PLOS ONE

RESEARCH ARTICLE

Variable Importance and Prediction Methods for Longitudinal Problems with Missing Variables

Iván Díaz[1]*, Alan Hubbard[2], Anna Decker[2], Mitchell Cohen[3]

1 Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, **2** Division of Biostatistics, University of California, Berkeley, CA, USA, **3** Department of Surgery, University of California San Francisco, San Francisco, CA, USA

* idiaz@jhu.edu

Citation: Díaz I, Hubbard A, Decker A, Cohen M (2015) Variable Importance and Prediction Methods for Longitudinal Problems with Missing Variables. PLoS ONE 10(3): e0120031. doi:10.1371/journal.pone.0120031

Academic Editor: Kewei Chen, Banner Alzheimer's Institute, UNITED STATES

Received: July 21, 2014

Accepted: January 23, 2015

Published: March 27, 2015

Data Availability Statement: The dataset used for this submission contains Personal Health Information (PHI) and can be linked to specific subjects, so this data could not be shared publicly. The raw data is available to to interested parties upon request, pending ethical approval. Please contact Mitchell Cohen at mcohen@sfghsurg.ucsf.edu to request access to the data.

Funding: Funding provided by PROMMTT - Prospective Observational Multi-Center Transfusion Study (UCSF). http://cetir-tmc.org/research/prommtt. The funders had no role in study design, data

Abstract

We present prediction and variable importance (VIM) methods for longitudinal data sets containing continuous and binary exposures subject to missingness. We demonstrate the use of these methods for prognosis of medical outcomes of severe trauma patients, a field in which current medical practice involves rules of thumb and scoring methods that only use a few variables and ignore the dynamic and high-dimensional nature of trauma recovery. Well-principled prediction and VIM methods can provide a tool to make care decisions informed by the high-dimensional patient's physiological and clinical history. Our VIM parameters are analogous to slope coefficients in adjusted regressions, but are not dependent on a specific statistical model, nor require a certain functional form of the prediction regression to be estimated. In addition, they can be causally interpreted under causal and statistical assumptions as the expected outcome under time-specific clinical interventions, related to changes in the mean of the outcome if each individual experiences a specified change in the variable (keeping other variables in the model fixed). Better yet, the targeted MLE used is doubly robust and locally efficient. Because the proposed VIM does not constrain the prediction model fit, we use a very flexible ensemble learner (the SuperLearner), which returns a linear combination of a list of user-given algorithms. Not only is such a prediction algorithm intuitive appealing, it has theoretical justification as being asymptotically equivalent to the oracle selector. The results of the analysis show effects whose size and significance would have been not been found using a parametric approach (such as stepwise regression or LASSO). In addition, the procedure is even more compelling as the predictor on which it is based showed significant improvements in cross-validated fit, for instance area under the curve (AUC) for a receiver-operator curve (ROC). Thus, given that 1) our VIM applies to any model fitting procedure, 2) under assumptions has meaningful clinical (causal) interpretations and 3) has asymptotic (influence-curve) based robust inference, it provides a compelling alternative to existing methods for estimating variable importance in high-dimensional clinical (or other) data.

collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Modern medical care is awash in a sea of data. The advent of new monitors, better diagnostics, electronic medical record keeping and the ideal of the quantified self has resulted in patients who are more completely measured than at any other time in medical history. While purported to allow for more complete evaluation and diagnosis our current data intense medical environment often leaves clinicians overwhelmed by the dimensionality and quantity of data. Indeed, despite the current ability to continuously measure and track multiple physiologic, demographic and biologic variables measure a patient's clinical history in detail, clinicians still make care decisions based on a few non relational variables (they can keep in their) head combined with rules of thumb and clinical gestalt. Medical decisions therefore fail to take into consideration the possible intricate relations between all the patient's underlying factors squandering data and missing important prognostic relationships. Fortunately, advances in the fields of biostatistics and bioinformatics have developed mathematical and computational tools that can allow optimal care decisions based on the entirety of the patient's data, which are beyond the computational ability of the clinician at the bedside.

A primary goal in evidence-based medicine is to design and implement prognosis tools that take into account an extremely large set of measured characteristics in order to predict a patient's most likely medical outcome. An equally important goal is to establish at each given time point which of these numerous measured characteristics is decisive in the development of the predicted outcome. These two goals have been traditionally called prediction and variable importance analysis, respectively. In addition to understanding the underlying biological mechanisms related to positive medical outcomes, the joint use of these tools can help doctors devise the optimal treatment plan according to the specific characteristics of the subject, simultaneously taking into account hundreds of variables collected for each patient.

Because of the large number of variables and the complexity of the relations between them, prediction and variable importance would be impossible to achieve without the use of complex statistical algorithms accompanied by powerful computers able to carry out a large number of computations. Such methods are aimed at helping doctors make the better treatment decisions using real time data.

Prediction and variable importance are different goals whose optimal achievement requires the use of different tools. The objective of a prediction algorithm is to accurately predict the outcome, whereas the objective for estimating variable importance measures (VIM) is to estimate the degree to which changes in the outcome are caused by changes in each of the predictor variables; VIM's optimally supply doctors with tools for making treatment decisions [1]. This difference between prediction and VIM has two main consequences. First, VIM problems are of a *causal* nature, whereas prediction problems do not try to distinguish the relative contribution of variables to the variation observed in an outcome. Second, in order to help the decision making process, VIM parameters should have, as much as possible, a clinically relevant interpretation, such as the expected change in the outcome under a given intervention. As explained below, a meaningful interpretation can only be obtained through an intelligible characterization of VIM as a statistical (or causal) parameter defined as a mapping from a tenable (thus usually very large) statistical model into simple low-dimensional parameter(s) (a Euclidean space, e.g., differences in "adjusted" means).

Current practice in biostatistics and bioinformatics involves the use of machine learning algorithms for prediction. However, measures of VIM have been via estimated coefficients, for instance, in a penalized regression (e.g., LASSO [2]). Alternatively, measures of VIM (such as returned by random forest or neural nets [3–6] are not necessarily specific to a particular model fitting routine, but are related to the added fit (or reduced estimated risk) of a "full"

model versus one without the variable. Those based on procedures like LASSO are based on arbitrary statistical models (e.g., linear with main effects), so the model is likely to misspecified and thus the interpretation of the resulting coefficients problematic. On the other hand, consider the case of regression and classification trees (e.g., random forests), where the VIM for a variable X is defined as the difference between prediction error when X is perturbed versus the prediction error otherwise [3,7]. Though such a measure does not require a relatively simple parametric model, the clinical relevance of this quantity as a measure of VIM is unclear because: 1) it does not represent a statistical or causal parameter, 2) it does not have a potential interpretation in terms of the mechanistic process that generates the data, and thus 3) the translation of the estimate into the impact of potential intervention is problematic. As an example of the technical difficulties arising from this practice, [6] discuss the "bias" of random forest VIM measures, missing the fact that bias can only be defined in terms of a statistical parameter, which is never specified in the random forest VIM analysis. We present a new VIM that 1) puts no constraints on the statistical model, so that the resulting estimate is relatively unbiased, 2) has an interpretation that is analogous to a "slope" (change in a mean for a change in the variable of interest, and 3) under further assumptions has a "causal" interpretation of great clinical relevance. Thus, it provides the virtue of more accessible regression coefficients, but is appropriate for the new powerful machine learning technology for creating more accurate diagnostic predictors.

