Population and individual level stability

Model discrimination was assessed using the AUROC, estimated on the test data both with and without correction for optimism. Optimism arises when a model overfits to training data, yielding performance estimates that are overly optimistic compared to what would be observed in new data. To correct for this, the 200 bootstrap models were used. For each iteration, the difference between performance on the bootstrap sample and on the original test data provides an estimate of optimism; averaging this across all iterations yields the overall optimism estimate, which is then subtracted from the apparent AUROC. This produces a more realistic estimate of out-of-sample performance. Learning curves were constructed for all prediction tasks across all model complexities.

To assess the potential clinical impact of individual instability, we calculated the mean risk of misclassification, defined as the proportion of bootstrap predictions that resulted in a different classification relative to the original model, using a decision threshold set to the outcome event rate.

Results

Predictor selection

Figure 1 displays the complexity of each final model in terms of the number of included predictors. The Full model retains the most predictors overall, with complexity decreasing across the 100-predictor model, IHT with AIC penalty, the constrained predictor set, the 50-, 25-, and 15-predictor models, IHT with BIC penalty, and the age + sex model.

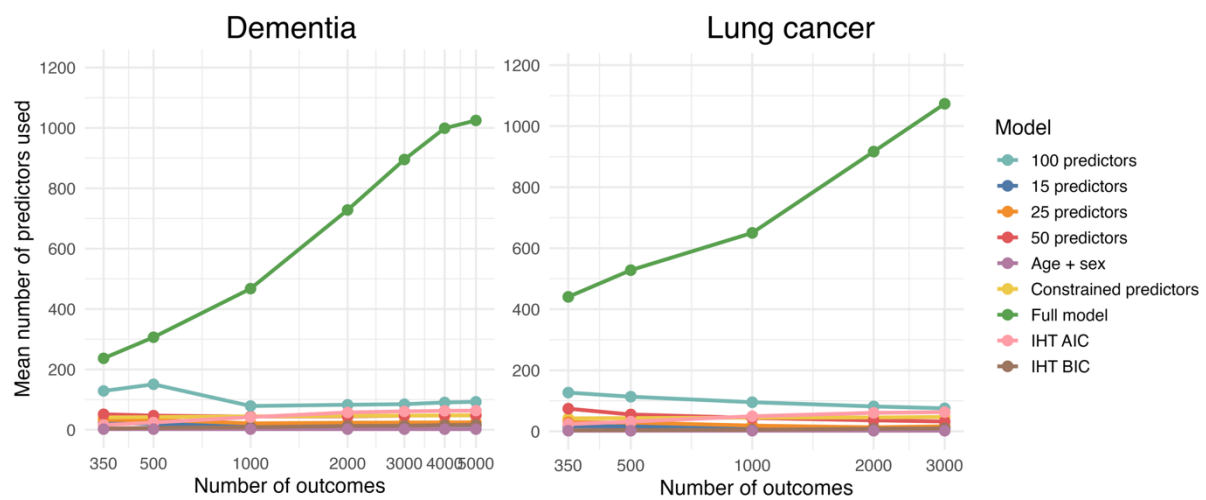


Figure 1 overview of model complexity, indicated by included predictors in the final model.

Overall performance

While models using the full set of predictors tend to yield the highest AUROC (Supplementary figure 1), they lose the most predictive performance after correcting for optimism (Figure 2).

