Title

A Novel Method for Identifying a Parsimonious and Accurate Predictive Model for Multiple Clinical Outcomes

Authors

Diaz-Ramirez, L Grisell
Lee, Sei J
Smith, Alexander K
et al.

Publication Date

2021-06-01

DOI

10.1016/j.cmpb.2021.106073

HHS Public Access

Author manuscript

Comput Methods Programs Biomed. Author manuscript; available in PMC 2021 June 01.

Published in final edited form as:

Comput Methods Programs Biomed. 2021 June ; 204: 106073. doi:10.1016/j.cmpb.2021.106073.

A Novel Method for Identifying a Parsimonious and Accurate Predictive Model for Multiple Clinical Outcomes

L. Grisell Diaz-Ramirez[a,b,*], **Sei J. Lee**[a,b], **Alexander K. Smith**[a,b], **Siqi Gan**[a,b], **W. John Boscardin**[a,b]

[a]Division of Geriatrics, University of California, San Francisco, 490 Illinois Street, Floor 08, Box 1265, San Francisco, CA 94143, United States

[b]San Francisco Veterans Affairs (VA) Medical Center, 4150 Clement Street, 181G, San Francisco, CA 94121, United States

Abstract

Background and Objective: Most methods for developing clinical prognostic models focus on identifying parsimonious and accurate models to predict a single outcome; however, patients and providers often want to predict multiple outcomes simultaneously. As an example, for older adults one is often interested in predicting nursing home admission as well as mortality. We propose and evaluate a novel predictor-selection computing method for multiple outcomes and provide the code for its implementation.

Methods: Our proposed algorithm selected the best subset of common predictors based on the minimum average normalized Bayesian Information Criterion (BIC) across outcomes: the Best Average BIC (baBIC) method. We compared the predictive accuracy (Harrell's C-statistic) and parsimony (number of predictors) of the model obtained using the baBIC method with: 1) a subset of common predictors obtained from the union of optimal models for each outcome (Union method), 2) a subset obtained from the intersection of optimal models for each outcome (Intersection method), and 3) a model with no variable selection (Full method). We used a case-study data from the Health and Retirement Study (HRS) to demonstrate our method and conducted a simulation study to investigate performance.

Results: In the case-study data and simulations, the average Harrell's C-statistics across outcomes of the models obtained with the baBIC and Union methods were comparable. Despite the similar discrimination, the baBIC method produced more parsimonious models than the Union method. In contrast, the models selected with the Intersection method were the most parsimonious, but with worst predictive accuracy, and the opposite was true in the Full method. In the simulations, the baBIC method performed well by identifying many of the predictors selected in

*Corresponding author at: San Francisco Veterans Affairs (VA) Medical Center, 4150 Clement Street, 181G, San Francisco, CA 94121, United States. grisell.diaz-ramirez@ucsf.edu (L.G. Diaz-Ramirez).

Declaration of Competing Interest

The authors declare that they have no competing interests.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2021.106073.

the baBIC model of the case-study data most of the time and excluding those not selected in the majority of the simulations.

Conclusions: Our method identified a common subset of variables to predict multiple clinical outcomes with superior balance between parsimony and predictive accuracy to current methods.

Keywords

backward elimination; Bayesian Information Criterion; prognostic models; survival analysis; variable selection

1. Introduction

One of the first steps in building a regression model is selecting a subset of predictors from a pool of many available predictors. Clinicians and researchers alike desire a model that explains the data in the simplest way–namely, a parsimonious model–with appropriate predictive accuracy. Parsimonious models offer the potential to save the time it takes to gather unnecessary predictors, and expense, either in visit time or in money.

Most current model development methods focus on accurate and parsimonious prediction of single outcomes. Popular methodologies that are easy to use and interpret include stepwise methods like backward elimination or criterion-based selection like the Akaike Information Criterion (AIC) [1] or the Bayesian Information Criterion (BIC) [2]. However, obtaining the most parsimonious and accurate model is more complex for the simultaneous prediction of multiple outcomes, a common scenario in clinical settings.

Several studies have demonstrated that older adults care not only about mortality, but also about their quality of life, specifically their ability to function independently [3,4]. In the realm of anticoagulation for atrial fibrillation, for example, clinicians may want to simultaneously predict risk of stroke and risk of a major gastrointestinal bleed [5,6]. In primary care, clinicians may want to balance risk of microvascular complications from diabetes against the risks of hypoglycemia and falls [7,8]. Yet, there is limited research on how best to develop clinical prognostic models that predict multiple outcomes simultaneously with accuracy and parsimony.

In this paper, we propose and evaluate a novel computing method for predictor selection in prognostic models of multiple clinical outcomes using the minimum average normalized BIC across outcomes, which we call the Best Average BIC (baBIC). To develop the proposed method, we use the Health and Retirement Study (HRS) data and a common set of health-related and demographic variables to predict time to: 1) Activities of Daily Living (ADL) Dependence, 2) Instrumental Activities of Daily Living (IADL) Difficulty, 3) Mobility Dependence, and 4) Death. We present the algorithm to compute the baBIC in both stepwise selection and LASSO settings. We compare the parsimony and predictive accuracy of this model with the models obtained using traditional approaches for variable selection in the clinical field. The proposed methodology provides a strategy to assess the incidence (i.e. probability of occurrence) of multiple outcomes simultaneously with appropriate predictive accuracy, while at the same time saving the clinical time and expense associated with

collecting unnecessary predictors. The data and SAS and R code for reproducing the results of this article are freely available at a Github repository [9].

This paper is organized as follows. Section 2 describes existing methods for variable selection and introduces the baBIC method. Section 3 presents the proposed baBIC methodology in details. Section 4 describes the case-study data and the implementation of our algorithm using these data. Section 5 presents the simulation study used to evaluate the performance of our method. The experimental results of the implementation of the baBIC method in the example data and simulations are described and discussed in Sections 6 and 7. Finally, we draw some conclusions in Section 8.

2. Background

Much of the research on variable selection for multiple outcomes has been done in the high-dimensional multivariate regression setting, where the number of predictors and, sometimes, the number of outcomes exceed the number of observations. Under this setting, the implementation of shrinkage or regularization methods is common [10–13]. Other authors have addressed variable selection for multivariate modelling using a Bayesian framework [14–16]. However, in clinical settings, where the sample size is frequently large relative to the number of predictors and outcomes, a simpler and easy-to-implement procedure that does not require complex software solutions could be of great utility. In this way, a recent clinical study identified a common set of predictors across several adverse outcomes, one of them being a composite of the other outcomes [17]. This method allowed the optimization of clinical resources by focusing on a single-combined outcome.

More broadly, outside biomedical studies, other authors have approached the issue of common variable selection using variations of orthogonal forward regression. For example, a study in the engineering field described the extended forward orthogonal regression (EFOR) algorithm to select a parsimonious common-structured model that would save time and money in system analysis and design [18]. In another study, these authors also developed an improved methodology where a common-structured model was identified using random subsampling and a multifold modelling (RSMM) approach with a multiple orthogonal search (MOS) algorithm [19].

An obvious approach (which we label Individual Outcome method) to address the multiple outcomes problem is to simply select a different subset of variables to predict each of the outcomes using selection methods for single outcomes. Although straightforward, this method could be time-consuming, expensive (due to the cost of acquiring multiple predictors), and potentially lead to overfitting and high variability [10,13].

A slight modification to this approach is the Union method. In this method, we take the separate models from the Individual Outcome method, and then force the union of the predictors from each model into the predictor set for each outcome. Like the Individual Outcome method, the Union method has the advantage of being a simple approach, and, additionally, it allows patients and clinicians to focus on a common subset of variables that can accurately predict their outcomes of interest simultaneously. Nevertheless, the Union

method could lack parsimony as it includes all variables that predict all outcomes well, including those that are only important for some of the outcomes.

On the other hand, to improve parsimony one could envision a method that only selects those common predictors obtained in all individual subset selections, namely the Intersection method. This method will likely be more parsimonious than the Union method, as only the variables important for all outcomes might be selected, but having too few predictors may not adequately describe the relationship between outcomes and predictors.

Finally, in the clinical field, sometimes researchers do not perform variable selection and instead choose a list of predictors defined a priori based on clinical reasons (which we call the Full method). This would be the simplest approach; however, when there are many candidate predictors, it may produce a non-parsimonious model that has redundant predictors and shows overfitting problems.

Our proposed method, the Best Average BIC (baBIC) method, selects the best subset of common predictors for M outcomes according to the baBIC. We compare our method with: 1) a method that selects individual subsets of predictors for each outcome (Individual Outcome method), 2) an enhanced method that creates a best subset of common predictors based on the union of individual subsets obtained in the Individual Outcome approach (Union method), 3) a method that selects a common subset based on the intersection of individual subsets (Intersection method), and 4) a method with no variable selection (Full method).

3. The Best Average BIC (baBIC) method

In the baBIC method, we averaged the normalized BIC (nBIC) across outcomes. Normalization was important to ensure that a change in BIC from a complex to a simpler model meant roughly the same across multiple outcomes; that is, the BICs were in a comparable scale. The nBIC was computed by dividing the absolute difference between the BIC of a particular model for a specific outcome and the BIC for the “best” individual model for that outcome by the difference between the BIC in the full model (i.e. with all candidate predictors) and the BIC in the best individual model:

$$nBIC\ (k) = \frac{(BIC(k) - BIC\ best\ individual\ model)}{(BIC\ full\ model - BIC\ best\ individual\ model)}$$

Where:

$$BIC\ (k) = -2logL + k\ \log\ (number\ of\ uncensored\ observations)$$

L: the maximized value of the likelihood function of the fitted model

k: number of parameters estimated by the fitted model

The nBIC thus ranges between 0 (for the best individual model) and 1 (for the model that contains all candidate predictors), with smaller being better. The nBIC can be larger than 1

for models that have a worse BIC value than the full model, but these models are typically not of interest in our setting. This normalization allowed us to average the nBIC across different outcomes and, at the same time, made this metric more interpretable.

