UvA-DARE (Digital Academic Repository)

A pragmatic approach to estimating average treatment effects from EHR data

The effect of prone positioning on mechanically ventilated COVID-19 patients

Izdebski, A.; Thoral, P.J.; Lalisang, R.C.A.; McHugh, D.M.; Elbers, P.W.G.; Cinà, G.; Dutch ICU Data Sharing Against COVID-19 Collaborators

[Link to publication](#)

A pragmatic approach to estimating average treatment effects from EHR data: the effect of prone positioning on mechanically ventilated COVID-19 patients

Adam Izdebski[1,2], Patrick J Thoral, MD, EDIC[2], Robbert C A Lalisang, MD[1], Dean M McHugh[3], Robert Entjes, MD[4], Nardo J M van der Meer, MD, PhD[5], Dave A Dongelmans, MD, PhD[2], Age D Boelens, MD[6], Sander Rigter, MD[7], Stefaan H A Hendriks, MD[8], Remko de Jong, MD[9], Marlijn J A Kamps, MD[10], Marco Peters, MD[11], A Karakus, MD[12], Diederik Gommers, MD, PhD[13], Dharmanand Ramnarain, MD[14], Evert-Jan Wils, MD, PhD[15], Sefanja Achterberg, MD, PhD[16], Ralph Nowitzky, MD[17], Walter van den Tempel[18], Cornelis P C de Jager, MD, PhD[19], Fleur G C A Nooteboom, MD[20], Evelien Oostdijk, MD, PhD[21], Peter Koetsier, MD[22], Alexander D Cornet, MD, PhD, FRCP[23], Auke C Reidinga, MD[24], Wouter de Ruijter, MD, PhD[25], Rob J Bosman, MD[26], Tim Frenzel, MD, PhD[27], Louise C Urlings-Strop, MD, PhD[28], Paul de Jong, MD[29], Ellen G M Smit, MD[30], Olaf L Cremer, MD, PhD[31], Frits H M van Osch, PhD[32], Harald J Faber, MD[33], Judith Lens, MD[34], Gert B Brunnekreef, MD[35], Barbara Festen-Spanjer, MD[36], Tom Dormans, MD, PhD[37], Bram Simons, MD[38], A A Rijkeboer, MD[39], Annemieke Dijkstra, MD[40], Sesmu Arbous, MD, PhD[41], Marcel Aries, MD, PhD[42], Menno Beukema, MD[43], Rutger van Raalte[44], Martijn van Tellingen, MD, EDIC[45], Niels C Gritters van den Oever, MD[46], Paul W G Elbers, MD, PhD, EDIC[2], Giovanni Cinà, PhD[1], and on behalf of Dutch ICU Data Sharing Against COVID-19 Collaborators

[1]Pacmed, Amsterdam, NL
[2]Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam UMC, Amsterdam, NL
[3]Institute of Logic, Language, and Computation, University of Amsterdam, NL
[4]Department of Intensive Care, Admiraal De Ruyter Ziekenhuis, Goes, NL
[5]Intensive Care, Amphia Ziekenhuis, Breda en Oosterhout, NL
[6]Anesthesiology, Antonius Ziekenhuis Sneek, Sneek, NL
[7]Department of Anesthesiology and Intensive Care, St. Antonius Hospital, Nieuwegein, NL
[8]Intensive Care, Albert Schweitzerziekenhuis, Dordrecht, NL

[9]Intensive Care, Bovenij Ziekenhuis, Amsterdam, NL

[10]Intensive Care, Catharina Ziekenhuis Eindhoven, Eindhoven, NL

[11]Intensive Care, Canisius Wilhelmina Ziekenhuis, Nijmegen, NL

[12]Department of Intensive Care, Diakonessenhuis Hospital, Utrecht, NL

[13]Department of Intensive Care, Erasmus Medical Center, Rotterdam, NL

[14]Intensive Care, ETZ Tilburg, Tilburg, NL

[15]Department of Intensive Care, Franciscus Gasthuis & Vlietland, Rotterdam, NL

[16]ICU, Haaglanden Medisch Centrum, Den Haag, NL

[17]Intensive Care, HagaZiekenhuis, Den Haag, NL

[18]Department of Intensive Care, Ikazia Ziekenhuis Rotterdam, Rotterdam, NL

[19]Department of Intensive Care, Jeroen Bosch Ziekenhuis, Den Bosch, NL

[20]Intensive Care, Laurentius Ziekenhuis, Roermond, NL

[21]ICU, Maasstad Ziekenhuis Rotterdam, Rotterdam, NL

[22]Intensive Care, Medisch Centrum Leeuwarden, Leeuwarden, NL

[23]Department of Intensive Care, Medisch Spectrum Twente, Enschede, NL

[24]ICU, SEH, BWC, Martiniziekenhuis, Groningen, NL

[25]Department of Intensive Care Medicine, Northwest Clinics, Alkmaar, NL

[26]ICU, OLVG, Amsterdam, NL

[27]Department of Intensive Care Medicine, Radboud University Medical Center, Nijmegen, NL

[28]Intensive Care, Reinier de Graaf Gasthuis, Delft, NL

[29]Department of Anesthesia and Intensive Care, Slingeland Ziekenhuis, Doetinchem, NL

[30]Intensive Care, Spaarne Gasthuis, Haarlem en Hoofddorp, NL

[31]Intensive Care, UMC Utrecht, Utrecht, NL

[32]Department of Clinical Epidemiology, VieCuri Medisch Centrum, Venlo, NL

[33]ICU, WZA, Assen, NL

[34]ICU, ICU, IJsselland Ziekenhuis, Capelle aan den IJssel, NL

[35]Department of Intensive Care, Ziekenhuisgroep Twente, Almelo, NL

[36]Intensive Care, Ziekenhuis Gelderse Vallei, Ede, NL

[37]Intensive care, Zuyderland MC, Heerlen, NL

[38]Intensive Care, Bravis Ziekenhuis, Bergen op Zoom en Roosendaal, NL

[39]ICU, Flevoziekenhuis, Almere, NL

[40]Department of Intensive Care Medicine, Het Van Weel-Bethesda Ziekenhuis, Dirksland, NL

[41]Intensivist, LUMC, Leiden, NL

[42]MUMC+, University Maastricht, Maastricht, NL

[43]Department of Intensive Care, Streekziekenhuis Koningin Beatrix, Winterswijk, NL

[44]Department of Intensive Care, Tergooi hospital, Hilversum, NL
[45]Department of Intensive Care Medicine, afdeling Intensive Care, ziekenhuis
Tjongerschans, Heerenveen, NL
[46]Intensive Care, Treant Zorggroep, Emmen, NL

September 15, 2021

Abstract

Background

Despite the recent progress in the field of causal inference, to date there is no agreed upon
methodology to glean treatment effect estimation from observational data. The consequence
on clinical practice is that, when lacking results from a randomized trial, medical personnel
is left without guidance on what seems to be effective in a real-world scenario. This article
showcases a pragmatic methodology to obtain preliminary estimation of treatment effect
from observational studies.

Methods

Our approach was tested to an open problem, the estimation of treatment effect of the
proning maneuver on oxygenation levels, on a cohort of COVID-19 Intensive Care patients.
We modeled our study design on a recent RCT for proning (the PROSEVA trial) in order
to select relevant covariates, outcome definition and inclusion criteria. Linear regression,
propensity score models such as blocking and doubly robust inverse probability weighting,
BART and two versions of Counterfactual Regression were employed to provide estimates
on observational data comprising first wave COVID-19 ICU patient data from 25 Dutch
hospitals.

Findings

6371 data points, from 745 mechanically ventilated patients, were included in the study.
Estimates for the early effect of proning – P/F ratio from 2 to 8 hours after proning – ranged
between 14·54 and 20·11 mm Hg depending on the model. Estimates for the late effect of
proning – oxygenation from 12 to 24 hours after proning – ranged between 13·53 and 15·26
mm Hg. All confidence interval being strictly above zero indicated that the effect of proning
on oxygenation for COVID-19 patient was positive and comparable in magnitude to the
effect on non COVID-19 patients.

Interpretation

These results provide further evidence on the effectiveness of proning on the treatment of
COVID-19 patients. This study, along with the accompanying open-source code, provides a
blueprint for treatment effect estimation in scenarios where RCT data is lacking.

Funding

SIDN fund, CovidPredict consortium, Pacmed.

