Issues in prognostic model building

ASA/SDSA Webinar

John Boscardin

May 8, 2025

Outline

- Introduction to our group
- Background/setting/use case
- Machine Learning vs. Traditional Regression
- Feature development and selection
- Internal validation
- Leveraging the leaderboard
- Stability of individual predictions

(Partial) Acknowledgement List

- Statistical Laboratory for Aging Research: Cenzer, Diaz-Ramirez, Espejo, Fung, Gan, Jeon, Jing, Lu, Patel, Shi
- Geriatrics: Covinsky, Deardorff, Greene, Growden, Harrison, Kotwal, Lam, Sei Lee, Alex Lee, Newman, Schwarz, Smith, Steinman, Sudore, Tang, Walter
- Pepper Friends: Bock, Bongiovanni, Chen, Cobert, Huang, Hunt, Lahue, Makam, Nouri, Oh, Whitlock, Wong
- VA Mental Health: Barnes, Byers, Kornblith, Yaffe
- Epi Mentees: Duchowny, Kim, Sims
- Research Admin Team: Haller, Kang, Ngo, Shahroodi, Shiff, Yu, Yuan
- Mt. Sinai ("UCSF East") P01 Team

UCSF Statistical Laboratory for Aging Research

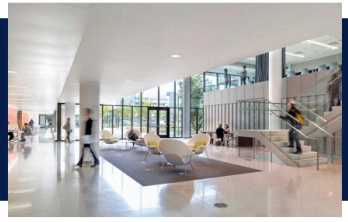
UCSF Division of Geriatrics
Department of Medicine

Search...

Programs | Innovation | News & Events | Resources | About Us

Home > Programs > Resource Cores > Data and Analysis Core > Statistical Laboratory

Statistical Laboratory



About Us

The Pepper Center Data and Analysis Core Statistical Laboratory comprises 9 stellar master's and PhD-level data analysts supervised by Pepper Center Data and Analysis Core Director W. John Boscardin, PhD.

The statistical laboratory supports investigators who are conducting quantitative research, including research using data resources supported by the Pepper Center. The laboratory has also pushed the envelope on advancing methods research by conducting original investigations statistical methods that have special relevance to aging research. Finally, the laboratory hosts a publicly available Github site where relevant statistical code from our investigations can be shared with other research teams.

Current Biostatisticians

W. John Boscardin, PhD
Professor of Medicine,
Data and Analysis Core Director
VIEW PROFILE

Irena Cenzer, PhD
Senior Statistician
VIEW PROFILE

Grisell Diaz-Ramirez, MS
Senior Statistician
VIEW PROFILE

Edie Espejo, MA
Statistician
VIEW PROFILE

Data and Analysis Core

- Statistical Laboratory
- Division of Geriatrics - CMS Data ReUse Initiative
- DAC Consultations
- Dataset Resource Library

UCSF Geriatrics Github Code Repository

Find the compendium our code repository developed by our biostatisticians and used to create our analysis for our manuscripts.

VISIT GITHUB

Kathy Fung, MS
Senior Statistician
VIEW PROFILE

Siqi Gan, MPH
Senior Statistician
VIEW PROFILE

Sun Young Jeon, MS, PhD
Senior Statistician
VIEW PROFILE

Bocheng Jing, MS
Senior Statistician
VIEW PROFILE

Kanan Patel, MPH, MBBS
Academic Specialist
VIEW PROFILE

Ying Shi, PhD
Senior Statistician
VIEW PROFILE

Model for Laboratory

- Grown from 2 to 10 core statisticians in lab (Cenzer, Diaz-Ramirez, Espejo, Fung, Gan, Jeon, Jing, Lu, Patel, Shi)
- Team-science framework with emphasis on deep, longitudinal collaboration
- *Statistical scientist*: statistician is key member of the research group and accumulates experience in data, methods, and substantive area
- Clinical researchers (PIs and trainees) are interested in the methodological details
- Projects/teams have more than one statistician in most cases
- Statistical and data science mentoring occurs in all directions (me, statisticians, PI, co-I, trainee)

Model for Consultation/Collaboration

- Faculty investigators in UCSF Division of Geriatrics/Pepper Center (and their mentees)
- Outside investigators currently supported by Pepper Center (Scholars or Pilot Awardees)
- Outside investigators formerly supported by Pepper Center
- Business model:
  - P30 Pepper Statistical Core (DAC): direct funding (I am Co-Director of the core)
  - P30 Pepper Pilot/Training Cores (PESC/REC): spending awards on statistical support
  - P01 Mt. Sinai/UCSF Statistical Core (RCB): direct funding (I am Co-Director of the core)
  - R and K and other funding from Geriatrics investigators
  - R and K and other funding from outside investigators (substantial component and key to financial stability)

Reasons to develop a predictive model

- Precision medicine: flagging high risk patients or those likely to benefit
- (Shared) decision making for patients, caretakers, physicians
- Case-mix adjustment
- Propensity score for subsequent analytic purposes

Point Scoring (Sullivan et al. 2004)

Risk factor	Categories	Points
Age		
	30–39	0
	40–49	2
	50–59	4
	60–69	6
	70–79	8
Sex		
	Female	0
	Male	5
Systolic blood pressure		
	<120	−1
	120–129	0
	130–139	1
	140–159	2
	$\geqslant 160$	3
Current smoker		
	No	0
	Yes	3

Point total	Estimate of risk
−1	0.0015
0	0.0020
1	0.0026
2	0.0035
3	0.0047
4	0.0062
5	0.0083
6	0.0110
7	0.0147
8	0.0195
9	0.0258
10	0.0341
11	0.0449
12	0.0590
13	0.0771
14	0.1002
15	0.1293
16	0.1652
17	0.2088
18	0.2602

Nomogram (Harrell)

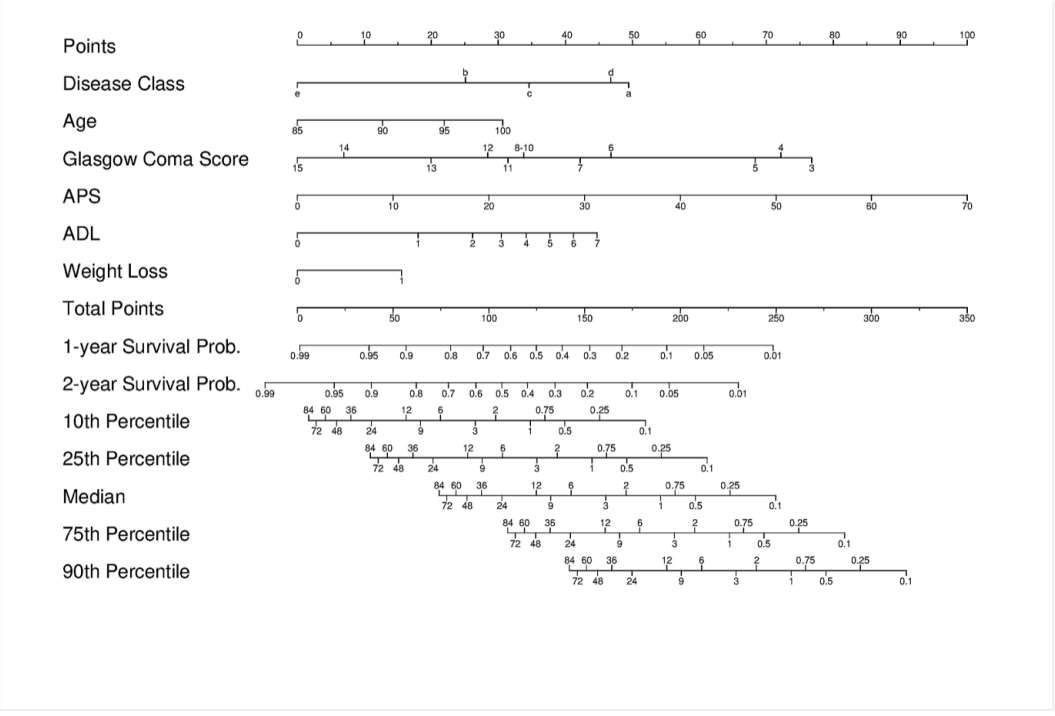
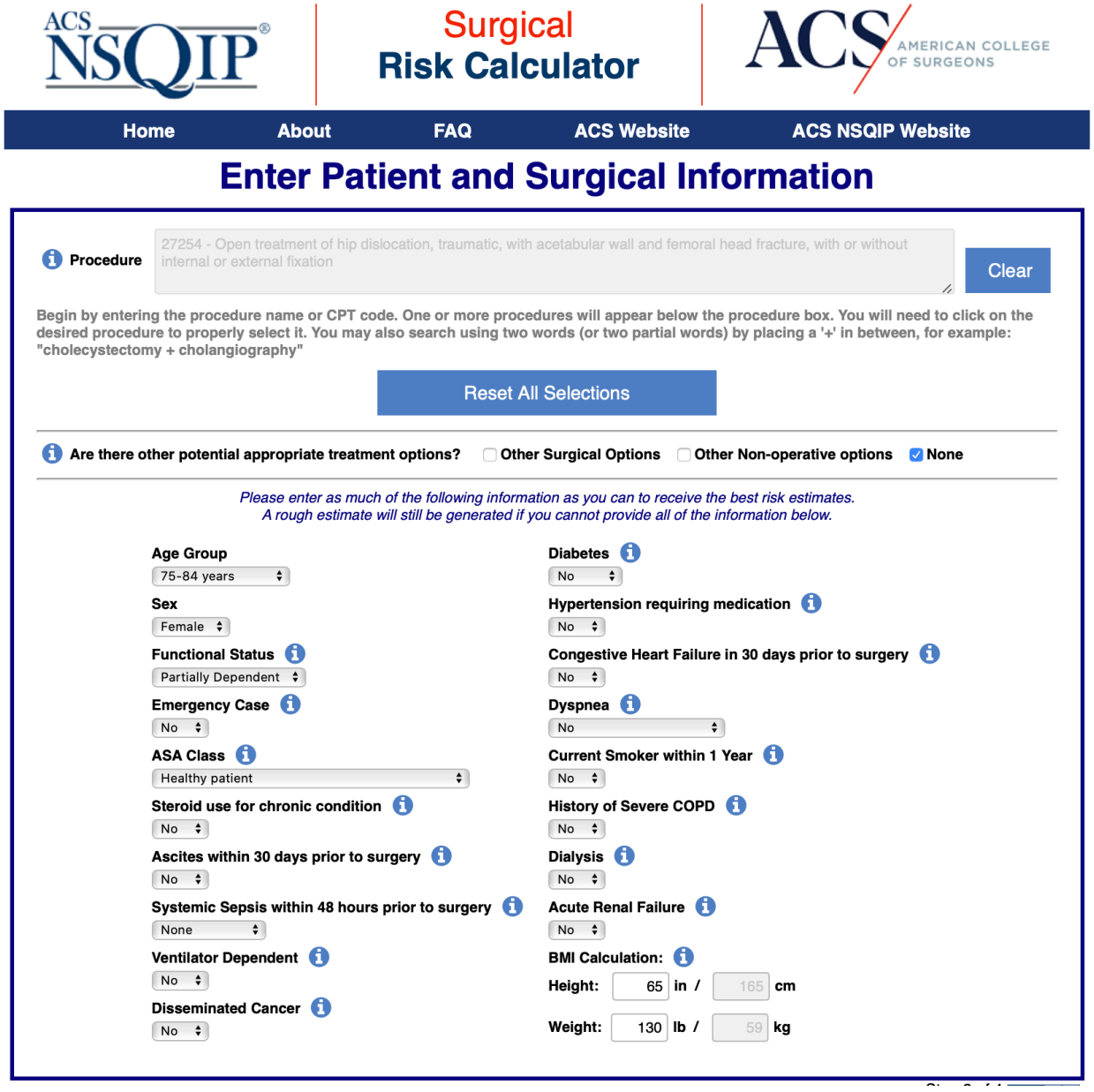