Explicitly, we defined the baBIC for a model with k parameters as:

$$baBIC\ (k) = \frac{1}{M}\sum_{m=i}^{M} nBIC\ (k; mth\ outcome)$$

Where:

m: enumerates the M total outcomes

The baBIC criterion has the flexibility to be incorporated into selection methods already available for single outcomes. Therefore, this method is not intrinsically linked to any particular method of variable selection and can be used to compare arbitrary sets of candidate models. In fact, a statistic similar to the baBIC was described in Wei and Billings' work [19] mentioned previously in Section 2. These authors proposed a statistic called the weighted average BIC (WABIC) to determine the number of common model terms across multiple regression models after using their random subsampling and multifold modelling approach. In their work, the WABIC included two main terms corresponding to the WABIC for the training and validation data sets, which were defined in a similar way as our baBIC equation above. However, one important difference between baBIC and WABIC is that the latter requires choosing and optimizing a weight coefficient, which does not appear to be trivial.

In order to compute the nBIC for specific outcome and then the baBIC across outcomes, we need to obtain the BIC of the full model and the BIC of the best individual model. The BICs of the full and best individual models can be found using stepwise regression methods like backward elimination or more current selection methods like the Least Absolute Shrinkage and Selection Operator (LASSO) [20]. Comparison of various methods for variable selection including best subset, stepwise, and LASSO remains an area of active investigation in the statistical literature, with no one method dominating the others across a variety of settings [21].

In software implementations of stepwise regression, the BIC is output at each step of the selection process so it is straightforward to find the BIC for the full model as well as the best individual model BIC value (see further details in Section 4.3). Similarly, selection based on minimum BIC can be directly incorporated into the LASSO setting [22,23]. LASSO regression shrinks the regression coefficients toward zero by penalizing the regression model with the sum of the absolute coefficients (L1 penalty). A tuning parameter called lambda (λ) controls the strength of the L1 penalty, so selecting a good value for λ is critical. After doing LASSO selection, we can compute BIC for each possible λ (as shown above), select the one that gives the minimum BIC as the optimal λ, and extract the corresponding BIC as the BIC of the best individual model. The BIC of the full model corresponds to the model with λ=0.

4. Implementation of the baBIC method in the case-study data

4.1. Description of the case-study data

To demonstrate our baBIC method, we created a nationally representative cohort of 5,531 community-dwelling seniors enrolled in the HRS, who were 70 years old or older at the time of their baseline interview in 2000. The HRS is an ongoing longitudinal survey of a representative sample of all persons in the United States over age 50 that examines changes in health and wealth [24]. Before each interview, HRS participants are provided with a written informed consent information document and give oral consent for their participation in the HRS. The institutional review boards of the University of California, San Francisco approved the present study.

The pool of predictors included 39 health-related and demographic categorical variables measured at baseline. We used 4 clinical outcomes encompassing 14 years of follow-up: 1) time to first ADL dependence (including five ADLs: bathing, dressing, toileting, transferring, and eating), 2) time to first IADL difficulty (including two IADLs: managing money and medication), 3) time to first mobility dependence, and 4) time to death.

The BICs used in the baBIC method were obtained from survival models. For time to death, we fitted Cox proportional hazards regression models [25]. For times to first ADL dependence, IADL difficulty, and mobility dependence, we fitted Fine and Gray competing-risk regression models to appropriately account for the risk of death [26].

4.2. Measures used to evaluate parsimony and predictive accuracy

In the final models, we evaluated parsimony with the number of predictors selected, and we measured predictive accuracy using Harrell's C-statistic [27]. Harrell's C-statistic is a goodness of fit measure for survival models, and it is calculated based on risk scores obtained after fitting the model. If the model has good predictive accuracy, subjects with higher risk scores have a shorter time-to-event. To compute the C-statistic, subjects are grouped in pairs and classified as concordant or discordant. Briefly, in concordant pairs, subjects with the event of interest have higher risk scores and shorter time-to-event. Then, the C-statistic is computed as the proportion of concordant pairs among all pairs. A value of C-statistic = 0.5 indicates a model with non-informative prediction, whereas C-statistic = 1 indicates perfect prediction.

Reporting predictive accuracy in the same data set used to develop the model can lead to an overestimate of model performance, termed "model optimism." A well-regarded approach [28,29] is to estimate and correct for the degree of optimism using bootstrapping. This is strongly preferable to single split sample validation and somewhat preferable to cross-validation.

We generated 500 bootstrap samples with replacement from the case-study data, each with the same sample size. We reported the average number of predictors with their 2.5th and 97.5th percentiles. We computed the bootstrap-based optimism-corrected C-statistic as follows:

1. Obtain final model and corresponding C-statistic on the case study data, namely C-statistic-apparent
2. Obtain final models of each bootstrap sample and compute the C-statistic of each bootstrap model, namely C-statistic-boot
3. Compute the C-statistic of each bootstrap model evaluated in the original case-study data, namely C-statistic-original
4. Calculate the optimism in the fit of each bootstrap sample as:

   C-statistic-boot - C-statistic-original
5. Average the optimism across 500 bootstrap sample, namely O
6. Compute the optimism-corrected C-statistic of the case-study data as:

   C-statistic-apparent - O

We then computed the location-shifted bootstrap confidence intervals of the optimism-corrected C-statistic by subtracting the optimism estimate from the 2.5^{th} and 97.5^{th} percentiles of the C-statistic-bootstrap distribution [30].

4.3. Implementation of the baBIC method

We compared the bootstrap average number of predictors and the optimism-corrected C-statistic of the models obtained with the Individual Outcome and Union methods using BIC backward elimination vs. LASSO selection based on optimal λ at the minimum BIC. We found that backward elimination based on minimum BIC produced more parsimonious models than LASSO selection, while maintaining very similar predictive accuracy (Appendix A). Consequently, we chose the BIC backward elimination to further illustrate the baBIC method in this setting.

We implemented the baBIC algorithm using BIC backward elimination as follows. The method started with all 39 (p) predictors and selected the subset of 38 (p-1) predictors with minimum baBIC. To select the subset of predictors with minimum baBIC, we fitted for each outcome all possible combinations of predictors obtained by removing 1 predictor at a time. We then computed the average of the nBICs across the 4 outcomes within each subset of predictors and selected the subset of 38 (p-1) with the minimum baBIC (Fig 1). The same process continued until there were only 2 variables left (i.e. “Age decile groups” and “Female”), which were forced in. Lastly, the method selected the final subset of predictors that had the minimum baBIC across all subsets of different number of predictors from p-1 to 2 (Fig. 2).

For the comparative methods, Individual Outcome, Union, and Intersection methods, we followed a similar approach as described above. The only difference being that the backward elimination was based on the minimum BIC for each individual outcome instead of the minimum baBIC across the 4 outcomes. We then obtained the final models of the Union and Intersection methods. The Union method produced a model that contained all the predictors that were in at least 1 of the 4 best subsets selected by the Individual Outcome method, whereas the model obtained by the Intersection method contained only a few predictors

selected in all four outcomes (Fig. 3). Fig. 4 shows the selection performance in terms of the baBIC and Individual Outcome selection against the number of predictors using the case-study data.

5. Simulation study

We conducted a simulation study to assess the performance and feasibility of the proposed baBIC method in the selection of a common subset of variables to predict multiple clinical outcomes with accuracy and parsimony.

The simulated survival times were generated using Harden and Kropko [31] method to simulate survival data for the Cox model. In brief, this method generates at each iteration of the simulation a unique baseline hazard by fitting a cubic spline to randomly-drawn points. This yields baseline hazards that can vary considerably and consequently simulated data with great heterogeneity. We used R version 4.0.3 random number generator with the default Mersenne Twister algorithm. The input seed was “20210109.”

We considered three data-generating mechanisms or scenarios. For each of the scenarios, the values of the non-zero β were set to be equal to those estimated in the case-study data. In scenario 1, for all the outcomes, we assumed the β=0 except for the ones corresponding to the 15 predictors selected with the baBIC method for the case-study data. On the other hand, in scenario 2, for each outcome we assumed the β=0 except the ones corresponding to the predictors selected with each Individual Outcome method for the case-study data. For example, to simulate time to first ADL dependence, we used β from the 10 predictors obtained with the Individual ADL Outcome method and set rest of β=0; whereas to simulate time to death we used 16 non-zero β and set rest of β=0 (Fig. 3). Finally, in scenario 3 we used the β corresponding to all 39 candidate predictors, thus all the β were non-zero. Appendix B shows the relationships between predictors and outcomes under the three scenarios.

The estimated betas used in the three scenarios were obtained from fitting Cox models instead of Competing-risk regression models. To do this, we used a modified version of the case-study data where those who died were treated as being censored at the longest possible time that any respondent was followed (i.e. 14 years) [32]. Of note, we obtained the same final subset of predictors for the case-study data set in all selection methods with and without this simplification.

Within each scenario we simulated 4 survival times with the same censoring as the case-study data (i.e. ADL= 66.55%, IADL=64.98%, WALK=81.90%, DEATH=31.87%) and with 25% censoring. For the three scenarios, the data were simulated on 5,531 respondents which was the sample size of the case-study.

We generated two sets of 500 training and 500 test simulated data for each scenario with the case-study censoring and 25% censoring. The training set was used for model selection and the test set was used to assess predictive accuracy. This number of simulations gave us a good balance between feasible computing time and acceptably small Monte Carlo Standard

Errors (SEs). After obtaining the simulated survival times, each simulated outcome data set was merged with the design matrix of 39 predictors of the case-study data set.

Appendix C shows the medians [IQR] and percentages of incidence for the marginal distributions of the survival outcomes in the case-study data and across simulations. As expected, the marginal distributions of the outcomes across scenarios showed great heterogeneity since at each iteration of the simulation we used a unique baseline hazard with a variety of shapes (e.g. unimodal, multimodal, monotonically increasing or decreasing) [31]. Fig. 5 and Fig. 6 show the predicted cumulative incidence by outcome at the mean of the predictors selected with the baBIC method in the case-study data using simulations of Scenario 1. The predicted cumulative incidence for the other two scenarios are shown in the Appendix D (Fig. D.1, D.2, D.3, D.4).

