

Early Cardiac Testing after Emergency-Department ECG: An Observational Target-Trial Emulation

Bjørn-Jostein Singstad

b.j.singstad@fys.uio.no

University of South-Eastern Norwa

Research Article

Keywords: Causal inference, Target-trial emulation, Doubly robust estimation, Machine 42 learning, Confounding by indication, Emergency department, Electrocardiography, Immortal 43 time bias

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10713210/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Early Cardiac Testing after Emergency-Department ECG: An Observational Target-Trial Emulation

Bjørn-Jostein Singstad^{1,2,3,4,*}

¹Institute of Clinical Medicine, University of Oslo, Oslo, Norway

²Medical Technology and E-Health, Akershus University Hospital, Lørenskog, Norway

³Department of Radiology, Vestfold Hospital Trust, Tønsberg, Norway

⁴Faculty of Technology, Natural Sciences and Maritime Sciences, University of South-Eastern Norway, Horten, Norway

*Corresponding author: Bjørn-Jostein Singstad, b.j.singstad@fys.uio.no

Abstract

Background: Because clinicians selectively request cardiac testing for emergency department (ED) patients judged to be at higher risk, a naive comparison of tested and untested patients confounds the effect of testing with its indication. We conducted an observational target-trial emulation to quantify how far adjustment for recorded covariates can resolve this confounding.

Methods: We linked ED encounter, patient, and ECG metadata and emulated a two-strategy trial: cardiac testing (stress testing or catheterisation) within 10 days of an ED visit versus none, evaluating the 30-day risk of a composite of positive cardiac troponin or death. The estimand was the marginal risk difference, estimated by five-fold cross-fitted augmented inverse-probability weighting (AIPW) with logistic and boosted-tree nuisance models. We also report patient-clustered inference, a one-encounter-per-patient restriction, a grace-period sensitivity analysis, a check on a covariate of doubtful baseline status, and an E-value for unmeasured confounding.

Results: The analytic cohort comprised 51,158 of 71,460 encounters (20,302 excluded), of whom 4,609 (9.0%) received testing and 4,747 (9.3%) had the outcome. The crude risk difference was 22.1 percentage points (pp). Cross-fitted AIPW gave 5.9 pp (95% CI 3.9-7.9) with logistic nuisance models and 5.1 pp (3.6-6.6) with boosted trees; patient clustering left the interval essentially unchanged. The estimate was less stable under two checks: it fell to 0.2 pp (95% CI -1.3 -1.7) when the exposure-ascertainment window shortened from 10 to 1 day, rising steadily back to 5.9 pp as the window widened, and rose to 11.1 pp (9.2-13.0) when a same-day AMI-diagnosis covariate (present in 42.8% of tested but 0.0% of untested encounters) was dropped from the adjustment set. The E-value for the main estimate was 2.66 (2.19 at the confidence limit closer to the null).

Conclusions: Adjustment substantially reduces, but does not eliminate, the crude contrast, and the residual estimate is not robust to the length of the exposure-ascertainment window or to the baseline status of covariates measured close to the treatment decision. The project illustrates both a causal machine-learning workflow and why a well-specified estimand does not, by itself, convert observational care data into treatment-effect evidence.

Keywords: Causal inference; Target-trial emulation; Doubly robust estimation; Machine learning; Confounding by indication; Emergency department; Electrocardiography; Immortal time bias

1 Background

ECGs and troponin measurements are central to the emergency department (ED) assessment of possible acute coronary syndrome. Their interpretation also influences whether clinicians pursue stress testing or catheterization. A naive comparison of outcomes among tested and untested patients is consequently misleading: testing is selectively given to patients whose presentation is thought to be riskier. This is the classic problem of confounding by indication

[1].

Here we use the available actionable variable, cardiac testing within 10 days, as the exposure, and use ECG and clinical measurements as pre-treatment predictors. The key contribution lies in providing a clear causal framework, rather than asserting that testing itself produces the observed difference in outcomes.

