



A Modular Framework for Treatment Effect Estimation in Latent Subgroups

Lars Lindhagen¹ · Hans Garmo² · Ollie Östlund¹

Received: 12 March 2024 / Accepted: 1 September 2025 / Published online: 9 September 2025
© The Author(s) 2025

Abstract

Latent subgroups arise when patients are randomized to an intended treatment, that can only be given for certain, treatable, patients. For biological efficacy, the relevant estimand is then the treatment effect in the subgroup of treatable patients, with the obvious issue that this subgroup is latent, identified only in the intervention arm. We present a modular framework for effect estimation in such latent subgroups. The framework consists of a core and three plug-in models, for subgroup membership and outcomes among treatable and non-treatable patients. The core computes maximum likelihood estimates using the EM algorithm, together with standard errors. It does so without any knowledge about the details of the plug-in models, giving the user great flexibility. The methods are implemented in an R package. The framework is validated in a simulation, where we also explore the use of predictors. Particularly intriguing are predictors of treatability, partly identifying the latent subgroup from baseline data. The results suggest that this can dramatically increase the power, while being robust against model misspecifications. Finally, the methods are applied to a prostate cancer trial.

Keywords Randomized trial · Latent subgroup · Instrumental variables · Maximum likelihood · EM algorithm

✉ Lars Lindhagen
lars.lindhagen@ucr.uu.se

Hans Garmo
hans.garmo@rccmellan.se

Ollie Östlund
ollie.ostlund@ucr.uu.se

¹ Uppsala Clinical Research Center, Uppsala University, Dag Hammarskjölds väg 38, 75185 Uppsala, Sweden

² Department of Surgical Sciences, Uppsala University, Uppsala, Sweden

1 Introduction

In some trials, patients are randomized to an intended treatment, that can only be given under certain conditions. A typical example is a screening trial, where individuals are randomized to a diagnostic test for, say, an infection, with subsequent eradication if positive. A similar situation arises in a cancer trial, if patients are randomized to surgery, which needs to be aborted if metastases are detected. Yet another example is a placebo-controlled drug trial, where only compliers take the drug.

All these examples have in common the existence of a *latent subgroup* of patients (infected; metastasis-free; compliers), who would receive treatment (eradication; surgery; the drug) if randomized to the intervention arm. We shall refer to these patients as *treatable*. The subgroup of treatable patients extends into both trial arms, but is identified only in the intervention arm, hence the word latent.

When estimating treatment effects in such trials, an important distinction has to be made between programmatic effectiveness and biological efficacy [1]. The former deals with strategies and guidelines: what would happen if the intervention was implemented on a large scale, e.g. by modifying a guideline? This can be assessed by so-called intention-to-treat (ITT) analyses. Non-treatable patients are then part of the population, and ITT properly accommodates the fact that they are not affected by the intervention. Biological efficacy, on the other hand, focuses on actual biological processes. The natural estimand is then the treatment effect among patients that can in fact be given the treatment at hand, the treatable subgroup. This is the topic of the present paper.

An obvious issue is the latency of this subgroup: we know who is treatable in the intervention arm, but not in the control arm. It may therefore perhaps seem surprising that it is even possible to say something about biological efficacy. Figure 1 gives an intuitive explanation of why this can be done. It shows a hypothetical trial with 1000 + 1000 patients. In the intervention arm, we observe 200 treatable and 800 non-treatable patients, with 50 and 100 deaths, respectively. In the control arm, we observe 200 deaths. By randomization, we would expect 800 non-treatable patients also in the control arm. Provided that randomized arm makes no difference for non-treatable patients, there should be 100 death also among these, and hence 100 deaths among the 200 non-treatable patients. The biological efficacy can be expressed as the relative risk among treatable patients: $RR_{BE} = 50/100 = 0.50$. This contrasts with the programmatic effectiveness, estimated using ITT: $RR_{PE} = 150/200 = 0.75$. The latter is weaker because the effect is diluted by patients not benefiting from the intervention. We stress that both these values are correct, but they answer different scientific questions.

The above computation is not the optimal way to compute biological efficacy, for one thing the estimate lacks a finite mean [2]. The maximum likelihood (ML) methods that we shall rely on are preferable. But the example does show that such things are possible, and also highlights the key underlying assumptions: randomization ensures symmetry between the trial arms; control patients never receive

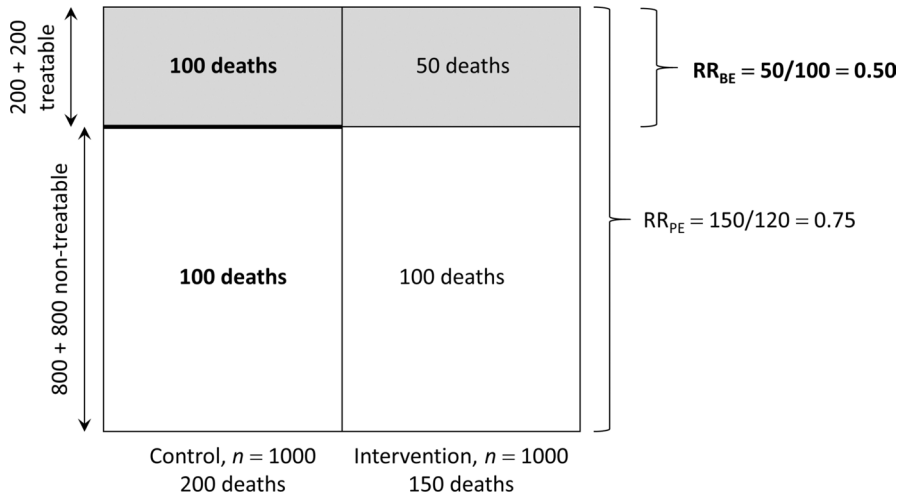


Fig. 1 A hypothetical trial with 1000 patients in each arm. Treatability is observed in the intervention arm, but not in the control arm. Unobserved (inferred) quantities are displayed using boldface text and thick lines. The latent subgroup of $200 + 200$ treatable patients is shaded. RR relative risk, BE biological efficacy, PE programmatic effectiveness

treatment, simplifying the relation between randomization and treatment; and, importantly, for non-treatable patients, it makes no difference which arm they are randomized to. These constitute the major part of the requirements on an instrumental variable, and are commonly referred to as exchangeability, monotonicity, and exclusion restriction. In many cases, they follow more or less directly from the study design, although exclusion restriction may need to be considered carefully, see Sect. 4 below for an example.