Previous work has taken the approach we outline here [1, 8]. Whereas that work accomplished the goals of a causal VIM in a semiparametric model, the predictors were required to be discrete, preferably binary. Defining VIM parameters in terms of potential causal associations for "continuous" variables poses additional technical challenges. When researchers using causal inference methods are faced with exposures of continuous nature, the most common approach is to dichotomize the continuous exposure and consider the effect of its binary version on the outcome. This approach suffers from various flaws. First, the causal parameter does not answer questions about plausible modifications to the data generating mechanism. For example, the additive causal effect of a dichotomized exposure compares an intervention by which the exposure is truncated below the dichotomization threshold with an intervention by which the exposure is truncated above it [9]. In addition, for most of the variables we discuss below, such interventions are not practical, so the resulting estimates too far removed from practicality.

In this paper we present variable importance measures that can be used to rank a list of both continuous and binary variables in terms of their importance on an outcome. The approach we present differs from other approaches in the literature [1] in that we also address estimation of VIM for continuous variables, as opposed to only binary ones. We use state-of-the-art machine learning coupled with causal inference methods to address VIM estimation and illustrate the use of our methods for analyzing the determinants of death in severe trauma patients.

The paper is organized as follows. We start by describing the problem that motivates the development of the tools presented in this paper, and then move to an introduction of the causal inference tools used to define the variable importance measures. We then present various estimators for the variable importance parameters previously defined and discuss their asymptotic properties; we also briefly describe the super learner [10], an ensemble learner whose asymptotic performance is optimal for prediction. We finish by presenting the details of the data analysis and some concluding remarks.

Background

Trauma is the leading cause of death between the ages of 1 and 44. The vast majority of these deaths take place quickly and much of the initial resuscitative and decision-making action takes place in the first minutes to hours after injury [11,12]. In addition, it is clear that as patients progress through their initial resuscitation, the relative attention paid to different physiologic and biologic parameters and indeed the interventions themselves are dynamic. Different variables are important and drive future outcome in the first few minutes after injury than at 24 hours when a patient has survived long enough to receive large volume resuscitation, operative intervention and ICU care. While these dynamics are intuitive, most practitioners do not have the ability to know which variables are important at any given time point. As a result, often the same vital signs and markers are followed throughout the patient's hospital course independent of whether they are currently relevant. This results in practitioners who are often left making care decisions without knowledge of the current patient physiologic state and which parameters are important at that moment. Left with this uncertainty and awash in constantly evolving multivariate data, practitioners make decisions based on clinical gestalt, a few favorite variables, and rules of thumb developed from clinical experience. To aid in prediction, the medical literature is filled with scoring systems and published associations between these variables (physiology, biomarker, demographic, etc.) and outcomes of interest [13–17]. While numerous, these published statistical associations, given the reported methodology, often report misspecified and overfit models. In addition most of these statistical predictive models do not account for the rapidly changing dynamics of a severely injured patient, and fail to take into account the statistical issues discussed in the previous paragraphs. An ideal system would mimic the clinical decision making of an experienced practitioner by providing dynamic prediction (changing prediction at iterative time points) while evaluating the dynamic importance of each variable over time [18]. This then would mimic the implicit understanding a clinician brings to a patient where it is clear that the necessary focus of care must change over time.

Trauma Data

The data that motivated the methods developed in this paper were collected as part of the Activation of Coagulation and Inflammation in Trauma (ACIT, see e.g., [19–21]) study, which is a prospective cohort study of severe trauma patients admitted to a single level 1 trauma center. Several physiological and clinical measurements were recorded at several time points for each patient after arrival to the emergency room. These variables include demographic variables (e.g., age, gender, etc.), baseline risk factors (e.g., asthma, chronic lung disease, Glasgow coma scale, diabetes, injury mechanism, injury severity score, etc.), longitudinally measured variables that account for the patient's treatment and health status history (e.g., respiratory and heart rate, platelets, coagulation measures like prothrombin time and INR, activated protein C, etc.), and an indicator of the occurrence of death at each time interval. Because these data are often collected in a high-stress environment, it is common that some variables are missing for some patients at a given time point. The list of variables we analyzed is presented in S1 Table Table of the Supporting Information.

Classical Regression Approaches

In order to estimate the effect of a variable A on an outcome Y controlling for a set of variables W, it is common practice among data analysts to estimate the parameter β in a parametric regression model $E(Y|A,W) = m(A,W|\beta)$ for a known function m, for example,

$$E(Y|A, W) = \beta_0 + \beta_1 A + \beta_2 W. \quad (1)$$

Under model (1), the estimate of β_1 is interpreted as the expected change in Y given a change of one unit in A:

$$\beta_1 = E\{E(Y|A+1, W) - E(Y|A, W)\}. \quad (2)$$

Small violations to the assumptions of model (1) (e.g., an interaction term) would yield an estimate of β_1 that cannot be interpreted as in (2). Therefore, in this paper we define parameters in terms of characteristics of the probability distribution of the data under a non-parametric model, as in Equation (2). This practice allows the definition of the parameter of interest independently of (possibly) misspecified parametric models, and avoids dealing with different interpretations of regression parameters under incorrect model specifications.

The causal interpretation of statistical parameters (e.g., Equation (2)) requires additional untestable assumptions about the distribution of counterfactual outcomes under hypothetical interventions. Such counterfactual outcomes may be defined using the potential outcomes framework of a structural equation model (NPSEM [22]). In the remaining of the section we describe the observed data, and define the variable importance measures using an NPSEM. If the assumptions encoded in the NPSEM do not hold, the estimates do not have a causal interpretation and must not be used to make treatment decisions. In that case, there are two main uses of these estimates. First, they can be used as tools for determining the best set of predictors variables by ruling out those whose with a zero non-significant variable importance. Second, they may be used as a tool for formulating causal hypothesis that may be tested in a subsequent randomized study or in an observational study in which the necessary causal assumptions are met.

Methods

Notation

Assume that observations on each patient are recorded at times $t_0,t_1,\ldots,t_J$, where $t_0 = 0$, and let T denote the time of death of a patient. The observed data for each patient is given by the random variable

$$O = (L_0, C_1, L_1, Y_1, \ldots, C_J, L_J, Y_J),$$

where L_0 denotes a set of baseline variables recorded at admission to the hospital, $L_j = (L_{j1},\ldots, L_{jK})$ denotes a set of variables measured at time t_j, $C_j = (C_{j1},\ldots,C_{jK})$ where C_{jk} denotes an indicator of missingness of L_{jk}, and $Y_j = I(t_j < T \leq t_{j+1})$ denotes an indicator of death occurring in the interval $(t_j,t_{j+1}]$, for $j = 0,\ldots,J-1$. Once death occurs the random variables in the remaining time points of the vector O become degenerate so that this structure is well defined.