2 Methods

2.1 Causal question and target-trial specification

The dataset is derived from Nightingale [2, 3] and comprises all ED visits to Brigham and Women’s Hospital from 2010 to 2015. Time zero is the start of the encounter. Let $A = 1$ denote any recorded cardiac test within 10 days and $A = 0$ denote no such test. In the dataset, a test within 10 days is defined as either a stress test or a catheterization performed within that timeframe and explicitly excludes troponin measurements. This is important because the outcome includes a troponin component, and we want to avoid linking the exposure and outcome through the same measurement. Confirming that troponin is not part of the exposure ensures that our main concern is confounding by indication rather than a definitional overlap between exposure and outcome. Let $Y = 1$ denote the composite of positive cardiac troponin or death within 30 days. The estimand is the marginal 30-day risk difference

$$\psi = \mathbb{E}[Y^{(1)} - Y^{(0)}],$$

where $Y^{(a)}$ denotes the outcome that would have been observed if the encounter had proceeded according to strategy a . This is the effect of an idealized strategy that ensures testing within 10 days versus no recorded testing; it is not the effect of one particular test modality and does not distinguish a stress test from catheterization.

Table 1 maps this design onto the standard target-trial protocol components [4], and

Table 1: Target-trial protocol components and their emulation in this study.

Component	Emulation
Eligibility	ED encounters with <code>exclude_modeling = FALSE</code> and a linked first ECG (51,158 of 71,460 repository encounters; Figure 2).
Treatment strategies	(1) Receive cardiac testing (stress test or catheterisation) within 10 days of time zero; (2) receive no such testing within 10 days.
Treatment assignment	Not randomised. Observed testing status is used directly and adjusted for X under conditional exchangeability; no attempt is made to construct treatment arms by cloning.
Time zero	Start of the ED encounter.
Follow-up	30 days from time zero for the outcome; up to 10 days from time zero for exposure ascertainment, which overlaps the outcome window (see immortal-time discussion below).
Outcome	Positive cardiac troponin or death within 30 days.
Causal contrast	Intention-to-treat marginal risk difference $\psi = \mathbb{E}[Y^{(1)} - Y^{(0)}]$.
Analysis plan	Cross-fitted AIPW with logistic and boosted-tree nuisance models, sensitivity analyses for exposure-window length, patient clustering, and a covariate of doubtful baseline status.

Protocol components follow the target-trial framework of Hernán and Robins [4].

Figure 1 provides the working causal graph. Measured baseline variables X include demographics, area-level income-bin probabilities, same-day troponin group/value, first linked ECG intervals, diagnostic flags and same-day AMI status. Latent clinical severity U represents symptoms, hemodynamics, comorbidity, prior coronary disease, and clinician judgment. U is a common cause of testing and outcomes. Importantly, variables such as discharge disposition and revascularization can occur after the treatment decision, so they are not adjusted for. We also treat same-day AMI status as a weaker instance of this issue: it appears in 42.8% of tested encounters but in 0.0% of untested encounters (Table 2), a pattern more in line with testing being triggered by an AMI diagnosis (or determined concurrently with it) than with a variable determined prior to the treatment decision. The results indicate the extent to which this single covariate contributes to the adjustment.

Assuming consistency, conditional exchangeability $(Y^{(0)}, Y^{(1)}) \perp A \mid X$, positivity $0 <$

$P(A = 1 \mid X) < 1$, accurate linkage/measurement, and independent encounters (or an encounter-level analysis that is valid), the backdoor criterion allows identification of ψ via standardization [5, 6]. Consistency is also demanding here because "testing" combines different investigations and timings. Repeated encounters for one patient may further violate independent-observation inference. The results section reports both cluster-robust inference and a one-encounter-per-patient restriction as checks.

A further identification concern specific to this design is immortal time. Because A is established over a 10-day window following time zero, while Y is ascertained over the overlapping 30-day window from the same time zero, an encounter effectively has to remain eligible to be classified $A = 1$ for as long as 10 days into the outcome-accrual period. This is the standard immortal-time-bias structure that arises whenever a forward-looking exposure-ascertainment window is used to define a baseline-anchored contrast [4, 7, 8]. The dataset does not record day-level timing for the general composite outcome, so we cannot implement the usual clone-censor-weight correction; instead, Section 3 reports a grace-period sensitivity analysis that shortens the exposure-ascertainment window as an empirical test of this bias.