1 Introduction

A central problem of causal inference is estimating treatment effects from observational data. A
vast body of literature has addressed the issue of finding the best method for this task, but there

is no established winner.[1] Most model comparisons are performed on synthetic data, and the race for developing new methodologies is ongoing.[2]

However, clinicians in practice are often faced with the problem of not having guidance on which treatment to use. A clear example of this phenomenon is the COVID-19 pandemic, during which clinical staff had to treat patients with a novel condition for which no randomized controlled trials (RCTs) on treatment effectiveness were available.[3] When there is no RCT data, clinicians are left to improvise based only on their clinical experience. This points to the need for a methodology to achieve preliminary yet robust conclusions from observational data, to guide clinicians' behaviour while randomized trials are ongoing and to determine future directions for RCTs. This need is highlighted by the increasing attention that regulatory bodies in the US and Europe are placing on the use of observational evidence.[4,5]

In this paper we propose a procedure to arrive at sensible treatment effect estimation from Electronic Health Records (EHRs). Our proposal is twofold. First, we suggest to emulate a target RCT, whose results provide a meaningful point of reference.[6,7] This is achieved by carefully replicating as much as possible the design of the past trial, i.e. using the same inclusion criteria, the same baseline covariates, and so forth. From a heuristic standpoint, this past RCT can act as a 'prior', indicating how much the results obtained from EHR data diverge from what was previously known. Second, we put forward a shortlist of methods to estimate treatment effect, including well-established ones and more recent Machine Learning techniques, with different theoretical guarantees. More than finding a single best-performing model, we intend to measure the level of agreement between the candidate models.

If our procedure is successful in both respects, meaning one can successfully emulate the design of similar RCTs from the past and the array of models provide relatively uniform estimations, we argue that the conclusions can be regarded as robust enough to guide clinical practice. This approach is 'pragmatic' in the sense that it can provide sought-after preliminary knowledge when an RCT is not available. To showcase the usefulness and applicability of this strategy, we test it on a real-world, large-scale use case, the estimation of the effect of proning on severely hypoxemic mechanically ventilated COVID-19 patients in the Intensive Care unit (ICU). The code to implement this strategy and to obtain the results described in the paper is fully available online.

2   Research in context

Prone positioning, consisting of turning the patient from the supine to prone position, is a commonly used technique for the treatment of severely hypoxemic mechanically ventilated patients with acute respiratory distress syndrome (ARDS). The current definition of ARDS uses the P/F ratio to classify disease severity. This ratio, defined as the partial pressure of O_2 in an arterial blood sample (PaO_2) divided by the fraction of inspired oxygen (FiO_2) ($\frac{PaO_2}{FiO_2}$ in mm Hg), subdivides patients in mild, moderate, and severe subgroups.

2.1 Evidence before this study

Meta-analysis of eight RCTs reported that prone positioning applied for at least 12 hours a day improved the P/F ratio on average by 23 mm Hg (95% CI: 12, 35) as compared to the supine (not-proned, control) group. In the PROSEVA trial, it was found that applying prone positioning for an average of 17 hours a day improved the P/F ratio on average by 15 mm Hg (95% CI: 3, 27) and therefore it is currently considered a cornerstone in the treatment of patients with ARDS and a P/F ratio of <150 mm Hg.

Prone positioning has been used widely as adjuvant therapy for respiratory failure in severe COVID-19 induced pneumonia similar to ARDS. However, increasing evidence suggests that mechanisms underlying COVID-19 respiratory failure may be different from 'classical' ARDS due to the frequent occurrence of pulmonary thrombosis. At the time of writing, no RCT has been conducted on the effect of proning on intubated patients with COVID-19 induced ARDS; furthermore, the available observational studies were conducted only on relatively small and single-center cohorts.

2.2 Added value of this study

In our observational study based on patients from the Dutch Data Warehouse (DDW) - a multi-center, large-scale ICU COVID-19 cohort comprising data from 25 Dutch hospitals - we found that applying prone positioning improved the P/F ratio, with estimates varying between 14·54 and 20·11 mm Hg (depending on the model) in the time window 2-8 hours after proning and between 13·53 and 15·26 mm Hg. in the time window 12-24 hours after proning (see Table 2). This aligns with results reported by previous RCTs and suggests a positive effect of prone positioning on patients with COVID-19 induced ARDS.

To ensure validity of the causal inference methods and put the estimates in context, the study design emulated the PROSEVA trial. To facilitate transportability of our results and replication on other cohorts, we made both the data processing pipeline as well as the suite of causal inference experiments available open source.

2.3 Implications of all the available evidence

Estimating treatment effects from observational data is a fundamental problem transcending disciplinary boundaries. While multiple methods and strategies to arrive at causal quantities of interests do exist, the task itself is claimed to be notoriously hard and not straightforward to implement. While all the applied methodologies were developed separately beforehand, we show how they can be combined together to arrive at clinically meaningful conclusions from unprocessed, unstructured and highly noisy EHR data.

3 Methods

We carefully designed our data set to emulate an RCT as much as possible, in order to provide the most meaningful term of comparison. This means using the same inclusion criteria, selecting the same covariates, measuring a similar outcome, and so forth. This strategy allowed us to compare ATE estimates with the literature, and we were able to achieve a reasonable level of confidence in the fulfillment of the identifiability assumptions, which are not testable from observational data (see Appendix C.1 for a discussion).

Since current guidelines for prone positing in ARDS are based mainly on the results of the PROSEVA trial,[8] we specifically chose to emulate this RCT. We used data from the DDW[9] to create a cohort with the same covariates and inclusion criteria as this trial. The DDW is the result of a nationwide, large-scale ICU data sharing collaboration in the Netherlands that was initiated during the COVID-19 pandemic. The collaboration and the data warehouse are described in detail in Appendix A. The steps followed to defined the cohort are outlined in Figure 1 and described in detail in the following sections.

We note that, since COVID-19 is a new disease, emulating an RCT does not mean replicating it. As far as we know, the treatment effect of proning for COVID-19 ARDS could be different from the one for standard ARDS, so the PROSEVA RCT should not be regarded as providing the "true" effects that the models are trying to estimate. Nonetheless, the two pathologies are still close enough that the PROSEVA trial can provide a meaningful term of comparison; indeed this similarity is why proning has been employed for COVID-19 ARDS. For this reason the PROSEVA trial can i) provide a blueprint for the kind of experiment we want to run for proning on COVID-19 patients – in terms of covariates, inclusion criteria, outcome definition, etc – and ii) serve to put the observational ATE estimates in context.

3.1 Constructing Observations

The data contained information of 2052 patients recorded in the DDW who were admitted to the ICU between March and October 2020. All patients were diagnosed with COVID-19. For all patients we extracted data concerning all supine and prone sessions. The data of each patient was sliced into prone and supine sessions as exemplified in Figure 2. All created sessions were later selected depending on adherence to the inclusion criteria (see Section 3.3).

We took sessions to be the observations of our study, meaning that a single patient can be represented by multiple observations. A session was defined as a time interval throughout which the patient was in the same position, either prone or supine.

Sessions were employed as observations to have a data point for every moment in which a treatment decision of proning could have been made. Artificial supine sessions were created for this reason, since during a supine session there are multiple occasions in which clinicians can decide to turn the patient to prone position. An additional artificial supine session was created if, at least 8h before the end of the original supine session, the patient's blood gas (PaO_2) and ventilator values (PEEP, FiO_2) were documented again. The newly created artificial session
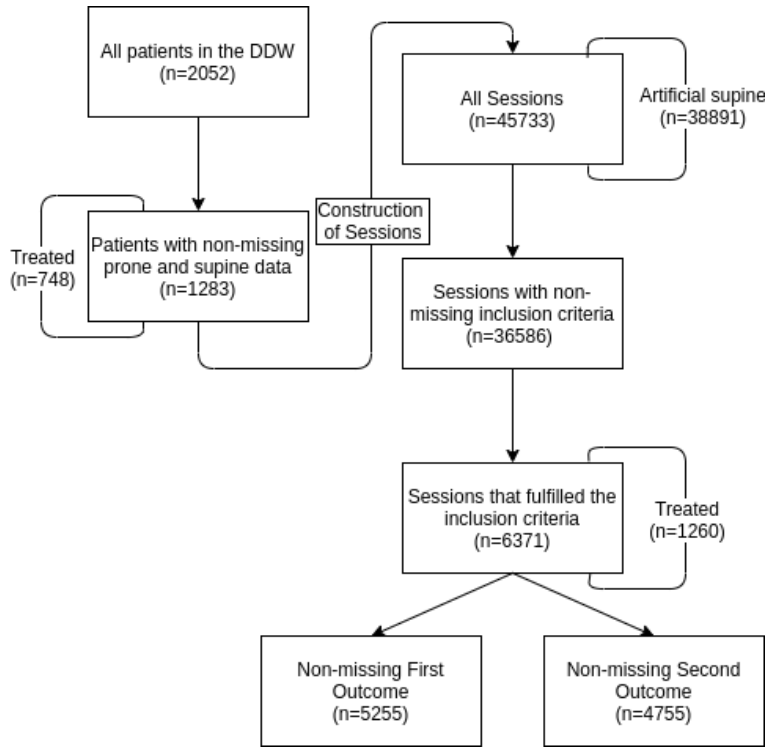
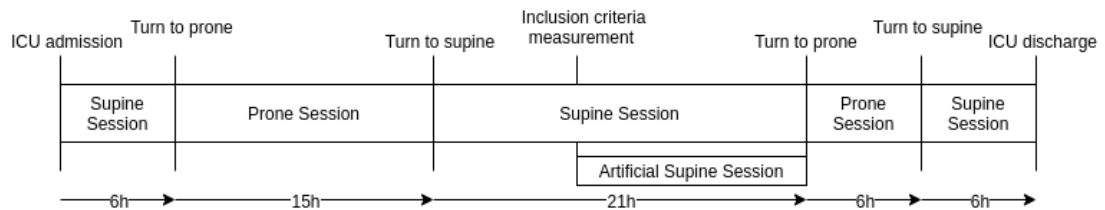Figure 1: Overview of our study design.

