Accuracy measures for prognostic survival models: tutorial

C. Jason Liang

National Institute of Allergy and Infectious Diseases (NIAID)
Biostatistics Research Branch (BRB)

2022, June 7th

Housekeeping

Can everyone access the following?

https://github.com/liangcj/KI-tutorial-2022

We will be using `actt.RData`

A note on data usage

We will be using recently collected human subjects data from a Covid-19 treatment trial.

Please be respectful of patient privacy.

Please delete the data after usage, and do not distribute beyond this class.

A note on data usage

3. Approved User and Accessing Institution agree to:

(a) Retain control of and agree not to distribute to any entity or individual not listed in the DAR, controlled-access clinical trial dataset(s) or any data derivatives obtained through the approval of the DAR. The approved DUA is not transferrable to another user or institution. For avoidance of doubt, controlled-access clinical trial dataset(s) or any data derivatives may not be distributed to any entity or individual not listed in the DAR even if that entity or individual is under the direct supervision of Approved User.

(b) Keep Data secure and confidential at all times and adhere to all data security practices, safeguard Data and protect participants' privacy, and adhere to the terms of use defined in this Agreement and Accessing Institution's IT security requirements and policies to prevent unauthorized use of or access to data including completing appropriate training related to data security and privacy.

(h) Not use the approved datasets, either alone or in concert with any other information or datasets, for any of the following purposes.

i. Identifying or contacting individual participants or their living relatives from whom Data was collected unless required by law to maintain public health and safety, or generating information (e.g., facial images or comparable representations) that could allow the identities of research participants to be readily ascertained.

ii. Any diagnostic, prognostic, or treatment purpose. Although an approved dataset cannot be used for a clinical diagnosis or treatment of an individual patient, this does not prohibit use of an approved dataset to support the development of diagnostic, prognostic, or therapeutic purposes.

iii. Any commercial purpose, including selling, commercial screening, or transferring Human Data to a third party for commercial purposes. "Commercial purpose" does not include using an approved dataset to develop commercial products or services as diagnostics or therapeutics interventions.

What does a (bio)statistician do? Define “biostatistician”.

What does a (bio)statistician do?

“Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.”

-Wikipedia[1][2][3]

1. "Statistics". Oxford Reference. Oxford University Press. January 2008
2. Romijn, Jan-Willem (2014). "Philosophy of statistics". Stanford Encyclopedia of Philosophy.
3. Cambridge Dictionary.

What does a (bio)statistician do?

“Guardian of scientific rigor”

-Scott Emerson, paraphrased

What does a (bio)statistician do?

“Extract knowledge from data”

-Patrick Heagerty, paraphrased

What does a (bio)statistician do?

“Justify sample sizes two days before the grant is due”

-Anonymous

https://www.youtube.com/watch?v=PbODigCZqL8

What does a (bio)statistician do?

“Whatever a data scientist does, with less pay”

-Twitter, paraphrased

What does a (bio)statistician do?

“Just as a physician treats a patient’s disease, a biostatistician treats a physician’s research study.”

-Me, talking to neighbors

Clarifying some terms

Diagnostic modeling: do I have the disease?

- e.g. Rapid antigen test

Prognostic modeling: will I get the disease?

- e.g. Framingham risk score, age, comorbidities

Predictive modeling: will this **treatment** improve/prevent disease?

- e.g. Oncotype DX for breast cancer

Ballman, K. V. (2015). Biomarker: predictive or prognostic?. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, *33*(33), 3968-3971.

Taxonomy of prognostic accuracy measures (2010)

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2 Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects.
Discrimination	C statistic	ROC curve	Rank order statistic; Interpretation for a pair of patients with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; Easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean(y) versus mean($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation related to 'shrinkage' of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification calibration		Compare observed and predicted within cross-classified categories
	Net Reclassification Index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right correction
	Integrated Discrimination Index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net Benefit (NB)	Cross-table	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)
	Decision curve analysis (DCA)	Decision curve	

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, *21*(1), 128.

Outline

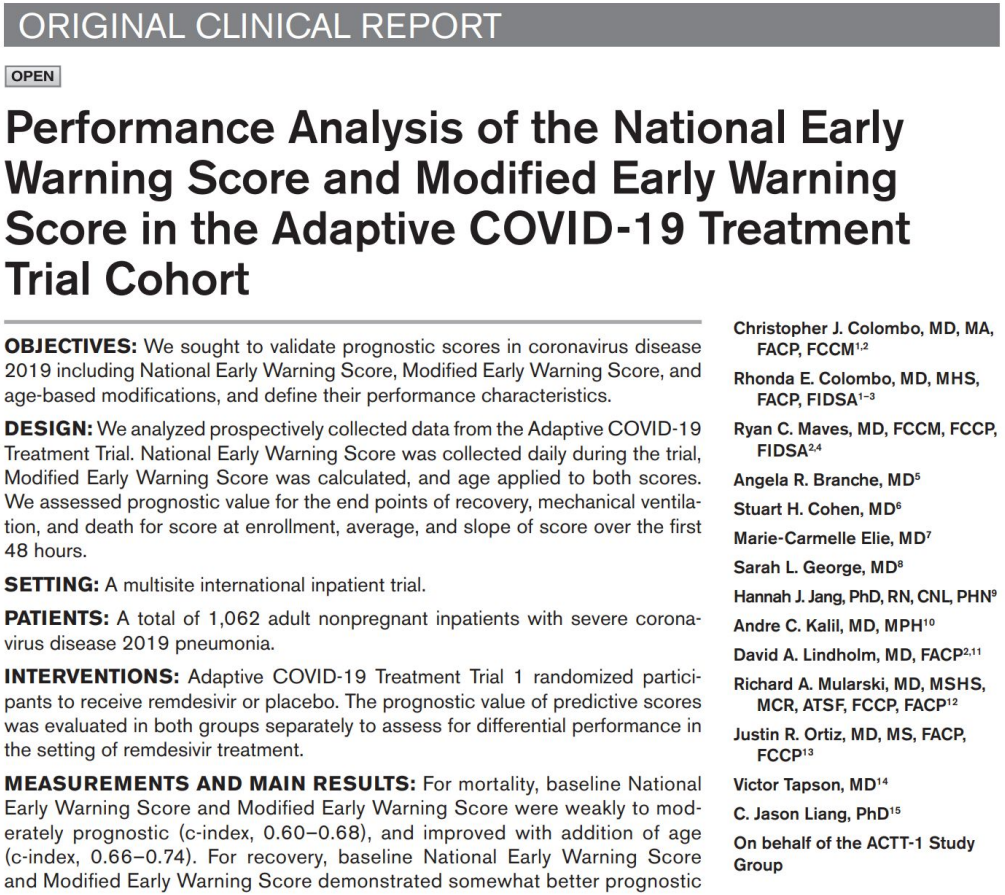
Tell story of the Critical Care Explorations (CCX) paper (2021)

- “Behind the scenes” look at why we made certain decisions and what I might have done differently

Dive into specific accuracy measures for survival models

- Interpretation and estimation
- Reproduce parts of the CCX paper

More general prognostic modeling practices



ORIGINAL CLINICAL REPORT

OPEN

Performance Analysis of the National Early Warning Score and Modified Early Warning Score in the Adaptive COVID-19 Treatment Trial Cohort

OBJECTIVES: We sought to validate prognostic scores in coronavirus disease 2019 including National Early Warning Score, Modified Early Warning Score, and age-based modifications, and define their performance characteristics.

DESIGN: We analyzed prospectively collected data from the Adaptive COVID-19 Treatment Trial. National Early Warning Score was collected daily during the trial, Modified Early Warning Score was calculated, and age applied to both scores. We assessed prognostic value for the end points of recovery, mechanical ventilation, and death for score at enrollment, average, and slope of score over the first 48 hours.

SETTING: A multisite international inpatient trial.

PATIENTS: A total of 1,062 adult nonpregnant inpatients with severe coronavirus disease 2019 pneumonia.

INTERVENTIONS: Adaptive COVID-19 Treatment Trial 1 randomized participants to receive remdesivir or placebo. The prognostic value of predictive scores was evaluated in both groups separately to assess for differential performance in the setting of remdesivir treatment.

MEASUREMENTS AND MAIN RESULTS: For mortality, baseline National Early Warning Score and Modified Early Warning Score were weakly to moderately prognostic (c-index, 0.60–0.68), and improved with addition of age (c-index, 0.66–0.74). For recovery, baseline National Early Warning Score and Modified Early Warning Score demonstrated somewhat better prognostic

Christopher J. Colombo, MD, MA, FACP, FCCM[1,2]

Rhonda E. Colombo, MD, MHS, FACP, FIDSA[1–3]

Ryan C. Maves, MD, FCCM, FCCP, FIDSA[2,4]

Angela R. Branche, MD[5]

Stuart H. Cohen, MD[6]

Marie-Carmelle Elie, MD[7]

Sarah L. George, MD[8]

Hannah J. Jang, PhD, RN, CNL, PHN[9]

Andre C. Kalil, MD, MPH[10]

David A. Lindholm, MD, FACP[2,11]

Richard A. Mularski, MD, MSHS, MCR, ATSF, FCCP, FACP[12]

Justin R. Ortiz, MD, MS, FACP, FCCP[13]

Victor Tapson, MD[14]

C. Jason Liang, PhD[15]

On behalf of the ACTT-1 Study Group

Colombo, C. J., Colombo, R. E., Maves, R. C., Branche, A. R., Cohen, S. H., Elie, M. C., ... & Liang, C. J. (2021). Performance Analysis of the National Early Warning Score and Modified Early Warning Score in the Adaptive COVID-19 Treatment Trial Cohort. Critical Care Explorations, 3(7).

Goals

Understand measures of discrimination for survival models

- Interpret and estimate the C-index
- Interpret and estimate Cumulative/Dynamic ROC curves and AUC
- Interpret and estimate Incident/Dynamic ROC curves and AUC

Awareness of the importance of calibration

Exposure to general practices and uses of prognostic modeling

List of resources and references

A biostatistician's case report

Jan 2020: Covid-19 enters public consciousness

Feb 2020: Adaptive Covid-19 Treatment Trial (ACTT-1) starts enrollment. Double blind RCT of Remdesivir (antiviral) vs. placebo.

May 2020: US FDA issues emergency use authorization for Remdesivir

May 2020: Wynants et al., via "living systematic review" in BMJ, reviewed 79 published models

August 2020: We kick-off secondary analysis of ACTT-1 data

May 2021: Manuscript accepted in Critical Care Explorations (CCX)

https://www.cdc.gov/museum/timeline/covid19.html

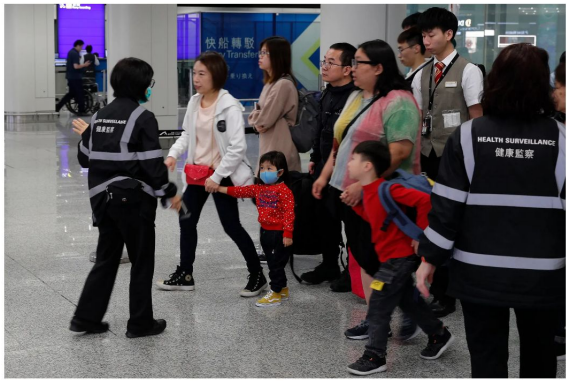
Jan 2020: Covid-19 enters public consciousness

The New York Times

China Grapples With Mystery Pneumonia-Like Illness

Beijing is racing to identify a new illness that has sickened 59 people as it tries to calm a nervous public.

Give this article 27



Health surveillance officers checked temperatures of passengers upon arrival at Hong Kong, on Saturday. Andy Wong/Associated Press

By Sui-Lee Wee and Vivian Wang

Published Jan. 6, 2020 Updated Jan. 21, 2020

Feb 2020: NIAID remdesivir trial begins (ACTT-1)

Adaptive Covid-19 Treatment Trial (ACTT-1) starts enrollment.

Double blind RCT of remdesivir (antiviral) vs. placebo.

NEWS RELEASES

Tuesday, February 25, 2020

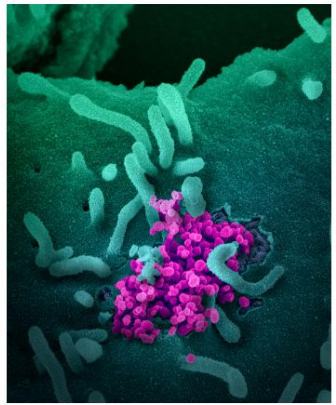
NIH clinical trial of remdesivir to treat COVID-19 begins

Study enrolling hospitalized adults with COVID-19 in Nebraska.

A randomized, controlled clinical trial to evaluate the safety and efficacy of the investigational antiviral remdesivir in hospitalized adults diagnosed with coronavirus disease 2019 (COVID-19) has begun at the University of Nebraska Medical Center (UNMC) in Omaha. The trial regulatory sponsor is the National Institute of Allergy and Infectious Diseases (NIAID), part of the National Institutes of Health. This is the first clinical trial in the United States to evaluate an experimental treatment for COVID-19, the respiratory disease first detected in December 2019 in Wuhan, Hubei Province, China.

The first trial participant is an American who was repatriated after being quarantined on the Diamond Princess cruise ship that docked in Yokohama, Japan and volunteered to participate in the study. The study can be adapted to evaluate additional investigative treatments and to enroll participants at other sites in the U.S. and worldwide.

There are no specific therapeutics approved by the Food and Drug Administration (FDA) to treat people with COVID-19, the disease caused by the newly emergent SARS-CoV-2 virus (formerly known as 2019-nCoV). Infection can cause mild to severe respiratory illness, and symptoms can include fever, cough and shortness of breath. As of February 24, the World Health Organization (WHO) has reported 77,262 confirmed cases of COVID-19 and 2,595 deaths in China and 2,069 cases of COVID-19 and 23 deaths in 29 other countries. There have been 14 confirmed COVID-19 cases reported in the United States and an additional 39 cases among persons repatriated to the United States, according to the Centers for Disease Control and Prevention (CDC).



Novel Coronavirus SARS-CoV-2 This scanning electron microscope image shows SARS-CoV-2 (round magenta objects) emerging from the surface of cells cultured in the lab. SARS-CoV-2, also known as 2019-nCoV, is the virus that causes COVID-19. The virus shown was isolated from a patient in the U.S. *NIAID-RML*

May 2020: US FDA authorizes remdesivir (EUA)

Modest improvement in recovery for hospitalized adults

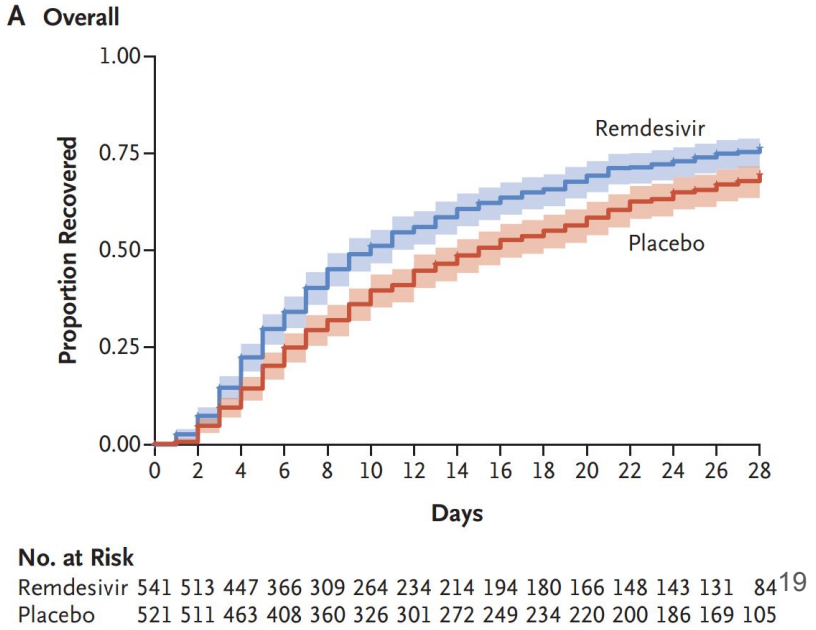
Suggestion of greater efficacy in those with moderate disease (as opposed to mild or severe)

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 NOVEMBER 5, 2020 VOL. 383 NO. 19

Remdesivir for the Treatment of Covid-19 — Final Report

J.H. Beigel, K.M. Tomashek, L.E. Dodd, A.K. Mehta, B.S. Zingman, A.C. Kalil, E. Hohmann, H.Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R.W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T.F. Patterson, R. Paredes, D.A. Sweeney, W.R. Short, G. Touloumi, D.C. Lye, N. Ohmagari, M. Oh, G.M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M.G. Kortepeter, R.L. Atmar, C.B. Creech, J. Lundgren, A.G. Babiker, S. Pett, J.D. Neaton, T.H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, and H.C. Lane, for the ACTT-1 Study Group Members*



Side note: more ACTT trials

ACTT-1: Placebo vs remdesivir

- Remdesivir superior

ACTT-2: Remdesivir vs remdesivir + baricitinib (JAK inhibitor)

- Remdesivir + baricitinib superior

ACTT-3: Remdesivir vs remdesivir + Interferon beta-1a

- Remdesirivr + Interferon beta-1a not superior

ACTT-4: Remdesivir + baricitinib vs remdesivir + dexamethasone (steroid)

- Stopped early for futility

Many other treatment trials outside of ACTT platform!

Aug 2020: Kick-off secondary analyses of ACTT-1 data

1) Mechanistic analysis: remdesivir works - how?
   a) Multistate modeling
2) Predictive biomarker: is there a subgroup that responds particularly well to treatment?
   a) Interaction between treatment and biomarker?
3) Prognostic biomarkers: validate existing risk scores
   a) ROC curves, AUC, C-index for survival data

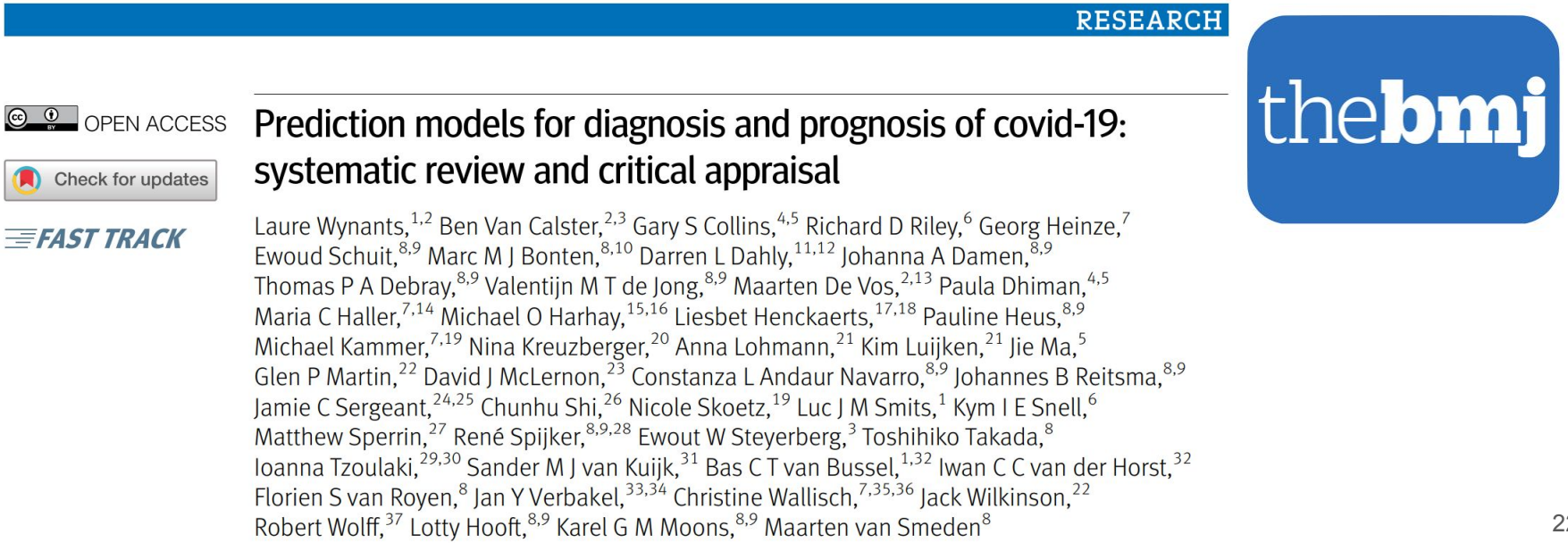
Fintzi, J., Bonnett, T., Sweeney, D. A., Huprikar, N. A., Ganesan, A., Frank, M. G., ... & Mehta, A. K. (2021). Deconstructing the Treatment Effect of Remdesivir in the Adaptive COVID-19 Treatment Trial-1: Implications for Critical Care Resource Utilization. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America.