2.2 Data source and cohort derivation

The repository contains 71,460 ED encounters, 44,713 patients, 112,900 ECG records, and an ECG-encounter linkage file with 103,952 links. We merged encounters to patient demographics and selected the chronologically first linked ECG for each encounter. All 71,460 encounters have a linked ECG, so the exclusions below are the only source of cohort exclusion.

Figure 2 shows how the analytic cohort was derived, following STROBE-style cohort accounting [9]. This study is reported in line with the STROBE guideline for observational studies, adapted for a target-trial emulation design. Applying `exclude_modeling = FALSE` left 51,158 encounters, all with a linked first ECG. The repository provides three named exclusion flags (`excl_flag_c_int`, `excl_flag_chronic`, `excl_flag_death`); their union accounts for 17,981 of the 20,302 excluded encounters, with 2,321 excluded under

other `exclude_modeling` criteria not further documented in the repository. The flags overlap (an encounter can trigger more than one), so the exclusions in Figure 2 are flag-positive totals rather than a partition. The binary ischaemic-ECG flag (ST abnormality, ST depression/elevation, or T-wave inversion) was present in 22.9% of the analytic cohort; it is a covariate rather than the causal exposure.

2.3 Statistical analysis

For a descriptive benchmark, we report the crude difference in observed outcome risks, and Table 2 reports baseline characteristics by testing status. For the causal estimators, we fit the propensity score $e(X) = P(A = 1 \mid X)$ [10] and outcome regressions $m_a(X) = \mathbb{E}(Y \mid A = a, X)$ using five-fold cross-fitting. Our primary estimator is the augmented inverse-probability-weighted (AIPW) doubly robust estimator [11, 12]:

$$\hat{\psi} = \frac{1}{n} \sum_{i=1}^n \left[\hat{m}_1(X_i) - \hat{m}_0(X_i) + \frac{A_i \{Y_i - \hat{m}_1(X_i)\}}{\hat{e}(X_i)} - \frac{(1 - A_i) \{Y_i - \hat{m}_0(X_i)\}}{1 - \hat{e}(X_i)} \right].$$

Consistency holds if at least one of the nuisance components is estimated consistently, given the causal assumptions. We evaluate two specifications for the nuisance models: (i) penalized logistic regression models with median imputation, feature scaling, and one-hot encoding; and (ii) histogram-based gradient-boosted classifiers with one-hot encoding [13]. Propensity scores were truncated to the interval $[0.01, 0.99]$ [14]. Before truncation, 5.6% of the cross-fitted logistic propensity scores and 20.8% of the boosted-tree propensity scores lay outside this interval, suggesting that the boosted-tree model depends much more heavily on extrapolated or truncated weights. Wald intervals use the empirical standard deviation of the cross-fitted efficient-influence-function values, divided by $\sqrt{n}$. Additionally, this paper reports a version of this variance, clustering on patients [15], since 51,158 encounters come from fewer than 44,713 distinct patients. Cross-fitting mitigates overfitting bias that can arise when employing flexible nuisance estimators [16].

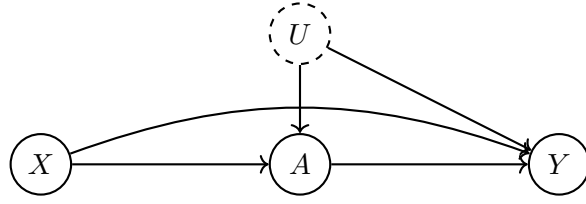


Figure 1: Working causal structure used for adjustment. X : recorded pre-treatment clinical, demographic, and ECG features. U (dashed, unobserved): unrecorded severity and clinician judgement, the principal threat to identification.