FIGURE 1:
Nomogram for obtaining predicted 1- and 2-year survival probabilities and the 10th, 25th, 50th, 75th, and 90th percentiles of survival time (in months) for individual patients in HELP. Disease class abbreviations: a=ARF/MOSF/Coma, b=all others, c=CHF, d=Cancer, e=Orthopedic. To use the nomogram, place a ruler vertically such that it touches the appropriate value on the axis for each predictor. Read off where the ruler intersects the 'Points' axis at the top of the diagram. Do this for each predictor, making a listing of the points. Add up all these points and locate this value on the 'Total Points' axis with a vertical ruler. Follow the ruler down and read off any of the predicted values of interest. APS is the APACHE III Acute Physiology Score.

ACS NSQIP input



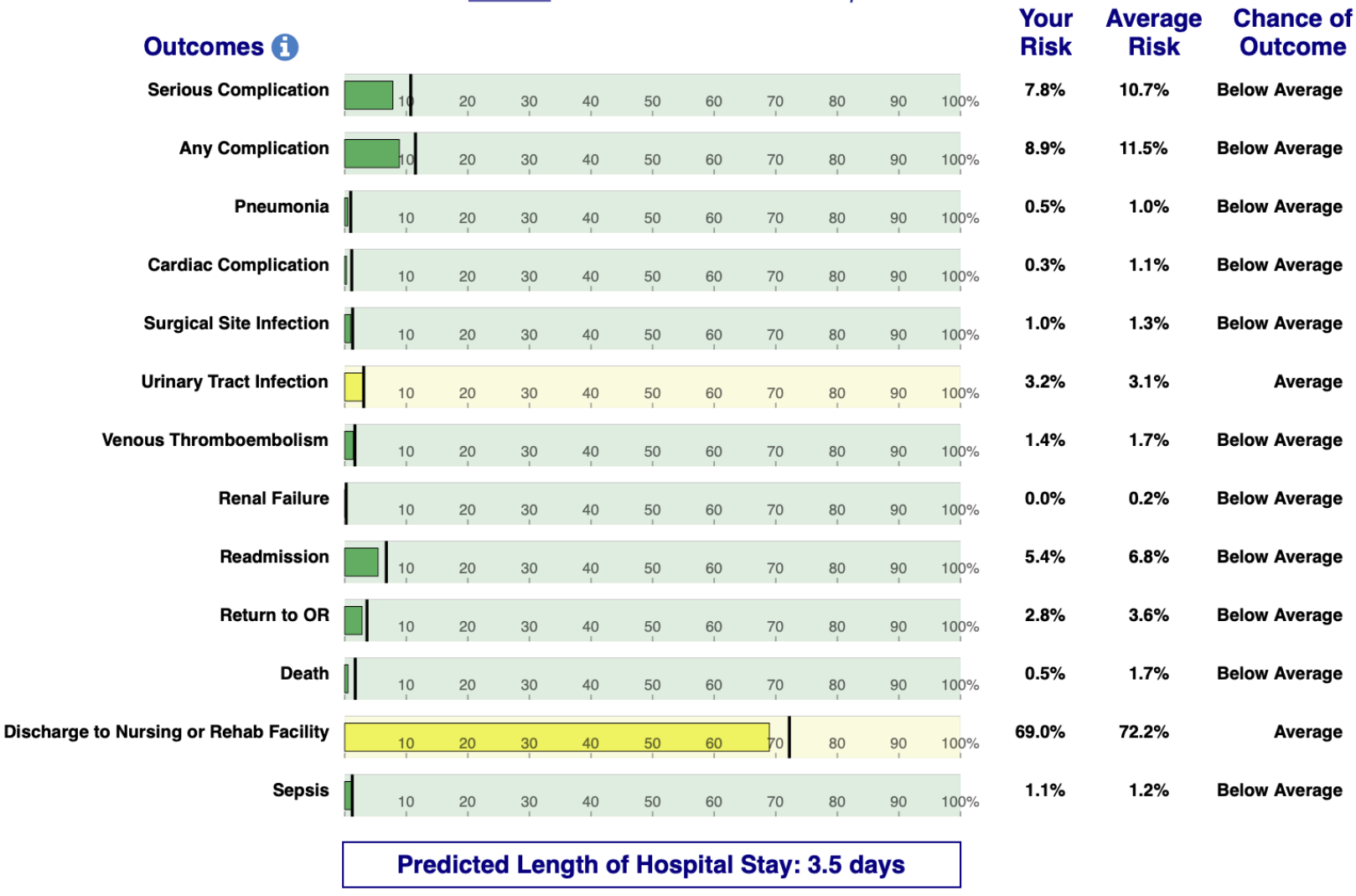
ACS NSQIP output

Procedure: 27254 - Open treatment of hip dislocation, traumatic, with acetabular wall and femoral head fracture, with or without internal or external fixation

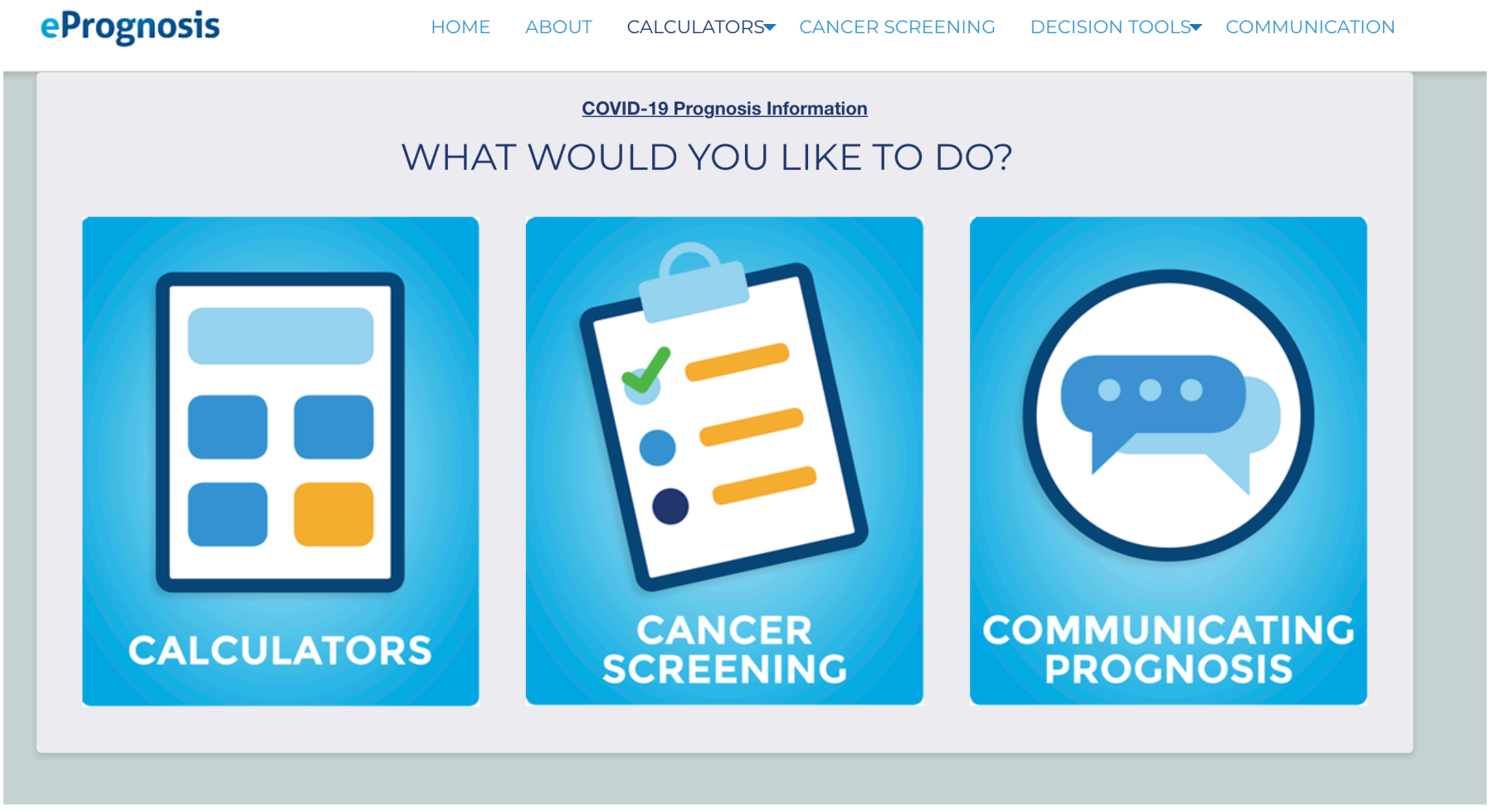
Risk Factors: 75-84 years, Partially dependent functional status, Mild systemic disease