Paules, C. I., Gallagher, S. K., Rapaka, R. R., Davey, R. T., Doernberg, S. B., Grossberg, R., ... & Benson, C. A. (2022). Remdesivir for the Prevention of Invasive Mechanical Ventilation or Death in Coronavirus Disease 2019 (COVID-19): A Post Hoc Analysis of the Adaptive COVID-19 Treatment Trial-1 Cohort Data. Clinical Infectious Diseases, 74(7), 1260-1264.

Colombo, C. J., Colombo, R. E., Maves, R. C., Branche, A. R., Cohen, S. H., Elie, M. C., ... & Liang, C. J. (2021). Performance Analysis of the National Early Warning Score and Modified Early Warning Score in the Adaptive COVID-19 Treatment Trial Cohort. Critical Care Explorations, 3(7).

A flood of Covid-19 prognostic models

By May 2020, at least 79 papers on Covid-19 prognostic models had been published.



RESEARCH

OPEN ACCESS

Check for updates

FAST TRACK

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Marc M J Bonten,[8,10] Darren L Dahly,[11,12] Johanna A Damen,[8,9] Thomas P A Debray,[8,9] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,13] Paula Dhiman,[4,5] Maria C Haller,[7,14] Michael O Harhay,[15,16] Liesbet Henckaerts,[17,18] Pauline Heus,[8,9] Michael Kammer,[7,19] Nina Kreuzberger,[20] Anna Lohmann,[21] Kim Luijken,[21] Jie Ma,[5] Glen P Martin,[22] David J McLernon,[23] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[24,25] Chunhu Shi,[26] Nicole Skoetz,[19] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[27] René Spijker,[8,9,28] Ewout W Steyerberg,[3] Toshihiko Takada,[8] Ioanna Tzoulaki,[29,30] Sander M J van Kuijk,[31] Bas C T van Bussel,[1,32] Iwan C C van der Horst,[32] Florien S van Royen,[8] Jan Y Verbakel,[33,34] Christine Wallisch,[7,35,36] Jack Wilkinson,[22] Robert Wolff,[37] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

thebmj

Covid-19 prognostic models (Jan 2021)

RESULTS

37 421 titles were screened, and 169 studies describing 232 prediction models were included. The review identified seven models for identifying people at risk in the general population; 118 diagnostic models for detecting covid-19 (75 were based on medical imaging, 10 to diagnose disease severity); and 107 prognostic models for predicting mortality risk, progression to severe disease, intensive care unit admission, ventilation, intubation, or length of hospital stay. The most frequent types of predictors included in the covid-19 prediction models are vital signs, age, comorbidities, and image features.

CONCLUSION

Prediction models for covid-19 are quickly entering the academic literature to support medical decision making at a time when they are urgently needed. This review indicates that almost all pubished prediction models are poorly reported, and at high risk of bias such that their reported predictive performance is probably optimistic. However, we have identified two (one diagnostic and one prognostic) promising models that should soon be validated in multiple cohorts, preferably through collaborative efforts and data sharing to also allow an investigation of the stability and heterogeneity in their performance across populations and settings. Details on all

RESEARCH

thebmj

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Marc M J Bonten,[8,10] Darren L Dahly,[11,12] Johanna A Damen,[8,9] Thomas P A Debray,[8,9] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,13] Paula Dhiman,[4,5] Maria C Haller,[7,14] Michael O Harhay,[15,16] Liesbet Henckaerts,[17,18] Pauline Heus,[8,9] Michael Kammer,[7,19] Nina Kreuzberger,[20] Anna Lohmann,[21] Kim Luijken,[21] Jie Ma,[5] Glen P Martin,[22] David J McLernon,[23] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[24,25] Chunhu Shi,[26] Nicole Skoetz,[19] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[27] René Spijker,[8,9,28] Ewout W Steyerberg,[3] Toshihiko Takada,[8] Ioanna Tzoulaki,[29,30] Sander M J van Kuijk,[31] Bas C T van Bussel,[1,32] Iwan C C van der Horst,[32] Florien S van Royen,[8] Jan Y Verbakel,[33,34] Christine Wallisch,[7,35,36] Jack Wilkinson,[22] Robert Wolff,[37] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

Table 2. PROBAST: Summary of Step 3—Assessment of Risk of Bias and Concerns Regarding Applicability*

1. Participants	2. Predictors	3. Outcome	4. Analysis
Signaling questions			
1.1. Were appropriate data sources used, e.g., cohort, RCT, or nested case–control study data?	2.1. Were predictors defined and assessed in a similar way for all participants?	3.1. Was the outcome determined appropriately?	4.1. Were there a reasonable number of participants with the outcome?
1.2. Were all inclusions and exclusions of participants appropriate?	2.2. Were predictor assessments made without knowledge of outcome data?	3.2. Was a prespecified or standard outcome definition used?	4.2. Were continuous and categorical predictors handled appropriately?
-	2.3. Are all predictors available at the time the model is intended to be used?	3.3. Were predictors excluded from the outcome definition?	4.3. Were all enrolled participants included in the analysis?
-	-	3.4. Was the outcome defined and determined in a similar way for all participants?	4.4. Were participants with missing data handled appropriately?
-	-	3.5. Was the outcome determined without knowledge of predictor information?	4.5. Was selection of predictors based on univariable analysis avoided?†
-	-	3.6. Was the time interval between predictor assessment and outcome determination appropriate?	4.6. Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?
		-	4.7. Were relevant model performance measures evaluated appropriately?
		-	4.8. Were model overfitting, underfitting, and optimism in model performance accounted for?†
-	-	-	4.9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?†

RESEARCH

thebmj

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., ... & PROBAST Group†. (2019). PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, *170*(1), 51-58.

Covid-19 prognostic models (Jan 2021)

Prediction models for diagnosis and pro systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Rich Ewoud Schuit,[8,9] Marc M J Bonten,[8,10] Darren L Dahly,[11,12] J Thomas P A Debray,[8,9] Valentijn M T de Jong,[8,9] Maarten De Maria C Haller,[7,14] Michael O Harhay,[15,16] Liesbet Henckaer Michael Kammer,[7,19] Nina Kreuzberger,[20] Anna Lohmann,[2] Glen P Martin,[22] David J McLernon,[23] Constanza L Andaur M Jamie C Sergeant,[24,25] Chunhu Shi,[26] Nicole Skoetz,[19] Luc J Matthew Sperrin,[27] René Spijker,[8,9,28] Ewout W Steyerberg, Ioanna Tzoulaki,[29,30] Sander M J van Kuijk,[31] Bas C T van Bu Florien S van Royen,[8] Jan Y Verbakel,[33,34] Christine Wallisch Robert Wolff,[37] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten

https://www.covprecise.org/living-review/

Show 40 entries Search:

REF	OUTCOME	RISK OF BIAS
Abbas, Abdelsamea, et al.	covid-19 diagnosis	High
Abdulaal, Patel et al	mortality (in hospital)	High
Abdulaal, Patel et al 2	mortality (in hospital)	High
Acar, Can et al	mortality (in hospital)	High
Al - Najjar, Al-Rousan	mortality (in or out of hospital)	High
Al - Najjar, Al-Rousan	other	High
Al Hassan, Cocks et al	other (composite)	High
Al Hassan, Cocks et al	other (composite)	High
Al Hassan, Cocks et al	other (composite)	High
Alafif, Alotaibi et al	other	High
Aliberti, Covinsky et al	#N/A	High
Allenbach, Saadoun et al	other (composite)	High
Altschul, Unda et al	mortality (in hospital)	High
Alvarez-Mon, Ortega et al	other (composite)	High
An, Lim et al	mortality (in or out of hospital)	High
An, Lim et al	mortality (in or out of hospital)	High
Angelov, Soares	covid-19 diagnosis	High
Anurag, Preetam	severe covid-19	High
Anurag, Preetam	severe covid-19	High
Anurag, Preetam	severe covid-19	High
Apostolopoulos, Aznaouridis et al	covid-19 diagnosis	High
Apostolopoulos, Mpesiana.	covid-19 diagnosis	High
Ardakani, Kanafi et al	covid-19 diagnosis	High
Arpan, Surya et al	covid-19 diagnosis	High
Artero, Madrazo et al	mortality (in hospital)	High
Artero, Madrazo et al	mortality (in hospital)	High
Artero, Madrazo et al	mortality (in hospital)	High
Artero, Madrazo et al	mortality (in hospital)	High

thebmj
Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

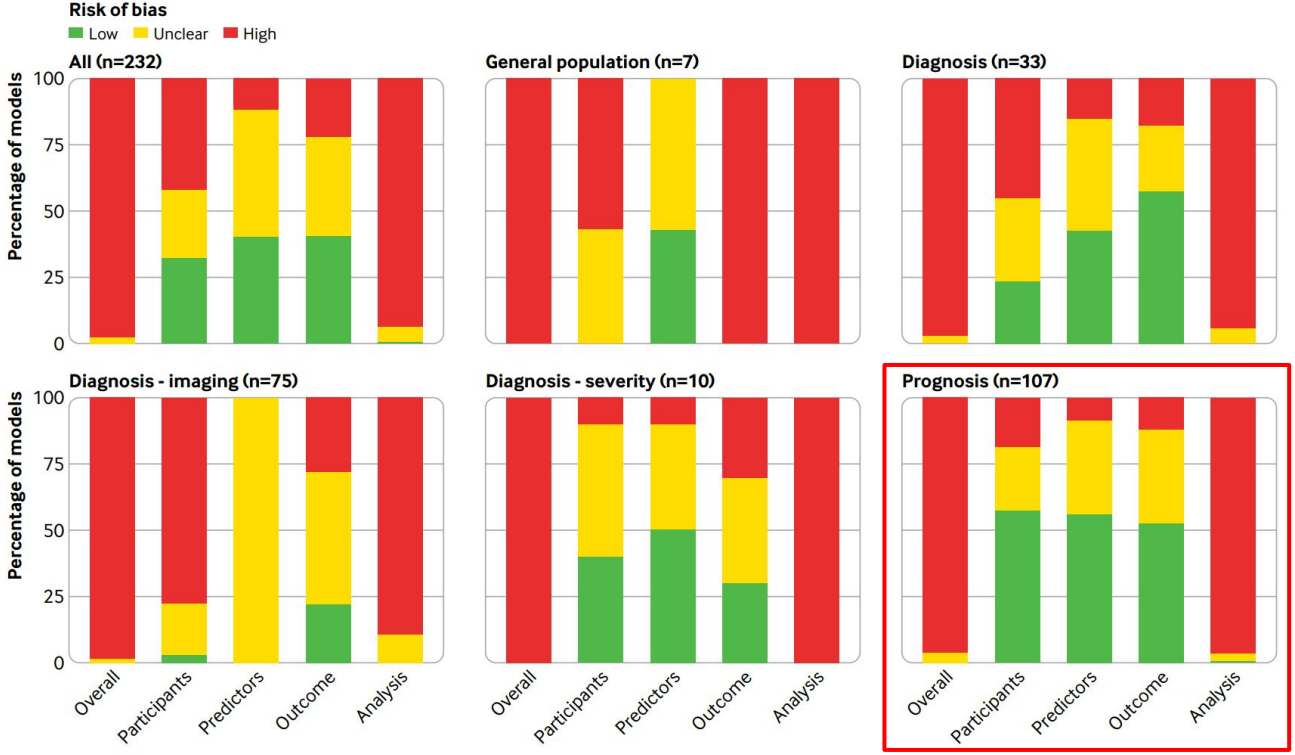


Fig 2 | PROBAST (prediction model risk of bias assessment tool) risk of bias for all included models combined (n=232) and broken down per type of model

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., ... & PROBAST Group†. (2019). PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, *170*(1), 51-58.

Early systematic external validation (Aug 2020)

ORIGINAL ARTICLE
PULMONARY INFECTIONS

Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study

Rishi K. Gupta[1,2], Michael Marks[2,3], Thomas H.A. Samuels[2], Akish Luintel[2], Tommy Rampling[2], Humayra Chowdhury[2], Matteo Quartagno[4], Arjun Nair[2], Marc Lipman[5], Ibrahim Abubakar[1], Maarten van Smeden[6], Wai Keong Wong[2], Bryan Williams[7,8] and Mahdad Noursadeghi[2,9], on behalf of The UCLH COVID-19 Reporting Group[10]

@ERSpublications
Oxygen saturation on room air and patient age are strong predictors of deterioration and mortality, respectively, among hospitalised adults with COVID-19. None of the 22 prognostic models evaluated in this study adds incremental value to these univariable predictors. https://bit.ly/2Hg24TO

ABSTRACT The number of proposed prognostic models for coronavirus disease 2019 (COVID-19) is growing rapidly, but it is unknown whether any are suitable for widespread clinical implementation.

We independently externally validated the performance of candidate prognostic models, identified through a living systematic review, among consecutive adults admitted to hospital with a final diagnosis of COVID-19. We reconstructed candidate models as per original descriptions and evaluated performance for their original intended outcomes using predictors measured at the time of admission. We assessed discrimination, calibration and net benefit, compared to the default strategies of treating all and no patients, and against the most discriminating predictors in univariable analyses.

We tested 22 candidate prognostic models among 411 participants with COVID-19, of whom 180 (43.8%) and 115 (28.0%) met the endpoints of clinical deterioration and mortality, respectively. Highest areas under receiver operating characteristic (AUROC) curves were achieved by the NEWS2 score for prediction of deterioration over 24 h (0.78, 95% CI 0.73–0.83), and a novel model for prediction of deterioration <14 days from admission (0.78, 95% CI 0.74–0.82). The most discriminating univariable predictors were admission oxygen saturation on room air for in-hospital deterioration (AUROC 0.76, 95% CI 0.71–0.81), and age for in-hospital mortality (AUROC 0.76, 95% CI 0.71–0.81). No prognostic model demonstrated consistently higher net benefit than these univariable predictors, across a range of threshold probabilities.

Admission oxygen saturation on room air and patient age are strong predictors of deterioration and mortality among hospitalised adults with COVID-19, respectively. None of the prognostic models evaluated here offered incremental value for patient stratification to these univariable predictors.

Systematic reviews

- Define what population/situation your model is for (Hospitalized patients? Low resource setting? Trial enrichment?)
- Do internal validation (address over optimism)
- Stop making new models and start validating existing models
- Don’t forget calibration

CCX paper

Unique RCT dataset - smaller and fewer variables than an EHR or most other convenience samples, but presumably higher quality and uniform follow-up.

Initial ideas:

1) validate known, easily measurable scores

2) make our own score!!!

CCX paper

After reading the systematic reviews:

1) Let’s focus on validating existing, easily measurable scores
2) Instead of brand new scores, let’s evaluate smaller tweaks to existing measures in pre-specified way

CCX paper

Let’s pre-register*

- Scientific benefit: Strengthens study.
- **Practical benefit: limits scope and off-the-cuff analysis requests**

Analyze each arm separately

- Unclear how fundamentally different the two arms were: conservatively treat as different populations

Because we were validating existing models, and we pre-registered, internal validation (cross-validation, training/test sets) was not necessary.

*In practice, the journal did not send out to statistical reviewers so I doubt anyone even noticed the pre-registration. In hindsight, should have emphasized this more.

CCX paper

Prognostic risk scores:

NEWS - existing score. Evidence NEWS2 was promising.

MEWS - related, existing score. Evidence NEWS2 was promising.

NEWS/MEWS + age - previously published, tweak to NEWS.

NEWS/MEWS over time* - take advantage of daily measurements

How well do those risk scores predict these endpoints:

Time to recovery

Time to mortality

Time to deterioration

*Hindsight - take deeper look at time-varying advantage of NEWS/MEWS

Pre-registration October 2020

- After literature review and several discussions with collaborators
- Before looking at any data (honor system - no way to verify)

We used https://aspredicted.org

There is also https://osf.io/

Validation of existing COVID-19 prognostic models using ACTT-1 trial data (#48985)

Created: 10/05/2020 11:45 PM (PT)
Public: 06/16/2021 10:27 AM (PT)

Author(s)
C. Jason Liang (National Institute of Allergy and Infectious Diseases) - jason.liang@nih.gov
Christopher Colombo (Madigan Army Medical Center) - christopher.j.colombo.mil@mail.mil
Rhonda Colombo (Madigan Army Medical Center) - rhonda.e.colombo.ctr@mail.mil
Ryan Maves (Naval Medical Center San Diego) - ryan.c.maves.mil@mail.mil

1) Have any data been collected for this study already?
It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?
Since the beginning of the COVID-19 pandemic, there have been at least 40 published prognostic models for predicting COVID-19 progression. A living systematic review (Wynants et al. 2020) has concluded that the majority of the published models are not suitable for clinical practice. A systematic validation of 22 prognostic models (Gupta et al. 2020) concluded that the new models did not appear to outperform the established NEWS2 score. Furthermore, the new models did not offer added prognostic performance over single variables such as age or oxygen saturation.

There is a need to both cull and validate the existing collection of COVID-19 prognostic risk models before the models can be implemented in practice. Potential applications include supporting clinical decision making and enriching treatment clinical trials to target higher risk individuals.

Additionally, as the standard of care evolves in response to the rapidly growing body of knowledge around COVID-19, there is a need to both monitor the durability of the accuracy of existing prognostic models and assess whether modifications to the models can improve predictive accuracy.

Question: How well do the established and best-performing of the existing prognostic risk scores, some of which were meant for more general settings, apply to a population of COVID-19 patients on placebo?

We will use data from the adaptive COVID-19 treatment trial (ACTT-1) of remdesivir vs. placebo to answer our question. Though the trial is completed, none of the co-authors have seen the data yet.

https://aspredicted.org/kq798.pdf

NEWS

National Early Warning Score

Developed for acute and critical care setting.

“It is also recommended that the NEWS is used as a surveillance system for all patients in hospitals, tracking their clinical condition, alerting the clinical team to any clinical deterioration and triggering a timely clinical response.”

PHYSIOLOGICAL PARAMETERS	3	2	1	0	1	2	3
Respiration Rate	≤8		9 - 11	12 - 20		21 - 24	≥25
Oxygen Saturations	≤91	92 - 93	94 - 95	≥96			
Any Supplemental Oxygen		Yes		No			
Temperature	≤35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	≥39.1	
Systolic BP	≤90	91 - 100	101 - 110	111 - 219			≥220
Heart Rate	≤40		41 - 50	51 - 90	91 - 110	111 - 130	≥131
Level of Consciousness				A			V, P, or U

Royal College of Physicians. National Early Warning Score (NEWS): Standardising the assessment of acute illness severity in the NHS. Report of a working party. London: RCP, 2012.