Figure 2: We describe how the data of a fictitious ICU patient would be converted into four supine and two prone sessions. An artificial supine session is created from a long supine session if there is a corresponding measurement allowing to check the inclusion criteria performed at least 8h before the end of the original session.

ended at the same time point at which the original supine session ended. Altogether, we created 45733 sessions of which 2762 were prone, 4080 were original supine sessions and 38891 were artificial supine sessions.

3.2 Constructing Features and Outcomes

28 covariates were constructed from the data collected in the DDW. They correspond to all covariates presented in the PROSEVA manuscript except for sepsis, McCabe and SAPS II scores, which were not available in the DDW. We included two additional covariates, one indicating morbid obesity (BMI > 35) and one for driving pressure, since plateau pressure measurements were missing for the majority of observations, because most patients in Dutch ICU's were ventilated in a pressure controlled mode. The list of included covariates can be found in Table 1.

For each session we constructed covariates from the DDW by taking the last measurement in the eight hour interval before the start of the session. Since in practice there can be some delay in the recording of such parameters, if a measurement was missing, the first measurement was taken in the 30 minutes interval following the start of the session.

We defined two outcomes. The first outcome was the last P/F ratio measurement recorded in the time interval from two hours up to eight hours from the start of the session but before the end of the session. This outcome corresponded to the measurement taken just after the prone/supine session started ('early proning effect'). The second outcome was the last P/F ratio measurement recorded in the time interval from 12h up to 24h from the start of the session but before end of the session. Given the average length of prone sessions in this cohort ($\approx$19 hours), this outcome corresponded to the measurement taken just before the session ends ('late proning effect'). Consequently, all sessions shorter than 12h will not have a defined late proning effect.

Because of the way observations were defined, the outcomes measured the average treatment effect of a single prone session on the P/F ratio, as opposed to measuring the treatment effect of repeated proning sessions applied for a fixed time interval (e.g. three days for the PROSEVA trial). Beside the medical relevance of intermediate outcomes, this methodological choice was due to the fact that in practice patients were not proned for an identical amount of hours for a few days in a row, so the exact administering of the proning sessions from the RCT cannot be replicated. Furthermore, the aforementioned way of constructing covariates achieved perfect separation between covariates and outcomes, i.e. covariates were always measured before the outcomes, allowing for valid causal inference.

3.3 Cohort Selection and Inclusion Criteria

All sessions that did not have corresponding PaO_2, FiO_2 and PEEP baseline measurements were discarded. Since non-invasive mechanical ventilation for COVID-19 patients was rare in the Netherlands (about 1%),[9] all sessions reporting the combination of PEEP and FiO_2 were considered to be sessions concerning mechanically ventilated patients. Furthermore, we

discarded all proning sessions longer than 96 hours, since such sessions were extremely rare, and from a medical perspective were most likely the result of data entry errors or omissions. Following this, the inclusion criteria from the PROSEVA trial were applied for the resulting sessions. These were P/F ratio $<$ 150 mm Hg, $FiO_2 \geq 60\%$ and PEEP $\geq$ 5 cm of water. We chose not to include the criterion 'tidal volume about 6 ml/kg', since the original PROSEVA paper did not specify explicit cutoff values for tidal volume. In addition, current Dutch guidelines for mechanical ventilation of COVID-19 patients already suggest to ventilate COVID-19 patients similarly to ARDS patients with tidal volumes of about 6 ml/kg.[10] For both outcomes, all observations with a missing outcome were discarded. Consequently, for the late proning effect we automatically discarded all sessions shorter than 12 hours, since they cannot have an outcome in the window from 12 to 24 hours. For all numerical variables we imputed data with mean values. All binary variables were imputed with the value *False*. Imputation was performed in such a way to prevent any form of data leakages between data splits. This selection resulted in 6371 observations (from 745 patients) included in our study, 1260 (from 493 patients) of which were prone, significantly more than the PROSEVA trial that included 466 observations (each corresponding to a patient), 237 of which were prone.

3.4   Feature Characteristics and comparison to PROSEVA Trial

All the previous steps of our study design were defined in order to create a sub-population from the available EHR data that emulates the PROSEVA trial. The characteristics of all observations that fulfilled the inclusion criteria are summarized in Table 1.

Compared to the PROSEVA trial, the average COVID-19 patient was slightly older, with a higher incidence of diabetes, chronic renal failure and COPD. The lower average SOFA score is in line with the observation that COVID-19 patients suffer primarily from respiratory failure but not from septic shock or multiple organ dysfunction syndrome that typically accompany ARDS patients. On average, observations in our study had higher PEEP at similar or lower FiO_2. Tidal volume was slightly above the recommended 6 ml/kg PBW with lower respiratory rates compared to the PROSEVA ARDS patients.

In our study higher levels of PEEP were applied and lower P/F ratios measured during prone sessions compared to PROSEVA, which may be explained by delayed proning in overwhelmed ICUs or uncertainty of effects of proning for COVID-19 patients. The most imbalanced variables between the prone and supine sessions of our study were FiO_2, PaO_2, P/F ratio, PEEP, SOFA Score and the usage of medications.

3.5   Models

A suite of causal inference models was employed to estimate the treatment effect of proning on oxygenation, including classical techniques as well as modern Machine Learning estimators:

- Linear regression (LR)
- Propensity score model: Blocking

- Propensity score model: doubly robust inverse probability weighting (DR-IPW)

- Bayesian Additive Regression Trees (BART)

- Treatment-Agnostic Representation Network (TARNet)

- Counterfactual Regression (CfR)

For an introduction to the potential outcome framework, details on the models as well as justification for their choice we refer the reader to the Appendix Section C. Clarifications on the notation can be found in Appendix B, while the specification on the models' hyper-parameters are reported in Appendix D.

4 Results

The goal of the study was to estimate the average treatment effect (ATE) of prone positioning on P/F ratio as measured after a session starts (early proning effect) and just before an average prone session ends (late proning effect). For each outcome the following procedure was instated. To estimate the ATE we performed a 80/20 train/test split. Each model was trained on 100 bootstrap re-samples of the train set (we repeatedly sampled 95% of the train set with replacements) and with each model an ATE estimate was calculated on the test set. We obtained the ATE estimate (Table 2) by taking the average of ATE estimates for every bootstrap sample of the train set. The 95% confidence intervals are calculated as bootstrap percentile intervals (as in Section 8 of Wasserman [12]). Prior to calculating ATE estimates, a hyper-parameter search was performed for some of the models. This search, as well as the strategy followed for the estimation of the propensity scores, is described in Appendix D. As we could not measure the accuracy of the estimated ATE, we evaluated the predictive performance for each model with RMSE on the test set for each of the 100 different bootstrap samples and reported the average. The 95% confidence intervals were calculated in the same way as for the ATE estimates.

The aim of the experiments was to assess whether the different frameworks for causal inference examined in Section C.2 provided uniform ATE estimates. Moreover, we checked whether the estimates were close to the ATE of 15 mm Hg reported by the PROSEVA trial. Comparing the models' ATE estimates with the result of the PROSEVA trial allowed us to gain further confidence in the robustness of the models' estimates.