Change Patient Risk Factors

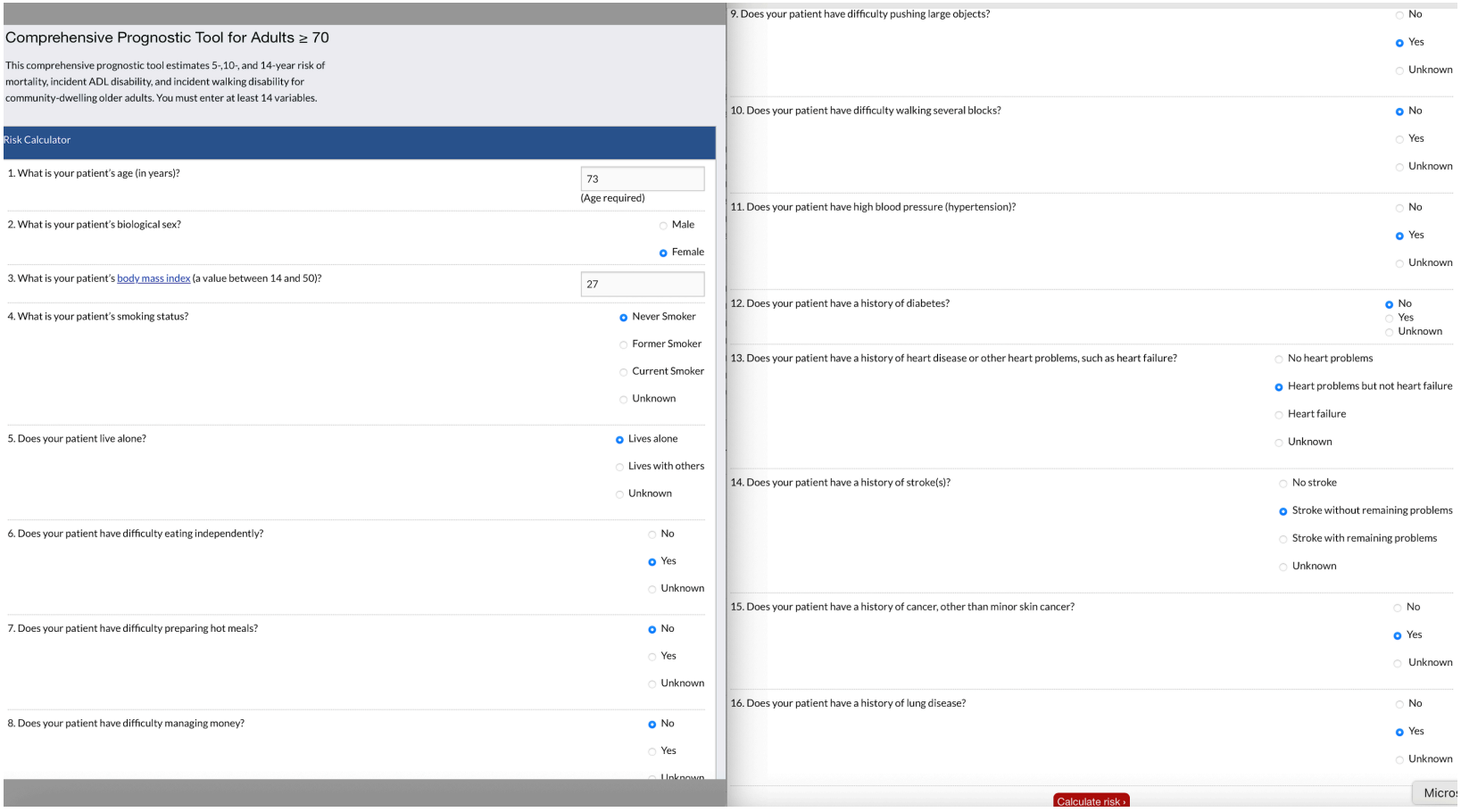
Note: Your Risk has been rounded to one decimal point.



ePrognosis (UCSF)



ePrognosis input example (A. Lee et al. ,2022)



Comprehensive Prognostic Tool for Adults ≥ 70

This comprehensive prognostic tool estimates 5-,10-, and 14-year risk of mortality, incident ADL disability, and incident walking disability for community-dwelling older adults. You must enter at least 14 variables.

Risk Calculator

1. What is your patient's age (in years)? 73 (Age required)
2. What is your patient's biological sex? Male / Female
3. What is your patient's body mass index (a value between 14 and 50)? 27
4. What is your patient's smoking status? Never Smoker / Former Smoker / Current Smoker / Unknown
5. Does your patient live alone? Lives alone / Lives with others / Unknown
6. Does your patient have difficulty eating independently? No / Yes / Unknown
7. Does your patient have difficulty preparing hot meals? No / Yes / Unknown
8. Does your patient have difficulty managing money? No / Yes / Unknown
9. Does your patient have difficulty pushing large objects? No / Yes / Unknown
10. Does your patient have difficulty walking several blocks? No / Yes / Unknown
11. Does your patient have high blood pressure (hypertension)? No / Yes / Unknown
12. Does your patient have a history of diabetes? No / Yes / Unknown
13. Does your patient have a history of heart disease or other heart problems, such as heart failure? No heart problems / Heart problems but not heart failure / Heart failure / Unknown
14. Does your patient have a history of stroke(s)? No stroke / Stroke without remaining problems / Stroke with remaining problems / Unknown
15. Does your patient have a history of cancer, other than minor skin cancer? No / Yes / Unknown
16. Does your patient have a history of lung disease? No / Yes / Unknown

Calculate risk ›

ePrognosis output example (A. Lee et al. ,2022)

ePrognosis

HOME ABOUT CALCULATORS CANCER SCREENING DECISION TOOLS COMMUNICATION

Comprehensive Prognostic Tool for Adults ≥ 70

This comprehensive prognostic tool estimates 5-,10-, and 14-year risk of mortality, incident ADL disability, and incident walking disability for community-dwelling older adults. You must enter at least 14 variables.

Scroll to the bottom for more detailed information.

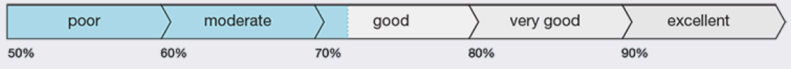
	Mortality		ADL Disability*		Walking Disability**	
	YOUR PATIENT	AVERAGE FOR AGE	YOUR PATIENT	AVERAGE FOR AGE	YOUR PATIENT	AVERAGE FOR AGE
5-year risk	**33%**	16%	**30%**	18%	**16%**	11%
10-year risk	**70%**	37%	**56%**	35%	**33%**	23%
14-year risk	**90%**	56%	**68%**	46%	**43%**	31%
Compare to others your patient's age your patient's risk at 10 year is:	**Higher than average**		**Higher than average**		**Higher than average**	

* ADL Disability: Needing help or unable to do 1 of the 5 ADLS

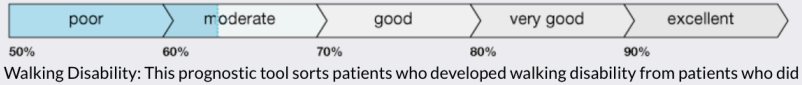
** Walking Disability: Needing help or unable to walk across the room

Finish

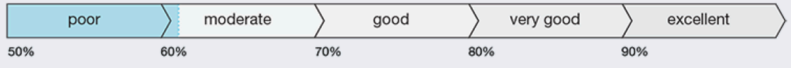
- The Comprehensive Prognostic Tool was developed in 6646 community-dwelling U.S. adults aged ≥70 years who were interviewed in the Health Retirement Survey in 2000 (mean age 78, 59% female, 86% white).
- This Tool was internally validated using bootstrapping with the same 6646 participants from the Health and Retirement Study.
- Discrimination:
  - Mortality: This prognostic tool sorts patients who died from patients who lived correctly 72% of the time (c-statistic).



  - ADL Disability: This prognostic tool sorts patients who developed ADL disability from patients who did not develop ADL disability 63% of the time (c-statistic).



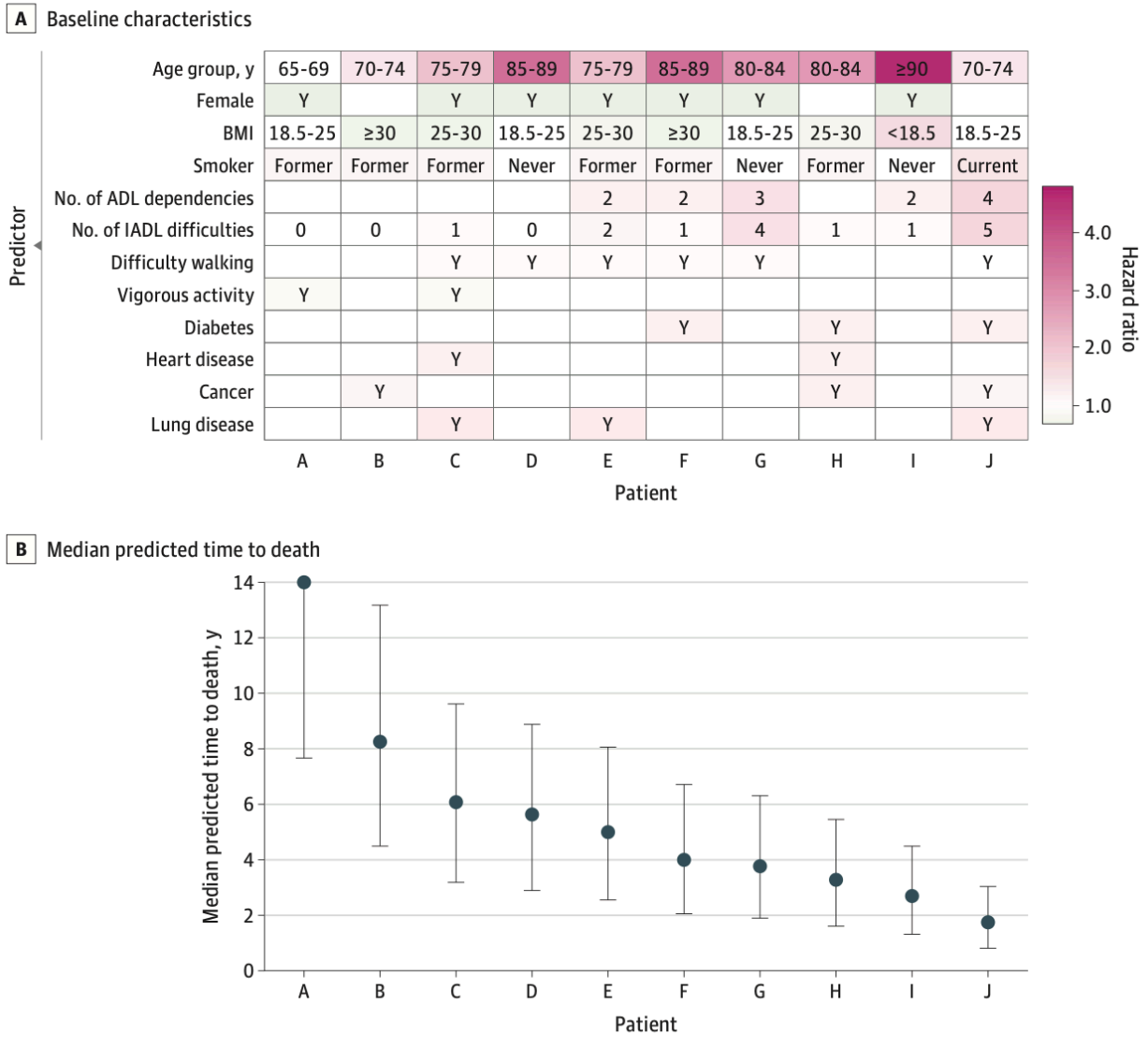
  - Walking Disability: This prognostic tool sorts patients who developed walking disability from patients who did not develop walking disability 61% of the time (c-statistic).