Instrumental variable methods were historically developed largely by econometricians, with a focus on numerical outcomes [3], and have later been extended to e.g. time to event data [4, 5]. Our work is a generalization of a paper by Altstein et al. [6], to which we refer for extensive historical references. They develop an accelerated failure time (AFT) model for time to event outcomes in the latent subgroup. The parameter vector then naturally splits into three parts: two for the AFT models of treatable and non-treatable patients, and an additional single parameter for the prevalence of treatability. ML estimates of these parameters are computed using the expectation-maximization (EM) algorithm [7].

The present paper extends [6] in several ways. First, we develop a modular framework, separating the computations into a core and three plug-in models, cf. Fig. 2 below. The core computes ML estimates without any knowledge about the internal structure of the plug-in models. Second, this framework allows for arbitrary outcome types, e.g. dichotomous or ordinal, as long as the plug-in models are fully parametric. Third, the flexible framework structure makes it easy to employ predictors of both outcome and treatability. The latter is particularly important, since even limited baseline information can mitigate the latency issue. Fourth, we derive closed-form expressions for asymptotic standard errors. And fifth, we provide the R

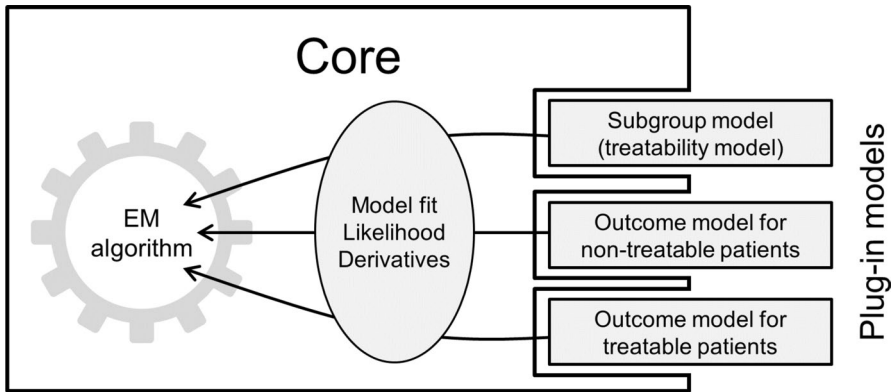


Fig. 2 The framework consists of a core that performs the computations using the EM algorithm, and three plug-in models for membership of the latent subgroup of treatable patients, and for outcomes of treatable and non-treatable patients. All information about data types and distributions are contained in the plug-in models

package **latent**, which implements the methods described in this paper, including a set of plug-in models for the most common outcomes. It can be downloaded from GitHub [8].

The paper is organized as follows: Sect. 2 presents the framework and the core algorithms, with technical details in an appendix. Section 3 validates the methods in a simulation process, after which we apply the framework to a data set from a prostate cancer trial in Sect. 4. Concluding remarks and a discussion of possible extensions can be found in Sect. 5.

2 The Framework

We propose a framework consisting of a core and three plug-in models. The core performs ML estimation of the treatment effect in the latent subgroup of treatable patients using the EM algorithm, and also computes standard errors. This is done without any knowledge of the assumed distributions of the outcome, or even its data type. All such issues are relayed to three statistical models supplied by the user: one for subgroup membership, and two for the outcome of treatable and non-treatable patients, respectively. We shall refer to these as plug-in models, since they are “plugged into” the core, as illustrated in Fig. 2.

2.1 Preliminaries

2.1.1 Vectors and Derivatives

For two vectors \mathbf{u} and \mathbf{v} in \mathbf{R}^d , the scalar product is $\mathbf{u} \cdot \mathbf{v} = \sum_i u_i v_i$, and the tensor product $\mathbf{u} \otimes \mathbf{v}$ is the matrix $A_{ij} = u_i v_j$. We also let $\mathbf{u}^{\otimes 2} = \mathbf{u} \otimes \mathbf{u}$ denote the tensor

square. For a smooth function $f : \mathbf{R}^d \rightarrow \mathbf{R}$, the Jacobian ∇f is the vector of derivatives $\partial f / \partial x_i$, and the Hessian $\nabla \nabla f$ is the matrix of second derivatives $\partial^2 f / \partial x_i \partial x_j$.

It will sometimes be more convenient to differentiate the likelihood itself, sometimes the log-likelihood. Logarithmic differentiation allows us to move back and forth between these. If $F : \mathbf{R}^d \rightarrow \mathbf{R}_+$ is a smooth function and $f = \log F$, then

$$\begin{cases} \nabla F = F \nabla f \\ \nabla \nabla F = F (\nabla \nabla f + (\nabla f)^{\otimes 2}) \end{cases} \quad (1)$$

and

$$\begin{cases} \nabla f = \frac{\nabla F}{F} \\ \nabla \nabla f = \frac{\nabla \nabla F}{F} - \left(\frac{\nabla F}{F} \right)^{\otimes 2}. \end{cases} \quad (2)$$

2.1.2 Notation

We shall use i as an index over patients. In order to ease the notational burden, this index will often be suppressed when there is no risk of confusion. Random variables are written with capital letters, and observations thereof in lower-case.

We let R denote randomization: $R = 1$ for intervention and $R = 0$ for control. Moreover, \mathbf{x} is a vector of patient characteristics known at baseline, including R , whereas G is the latent subgroup indicator: $G = 1$ for treatable patients, and $G = 0$ for non-treatable ones. After randomization and treatment, an outcome y is recorded. All these variables are observed, except G in the control arm.

2.1.3 Likelihoods

We shall use p for discrete probabilities or probability densities, depending on the type of distribution. In our framework, likelihoods exist on two levels. The first level (the core) handles the total likelihood of a patient, L_i . The full likelihood of the data is $L_{\text{Full}} = \prod_i L_i$. On the second level, there are likelihoods of the plug-in models. To avoid confusion, the latter will be denoted M , not L .

A comment on notation: likelihoods are really functions of model parameters given data, say $M(\boldsymbol{\theta}) = p(y|\boldsymbol{\theta})$. However, we will often use the same notation to express probabilities of data given parameters, writing things like $M(y) = M(y|\boldsymbol{\theta}) := p(y|\boldsymbol{\theta})$. This should cause no confusion.