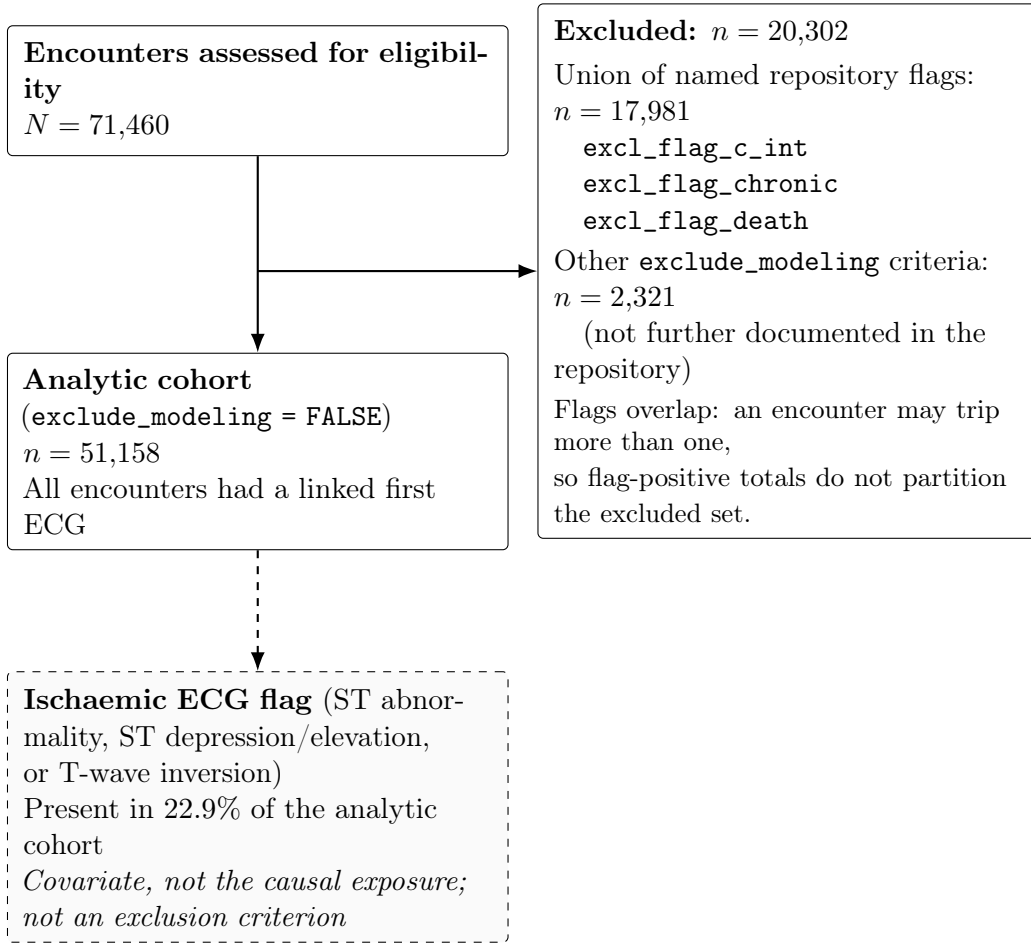


Figure 2: Derivation of the analytic cohort ($n = 51,158$ of 71,460 encounters). Exclusion counts are flag-positive totals; the three named flags overlap and do not partition the 20,302 excluded encounters. The ischaemic-ECG flag (dashed) is a covariate describing the analytic cohort, not an exclusion step.

We report four further checks on the primary (logistic) AIPW estimate. First, a one-
 encounter-per-patient restriction (earliest encounter only) removes within-patient correlation
 entirely rather than only re-estimating its effect on the standard error. Second, a grace-
 period sensitivity analysis re-defines A using the earlier of days to stress test and days to
 catheterization truncated at 1, 3, 5, and 10 days, while holding the outcome window fixed
 at 30 days. This varies the length of the exposure-identification window discussed above
 without needing day-level outcome timing. Third, given the near-deterministic imbalance in
 same-day AMI status noted above, we refit the logistic AIPW estimator, dropping that one
 covariate from X . Fourth, we compute an E-value [17] for the primary estimate and for its
 confidence limit closer to the null, converting the risk difference to an approximate risk ratio
 using the AIPW-implied standardized risk under no testing, $\hat{R}_0 = \frac{1}{n} \sum_i \hat{m}_0(X_i)$.