NEWS + Age

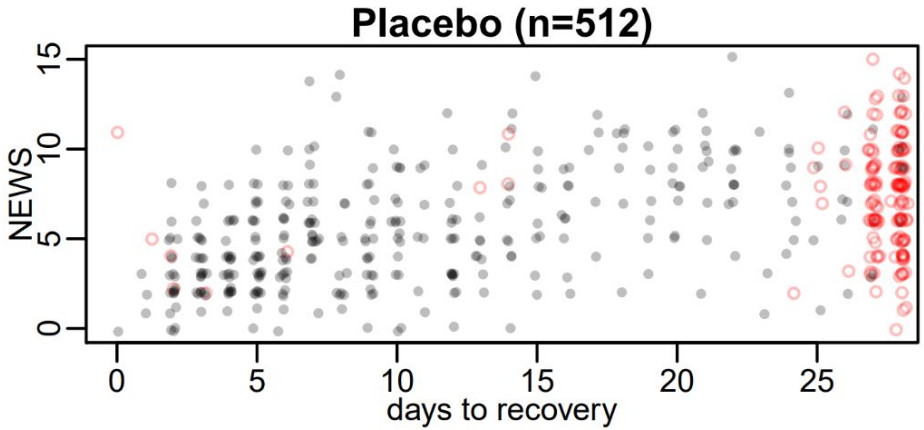
National Early Warning Score

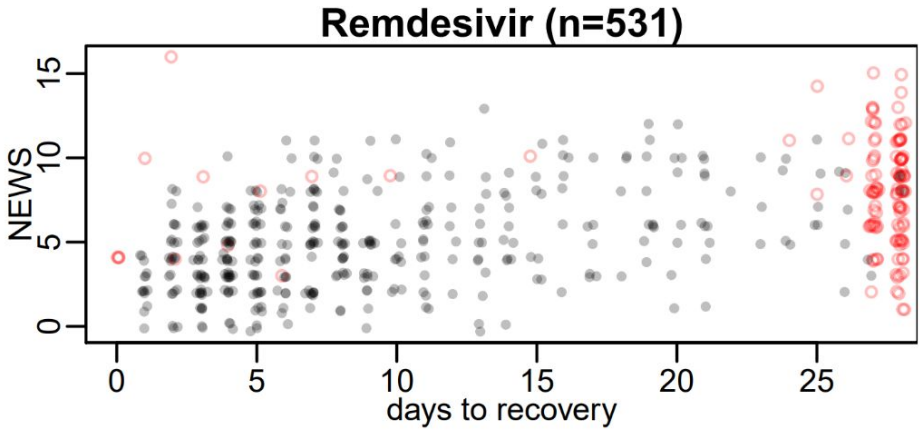
+ Age

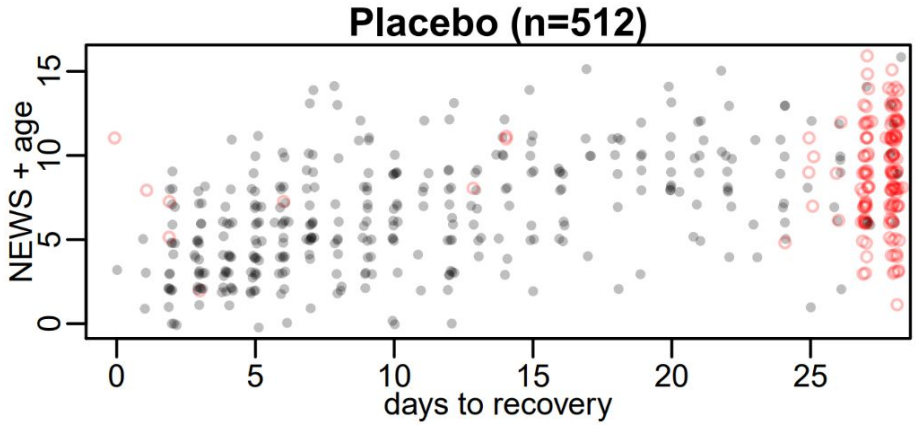
PHYSIOLOGICAL PARAMETERS	3	2	1	0	1	2	3
Respiration Rate	≤8		9 - 11	12 - 20		21 - 24	≥25
Oxygen Saturations	≤91	92 - 93	94 - 95	≥96			
Any Supplemental Oxygen		Yes		No			
Temperature	≤35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	≥39.1	
Systolic BP	≤90	91 - 100	101 - 110	111 - 219			≥220
Heart Rate	≤40		41 - 50	51 - 90	91 - 110	111 - 130	≥131
Level of Consciousness				A			V, P, or U
Age	≥65			<65			

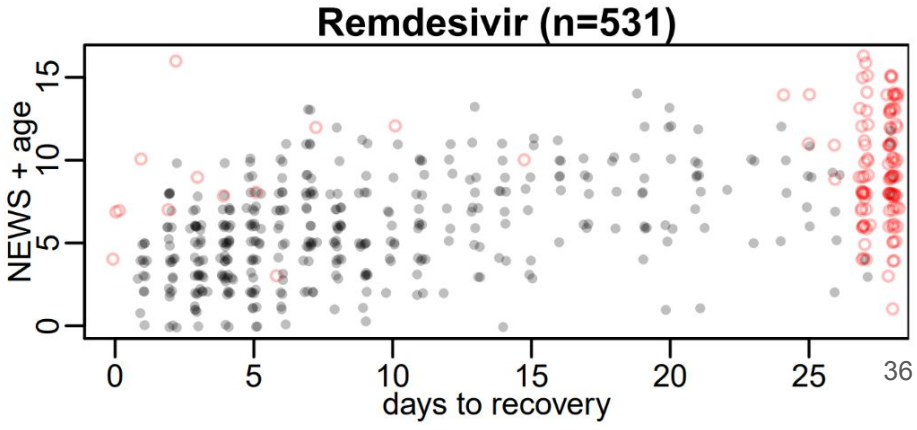
Liao, X., Wang, B., & Kang, Y. (2020). Novel coronavirus infection during the 2019–2020 epidemic: preparing intensive care units—the experience in Sichuan Province, China. *Intensive care medicine*, *46*(2), 357-360.

CCX paper


Placebo (n=512)
NEWS
0
5
10
15
0
5
10
15
20
25
days to recovery


Remdesivir (n=531)
NEWS
0
5
10
15
0
5
10
15
20
25
days to recovery


Placebo (n=512)
NEWS + age
0
5
10
15
0
5
10
15
20
25
days to recovery


Remdesivir (n=531)
NEWS + age
0
5
10
15
0
5
10
15
20
25
days to recovery

CCX paper

	Placebo C-Index	Remdesivir C-Index	Placebo 14-d AUC	Remdesivir 14-d AUC	Placebo 28-d AUC	Remdesivir 28-d AUC
Risk score for mortality end point						
NEWS	0.60 (0.54–0.66)	0.68 (0.61–0.74)	0.59 (0.51–0.66)	0.70 (0.60–0.79)	0.61 (0.55–0.68)	0.68 (0.61–0.76)
NEWS+age	0.66 (0.60–0.72)	0.73 (0.67–0.79)	0.65 (0.58–0.71)	0.77 (0.69–0.84)	0.67 (0.61–0.74)	0.74 (0.68–0.81)
MEWS	0.59 (0.53–0.65)	0.66 (0.60–0.73)	0.56 (0.48–0.64)	0.65 (0.56–0.73)	0.60 (0.53–0.67)	0.67 (0.60–0.74)
MEWS+age	0.67 (0.61–0.73)	0.74 (0.68–0.80)	0.64 (0.57–0.71)	0.76 (0.68–0.83)	0.69 (0.63–0.75)	0.75 (0.68–0.81)
NEWS avg	0.66 (0.60–0.71)	0.71 (0.64–0.78)	0.66 (0.58–0.72)	0.73 (0.63–0.82)	0.67 (0.61–0.73)	0.72 (0.64–0.79)
NEWS slope	0.63 (0.56–0.69)	0.55 (0.47–0.63)	0.67 (0.59–0.76)	0.54 (0.42–0.65)	0.62 (0.56–0.70)	0.56 (0.47–0.64)
MEWS avg	0.65 (0.59–0.70)	0.71 (0.64–0.77)	0.64 (0.56–0.70)	0.71 (0.61–0.79)	0.66 (0.60–0.72)	0.72 (0.65–0.78)
MEWS slope	0.57 (0.49–0.64)	0.53 (0.45–0.61)	0.62 (0.54–0.70)	0.60 (0.51–0.71)	0.56 (0.48–0.64)	0.53 (0.44–0.61)
Risk score for recovery end point						
NEWS	0.68 (0.65–0.71)	0.69 (0.67–0.72)	0.76 (0.72–0.81)	0.79 (0.74–0.83)	0.67 (0.62–0.71)	0.76 (0.71–0.81)
NEWS+age	0.70 (0.67–0.72)	0.71 (0.69–0.74)	0.78 (0.74–0.82)	0.80 (0.76–0.84)	0.70 (0.65–0.75)	0.78 (0.73–0.82)
MEWS	0.65 (0.62–0.68)	0.67 (0.64–0.70)	0.72 (0.68–0.77)	0.76 (0.72–0.08)	0.66 (0.60–0.71)	0.74 (0.68–0.79)
MEWS+age	0.68 (0.65–0.71)	0.69 (0.67–0.72)	0.75 (0.71–0.79)	0.78 (0.73–0.81)	0.70 (0.65–0.75)	0.76 (0.70–0.81)
NEWS avg	0.72 (0.69–0.74)	0.73 (0.71–0.76)	0.82 (0.78–0.85)	0.82 (0.78–0.85)	0.72 (0.67–0.77)	0.8 (0.75–0.84)
NEWS slope	0.58 (0.55–0.61)	0.56 (0.53–0.59)	0.60 (0.55–0.65)	0.56 (0.51–0.61)	0.62 (0.56–0.68)	0.56 (0.49–0.61)
MEWS avg	0.70 (0.67–0.72)	0.72 (0.69–0.74)	0.79 (0.75–0.82)	0.81 (0.77–0.85)	0.72 (0.67–0.77)	0.79 (0.74–0.83)
MEWS slope	0.56 (0.53–0.59)	0.53 (0.50–0.56)	0.57 (0.52–0.62)	0.52 (0.47–0.58)	0.58 (0.52–0.63)	0.52 (0.46–0.58)
Risk score for deterioration end point						
NEWS	0.69 (0.64–0.75)	0.65 (0.58–0.71)	0.71 (0.65–0.78)	0.65 (0.58–0.73)	0.72 (0.65–0.78)	0.65 (0.58–0.72)
NEWS+age	0.70 (0.65–0.75)	0.66 (0.59–0.73)	0.73 (0.66–0.79)	0.67 (0.59–0.75)	0.74 (0.67–0.80)	0.68 (0.61–0.74)
MEWS	0.62 (0.56–0.68)	0.59 (0.53–0.66)	0.62 (0.56–0.69)	0.61 (0.54–0.68)	0.64 (0.57–0.71)	0.60 (0.53–0.67)
MEWS+age	0.64 (0.58–0.70)	0.63 (0.57–0.70)	0.66 (0.59–0.73)	0.65 (0.56–0.72)	0.68 (0.61–0.74)	0.65 (0.57–0.72)
NEWS avg	0.78 (0.73–0.83)	0.71 (0.65–0.77)	0.82 (0.76–0.87)	0.73 (0.66–0.79)	0.82 (0.76–0.87)	0.73 (0.66–0.80)
NEWS slope	0.61 (0.55–0.67)	0.59 (0.51–0.66)	0.63 (0.56–0.70)	0.59 (0.51–0.68)	0.63 (0.56–0.70)	0.59 (0.51–0.67)
MEWS avg	0.71 (0.65–0.77)	0.67 (0.60–0.74)	0.73 (0.66–0.79)	0.68 (0.60–0.75)	0.73 (0.66–0.80)	0.68 (0.60–0.76)
MEWS slope	0.57 (0.51–0.64)	0.56 (0.49–0.64)	0.60 (0.52–0.68)	0.56 (0.47–0.63)	0.58 (0.51–0.66)	0.56 (0.48–0.64)

CCX paper

	Placebo C-Index	Remdesivir C-Index	Placebo 14-d AUC	Remdesivir 14-d AUC	Placebo 28-d AUC	Remdesivir 28-d AUC
Risk score for recovery end point						
NEWS	0.68 (0.65–0.71)	0.69 (0.67–0.72)	0.76 (0.72–0.81)	0.79 (0.74–0.83)	0.67 (0.62–0.71)	0.76 (0.71–0.81)
NEWS+age	0.70 (0.67–0.72)	0.71 (0.69–0.74)	0.78 (0.74–0.82)	0.80 (0.76–0.84)	0.70 (0.65–0.75)	0.78 (0.73–0.82)

CCX paper

NEWS is ok

NEWS + age is better

Averaging NEWS from first two days is also better

Not a crystal ball by any means

ORIGINAL CLINICAL REPORT

OPEN

Performance Analysis of the National Early Warning Score and Modified Early Warning Score in the Adaptive COVID-19 Treatment Trial Cohort

OBJECTIVES: We sought to validate prognostic scores in coronavirus disease 2019 including National Early Warning Score, Modified Early Warning Score, and age-based modifications, and define their performance characteristics.

DESIGN: We analyzed prospectively collected data from the Adaptive COVID-19 Treatment Trial. National Early Warning Score was collected daily during the trial, Modified Early Warning Score was calculated, and age applied to both scores. We assessed prognostic value for the end points of recovery, mechanical ventilation, and death for score at enrollment, average, and slope of score over the first 48 hours.

SETTING: A multisite international inpatient trial.

PATIENTS: A total of 1,062 adult nonpregnant inpatients with severe coronavirus disease 2019 pneumonia.

INTERVENTIONS: Adaptive COVID-19 Treatment Trial 1 randomized participants to receive remdesivir or placebo. The prognostic value of predictive scores was evaluated in both groups separately to assess for differential performance in the setting of remdesivir treatment.

MEASUREMENTS AND MAIN RESULTS: For mortality, baseline National Early Warning Score and Modified Early Warning Score were weakly to moderately prognostic (c-index, 0.60–0.68), and improved with addition of age (c-index, 0.66–0.74). For recovery, baseline National Early Warning Score and Modified Early Warning Score demonstrated somewhat better prognostic

Christopher J. Colombo, MD, MA, FACP, FCCM[1,2]

Rhonda E. Colombo, MD, MHS, FACP, FIDSA[1–3]

Ryan C. Maves, MD, FCCM, FCCP, FIDSA[2,4]

Angela R. Branche, MD[5]

Stuart H. Cohen, MD[6]

Marie-Carmelle Elie, MD[7]

Sarah L. George, MD[8]

Hannah J. Jang, PhD, RN, CNL, PHN[9]

Andre C. Kalil, MD, MPH[10]

David A. Lindholm, MD, FACP[2,11]

Richard A. Mularski, MD, MSHS, MCR, ATSF, FCCP, FACP[12]

Justin R. Ortiz, MD, MS, FACP, FCCP[13]

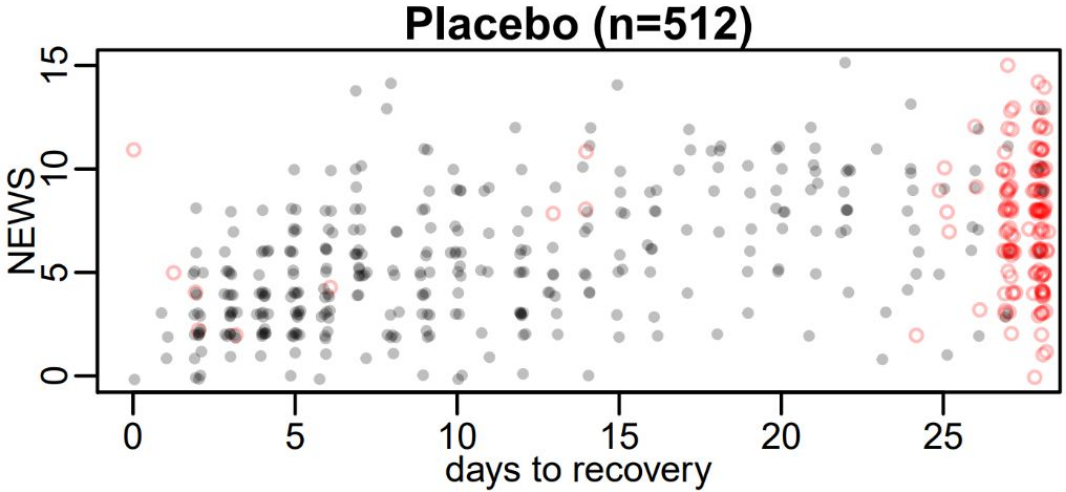
Victor Tapson, MD[14]

C. Jason Liang, PhD[15]

On behalf of the ACTT-1 Study Group

C-index

	Placebo C-Index	Remdesivir C-Index	Placebo 14-d AUC
Risk score for recovery end point			
NEWS	0.68 (0.65–0.71)	0.69 (0.67–0.72)	0.76 (0.72–0.81)
NEWS+age	0.70 (0.67–0.72)	0.71 (0.69–0.74)	0.78 (0.74–0.82)



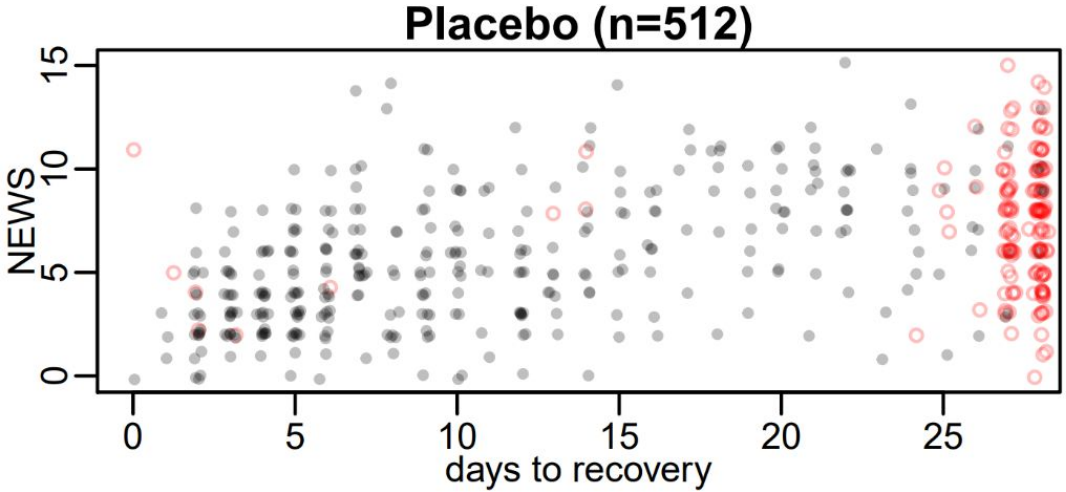
$$\text{C-index} = P(NEWS_i < NEWS_j | T_i < T_j)$$

$$\hat{P}(NEWS_i < NEWS_j | T_i < T_j) = 0.68$$

- When selecting a pair of individuals, we estimate a 68% chance the person who recovered earlier will have the lower NEWS score.
- If presented with two random individuals, the NEWS score will correctly order them 68% of the time.

C-index

	Placebo C-Index	Remdesivir C-Index	Placebo 14-d AUC
Risk score for recovery end point			
NEWS	0.68 (0.65–0.71)	0.69 (0.67–0.72)	0.76 (0.72–0.81)
NEWS+age	0.70 (0.67–0.72)	0.71 (0.69–0.74)	0.78 (0.74–0.82)

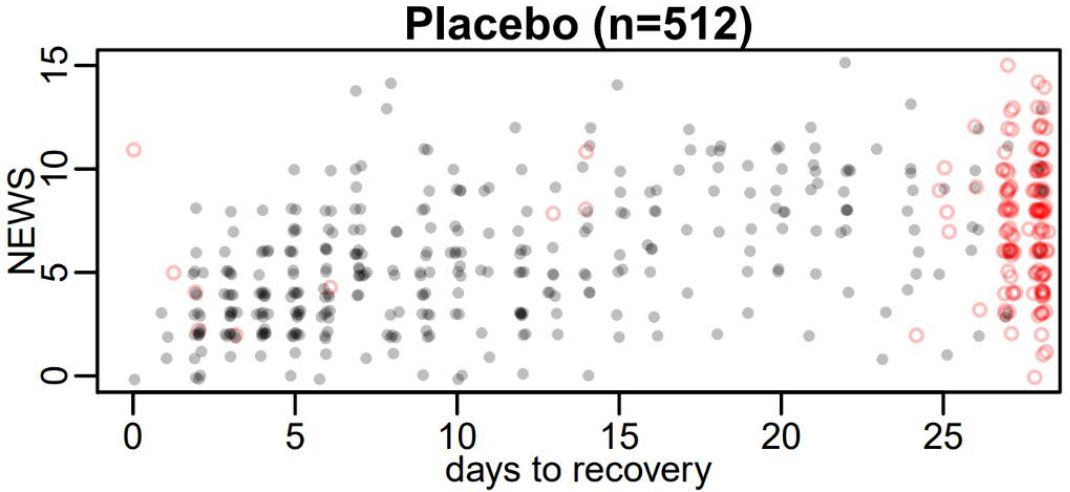


R package: `Hmisc::rcorr.cens()`

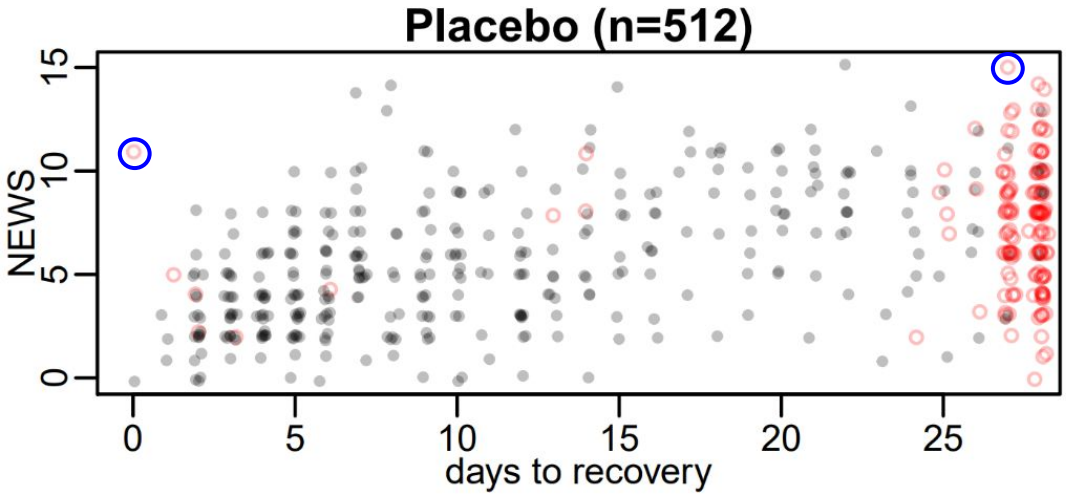
```
with(acttp, rcorr.cens(news, Surv(time.r, status.r)))
```

Exercise 1: Using the placebo arm, NEWS, and days to recovery, "hand code" an estimate for the C-index. Compare to results from `Hmisc::rcorr.cens()`. How are ties handled? How are censored observations handled?