- Calibration:
  - The mortality model was well calibrated across all risk levels, with less than 7% difference between estimated and actual 10-year mortality rates.
  - The ADL disability model was well calibrated across all risk levels, with less than 6% difference between estimated and actual 10-year ADL disability rates.
  - The walking disability model was well calibrated across all risk levels, with less than 5% difference between estimated and actual 10-year walking disability rates.

Example from Deardorff et al. (2022)

Figure 2. Baseline Characteristics and Median Predicted Time to Death in Years of 10 Randomly Selected Individuals With Dementia From the Health and Retirement Study Within Each Decile of Predicted Risk



Setting 1 for Predictive Modeling

- Outcome (survival, functional decline, nursing home admission) data on adults age 70+ ($n \approx 1000$, e.g.).
- Have maybe $P = 50$ characteristics potentially predicting outcome
- Goal: build a reasonably parsimonious ($p = 10$ or $p = 15$ predictors), clinically practical and sensible model that has good discrimination and calibration
- Move from statistical model (odds ratios or hazard ratios) to something simple and clinically useful
- Examples: 5 year survival probabilities, median life expectancy, probability of functional decline before death, time spent in nursing home
- (Our group: Health and Retirement Study or NHATS quite often)

Setting 2 for Predictive Modeling

- Same as first typical setting except...
- Number of subjects is much larger ($n \approx 1,000,000$, e.g.)
- Number of potential predictors much larger ($n \approx 1000$, e.g.)
- (Our group: VA EHR data or Medicare claims data)

Setting 3 for Predictive Modeling

- Setting 1/2 plus any of:
  - highly irregular longitudinal data
  - image data
  - text data
- (we are doing a lot more of this in last couple of years but not focus for today)

Statistical models for risk prediction

- Logistic regression (or other binary regression)
- Cox regression (or other time-to-event models)
- Multinomial regression (for nominal outcomes)
- Multi-state models (for longitudinal or survival data with multiple event types)

Some useful papers/books on prediction modeling

- Harrell, Lee, Mark (1996) or Harrell's RMS book (2015)
- Steyerberg et al. (2010) or Steyerberg's CPM book (2019)
- **Moons et al. (2015)**

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

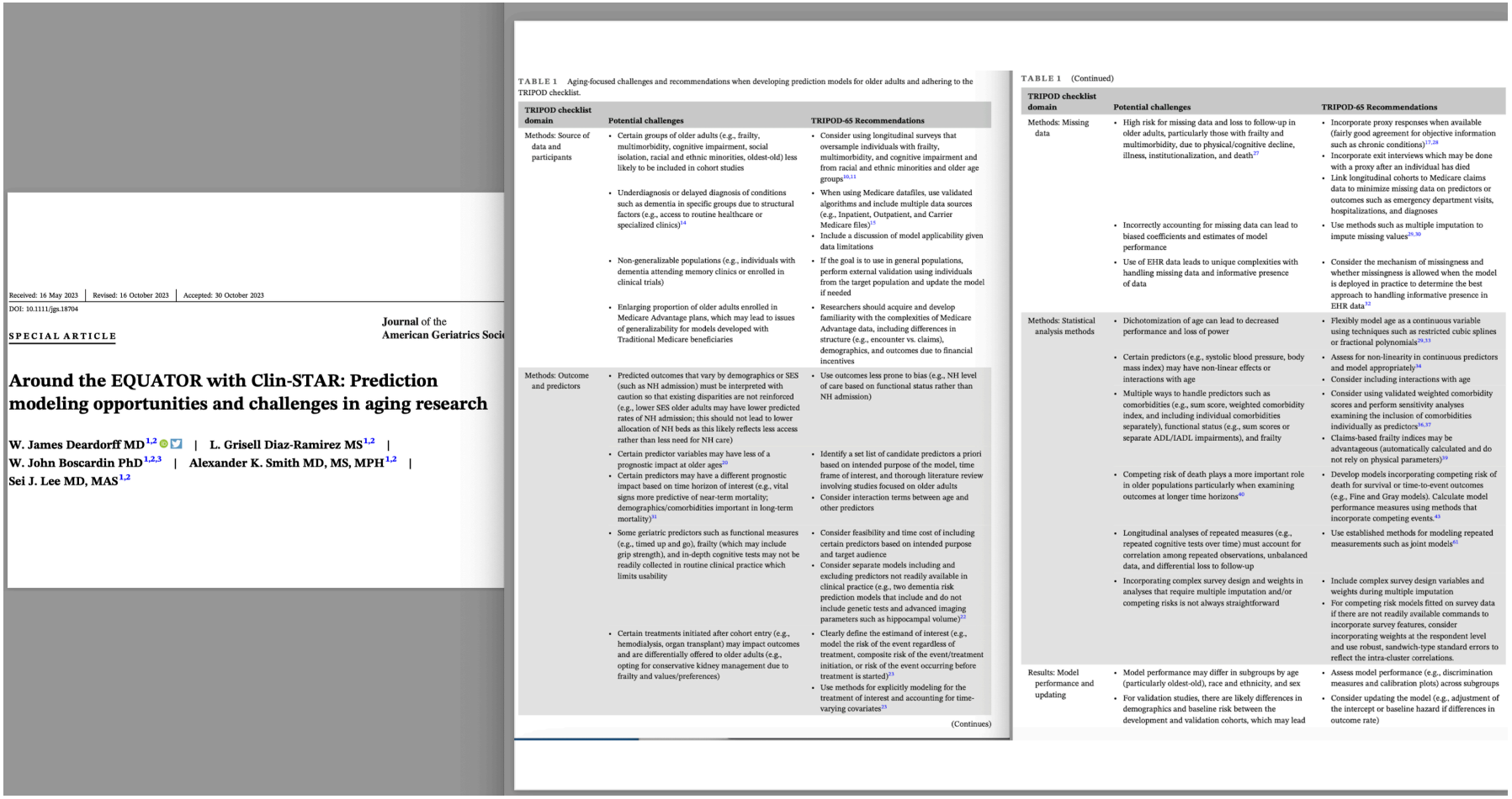
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement includes a 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. This explanation and elaboration document describes the rationale; clarifies the meaning of each item; and discusses why transparent reporting is important, with a view to assessing risk of bias and clinical usefulness of the prediction model. Each checklist item of the TRIPOD Statement is explained in detail and accompanied by published examples of good reporting. The document also provides a valuable reference of issues to consider when designing, conducting, and analyzing prediction model studies. To aid the editorial process and help peer reviewers and, ultimately, readers and systematic reviewers of prediction model studies, it is recommended that authors include a completed checklist in their submission. The TRIPOD checklist can also be downloaded from www.tripod-statement.org.

Ann Intern Med. 2015;162:W1-W73. doi:10.7326/M14-0698 **www.annals.org**
For author affiliations, see end of text.
For members of the TRIPOD Group, see the **Appendix**.

TRIPOD-65: updating TRIPOD for work in older populations (Deardorff et al, 2023)



Received: 16 May 2023 | Revised: 16 October 2023 | Accepted: 30 October 2023
DOI: 10.1111/jgs.18704

Journal of the American Geriatrics Soci

SPECIAL ARTICLE

Around the EQUATOR with Clin-STAR: Prediction modeling opportunities and challenges in aging research

W. James Deardorff MD[1,2] | L. Grisell Diaz-Ramirez MS[1,2] | W. John Boscardin PhD[1,2,3] | Alexander K. Smith MD, MS, MPH[1,2] | Sei J. Lee MD, MAS[1,2]

TABLE 1 Aging-focused challenges and recommendations when developing prediction models for older adults and adhering to the TRIPOD checklist.

TRIPOD checklist domain	Potential challenges	TRIPOD-65 Recommendations
Methods: Source of data and participants	• Certain groups of older adults (e.g., frailty, multimorbidity, cognitive impairment, social isolation, racial and ethnic minorities, oldest-old) less likely to be included in cohort studies	• Consider using longitudinal surveys that oversample individuals with frailty, multimorbidity, and cognitive impairment and from racial and ethnic minorities and older age groups[10,11]
	• Underdiagnosis or delayed diagnosis of conditions such as dementia in specific groups due to structural factors (e.g., access to routine healthcare or specialized clinics)[14]	• When using Medicare datafiles, use validated algorithms and include multiple data sources (e.g., Inpatient, Outpatient, and Carrier Medicare files)[15] • Include a discussion of model applicability given data limitations
	• Non-generalizable populations (e.g., individuals with dementia attending memory clinics or enrolled in clinical trials)	• If the goal is to use in general populations, perform external validation using individuals from the target population and update the model if needed
	• Enlarging proportion of older adults enrolled in Medicare Advantage plans, which may lead to issues of generalizability for models developed with Traditional Medicare beneficiaries	• Researchers should acquire and develop familiarity with the complexities of Medicare Advantage data, including differences in structure (e.g., encounter vs. claims), demographics, and outcomes due to financial incentives
Methods: Outcome and predictors	• Predicted outcomes that vary by demographics or SES (such as NH admission) must be interpreted with caution so that existing disparities are not reinforced (e.g., lower SES older adults may have lower predicted rates of NH admission; this should not lead to lower allocation of NH beds as this likely reflects less access rather than less need for NH care)	• Use outcomes less prone to bias (e.g., NH level of care based on functional status rather than NH admission)
	• Certain predictor variables may have less of a prognostic impact at older ages[10] • Certain predictors may have a different prognostic impact based on time horizon of interest (e.g., vital signs more predictive of near-term mortality; demographics/comorbidities important in long-term mortality)[11]	• Identify a set list of candidate predictors a priori based on intended purpose of the model, time frame of interest, and thorough literature review involving studies focused on older adults • Consider interaction terms between age and other predictors
	• Some geriatric predictors such as functional measures (e.g., timed up and go), frailty (which may include grip strength), and in-depth cognitive tests may not be readily collected in routine clinical practice which limits usability	• Consider feasibility and time cost of including certain predictors based on intended purpose and target audience • Consider separate models including and excluding predictors not readily available in clinical practice (e.g., two dementia risk prediction models that include and do not include genetic tests and advanced imaging parameters such as hippocampal volume)[22]
	• Certain treatments initiated after cohort entry (e.g., hemodialysis, organ transplant) may impact outcomes and are differentially offered to older adults (e.g., opting for conservative kidney management due to frailty and values/preferences)	• Clearly define the estimand of interest (e.g., model the risk of the event regardless of treatment, composite risk of the event/treatment initiation, or risk of the event occurring before treatment is started)[23] • Use methods for explicitly modeling for the treatment of interest and accounting for time-varying covariates[24]