The code ¹ performs the analysis with a random seed, writes numerical outputs to an
 output folder and generates the overlap plot and the baseline characteristics table. This
 analysis uses only the repository’s de-identified CSV files.

3 Results

Table 2 shows how the tested and untested groups differ on every recorded characteristic.
 Tested encounters are older (61.2 vs. 53.5 years), less often female (46.3% vs. 58.1%), more
 likely to have an ischaemic ECG pattern (33.9% vs. 21.8%), and are far more likely to have
 a same-day troponin value recorded at all (83.9% vs. 34.8% non-missing). Most strikingly,
 42.8% of tested encounters have a same-day AMI diagnosis versus 0.0% of untested encounters,
 a near-deterministic split that is difficult to reconcile with treating this variable as fixed
 before the testing decision.

Table 3 shows the resulting baseline selection problem in the observed risks. Only 9.0% of
 encounters received testing, but their observed outcome risk was 29.4%, compared with 7.3%
 without testing. The 22.1 pp crude contrast is therefore not credible as a treatment effect.

¹<https://github.com/Bsingstad/Causality-and-MI>

Table 2: Baseline characteristics by testing status.

Characteristic	Tested (n=4,609)	Not tested (n=46,549)
Age, years	61.2 (14.3)	53.5 (19.4)
Female sex	46.3%	58.1%
Black race	19.7%	23.9%
Hispanic ethnicity	13.2%	17.3%
White race	59.8%	50.7%
Same-day troponin missing	16.1%	65.2%
Same-day troponin value, mean (SD)	0.2 (1.4)	0.0 (0.1)
Ischaemic ECG pattern	33.9%	21.8%
Ventricular rate, bpm	78.1 (19.8)	80.5 (18.9)
QTc interval, ms	441.2 (38.5)	437.7 (34.6)
AMI diagnosed same day as visit	42.8%	0.0%

Continuous variables are reported as mean (SD); all others are percentages.

Table 3: Analytic cohort and observed 30-day outcome risk, by testing status.

Group	Encounters	Outcome risk
No cardiac test within 10 days	46,549	7.3%
Cardiac test within 10 days	4,609	29.4%
All eligible encounters	51,158	9.3%

The outcome is a composite of positive cardiac troponin or death within 30 days.

The propensity-score distributions in Figure 3 overlap appreciably, though testing is concentrated at higher estimated propensity. The substantial remaining overlap supports estimation in the observed covariate space; it cannot diagnose a lack of overlap on unmeasured severity.

Both adjustment specifications greatly attenuate the crude contrast but remain positive (Table 4). The logistic AIPW estimate is 5.9 pp (95% CI 3.9-7.9). The boosted-tree estimate is 5.1 pp (3.6-6.6). Agreement across nuisance specifications is useful as a model-dependence check, but it is not evidence against unmeasured confounding.

Table 5 reports the robustness checks. Patient clustering left the interval essentially unchanged (3.9-7.9), and restricting to one encounter per patient gave a similar, though slightly smaller, estimate (4.6 pp, 95% CI 2.5-6.8), so within-patient correlation and repeated encounters are not driving the result. The other two checks are more consequential. Shortening