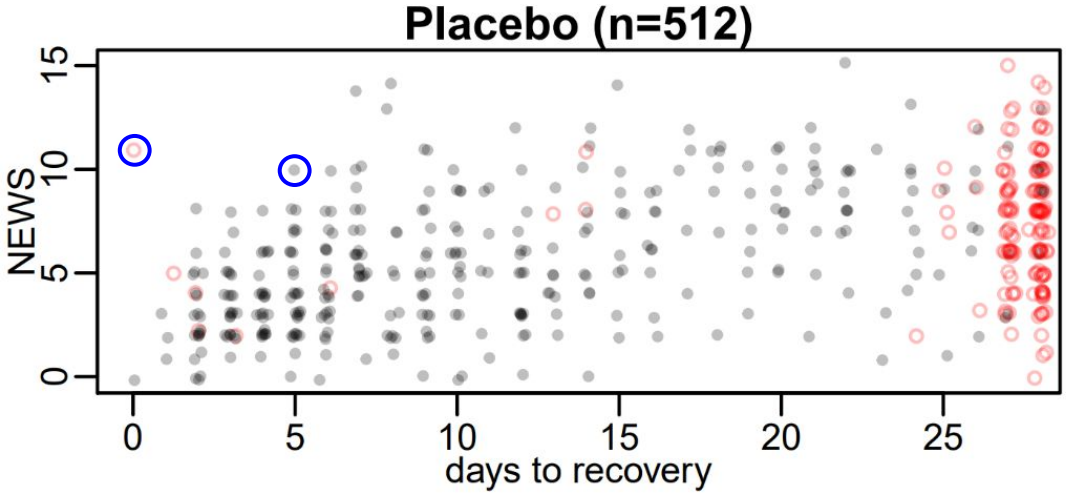
C-index estimation



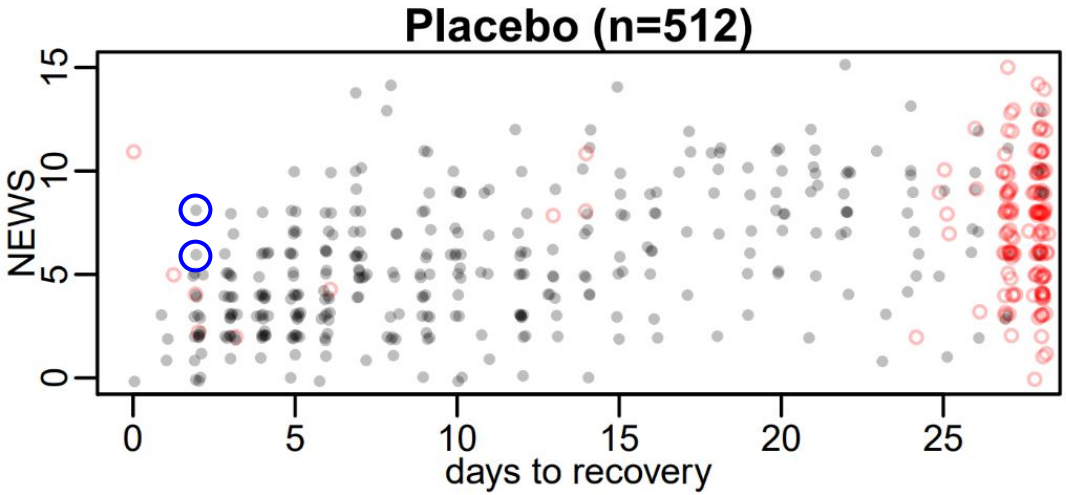
$$\hat{C} = \frac{\#\ \text{Concordant Pairs}}{\#\ \text{Relevant Pairs}}$$



Two censored observations
Not relevant pair: denominator +0
Concordance not relevant: numerator +0

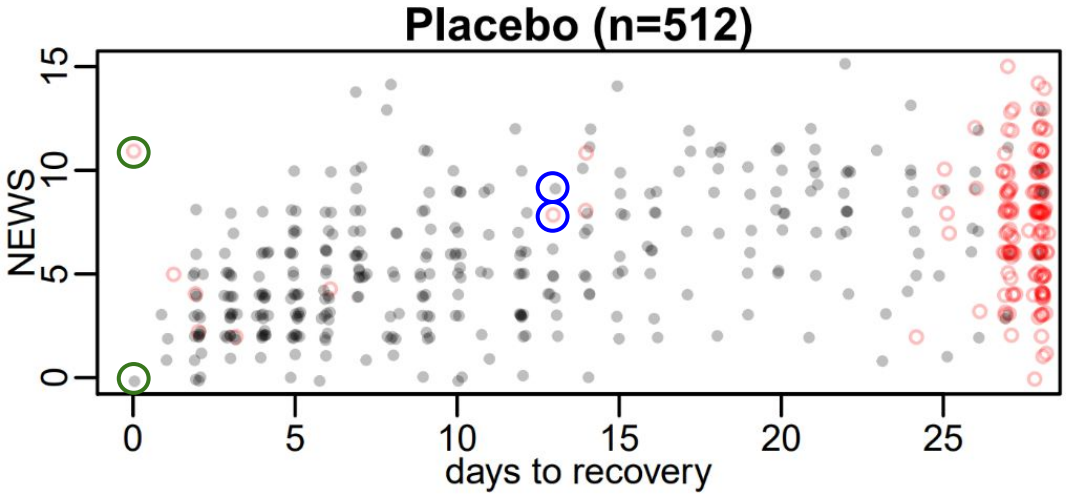


Censored observation **before** event observation
Not relevant pair: denominator +0
Concordance not relevant: numerator +0



Two events at same time

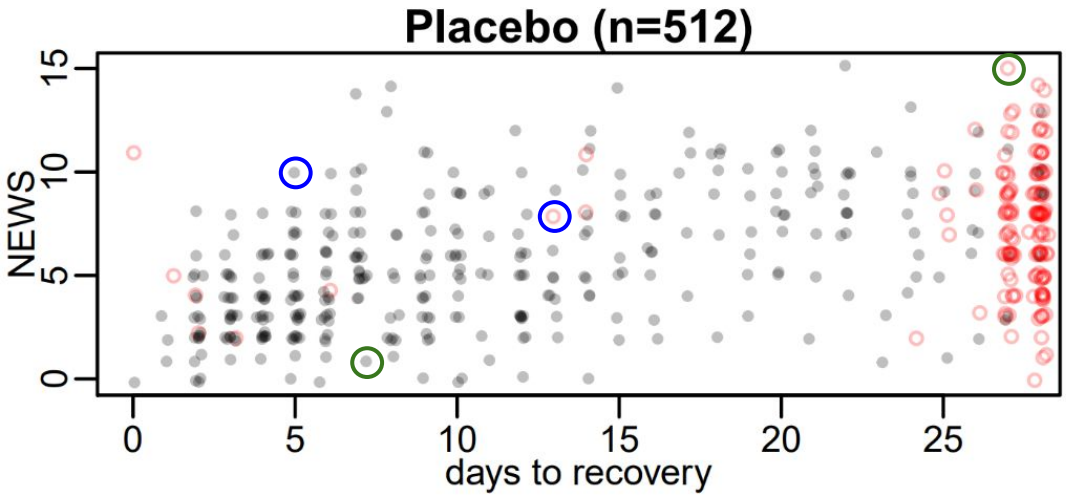
Not relevant pair: denominator +0
Concordance not relevant: numerator +0

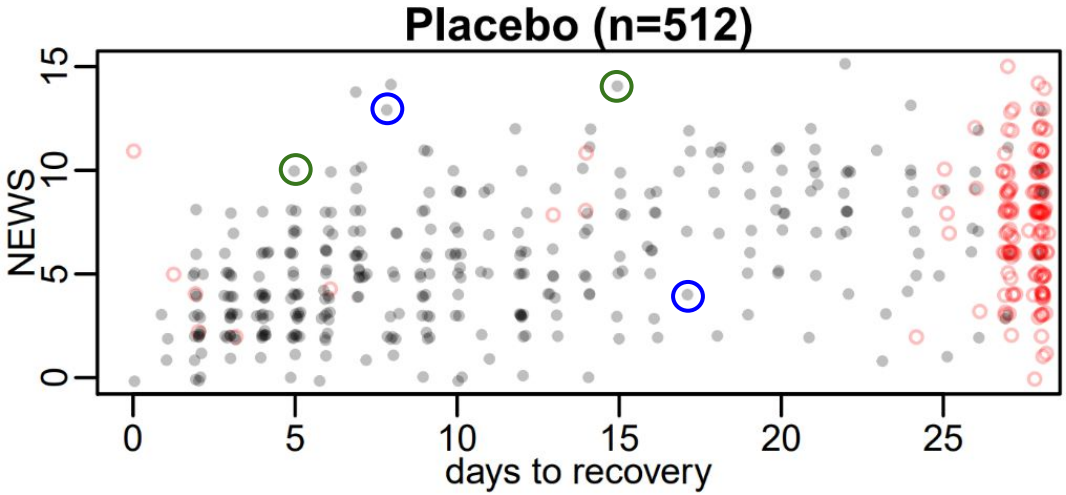


One censored, one event, at same time

Relevant pair: denominator +1
Assess concordance:

- If NEWS ordered correctly, numerator +1
- If NEWS ordered incorrectly, numerator +0
- If NEWS tied, numerator +0.5





Censored observation **after** event observation

Relevant pair: denominator +1
Assess concordance:

- If NEWS ordered correctly, numerator +1
- If NEWS ordered incorrectly, numerator +0
- If NEWS tied, numerator +0.5

Two events

Relevant pair: denominator +1
Assess concordance:

- If NEWS ordered correctly, numerator +1
- If NEWS ordered incorrectly, numerator +0
- If NEWS tied, numerator +0.5

Harrell’s c-index (1982) [1]

“Some of the examples of measures that are commonly used but are not needed in this setting are the c-index…”

-Frank Harrell (2019) [2], on the c-index for quantifying added value of biomarkers

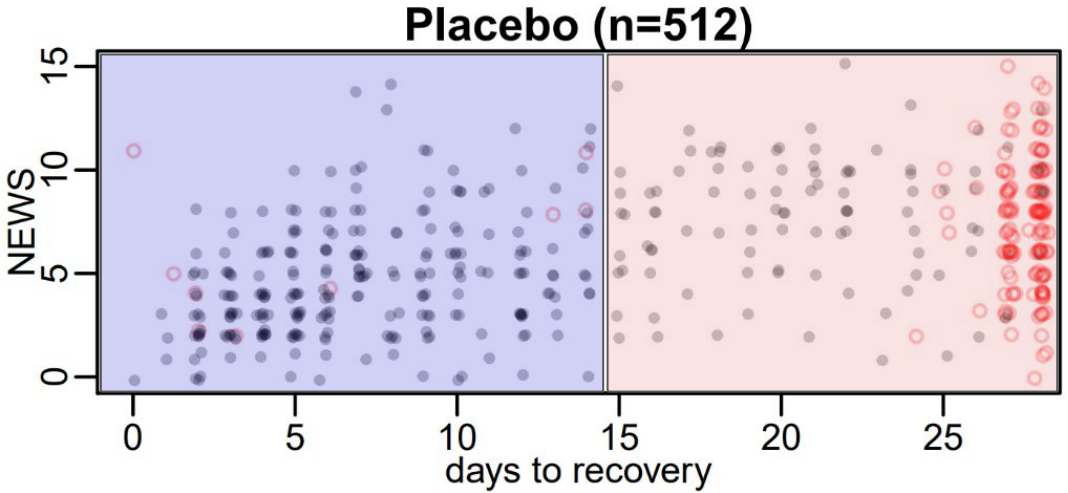
1. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543-2546.
2. https://www.fharrell.com/post/addvalue/

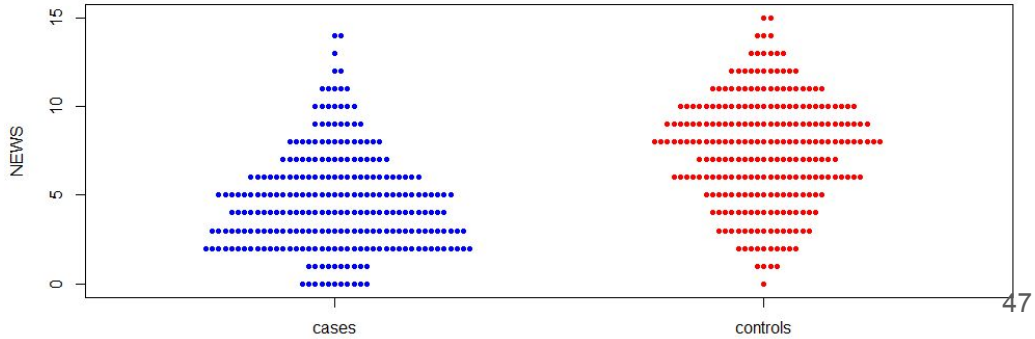
Final note on the c-index, and review of binary outcomes

Somewhat confusingly, the "c-index" can refer to both the above case with **censored data**, but is also used with binary outcome data.

With **binary outcomes**, the c-index = area under the receiver operating characteristic (ROC) curve (AUC)

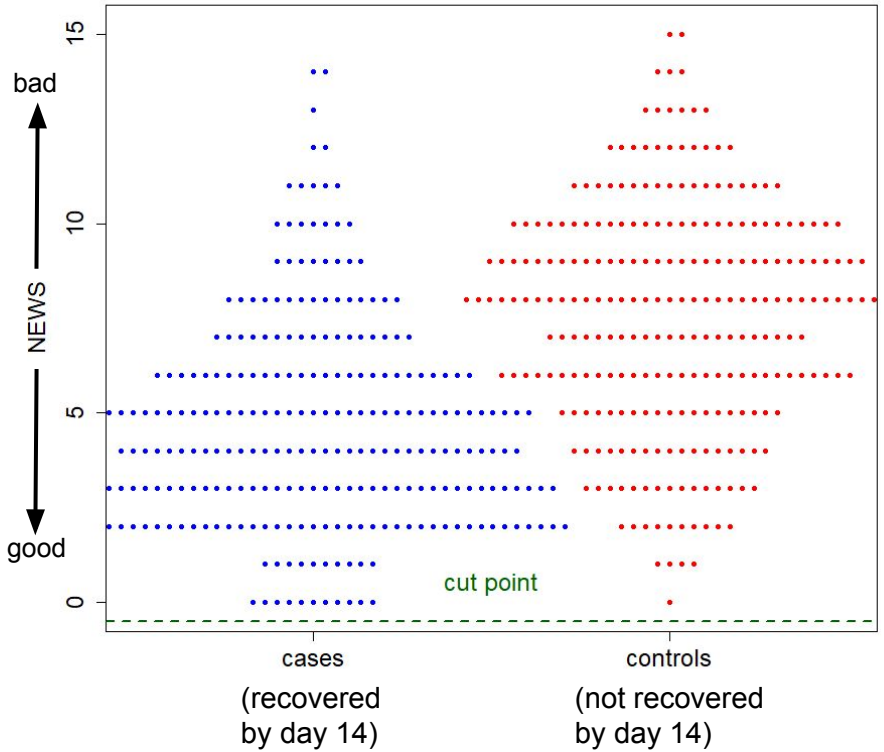
Example: **ignoring censoring***, suppose we dichotomize

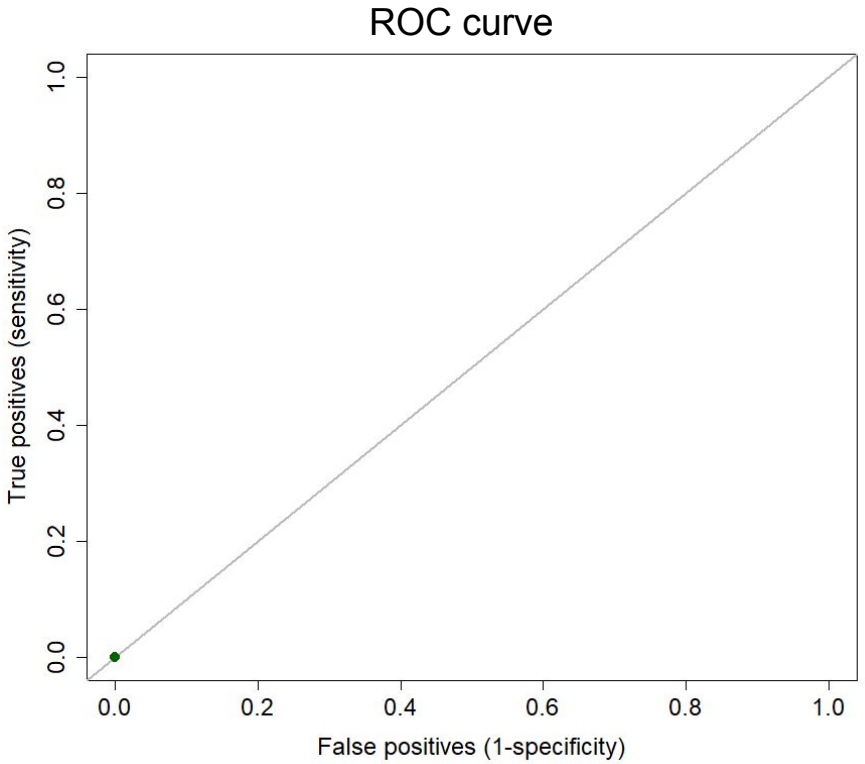




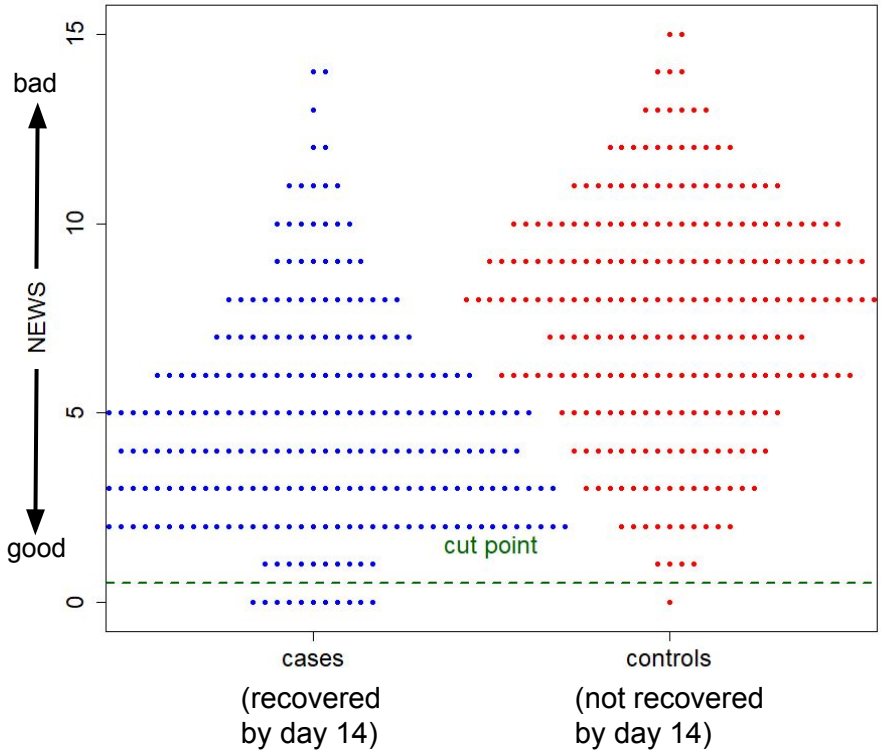
*** For illustrative purposes.**
In practice, please don't ignore censoring.