For the analysis of the ACIT data we have classified the variables L_{jk} in two non-mutually exclusive categories: baseline and treatment variables. Baseline variables (L_0) are causally related to the outcome but can seldom be manipulated by the physician and are rarely of interest as possible care targets. Although baseline variables are not of interest in themselves, controlling for them is crucial when estimating the effect of treatment variables, which are often longitudinal variables that represent possible targets for clinical care. The label of each variable according to this classification is shown in S1 Table Table of the Supporting Information.

We define VIM measures in terms of the effect of L_{jk} on $Y_{j'}$, for all $j' \geq j$ and for all k. That is, we are interested in importance of a variable recorded at time point t_j on the hazard of death in each of the subsequent time intervals $(t_j,t_{j+1}],\ldots,(t_{J-1},t_J]$. This approach has the advantage that VIM can be seen as a dynamic process in which the factors that are decisive for developing/predicting a clinical outcome change as a function of time.

Causal model

We encode the assumptions necessary to make causal claims in terms of the non-parametric structural equation model [22]:

$$\begin{aligned} L_0 &= f_{L_0}(U_{L_0}) \\ C_{jk} &= f_{C_{jk}}(C_{j-1}, L_{j-1}, L_0, U_{C_j}) \qquad j = 1, \ldots, J;\ k = 1, \ldots, K \\ L_{jk} &= C_{jk} f_{L_{jk}}(C_{j-1}, L_{j-1}, L_0, U_{L_j}) \qquad j = 1, \ldots, J;\ k = 1, \ldots, K \\ Y_j &= f_{Y_j}(\bar{C}_j, \bar{L}_j, L_0, U_{Y_j}) \qquad j = 1, \ldots, J, \end{aligned} \tag{3}$$

where, for a random variable X, f_X denotes an unknown but fixed function, U_X denotes all the unmeasured factors that are causally related to X, and $\bar{X}_j = (X_1, \ldots, X_j)$ denotes the history of X up until time t_j. As pointed out by [22], this model assumes that the data O for each patient are generated by the mechanistic process implied by the functions f_{X_j} with a temporal order dictated by the ordering of the time points t_j. In addition, this NPSEM encodes two important conditional independence assumptions:

$$L_{jk} \perp\!\!\!\perp L_{jk^*} | (L_0, L_{j-1}) \quad \forall j,\ k^* \neq k, \tag{4}$$

$$L_{jk} \perp\!\!\!\perp \bar{L}_{j-2} | (L_0, L_{j-1}) \quad \forall j,\ k. \tag{5}$$

Assumption (4) means that the variables L_{jk} at time t_j are drawn simultaneously as a function of the past only, and that, given the past, contemporary variables do not interact with each other. This is an assumption that must be taken with caution since the causal structure of the relation between contemporary measurements of physiological characteristics of trauma patients is not well understood yet. Assumption (5) is a standard Markov independence assumption stating that L_{jk} is independent on covariate past conditional on the most recent measurements L_{j-1} and baseline L_0.

As a consequence of these assumptions, the problem of estimating the causal effect of each L_{jk} on each $Y_{j'}$ for $j' \geq j$ can be seen as a series of cross-sectional problems as follows. Note that L_{jk^*} for $k^* \neq k$ are not confounders of the causal relation between L_{jk} and $Y_{j'}$. To illustrate this, consider the NPSEM encoded in the directed acyclic graph of Fig. 1, in which for simplicity we assume that all covariates are observed (i.e., C variables are not present) and that $J = K = 2$. It stems from the graph that the variable L_{22} plays no role as a confounder of the causal effect of L_{21} on Y_2. The problem of variable importance for these data can thus be transformed into a series of cross-sectional problems as follows. For each patient still at risk at $t_{j'}$, denote

$$\begin{aligned} A &\equiv L_{jk} \\ C &\equiv C_{jk} \\ W &\equiv (L_0, C_{j-1}, L_{j-1}) \\ Y &\equiv Y_{j'}, \end{aligned} \tag{6}$$

and

$$\bar{Q}(A, C, W) \equiv E(Y|A, C, W),\ g(A|C, W) \equiv P(A|C, W)$$
$$\phi(C|W) \equiv P(C|W),\ Q_W(W) \equiv P(W).$$

Without loss of generality we assume that the variable A is either binary or continuous in the interval (0,1). For fixed $j, j' \geq j$, and k, and for each patient still at risk at $t_{j'}$, using the notation

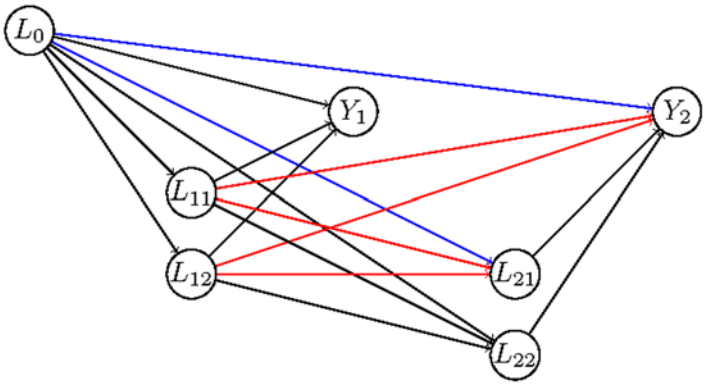

Fig 1. Directed acyclic graph, the arrows in blue and red denote the relations that confound the causal effect of L_{21} on Y_2.

doi:10.1371/journal.pone.0120031.g001

introduced in (6), it suffices to consider the following simplified NPSEM:

$$W = f_W(U_W),\ \ C = f_C(W, U_C),\ \ A = Cf_A(W, U_A),\ \ Y = f_Y(A, C, W, U_Y), \tag{7}$$

where the U variables denote all the exogenous, unobserved factors associated to each of the observed variables, and the functions f are deterministic but unknown and completely unspecified. Some additional consequences of NPSEM (7) are:

1. The missingness indicator C is allowed to depend on the covariates measured in the previous time point. In this way we take into account that a variable can be missing as a result of the previous health status of the patient, and also that it can be correlated with previous missingness indicators. Other approaches to handle missing data such as multiple imputation may be used here, each requiring their own assumptions. However, we favor this approach because i) it is embedded within the same causal inference framework in which we define the VIM parameters, and ii) the validity of the results does not depend on the correct specification of a multiple imputation model.

2. Missingness is informative. A patient's missingness indicator C is allowed to affect the way Y is generated, therefore acknowledging that missingness can contain information about the health outcome (e.g., sicker patients who die earlier might be more likely to have missing values because information stops being recorded during life-threatening situations).