For each training data, we obtained the final models corresponding to the baBIC, Individual Outcome, Union, and Intersection methods. Then, each of the final models was fitted on the test set, and the averages, Monte Carlo SEs, and 2.5^{th} and 97.5^{th} percentiles of the Harrell's C-statistic computed across simulations (adding the Full method). For the number of predictors, we computed the averages, Monte Carlo SEs, and 2.5^{th} and 97.5^{th} percentiles over the 500 training data sets for each model. Additionally, for the final models obtained with the baBIC method in the simulations, we calculated the percentage of times that each of the variables selected with the baBIC for the case-study data appeared in the simulations, and the percentage of times that each of the variables that were not selected with the baBIC for the case-study data appeared in the simulations. Finally, we computed the average percentage of inclusion of predictors selected and not selected with the baBIC for the case-study data, and the percentages of models with 3 to 10+ selected predictors.

All the analyses were performed with SAS/STAT® 15.1 (Copyright © 2016 by SAS Institute Inc., Cary, NC, USA) and R version 4.0.3 (Copyright © 2020 The R Foundation for Statistical Computing). The data and SAS and R code for reproducing the results of this article are freely available at a Github repository [9].

6. Results

Fig. 3 shows the selection of the common subset of predictors of the Union method using the predictors obtained in the Individual Outcome method for the case-study data. The number of predictors selected in the Individual Outcome method ranged from 7 to 16. The model obtained with the Union method, which contained all the predictors found in at least 1 of the 4 best individual models, had 23 predictors, and most of them came from 1 or 2 individual models. By contrast, the model obtained with the baBIC method with 15 predictors was more parsimonious than the one selected with the Union method, and all the predictors selected by the baBIC were also obtained with the Union method. The Intersection method that selects a final subset of predictors that were in all four individual subsets produced the most parsimonious model with only 3 predictors, from which 2 of them were forced by all selection methods into the final models.

The results above were also confirmed in the case-study bootstrap samples and simulation study. In these data, the Union method consistently produced models with a higher average number of predictors than the baBIC method, whereas the Intersection method always produced the most parsimonious models (Fig. 7). As expected, regardless of the selection method the number of predictors obtained was significantly smaller than the initial pool of candidate predictors in the Full method. The difference in the average numbers of predictors obtained between the Union and baBIC methods was more subtle in scenario 1 where the simulated survival times were generated using the common set of predictors from the baBIC model in the case-study data (Appendix E; e.g. case-study censoring Union method 15 [2.5^{th}, 97.5^{th} : 12–17] vs. baBIC method 10 [2.5^{th}, 97.5^{th} : 7–13]; 25% censoring Union method 15 [2.5^{th}, 97.5^{th}: 14–16] vs. baBIC method 13 [2.5^{th}, 97.5^{th}: 11–15]). In contrast, scenario 2 and 3–where the survival times were obtained using the individual best sets of predictors or all the 39 candidate predictors of the case-study data–showed a more evident difference between the Union and baBIC methods.

In the case-study bootstrap data and simulations, the C-statistics of the models obtained with the Individual Outcome, Union, and baBIC methods were clinically similar within each outcome. By contrast, the final models of the Intersection and Full methods had the lowest and highest C-statistics respectively (Fig. 8). The average optimism-corrected C-statistics across outcomes of the Union and the baBIC methods were very similar in the case-study data (both 0.65). Likewise, in the simulations the average predictive accuracies of the Union and baBIC methods were alike regardless of the scenario and censoring (e.g. scenario 2 with case-study censoring, Union method: 0.62 vs. baBIC method 0.61) (Appendix F).

When using the baBIC method in the simulations, many of the predictors present in the baBIC model of the case-study data were correctly identified. On average in scenario 1, this method selected the same predictor obtained with the baBIC in the case-study data 58.3% and 83.9% of the times for case-study censoring and 25% censoring respectively. In simulation scenario 2, the baBIC method selected on average these predictors 50.5% of the times for case-study censoring and 64.1% of the times for 25% censoring; whereas in scenario 3 these percentages were 34.4% and 56.7% respectively. Of the 15 predictors selected by the baBIC method in the case-study data, a range of 3 to 15 of these predictors were present in the baBIC models of the simulations, and the percentage of models with 10 or more of these predictors ranged from 2.6% (scenario 3, case-study censoring) to 99.8% (scenario 1, 25% censoring). Finally, the percentage of predictors not included in the baBIC model of the case-study but present in the baBIC models of the simulations was less than 19% across scenarios (Table 1).

7. Discussion

The baBIC selection method produced a model with a good balance between parsimony and predictive accuracy. In both the case-study data and the simulations, this model was more parsimonious than the one obtained with the Union method or the Full method, and it showed minimal loss of predictive discrimination, as opposed to the Intersection method. A good compromise between parsimony and accuracy is important since models that are simpler to understand and explain and that predict outcomes well are more likely to be

implemented. Models with too few predictors cannot adequately describe the relationship between outcomes and predictors, whereas those with too many predictors can cause overfitting problems. Moreover, as the number of predictors in the model increases, the time and cost of collecting them could also increase. From a practical perspective, busy clinicians are unlikely to use a prognostic model with a daunting list of predictors to collect and enter. Although we did not formally incorporate a penalization associated with the cost of the predictors, other authors have explicitly balanced predictive accuracy against cost of the predictors [14,33].

In scenario 1, where the simulated survival times were generated using only the common set of predictors from the original baBIC model, the simulated baBIC models were still more parsimonious than those obtained with the Union method (by about 2–5 predictors on average). The selection method intrinsically favored the predictors that were used during the data-generating mechanism. Consequently, during the individual selection process, the 4 outcomes ended up having more common predictors, which in turn reduced the overall number of predictors selected with the Union method. On the other hand, scenario 2 assumed that each outcome had an individual best set of predictors, whereas scenario 3 incorporated all 39 candidates predictors for all outcomes. This markedly increased the number of predictors in the simulated models with the Union method while maintaining comparable parsimony in the simulated baBIC models to those from scenario 1.

In the simulations, we found that the baBIC method performed well by selecting on average a high percentage of the predictors included in the final baBIC model of the case-study data, while keeping a low percentage of the predictors that were not in the baBIC model. Overall, scenarios with 25% censoring included more predictors (both selected and not selected in the baBIC model of case-study data). This suggested that we had higher power in the selection of the predictors used during the data-generating mechanism with lower censoring rates (i.e. 25% censoring). Others have found similar results [34].

As noted in Section 2, several studies have used penalized regression under the high dimensional multivariate regression setting, where the numbers of predictors and outcomes may be large compared to the sample size. Regularization methods are particularly suitable for the study of genetic pathways or genome-wide association analysis, where high dimension, low sample size settings are very common [13,35,36]. In clinical settings, researchers are usually interested in interpretable effect estimates in addition to good predictive performance. Regression coefficients estimated by regularization schemes like those that are an extension of LASSO can be biased, making their interpretation more difficult [37]. Furthermore, in our case-study data we obtained less parsimonious models with the Individual and Union LASSO selection methods compared to backward elimination, while maintaining very similar predictive accuracy. Likewise, in an extensive simulation study, Hastie et al. [21] have recently noted that neither LASSO nor best subset selection nor stepwise regression was dominant across a variety of problem settings, and that no method had a large difference in variation explained. Thus, they suggested favoring methods that are easy to compute.

Consequently, we believe that in the clinical practice where the sample size is usually large compared with the number of outcomes and predictors, our baBIC method, which extends the use of popular (non-regularized) variable selection methods to the multivariate settings, has the benefit of easier implementation and interpretation as well as good predictive performance and parsimony.

It is worth mentioning that our method focused on the selection of a common set of variables to predict multiple outcomes accurately based on the assumption that all the outcomes should be associated (to some degree) with some of the predictors used in the pool of candidate predictors. We envision our method to be particularly useful in the clinical field where practitioners collect a number of health-related and demographic variables known to be important in the prediction of related outcomes. The extent to which the correlation between the outcomes impacts the selection of the predictors could be investigated in future projects. In this way, several studies have developed methods for variable selection explicitly accounting for the correlation among multiple outcomes [11,12,15,35,36].

In this study, we aimed to select a common subset of variables from a pool of many available predictors rather than identify a final predictive model. Thus, we assumed that all aspects of model building are fixed, except the selection of the predictors. In the actual application of this method, researchers will need to consider the rest of the aspects involved in model building; for example, possible inclusion of non-linear terms, interaction and multi-collinearity between predictors, and for survival models, validity of the proportional hazard assumption. Additionally, it will be important to assess the performance of the final model using both calibration and discrimination techniques, as well as conducting model validation by using a training-test split, internal cross validation (bootstrapping), and external validation [28]. In a real life application, our method could be fully incorporated during the process of model development and validation.

There are some limitations to our baBIC method. In the case-study data with a sample size of 5,531 observations, 4 outcomes, and 39 candidate predictors, the baBIC method took approximately 19 hours to complete. This is mainly because of the computational constraint of fitting hundreds of competing-risk regression models. Competing risk models have longer convergence time than Cox models. For example, a competing-risk model with 39 predictors and 5,531 observations takes approximately 40 seconds to converge, whereas a Cox model only takes less than 1 second. Thus, when we fit hundreds of Cox models in the baBIC method, the run time reduced significantly to 21 minutes. We recommend that for exploratory and simulations studies researchers use Wolbers [32] approximation to the Competing-risk setting to fit Cox models instead. In fact, as mentioned previously for model selection we obtained the same predictors using Competing-risk regression or Cox regression. Additionally, to avoid instability of selection, the number of candidate predictors relative to the effective sample size is an important consideration. A ratio of 50 events per variable (EPV) has been suggested by some authors for reliable selection [38]. Finally, our method will be more useful in a setting where the number of predictors and outcomes are small to moderate (e.g. less than 10 outcomes and less than 50 candidate predictors) which is usually the case in the clinical field. In high-dimension setting, where the number of

predictors and outcomes is much larger than the sample size, regularization methods are preferable.

8. Conclusions

Our baBIC method implemented a straightforward approach to obtain a common set of variables for the prediction of several outcomes. Researchers will be able to use our algorithm and code to develop prognostic models that are both accurate and parsimonious, potentially saving the clinical time and expense associated with gathering additional unnecessary predictors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to the reviewers for their valuable comments which helped improved this paper. We greatly appreciate Dr. Jeffrey Harden's insight and advice on the use of the R package coxed to simulate survival data. We thank Regina Anavy for proofreading this paper. This work was supported by the National Institute on Aging (grant numbers: R01AG047897, R01AG057751, K24AG066998, K24AG068312, P01AG066605) and UCSF Claude D. Pepper Older Americans Independence Center funded by National Institute on Aging (grant number: P30AG044281). The funding agency did not participate in the design of the study, collection, analysis, interpretation of data, or in writing the manuscript.