4.1 Average Treatment Effect Estimation

The estimated mean ATEs ranged from 14·54 to 20·11 mm Hg for the early proning effect, and from 13·53 to 15·26 mm Hg for the late proning effect. Confidence intervals for both outcomes being strictly above zero indicated a positive treatment effect of proning on P/F ratio. A summary of all estimates can be found in Table 2 and Figure 3.
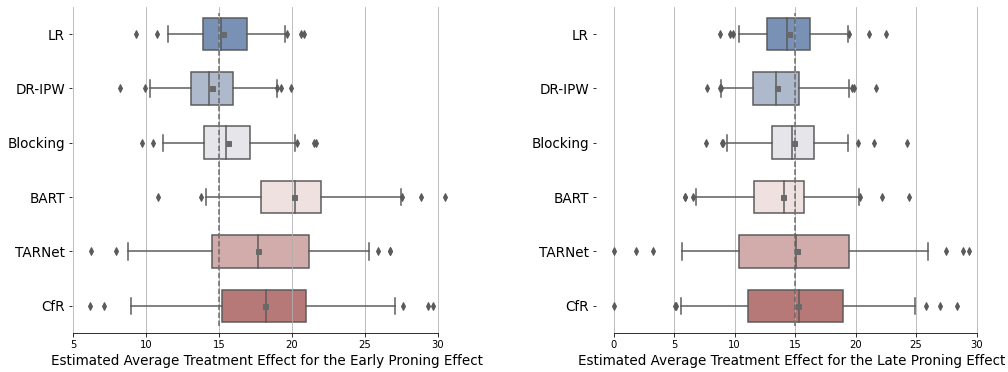
Figure 3: Distribution of the average treatment effects estimated on the test set for the early proning effect (Left) and the late proning effect (Right). Each boxplot displays the beginning of the 95% CI, the first quartile, the median, the third quartile and the end of the 95% CI of all obtained ATE estimates. The black square indicates the sample mean, which is the final ATE estimate reported in Table 2. The vertical, dotted line indicates the ATE (=15) from PROSEVA trial.

4.1.1 Early Proning Effect

The unadjusted treatment effect, meaning the difference in the mean outcome in the treated and control group, was 12·89 (95% CI: 7·88, 17·22). Linear Regression (LR) and propensity score models (DR-IPW, blocking) provided the narrowest 95% CIs for the ATE while deep learning methods (TARNet, CfR) the widest ones. This was probably due to the fact that the latter models are more prone to overfitting on the different bootstrap samples. Linear Regression, DR-IPW weighting and Blocking reported the ATE estimates closest to the PROSEVA trial, with BART, TARNet and CfR giving slightly higher estimates.

The best performing model with respect to predictive performance was BART (RMSE: 48·62), slightly outperforming TARNet and CfR (RMSE: 51·72 and 51·82 respectively). The performance of the selected models was in line with experiments on synthetic data (Table 1 in Shalit, Johansson, and Sontag[13]), where BART, CfR and TARNet reported comparable errors with respect to ATE estimation (ε_{ATE}) in out-of-sample testing on both IHDP and Jobs data sets. Linear Regression and propensity models were outperformed by more recent techniques with respect to predictive performance, probably due to the greater flexibility of the latter methods. This might explain also why the ATE estimates were slightly higher for the non-linear models.

4.1.2 Late Proning Effect

The unadjusted treatment effect was equal to 11·88 (95% CI: 7·88, 16·91). Compared to the early proning effect, all models provided lower ATE estimates. The estimates reported for the

11

two outcomes differed significantly, with BART reporting the wider gap (20·11 vs. 13·99) as opposed to Linear Regression giving more similar estimates (15·31 vs. 14·47). Compared to the early proning effect, all estimates were closer to the treatment effect from the PROSEVA trial.

The predictive performance of BART, TARNet and CfR remained the strongest with similar RMSE as for the early proning effect. Obtaining ATE estimates for the late proning effect seemed to be an easier task for the Linear Regression and propensity models as they reported smaller RMSE compared to the early proning effect. While non-linear models still outperformed linear ones, the difference was much smaller in RMSE, indicating that perhaps non-linear effects played less of a role for this outcome.

Models of a similar kind, both Linear Regression and propensity models, and CfR and TARNet reported comparable ATE estimates and predictive performances. This was somewhat surprising, since one would expect that the correction that propensity score models and CfR provide on top of Linear Regression and TARNet (respectively) would improve their performance and possibly modify their ATE. One of the possible explanations is that our data was rather balanced, with high treatment assignment variation. As there was no observable metric to measure the accuracy of CATE estimates and RCTs do not provide CATE estimates to compare with, it remains an open question to what extent these models could be used for CATE estimation.

5   Discussion

In this article we addressed an urgent problem, namely the lack of guidance for clinicians when RCT results are not available, by describing an approach to obtain information on treatment effectiveness from EHR data. We suggested that a multi-viewpoint approach, where different models are tested in parallel, coupled with an RCT-like experimental design can be employed to reach a reasonable level of confidence in the result of an observational study. This approach was tested on an important use case, namely treatment effect of the proning maneuver on COVID-19 patients on the ICU. Our experiments showed that our diverse set of models reached a relatively uniform conclusion. Moreover, our findings suggested that mechanically ventilated COVID-19 patients benefit from proning to a similar extent as established by the PROSEVA trial. This provides important medical knowledge in a situation when RCTs on proning for mechanically ventilated COVID-19 patients are not yet available and observational evidence is limited to single-center and relatively small cohorts.[14,15,16,17,18]

The extent to which results such as the ones presented here can be used to directly affect clinical decision making, or can be contrasted with RCTs, should be a matter of further discussion, as well as scrutiny from a regulatory perspective.[19] These results cannot be considered definitive and a randomized trial is required to formally establish the effectiveness of the treatment. Repetition of these experiments on another observational COVID-19 cohort may also help in assessing the robustness of the estimates; the fully documented and open-source code released with this article facilitates and encourages such follow-up analyses. Despite their potential brittleness and biases, observational studies such as the one presented here can help provide temporary answers for clinicians while RCTs are developed, as well as accelerate and guide prospective studies.[20] Lastly, further improvement to the models' predictive performance could warrant the

investigation of the effect of proning on subgroups or of the threshold on P/F ratio for which proning is beneficial.

Limitations   We now come to the limitations that should be considered when weighing the result of this study. First, we do not compare randomized experiments and observational studies with identical designs, such as in Shadish, Clark, and Steiner,[21] since from the DDW data it is clear that in practice the proning maneuver is not administered exactly as in the RCTs. This also holds for the outcome: the difference in treatment led us to consider oxygenation after proning, which differs from the oxygenation outcome after three days of repeated proning as defined in the PROSEVA trial. It should be noted however that RCTs already display a great deal of variability. The proning maneuver itself is performed differently in all RCTs, and the outcomes are also measured at different time points, thus the discrepancy between our study and the RCTs is arguably not greater than the one existing between RCTs. It also worth noting that the usage of sessions introduces dependence between samples from the same patient and might lead to over-representation of some individuals; sicker patients, for example, might stay longer in the ICU and generate more sessions. However, we do not register any additional imbalance in the distribution of baseline covariates, as reported in Table 1. Furthermore, we sought to mitigate this problem by employing bootstrap samples. We note that the construction of multiple data points per patients opens the possibility to study repeated or time-varying treatment by means of causal inference methods for longitudinal data; this line of enquiry is left for future work.

Second, all models assume the absence of unmeasured confounders. To facilitate comparison with previous RCTs, we decided to limit the set of covariates to those in the study by Guérin and colleagues.[8] While this expert-picked list ensures that most relevant variables are considered, hence giving some measure of reassurance that the assumption is fulfilled, it is already known that it might be relevant to add a few more covariates to have a more complete picture of the patient, e.g. fluid balance. We believe that for our case study it is possible to minimize the impact of this assumption, since the ICU routinely monitors patients in great detail. It is therefore reasonable to assume that the vast majority, if not all, of relevant parameters are recorded in EHRs, meaning that a future iteration of this study on a larger set of covariates could provide more robust results.

Third, it could be argued that a different or wider selection of models is more appropriate. The guiding principle of our choice of models was to represent both traditional methods used in the medical community as well as top performers from previous causal inference competitions, to arrive at modern Machine Learning techniques with strong theoretical guarantees. The extension of this selection, which could provide additional information (as long as the list of models is fixed prior to the experiments), will be investigated in the future.

Finally, a further point of improvement for our approach is the enhancement of the comparison between the different models. We adopted simple and well-known metrics such as RMSE, but more advanced techniques are becoming available for observational studies, such as those by Alaa and van der Schaar,[22] allowing for a more fine-grained analysis of the models' differences.

In conclusion, these results provide further evidence on the effectiveness of proning for the treatment of mechanically ventilated COVID-19 patients. The design of this study, implemented in fully open-source code, allows for reproducibility on other COVID-19 cohorts and provides a

blueprint for treatment effect estimation in scenarios where RCT data is lacking.

6 Data Sharing

Within the boundaries of privacy laws and ethics, access to the DDW can be requested on the portal of the consortium: `www.icudata.nl`. The code used to process the data, implement the causal inference models and perform the experiments is available online at `github.com/Pacmed/causal_inference`.