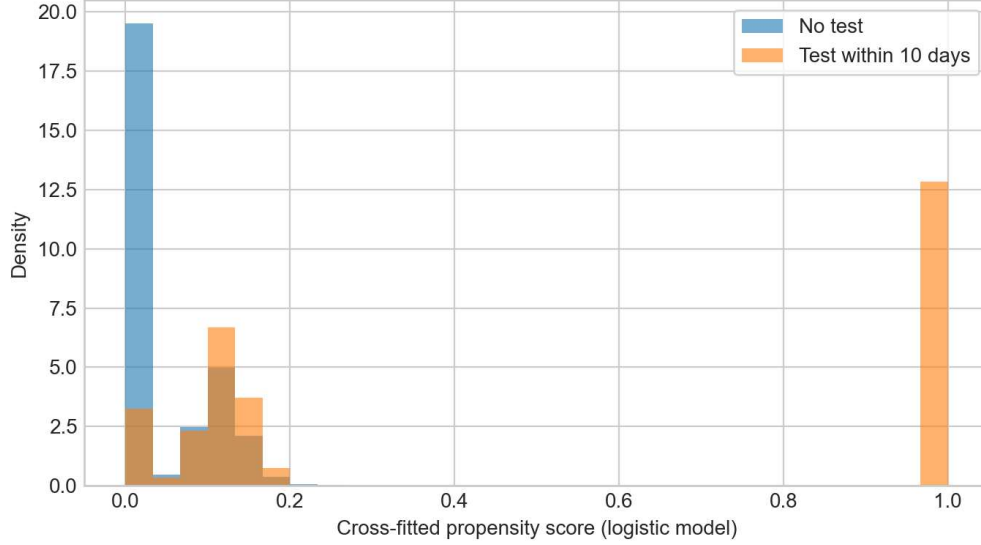


Figure 3: Cross-fitted logistic propensity-score distributions by observed testing status. Scores were truncated only for AIPW estimation, not for this display.

Table 4: Risk difference for cardiac testing within 10 days versus no testing.

Estimator	Risk difference	95% CI
Crude observed contrast	22.1 pp	—
Cross-fitted AIPW, logistic nuisances	5.9 pp	3.9, 7.9
Cross-fitted AIPW, boosted-tree nuisances	5.1 pp	3.6, 6.6

A positive value indicates higher 30-day outcome risk under the observed testing strategy. AIPW: augmented inverse-probability weighting; CI: confidence interval.

the exposure ascertainment window from 10 to 1 day, which shrinks the immortal-time overlap 172
between exposure ascertainment and outcome accrual, reduced the estimate to 0.2 pp (95% 173
CI -1.3 - 1.7), and it rose roughly monotonically back to 5.9 pp as the window widened to 174
10 days. Dropping the same-day AMI covariate from the adjustment set nearly doubled the 175
estimate to 11.1 pp (95% CI 9.2-13.0). The E-value for the primary estimate was 2.66, and 176
2.19 at the confidence limit closer to the null: an unmeasured confounder would need to be 177
associated with both testing and the outcome by a risk ratio of at least this size, beyond the 178
measured covariates, to fully explain away the adjusted association. 179

A positive adjusted association may still reflect residual severity: clinicians may test 180
patients with features absent from the tables precisely because they are at higher risk. The 181

Table 5: Robustness checks on the primary (logistic AIPW) risk-difference estimate.

Check	Risk difference	95% CI
Patient-clustered standard error	5.9 pp	3.9, 7.9
One encounter per patient (n=35,595)	4.6 pp	2.5, 6.8
Exposure window = 1 day	0.2 pp	-1.3, 1.7
Exposure window = 3 days	1.9 pp	0.2, 3.7
Exposure window = 5 days	3.2 pp	1.2, 5.1
Exposure window = 10 days (main analysis)	5.9 pp	3.9, 7.9
Drop same-day-AMI covariate	11.1 pp	9.2, 13.0

Main estimate: 5.9 pp, 95% CI 3.9-7.9. AIPW: augmented inverse-probability weighting; CI: confidence interval.

grace-period and covariate-drop checks show that, in addition, part of the "well-adjusted" 10-day estimate is sensitive to design choices about the exposure-ascertainment window and about which measurements count as baseline, choices that are conceptually distinct from, but as consequential as, unmeasured confounding.

4 Discussion

This project translates a clinically meaningful but observational workflow into an explicit causal estimand. It also changes the original ECG-ischaemia framing into one with an actionable exposure. The numerical pattern is exactly what one expects under confounding by indication: high-risk patients are preferentially tested, and adjustment explains much, but not all, of their observed excess risk.