References

[1]. Akaike H, Petrov BN, Csaki F, Information theory and an extension of the maximum likelihood principle, in: Second international symposium on information theory, Budapest, Hungary, Akadémiai Kiado, 1973, pp. 267–281. 10.1007/978-1-4612-1694-0_15.

[2]. Schwarz G, Estimating the dimension of a model, Ann Statist 6 (1978) 461–464 10.1214/aos/1176344136.

[3]. Steinhauser KE, Christakis NA, Clipp EC, McNeilly M, McIntyre L, Tulsky JA, Factors considered important at the end of life by patients, family, physicians, and other care providers, JAMA 284 (2000) 2476–2482, doi:10.1001/jama.284.19.2476. [PubMed: 11074777]

[4]. Fried TR, Bradley EH, Towle VR, Phil M, Allore H, Understanding the treatment preferences of seriously ill patients, N Engl J Med 346 (2002) 1061–1066, doi:10.1056/NEJMsa012528. [PubMed: 11932474]

[5]. Singer DE, Chang Y, Fang MC, et al., The net clinical benefit of warfarin anticoagulation in atrial fibrillation, Ann Intern Med 151 (2009) 297–305, doi:10.7326/0003-4819-151-5-200909010-00003. [PubMed: 19721017]

[6]. Fang MC, Go AS, Chang Y, et al., A new risk scheme to predict warfarin-associated hemorrhage, J Am Coll Cardiol 58 (2011) 395–401, doi:10.1016/j.jacc.2011.03.031. [PubMed: 21757117]

[7]. Kirkman MS, Briscoe VJ, Clark N, et al., Diabetes in older adults: a consensus report, J Am Geriatr Soc 60 (2012) 2342–2356, doi:10.1111/jgs.12035. [PubMed: 23106132]

[8]. Moreno G, Mangione CM, Kimbro L, Vaisberg E, American Geriatrics Society Expert Panel on Care of Older Adults with Diabetes Mellitus. Guidelines abstracted from the American Geriatrics Society Guidelines for Improving the Care of Older Adults with Diabetes Mellitus: 2013 update, J Am Geriatr Soc. 61 (2013) 2020–2026, doi:10.1111/jgs.12514. [PubMed: 24219204]

[9]. Diaz-Ramirez LG, Lee SJ, Smith AK, Gan S, Boscardin WJ, A Novel Method for Identifying a Parsimonious and Accurate Predictive Model for Multiple Clinical Outcomes, 2020. https://github.com/UCSFGeriatrics/multiple-outcomes-selection.

[10]. Turlach BA, Venables WN, Wright SJ, Simultaneous variable selection, Technometrics 47 (2005) 349–363, doi:10.1198/004017005000000139.

[11]. Kim S, Sohn K-A, Xing EP, A multivariate regression approach to association analysis of quantitative trait network, Bioinformatics 25 (2009) i204–i212, doi:10.1093/bioinformatics/btp218. [PubMed: 19477989]

[12]. Rothman AJ, Levina E, Zhu J, Sparse multivariate regression with covariance estimation, J Comput Graph Statist 19 (2010) 947–962, doi:10.1198/jcgs.2010.09188.

[13]. Peng J, Zhu J, Bergamaschi A, et al., Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer, Ann Appl Statist 4 (2010) 53–77 10.1214/09-AOAS271SUPP.

[14]. Brown PJ, Fearn T, Vannucci M, The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach, Biometrika 86 (1999) 635–648, doi:10.1093/biomet/86.3.635.

[15]. Lee KH, Tadesse MG, Baccarelli AA, Schwartz J, Coull BA, Multivariate Bayesian variable selection exploiting dependence structure among outcomes: Application to air pollution effects on DNA methylation, Biometrics 73 (2016) 232–241, doi:10.1111/biom.12557. [PubMed: 27377873]

[16]. Kundu D, Mitra R, Gaskins JT, Bayesian variable selection for multioutcome models through shared shrinkage, Scand J Statist (2020) 1–26, doi:10.1111/sjos.12455.

[17]. Kabue S, Liu V, Dyer W, Raebel M, Nichols G, Schmittdiel J, Identifying Common Predictors of Multiple Adverse Outcomes Among Elderly Adults With Type-2 Diabetes, Med Care 57 (2019) 702–709, doi:10.1097/MLR.0000000000001159. [PubMed: 31356411]

[18]. Wei HL, Lang ZQ, Billings SA, Constructing an overall dynamical model for a system with changing design parameter properties, Int J Model Identif Control 5 (2008) 93–104 10.1504/IJMIC.2008.022014.

[19]. Wei HL, Billings SA, Improved model identification for non-linear systems using a random subsampling and multifold modelling (RSMM) approach, Int J Control 82 (2009) 27–42, doi:10.1080/00207170801955420.

[20]. Tibshirani R, Regression shrinkage and selection via the Lasso, J Roy Statist Soc Ser B 58 (1996) 267–288 www.jstor.org/stable/2346178.

[21]. Hastie T, Tibshirani R, Tibshirani R, Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons, Stat Sci 35 (2020) 579–592, doi:10.1214/19-STS733.

[22]. Zhou H, Hastie T, Tibshirani R, On the “degrees of freedom” of the LASSO, Ann Statist 35 (2007) 2173–2192, doi:10.1214/009053607000000127.

[23]. Ahrens A, Hansen CB, Schaffer ME, lassopack: Model selection and prediction with regularized regression in Stata, Stata Journal 20 (2020) 176–235, doi:10.1177/1536867X20909697.

[24]. Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JW, Weir DR, Cohort profile: the Health and Retirement Study (HRS), Int J Epidemiol 43 (2014) 576–585, doi:10.1093/ije/dyu067. [PubMed: 24671021]

[25]. Cox DR, Regression models and life tables, J R Stat Soc Series B 34 (1972) 187–220 https://www.jstor.org/stable/2985181.

[26]. Fine JP, Gray RJ, A proportional hazards model for the subdistribution of a competing risk, J Am Stat Assoc 94 (1999) 496–509, doi:10.1080/01621459.1999.10474144.

[27]. Harrell FE Jr. The PHGLM Procedure. In: SUGI Supplemental Library Users Guide; 1986 Version 5 Edition:437–466. SAS Institute Inc., Cary, NC.

[28]. Harrell FE Jr, Lee KL, Mark DB, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Statist. Med 15 (1996) 361–387 10.1002/(SICI)1097-0258(19960229)15. 4<361::AID-SIM168>3.0.CO;2-4.

[29]. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD, Internal validation of predictive models: efficiency of some procedures for logistic regression analysis, J Clin Epidemiol 54 (2001) 774–781 10.1016/s0895-4356(01)00341-9. [PubMed: 11470385]

[30]. Noma H, Shinozaki T, Iba K, Teramukai S, Furukawa TA. Confidence intervals of prediction accuracy measures for multivariable prediction models based on the bootstrap-based optimism

correction methods. arXiv preprint arXiv:2005.01457. https://arxiv.org/ftp/arxiv/papers/2005/2005.01457.pdf.

[31]. Harden JJ, Kropko J, Simulating Duration Data for the Cox Model, PSRM 7 (2019) 921–928, doi:10.1017/psrm.2018.19.

[32]. Wolbers M, Koller MT, Witteman JC, Steyerberg EW, Prognostic models with competing risks: methods and application to coronary risk prediction, Epidemiology 20 (2009) 555–561, doi:10.1097/EDE.0b013e3181a39056. [PubMed: 19367167]

[33]. Lee SJ, Smith AK, Ramirez-Diaz LG, Covinsky KE, Gan S, Chen CL, Boscardin WJ, A Novel Metric for Developing Easy-To-Use and Accurate Clinical Prediction Models: The Time-Cost Information Criterion, Med Care (2021), doi:10.1097/MLR.0000000000001510.

[34]. Jiang H, Symanowski J, Paul S, Qu Y, Zagar A, Hong S, The type I error and power of non-parametric logrank and Wilcoxon tests with adjustment for covariates—A simulation study, Statist Med 27 (2008) 5850–5860, doi:10.1002/sim.3406.

[35]. Sofer T, Dicker L, Lin X, Variable selection for high dimensional multivariate outcomes, Stat Sin 24 (2014) 1633–1654 10.5705/ss.2013.019. [PubMed: 28642637]

[36]. Zhang H, Zheng Y, Yoon G, et al., Regularized estimation in sparse high dimensional multivariate regression, with application to a DNA methylation study, Stat Appl Genet Mol Biol 16 (2017) 159–171, doi:10.1515/sagmb-2016-0073. [PubMed: 28734115]

[37]. Heinze G, Wallisch C, Dunkler D, Variable selection - A review and recommendations for the practicing statistician, Biom J 60 (2018) 431–449 10.1002/bimj.201700067. [PubMed: 29292533]

[38]. Steyerberg EW, Gail M, Tsiatis A, Krickeberg K, Wong W, Sarnet J, Disadvantages of Stepwise Methods, in: Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating, Springer, 2009, pp. 197–204.