7 Acknowledgements

References

[1] V. Dorie, J. Hill, U. Shalit, M. Scott, D. Cervone, *et al.*, "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition," *Statistical Science*, vol. 34, no. 1, pp. 43–68, 2019.

[2] I. Bica, A. M. Alaa, C. Lambert, and M. van der Schaar, "From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges," *Clinical Pharmacology & Therapeutics*, vol. 109, no. 1, pp. 87–100, 2021.

[3] C. Guérin, R. K. Albert, J. Beitler, L. Gattinoni, S. Jaber, J. J. Marini, L. Munshi, L. Papazian, A. Pesenti, A. Vieillard-Baron, *et al.*, "Prone position in ards patients: why, when, how and for whom," *Intensive care medicine*, pp. 1–12, 2020.

[4] A. Cave, X. Kurz, and P. Arlett, "Real-world data for regulatory decision making: challenges and possible solutions for europe," *Clinical Pharmacology & Therapeutics*, vol. 106, no. 1, pp. 36–39, 2019.

[5] FDA, "Us food and drug administration. real-world evidence," 2019.

[6] M. A. Hernán and J. M. Robins, "Using big data to emulate a target trial when a randomized trial is not available," *American journal of epidemiology*, vol. 183, no. 8, pp. 758–764, 2016.

[7] B. Gershman, D. P. Guo, and I. J. Dahabreh, "Using observational data for personalized medicine when clinical trial evidence is limited," *Fertility and Sterility*, vol. 109, no. 6, pp. 946–951, 2018.

[8] C. Guérin, J. Reignier, J.-C. Richard, P. Beuret, A. Gacouin, T. Boulain, E. Mercier, M. Badet, A. Mercat, O. Baudin, *et al.*, "Prone positioning in severe acute respiratory distress syndrome," *New England Journal of Medicine*, vol. 368, no. 23, pp. 2159–2168, 2013.

[9] L. M. Fleuren, D. P. de Bruin, M. Tonutti, R. C. Lalisang, and P. W. Elbers, "Large-scale icu data sharing for global collaboration: the first 1633 critically ill covid-19 patients in the dutch data warehouse," *Intensive Care Medicine*, 2021.

[10] L. M. A. Heunks, R. Endeman, and J. van der Hoeven, "Mechanical ventilation in covid-19. dutch society for intensive care medicine (nvic) guideline," Oct 2020.

[11] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[12] L. Wasserman, *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[13] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," 2017.

[14] C. Pan, L. Chen, C. Lu, W. Zhang, J.-A. Xia, M. C. Sklar, B. Du, L. Brochard, and H. Qiu, "Lung recruitability in covid-19–associated acute respiratory distress syndrome: a single-center observational study," *American journal of respiratory and critical care medicine*, vol. 201, no. 10, pp. 1294–1297, 2020.

[15] F. Perier, S. Tuffet, T. Maraffi, G. Alcala, M. Victor, A.-F. Haudebourg, N. De Prost, M. Amato, G. Carteaux, and A. Mekontso Dessap, "Effect of positive end-expiratory pressure and proning on ventilation and perfusion in covid-19 acute respiratory distress syndrome," *American Journal of Respiratory and Critical Care Medicine*, vol. 202, no. 12, pp. 1713–1717, 2020.

[16] M. Berrill, "Evaluation of oxygenation in 129 proning sessions in 34 mechanically ventilated covid-19 patients," *Journal of Intensive Care Medicine*, vol. 36, no. 2, pp. 229–232, 2021.

[17] T. T. Weiss, F. Cerda, J. B. Scott, R. Kaur, S. Sungurlu, S. H. Mirza, A. A. Alolaiwat, A. E. Augustynovich, and J. Li, "Prone positioning for patients intubated for severe acute respiratory distress syndrome (ards) secondary to covid-19: a retrospective observational cohort study," *British journal of anaesthesia*, vol. 126, no. 1, pp. 48–55, 2021.

[18] M. C. Shelhamer, P. D. Wesson, I. L. Solari, D. L. Jensen, W. A. Steele, V. G. Dimitrov, J. D. Kelly, S. Aziz, V. P. Gutierrez, E. Vittinghoff, *et al.*, "Prone positioning in moderate to severe acute respiratory distress syndrome due to covid-19: A cohort study and analysis of physiology," *Journal of Intensive Care Medicine*, vol. 36, no. 2, pp. 241–252, 2021.

[19] H.-G. Eichler, F. Koenig, P. Arlett, H. Enzmann, A. Humphreys, F. Pétavy, B. Schwarzer-Daum, B. Sepodes, S. Vamvakas, and G. Rasi, "Are novel, nonrandomized analytic methods fit for decision making? the need for prospective, controlled, and transparent validation," *Clinical Pharmacology & Therapeutics*, vol. 107, no. 4, pp. 773–779, 2020.

[20] B. K. Beaulieu-Jones, S. G. Finlayson, W. Yuan, R. B. Altman, I. S. Kohane, V. Prasad, and K.-H. Yu, "Examining the use of real-world evidence in the regulatory process," *Clinical Pharmacology & Therapeutics*, vol. 107, no. 4, pp. 843–852, 2020.

[21] W. R. Shadish, M. H. Clark, and P. M. Steiner, "Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments," *Journal of the American statistical association*, vol. 103, no. 484, pp. 1334–1344, 2008.

[22] A. Alaa and M. Van Der Schaar, "Validating causal inference models via influence functions," in *International Conference on Machine Learning*, pp. 191–201, PMLR, 2019.

[23] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[24] J. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, pp. 217–240, 03 2011.

[25] M. A. Hernán and J. M. Robins, *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC, 2020.

[26] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[27] G. W. Imbens, "Nonparametric estimation of average treatment effects under exogeneity: A review," *Review of Economics and statistics*, vol. 86, no. 1, pp. 4–29, 2004.

[28] J. D. Kang, J. L. Schafer, *et al.*, "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical science*, vol. 22, no. 4, pp. 523–539, 2007.

[29] H. A. Chipman, E. I. George, R. E. McCulloch, *et al.*, "Bart: Bayesian additive regression trees," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266–298, 2010.

[30] Q. Zhou and J. S. Liu, "Extracting sequence features to predict protein–dna interactions: a comparative study," *Nucleic acids research*, vol. 36, no. 12, pp. 4137–4148, 2008.

[31] G. Blattenberger and R. Fowles, "Avalanche forecasting: Using bayesian additive regression trees (bart)," in *Demand for Communications Services–Insights and Perspectives*, pp. 211–227, Springer, 2014.

[32] C. Villani, *Optimal transport: old and new*, vol. 338. Springer Science & Business Media, 2008.

[33] F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag, "Generalization bounds and representation learning for estimation of potential outcomes and causal effects," 2020.

[34] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[35] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[36] A. Kapelner and J. Bleich, "bartMachine: Machine learning with Bayesian additive regression trees," *Journal of Statistical Software*, vol. 70, no. 4, pp. 1–40, 2016.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

A   Dutch Data Warehouse

The DDW was constructed by combining, unifying, and structuring EHR data from 25 different ICUs in the Netherlands. A dedicated team of ICU clinicians, data scientists, and IT staff put great effort in the medical validation of the data by assessing data quality at several stages in the extensive data validation process.[9] Since raw EHR data contains all routinely and frequently captured clinical information, this whole endeavour eventuated in a unique dataset containing highly granular data of COVID-19 patients throughout their entire ICU stay.

The DDW includes data on demographics, comorbidities, monitoring and life support devices, laboratory results, clinical observations, medications, fluid balance, and outcomes. While the DDW is continuously growing, it contained over 400 million data points from 2052 COVID-19 patients at the moment of accessing.

B   Notation

Throughout the text we associate each observation with an index $i, j, k \in \mathbb{N}$. The covariate (feature) space $\mathcal{X} \subseteq \mathbb{R}^d$ and outcome space $\mathcal{Y} \subseteq \mathbb{R}$ are defined as usual. We denote random variables with capital letters X, Y, Z. We define the covariate vector X, outcome vector Y and treatment indicator T and write X_i, Y_i, T_i, if they correspond to an observation i. We denote realizations of random variables with corresponding lower case letters x, y, z. If a realization of a random variable X is associated with the observation i, then we add a subscript x_i. When not necessary, we omit the corresponding subscripts. When a random variable X follows a distribution p, we will write $\mathbb{E}[X]$ instead of $\mathbb{E}_{x \sim p(x)}[x]$ to denote the expected value.