2.2 Plug-In Models

Our framework rests on three statistical models. We shall use $\boldsymbol{\theta}$ to denote model-specific parameters, reserving $\boldsymbol{\theta}$ for the totality of parameters.

Subgroup model: Membership of the latent subgroup of treatable patients is modeled using a set of parameters $\boldsymbol{\vartheta}_S$:

$$M_S(g|\mathbf{x}, \boldsymbol{\vartheta}_S) = p(g|\mathbf{x}, \boldsymbol{\vartheta}_S), \quad g = 0, 1.$$

Outcome model for non-treatable patients: Outcomes in the subgroup of non-treatable patients ($G = 0$) are modeled using parameters $\boldsymbol{\vartheta}_0$:

$$M_0(y|\mathbf{x}, \boldsymbol{\vartheta}_0) = p(y|\mathbf{x}, G = 0, \boldsymbol{\vartheta}_0).$$

Outcome model for treatable patients: Similarly, for the treatable subgroup,

$$M_1(y|\mathbf{x}, \boldsymbol{\vartheta}_1) = p(y|\mathbf{x}, G = 1, \boldsymbol{\vartheta}_1).$$

The outcome models will be indexed by $m = 0, 1$. Thus, the latter two items can be summarized as $M_m(y|\mathbf{x}, \boldsymbol{\vartheta}_m) = p(y|\mathbf{x}, G = m, \boldsymbol{\vartheta}_m)$. Finally, $\boldsymbol{\theta} = (\boldsymbol{\vartheta}_S, \boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}_1)$ is the total parameter vector.

The response g of the subgroup model is always dichotomous, whereas the outcome models use whatever outcome was recorded in the trial. Note that the two outcome models are completely independent of each other. Not only do they have disjoint parameter vectors $\boldsymbol{\vartheta}_0$ and $\boldsymbol{\vartheta}_1$, but they may also in principle have different underlying distributions, and use different parts of the covariate vector \mathbf{x} . For example, the outcome model for treatable patients depends on the randomization R , whereas the other model does not (exclusion restriction).

In addition to this, all three models can benefit from variables that are predictive of their respective response. This is particularly interesting for the subgroup model, since it means a step towards identification of the latent subgroup from baseline data. We shall explore this idea further in Sect. 3.2.

2.3 Treatability

As discussed, subgroup membership (treatability) is known for $R = 1$, but not for $R = 0$. In the latter case, we instead specify probabilities. We shall use both prior and posterior probabilities, i.e. with and without conditioning on the outcome. To this end, let $D_{\text{obs}}^{(\text{BL})}$ comprise all observed baseline data, i.e. \mathbf{x} together with g , when observed, but not the outcome (recall that R is contained in \mathbf{x}). Let π denote the prior probability of treatability:

$$\pi = P\left(G = 1 \mid D_{\text{obs}}^{(\text{BL})}, \boldsymbol{\vartheta}_S\right) = \begin{cases} M_S(1|\mathbf{x}, \boldsymbol{\vartheta}_S), & R = 0, \\ g, & R = 1. \end{cases}$$

The ($R = 1$) case expresses the fact that subgroup membership is known in the intervention arm: $\pi = g \in \{0, 1\}$.

After observing the outcome y , we update this to a posterior probability, which we, for reasons to become clear shortly, denote w . Thus, $w = P(G = 1 | D_{\text{obs}}, \boldsymbol{\theta})$, where $D_{\text{obs}} = D_{\text{obs}}^{(\text{BL})} \cup \{y\}$ is all observed data. By Bayes' rule, we have, for $m = 0, 1$,

$$P(G = m | D_{\text{obs}}, \theta) \propto P(G = m | D_{\text{obs}}^{(\text{BL})}, \theta) \times p(y | \mathbf{x}, G = m, \theta).$$

In other words,

$$\begin{cases} P(G = 0 | D_{\text{obs}}, \theta) \propto (1 - \pi) \times M_0(y | \mathbf{x}, \boldsymbol{\vartheta}_0) =: \tilde{w}_0, \\ P(G = 1 | D_{\text{obs}}, \theta) \propto \pi \times M_1(y | \mathbf{x}, \boldsymbol{\vartheta}_1) =: \tilde{w}_1. \end{cases}$$

The posterior probability of treatability is computed from this by normalization:

$$w = \frac{\tilde{w}_1}{\tilde{w}_0 + \tilde{w}_1}. \quad (3)$$

Just like for the prior probabilities, $w = g \in \{0, 1\}$ if G is observed to have the value g .

2.4 Point Estimation with the EM Algorithm

The EM algorithm [7] is an iterative method to perform ML estimation in the presence of missing data. In our case, the missing data are $D_{\text{mis}} = \{G | R = 0\}$, i.e. treatability in the control arm. In each EM iteration, old parameters $\theta^{(t)}$ are updated to new ones $\theta^{(t+1)}$, as described below.

2.4.1 E step

The EM algorithm begins with the likelihood of complete data $D_{\text{com}} := D_{\text{obs}} \cup D_{\text{mis}}$. Since G is then fully known, the likelihood is

$$\begin{aligned} p(D_{\text{com}} | \theta) &= \prod_i p(y_i | g_i, \mathbf{x}_i, \theta) p(g_i | \mathbf{x}_i, \theta) \\ &= \prod_i \left\{ \begin{array}{l} M_0(y_i | \mathbf{x}_i, \boldsymbol{\vartheta}_0) M_S(0 | \mathbf{x}_i, \boldsymbol{\vartheta}_S), \quad g_i = 0 \\ M_1(y_i | \mathbf{x}_i, \boldsymbol{\vartheta}_1) M_S(1 | \mathbf{x}_i, \boldsymbol{\vartheta}_S), \quad g_i = 1 \end{array} \right\}. \end{aligned}$$

Taking logarithms, this can be written as

$$\begin{aligned} \log p(D_{\text{com}} | \theta) &= \sum_i \left\{ (1 - g_i) \times [\log M_0(y_i | \mathbf{x}_i, \boldsymbol{\vartheta}_0) + \log M_S(0 | \mathbf{x}_i, \boldsymbol{\vartheta}_S)] \right. \\ &\quad \left. + g_i \times [\log M_1(y_i | \mathbf{x}_i, \boldsymbol{\vartheta}_1) + \log M_S(1 | \mathbf{x}_i, \boldsymbol{\vartheta}_S)] \right\}. \end{aligned} \quad (4)$$

The second part of the E step is to take the expectation of this over D_{mis} , i.e. unobserved g 's, conditioning on observed data and old model parameters $\theta^{(t)}$. The resulting expression is usually denoted Q . Since everything is linear in g_i , this simply amounts to replacing g_i by its posterior probability of being unity, w_i , as computed in (3) with $\theta = \theta^{(t)}$. Moreover, $w_i = g_i$ whenever g_i is observed, so it doesn't hurt to do this for the observed g 's as well. Thus, Q arises by simply substituting w_i for g_i in (4).