The most serious limitation is conditional exchangeability. The dataset lacks key variables that influence both testing and outcome, including presenting symptoms, vital signs, detailed medical history, timing of ECG/troponin results relative to testing, clinician assessment, and institutional capacity. The E-value of 2.66 puts a number on this: an unmeasured factor as strongly associated with both testing and the outcome as, say, moderate-to-severe presenting symptoms plausibly are, would be enough to erase the adjusted association. Given how directly clinical severity drives both the decision to test and the risk of the composite outcome,

an E-value in this range should not be read as reassuring.

199

Two further findings complicate a causal reading beyond unmeasured confounding. First, the grace-period sensitivity analysis shows the adjusted estimate is highly sensitive to the length of the exposure-ascertainment window: it is not distinguishable from zero at 1 day and grows to 5.9 pp at 10 days. This pattern is consistent with immortal-time bias from defining the exposure over a window that overlaps the outcome-accrual period, but it is also consistent with a genuine relationship between delay-to-testing and severity (patients tested later may be a systematically different, and possibly sicker, mix over a longer at-risk period than patients tested on day 0–1). We cannot fully separate these explanations without day-level outcome timing, which the repository does not provide; either way, the practical implication is the same, the headline 10-day estimate should not be treated as a stable quantity, and a follow-up analysis should implement the clone-censor-weight approach to target-trial emulation rather than a single fixed window. Second, same-day AMI diagnosis is present in 42.8% of tested but 0.0% of untested encounters, and removing it from the adjustment set nearly doubles the estimate (to 11.1 pp). A covariate that is completely absent in one exposure arm is not functioning as an ordinary baseline confounder in the propensity model; it is more plausibly a concurrent or downstream marker of the same clinical process that leads to testing. Its inclusion is defensible only insofar as the diagnosis genuinely precedes the testing decision within the encounter, which we cannot verify from the available data, so we regard the two estimates it produces (5.9 pp with it, 11.1 pp without it) as bounding our uncertainty about the adjustment rather than as a resolved question.

200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219

Other limitations remain as previously noted. The outcome includes positive troponin, which may be more likely to be detected among tested patients because testing pathways prompt repeat troponin draws (differential identification), even though the exposure itself is defined only by stress testing or catheterization and not by troponin measurement. Testing is heterogeneous, follow-up timing is coarsened, and encounters rather than patients are conceptually the unit of interest, though Table 5 shows that patient clustering has little

220
221
222
223
224
225

practical effect here.

Accordingly, the adjusted estimates should be read as doubly robust estimates under a strong, likely violated identifying model, not as evidence that testing raises harm. A stronger follow-up study would define a narrower treatment (for example, early stress testing among initially low-risk patients), obtain day-level outcome timing so that a clone-censor-weight emulation can replace the fixed 10-day window, add clinical history and ED severity measures, verify the temporal provenance of same-day diagnostic covariates before including them as baseline, use patient-clustered inference by default, and report an E-value alongside every point estimate.

5 Conclusions

Using linked ED, patient, and ECG metadata, we implemented a reproducible target-trial-style AIPW analysis of early cardiac testing. Flexible and parametric nuisance models both reduced a 22.1 pp crude contrast to roughly 5–6 pp, and this reduction was essentially unaffected by patient clustering. However, the same estimate collapsed to a null result under a 1-day exposure-ascertainment window and nearly doubled when a covariate of doubtful baseline status was dropped, and its E-value (2.66) is modest relative to the plausible strength of unmeasured clinical severity. The exercise demonstrates the practical value of DAGs, cross-fitting, overlap assessment, and doubly robust estimation, while its central lesson is caution: causal machine learning can improve adjustment for recorded variables but cannot repair absent clinical confounders, an imprecise intervention window, or covariates whose baseline status is itself doubtful.

Abbreviations

AIPW: Augmented inverse-probability weighting; AMI: Acute myocardial infarction; CI: Confidence interval; DAG: Directed acyclic graph; ECG: Electrocardiogram; ED: Emergency

department; pp: Percentage points; QTc: Corrected QT interval; STROBE: Strengthening
the Reporting of Observational Studies in Epidemiology.