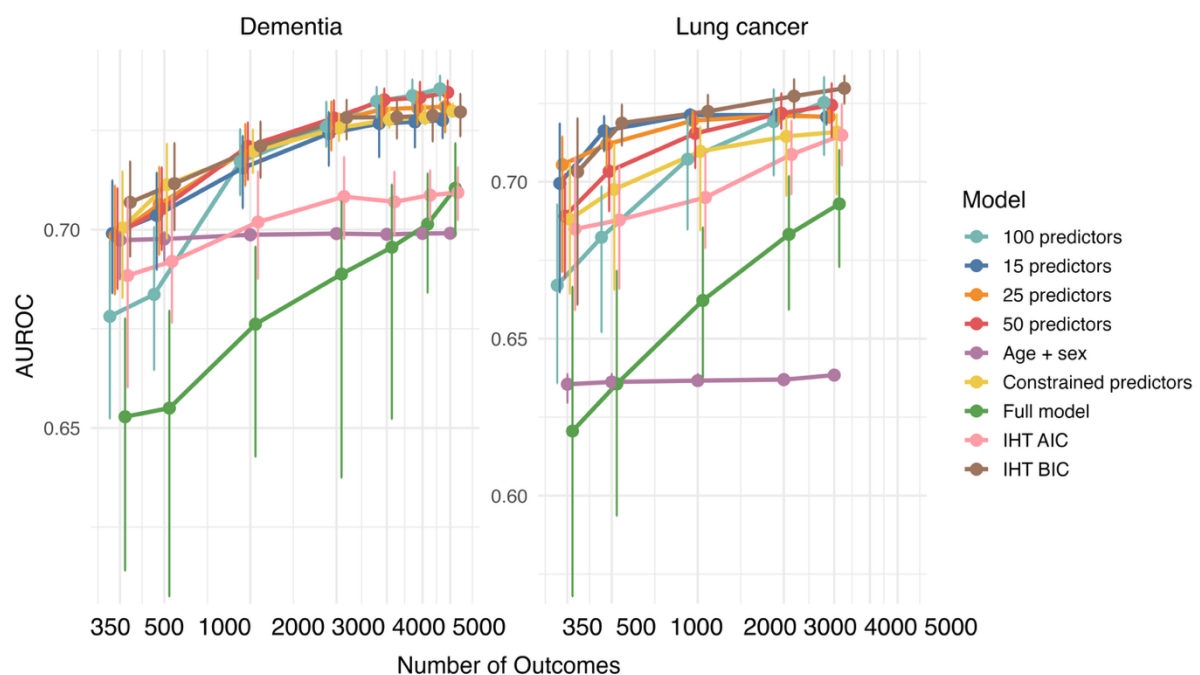


Figure 2 Learning curves of optimism corrected AUROC.

Trade off AUROC vs Decision instability

Figure 3 illustrates the relationship between optimism-corrected AUROC and individual instability as a source of decision uncertainty across increasing development sample sizes. In the figure, higher AUROC values appear toward the center, while greater decision uncertainty is reflected by a wider spread of points. Although the age + sex model exhibits the lowest decision uncertainty, it does not achieve the highest AUROC. Conversely, the Full model, particularly at smaller sample sizes, shows both the lowest optimism-corrected AUROC and the highest decision uncertainty.