2.4.2 M Step

The M step proceeds by maximizing Q from the E step over $\theta = (\boldsymbol{\vartheta}_S, \boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}_1)$. Recall that the expression to be maximized is (4) with g_i replaced by w_i . Since the latter depend only on $\theta^{(t)}$, they can be treated as constants. The maximization then separates into three problems:

Maximizing over $\boldsymbol{\vartheta}_S$: There are two terms in (4) that contain $\boldsymbol{\vartheta}_S$. We therefore need to maximize

$$\sum_i \{(1 - w_i) \log M_S(0|\mathbf{x}_i, \boldsymbol{\vartheta}_S) + w_i \log M_S(1|\mathbf{x}_i, \boldsymbol{\vartheta}_S)\}.$$

This is equivalent to solving a weighted ML problem for the subgroup model based on an augmented data set, where each patient i contributes two rows: one row with response $g_i = 0$ and weight $(1 - w_i)$, and one row with $g_i = 1$ and weight w_i . Since $w_i \in \{0, 1\}$ in the intervention arm, only the "correct" row will contribute there. This weighted problem was the rationale for denoting the posterior probabilities w_i .

Maximizing over $\boldsymbol{\vartheta}_0$: Since there is only one term in (4) that contains $\boldsymbol{\vartheta}_0$, this reduces to maximizing

$$\sum_i (1 - w_i) \log M_0(y_i|\mathbf{x}_i, \boldsymbol{\vartheta}_0),$$

a weighted ML problem with weights $(1 - w_i)$.

Maximizing over $\boldsymbol{\vartheta}_1$: Similar to the above, we now maximize

$$\sum_i w_i \log M_1(y_i|\mathbf{x}_i, \boldsymbol{\vartheta}_1),$$

again a weighted ML problem, but with weights w_i .

These problems can usually be solved using standard statistical software. The solution gives us updated model parameters $\theta^{(t+1)}$, completing the iteration.

2.4.3 Initial Values and Stopping Criterion

As any iterative procedure, the EM algorithm depends on a sensible start guess. We suggest first fitting the subgroup model in the intervention arm, where the subgroup is identified. Based on this fit, the probability of treatability is computed for the controls. Finally, the two outcome models are fitted on the entire dataset using weights based on these probabilities.

To end the iteration, we suggest a likelihood-based stopping rule: the iteration is terminated if the log-likelihood increase between two consecutive iterations is less than, say, 10^{-5} .

2.5 Closed-Form Standard Errors

Having arrived at a point estimate $\hat{\theta}$ from the EM algorithm, we now turn to standard errors. By general likelihood theory, the covariance matrix of $\hat{\theta}$ is asymptotically given by the inverse observed Fisher information matrix, $(-\nabla\nabla \log L_{\text{full}}(\hat{\theta}))^{-1}$, where the derivatives are taken with respect to the full parameter vector θ . Since $\log L_{\text{full}}(\theta) = \sum_i \log L_i(\theta)$, we can compute the Hessian for one patient at a time, and then sum the results. Doing so, we shall suppress the index i , writing $L(\theta)$ rather than $L_i(\theta)$.

2.5.1 The Likelihood of a Single Patient

The likelihood of one patient is the probability of observing y and, if observed, g , conditional on x and θ . By the law of total probability, this is

$$\begin{aligned} L(\theta) &= p(y, \{g \text{ if observed}\} | x, \theta) \\ &= \begin{cases} p(y|G=0, x, \vartheta_0)p(G=0|x, \vartheta_S) + p(y|G=1, x, \vartheta_1)p(G=1|x, \vartheta_S), & R=0, \\ p(y|G=g, x, \vartheta_g)p(G=g|x, \vartheta_S), & R=1 \end{cases} \\ &= \begin{cases} \sum_m M_m(y|x, \vartheta_m)M_S(m|x, \vartheta_S), & R=0, \\ M_g(y|x, \vartheta_g)M_S(g|x, \vartheta_S), & R=1, \end{cases} \end{aligned} \quad (5)$$

where the sums run over $m \in \{0, 1\}$.

To ease the notation, we suppress the parameters, writing e.g. M_1 for $M_1(y|x, \vartheta_1)$. With this reduced notation, (5) can be shortened to

$$L(\theta) = \begin{cases} \sum_m M_m M_S(m), & R=0, \\ M_g M_S(g), & R=1. \end{cases} \quad (6)$$

Finally, we define 0/1 variables δ_m as $\delta_m = 1$ if $G = m$ is possible (g unobserved, or observed to be m) and $\delta_m = 0$ otherwise. Equation (6) can then be further simplified to

$$L(\theta) = \sum_m \delta_m M_m M_S(m). \quad (7)$$

2.5.2 Derivatives of the Likelihood

Differentiating (7) twice yields

$$\nabla L(\theta) = \sum_m \delta_m \{M_S(m)\nabla M_m + M_m \nabla M_S(m)\} \quad (8)$$

and

$$\nabla \nabla L(\theta) = \sum_m \delta_m \left\{ M_S(m) \nabla \nabla M_m + \nabla M_m \otimes \nabla M_S(m) + \nabla M_S(m) \otimes \nabla M_m + M_m \nabla \nabla M_S(m) \right\}, \tag{9}$$

where all derivatives are taken with respect to θ . Since $\theta = (\vartheta_S, \vartheta_0, \vartheta_1)$, it is natural to write this in block matrix form. Several derivatives then vanish, for example $\nabla_{\vartheta_0} M_S(m) = \mathbf{0}$ since M_S depends only on ϑ_S . By straightforward calculations, (8) and (9) can be written as