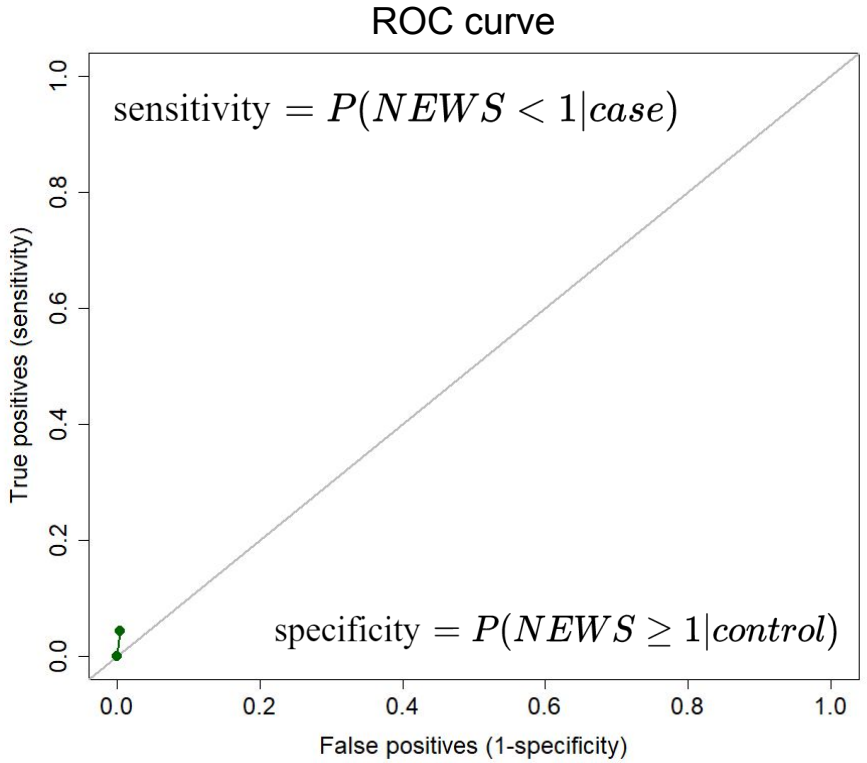
Review of ROC curves and AUC for **binary outcomes**



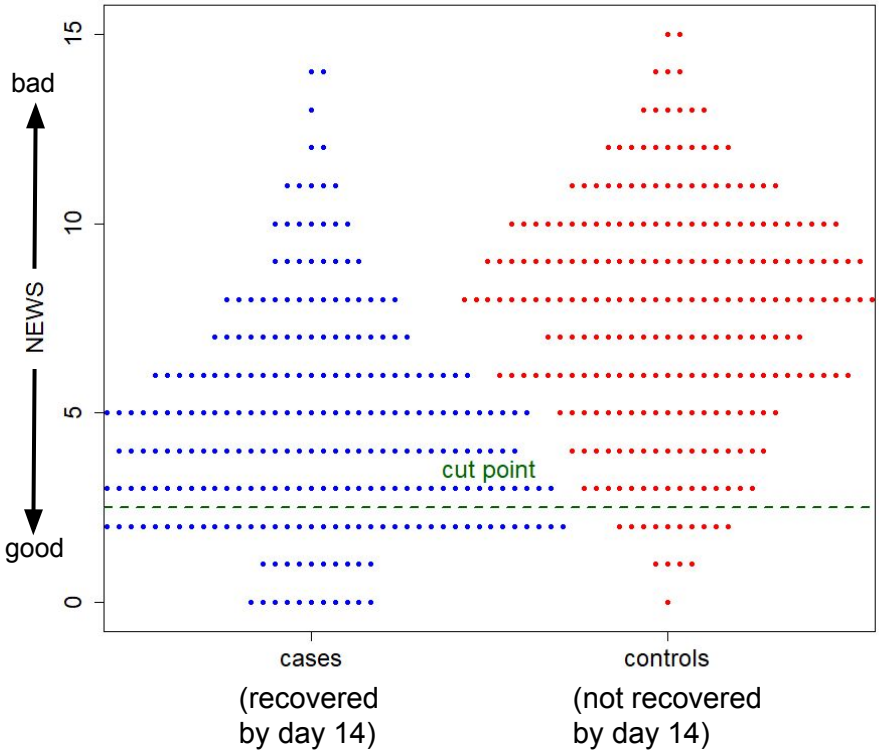


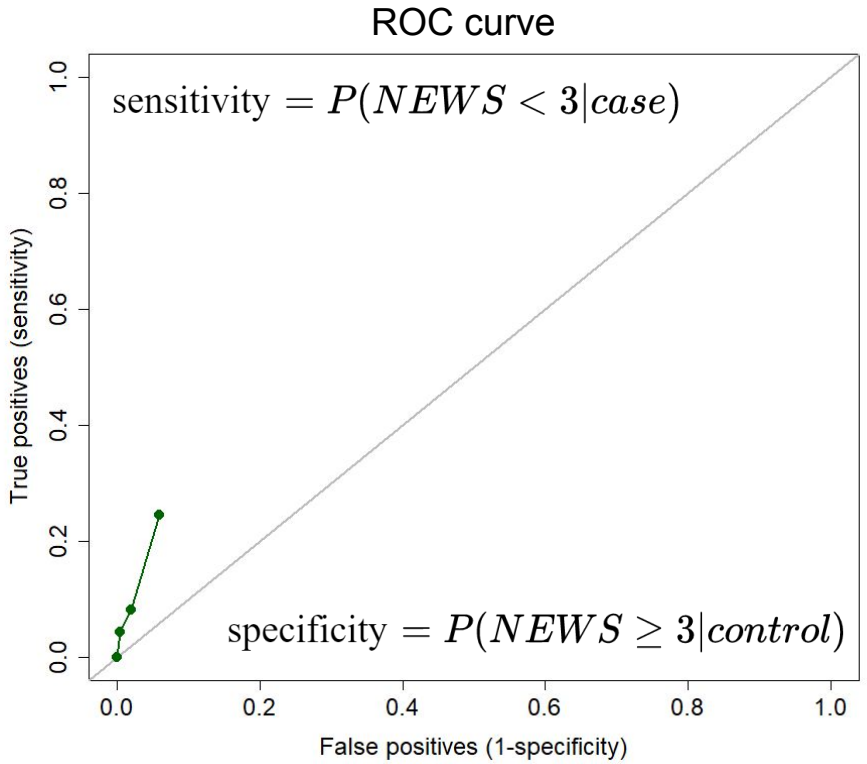
Review of ROC curves and AUC for binary outcomes



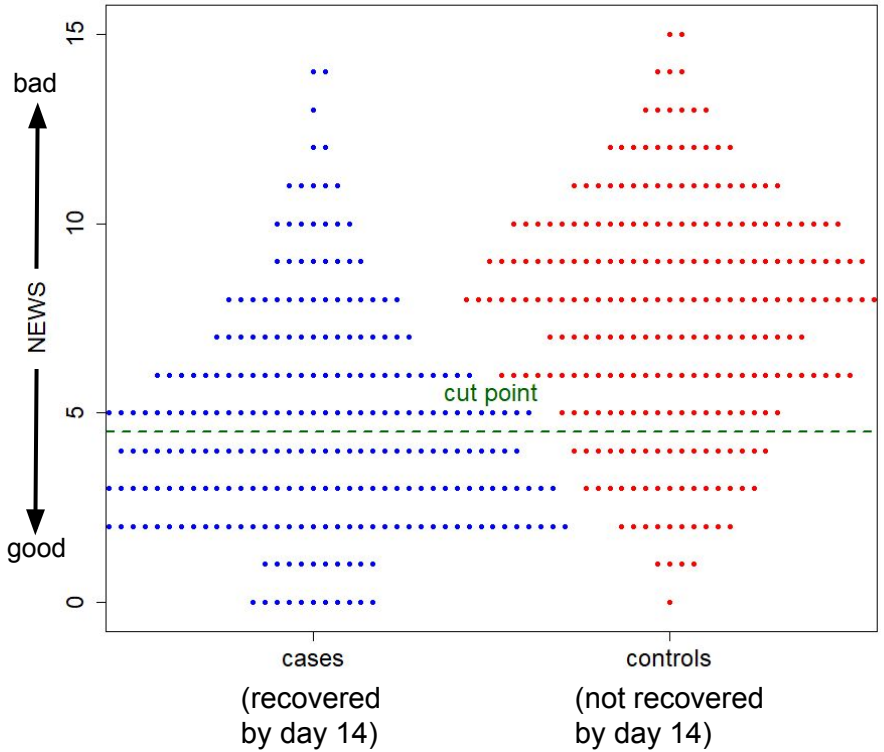


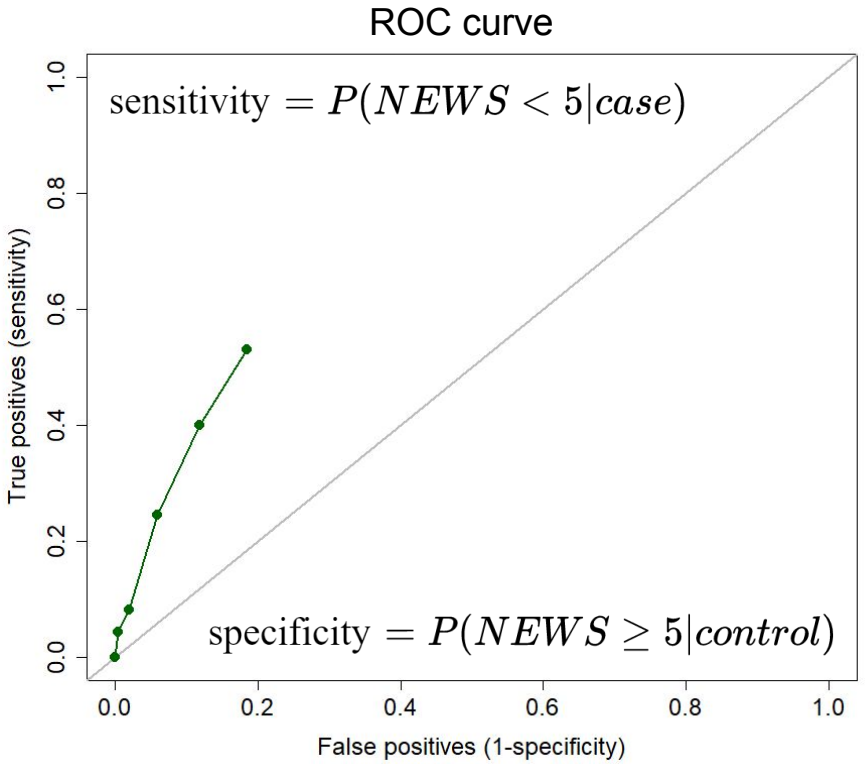
Review of ROC curves and AUC for binary outcomes



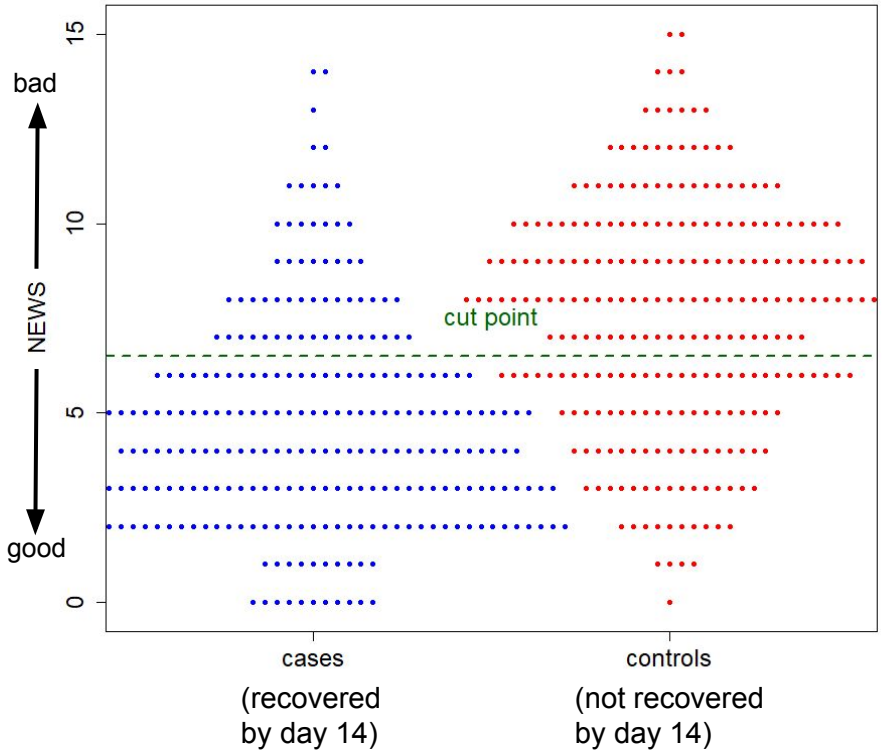


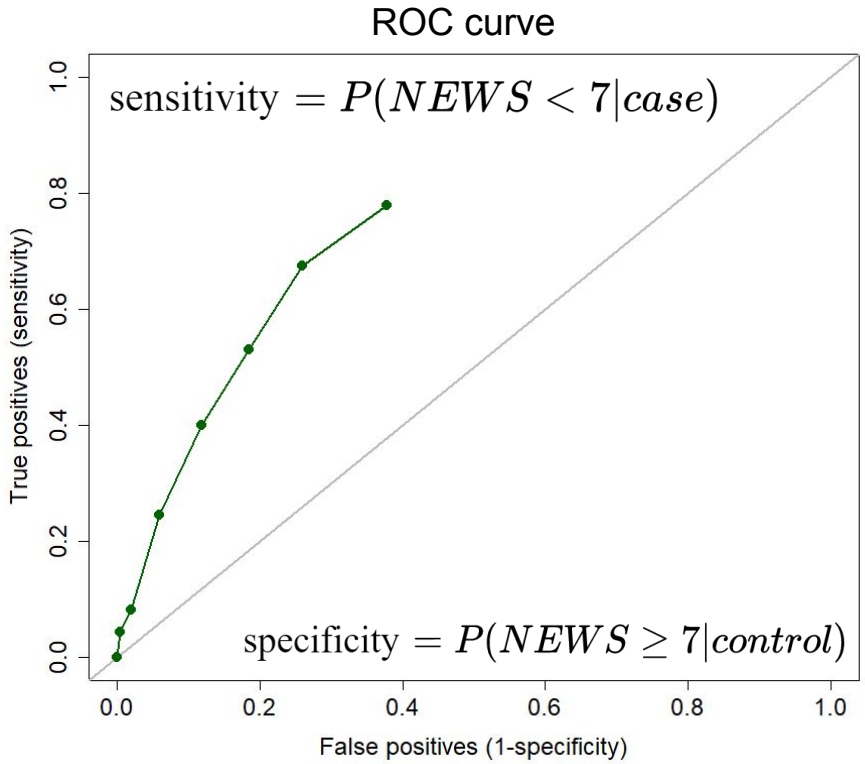
Review of ROC curves and AUC for binary outcomes



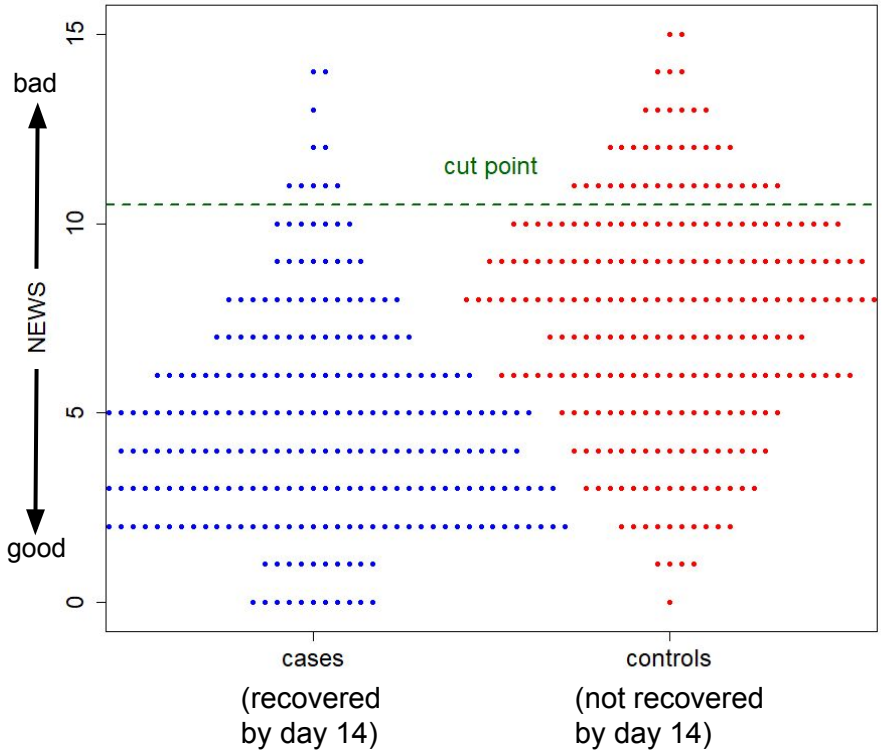


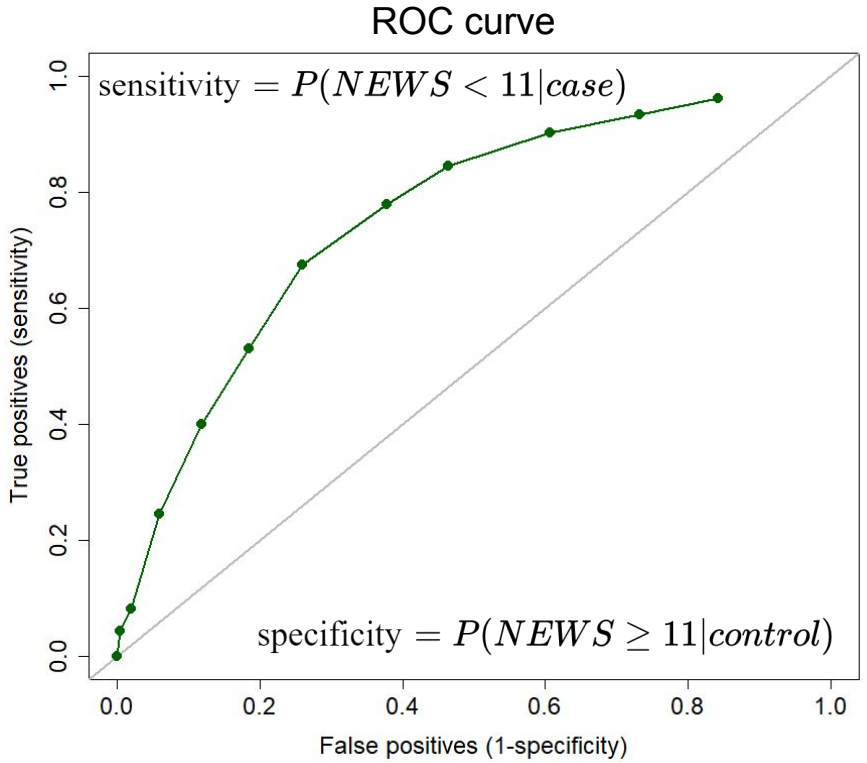
Review of ROC curves and AUC for binary outcomes



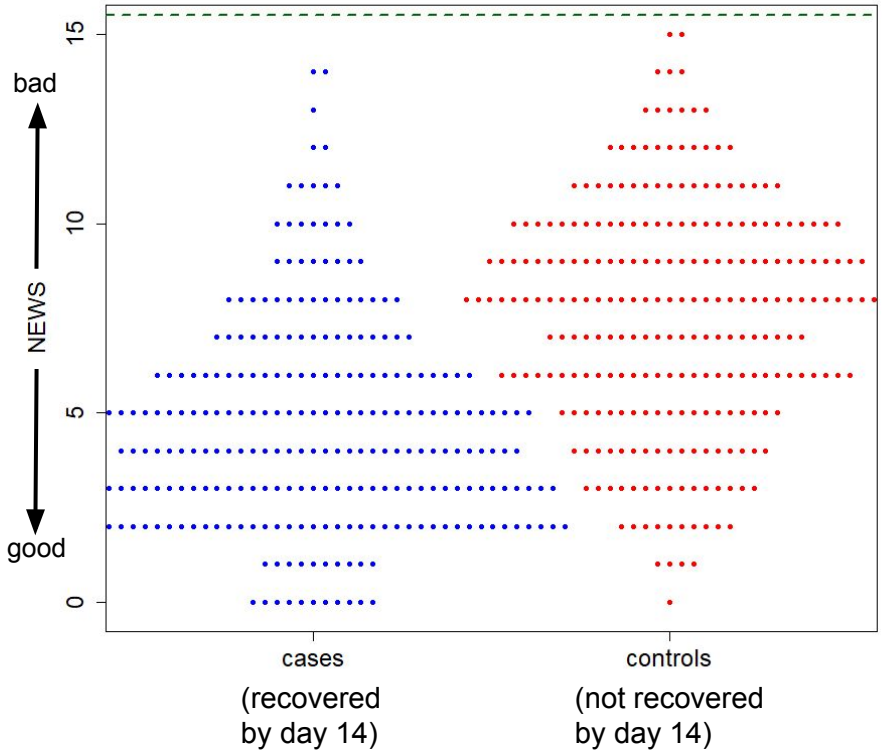


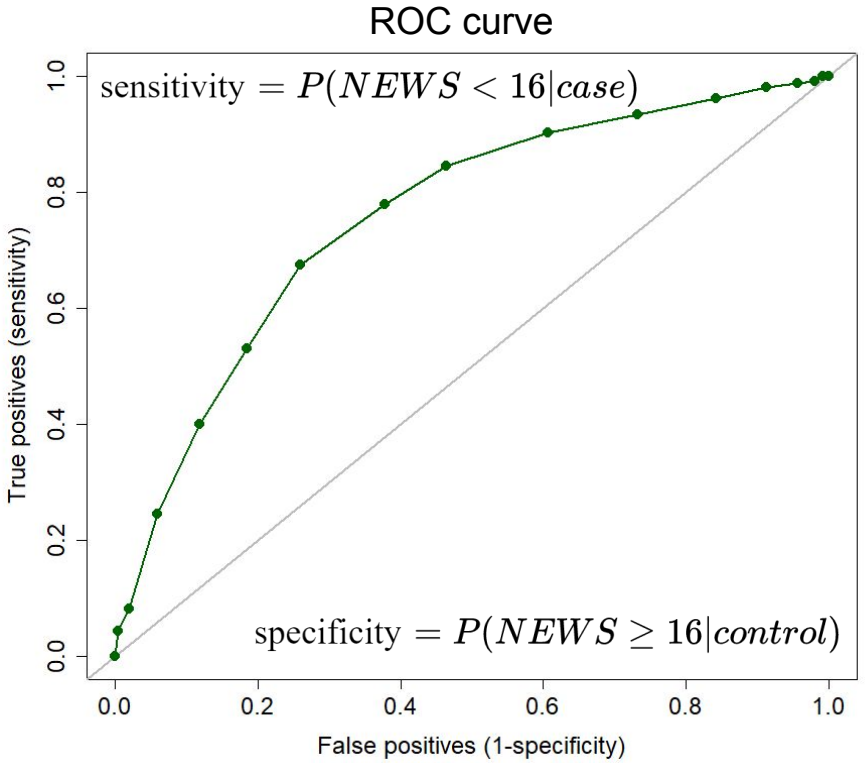
Review of ROC curves and AUC for binary outcomes