We now define the variable importance for continuous and binary outcomes separately.

Continuous Variables. Consider an intervened system in which the variables are generated by the following system of equations

$$\begin{aligned} W &= f_W(U_W) \\ C^I &= 1 \\ A^I &= f_A(W, U_A) + \delta \\ Y^I &= f_Y(A^I, C^I, W, U_Y), \end{aligned} \tag{8}$$

which, for a small positive δ, can be interpreted as a model in which there is no missingness, and the distribution of the exposure variable A^I is shifted to the right by δ units. This type of intervention has been previously discussed in the literature [23], and belongs to a wider class of interventions known as stochastic interventions [24–26]. The parameter $E(Y^I)$–$E(Y)$ can be causally interpreted as the expected reduction in mortality rate gained by an increase of δ units

in the variable A for each patient. Since the counterfactual data $O^I = (W,C^I,A^I,Y^I)$ are not observed, $E(Y^I)$ is not estimable without further untestable assumptions. Under the randomization assumption [22, 27]

$$(C, A) \perp Y^I | W, \tag{9}$$

and the positivity assumption

$$g_0(A|W) > 0, \quad \text{and} \quad \phi_0(1|W) > 0 \quad \text{for all } A \text{ and } W, \tag{10}$$

the expectation $E(Y^I)$ is identified as $E(Y^I) = E_W E\{\bar{Q}(A+\delta, C, W)|C=1, W\}$, and the parameter of interest is defined as

$$\Psi_c(\bar{Q}, Q_W, g) \equiv E_W E_g\{\bar{Q}(A+\delta, 1, W)|C=1, W\} - E(Y). \tag{11}$$

A proof of this result under the randomization assumption is presented in [23]. That proof follows the arguments for identification of general causal parameters given in [22], who provides a unified framework for identification of counterfactual parameters as function of the observed data generating mechanism.

Binary Variables. For binary variables, following the structural causal model described in (7), the VIM parameter is defined according to the following intervened system:

$$\begin{aligned} W &= f_W(U_W) \\ C^I &= 1 \\ A^I &= \begin{cases} 1 & \text{with probability } g(1|1, W) + \delta \\ 0 & \text{with probability } g(0|1, W) - \delta \end{cases} \\ Y^I &= f_Y(A^I, C^I, W, U_Y), \end{aligned}$$

where $0 < \delta < \sup_w g(0|1,w)$ is a user-given value. Under randomization assumption (9), and the positivity assumption

$$0 < g_0(1|W) < 1, \quad \text{and} \quad \phi_0(1|W) > 0 \quad \text{for all } W, \tag{12}$$

the expectation of Y^I is identified as a function of the observed data generating mechanism as $E(Y^I) = E_W E\{\bar{Q}(A, C, W)|C=1, W\} + \delta\{E[\bar{Q}(1,1,W) - \bar{Q}(1,0,W)]\}$, and the parameter of interest is defined as

$$\begin{aligned} \Psi_b(\bar{Q}, Q_W, g) \equiv E_W E_g\{\bar{Q}(A, C, W)|C=1, W\} \\ + \delta\{E[\bar{Q}(1,1,W) - \bar{Q}(1,0,W)]\} - E(Y), \end{aligned} \tag{13}$$

The true values of Ψ_c and Ψ_b will be denoted $\psi_{c,0}$ and $\psi_{b,0}$, respectively.

Comparability. Since the previous variable importance will be used to provide a ranking of the variables mixing continuous and binary ones, we provide a heuristic argument that they are comparable up to first order. First of all, note that, under the appropriate differentiability assumptions, for continuous A we have

$$\begin{aligned} \Psi_c(\bar{Q}, Q_W, g) \approx E_W\{\bar{Q}(A, 1, W)|C=1, W\} \\ + \delta \frac{d}{d\delta} E_W E\{\bar{Q}(A+\delta, 1, W)|C=1, W\}|_{\delta=0}. \end{aligned} \tag{14}$$

This expression and (13) both have the form $a+\delta \times b$, where b can be seen as the appropriate slope of $E\{\bar{Q}(A, C, W)\}$ as a function of A.

To illustrate this, let us consider the following scenarios. First, consider a continuous variable A_1 distributed as $Beta(2,2)\times 0.8+0.1$, and a binary variable A_2 distributed as $Ber(A_1)$, where $Ber(p)$ is the Bernoulli distribution with probability p. If the outcome is continuous, with no missingness, and $\bar{Q}(A_1, A_2) = A_1 + A_2$, it is straightforward to see that $\psi_{c,0} = \psi_{b,0} = \delta$, for any $0 < \delta < 0.1$. Equality of these parameters is a consequence of the comparability argument above, and the fact that the approximation (14) is exact for linear $\bar{Q}$. Consider now a nonlinear case, with Y binary distributed as $Ber(\text{expit}(A_1+A_2))$. In this case, Monte Carlo computation of the integrals involved yields $\psi_{c,0} = 0.0019$ and $\psi_{b,0} = 0.0020$ for $\delta = 0.01$, providing an example of the comparability argument described above.

In the following sections we discuss doubly robust estimation methods for these parameters for continuous and binary variables.

VIM estimation

In order to define semi-parametric VIM estimates that have optimal asymptotic properties we first need to talk about the efficient influence function. The efficient influence function is a known function D of the data O and P_0, and is a key element in semi-parametric efficient estimation, since it defines the linear approximation of all efficient regular asymptotically linear estimators [28]. This means that the variance of the efficient influence function provides a lower bound for the variance of all regular asymptotically linear estimators, analogously to the Cramer-Rao lower bound in parametric models. The efficient influence functions of parameters Ψ_c and Ψ_b are presented in Section S1 TMLE Algorithm of the Supporting Information.

We use targeted minimum loss based estimators (TMLE)[29,30] of the parameters Ψ_c and Ψ_b. TMLE is a substitution/plug-in estimation method that, given initial estimators $(\bar{Q}_n, Q_{W,n}, g_n)$ of $(\bar{Q}, Q_W, g)$, finds updated estimators $(\bar{Q}^*_n, Q^*_{W,n}, g^*_n)$ to define the estimator of Ψ as

$$\psi_n = \Psi(\bar{Q}^*_n, Q^*_{W,n}, g^*_n).$$

TMLE is an estimation method that enjoys the best properties of both G-computation estimators [31] and the estimating equation methodology (see e.g., [32, 33]). On one hand, TMLE is similar to G-computation estimators (e.g., $\Psi(\bar{Q}_n, Q_{W,n}, g_n)$) in that it is a plug-in estimator, and therefore produces estimates that are always within the range of the parameter of interest (e.g., it is always in the interval [0, 1] if the estimand is a probability). On the other hand, under regularity conditions and consistency of $(\bar{Q}_n, g_n, \phi_n)$, it is asymptotically linear with influence function equal to the efficient influence function:

$$\psi_n - \psi_0 = \sum_{i=1}^{n} D(P_0)(O_i) + o_p(1/\sqrt{n}).$$

As a consequence, TMLE has the following properties:

- It respects the known bounds of the target parameter.
- It is efficient if $\bar{Q}_n, g_n$, and ϕ_n are consistent for $\bar{Q}_0, g_0$, and ϕ_0, respectively.
- It is consistent if either $\bar{Q}_n$ or both g_n and ϕ_n are consistent. This property is referred to as double robustness.
- It is more robust to empirical violations of the positivity assumptions (10) and (12).