(Continues)

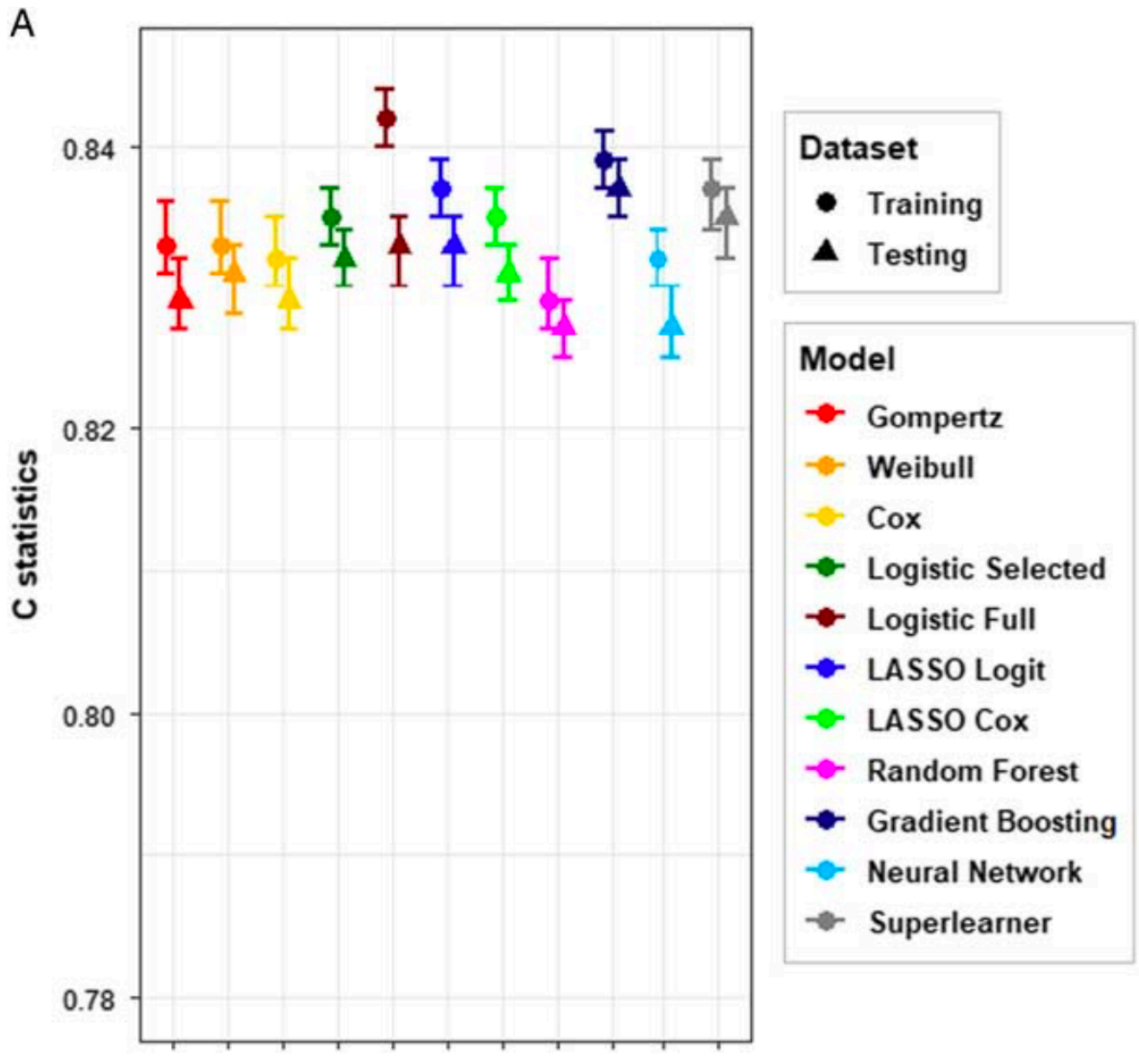
TABLE 1 (Continued)

TRIPOD checklist domain	Potential challenges	TRIPOD-65 Recommendations
Methods: Missing data	• High risk for missing data and loss to follow-up in older adults, particularly those with frailty and multimorbidity, due to physical/cognitive decline, illness, institutionalization, and death[27]	• Incorporate proxy responses when available (fairly good agreement for objective information such as chronic conditions)[17,28] • Incorporate exit interviews which may be done with a proxy after an individual has died • Link longitudinal cohorts to Medicare claims data to minimize missing data on predictors or outcomes such as emergency department visits, hospitalizations, and diagnoses
	• Incorrectly accounting for missing data can lead to biased coefficients and estimates of model performance	• Use methods such as multiple imputation to impute missing values[29,30]
	• Use of EHR data leads to unique complexities with handling missing data and informative presence of data	• Consider the mechanism of missingness and whether missingness is allowed when the model is deployed in practice to determine the best approach to handling informative presence in EHR data[32]
Methods: Statistical analysis methods	• Dichotomization of age can lead to decreased performance and loss of power	• Flexibly model age as a continuous variable using techniques such as restricted cubic splines or fractional polynomials[29,33]
	• Certain predictors (e.g., systolic blood pressure, body mass index) may have non-linear effects or interactions with age	• Assess for non-linearity in continuous predictors and model appropriately[34] • Consider including interactions with age
	• Multiple ways to handle predictors such as comorbidities (e.g., sum score, weighted comorbidity index, and including individual comorbidities separately), functional status (e.g., sum scores or separate ADL/IADL impairments), and frailty	• Consider using validated weighted comorbidity scores and perform sensitivity analyses examining the inclusion of comorbidities individually as predictors[36,37] • Claims-based frailty indices may be advantageous (automatically calculated and do not rely on physical parameters)[39]
	• Competing risk of death plays a more important role in older populations particularly when examining outcomes at longer time horizons[40]	• Develop models incorporating competing risk of death for survival or time-to-event outcomes (e.g., Fine and Gray models). Calculate model performance measures using methods that incorporate competing events.[43]
	• Longitudinal analyses of repeated measures (e.g., repeated cognitive tests over time) must account for correlation among repeated observations, unbalanced data, and differential loss to follow-up	• Use established methods for modeling repeated measurements such as joint models[45]
	• Incorporating complex survey design and weights in analyses that require multiple imputation and/or competing risks is not always straightforward	• Include complex survey design variables and weights during multiple imputation • For competing risk models fitted on survey data if there are not readily available commands to incorporate survey features, consider incorporating weights at the respondent level and use robust, sandwich-type standard errors to reflect the intra-cluster correlations.
Results: Model performance and updating	• Model performance may differ in subgroups by age (particularly oldest-old), race and ethnicity, and sex	• Assess model performance (e.g., discrimination measures and calibration plots) across subgroups
	• For validation studies, there are likely differences in demographics and baseline risk between the development and validation cohorts, which may lead	• Consider updating the model (e.g., adjustment of the intercept or baseline hazard if differences in outcome rate)

Machine Learning vs. Traditional Regression

- Comparison of traditional statistical modeling (TR) and machine learning (ML) in various scenarios (Jing, Boscardin, Deardorff, Jeon, Lee, Donovan, Lee 2022)
- In Settings 1 and 2 (large rectangular data) we and others (e.g. Austin, Harrell, Steyerberg 2022) have found that TR is extremely competitive with ML methods and much easier to begin to understand

Jing et al. 2022



Goals for our (TR) prognostic models

- Predictive accuracy (discrimination and calibration)
- Lack of overfitting
- Parsimony
- Interpretability
- Stability of Individual Predictions

Assessing overfitting in TR modeling

Three main sources of overfitting

- *Feature engineering* (make Table 1 before make model and use it to make decisions)
- *Variable selection* (LASSO or other selection PLUS some other hand-tooling)
- *Parameter estimation* (coefficients are optimized for the data at hand)

Overfitting can occur in all three parts

Internal validation to account for overfitting

- People used to (and still do) use single split-sample for development and validation
- This is uniformly recognized as bad idea (Collins 2024; Steyerberg)
- With a single split sample, you can't separate random variability from systematic overfitting
- Better to use bootstrap internal validation (or cross-validation)

Feature engineering

- Categorizing continuous variables
- Grouping levels of categorical variables
- Choosing to include spline terms for continuous variables
- Deciding whether to look at interactions

Selection with LASSO

- *Always include variables*: in `glmnet` can use the *penalty.factor* option with 0 for variables you do not want to be penalized and 1 for variables that you want LASSO applied to
- *Constraints*: can provide bounds on allowable coefficient estimates on a predictor-by-predictor basis. In textttglmnet can set *lower.limits* to 0 to ensure a variable can only enter as a risk factor
- *Grouping*: can tell grouped LASSO variant that should shrink or kill at a group level. Useful for categorical predictors
- *Other shrinkage targets*: can shrink not towards zero but in other directions (e.g. towards principal components of groups of variables)
- *Less abrubt behavior*: can combine an L1 penalty with an L2 penalty (*elastic net*), Still gives shrink or kill behavior