C   Potential outcome framework and models

C.1   Average Treatment Effect Estimation

We frame our problem of estimating the average treatment effect of proning using Rubin's potential outcome framework.[23] Let X be a vector of observed *covariates* (*features*) and T a

binary *treatment* variable, where $T = 1$ means treated. In the potential outcomes framework we introduce two new random variables Y^1, Y^0 called *potential outcomes*. We call the observed potential outcome the *factual outcome Y* and the unobserved one the *counterfactual outcome Y^**. We assume that the data is generated from a fixed and unknown distribution $p(X, T, (Y^1, Y^0))$ such that the standard *identifiability* criteria hold, namely consistency $Y = TY^1 + (1-T)Y^0$, positivity $p(t \mid x) > 0$, for all x, t and ignorability $\left(T \perp\!\!\!\perp (Y^1, Y^0)\right) \mid X$. In our observational setting, for each sample we observe only the factual outcome, hence the data is of the form $(X_1, T_1, Y_1), (X_2, T_2, Y_2), \ldots, (X_n, T_n, Y_n)$.

We define the average treatment effect as $ATE = \mathbb{E}[Y^1 - Y^0]$ and the conditional average treatment effect (CATE) as $\tau(x) = \mathbb{E}[Y^1 - Y^0 \mid X = x]$. Under the identifiability criteria we can treat observational data as a RCT on covariate values and CATE can be estimated from data as

$$\tau(x) = \mathbb{E}[Y \mid X = x, T = 1] - \mathbb{E}[Y \mid X = x, T = 0].$$

C.1.1 Model Evaluation

In the task of estimating the average treatment effect of proning we are interested in assessing how close our estimate $\widehat{ATE}$ is to the true effect ATE, as measured by the error $\varepsilon_{ATE} = |ATE - \widehat{ATE}|$ or similarly for the conditional effect $\varepsilon_{CATE}(\hat{\tau}) = \mathbb{E}_x[L(\tau(x), \hat{\tau}(x))]$, where L is an arbitrary loss function, $\hat{\tau}$ is the estimated conditional effect and τ is the true conditional effect. In our setting we never observe the true effect, hence we cannot use standard causal inference evaluation metrics ε_{ATE} and ε_{CATE} used on synthetic and semi-synthetic data sets like IHDP.[24] There is no widely established way to evaluate causal inference models on real-world observational data. We can however leverage two other markers to understand whether models are providing sensible estimates.

First, we compare ATEs estimated by our models to the ATE estimated by a randomized control trial that our study emulates. By emulating the RCT's design and by ensuring that the identifiability criteria are satisfied in the observational study, the estimated ATEs should be comparable to the RCT's estimate. In practice there can be some discrepancies, since e.g. the RCT's population can have a different composition from what is present in real-world data, but nonetheless the RCT's estimate can serve as a surrogate for the ground truth effect. Second, as the goal of causal inference methods is not only to provide an estimate close to the RCT's estimate but also to accurately predict the outcomes, we additionally evaluate the predictive performance by calculating the root-mean-squared-error (RMSE) of each method. For each model f, we calculate RMSE for factual outcomes y_i hence $RMSE(f) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i, t_i) - y_i)^2}$.

C.2 Models

We choose models that rely on a diverse set of modeling assumptions, with the goal of estimating the average treatment effect of prone positioning. We introduce traditional methods for causal inference: linear regression and propensity score based models in Section C.2.1 and Section C.2.2, the popular Bayesian regression model BART in Section C.2.3 and a recent deep-learning based

framework in Section C.2.4. The way we choose hyper-parameters for each of the models, as well as details on implementation, is described in Appendix D.

C.2.1 Linear Regression

Linear regression (LR) is the simplest parametric approach to estimating average treatment effect from observational data.[25] We add it to the shortlist of models since it is widely used in the literature, and it is a valid way to adjust for confounders under the following assumption.[25] Linear regression assumes that the conditional expectation of the outcome Y can be expressed as

$$\mathbb{E}[Y \mid X = x, T = t] = \alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot x. \tag{1}$$

We estimate weights in Equation 1 by fitting a linear regression model and use the coefficient of the treatment effect indicator α_1 as the estimated average treatment effect. This is because by Equation 1 it holds that

$$\tau(x) = \mathbb{E}[Y \mid X = x, T = 1] - \mathbb{E}[Y \mid X = x, T = 0] = \alpha_1.$$

Therefore, not only is the outcome Y a linear combination of the treatment T and covariates X, but also the treatment effect is constant across different values of covariates X.

C.2.2 Models Based on Propensity Score

A propensity score, introduced by Rosenbaum and Rubin,[26] is defined as the conditional probability of receiving the treatment $e(x) = \mathbb{P}(T = 1 \mid X = x)$.

In our setting the true propensity score is unknown and needs to be estimated from data using a *propensity score model*. Such model is used to correct the linear regression model in to two ways: blocking and inverse probability weighting (IPW).[27]

First, we use blocking (sometimes referred to as 'stratification') to partition the observational data into subsets (blocks) containing observations with an estimated propensity score in a given range. We estimate the treatment effect within each block by fitting an linear regression to the data within the block. We obtain the final ATE estimate by taking the average of block-wise estimates, weighted by the proportion of data points in each block.

Second, following Kang and colleagues,[28] we define a doubly robust estimator that combines inverse probability weighting with linear regression (DR-IPW) as follows. For each observation with covariate value x and treatment indicator t, we use the estimated propensity score $\hat{e}(x)$ to define a weight

$$w(x, t) = \left[\hat{e}(x)^t \cdot (1 - \hat{e}(x))^{(1-t)} \right]^{-1}.$$

We then fit an linear regression weighted by w to our data, giving more importance to observations with an unlikely treatment assignment. The ATE estimate is the treatment indicator's coefficient of the weighted model, as for the linear regression.

The idea behind both models is to fit an linear regression to a modified version of the original data, using the propensity scores to balance the difference between the treated and control group and to correct the fact that treatment is not independent of covariates. This can be beneficial in a variety of theoretical scenarios (for a detailed discussion of the above models and their statistical properties, see Imbens[27] and Kang et al.[28]).

C.2.3 BART

Bayesian Additive Regression Trees (BART) is a nonparametric, Bayesian method for estimating functions using regression trees.[29] It is a popular method for causal inference,[1] used often as a baseline comparison for state-of-the-art deep learning methods. The model's popularity is based on its outstanding performance in multiple real world scenarios,[30,31] as well as simulated ones.[1]

BART assumes that the conditional expectation of the outcome Y can be expressed as

$$\mathbb{E}[Y \mid X = x, T = t] = g(x,t) + \varepsilon; \quad \varepsilon \sim \mathcal{N}(0,\sigma^2),$$

where g is an arbitrary function, ε is an error term and σ^2 is a constant. BART approximates function g with a sum $\sum_{m=1}^{M} \mathscr{T}_m(x,t)$ of M-many regression trees $\mathscr{T}_1, \ldots, \mathscr{T}_M$. We obtain a CATE estimate by taking the difference $\hat{\tau}(x) = \sum_{m=1}^{M} \left(\mathscr{T}_m(x,1) - \mathscr{T}_m(x,0) \right)$. The ATE estimate is obtained by averaging over CATE estimates computed on the test set $\widehat{ATE} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}(x_i)$.

C.2.4 Counterfactual Regression

Counterfactual Regression (CfR) is a general method for estimating CATE that is based on a theoretical framework developed by Shalit, Johansson, and Sontag.[13] CfR is a regularized minimization procedure that simultaneously tries to predict the correct observed outcomes and to find a representation that balances the treated and control group. Thus CfR fits a representation $\Phi : \mathscr{X} \to \mathscr{X}$ and a regression function $f : \mathscr{X} \times \{0,1\} \to \mathscr{Y}$, where $\mathscr{X}$ is the covariate space and $\mathscr{Y}$ is the outcome space, in order to minimize the objective

$$\min_{f, \Phi: \|\Phi\|=1} \frac{1}{n} \sum_{i=1}^{n} w_i \cdot (f(\Phi(x_i), t_i) - y_i)^2 + \lambda \cdot \mathscr{R}(f)$$
$$+ \alpha \cdot Wass\left(\{\Phi(x_i)\}_{t_i=1}, \{\Phi(x_i)\}_{t_i=0} \right), \tag{2}$$

where $w_i = \frac{1}{2}\left(\frac{t_i}{u} + \frac{1-t_i}{1-u} \right)$ is a weight accounting for the treatment imbalance, $u = \sum_{i=1}^{n} \frac{t_i}{n}$ is the fraction of observations being treated, α and $\lambda \cdot \mathscr{R}(f)$ are arbitrary regularization terms and $Wass$ is the Wasserstein distance defined by Villani.[32] Theorem 3 of Johansson et al.[33] guarantees that by minimizing Equation 2 we obtain f, Φ such that $\hat{\tau}(x) = f(\Phi(x),1) - f(\Phi(x),0)$ is a consistent CATE estimator. We also employ a special case of CfR, called TARNet, whose objective is given by Equation 2 with $\alpha = 0$. We follow CfR's and TARNet's implementation proposed by Shalit, Johansson, and Sontag.[13] We use CfR and TARNet to obtain the ATE estimate by averaging over CATE estimates computed on the test set $\widehat{ATE} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}(x)$.