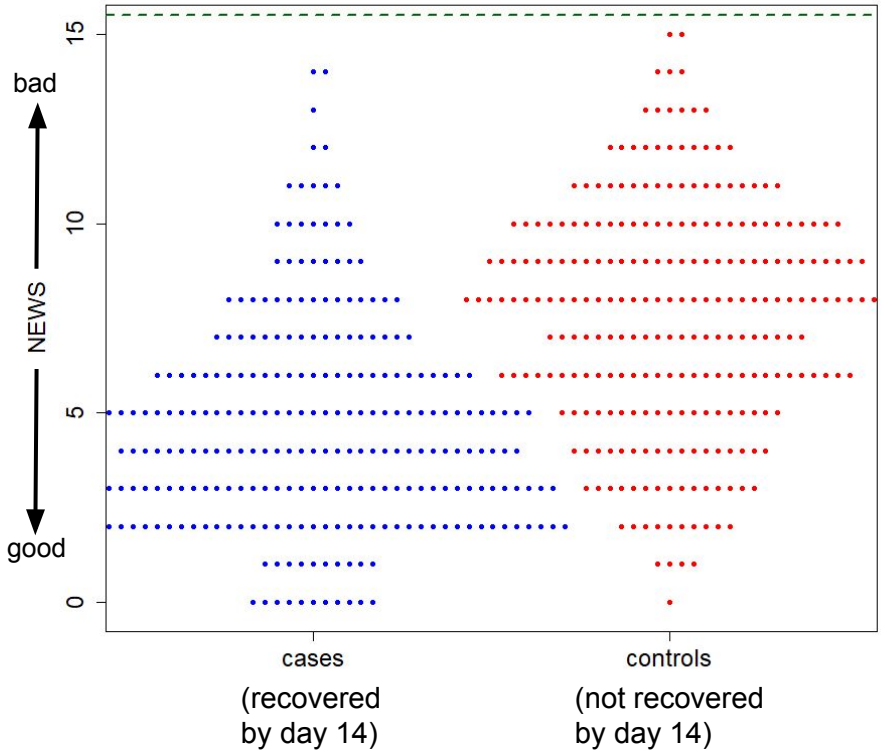


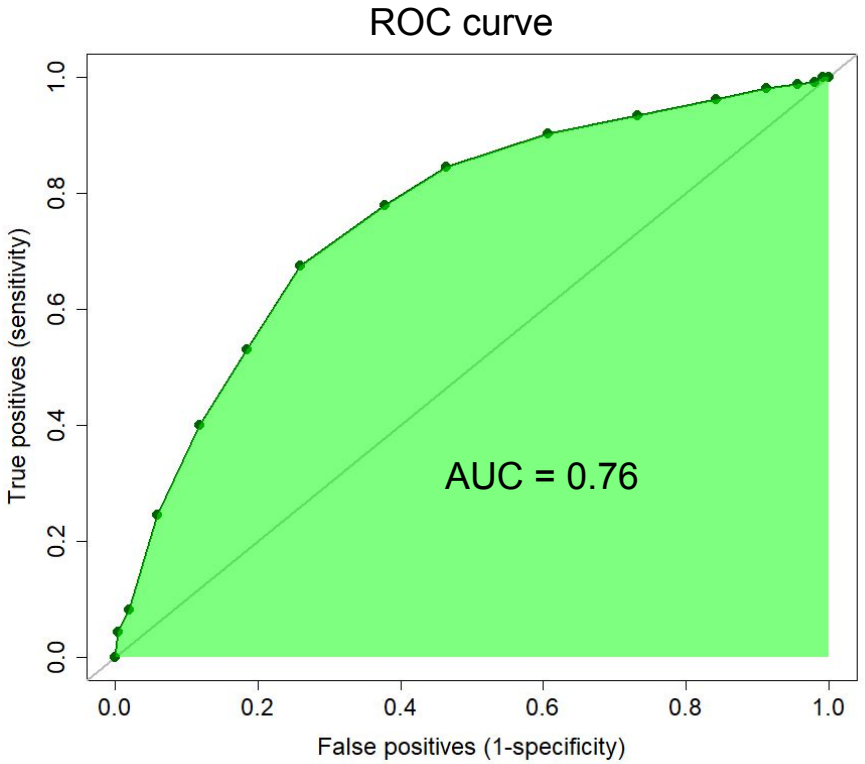
Review of ROC curves and AUC for binary outcomes





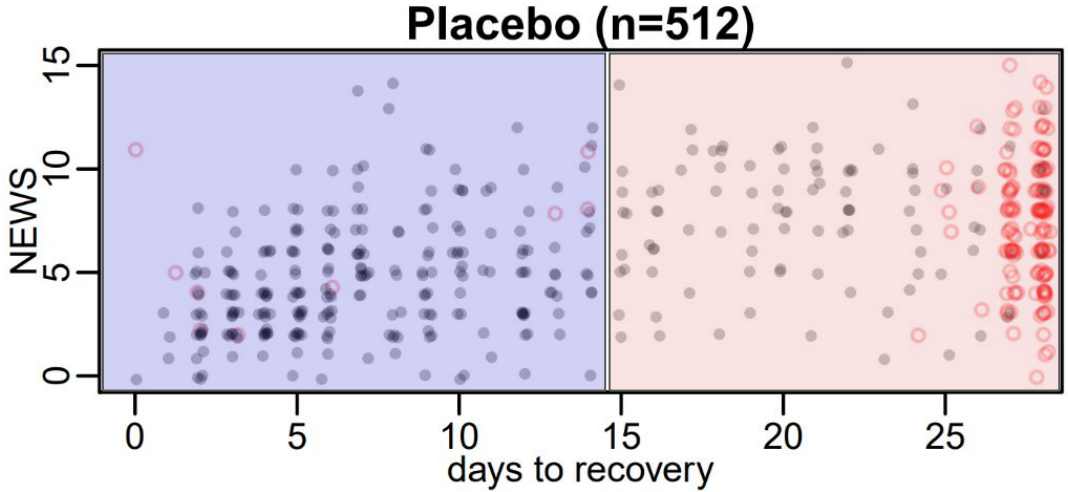
Review of ROC curves and AUC for binary outcomes





Cumulative/Dynamic ROC curves and AUC

	Placebo C-Index	Remdesivir C-Index	Placebo 14-d AUC
Risk score for recovery end point			
NEWS	0.68 (0.65–0.71)	0.69 (0.67–0.72)	0.76 (0.72–0.81)
NEWS+age	0.70 (0.67–0.72)	0.71 (0.69–0.74)	0.78 (0.74–0.82)



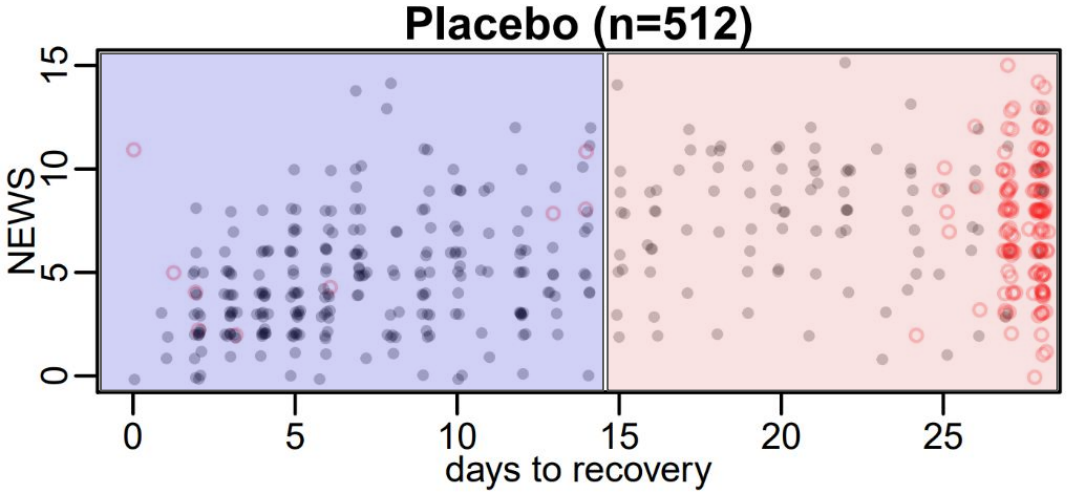
$$AUC^{C/D}(14) = P(NEWS_i < NEWS_j | T_i \leq 14, T_j > 14)$$

$$\widehat{AUC}^{C/D}(14) = 0.76$$

- At day 14, when comparing someone who had already recovered to someone who still has not, we estimate a 76% chance that the recovered person will have a lower NEWS score.
- Presented with someone from the blue box and someone from the red box, NEWS will correctly order them 76% of the time.
- A dot from the blue box (cases) has a 76% chance of being lower than a dot from the red box (controls).

Cumulative/Dynamic ROC curves and AUC

	Placebo C-Index	Remdesivir C-Index	Placebo 14-d AUC
Risk score for recovery end point			
NEWS	0.68 (0.65–0.71)	0.69 (0.67–0.72)	0.76 (0.72–0.81)
NEWS+age	0.70 (0.67–0.72)	0.71 (0.69–0.74)	0.78 (0.74–0.82)

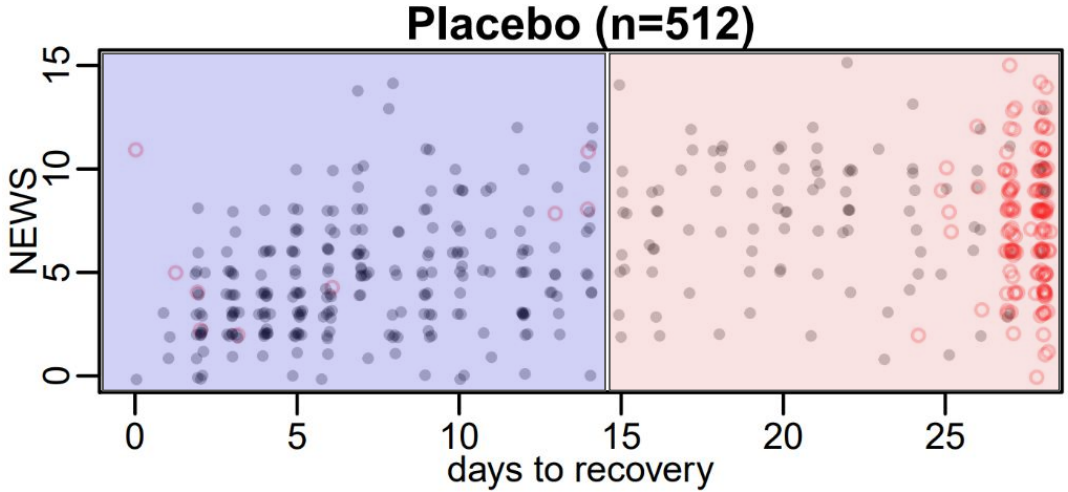


Once dichotomization time is chosen (here, day 14), the cumulative/dynamic AUC is conceptually very similar to the usual ROC curves for binary outcomes. Key here is graceful handling of censoring.

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, *56*(2), 337-344.

Cumulative/Dynamic ROC curves and AUC

	Placebo C-Index	Remdesivir C-Index	Placebo 14-d AUC
Risk score for recovery end point			
NEWS	0.68 (0.65–0.71)	0.69 (0.67–0.72)	0.76 (0.72–0.81)
NEWS+age	0.70 (0.67–0.72)	0.71 (0.69–0.74)	0.78 (0.74–0.82)



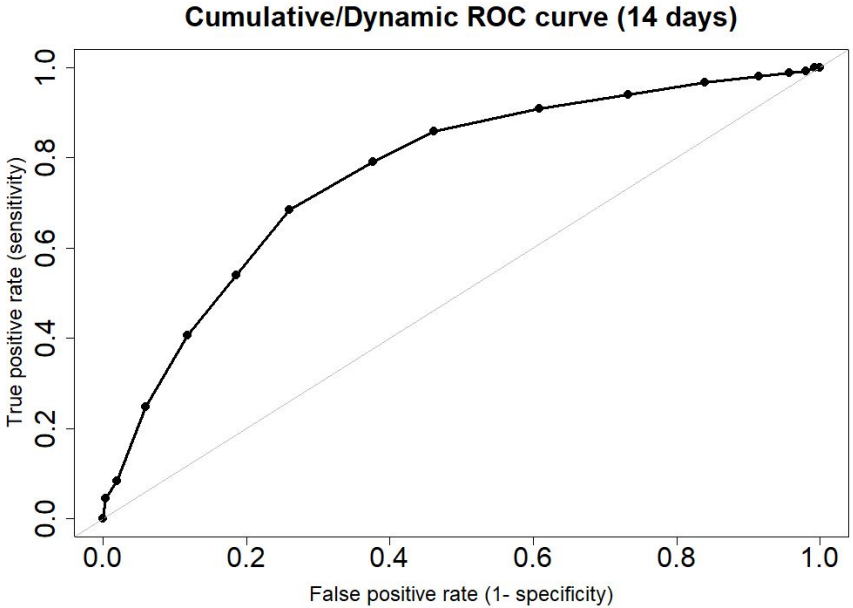
Nonparametric estimation: `survivalROC::survivalROC()` package

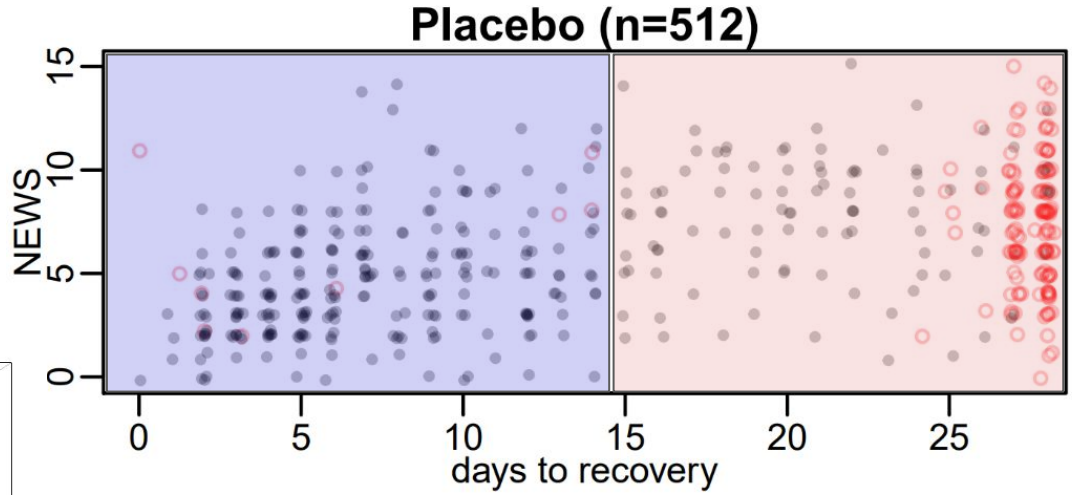
```
with(acttp, survivalROC(Stime=time.r, status=status.r,
marker=-news, predict.time=14, method="KM"))
```

Also consider `method="NNE"` - see below reference for details on why

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, *56*(2), 337-344.

Cumulative/Dynamic ROC curves and AUC



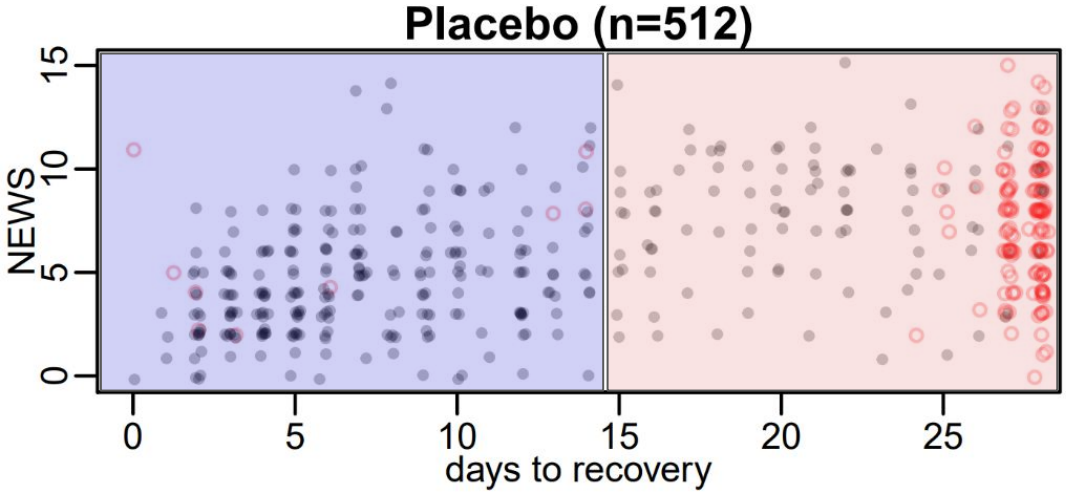


	Placebo C-Index	Remdesivir C-Index	Placebo 14-d AUC
Risk score for recovery end point			
NEWS	0.68 (0.65–0.71)	0.69 (0.67–0.72)	0.76 (0.72–0.81)
NEWS+age	0.70 (0.67–0.72)	0.71 (0.69–0.74)	0.78 (0.74–0.82)

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, *56*(2), 337-344.

Cumulative/Dynamic ROC curves and AUC

	Placebo C-Index	Remdesivir C-Index	Placebo 14-d AUC
Risk score for recovery end point			
NEWS	0.68 (0.65–0.71)	0.69 (0.67–0.72)	0.76 (0.72–0.81)
NEWS+age	0.70 (0.67–0.72)	0.71 (0.69–0.74)	0.78 (0.74–0.82)

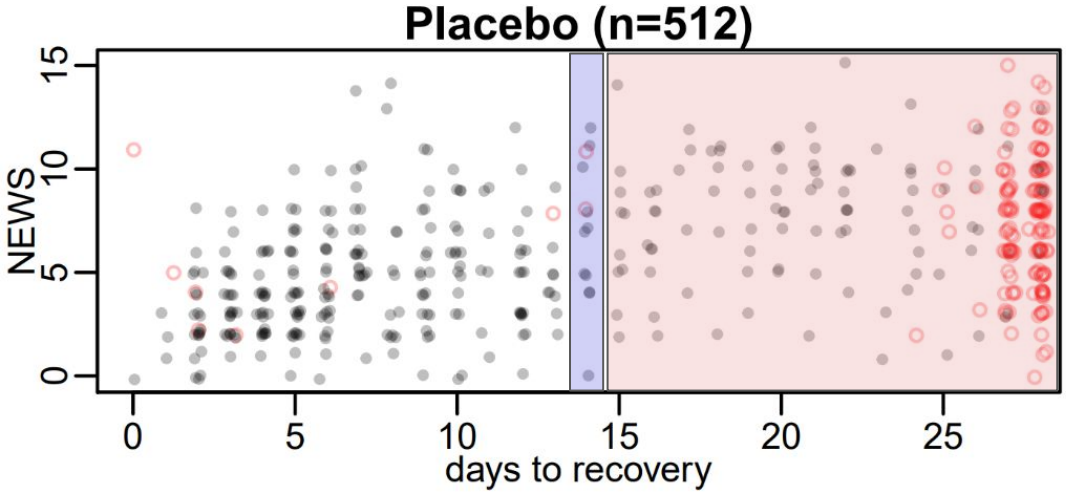


Exercise 2: construct quantile bootstrap 95% confidence intervals for AUC^C/D(14) using `survivalROC::survivalROC()`

Note: in this exercise, for `survivalROC()` options, use `method="KM"`; may also need to use negative NEWS (`marker=-news`) - why?