Constrained LASSO

	A	C	D	E	G	H
	Unconstrained LASSO			**Constrained LASSO (comborbidities OR > 1.0 only)**		
		Odds Ratio			Odds Ratio	
0	Female	0.76		Female	0.76	
3	TreatedInICU	1.30		TreatedInICU	1.33	
9	dialysis	1.04		dialysis	1.13	
0	AplasticAnemia	2.67		AplasticAnemia	2.74	
1	ASTHMA			ASTHMA		
2	CANCER	3.58		CANCER	3.54	
3	CHRONICPAIN	1.15		CHRONICPAIN	1.15	
4	CKD	1.22		CKD	1.21	
5	Coccidioidomycosis	0.93		Coccidioidomycosis		
5	ConnectiveTissueDisorder			ConnectiveTissueDisorder		
7	COPD	1.27		COPD	1.24	
3	CVDArrhythmia	1.36		CVDArrhythmia	1.37	
9	CVDCerebrovascular			CVDCerebrovascular		
0	CVDCHF	1.17		CVDCHF	1.15	
1	CVDHTN			CVDHTN		
2	CVDIHD			CVDIHD		
3	CVDNOS	1.01		CVDNOS	1.06	
4	CVDPVD	1.13		CVDPVD	1.13	
5	CVDThromboembolic			CVDThromboembolic		
5	CVDValvular	1.08		CVDValvular	1.11	
7	Dementia_Parkinson	0.94		Dementia_Parkinson		
3	DIABETES	1.17		DIABETES	1.16	
9	Dyslipidemia			Dyslipidemia		
0	ESLD	3.29		ESLD	3.23	
1	Glaucoma	0.91		Glaucoma		
2	HCV	1.03		HCV	1.06	
3	IMMUNOSUPPRESSED	1.52		IMMUNOSUPPRESSED	1.53	
4	OSTEOPOROSIS	0.93		OSTEOPOROSIS		
5	Ostomies	1.43		Ostomies	1.49	
5	PulmonaryFibrosis	2.45		PulmonaryFibrosis	2.51	
7	Seizures			Seizures		
3	SleepApnea	0.87		SleepApnea		
9	SolitaryPulmonaryNodule			SolitaryPulmonaryNodule		
0	UnspecifiedHepatitis	1.30		UnspecifiedHepatitis	1.35	
1						
2		Both models have same c-statistic to third decimal place and excellent calibration				

Assessing overfitting (Collins et al, 2024)

Box 2: Using bootstrapping for internal validation

The steps to calculate optimism corrected performance using bootstrapping are:

1. Develop the prediction model using the entire original data and calculate the apparent performance.
2. Generate a bootstrap sample (of the same size as the original data), by sampling individuals with replacement from the original data.
3. Develop a bootstrap model using the bootstrap sample (applying all the same modelling and predictor selection methods, as in step 1):
   a. Determine the apparent performance (eg, c statistic, calibration slope) of this model on the bootstrap sample (bootstrap performance).
   b. Determine the performance of the bootstrap model in the original data (test performance).
4. Calculate the optimism as the difference between the bootstrap performance and the test performance.
5. Repeat steps 2 to 4 many times (eg, 500 times).
6. Average the estimates of optimism in step 5.
7. Subtract the average optimism (from step 6) from the apparent performance obtained in step 1 to obtain an optimism corrected estimate of performance.

Bootstrap internal validation

- To fully account for overfitting in internal validation need to replicate the feature engineering, variable selection, coefficient estimation in each bootstrap sample
- So need to algorithmize each component
- Selection and estimation are typically straightforward (e.g. with LASSO)
- Feature engineering might mimic with ad hoc rules, unsupervised thresholding, or supervised knot finding

Large sample setting LASSO (Zhao et al., 2021)

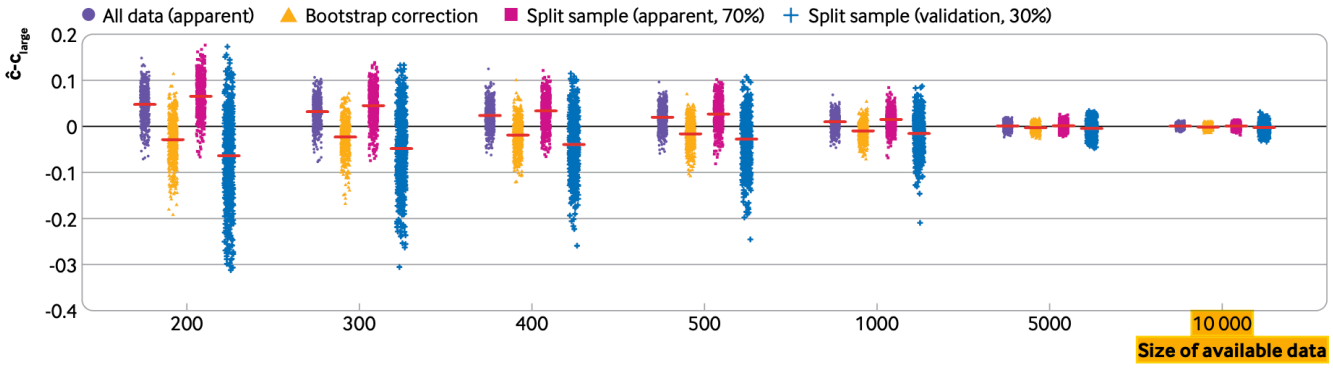
In Defense of the Indefensible: A Very Naïve Approach to High-Dimensional Inference

Sen Zhao, Daniela Witten and Ali Shojaie

Abstract. A great deal of interest has recently focused on conducting inference on the parameters in a high-dimensional linear model. In this paper, we consider a simple and very naïve two-step procedure for this task, in which we (i) fit a lasso model in order to obtain a subset of the variables, and (ii) fit a least squares model on the lasso-selected set. Conventional statistical wisdom tells us that we cannot make use of the standard statistical inference tools for the resulting least squares model (such as confidence intervals and p-values), since we peeked at the data twice: once in running the lasso, and again in fitting the least squares model. However, in this paper, we show that under a certain set of assumptions, with high probability, the set of variables selected by the lasso is identical to the one selected by the noiseless lasso and is hence deterministic. Consequently, the naïve two-step approach can yield asymptotically valid inference. We utilize this finding to develop the *naïve*

- Large data setting, OK to LASSO select then refit for 95%CI

Large sample overfitting? (Collins et al., 2024)



- In large sample setting, TR does not lead to substantial overfitting

Leaderboard vs. Best Model

- Problems inherent in focus on single best model
- "Essentially all models are wrong, but some are useful" (George E. P. Box)
- "Your model is not that special": in our experience, many models are good fit for the data (similar calibration, acceptable discrimination)
- Better to think about the "leaderboard": a large collection of good models some of which may be useful
- Original SAS implementation of best subset makes a "leaderboard"
- Bayesian or Bootstrap selection also get at leaderboard idea

Constrained LASSO (revisited)

	A	C	D	E	G	H
	Unconstrained LASSO			**Constrained LASSO (comborbidities OR > 1.0 only)**		
		Odds Ratio			Odds Ratio	
0	Female	0.76		Female	0.76	
3	TreatedInICU	1.30		TreatedInICU	1.33	
9	dialysis	1.04		dialysis	1.13	
0	AplasticAnemia	2.67		AplasticAnemia	2.74	
1	ASTHMA			ASTHMA		
2	CANCER	3.58		CANCER	3.54	
3	CHRONICPAIN	1.15		CHRONICPAIN	1.15	
4	CKD	1.22		CKD	1.21	
5	Coccidioidomycosis	0.93		Coccidioidomycosis		
5	ConnectiveTissueDisorder			ConnectiveTissueDisorder		
7	COPD	1.27		COPD	1.24	
3	CVDArrhythmia	1.36		CVDArrhythmia	1.37	
9	CVDCerebrovascular			CVDCerebrovascular		
0	CVDCHF	1.17		CVDCHF	1.15	
1	CVDHTN			CVDHTN		
2	CVDIHD			CVDIHD		
3	CVDNOS	1.01		CVDNOS	1.06	
4	CVDPVD	1.13		CVDPVD	1.13	
5	CVDThromboembolic			CVDThromboembolic		
5	CVDValvular	1.08		CVDValvular	1.11	
7	Dementia_Parkinson	0.94		Dementia_Parkinson		
3	DIABETES	1.17		DIABETES	1.16	
9	Dyslipidemia			Dyslipidemia		
0	ESLD	3.29		ESLD	3.23	
1	Glaucoma	0.91		Glaucoma		
2	HCV	1.03		HCV	1.06	
3	IMMUNOSUPPRESSED	1.52		IMMUNOSUPPRESSED	1.53	
4	OSTEOPOROSIS	0.93		OSTEOPOROSIS		
5	Ostomies	1.43		Ostomies	1.49	
5	PulmonaryFibrosis	2.45		PulmonaryFibrosis	2.51	
7	Seizures			Seizures		
3	SleepApnea	0.87		SleepApnea		
9	SolitaryPulmonaryNodule			SolitaryPulmonaryNodule		
0	UnspecifiedHepatitis	1.30		UnspecifiedHepatitis	1.35	
1						
2		Both models have same c-statistic to third decimal place and excellent calibration				

SAS best subsets leaderboard (Miao et al. 2013)

Number of Variables in Original Model	Variables in Original Model	Number of Variables in Complete Model	Variables in Complete Model	AIC with Covariates in Complete Model	SC with Covariates in Complete Model	Harrell's *c* Statistic
12	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM	15	RACEETH1 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	***548.5637*** ***[Best AIC Model]***	626.9754	0.848265
13	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM	16	RACEETH1 MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	549.5971	632.9095	0.847923
14	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM	16	RACEETH1 MALE SMOKE EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	549.6694	632.9818	0.847757
11	AGECAT3 AGECAT5 AGECAT6 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM	14	MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	549.9432	623.4542	0.843986
15	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM	17	RACEETH1 MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	550.4291	638.6423	***0.850518*** ***[Best Harrell's c]***
11	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM	14	RACEETH1 MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	550.6212	624.1774	0.846005
15	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM	17	RACEETH1 EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	550.8061	639.0011	0.847096
15	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM	17	RACEETH1 MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	550.8468	639.0599	0.848321
16	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM	18	RACEETH1 EDUCATION MALE SMOKE DRESS EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	551.5204	644.6152	0.849145
16	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM	18	RACEETH1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	551.6417	644.7556	0.849985
16	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM	18	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	551.6501	644.7448	0.849034
12	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT HYPERTEN DIABETES CANCER LUNG WALKROOM	15	RACEETH1 MALE SMOKE EAT HYPERTEN DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	551.6578	630.1178	0.847598
12	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM	15	RACEETH1 MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	551.7328	630.1927	0.846240
10	AGECAT3 AGECAT5 AGECAT6 MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM	13	MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	551.8301	620.4825	0.841289

Kaggle competition (100's of entries with $c > 0.900$)