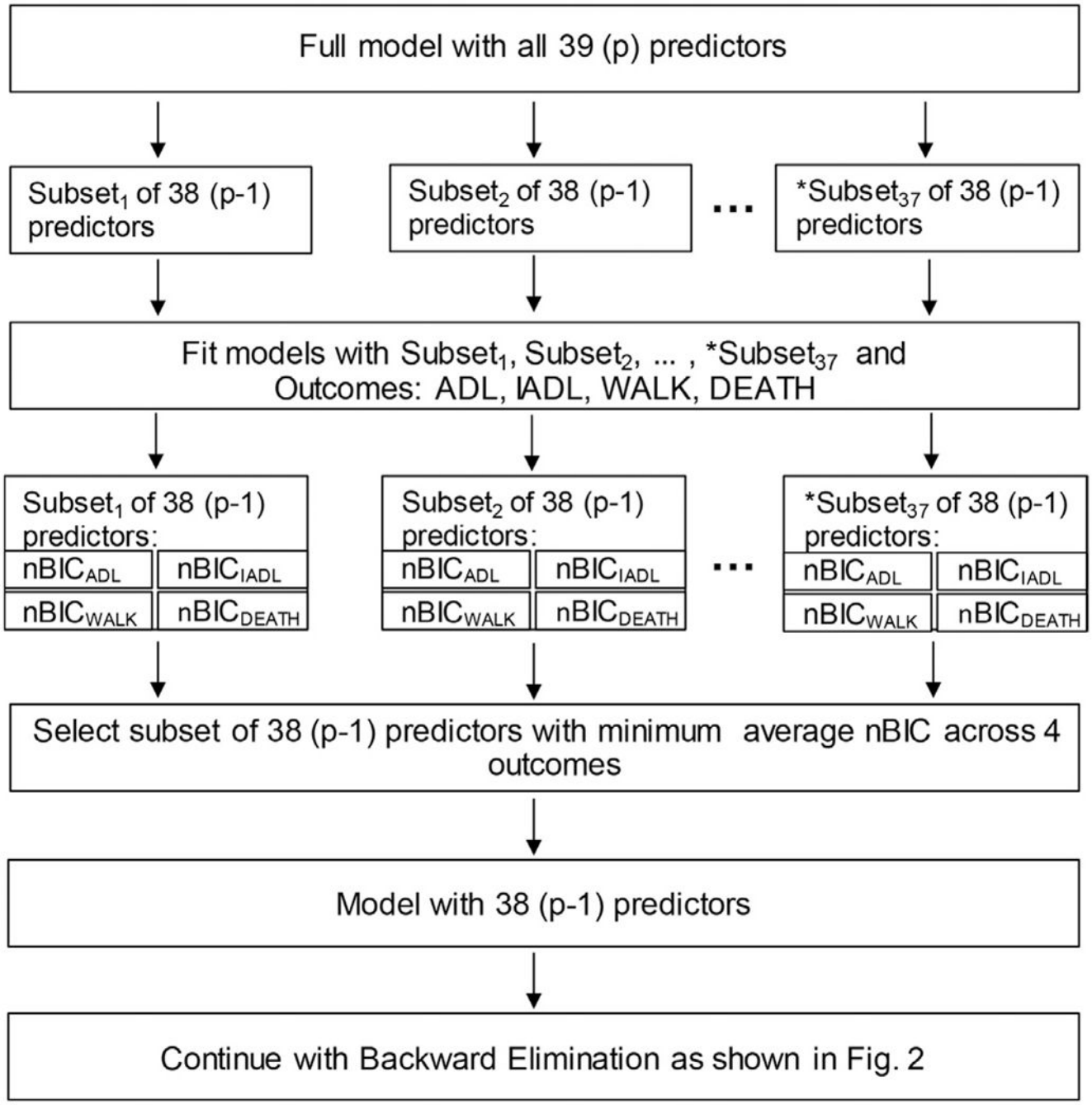

Fig. 1.

Overview of Algorithm for the Selection of Subset of (p-1) Predictors with Minimum Average Normalized BIC across 4 Outcomes.

ADL: time to first Activities of Daily Living (ADL) dependence.

DEATH: time to death.

IADL: time to first Instrumental Activities of Daily Living (IADL) difficulty.

nBIC: normalized Bayesian Information Criterion.

p: number of predictors.

$Subset_1$, $Subset_2$, $Subset_{37}$: combination of predictors obtained by removing 1 predictor at a time.

*$Subset_{37}$: Maximum number of subsets of predictors fitted in the first step of backward elimination. In the first step, the full model has 39 predictors, and there are 2 predictors that are forced into all models. Thus, the maximum number of subsets fitted by removing one predictor at a time is 37 since the initial pool contains 37 available predictors.

WALK: time to first mobility dependence.

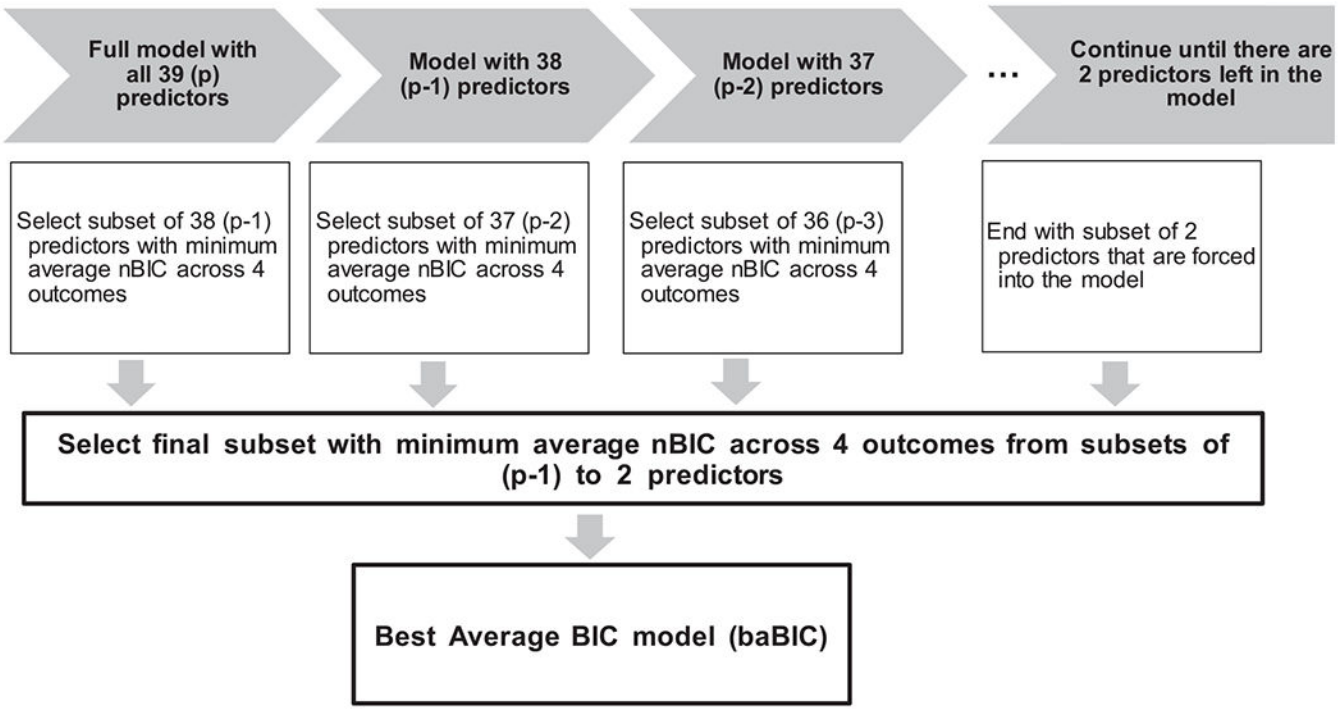

Fig. 2.

Overview of Algorithm for the Selection of Final Subset of Predictors with Minimum Average Normalized BIC across 4 outcomes.

nBIC: normalized Bayesian Information Criterion.

p: number of predictors.

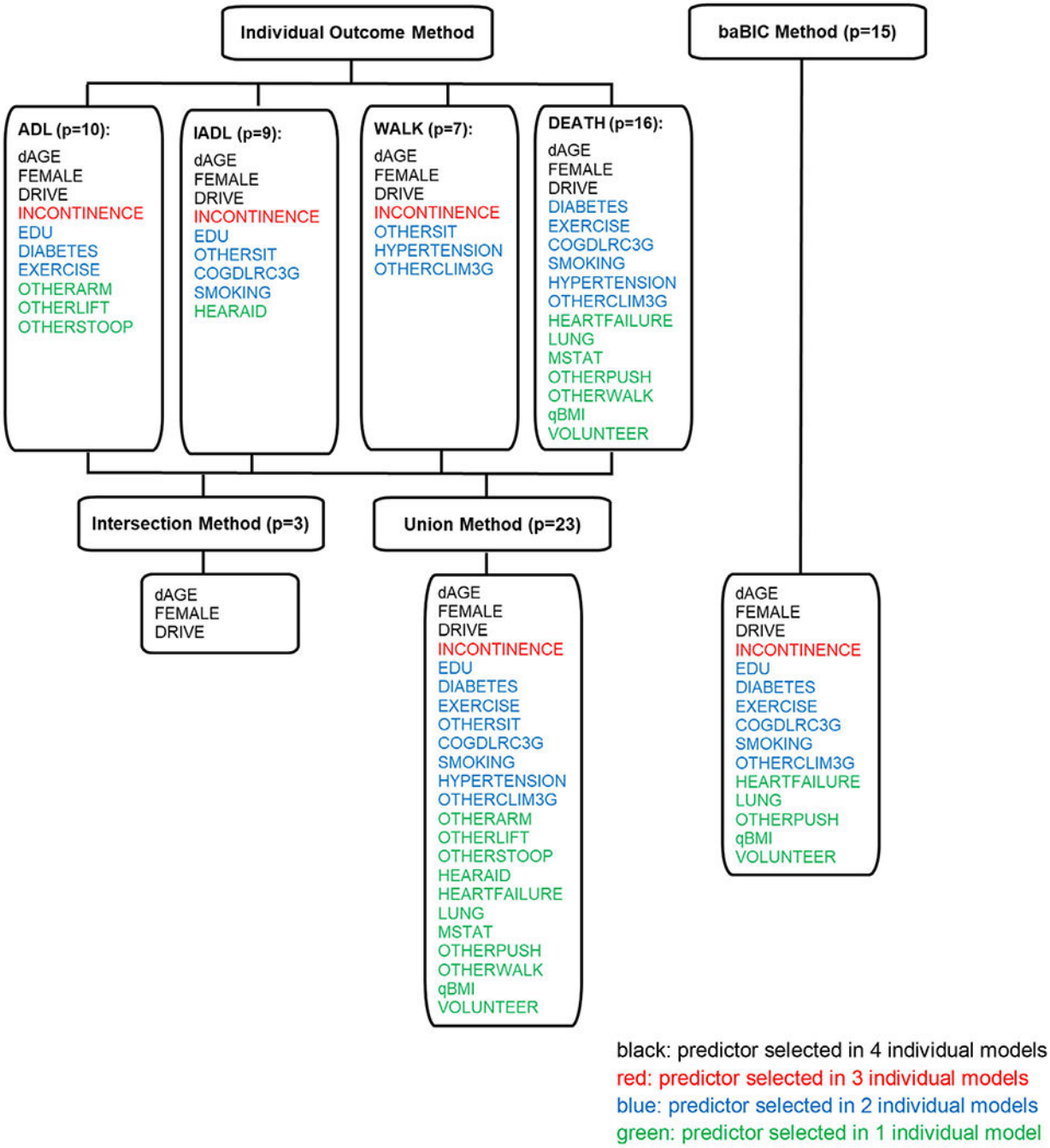

Fig. 3.