D Algorithmic Details

D.0.1 Propensity Score Models

We evaluate three different strategies to estimate the propensity score. We fit a logistic regression model to 1) confounders chosen by ICU doctors, 2) all covariates, 3) all covariates with interactions. The second model is preferable to the first as adjusting with the second model results in lower imbalance, when measured by the absolute value of the normalized mean difference defined by Imbens.[11] We choose the second model over the third, despite a slightly lower imbalance, because the propensity scores estimated by the third model tend to extreme values, making it less useful for calculating IPW weights. We use a scikit-learn implementation for the logistic regression with no regularization and with weights adjusting for class imbalance.[34] After estimating the propensity scores (Figure 4) we perform propensity score clipping ($= 0.1$) to account for the lack of overlap in the region with values below 0.1. For the linear regression models we use the implementation proposed in *statsmodels* package.[35]
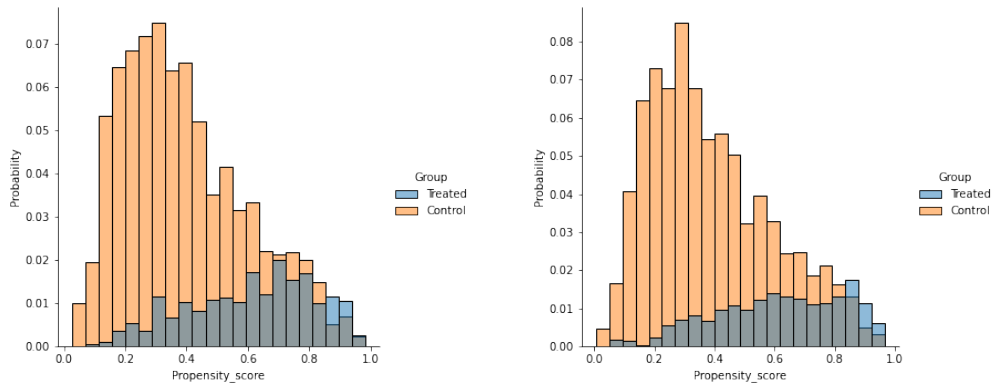
Figure 4: Distribution of the estimated propensity scores for the data set corresponding to the early proning effect (Left) and the late proning effect (Right). The plots show a significant overlap for the estimated propensity scores above 0.1.

D.0.2 BART

We use BART as implemented by Kapelner and Bleich.[36] We train BART with $m = 50$ trees and use 5-fold CV (with respect to the lowest RMSE) on the train data to choose hyperparameters: $k = 2$, $\nu = 3$, $q = 0.9$ controlling the model regularization. These hyperparameters are reported as default by Chipman and colleages.[29] Additionally, we use default parameters $\alpha = 0.95, \beta = 2$ specifying the prior over trees' structure.

D.0.3 CFR and TARNet

We follow CfR's and TARNet's implementation proposed by Shalit, Johansson, and Sontag.[13] They implement CfR and TARNet as a feed-forward neural network with three fully-connected layers and with ELU activation functions for both the representation Φ and for the regression functions $f_1 = f(x, 1)$ and $f_0 = f(x, 0)$. We use default hyperparameters with layer size of 200 for the representation and 100 for outcome networks. The model is trained using Adam optimizer and ℓ_2 weight decay for outcome networks.[37] We choose default regularization hyper-parameters: $\lambda = 0\cdot0001$ and $\alpha = 1$ ($\alpha = 0$ for TARNet). We perform early stopping w.r.t. surrogate mean-squared-error defined by Shalit, Johansson, and Sontag and calculated on a held-out validation set.[13] We perform a grid search evaluated on the validation set in order to access whether to use different than default hyperparameters for TARNet and CfR. We decided to use the default hyperparameters for all models as improvement w.r.t. their predictive performance was negligible.

E Medical Abbreviations

COVID-19: Coronavirus disease 2019
ICU: Intensive care unit
EHR: Electronic health record
RCT: Randomized controlled trial
PaO2: Partial pressure of O2 in an arterial blood sample
FiO2: Fraction of inspired oxygen
P/F ratio: Ratio between PaO2 and FiO2 (PaO2/FiO2)
ARDS: Acute respiratory distress syndrome
PEEP: Positive end-expiratory pressure
COPD: Chronic obstructive pulmonary disease
SOFA score: Sequential organ failure assessment score
BMI: Body mass index
PBW: Predicted body weight
BW: Body weight

F Dutch ICU Data Sharing Against COVID-19 Collaborators

From collaborating hospitals having shared data:

D. Jannet Mehagnoul-Schipper, MD, PhD, Intensive Care, VieCuri Medisch Centrum, Venlo, The Netherlands,

Julia Koeter, MD, Intensive Care, Canisius Wilhelmina Ziekenhuis, Nijmegen, The Netherlands,

Roger van Rietschote, Business Intelligence, Haaglanden MC, Den Haag,The Netherlands,

Thijs C.D. Rettig, MD, PhD, Department of Intensive Care, Amphia Ziekenhuis, Breda, The Netherlands,

M.C. Reuland, MD, Department of Intensive Care Medicine, Amsterdam UMC, Universiteit van Amsterdam, Amsterdam, The Netherlands,

Laura van Manen, MD, Department of Intensive Care, BovenIJ Ziekenhuis, Amsterdam, The Netherlands,

Leon Montenij, MD, PhD, Department of Anesthesiology, Pain Management and Intensive Care, Catharina Ziekenhuis Eindhoven, Eindhoven, The Netherlands,

Jasper van Bommel, MD, PhD, Department of Intensive Care, Erasmus Medical Center, Rotterdam, The Netherlands,

Roy van den Berg, Department of Intensive Care, ETZ Tilburg, Tilburg, The Netherlands,

Ellen van Geest, Department of ICMT, Haga Ziekenhuis, Den Haag, The Netherlands,

Anisa Hana, MD, PhD, Intensive Care, Laurentius Ziekenhuis, Roermond, The Netherlands,

B. van den Bogaard, MD, PhD, ICU, OLVG, Amsterdam, The Netherlands,

Prof. Peter Pickkers, Department of Intensive Care Medicine, Radboud University Medical Centre, Nijmegen, The Netherlands,

Pim van der Heiden, MD, PhD, Intensive Care, Reinier de Graaf Gasthuis, Delft, The Netherlands,

Claudia (C.W.) van Gemeren, MD, Intensive Care, Spaarne Gasthuis, Haarlem en Hoofddorp, The Netherlands,

Arend Jan Meinders, MD, Department of Internal Medicine and Intensive Care, St Antonius Hospital, Nieuwegein, The Netherlands,

Martha de Bruin, MD, Department of Intensive Care, Franciscus Gasthuis and Vlietland, Rotterdam, The Netherlands,

Emma Rademaker, MD, MSc, Department of Intensive Care, UMC Utrecht, Utrecht, The Netherlands,

Martijn de Kruif, MD, PhD, Department of Pulmonology, Zuyderland MC, Heerlen, The Netherlands,

Nicolas Schroten, MD, Intensive Care, Albert Schweitzerziekenhuis, Dordrecht, The Netherlands,

Klaas Sierk Arnold, MD, Anesthesiology, Antonius Ziekenhuis Sneek, Sneek, The Netherlands,

J.W. Fijen, MD, PhD, Department of Intensive Care, Diakonessenhuis Hospital, Utrecht, The Netherland,

Jacomar J.M. van Koesveld, MD, ICU, IJsselland Ziekenhuis, Capelle aan den IJssel, The Netherlands,

Koen S. Simons, MD, PhD, Department of Intensive Care, Jeroen Bosch Ziekenhuis, Den Bosch, The Netherlands,

Joost Labout, MD, PhD, ICU, Maasstad Ziekenhuis Rotterdam, The Netherlands,

Bart van de Gaauw, MD, Martiniziekenhuis, Groningen, The Netherlands,

Michael Kuiper, Intensive Care, Medisch Centrum Leeuwarden, Leeuwarden, The Netherlands,

Albertus Beishuizen, MD, PhD, Department of Intensive Care, Medisch Spectrum Twente, Enschede, The Netherlands,

Dennis Geutjes, Department of Information Technology, Slingeland Ziekenhuis, Doetinchem, The Netherlands,

Johan Lutisan, MD, ICU, WZA, Assen, The Netherlands,

Bart P. Grady, MD, PhD, Department of Intensive Care, Ziekenhuisgroep Twente, Almelo, The Netherlands,