Binary Prediction with a Rainfall Dataset

Late Submission ···

Overview Data Code Models Discussion Leaderboard Rules

#	△	Team	Members	Score	Entries	Last	Solution
1	▲ 812	Guillaume HIMBERT		0.90654	56	1mo	
2	▼ 1	Chris Deotte		0.90604	17	1mo	
3	▲ 2251	AndNov		0.90583	2	1mo	
4	▲ 815	Daniel Halwell		0.90575	9	1mo	
5	▲ 1555	MonicaWatashi		0.90534	15	1mo	
6	▲ 2415	kgmuzu		0.90534	13	1mo	
7	▲ 753	anonymous		0.90526	1	2mo	
8	▲ 302	Arko Bera		0.90518	47	2mo	
9	▲ 2888	sam114119		0.90491	2	1mo	
10	▲ 2431	Ranapratap Deshmukh		0.90489	5	1mo	
11	▲ 1595	Howard Liao		0.90464	8	2mo	
12	▲ 1542	czyi28		0.90421	35	2mo	
13	▲ 2238	Saurav Das		0.90410	6	2mo	
14	▲ 3037	Pruthvinath Jeripity Venkata		0.90406	1	1mo	

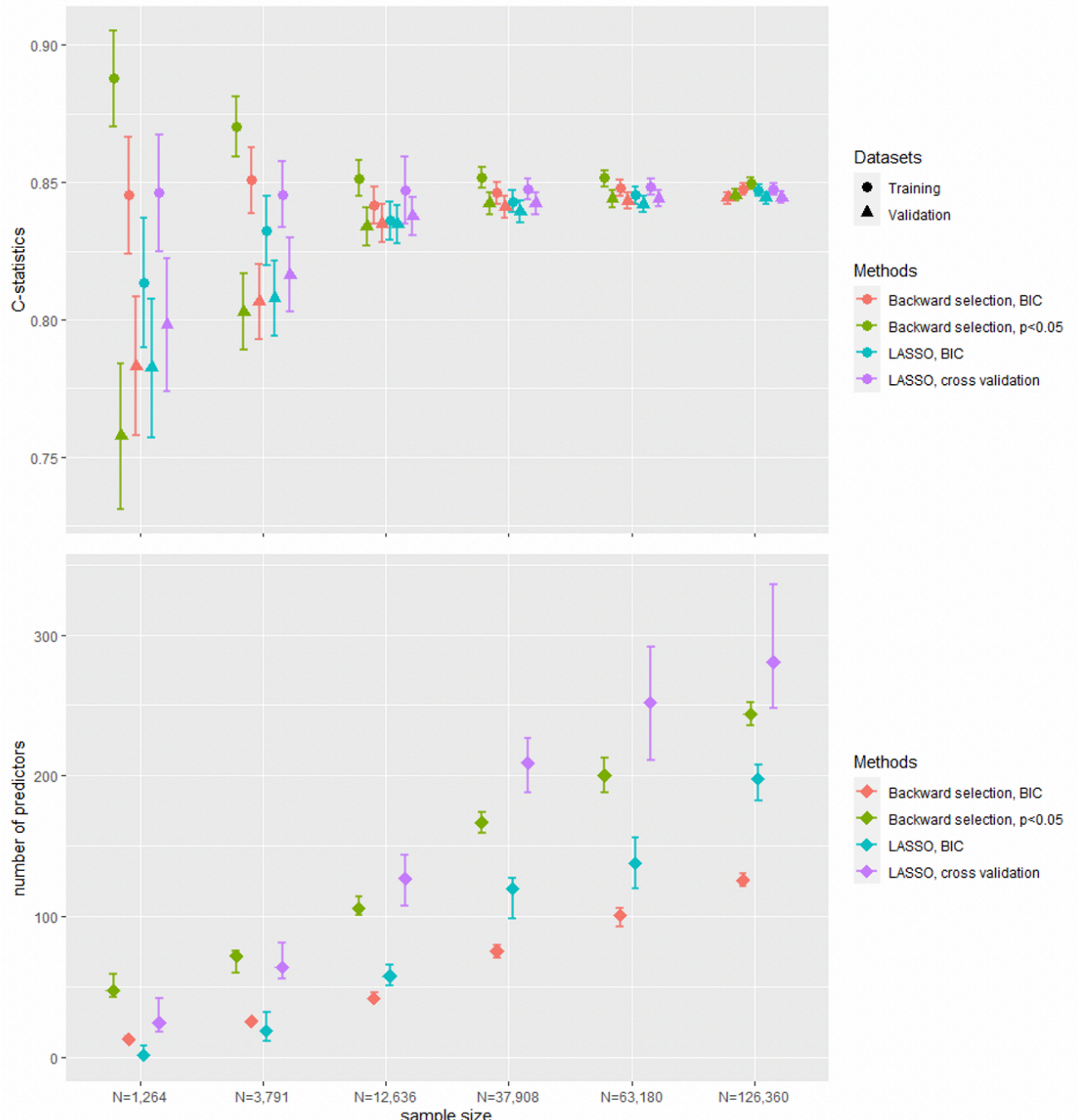
Taking stock

- Can LASSO methods help to accomplish all the goals?
- Predictive accuracy (extremely competitive)
- Minimal overfitting (LASSO is good at this in settings 1 and 2)
- Interpretability (regression method *and* can also require only positive coefficients for subset of terms *and* can force in some terms)
- Parsimony? (does selection but maybe not enough as discussed next)
- But what about stability? (more on this in moment)

Parsimony and LASSO

- LASSO vs. stepwise vs. best subset in practice (Jeon, Lee, Ding, Jing, Deardorff, Boscardin, under review, 2025)
- LASSO picks a much less parsimonious model that does not perform any better in Setting 2
- Similar ideas noted in Hastie, Tibshirani, and Tibshirani (2020)

Jeon et al. 2025



Possible solutions

- Can think about using L0 regression and variants
- Broken Adaptive Ridge (BAR) package is very promising

BAR paper

Journal of Statistical Planning and Inference 213 (2021) 262–281

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A scalable surrogate L_0 sparse regression method for generalized linear models with applications to large scale data

Ning Li [a], Xiaoling Peng [b], Eric Kawaguchi [c], Marc A. Suchard [d], Gang Li [e,*]

[a] Department of Medicine Statistics Core, University of California - Los Angeles, CA, USA
[b] Division of Science and Technology, Beijing Normal University - Hong Kong Baptist University United International College, Zhuhai, China
[c] Division of Biostatistics and Epidemiology, Keck School of Medicine, University of Southern California, CA, USA
[d] Departments of Computational Medicine, Biostatistics and Human Genetics, University of California - Los Angeles, CA, USA
[e] Department of Biostatistics, University of California - Los Angeles, CA, USA

ARTICLE INFO

Article history:
Received 6 February 2020
Received in revised form 1 December 2020
Accepted 4 December 2020
Available online 17 December 2020

Keywords:
Generalized linear models
High dimensional massive sample size data
L_0 penalty
Ridge regression
Variable selection

ABSTRACT

This paper rigorously studies large sample properties of a surrogate L_0 penalization method via iteratively performing reweighted L_2 penalized regressions for generalized linear models and develop a scalable implementation of the method for **s**parse **h**igh **d**imensional **m**assive **s**ample **s**ize (sHDMSS) data. We show that for generalized linear models, the limit of the algorithm, referred to as the broken adaptive ridge (BAR) estimator, is consistent for variable selection, enjoys an oracle property for parameter estimation, and possesses a grouping property for highly correlated covariates. We further demonstrate that by taking advantage of an existing efficient implementation of massive L_2-penalized generalized linear models, the proposed BAR method can be conveniently implemented for sHDMSS data. An illustration is given using a large sHDMSS data from the Truven MarketScan Medicare (MDCR) database to investigate the safety of dabigatran versus warfarin for treatment of nonvalvular atrial filbrillation in elder patients.

© 2020 Elsevier B.V. All rights reserved.

BAR result

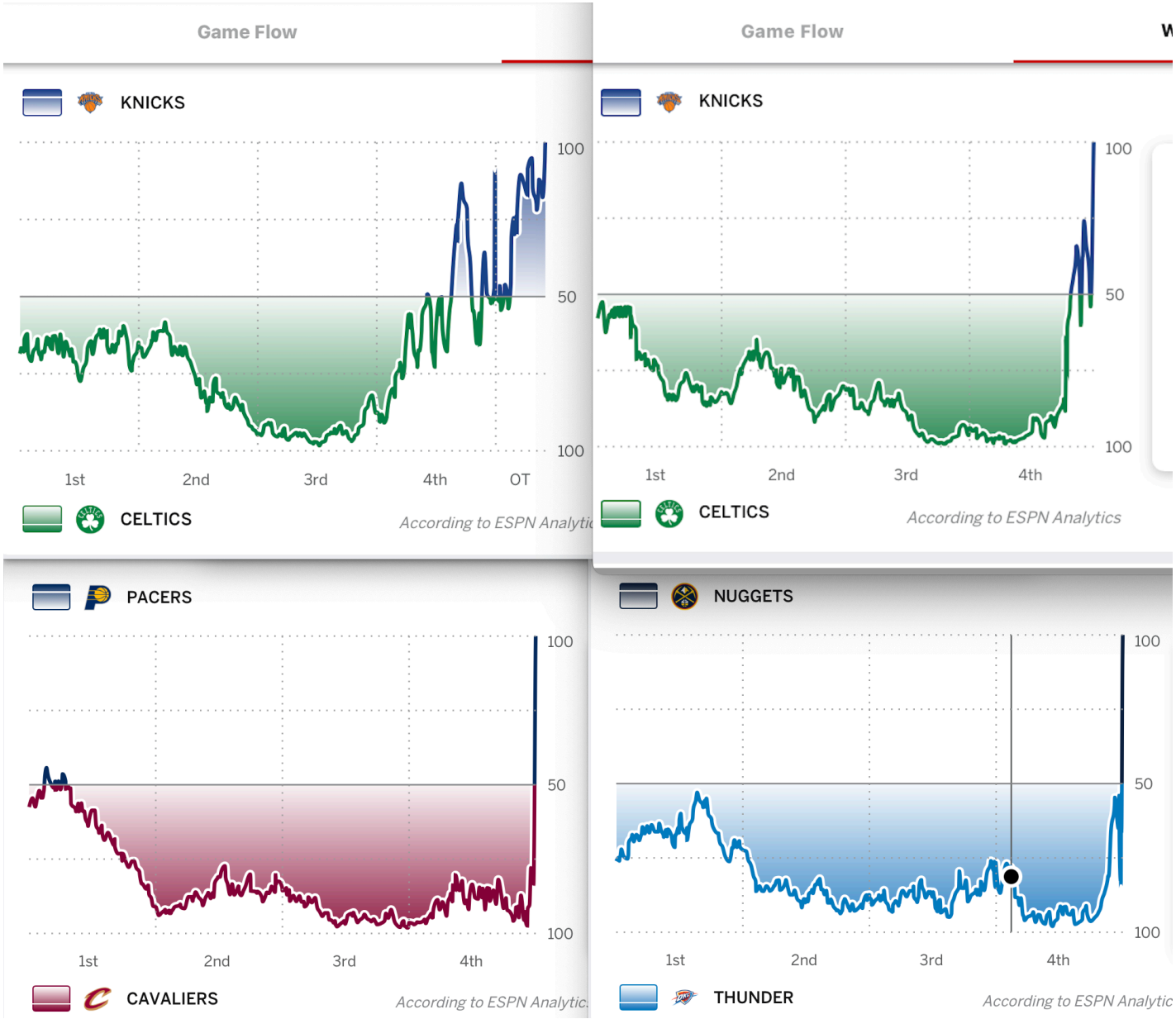
TABLE 3 (Pediatric National Trauma Data Bank (NTDB) data) Comparison of mCox-LASSO and massive Cox's regression for broken adaptive ridge (mBAR) regression for the pediatric NTDB data. (mCox-LASSO cross-validation (CV) and mCox-LASSO Bayesian information criterion (BIC) correspond to mCox-LASSO using cross validation and BIC selection criterion, respectively. mBAR (BIC) denotes mBAR using the BIC selection criterion while fixing $\xi_n = \log(p_n)$. The training set has a sample size of 168 000, while the test set used for the *c*-index has a sample size of 45 555)