Subsets of Predictors selected with Individual Outcome, Union, baBIC, and Intersection Methods using the Case-study Data.

ADL: time to first Activities of Daily Living (ADL) dependence.

baBIC Method: best Average BIC method, selects best subset of predictors based on the minimum average normalized BIC across the 4 outcomes.

BIC: Bayesian Information Criterion.

COGDLRC3G: number of words from 10-word list recalled correctly after 5 minutes.

dAGE: age deciles groups.
DEATH: time to death.
DIABETES: whether has diabetes with and without medicine.
DRIVE: whether able to drive.
EDU: education 12+ years.
EXERCISE: exercise frequency.
FEMALE: whether female.
HEARAID: whether wears hearing aid.
HEARTFAILURE: whether has heart failure or others heart problems (e.g. angina, heart attack, heart disease).
HYPERTENSION: whether has hypertension.
IADL: time to first Instrumental Activities of Daily Living (IADL) difficulty.
INCOTINENCE: whether has incontinence.
Individual Outcome Method: selects final subset of predictors based on the minimum BIC for each individual outcome.
Intersection Method: selects final subset of predictors that were in all 4 final subsets based on the minimum BIC for each individual outcome.
LUNG: chronic lung disease.
MSTAT: marital status.
OTHERARM: having difficulty reaching above shoulder.
OTHERCLIM3G: having difficulty climbing stairs.
OTHERLIFT: having difficulty with lifting weights over 10 pounds.
OTHERPUSH: having difficulty with pushing large objects.
OTHERSIT: having difficulty with sitting for 2 hours.
OTHERSTOOP: having difficulty with stooping, kneeling, or crouching.
OTHERWALK: having difficulty with walking one block or in the room.
p: number of predictors.
qBMI: quintile groups.
SMOKING: whether smokes.
Union Method: selects final subset of all the predictors that were in at least 1 of the 4 final subsets based on the minimum BIC for each individual outcome.
VOLUNTEER: whether helps as volunteer.
WALK: time to first mobility dependence.

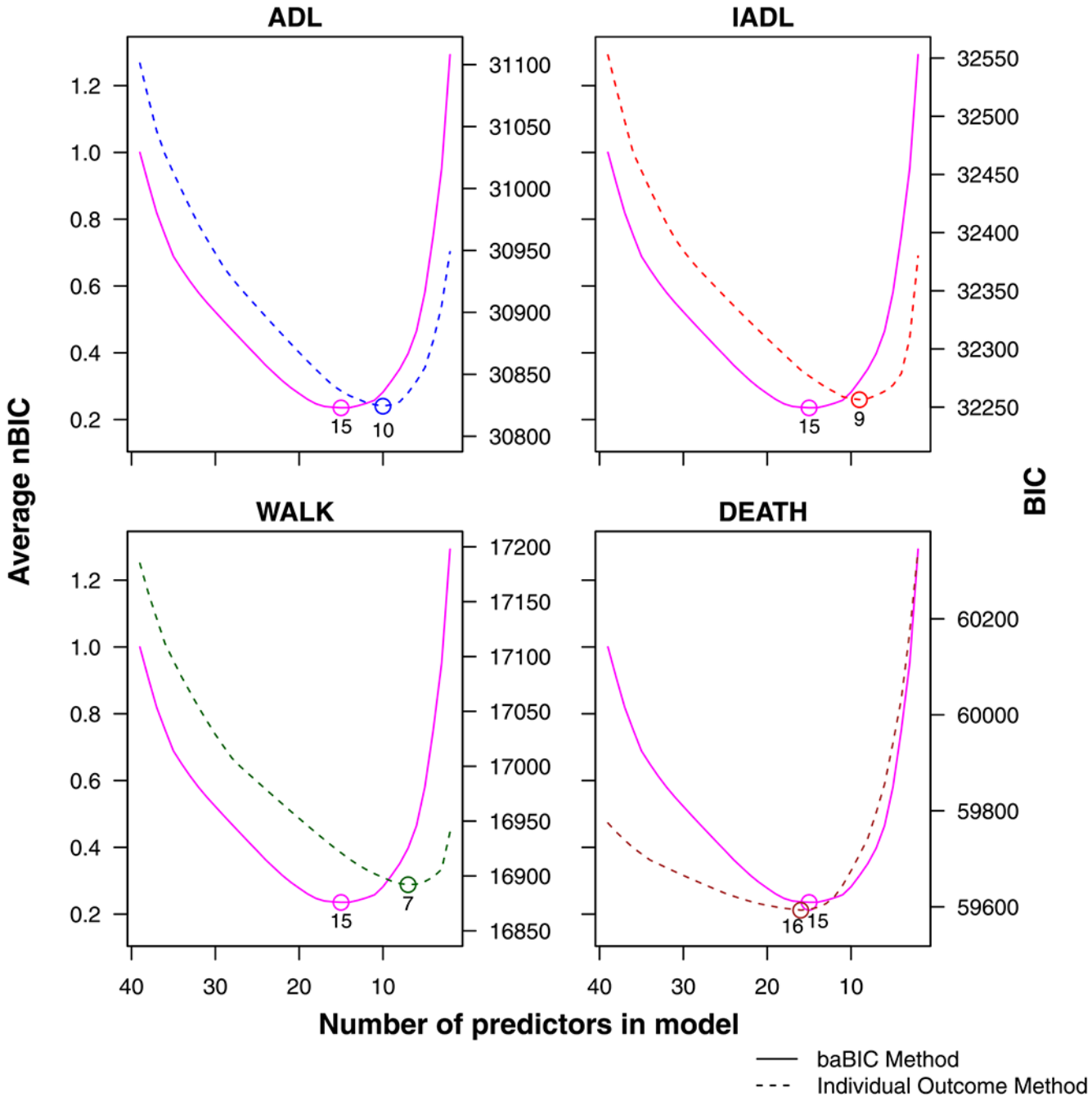

Fig. 4.

Selection with the baBIC Method and Individual Outcome Methods in the Case-study Data.

ADL: time to first Activities of Daily Living (ADL) dependence.

baBIC Method: best Average BIC method, selects best subset of predictors based on the minimum average normalized BIC across the 4 outcomes.

BIC: Bayesian Information Criterion.

DEATH: time to death.

IADL: time to first Instrumental Activities of Daily Living (IADL) difficulty.

Individual Outcome Method: selects final subset of predictors based on the minimum BIC for each individual outcome.
nBIC: normalized Bayesian Information Criterion.
WALK: time to first mobility dependence.

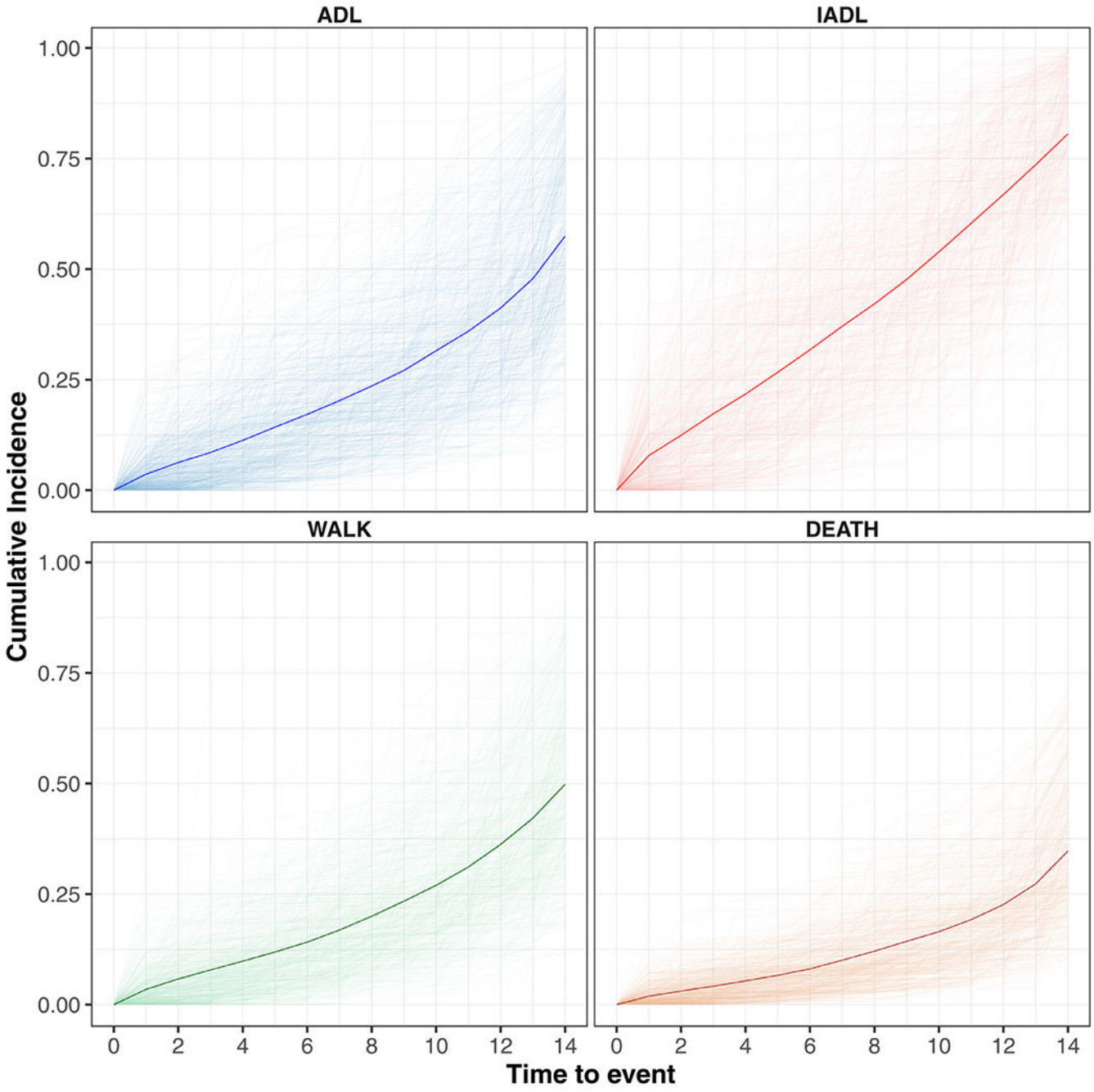

Fig. 5.