In Section S1 TMLE Algorithm of the Supporting Information we describe an iterative procedure that transforms the initial estimates $\bar{Q}_n$ and g_n into targeted estimates $\bar{Q}^*_n$ and g^*_n such that

$\Psi(\bar{Q}_n^*, g_n^*, Q_{W,n}^*)$ is a TMLE of $\Psi(\bar{Q}_0, g_0, Q_{W,0})$, and discuss in more detail the properties of the TMLE.

Estimating equation (EE), Gcomp/IPMW, and unadjusted estimators. For comparison, we compute three additional estimates of the VIM. The first estimator, based on the estimating equation (EE) methodology, is an estimator that uses the efficient influence function of the parameter in order to define the estimator as the solution of the corresponding estimating equation. Because the EE is also asymptotically linear with influence function equal to the efficient influence function, it is consistent and asymptotically efficient under regularity and consistency conditions on $(\bar{Q}_n, Q_{W,n}, g_n)$. However, the estimating equation that defines the EE may not have a solution in the parameter space, in which case the EE does not exist. The second estimator, a mixture of the G-computation formula and the inverse probability of missingness weighted estimator IPMW (Gcomp/IPMW) represents a choice that could have been made in common practice in statistics. The Gcomp/IPMW estimator uses initial estimators ϕ_n and $\bar{Q}_n$ of ϕ_0 and $\bar{Q}_0$ obtained through step-wise regression, and is defined as

$$\begin{aligned}\psi_{c,n,GI} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{C_i}{\phi_n(W_i)}\bar{Q}_n(A_i+\delta,1,W_i)-Y_i\right\}\\ \psi_{b,n,GI} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{C_i}{\phi_n(W_i)}\bar{Q}_n(A_i,1,W_i)+\delta[\bar{Q}_n(1,1,W_i)-\bar{Q}_n(0,1,W_i)]-Y_i\right\},\end{aligned}$$

for Ψ_c and Ψ_b, respectively. This estimator is consistent only if both the model for ϕ_0 and the model for $\bar{Q}_0$ have been correctly specified. The unadjusted estimator is identical to the Gcomp/IPMW estimator but including only the intercept term in the vector W.

Since the consistency of the initial estimators of $\bar{Q}_0$, g_0 and ϕ_0 is key to attain estimators with optimal statistical properties (i.e., consistency and efficiency), we carefully discuss the construction of such estimators in the next subsection. In particular, the next subsection deals with the construction of an estimator for $\bar{Q}_0$, the predictor of death in our working example.

Prediction via SuperLearner

As explained in the previous section, the consistency of the initial estimators $\bar{Q}_n$, g_n and ψ_n determine the statistical properties of the estimators of $\psi_{c,0}$ and $\psi_{b,0}$. Common practice in statistics involves the estimation of models like

$$\text{logit}\,\bar{Q}(A,W)=\beta_0+\beta_1 A+\beta_2 W+\beta_3 AW. \tag{15}$$

This approach that has gained popularity among researchers in epidemiology and biostatistics, partly because of the analysis of its statistical properties requires simple mathematical methods, and partly because it is readily available in statistical software. Nevertheless, as it is also well known among their users, parametric models of the type described by (15) are rarely correct, and their choice is often based on their continence and other subjective criteria. This practice leads to estimator whose usefulness is highly questionable given that the assumptions it entails (linearity, normality, link function, etc.) will rarely be theoretically supported.

In this paper we use the super learner [10] for estimation of $\bar{Q}_0$, g_0, and ψ_0. Super learner is a methodology that uses cross-validated risks to find an optimal combination of a list of user-supplied estimation algorithms. One of its most important theoretical properties is that its solution converges to the oracle estimator (i.e., the candidate in the library that minimizes the loss function with respect to the true probability distribution), thus providing the closest approximation to the real data generating mechanism. Proofs and simulations regarding these and other asymptotic properties of the super learner can be found in [34] and [35].

To implement the super learner predictor it is necessary to specify a library of candidate predictors algorithm. In the case of the conditional expectations $\bar{Q}_0$, ϕ_0, and g_0 for binary A, the candidates can be any regression or classification algorithm. Examples include random forests, logistic regression, k nearest neighbors, Bayesian models, etc. For estimation of the conditional densities g_0 we also use the super learner, with candidates given by several histogram density estimators, which yields a piece-wise constant estimator of the conditional density. The choice of the number of bins and their location is indexed by two tuning parameters. The implementation of this density estimator is discussed in detail in [36], and is omitted in this paper.

Details of Data Analysis

The sample size was $n = 918$ patients, and measurements of the variables described in S1 Table Table of the Supporting Information were taken at 6, 12, 24, 48, and 72 hours after admission to the emergency room.

The main objective of the study was the construction of prediction models for the risk of death of a patient in a certain time interval given the variables measured up to the start of the interval, as well as the definition and estimation of VIM measures that provide an account of the longitudinal evolution of the relation between these physiological and clinical measurements and the risk of death at a certain time point.

The data set was partitioned in 6 different data sets according to the time intervals defined by the time points in which measurements were taken, each of these 6 datasets contained only the patients that were at risk of death (alive) at the start of the time interval. Each of the continuous covariates was rescaled by subtracting the minimum and dividing by the range so that all of the covariates range between zero and one. The methods described in the previous sections were applied to each variable in each of these datasets.

The candidate algorithms for prediction of death used in the super learner predictor are:

- Logistic regression with main terms (GLM)
- Stepwise logistic regression (SW)
- Bayesian logistic regression (BLR) [37]
- Generalized additive models (GAM) [38]
- MARS (MARS) [39]
- Sample mean (MEAN).

The first three represent common practice in epidemiology and statistics, and the GAM and MARS algorithms intend to capture nonlinearities in relations between the data. The sample mean is included for contrast.

The super learner is a "black box" prediction algorithm, constructed to minimize the prediction error. As such, the coefficients of each candidate do not have any meaningful interpretation. To see this, consider a hypothetical situation in which two prediction candidates provide highly correlated predictions. In that case, the coefficients of the two candidates in the library will be highly variable, but the prediction error and the super learner will not be affected by such variability.

Results

Table 1 shows the coefficients of each candidate algorithm in the super learner predictor of $E(Y_j|\bar{L}_j, \bar{C}_j, L_0)$. The variability in these coefficients shows that no single algorithm is optimal for prediction at each time point, and that each algorithm describes certain features of the data

Table 1. Coefficients in the Super Learner.