Method	# Selected	BIC score	*c*-index	Runtime (hours)
mBAR ($\lambda_n = 0.5\log(p_n)$)	45	51 613.52	0.91	8
mBAR ($\lambda_n = \log(p_n)$)	21	52 182.90	0.89	8
mBAR (BIC)	83	51 269.43	0.93	97
mCox-LASSO (BIC)	100	52 544.90	0.91	25
mCox-LASSO (CV)	253	53 165.44	0.92	41

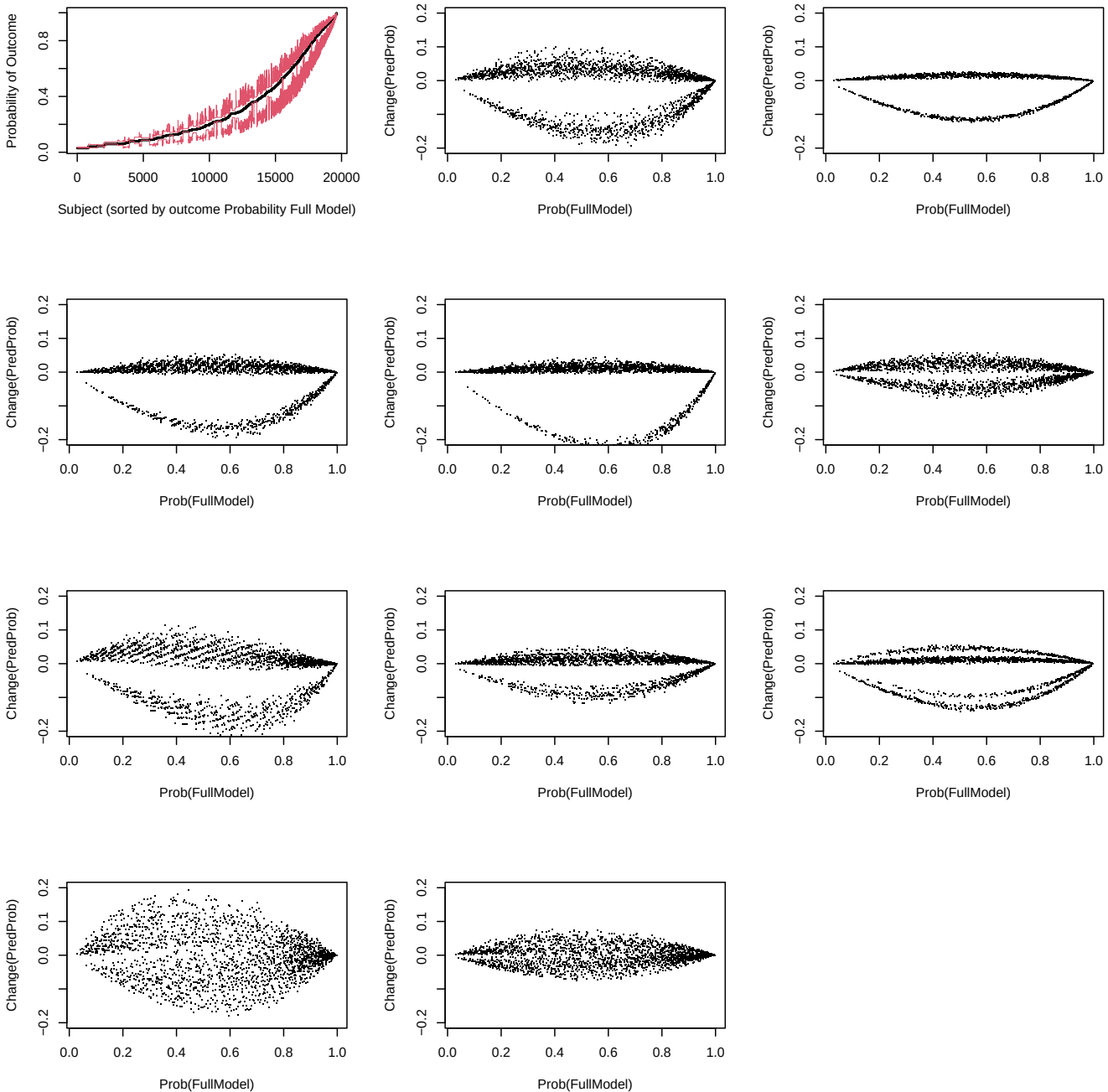
Stability of predictions

- Many models can have similar discrimination and excellent calibration
- But for any given pair of models, an individual might have very different predictions
- This is very undesirable for our use cases
- Idea has been called predictive multiplicity in ML literature (e.g. Watson-Daniels et al., 2023)
- Binary version of this idea leads to reclassification metrics (NRI, IDI)
- Variant of this issue looks at instability in individual predictions using same model but in replicate data (Riley et al., 2023)

Instability of in-game win probabilities



Predictive multiplicity (leave out 1 predictor and refit)



Tradeoff of parsimony and stability

- Parsimonious models are most subject to predictive multiplicity
- This suggests using modestly parsimonious constrained LASSO or even elastic net could help avoid instability of individual predictions
- A non-parsimonious penalized regression model may be a good "reference" model to check instability against

Summary

- Many competing goals in building prognostic models with use case involving individual predicted probabilities
- Leverage the top-heavy leaderboard
- Constrained LASSO models seem to satisfy many of the goals
- For larger sample sizes, not particularly parsimonious (L0 methods may be preferred)
- Parsimonious models may suffer from predictive multiplicity (i.e. individual predictions from these models may differ qualitatively from another model with equally good overall fit)
  Thanks to audience for listening and to SDSA for the invitation!

References

- Aliberti MJR, Kotwal AA, Smith AK, Lee SJ, Banda S, Boscardin WJ (2021). Pre-estimating subsets: A new approach for unavailable predictors in prognostic modeling. *J Am Geriatr Soc*.
- Austin PC, Harrell FE, Steyerberg EW (2021). Predictive performance of machine and statistical learning methods in the large N, small p setting. *Stat Meth Med Res*.
- Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B et al. (2024. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*.
- Diaz-Ramirez LG, Lee SJ, Smith AK, Gan S, Boscardin WJ (2021). A Novel Method for Identifying a Parsimonious and Accurate Predictive Model for Multiple Clinical Outcomes. *Comput Methods Programs Biomed*.
- Harrell FE, Lee KL, Mark DB (1996). Tutorial in Biostatistics: Multivariable prognostic models. *Stat Med*.
- Harrell FE (2015). Regression Modeling Strategies (second edition).
- Hastie T, Tibshirani R, Tibshirani R (2020). Best subset, forward stepwise or lasso? *Stat Sci*.

References (continued)

- Jing B, Boscardin WJ, Deardorff WJ, Jeon SY, Lee AK, Donovan AL, Lee SJ (2022). Comparing Machine Learning to Regression Methods for Mortality Prediction Using VA EHR Clinical Data. *Med Care*.
- Jeon S, Lee SJ, Jing B, Lee AK, Deardorff WJ, Boscardin WJ (2025, under review). Comparing the Accuracy and Parsimony of LASSO and Backward Selection Prediction Models Across Varying Sample Sizes.
- Lee SJ, Smith AK, Diaz-Ramirez LG, Covinsky KE, Gan S, Chen CL, Boscardin WJ (2021). A Novel Metric for Developing Easy-to-Use and Accurate Clinical Prediction Models: The Time-cost Information Criterion. *Med Care*.
- Lee AK, Diaz-Ramirez LG, Boscardin WJ, Smith AK, Lee SJ (2022). A comprehensive prognostic tool for older adults: Predicting death, ADL disability, and walking disability simultaneously. *J Am Geriatr Soc*.
- Li N, Peng X, Kawaguchi E, Suchard MA, Li G (2021). A scalable surrogate L0 sparse regression method for generalized linear models with applications to large scale data. *J Stat Plan Inf*.
- Miao Y, Cenzer IS, Kirby KA, Boscardin WJ (2013). Estimating Harrell's Optimism on Predictive Indices using Bootstrap Samples. *SAS Global Forum Proceedings*.

References (continued)

- Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Int Med*.
- Riley RD, Ensor J, Snell KI, Harrell FE, Martin GP, Reitsma JB, Moons KG, Collins G, Van Smeden M (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*.
- Riley RD, Pate A, Dhiman P, Archer L, Martin GP, Collins GS (2023). Clinical prediction models and the multiverse of madness. *BMC medicine*.
- Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidem*.
- Steyerberg EW (2019). Clinical prediction models (second edition)
- Sullivan LM, Massaro JM, D'Agostino RB (2004), Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med*, **23**.
- Watson-Daniels J, Parkes DC, Ustun B (2023). Predictive Multiplicity in Probabilistic Classification. *Proceedings of AAAI*
- Zhao S, Witten D, Shojaie A (2021). In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference. *Stat Sci*.