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, *56*(2), 337-344.

Incident/Dynamic ROC curves (not used in CCX paper)

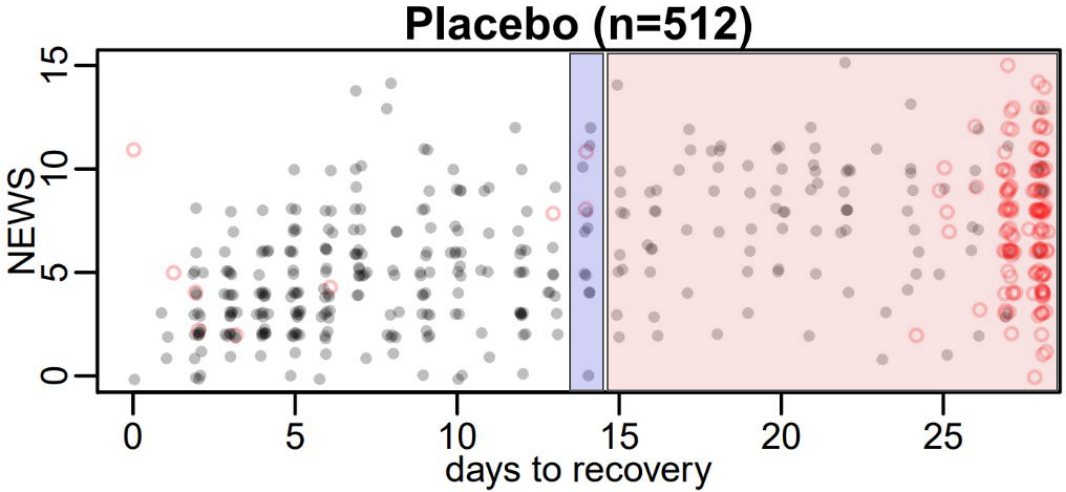


$$AUC^{I/D}(14) = P(NEWS_i < NEWS_j | T_i = 14, T_j > 14)$$

$$\widehat{AUC}^{I/D}(14) = 0.61$$

- When comparing someone who recovered **at day 14** to someone who still has not, we estimate a 61% chance that the recovered person will have a lower NEWS score.
- Presented with someone from the blue line and someone from the red box, NEWS will correctly order them 61% of the time.
- A dot from the blue line (cases) has a 61% chance of being lower than a dot from the red box (controls).

Incident/Dynamic ROC curves (not used in CCX paper)



$$AUC^{I/D}(14) = P(NEWS_i < NEWS_j | T_i = 14, T_j > 14)$$

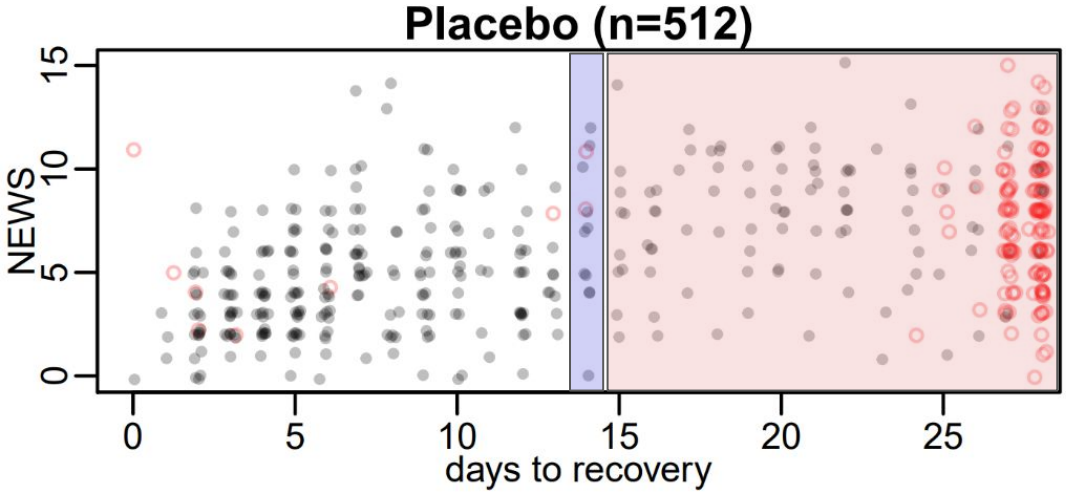
$$\widehat{AUC}^{I/D}(14) = 0.61$$

Unlike the previous AUC^C/D, this measure is more useful when evaluated at all (or a meaningful subset) times.

Quantifies how performance of NEWS, measured at baseline, **varies over time.**

We won't cover this but is easily adapted for **time-varying covariates**.

Incident/Dynamic ROC curves (not used in CCX paper)



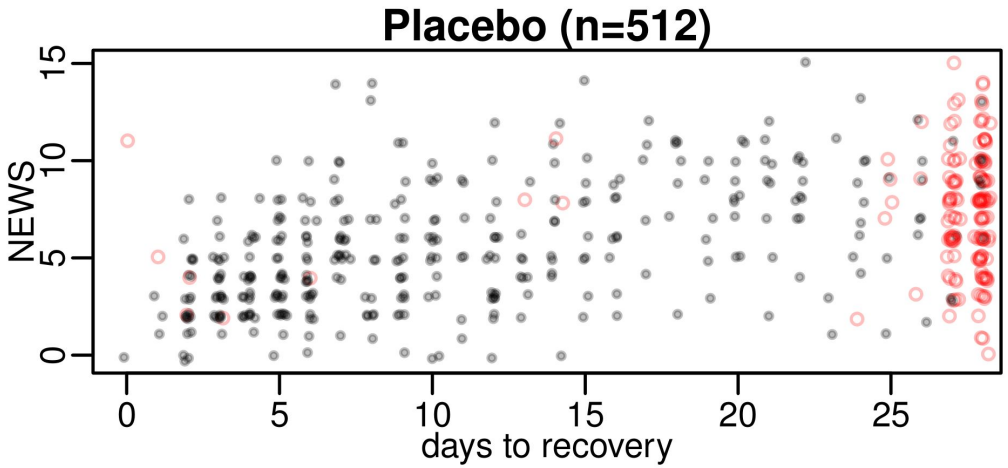
Estimation

- R (semiparametric [1]): `risksetROC::risksetROC()`
- **R (nonparametric [2]): Weighted Mean Rank estimator** (http://faculty.washington.edu/abansal/software.html)

1. Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, *61*(1), 92-105.
2. Saha-Chaudhuri, P., & Heagerty, P. J. (2013). Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics*, *14*(1), 42-59.

Weighted Mean Rank estimator of AUC^I/D

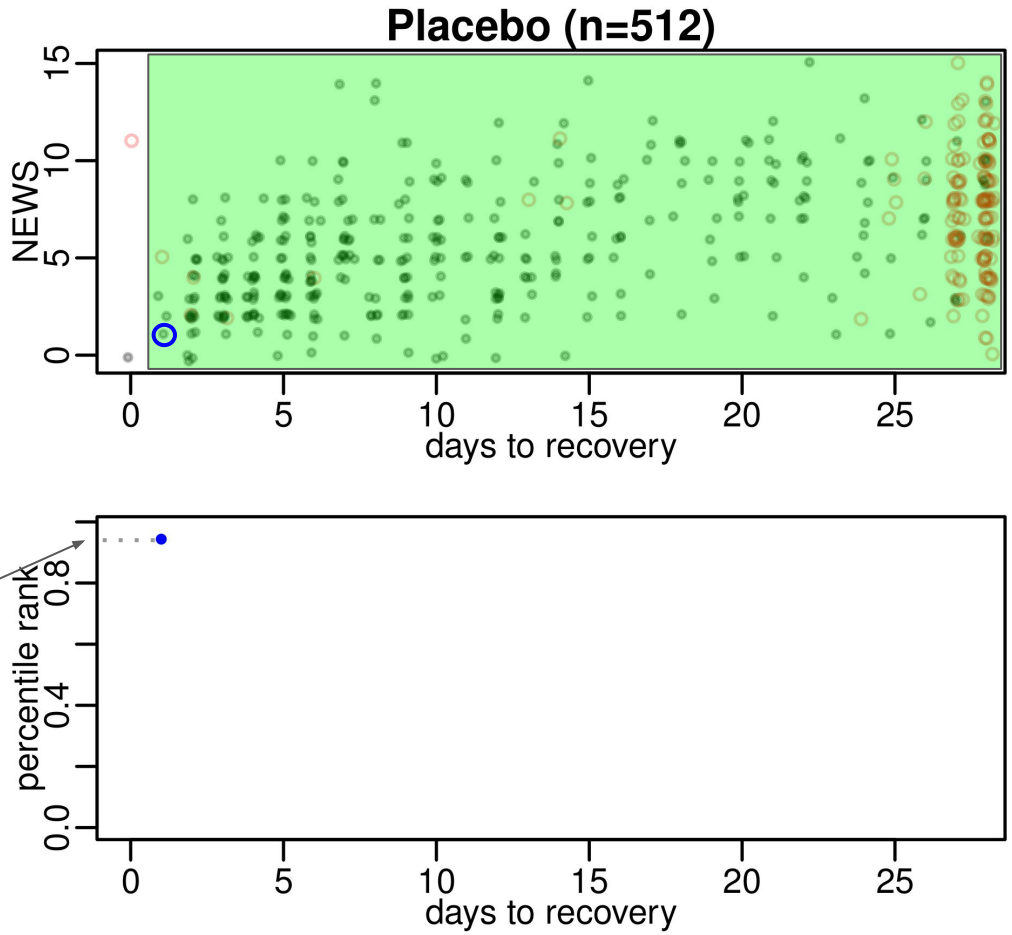
1. Transform all events (just black dots) from NEWS scale to a “percentile rank” scale.
2. Locally smooth these new percentile ranks over time to get the WMR estimator



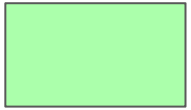
Weighted Mean Rank estimator of AUC^I/D

holds the risk set of

The NEWS score of is lower than 94% of the others in the risk set

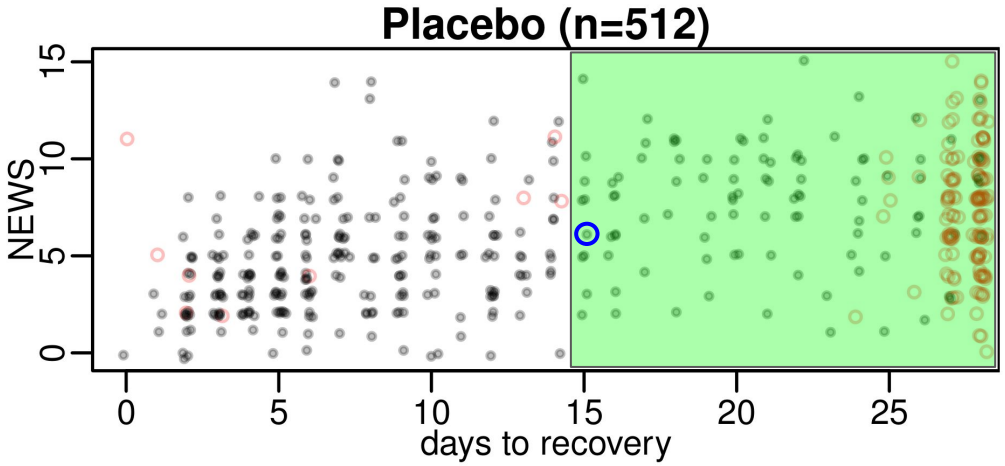


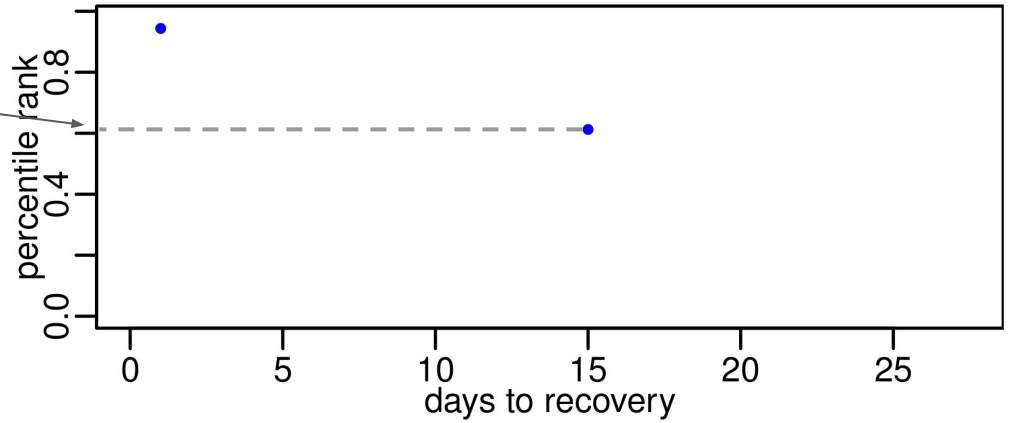
Weighted Mean Rank estimator of AUC^I/D



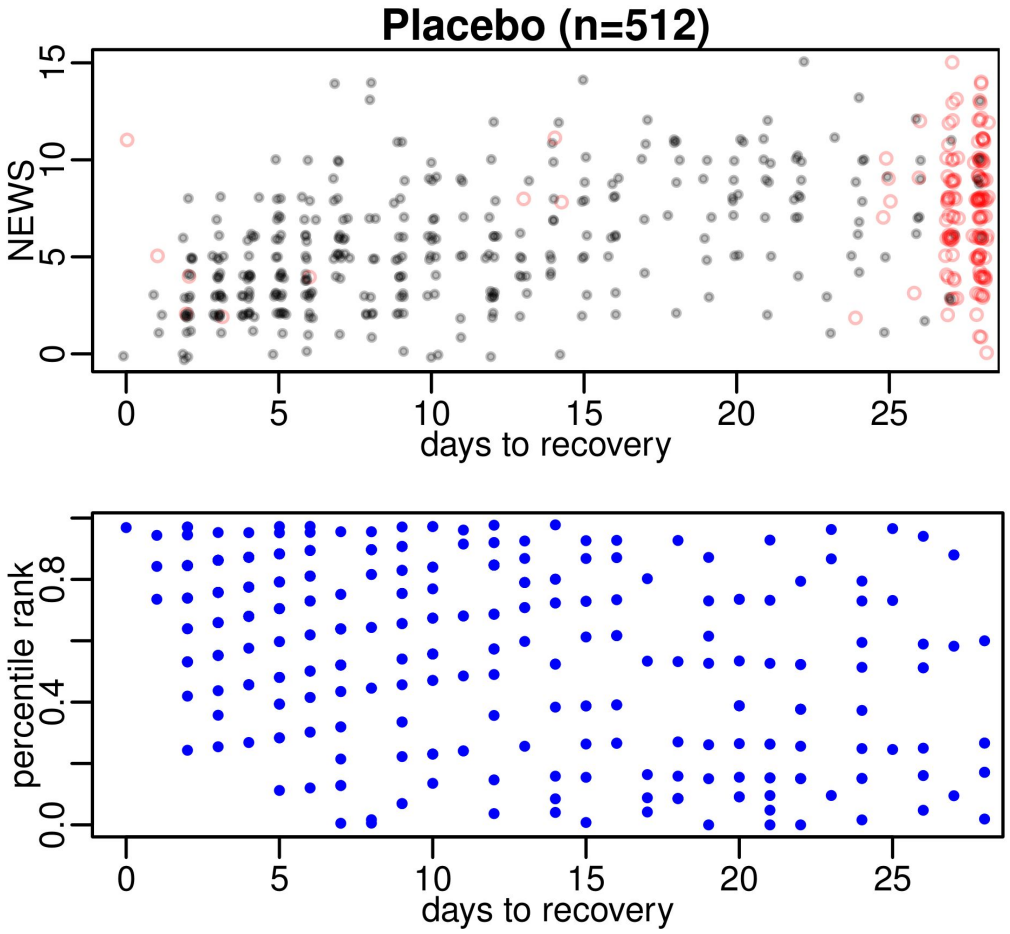
holds the risk set of ○

The NEWS score of ○ is lower than 61% of the others in the risk set





Weighted Mean Rank estimator of AUC^I/D



Weighted Mean Rank estimator of AUC^I/D

Recall:

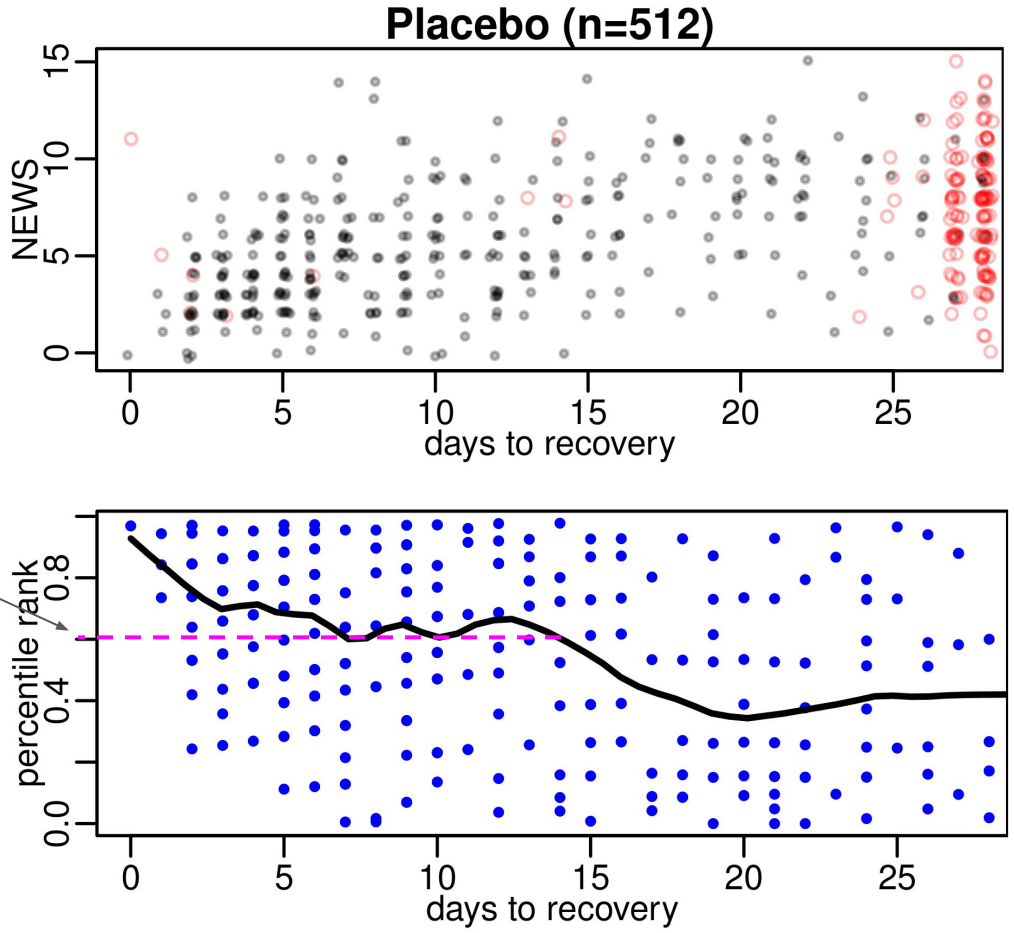
When comparing someone who recovered **at day 14** to someone who still has not, we estimate a 61% chance that the recovered person will have a lower NEWS score.

$$\widehat{AUC}^{I/D}(14) = 0.61$$

But interpreting entire line probably more meaningful.