Predicted Cumulative Incidence by Outcome at the Mean of the Predictors Selected with the baBIC Method in the Case-study Data using simulations (lighter color) of Scenario 1 with Case-study Levels of Censoring (darker color: mean of simulations).

ADL: time to first Activities of Daily Living (ADL) dependence.

Case-study levels of censoring: ADL= 66.55%, IADL= 64.98%, WALK=81.90%, DEATH=31.87%.

DEATH: time to death.

IADL: time to first Instrumental Activities of Daily Living (IADL) difficulty.
Scenario 1: simulated data generated using 15 non-zero coefficients corresponding to the common subset of predictors obtained with the baBIC method in the case-study data.
WALK: time to first mobility dependence.

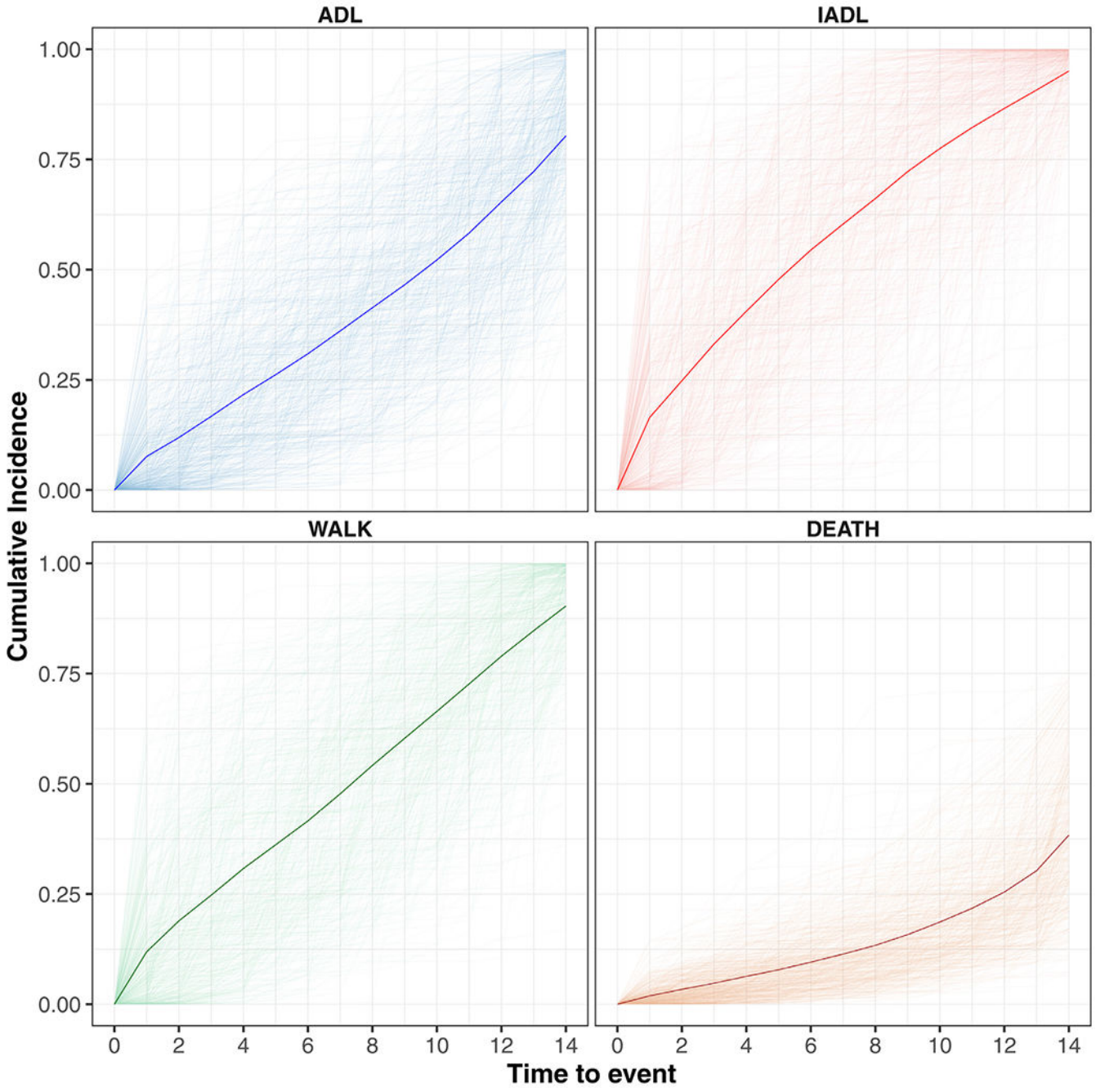

Fig. 6.

Predicted Cumulative Incidence by Outcome at the Mean of the Predictors Selected with the baBIC Method in the Case-study Data using simulations (lighter color) of Scenario 1 with 25% Censoring (darker color: mean of simulations).

ADL: time to first Activities of Daily Living (ADL) dependence.

DEATH: time to death.

IADL: time to first Instrumental Activities of Daily Living (IADL) difficulty.

Scenario 1: simulated data generated using 15 non-zero coefficients corresponding to the common subset of predictors obtained with the baBIC method in the case-study data.

WALK: time to first mobility dependence.

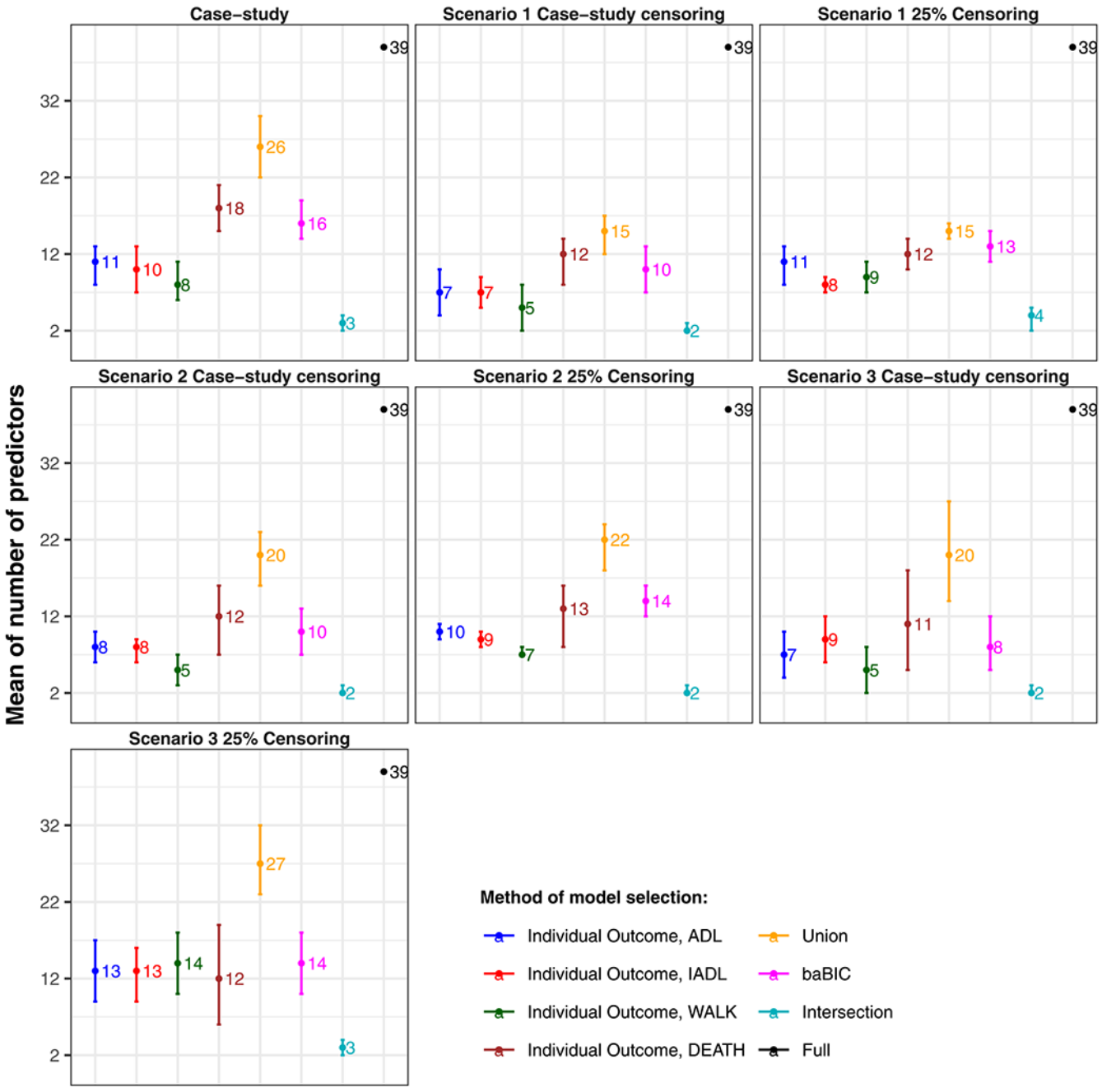

Fig. 7.

Comparison of Number of Predictors Selected (mean, 2.5th -97.5th percentiles) Across Case-study Bootstrap Data and Simulations with Case-study Levels of Censoring and 25% Censoring.