Declarations

Ethics approval and consent to participate. Not applicable as this project uses only
the de-identified, pre-existing CSV extracts supplied with the course repository; the author
did not access identifiable data or interact with participants.

Trial registration. Not applicable. This investigation is a retrospective observational
analysis (target-trial emulation) based on pre-existing, de-identified data and does not include
the prospective allocation of participants to any intervention. Consequently, it does not fulfill
the ICMJE/WHO criteria for a clinical trial that mandates registration.

Consent for publication. Not applicable.

Availability of data and materials. The de-identified CSV files and analysis code are
available in the project repository at <https://docs.ngsci.org/datasets/ed-bwh-ecg/>.

Competing interests. The author declares no competing interests.

Funding. No funding was received for conducting this study.

Authors' contributions. B.-J.S. is the only author of this manuscript and therefore did
all the work.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative AI tools under full human supervision:

- Generate ideas on how to present the results
- Proofreading and language editing
- Suggest reformatting

The generative artificial intelligence (GAI) tool used was: Claude and Writefull. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by: Bjørn-Jostein Singstad ²

References

- [1] Alexander M. Walker. Confounding by indication. *Epidemiology*, 7(4):335–336, 1996.
- [2] Philip D Anderson, Zoey Li, Nick Foster, Senthil Nachimuthu, Josh Risley, and Ziad Obermeyer. Assessing Heart Attack Risk. URL <https://docs.ngsci.org/datasets/ed-bwh-ecg/>.
- [3] Sendhil Mullainathan and Ziad Obermeyer. Solving medicine’s data bottleneck: Nightingale Open Science. *Nature Medicine*, 28(5):897–899, May 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01804-4. URL <https://www.nature.com/articles/s41591-022-01804-4>.

²Generated using: GAIDeT Declaration Generator

- [4] Miguel A. Hernán and James M. Robins. Using big data to emulate a target trial when
a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764,
2016. doi: 10.1093/aje/kwv254.
- [5] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman &
Hall/CRC, 2020.
- [6] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference:
Foundations and Learning Algorithms*. MIT Press, 2017.
- [7] Samy Suissa. Immortal time bias in pharmacoepidemiology. *American Journal of
Epidemiology*, 167(4):492–499, 2008. doi: 10.1093/aje/kwm324.
- [8] Miguel A. Hernán, Brian C. Sauer, Sonia Hernández-Díaz, Robert Platt, and Ian
Shrier. Specifying a target trial prevents immortal time bias and other self-inflicted
injuries in observational analyses. *Journal of Clinical Epidemiology*, 79:70–75, 2016. doi:
10.1016/j.jclinepi.2016.04.014.
- [9] Erik von Elm, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche,
and Jan P. Vandenbroucke. The strengthening the reporting of observational studies
in epidemiology (strobe) statement: guidelines for reporting observational studies. *The
Lancet*, 370(9596):1453–1457, 2007. doi: 10.1016/S0140-6736(07)61602-X.
- [10] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in
observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/
biomet/70.1.41.
- [11] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression
coefficients when some regressors are not always observed. *Journal of the American
Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818.

- [12] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and
causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.
00377.x. 311 312 313
- [13] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. 314
Annals of Statistics, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451. 315
- [14] Stephen R. Cole and Miguel A. Hernán. Constructing inverse probability weights for 316
marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008. 317
doi: 10.1093/aje/kwn164. 318
- [15] A. Colin Cameron and Douglas L. Miller. A practitioner’s guide to cluster-robust 319
inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi: 10.3368/jhr.50.2.317. 320
- [16] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, 321
Whitney Newey, and James Robins. Double/debiased machine learning for treatment 322
and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10. 323
1111/ectj.12097. 324
- [17] Tyler J. VanderWeele and Peng Ding. Sensitivity analysis in observational research: 325
introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274, 2017. doi: 10. 326
7326/M16-2607. 327