$$\nabla L(\theta) = \left[\begin{array}{c} \sum_m \delta_m M_m \nabla M_S(m) \\ \delta_0 M_S(0) \nabla M_0 \\ \delta_1 M_S(1) \nabla M_1 \end{array} \right]$$

and

$$\nabla \nabla L(\theta) = \left[\begin{array}{c|c|c} \sum_m \delta_m M_m \nabla \nabla M_S(m) & (*) & (*) \\ \hline \delta_0 \nabla M_0 \otimes \nabla M_S(0) & \delta_0 M_S(0) \nabla \nabla M_0 & (*) \\ \hline \delta_1 \nabla M_1 \otimes \nabla M_S(1) & \mathbf{0} & \delta_1 M_S(1) \nabla \nabla M_1 \end{array} \right],$$

where (*) is whatever it takes to form a symmetric matrix. These derivatives are to be evaluated at the point estimate $\hat{\theta}$, and are understood to involve only the parameters relevant to the model at hand, e.g. $\nabla M_0 := \nabla_{\vartheta_0} M_0$.

To round things up, the Hessian of the log-likelihood of a single patient, $\nabla \nabla \log L$, can be found from these expressions together with logarithmic differentiation (2). Summing over patients yields the full Hessian $\nabla \nabla \log L_{\text{full}}$, from which standard errors for individual model parameter are found as the square root of the diagonal elements of $(-\nabla \nabla \log L_{\text{full}})^{-1}$. From this, Wald confidence intervals and p -values can be computed in the usual way. For example, a 95% confidence interval for the k :th model parameter is given by $\hat{\theta}_k \pm 1.96 \times \text{SE}_k$.

2.6 Developing Plug-In Models

Sections 2.4 and 2.5 describe the computations of the core. As can be seen, no knowledge about the internal structure of the plug-in models is needed for this, not even about the type of the outcome. In fact, only two things are required of the models: (1) fitting weighted regression models using ML, and (2) computing likelihoods and their first two derivatives. This decoupling of the core and the plug-in models a major strength of our framework. It makes it easy for the user to tailor the analysis by modifying the plug-in models, without having to modify the core.

In Appendix 1, we work out the details for popular models for numerical, dichotomous, count, ordinal, and time to event data. We stress that the framework is in no way restricted to these models. On the contrary, any parametric model for any type of outcome can be used, as long as the likelihood and its derivatives can be derived.

2.7 The latent Package

All methods described in this section, including the plug-in models of Appendix 1, have been implemented in the R package **latent**, which is available at GitHub [8]. The package vignette gives several illustrations of how to use it, a simple example is also provided in Appendix 2.

3 Simulation

This section validates the methods and the R package in a simulation. Since estimates computed via the EM algorithm are maximum likelihood estimates, albeit in a non-trivial setting, the general likelihood theory guarantees asymptotic unbiasedness and correct confidence interval (CI) coverage.

A note on sample size: our target parameter is a treatment effect in a (latent) subgroup. Hence the effective sample size is the size of this subgroup, $n\pi$, where n is the number of patients, and π is the prevalence of treatability. For fixed subgroup size, one could expect larger π and smaller n to be preferable, since this reduces the uncertainty as to who is treatable. Compare, for example, ($n = 2000$, $\pi = 50\%$) with ($n = 1000$, $\pi = 100\%$). In both cases, the effective sample size is $n\pi = 1000$, but in the latter case, the subgroup is completely identified.

For each of the scenarios described below, 25,000 simulations were run. Power was defined as the proportion of simulations with a p -value below 0.05 for the null hypothesis of no treatment effect. In the scenarios where that is true ($\psi = 0$), this is the type I error rate. The Monte Carlo errors of power and CI coverage are typically a few tenths of a percentage point.

3.1 Proof of Concept

To begin with, we ran a series of basic simulations, varying the sample size, subgroup prevalence, and treatment effect as follows. Sample size: $n = 500$ and 3000 . Prevalence: $\pi = 10\%$, 25% , and 60% . Treatment effect: $\psi = 0$, -0.2 , and -0.4 . Negative treatments effects reflect the idea that small outcome values are beneficial, $\psi = 0$ corresponds to no treatment effect. The values were chosen so that ($n = 500$, $\pi = 60\%$) and ($n = 3000$, $\pi = 10\%$) both give the effective sample size of 300. The subgroup model was logistic regression without covariates. Four outcome types were studied: numerical, dichotomous, ordinal, and time to event.

Numerical outcomes were simulated according to the model

$$\begin{cases} Y = \alpha_0 + \epsilon, & G = 0, \\ Y = \alpha_1 + \psi R + \epsilon, & G = 1, \end{cases} \quad (10)$$

where ϵ are centered normal residuals with standard deviations $\sigma_0 = 2$ and $\sigma_1 = 2.5$. We also let $\alpha_1 = \alpha_0 + 0.4$ (the value of α_0 is immaterial), corresponding to a

situation where treatable patients are worse off, but with a maximal treatment effect ($\psi = -0.4$) they improve to the levels of non-treatable ones. The σ 's were chosen different to illustrate the fact that the two outcome models are fully independent.

Dichotomous outcomes were handled in a similar way. The event rate without treatment was set to 20% in the non-treatable group and 27.2% in the treatable group, again corresponding to $\alpha_1 = \alpha_0 + 0.4$ with the logit link.

For *ordinal* outcomes, we used four levels with probabilities 20%, 40%, 30%, and 10% for non-treatable patients. For treatable patients, these probabilities were modified by increasing the three intercepts by 0.4 in a proportional odds model.

Finally, *time to event* outcomes were generated according to the Royston-Parmar model (14) with three knots, including the boundary knots. The spline coefficients γ_j were chosen to approximate a Gompertz distribution with shape 1.5 and rate 1. Again, the treatable subgroup had its intercept γ_0 increased by 0.4. Censoring times were generated according to the same distribution, resulting in 50% censorings.

The simulation results are shown in Table 1. Overall, they are satisfactory, with limited bias and close to nominal (95%) CI coverage. Comparing the two scenarios with the same effective sample size, we see that, as anticipated, the larger prevalence ($n = 500, \pi = 60\%$ vs. $n = 3000, \pi = 10\%$) gives better results in terms of smaller bias, CI coverage closer to nominal, and larger power.