ADL: time to first Activities of Daily Living (ADL) dependence.

baBIC Method: best Average BIC method, selects best subset of predictors based on the minimum average normalized BIC across the 4 outcomes.

Case-study levels of censoring: ADL= 66.55%, IADL= 64.98%, WALK=81.90%, DEATH=31.87%.

DEATH: time to death.
BIC: Bayesian Information Criterion.
DEATH: time to death.
Full Method: includes all 39 candidate predictors of the case-study data.
IADL: time to first Instrumental Activities of Daily Living (IADL) difficulty.
Individual Outcome Method: selects final subset of predictors based on the minimum BIC for each individual outcome.
Intersection Method: selects final subset of predictors that were in all 4 final subsets based on the minimum BIC for each individual outcome.
Scenario 1: simulated data generated using 15 non-zero coefficients corresponding to the common subset of predictors obtained with the baBIC method in the case-study data.
Scenario 2: simulated data generated using the outcome specific non-zero coefficients corresponding to those obtained with the Individual Outcome method in the case-study data.
Scenario 3: simulated data generated using non-zero coefficients for all 39 candidate predictors using estimates from the case-study data.
Union Method: selects final subset of all the predictors that were in at least 1 of the 4 final subsets based on the minimum BIC for each individual outcome.
WALK: time to first mobility dependence.

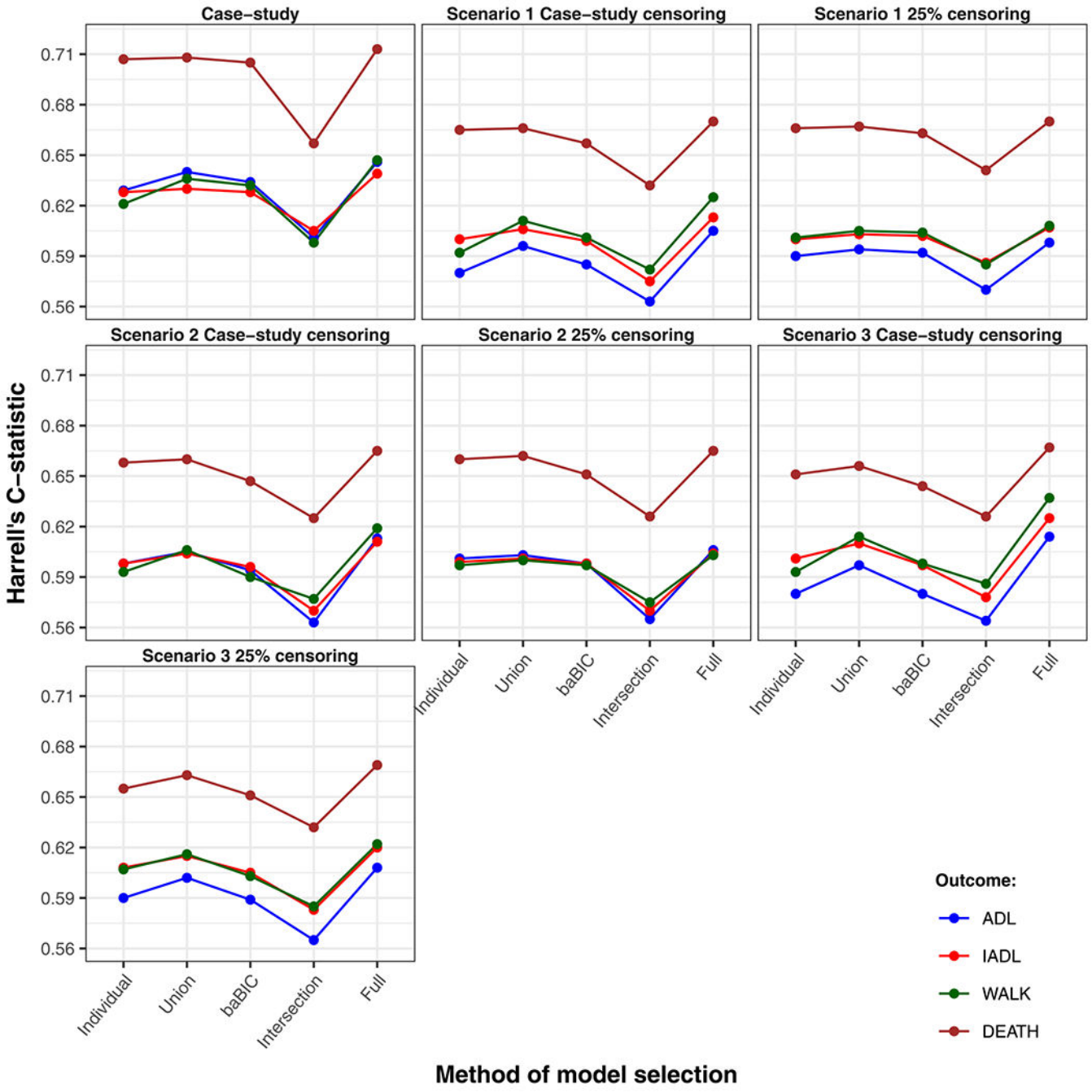

Fig. 8.

Comparison of Mean Harrell's C-statistic Across Case-study Bootstrap Data and Simulations with Case-study Levels of Censoring and 25% Censoring.

ADL: time to first Activities of Daily Living (ADL) dependence.

baBIC Method: best Average BIC method, selects best subset of predictors based on the minimum average normalized BIC across the 4 outcomes.

BIC: Bayesian Information Criterion.

Case-study levels of censoring: ADL= 66.55%, IADL= 64.98%, WALK=81.90%, DEATH=31.87%.

DEATH: time to death.

Full Method: includes all 39 candidate predictors of the case-study data.

IADL: time to first Instrumental Activities of Daily Living (IADL) difficulty.

Individual Outcome Method: selects final subset of predictors based on the minimum BIC for each individual outcome.

Intersection Method: selects final subset of predictors that were in all 4 final subsets based on the minimum BIC for each individual outcome.

Scenario 1: simulated data generated using 15 non-zero coefficients corresponding to the common subset of predictors obtained with the baBIC method in the case-study data.

Scenario 2: simulated data generated using the outcome specific non-zero coefficients corresponding to those obtained with the Individual Outcome method in the case-study data.

Scenario 3: simulated data generated using non-zero coefficients for all 39 candidate predictors using estimates from the case-study data.

Union Method: selects final subset of all the predictors that were in at least 1 of the 4 final subsets based on the minimum BIC for each individual outcome.

WALK: time to first mobility dependence.

Table 1

Percentage of Predictor Inclusion with the baBIC Method in Simulations with Case-study Levels of Censoring and 25% Censoring.

	Inclusion Percentage in Simulations					
	Scenario 1		Scenario 2		Scenario 3	
Censoring	Case-study	25%	Case-study	25%	Case-study	25%
Inclusion Percentage for the 15 individual predictors selected in the case-study data						
dAGE[a]	100.0	100.0	100.0	100.0	100.0	100.0
FEMALE[b]	100.0	100.0	100.0	100.0	100.0	100.0
DRIVE	92.2	100.0	91.4	100.0	36.6	93.4
INCONTINENCE	73.6	100.0	83.6	100.0	49.0	99.8
EDU	59.4	100.0	80.6	100.0	48.4	91.0
DIABETES	98.8	99.6	99.2	99.4	66.0	72.6
EXERCISE	68.2	99.8	62.4	88.0	25.6	82.8
COGDLRC3G	99.4	100.0	88.8	100.0	92.0	100.0
SMOKING	100.0	100.0	98.2	99.8	98.6	100.0
OTHERCLIM3G	40.4	100.0	12.4	97.2	9.2	42.2
HEARTFAILURE	24.0	42.2	5.8	7.0	3.8	8.4
LUNG	7.8	18.8	5.2	4.2	11.4	7.4
OTHERPUSH	38.0	88.8	5.6	9.0	0.6	2.2
qBMI	11.4	80.4	3.4	6.4	3.8	32.0
VOLUNTEER	44.8	61.2	20.0	22.0	2.8	5.2
Average inclusion percentage for the 15 predictors selected in the case-study data						
	58.3	83.9	50.5	64.1	34.4	56.7
Average inclusion percentage for the 24 predictors not selected in the case-study data						
	0.0	0.0	4.6	13.4	5.1	18.1
Number of predictors chosen from the 15 predictors selected in the case-study data						
3					1.4	
4			0.2		5.6	
5	0.2		1.6		18.0	
6	1.8		4.6		27.8	0.8
7	11.2		13.0		25.4	9.0

	Inclusion Percentage in Simulations					
	Scenario 1		Scenario 2		Scenario 3	
Censoring	Case-study	25%	Case-study	25%	Case-study	25%
8	19.4		29.2	0.6	13.8	21.6
9	19.4	0.2	31.6	9.0	5.4	23.8
10+	48.0	99.8	19.8	90.4	2.6	44.8

baBIC method: best Average BIC method, selects best subset of predictors based on the minimum average normalized BIC across the 4 outcomes.

BIC: Bayesian Information Criterion.

Case-study levels of censoring: ADL= 66.55%, IADL= 64.98%, WALK=81.90%, DEATH=31.87%. COGDLRC3G: number of words from 10-word list recalled correctly after 5 minutes.

dAGE[a] (age deciles groups), FEMALE[b] (whether female): predictors that are forced into all models.

DIABETES: whether has diabetes with and without medicine.

DRIVE: whether able to drive.

EDU: education 12+ years.

EXERCISE: exercise frequency.

HEARTFAILURE: whether has heart failure or others heart problems (e.g. angina, heart attack, heart disease).

INCOTINENCE: whether has incontinence.

LUNG: chronic lung disease.

OTHERCLIM3G: having difficulty climbing stairs.

OTHERPUSH: having difficulty with pushing large objects.

qBMI: quintile groups.

Scenario 1: simulated data generated using 15 non-zero coefficients corresponding to the common subset of predictors obtained with the baBIC method in the case-study data.

Scenario 2: simulated data generated using the outcome specific non-zero coefficients corresponding to those obtained with the Individual Outcome method in the case-study data.

Scenario 3: simulated data generated using non-zero coefficients for all 39 candidate predictors using estimates from the case-study data.

SMOKING: whether smokes.

VOLUNTEER: whether helps as volunteer.