Remko van den Akker, Intensive Care, Adrz, Goes, The Netherlands,

Tom A. Rijpstra, MD, Department of Intensive Care, Amphia Ziekenhuis, Breda, The Netherlands,

Roos Renckens, MD, PhD, Department of Internal Medicine, Northwest Clinics, Alkmaar, the Netherlands,

From collaborating hospitals having signed the data sharing agreement:

Daniël Pretorius, MD, Department of Intensive Care Medicine, Hospital St Jansdal, Harderwijk, The Netherlands,

From the Laboratory for Critical Care Computational Intelligence:

Tariq A. Dam, MD, Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands,

Lucas M. Fleuren, MD, Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands,

Martin E. Haan, MD, Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands,

Mark Hoogendoorn, PhD, Quantitative Data Analytics Group, Department of Computer Science,

Faculty of Science, VU University, Amsterdam, The Netherlands,

Armand R.J. Girbes, MD, PhD, EDIC, Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands,

Dagmar M. Ouweneel, PhD, Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands,

Ronald Driessen, Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands,

Jan Peppink, Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands,

H.J. de Grooth, MD, PhD, Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands,

From Pacmed:

Michele Tonutti, MRes, Pacmed, Amsterdam, The Netherlands,

Daan P. de Bruin, MSc, Pacmed, Amsterdam, The Netherlands,

Sebastiaan J.J. Vonk, MSc, Pacmed, Amsterdam, The Netherlands,

Mattia Fornasa, PhD, Pacmed, Amsterdam, The Netherlands,

Tomas Machado, Pacmed, Amsterdam, The Netherlands,

Michael de Neree tot Babberich, Pacmed, Amsterdam, The Netherlands,

Olivier Thijssens, MSc, Pacmed, Amsterdam, The Netherlands,

Lot Wagemakers, Pacmed, Amsterdam, The Netherlands,

Hilde G.A. van der Pol, Pacmed, Amsterdam, The Netherlands,

Tom Hendriks, Pacmed, Amsterdam, The Netherlands,

Julie Berend, Pacmed, Amsterdam, The Netherlands,

Virginia Ceni Silva, Pacmed, Amsterdam, The Netherlands,

Robert F.J. Kullberg, MD, Pacmed, Amsterdam, The Netherlands,

Taco Houwert, MSc, Pacmed, Amsterdam, The Netherlands,

Hidde Hovenkamp, MSc, Pacmed, Amsterdam, The Netherlands,

Roberto Noorduijn Londono, MSc, Pacmed, Amsterdam, The Netherlands,

Davide Quintarelli, MSc, Pacmed, Amsterdam, The Netherlands,

Martijn G. Scholtemeijer, MD, Pacmed, Amsterdam, The Netherlands,

Aletta A. de Beer, MSc, Pacmed, Amsterdam, The Netherlands,

Willem E. Herter, BSc, Pacmed, Amsterdam, The Netherlands,

From RCCnet:

Leo Heunks, MD, PhD, Department of Intensive Care Medicine, Amsterdam Medical Data
Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands,

Nicole Juffermans, MD, PhD, ICU, OLVG, Amsterdam, The Netherlands,

Arjen J.C. Slooter, MD, PhD, Department of Intensive Care Medicine, UMC Utrecht, Utrecht
University, Utrecht, the Netherlands,

From other collaborating partners:

Martijn Beudel, MD, PhD, Department of Neurology, Amsterdam UMC, Universiteit van
Amsterdam, Amsterdam, The Netherlands,

Table 1: **Characteristics of observations that Fulfilled the Inclusion Criteria.** We report mean values with standard deviation for numerical variables and frequency for binary variables, ignoring missing values. Column Abs. Norm. Diff. indicates the absolute value of the normalized difference between prone and supine group, as defined by Imbens[11] – higher values indicate higher imbalance.

Characteristic	Prone Sessions (N=1260)	Supine Sessions (N=5111)	Abs. Norm. Diff.
Age – yr.	$63{\cdot}6 \pm 10{\cdot}6$	$63{\cdot}5 \pm 10{\cdot}3$	$0{\cdot}01$
Male sex – %	73%	77%	$0{\cdot}09$
BMI	$28{\cdot}3 \pm 5{\cdot}2$	$28{\cdot}8 \pm 5{\cdot}6$	$0{\cdot}09$
SOFA Score	$7{\cdot}5 \pm 3{\cdot}2$	$8{\cdot}0 \pm 3{\cdot}4$	$0{\cdot}15$
Diabetes – %	25%	25%	$0{\cdot}00$
Renal failure – %	19%	20%	$0{\cdot}03$
Hepatic disease – %	5%	6%	$0{\cdot}06$
Coronary artery disease – %	5%	7%	$0{\cdot}09$
Cancer – %	8%	9%	$0{\cdot}05$
COPD – %	19%	16%	$0{\cdot}09$
Immunodeficiency – %	14%	15%	$0{\cdot}02$
Morbid Obesity – %	9%	13%	$0{\cdot}12$
Vasopressors – %	64%	56%	$0{\cdot}17$
Neuromuscular blockers – %	46%	25%	$0{\cdot}43$
Renal–replacement therapy – %	7%	11%	$0{\cdot}14$
Glucocorticoids – %	5%	10%	$0{\cdot}19$
Tidal volume – ml	$454{\cdot}7 \pm 124{\cdot}0$	$473{\cdot}3 \pm 134{\cdot}4$	$0{\cdot}14$
Tidal volume – ml per kg of PBW	$6{\cdot}7 \pm 1{\cdot}8$	$6{\cdot}9 \pm 1{\cdot}9$	$0{\cdot}11$
Respiratory rate	$23{\cdot}2 \pm 8{\cdot}7$	$24{\cdot}3 \pm 9{\cdot}2$	$0{\cdot}12$
PEEP (cm H_2O)	$13{\cdot}1 \pm 3{\cdot}6$	$12{\cdot}4 \pm 3{\cdot}7$	$0{\cdot}19$
FiO$_2$ – %	$79{\cdot}0 \pm 14{\cdot}8$	$71{\cdot}0 \pm 12{\cdot}5$	$0{\cdot}58$
Plateau pressure – cm H_2O	$25{\cdot}8 \pm 6{\cdot}6$	$26{\cdot}7 \pm 6{\cdot}7$	$0{\cdot}14$
Driving pressure – cm H_2O	$13{\cdot}8 \pm 5{\cdot}6$	$14{\cdot}0 \pm 5{\cdot}5$	$0{\cdot}04$
PaO$_2$ – mm Hg	$69{\cdot}4 \pm 12{\cdot}8$	$71{\cdot}5 \pm 11{\cdot}7$	$0{\cdot}17$
P/F ratio – mm Hg	$91{\cdot}0 \pm 22{\cdot}9$	$103{\cdot}3 \pm 22{\cdot}0$	$0{\cdot}55$
PaCO$_2$ – mm Hg	$56{\cdot}7 \pm 15{\cdot}2$	$56{\cdot}6 \pm 16{\cdot}4$	$0{\cdot}01$
Arterial pH	$7{\cdot}3 \pm 0{\cdot}1$	$7{\cdot}4 \pm 0{\cdot}1$	$1{\cdot}00$
Lung compliance static	$42{\cdot}7 \pm 31{\cdot}0$	$43{\cdot}4 \pm 32{\cdot}0$	$0{\cdot}02$

Table 2: **ATE estimates for early and late proning effect.** All models estimate an ATE that approximates or is slightly higher than the ATE of 15 reported by the PROSEVA trial. ATE and RMSE are reported with mean and 95% CIs, for both outcomes.

| Model | Early Proning Effect | | Late Proning Effect | |
	ATE	RMSE	ATE	RMSE
LR	15·31(11·69, 19·80)	58·48(57·91, 59·11)	14·47(10·34, 19·34)	53·88(53·41, 54·58)
Weighting	14·54 (10·29, 19·00)	59·09 (58·15, 60·10)	13·53 (8·88, 19·45)	54·14 (53·58, 54·96)
Blocking	15·59 (11·44, 20·29)	60·02 (58·27, 61·71)	14·87 (9·39, 19·34)	53·22 (51·88, 55·14)
BART	20·11 (14·17, 27·50)	48·62 (45·12, 56·81)	13·99 (6·70, 20·34)	50·44 (47·35, 55·40)
TARNet	17·70 (8·80, 25·60)	51·79 (50·74, 53·53)	15·13 (5·66 25·97)	50·61 (48·80, 52·01)
CfR	18·14 (9·28, 27·35)	51·82 (50·83, 53·52)	15·26 (5·54, 24·94)	50·59 (48·58, 51·54)