For very small trials, especially ($n = 500, \pi = 10\%$), with an effective sample size of 50 and about 10 events on average, things tend to deteriorate, with substantial bias and incorrect CI coverage. This is particularly true for dichotomous outcomes, with a bias away from zero, and a CI coverage of almost 100%. This small-sample bias is a well-known shortcoming of the logistic regression model [9], and remedies via penalization have been proposed [10]. Although we shall not pursue this, we note that thanks to the modularity of the framework, it is straightforward to implement such things, it is just another plug-in model. Similar small-sample problems, although to a lesser degree, can also be seen for numerical and time to event outcomes. Interestingly, the bias is then towards the null, with a subnominal CI coverage.

3.2 Predictors of Treatability and Outcome

A natural idea is that baseline information that is predictive of treatability should be helpful, making the latent subgroup of treatable patients less latent, so to speak. Our framework makes it easy to exploit such information, one just needs to add covariates to the subgroup model. The same idea can be applied to the outcome, where it is well-known that predictors can increase the power of randomized trials [11].

To examine this, another set of simulations were run, where predictors of treatability and/or the outcome were added. As an example, the model for numerical outcomes with both kinds of predictors is

$$\begin{cases} \text{logit } \pi = \alpha_S + \beta_S x_S, \\ Y = \alpha_0 + \beta_0 x_y + \epsilon, & G = 0, \\ Y = \alpha_1 + \beta_1 x_y + \psi R + \epsilon, & G = 1, \end{cases}$$

Table 1 Simulation results, proof of concept

True ψ	n	π	Dichotomous				Ordinal				Time to event			
			Numerical				Ordinal				Time to event			
			med $\hat{\psi}$	CI coverage	Power	med $\hat{\psi}$	CI coverage	Power	med $\hat{\psi}$	CI coverage	Power	med $\hat{\psi}$	CI coverage	Power
0	500	10%	0.07	88.4%	11.6%	-0.02	99.9%	0.1%	0.06	96.8%	3.2%	0.19	92.4%	7.6%
		25%	0.00	93.4%	6.6%	0.00	97.2%	2.8%	0.01	93.7%	6.3%	0.05	92.2%	7.8%
		60%	0.00	94.9%	5.1%	0.00	95.7%	4.3%	0.00	95.4%	4.6%	0.03	94.4%	5.6%
	3000	10%	0.01	93.4%	6.6%	-0.04	97.0%	3.0%	0.02	92.2%	7.8%	0.07	90.6%	9.4%
		25%	0.00	94.8%	5.2%	0.00	96.3%	3.7%	0.00	94.7%	5.3%	0.04	93.7%	6.3%
		60%	0.00	95.0%	5.0%	0.00	94.9%	5.1%	0.00	95.2%	4.8%	0.03	93.8%	6.2%
-0.2	500	10%	-0.12	88.5%	11.8%	-0.26	99.9%	0.2%	-0.15	97.0%	3.7%	0.00	92.1%	9.4%
		25%	-0.21	93.5%	7.8%	-0.21	97.3%	5.5%	-0.19	93.9%	7.3%	0.00	92.1%	10.1%
		60%	-0.20	94.6%	9.2%	-0.20	95.9%	9.6%	-0.20	94.9%	11.0%	-0.17	94.1%	13.3%
	3000	10%	-0.20	93.6%	8.2%	-0.23	97.0%	6.5%	-0.19	92.4%	9.1%	-0.12	90.3%	11.2%
		25%	-0.20	95.0%	11.2%	-0.20	96.4%	11.6%	-0.20	94.9%	12.3%	-0.16	93.7%	13.0%
		60%	-0.20	95.0%	30.2%	-0.20	95.4%	32.1%	-0.20	95.1%	43.6%	-0.17	93.7%	50.2%
-0.4	500	10%	-0.36	88.6%	12.7%	-0.49	99.9%	0.4%	-0.36	97.4%	4.4%	-0.21	92.6%	11.2%
		25%	-0.40	93.3%	10.7%	-0.41	97.2%	9.9%	-0.39	93.6%	10.7%	-0.35	91.6%	15.5%
		60%	-0.40	94.8%	21.8%	-0.40	96.0%	23.0%	-0.40	95.3%	30.4%	-0.37	94.2%	40.7%
	3000	10%	-0.40	93.3%	12.4%	-0.44	97.0%	11.1%	-0.39	92.1%	12.9%	-0.32	89.9%	16.6%
		25%	-0.40	94.7%	30.1%	-0.40	96.3%	27.6%	-0.40	94.9%	33.2%	-0.36	93.8%	40.3%
		60%	-0.40	95.3%	82.5%	-0.40	95.3%	81.9%	-0.40	94.7%	94.6%	-0.37	93.4%	98.7%

The table shows the median estimate $\hat{\psi}$, 95% confidence interval coverage rate, and power for different values of true treatment effect ψ , sample size n , latent subgroup prevalence π , and outcome type. Under no treatment effect ($\psi = 0$), the power is the type I error rate

where $\text{logit}(p) = \log(p/(1 - p))$ and $\pi = P(G = 1)$. Here, x_S and x_y are predictors of subgroup membership and outcome. Some care must be taken not to alter the treatability prevalence due to non-collapsibility [12]. We used the methods of Lindhagen et al. [13] to modify α_S accordingly. Similarly, the standard deviations of ϵ were decreased to maintain the marginal outcome standard deviations.

We ran models without treatability predictors, along with models with a weak ($\beta_S = 0.5$) and strong ($\beta_S = 1$) predictor. In an additional scenario, the subgroup was fully identified (perfect prediction). Although not realistic in itself, this serves as a benchmark, indicating the maximum possible gain from prediction. The x 's were standard normal, $\beta_0 = \beta_1 = 0.5$, and the simulations were confined to $n = 3000$, $\pi = 25\%$, $\psi \in \{0, -0.4\}$, and to numerical and dichotomous outcomes.

Non-collapsibility issues also arise when adjusting for x_y for dichotomous outcomes. In order to maintain the marginal event rates, the regression coefficient was changed from its nominal value $\psi = -0.40$ to a "true" value $\psi = -0.42$. An unbiased estimate should converge to this.