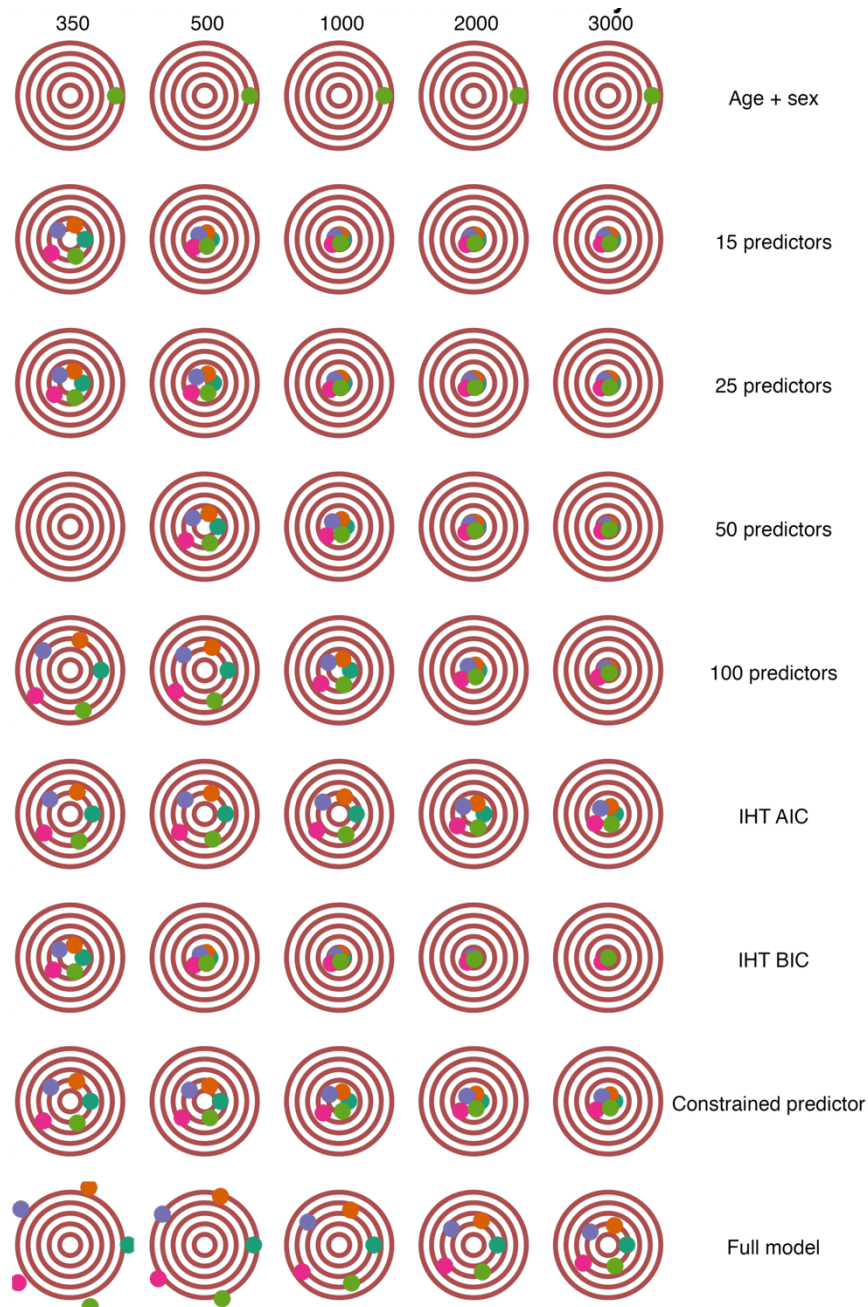


Figure 3 Dartboard visualizing the tradeoff between optimism corrected AUROC and decision uncertainty. With dots more in the middle indicating higher AUROC, and dots more spread out indicating more instability.