	0–6hr	6–12hr	12–24hr	24–48hr	48–72hr	72+hr
GLM	0.0000	0.0000	0.0000	0.0318	0.0259	0.0000
SW	0.0000	0.1889	0.0000	0.0000	0.2073	0.1787
BGLM	0.3318	0.0586	0.1049	0.1329	0.0313	0.2750
GAM	0.5118	0.7525	0.8951	0.8353	0.7201	0.2487
MARS	0.1563	0.0000	0.0000	0.0000	0.0154	0.1298
MEAN	0.0000	0.0000	0.0000	0.0000	0.0000	0.1678

doi:10.1371/journal.pone.0120031.t001

generating mechanism that the others are not capable of unveiling, advocating for the need of an automated method to choose between them.

Fig. 2 presents the ROC curves for the cross-validated super learning predictions of death, as well as the cross-validated predictions based on a logistic model with AIC-based stepwise selection of variables, for comparison with common practice. The super learner prediction method outperforms the stepwise prediction in all cases, with AUC ROC (area under the ROC curve) differences ranging from 0.02 to 0.07. Though this differences might be small, an interpretation of their meaning reveals the clinical relevance of a slight improvement in prediction. The AUC ROC can be interpreted as the proportion of times that a patient who dies obtains a higher prediction score than a patient who survives. In practice, an AUC ROC difference of 0.02 means that in 100 pairs of patients (pairs formed by one patient who dies and one who

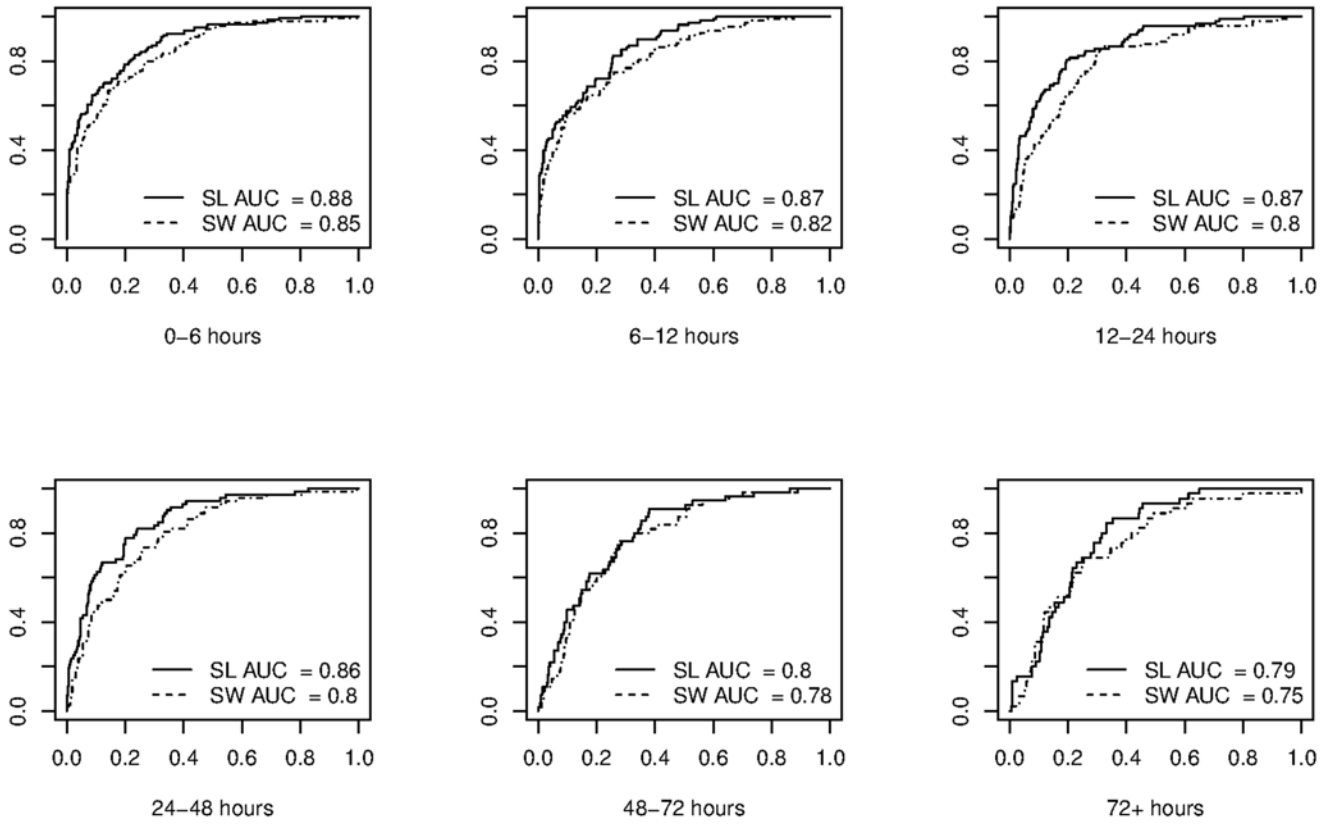

Fig 2. ROC curves of cross-validated prediction for the super learner (SL) and the logistic step-wise regression (SW), for different time intervals.

doi:10.1371/journal.pone.0120031.g002

Table 2. VIM estimates for the most important variables for prediction of death at each time interval according to TML estimate (p-values in parentheses and truncated at 0.001).