The results are presented in Table 2. It can be seen that there is a substantial potential for increased power if the latent subgroup could be identified (perfect prediction), roughly from 30% to 60% in this example. Our predictors of course did not attain this, but still the power increased to about 40% when using the strong predictor. There is also a moderate increase in power due to outcome prediction.

Table 2 Simulation results, predictors of treatability and outcome

Nominal ψ	Outcome prediction	Treatability prediction	Numerical			Dichotomous			
			med $\hat{\psi}$	CI coverage	Power	True ψ	med $\hat{\psi}$	CI coverage	Power
0	No	None	0.00	94.8%	5.2%	0	0.00	96.3%	3.7%
		Weak	0.00	94.9%	5.1%	0	0.00	95.8%	4.2%
		Strong	0.00	95.0%	5.0%	0	0.00	95.6%	4.4%
		Perfect	0.00	95.1%	4.9%	0	0.00	94.9%	5.1%
	Yes	None	0.00	94.7%	5.3%	0	0.00	96.2%	3.8%
		Weak	0.00	95.0%	5.0%	0	0.00	95.6%	4.4%
		Strong	0.00	94.9%	5.1%	0	0.00	95.4%	4.6%
		Perfect	0.00	94.9%	5.1%	0	0.00	95.1%	4.9%
-0.4	No	None	-0.40	94.7%	30.1%	-0.40	-0.40	96.3%	27.6%
		Weak	-0.39	94.7%	32.1%	-0.40	-0.40	95.9%	31.0%
		Strong	-0.40	94.8%	38.4%	-0.40	-0.40	95.3%	38.3%
		Perfect	-0.40	94.9%	59.3%	-0.40	-0.40	94.9%	63.7%
	Yes	None	-0.40	94.7%	31.4%	-0.42	-0.42	96.2%	28.4%
		Weak	-0.40	94.9%	34.1%	-0.42	-0.42	95.5%	32.3%
		Strong	-0.40	94.8%	40.2%	-0.42	-0.42	95.4%	38.5%
		Perfect	-0.40	94.8%	61.0%	-0.42	-0.42	95.0%	65.2%

The table shows the median estimate $\hat{\psi}$, 95% confidence interval coverage rate, and power, with and without predictors. The sample size was $n = 3000$ and the treatability prevalence $\pi = 25\%$. When using predictors of dichotomous outcomes, the treatment effect ψ was modified due to non-collapsibility. Under no treatment effect ($\psi = 0$), the power is the type I error rate

Throughout, the type I error rate was close to nominal, again with somewhat poorer results for dichotomous outcomes.

In conclusion, the simulation suggests a considerable benefit from treatability prediction. An important question is then whether this is associated with a risk. Could, for example, the type I error rate increase if the subgroup model is misspecified? This is the topic of the next subsection.

3.3 Misspecified Subgroup Models

To study the consequences of misspecified subgroup models, a third set of simulations was run, with three kinds of misspecifications. The first was a missed non-linearity, generating subgroup data according to

$$\text{logit } \pi = \alpha + e^{\beta x}.$$

The second was an incorrect link function, where data were simulated using the complementary log-log link:

$$\log(-\log(1 - \pi)) = \alpha + \beta x.$$

Finally, a missed interaction was studied, generating data by

$$\text{logit } \pi = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2.$$

Data were subsequently analyzed using the (incorrect) model $\text{logit } \pi = \alpha + \beta x$ or, for interactions, $\text{logit } \pi = \alpha + \beta_1 x_1 + \beta_2 x_2$.

The x variables were standard normal, all β 's were set to 0.5, and the α 's were determined numerically to get the desired prevalence $\pi = 25\%$. Moreover, $n = 3000$, $\psi \in \{0, -0.4\}$, and the outcomes were numerical or dichotomous.

The results are shown in Table 3. No notable problems are evident: the power still increases when predictors are included, maintaining close to nominal CI coverage and type I error rate. If anything, the latter improves when adding the predictors, despite the model misspecification. In summary, the framework seems relatively robust to subgroup model misspecifications.

4 Application to a Prostate Cancer Trial

We illustrate the use of our framework on data from a prostate cancer trial. In Scandinavian Prostate Cancer Group Study Number 4 (SPCG-4), men with clinically localized prostate cancer (PCa) were randomized to surgery (radical prostatectomy; RP) or conservative treatment (watchful waiting; WW). The trial included 695 men diagnosed between 1989 and 1999. A 29-year follow-up analysis was published in 2018 [14].

In the RP arm, the surgery procedure began by removal and histological examination of lymph nodes. If lymph node metastases were detected (node-positive PCa), the patient was deemed non-treatable, surgery was aborted, and androgen

Table 3 Simulation results, misspecified subgroup models

Misspecification	True ψ	Treatability prediction	Numerical			Dichotomous		
			med $\hat{\psi}$	CI coverage	Power	med $\hat{\psi}$	CI coverage	Power
Unhandled non-linearity	0	No	0.00	95.0%	5.0%	0.00	96.3%	3.7%
		Yes	0.00	95.0%	5.0%	-0.01	95.5%	4.5%
	-0.4	No	-0.40	94.7%	30.2%	-0.40	96.3%	27.2%
		Yes	-0.40	95.0%	36.5%	-0.41	95.1%	37.2%
Incorrect link	0	No	0.00	94.9%	5.1%	0.00	96.5%	3.5%
		Yes	0.00	94.8%	5.2%	0.00	95.9%	4.1%
	-0.4	No	-0.40	94.8%	29.8%	-0.40	96.2%	27.1%
		Yes	-0.40	95.0%	33.8%	-0.40	95.8%	31.6%
Missed interaction	0	No	0.00	95.2%	4.8%	0.00	96.2%	3.8%
		Yes	0.00	94.9%	5.1%	0.00	95.4%	4.6%
	-0.4	No	-0.41	94.9%	30.5%	-0.40	96.3%	27.7%
		Yes	-0.40	94.9%	36.4%	-0.41	95.4%	36.7%

The table shows the median estimate $\hat{\psi}$, 95% confidence interval coverage rate, and power for three kinds of subgroup model misspecifications. The sample size was $n = 3000$ and the treatability prevalence $\pi = 25\%$. No outcome predictors were used. Under no treatment effect ($\psi = 0$), the power is the type I error rate

deprivation therapy (ADT) was initiated. This happened for 23 of the 347 men in the RP arm (6.6%). As a consequence, SPCG-4 has the latent subgroup structure that our framework is designed to deal with. Namely, there is a subgroup of patients (node-negative ones), who can receive the intended treatment (surgery). This subgroup is identified only in the RP arm, and is hence latent. While the ITT analysis estimates the effect of a surgery strategy on the total population (programmatic effectiveness), it is natural to also ask for the effect of actual surgery among treatable patients (biological efficacy). To assess this, we shall re-analyze the trial data, applying our framework to the endpoint death from PCa, with 181 events.