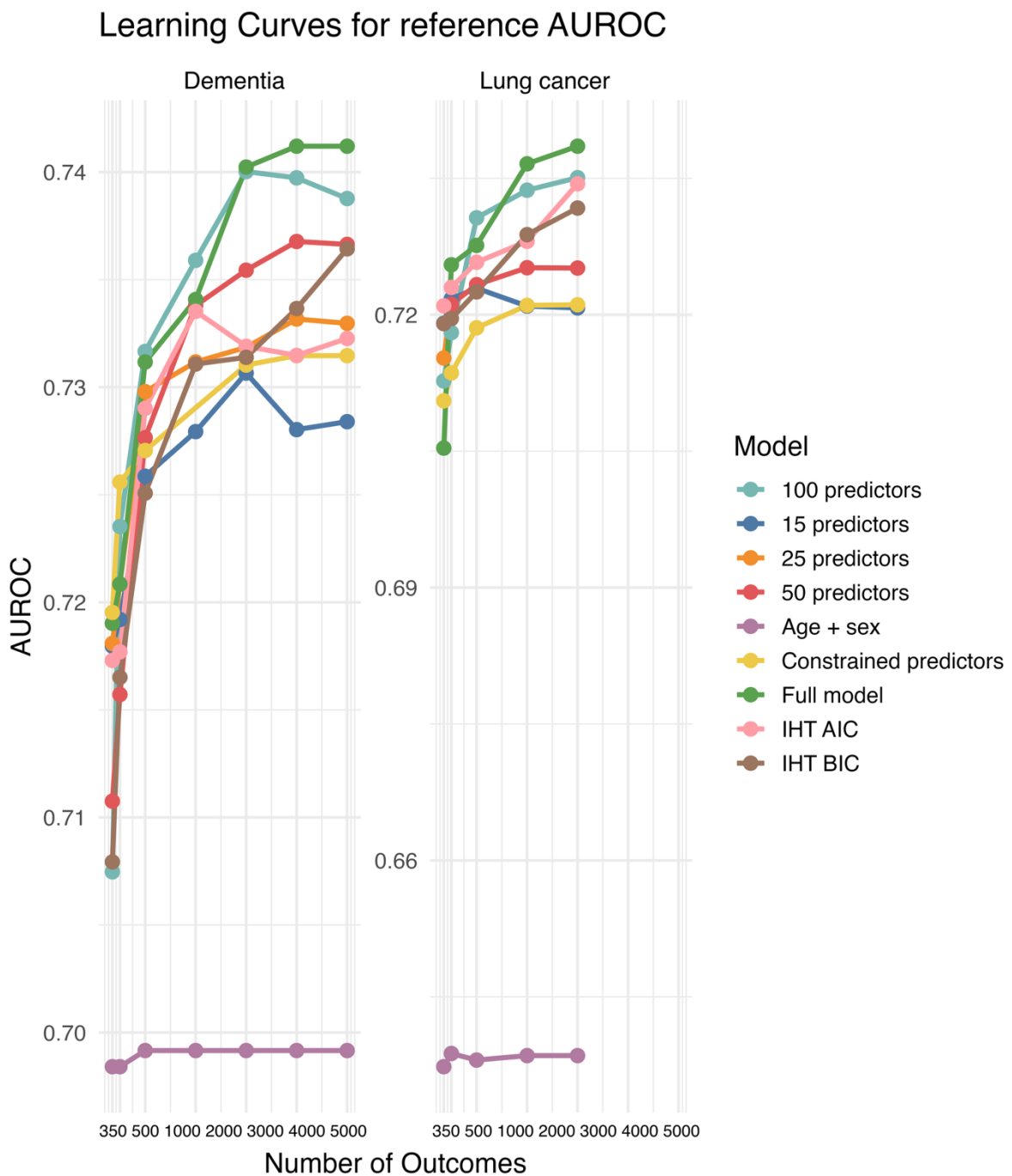
Conclusion

The Full model, which places no restriction on the number of predictors, produced overly optimistic performance estimates and greater instability. Models restricted to 15 or 25 predictors and IHT with BIC penalty achieved the best complexity-performance tradeoff at small development sample sizes. As sample size increases, a comparable tradeoff can be achieved with a larger number of predictors.

References

1. Steyerberg, E.W., et al., *Equally valid models gave divergent predictions for mortality in acute myocardial infarction patients in a comparison of logistic [corrected] regression models*. J Clin Epidemiol, 2005. **58**(4): p. 383-90.
2. Clyde, M. and E.I. George, *Model Uncertainty*. Statistical Science, 2004. **19**(1).
3. John, L.H., et al., *Logistic regression models for patient-level prediction based on massive observational data: Do we need all data?* Int J Med Inform, 2022. **163**: p. 104762.
4. Riley, R.D., et al., *Calculating the sample size required for developing a clinical prediction model*. BMJ, 2020. **368**: p. m441.
5. Vereijken, F., et al., *Loss function influence on hyperparameter optimization for observational healthcare prediction models*. J Am Med Inform Assoc, 2026.
6. D'Amour, A., et al., *Underspecification Presents Challenges for Credibility in Modern Machine Learning*. Journal of Machine Learning Research, 2022. **23**: p. 1-61.
7. Reps, J.M., et al., *Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data*. J Am Med Inform Assoc, 2018. **25**(8): p. 969-975.
8. de Ridder, M.A.J., et al., *Data Resource Profile: The Integrated Primary Care Information (IPCI) database, The Netherlands*. Int J Epidemiol, 2022. **51**(6): p. e314-e323.
9. Overhage, J.M., et al., *Validation of a common data model for active safety surveillance research*. J Am Med Inform Assoc, 2012. **19**(1): p. 54-60.
10. Tibshirani, R., *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological), 1996. **58**(1): p. 267-288.
11. Suchard, M.A., et al., *Massive parallelization of serial inference algorithms for a complex generalized linear model*. ACM Trans Model Comput Simul, 2013. **23**(1).
12. Blumensath, T. and M. Davies, *Iterative Hard Thresholding for Compressed Sensing*. Applied and Computational Harmonic Analysis, 2009. **27**: p. 265-274.
13. Reps, J.M., et al., *Finding a constrained number of predictor phenotypes for multiple outcome prediction*. BMJ Health Care Inform, 2025. **32**(1).

Supplementary



Supplementary figure 1 AUROC learning curves.