Time of death	Var. Name	Var. Time	TMLE	EE	G-comp/IPMW	Unadjusted
0–6 hrs.	APC	00	0.0205(0.023)	0.0183(0.043)	0.0235(0.386)	0.1063(< 0.001)
	INR	00	0.0216(< 0.001)	0.0193(0.002)	−0.0011(0.722)	0.0345(< 0.001)
	PT	00	0.0248(< 0.001)	0.0248(< 0.001)	−0.0011(0.698)	0.0400(< 0.001)
	ISS	00	−0.0314(< 0.001)	−0.0319(< 0.001)	−0.0242(< 0.001)	−0.0244(0.002)
	SBP	00	0.0041(0.010)	0.0041(0.011)	0.0020(0.108)	0.0340(< 0.001)
6–12 hrs.	PT	00	0.0076(0.001)	0.0074(0.001)	0.0019(0.426)	0.0376(< 0.001)
	BDE	00	0.0098(0.028)	0.0114(0.011)	0.0120(0.004)	0.0796(< 0.001)
	FV	06	−0.0139(0.047)	−0.0394(0.018)	−0.0394(0.141)	0.0199(0.349)
	ATIII	06	−0.0160(0.026)	−0.0323(0.009)	−0.0434(0.072)	0.0199(0.340)
12–24 hrs.	SBP	00	0.0031(0.034)	0.0030(0.037)	0.0018(0.220)	0.0317(< 0.001)
	PT	00	0.0051(0.025)	0.0048(0.031)	0.0019(0.377)	0.0357(< 0.001)
	FV	06	−0.0134(0.050)	−0.0246(0.008)	−0.0327(0.562)	0.0276(0.140)
	DDIM	00	0.0140(0.042)	0.0134(0.054)	0.0142(0.331)	0.0955(< 0.001)
	PC	06	−0.0224(0.001)	−0.0394(0.008)	−0.0286(0.349)	0.0262(0.221)
24–48 hrs.	PT	00	0.0080(< 0.001)	0.0080(< 0.001)	0.0026(0.124)	0.0313(< 0.001)
	DDIM	00	0.0134(0.026)	0.0133(0.028)	0.0144(0.504)	0.0770(< 0.001)
	ISS	00	−0.0229(< 0.001)	−0.0232(< 0.001)	−0.0211(< 0.001)	−0.0155(0.020)
	HR	12	0.0346(< 0.001)	0.0136(0.118)	−0.0030(0.922)	0.0570(< 0.001)
	APC	00	0.0429(< 0.001)	0.0432(< 0.001)	0.0226(0.310)	0.0761(< 0.001)
48–72 hrs.	CREA	00	0.0028(0.030)	0.0028(0.030)	0.0007(0.301)	0.0204(< 0.001)
	PT	00	0.0089(< 0.001)	0.0089(< 0.001)	0.0017(0.210)	0.0241(< 0.001)
	DDIM	12	0.0140(0.049)	0.0122(0.095)	0.0481(0.885)	0.0605(< 0.001)
	PC	06	−0.0164(0.012)	−0.0190(0.010)	−0.0252(0.075)	0.0381(0.035)
	RR	24	0.0187(0.002)	0.0148(0.012)	0.0057(0.749)	0.0644(< 0.001)
72+ hrs.	CREA	00	0.0027(0.002)	0.0027(0.002)	0.0007(0.291)	0.0168(< 0.001)
	ISS	00	−0.0142(0.005)	−0.0149(0.003)	−0.0145(0.008)	−0.0085(0.124)
	PTT	00	0.0220(< 0.001)	0.0219(< 0.001)	0.0017(0.012)	0.0220(< 0.001)

doi:10.1371/journal.pone.0120031.t002

does not) the super learner, on average, classifier correctly classifies two pairs more than the step-wise classifier, which could potentially lead to live-saving treatments for these two patients.

The VIM-TMLE measures that were significant at 0.05 were ranked according to their magnitude. Table 2 presents the first five (whenever five or more were significant) most important variables for prediction of death at each time interval, according to the TML estimator previously introduced. Recall that all the continuous variables were re-scaled between zero and one; the value $\delta = 0.01$ was used for all the estimates. The interpretation of the values in the first row of Table 2, for example, is that if APC were to increase by 1% for every patient, the mortality rate in the first time interval would, on average, increase by 2%. The TMLE and the EE produced generally similar results, whereas the Gcomp/IPMW estimator produced results that are somewhat different and with greater estimated variability. Note that several of the Gcomp/IPMW point estimates coincide with the TMLE and EE, but their p-values are generally larger. This could be due to the fact that the TMLE and EE are locally efficient estimators, and therefore provide more powerful hypothesis tests. In light of the superior theoretical properties of the TMLE and EE, we prefer to rely on estimates obtained through these methods.

Discussion

In this paper we address the problem of estimating variable importance parameters for longitudinal data that are subject to missingness. We present variable importance parameters that have a clear interpretation either as purely statistical parameters or as causal effects, depending on the assumptions about the data generating mechanism that the researcher is willing to make. These are important characteristics that advance the field in various fronts. First, unlike VIMs derived from machine learning and data-adaptive predictors (e.g., random forests), the VIMs defined in this paper have a concise definition as statistical parameters, which allowed the study of its asymptotic statistical properties and ultimately led to the construction of estimators with desirable statistical properties like consistency, efficiency, and asymptotic normality. Second, the assumptions required to give a causal interpretation to statistical parameters are often concealed, and the language used attempts to imply causal relations without clearly stating the necessary assumptions. The framework we present endows the user with the necessary tools to decide whether it is correct or not to interpret the estimates in terms of causal relations. Additionally, the parameters that we present have a purely statistical interpretation as a measure of conditional dependence, interpretation that must be used when there is not enough knowledge about the causal structure. We provide a methodology that can be used to compare continuous and binary variables in terms of their effect on an outcome, guaranteeing that the results are mathematically comparable.

We illustrate the use of the methods through the analysis of the drivers of recovery after sever trauma. These analyses provide a significant contribution to the field of trauma injury, by bringing state-of-the-art statistical methods to a field in which the large dimensionality of the problem constitutes a limiting factor for understanding the intricate relations between the variables involved. We propose a "black-box" prognosis algorithm (super learner) that can take into account the complexity of the problem, and represents an alternative to the scoring methods based on rules of thumb that are currently used in this setting. The results of the variable importance analysis corroborate the hypothesis that recovery after severe trauma is a dynamic process in which the decisive factors change over time, and provides provisional answers to various questions about recovery after severe trauma. Because there is not certainty that the structural causal assumptions required are met, the estimated VIMs can only be used as predictive performance measures and used to postulate hypothesis about causal relations that can be tested in more carefully designed studies. An additional advantage of a more carefully designed study is the possibility of performing a detailed comparison of the trajectories of each variable, using data that is not subject to missingness, or in which the amount of missingness is controlled.

We proposed a TMLE and an estimating equation estimator. Both of these estimators are doubly robust and efficient under certain regularity and consistency conditions of the initial estimators, but the TMLE has the additional advantages of a plug-in estimator. However, we did not observe any relevant difference between them in the illustration example. Various authors [23] have already compared these two estimators through a simulation study under no missingness of the treatment variable, finding no difference between them. We proposed the G-comp/IPMW, an additional estimator that represents an easy alternative to the TMLE or EE. Although we found various differences in the magnitude of the estimates between the TMLE and the G-comp/EE, the main discrepancy was with respect to the standard errors and p-values. We conjecture that these differences are a consequence of the inefficiency of the G-comp/IPMW, which results in hypothesis tests with less power. As discussed in Section S1 TMLE Algorithm of the Supporting Information, the proposed TMLE is defined by an iterative procedure that involves numerical integration of super learning predictions at each step. This

represents a drawback of the estimator in terms of scalability when compared to the estimating equation (EE) methodology. Because the estimating equation methodology involves only computation of predicted values from three super learners (outcome, missingness mechanism, and treatment mechanism), and averaging of a function of these predictions, its computational time is expected to be of the order of the computational time of the learners considered in the library.

Finally, given the flexibility of this general approach, and the ability to automate the algorithms, these type of variable importance measures promise great benefit in high dimensional clinical and other longitudinal settings.

Supporting Information

S1 Table. Variables in the ACIT data set.
(PDF)

S1 TMLE Algorithm. TMLE Algorithm.
(PDF)

Author Contributions

Conceived and designed the experiments: MC ID AH. Performed the experiments: MC ID AH. Analyzed the data: ID AH AD. Wrote the paper: ID AH MC.