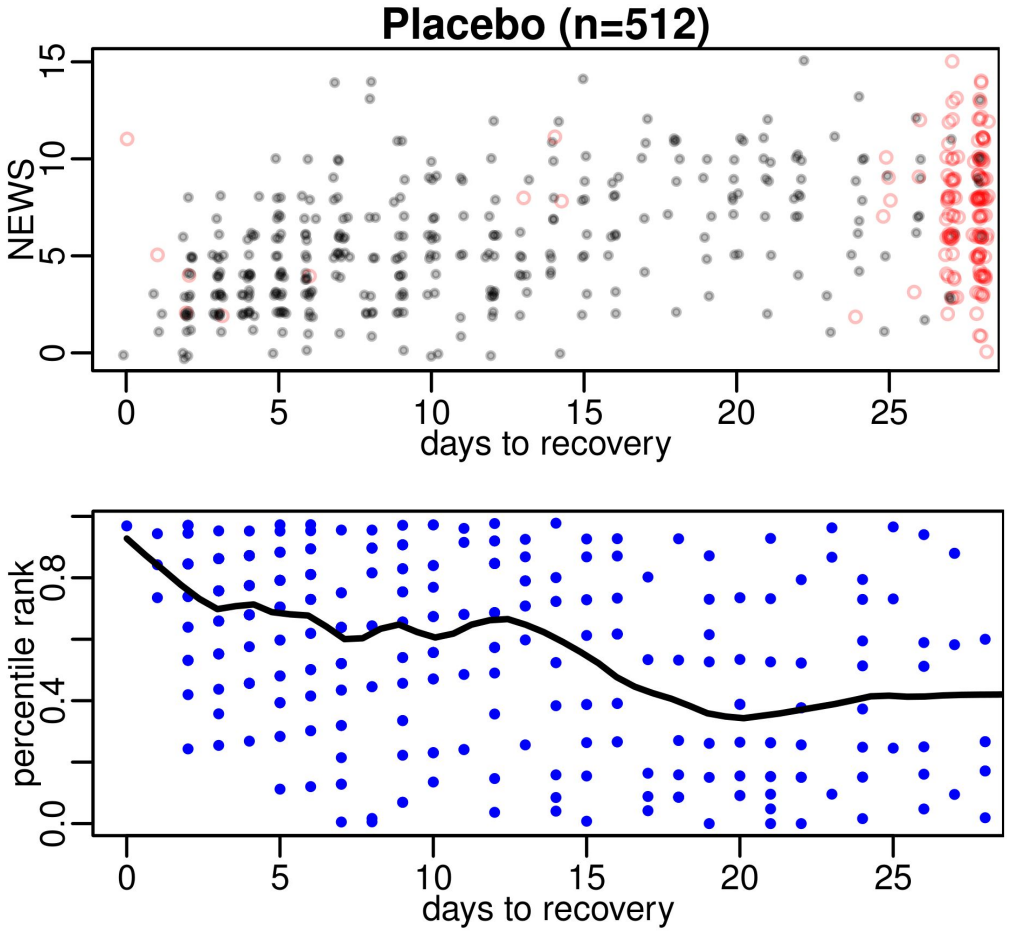
Suggests NEWS is most useful within the first few days. After two weeks it loses most of its ability to prognosticate recovery.

In practice, often see this “decaying” accuracy, which lines up with intuition.



Weighted Mean Rank estimator of AUC^I/D

Exercise 3: “manually” code weighted mean rank estimator of AUC^I/D. Use placebo arm, NEWS, and days to recovery. Refer to last sequence of slides. See Saha-Chaudhuri & Heagerty (2013) if necessary.



Saha-Chaudhuri, P., & Heagerty, P. J. (2013). Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics*, *14*(1), 42-59.

Discrimination measure for survival models: summary

C-index

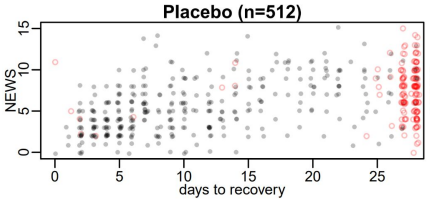
- Global, single number summary overall performance
- R: Hmisc::rcorr.cens()
- Python: scikit-survival

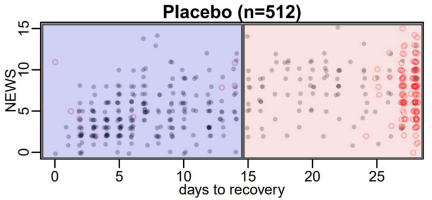
Cumulative/Dynamic ROC curves and AUC

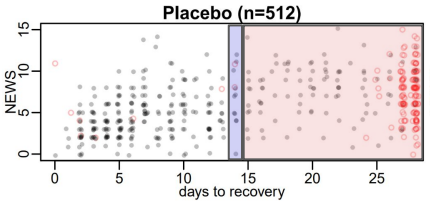
- Dichotomize time-to-event outcome at particular time
- Estimate ROC and AUC while incorporating censoring
- Typically will just calculate for a single or handful times of interest
- R: survivalROC::survivalROC()
- Python: scikit-survival

Incident/Dynamic ROC curves and AUC

- Time-varying prognostic performance (e.g. how quickly does my biomarker's prognostic value decay)
- Handles time-varying covariates
- R (semiparametric): risksetROC::risksetROC()
- R (nonparametric, Aasthaa Bansal): http://faculty.washington.edu/abansal/software.html
- Python (nonparametric, Diego Seira): https://github.com/dseira95/Weighted_Mean_Rank
- Julia (nonparametric, anyone?): Email me! liangcj@nih.gov







See Bansal & Heagerty (2018) for more details

Bansal, A., & Heagerty, P. J. (2018). A Tutorial on Evaluating the Time-Varying Discrimination Accuracy of Survival Models Used in Dynamic Decision Making. Medical Decision Making, 38(8), 904–916. https://doi.org/10.1177/0272989X18801312

Discrimination measure for survival models: summary

Not covered:

Other extensions [1]: other meaningful ways to define cases and controls

Risk-based alternatives to ROC/AUC [2]: ties create ambiguities; risk-based measures less prone AND are arguably more flexible measures of discrimination

Competing risks setting [3-5]

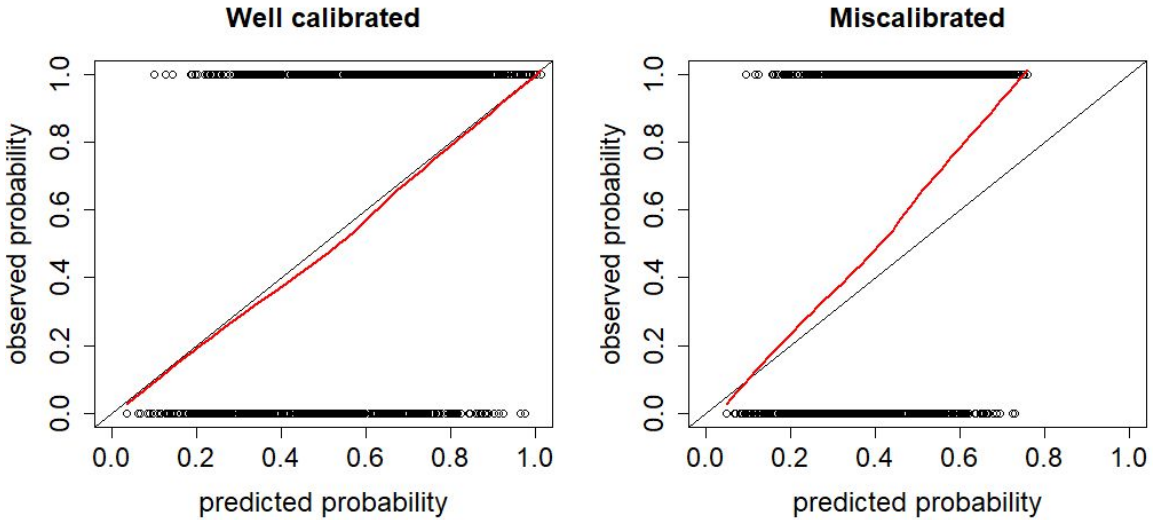
1. Zheng, Y., & Heagerty, P. J. (2007). Prospective accuracy for longitudinal markers. *Biometrics*, *63*(2), 332-341.
2. Liang, C. J., & Heagerty, P. J. (2017). A risk-based measure of time-varying prognostic discrimination for survival models. Biometrics, 73(3), 725-734.
3. Saha, P., & Heagerty, P. J. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*, *66*(4), 999-1011.
4. Wolbers, M., Blanche, P., Koller, M. T., Witteman, J. C., & Gerds, T. A. (2014). Concordance for prognostic models with competing risks. *Biostatistics*, *15*(3), 526-539.
5. van Geloven, N., Giardiello, D., Bonneville, E. F., Teece, L., Ramspek, C. L., van Smeden, M., ... & Steyerberg, E. (2022). Validation of prediction models in the presence of competing risks: a guide through modern methods. *bmj*, *377*.

Calibration (not assessed in CCX paper)

Consider again the binary outcome scenario. Suppose we have a model that outputs predictions between 0 and 1 (e.g. logistic regression, random forest).

Calibration: how close are predicted risks to true risks?

We never know the true risk but we can estimate it (red lines) for different buckets of predicted probabilities.



Calibration

Table 2. A hierarchy of calibration levels for risk prediction models

Level	Definition	Assessment
Mean	Observed event rate equals average predicted risk; "calibration-in-the-large"	*Compare event rate with average predicted risk; *Evaluate $a \mid b_L = 1$ (with 1 df test $a \mid b_L = 1 = 0$)
Weak	No systematic overfitting or underfitting and/or overestimation or underestimation of risks; "logistic calibration"	Logistic calibration analysis to evaluate $a \mid b_L = 1$ and b_L (with Cox recalibration test: a 2 df test of the null hypothesis that $a \mid b_L = 1 = 0$ and $b_L = 1$)
Moderate	Predicted risks correspond to observed event rates	Calibration plot (eg, using loess or splines), or analysis by grouped predictions (including Hosmer-Lemeshow test)
Strong	Predicted risks correspond to observed event rates for each and every covariate pattern	Scatter plot of predicted risk and observed event rate per covariate pattern; impossible when continuous predictors are involved

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology*, *74*, 167-176.

Calibration

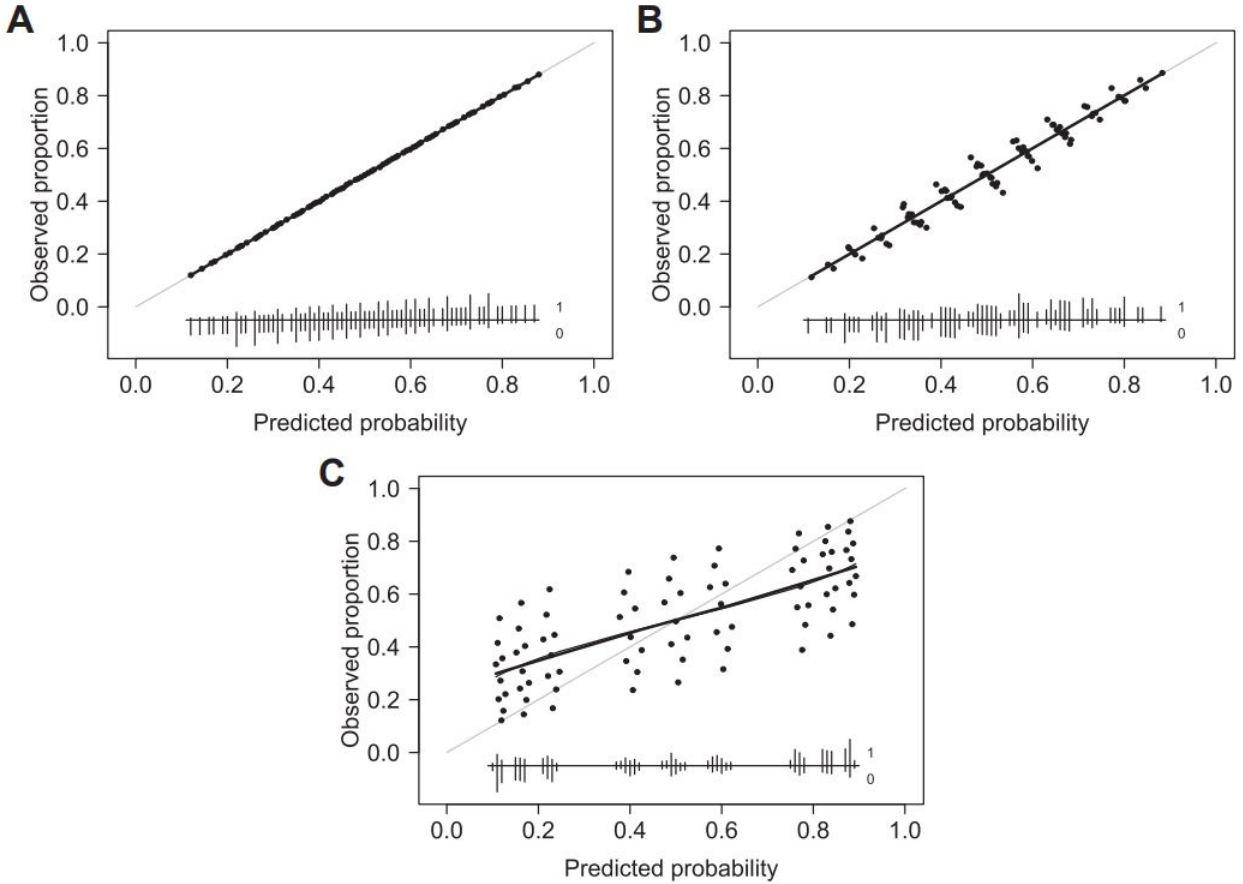
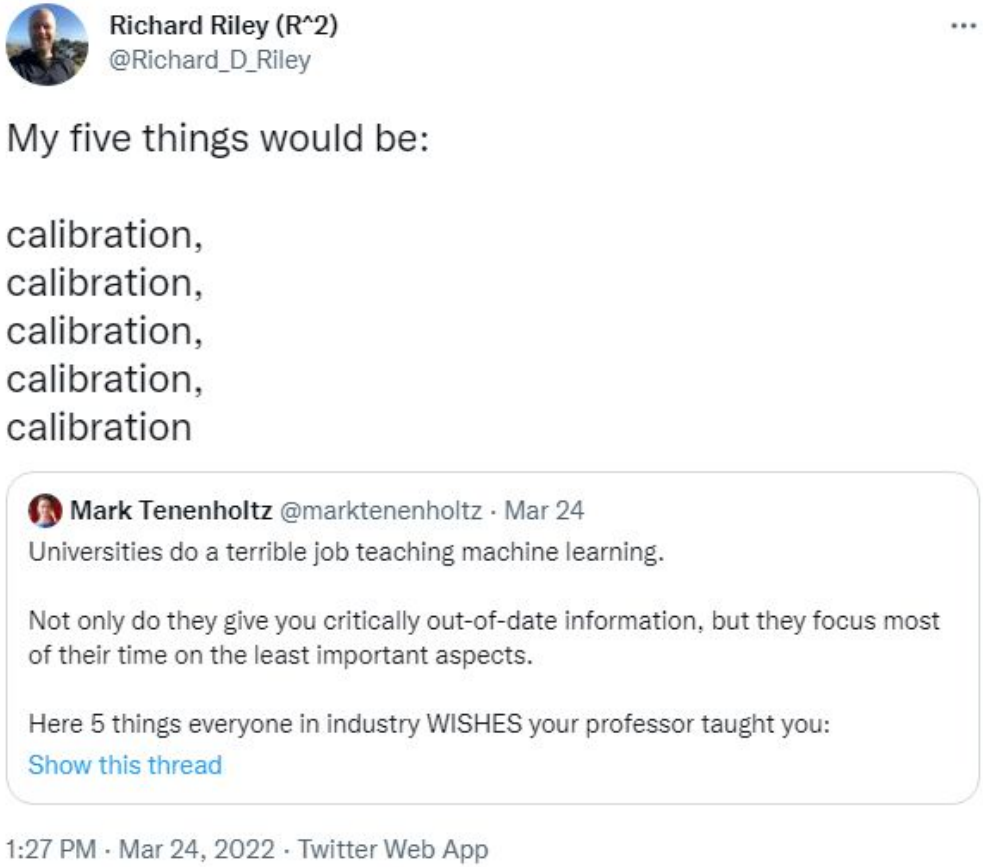


Fig. 3. Calibration plots illustrating (A) strong calibration, (B) moderate but not strong calibration, and (C) miscalibration.

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology*, *74*, 167-176.

Calibration



https://twitter.com/richard_d_riley/status/1507046379105013760

Calibration

OPINION Open Access

Calibration: the Achilles heel of predictive analytics

Ben Van Calster[1,2,6*], David J. McLernon[3,6], Maarten van Smeden[2,4,6], Laure Wynants[1,5], Ewout W. Steyerberg[2,6]
On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative[6]

Abstract

Background: The assessment of calibration performance of risk prediction models based on regression or more flexible machine learning algorithms receives little attention.

Main text: Herein, we argue that this needs to change immediately because poorly calibrated algorithms can be misleading and potentially harmful for clinical decision-making. We summarize how to avoid poor calibration at algorithm development and how to assess calibration at algorithm validation, emphasizing balance between model complexity and the available sample size. At external validation, calibration curves require sufficiently large samples. Algorithm updating should be considered for appropriate support of clinical practice.

Conclusion: Efforts are required to avoid poor calibration when developing prediction models, to evaluate calibration when validating models, and to update models when indicated. The ultimate aim is to optimize the utility of predictive analytics for shared decision-making and patient counseling.

Keywords: Calibration, Risk prediction models, Predictive analytics, Overfitting, Heterogeneity, Model performance

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC medicine*, *17*(1), 1-7.

Calibration

See Riley et al. (2019 Statistics in Medicine) and related series of articles for thoughtful recommendations on sample size estimation for prognostic models

https://onlinelibrary.wiley.com/doi/full/10.1002/sim.9025

If you REALLY want to get in the weeds on calibration and proper scoring rules, see Gneiting and Raftery (2007 JASA).

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. *Statistics in medicine*, *38*(7), 1276-1296.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359-378.

Calibration

Exercise 4: NEWS is on the 0-20 scale. Should we assess calibration? If we want to, how could we do it? How would we account for the outcome being time-to-event and not binary?

Calibration and Discrimination

Simple weather example: Seattle tends to get 150 rainy days each year (41%).

If my model predicted a 41% chance of rain every day in Seattle, the model would be **“properly calibrated”** but **poor at discrimination.**

If every rainy day my model assigned a 10% chance of rain while every non-rainy was assigned a 5% chance, that’s **perfect discrimination but poor calibration**.

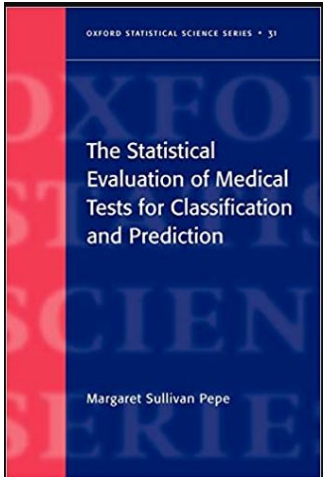
Additional resources

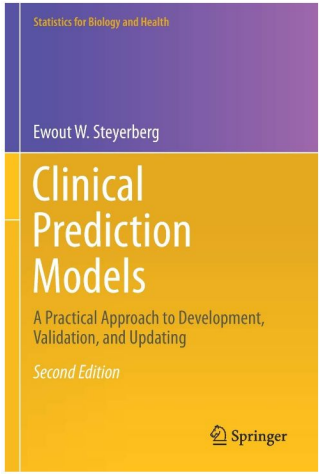
Margaret Pepe - The Statistical Evaluation of Medical Tests for Classification and Prediction (2004)

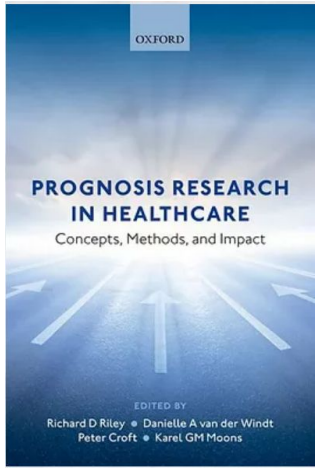
Ewout Steyerberg - Clinical Prediction Models (2020)

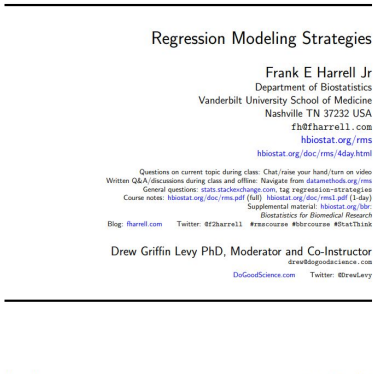
Richard Riley et al. - Prognosis Research in Healthcare (2019)

Frank Harrell - Regression Modeling Strategies (2022)









Exercises

Exercise 1 (c-index estimation): slide 43

Exercise 2 (boostrap CI for cumulative/dynamic AUC): slide 62

Exercise 3 (weighted mean rank estimation): slide 71

Exercise 4 (calibration): slide 80