References

1. van der Laan MJ. Statistical Inference for Variable Importance. International Journal of Biostatistics. 2006; 2(1). doi: 10.2202/1557-4679.1008
2. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;p. 267–288.
3. Breiman L. Random Forests. Machine Learning. 2001; 45:5–32. doi: 10.1023/A:1010933404324
4. Olden JD, Jackson DA. Illuminating "the black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling. 2002; 154:135 – 150. doi: 10.1016/S0304-3800(02)00064-9
5. Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling. 2004; 178:389 – 397. doi: 10.1016/j.ecolmodel.2004.03.013
6. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics. 2007; 8(1):25. doi: 10.1186/1471-2105-8-25 PMID: 17254353
7. Ishwaran H. Variable importance in binary regression trees and forests. Electronic Journal of Statistics. 2007;p. 519–537.
8. Tuglus C, van der Laan MJ. Targeted methods for biomarker discovery. In: Targeted Learning. Springer; 2011. p. 367–382.
9. Stitelman OM, Hubbard AE, Jewell NP. The Impact Of Coarsening The Explanatory Variable Of Interest In Making Causal Inferences: Implicit Assumptions Behind Dichotomizing Variables. UC Berkeley Division of Biostatistics Working Paper Series. 2010;Working Paper 264.
10. van der Laan MJ, Polley E, Hubbard A. Super Learner. Statistical Applications in Genetics and Molecular Biology. 2007; 6(25).
11. Hess JR, Holcomb JB, Hoyt DB. Damage control resuscitation: the need for specific blood products to treat the coagulopathy of trauma. Transfusion. 2006 May; 46:685–686. doi: 10.1111/j.1537-2995.2006.00816.x PMID: 16686833
12. Holcomb JB, McMullin NR, Pearse L, Caruso J, Wade CE, Oetjen-Gerdes L, et al. Causes of death in US Special Operations Forces in the global war on terrorism: 2001–2004. Annals of surgery. 2007; 245 (6):986. doi: 10.1097/01.sla.0000259433.03754.98 PMID: 17522526

13. Krumrei NJ, Park MS, Cotton BA, Zielinski MD. Comparison of massive blood transfusion predictive models in the rural setting. The Journal of Trauma and Acute Care Surgery. 2012; 72(1):211. PMID: 22310129
14. Lesko MM, Jenks T, O'Brien S, Childs C, Bouamra O, Woodford M, et al. Comparing Model Performance for Survival Prediction Using Total GCS and Its Components in Traumatic Brain Injury. Journal of Neurotrauma. 2012;(ja).
15. MacFadden LN, Chan PC, Ho KHH, Stuhmiller JH. A model for predicting primary blast lung injury. The Journal of Trauma and Acute Care Surgery. 2012;.
16. Nunez TC, Voskresensky IV, Dossett LA, Shinall R, Dutton WD, Cotton BA. Early prediction of massive transfusion in trauma: simple as ABC (assessment of blood consumption)? The Journal of Trauma and Acute Care Surgery. 2009; 66(2):346–352. doi: 10.1097/TA.0b013e3181961c35
17. Schöchl H, Cotton B, Inaba K, Nienaber U, Fischer H, Voelckel W, et al. FIBTEM provides early prediction of massive transfusion in trauma. Crit care. 2011; 15(6):R265. doi: 10.1186/cc10539 PMID: 22078266
18. Buchman TG. Novel representation of physiologic states during critical illness and recovery. Crit Care. 2010; 14:127. doi: 10.1186/cc8868 PMID: 20236462
19. Bir N, Lafargue M, Howard M, Goolaerts A, Roux J, Carles M, et al. Cytoprotective-Selective Activated Protein C Attenuates Pseudomonas aeruginosa–Induced Lung Injury in Mice. American journal of respiratory cell and molecular biology. 2011; 45(3):632. doi: 10.1165/rcmb.2010-0397OC PMID: 21257925
20. Cohen M, Brohi K, Calfee C, Rahn P, Chesebro B, Christiaans S, et al. Early release of high mobility group box nuclear protein 1 after severe trauma in humans: role of injury severity and tissue hypoperfusion. Critical Care. 2009; 13(6):R174. doi: 10.1186/cc8152 PMID: 19887013
21. Cohen MJ, Bir N, Rahn P, Dotson R, Brohi K, Chesebro BB, et al. Protein C depletion early after trauma increases the risk of ventilator-associated pneumonia. The Journal of Trauma and Acute Care Surgery. 2009; 67(6):1176–1181. doi: 10.1097/TA.0b013e3181c1c1bc
22. Pearl J. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge; 2000.
23. Díaz, I, van der Laan M. Population Intervention Causal E ects Based on Stochastic Interventions. Biometrics. 2011;p. In press. Available from: http://dx.doi.org/10.1111/j.1541-0420.2011.01685.x.
24. Korb K, Hope L, Nicholson A, Axnick K. Varieties of Causal Intervention. In: Zhang C, W Guesgen H, Yeap WK, editors. PRICAI 2004: Trends in Artificial Intelligence. vol. 3157 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg; 2004. p. 322–331.
25. Didelez V, Dawid AP, Geneletti S. Direct and Indirect E ects of Sequential Treatments. In: UAI; 2006..
26. Dawid, AP, Didelez V. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. CoRR. 2010;abs/1010.3425.
27. Rubin DB. Bayesian Inference for causal e ects: the role of randomization. Annals of Statistics. 1978; 6:34–58. doi: 10.1214/aos/1176344064
28. Bickel PJ, Klaassen CAJ, Ritov Y, Wellner J. E cient and Adaptive Estimation for Semiparametric Models. Springer-Verlag; 1997.
29. van der Laan MJ, Rubin D. Targeted Maximum Likelihood Learning. The International Journal of Biostatistics. 2006; 2(1). doi: 10.2202/1557-4679.1043
30. van der Laan MJ, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer; 2011.
31. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor e ect. Mathematical Modelling. 1986; 7:1393–1512. doi: 10.1016/0270-0255(86)90088-6
32. van de Geer SA. Empirical Processes in M-Estimation. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press; 2000.
33. van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. Springer, New York; 2003.
34. van der Laan, MJ, Dudoit, S, Keles, S. Asymptotic optimality of likelihood-based cross-validation. Statistical Applications in Genetics and Molecular Biology. 2004;3.
35. van der Laan MJ, Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Division of Biostatistics, University of California, Berkeley; 2003.
36. Díaz I, van der Laan M. Super Learner Based Conditional Density Estimation with Application to Marginal Structural Models. The International Journal of Biostatistics. 2011; 7(1):38.

37. Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics. 2008;p. 1360–1383.

38. Hastie T, Tibshirani R. Generalized additive models. Statistical science. 1986;p. 297–310.

39. Friedman JH. Multivariate adaptive regression splines. The annals of statistics. 1991;p. 1–67.