We first need to consider the instrumental variable assumptions, in particular exclusion restriction, i.e. the statement that non-treatable patients are unaffected by attempted surgery. This can be called into question in SPCG-4, since node-positive patients in the RP arm received ADT already upon detection of the metastases. However, WW patients actually received ADT almost as fast as RP patients; the proportions were similar already after one year. In addition, the effect of ADT is probably not dramatic. A trial from 2004 [15] reported a hazard ratio of 1.23 for mortality in delayed vs. immediate ADT in node-positive PCa patients. In conclusion, there will likely be a bias due to violation of exclusion restriction, but it can be expected to be minor.

Cumulative incidence curves for the two trial arms are presented in Panel a of Fig. 3 (solid lines), suggesting a clear benefit from surgery. Our first goal is to produce similar curves for treatable patients. For the RP arm, this is straightforward; simple stratification gives curves for treatable (dashed line) and non-treatable (gray line) patients. Unsurprisingly, the prognosis of non-treatable patients is very poor.

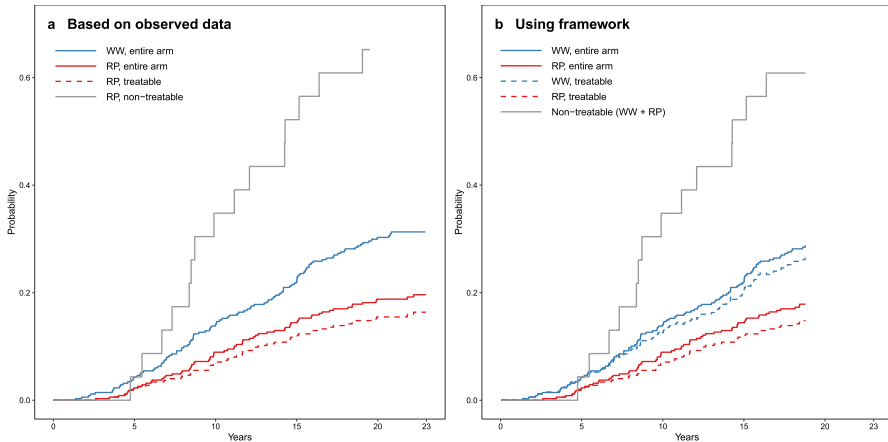


Fig. 3 Cumulative incidence curves for death from prostate cancer in the SPCG-4 trial. The curves in Panel **a** come directly from observed data, whereas those of Panel **b** were estimated using the framework of this paper. The four curves in Panel **a** are unchanged in Panel **b**

Moreover, the treatable subgroup is fairly similar to the entire RP arm, since most patients are treatable. We next applied our framework to execute a similar split of the WW curve, using a trick that works as long as we have complete follow-up, 19 years in this case. The idea is to perform a separate analysis at each time point. In the absence of censorings, we know who has died from PCa and who has not (patients who have died from other causes are treated like survivors [16]). At a fixed time point, we therefore have dichotomous outcome data, and can use the framework with three logistic regression models. The subgroup model and the outcome model for non-treatable patients contained only an intercept, whereas the outcome of treatable patients depended on trial arm. These models estimate the prevalence of treatability, the risk of PCa death for non-treatable patients (common value for both arms; exclusion restriction), and the same risk for treatable patients, per trial arm. The results are shown as dashed and gray lines in Panel b of Fig. 3. From these, solid curves for the full trial arms were computed as weighted means. Although all five curves in Panel b have been computed from models, the four observable ones are identical to those in Panel a. So what Panel b really contributes is the curve for treatable patients in the WW arm. Slightly disturbingly, this curve is not monotonic. Rather, it decreases at each non-treatable RP event, since a larger mortality among non-treatable patients makes the model more inclined to attribute WW deaths to the non-treatable subgroup. This anomaly could be removed by ad hoc methods, such as replacing decreased values by the value before the event.

We turn to the statistical analysis. The 29-year paper applied a Cox model, giving a hazard ratio (HR) of 0.55. To get comparable results, we loaded our framework with a parametric analogue of the Cox model, the flexible spline model of Royston and Parmar (Appendix 1.3). We used 4 knots for treatable patients and 2 knots for non-treatable ones; the results were very insensitive to the number and placing of knots. With logistic regression for the subgroup model, we found $HR = 0.50$

(Table 4, top row). The interpretation is that the effect of surgery, when it can actually be performed, is $HR = 0.50$ (biological efficacy). This is what would be seen in an “ideal” trial, with only treatable patients. In SPCG-4, this effect has been diluted by the inclusion of node-positive patients, who cannot receive surgery, resulting in $HR = 0.55$. The latter is a measure of the effect of a surgery strategy (programmatic effectiveness). The difference between the two is modest, since lymph node metastases are rare.

We conclude with a discussion of covariate adjustment. The Cox model and the outcome model for treatable patients were both adjusted for age and prostate-specific antigen (PSA) using quadratic terms, and for tumor stage (T1b, T1c, and T2). Due to the small number of such patients, the outcome model for non-treatable patients was adjusted only for PSA (linear term). The results are shown in the second row of Table 4. As expected from non-collapsibility, the adjusted hazard ratios are further from unity. Since the standard errors have only increased slightly, this indicates a substantial power gain, in agreement with the simulations in Sect. 3.2. Finally, we consider predictors of treatability. Due to the limited number of non-treatable patients, we again adjusted only for PSA (linear term), leading to the third row in Table 4. Unlike the simulation in Sect. 3.2, this did not reduce the standard errors noticeably, despite PSA being a strong predictor of metastases. It is not hard to understand why: since the vast majority (93%) of the patients are treatable, this latent subgroup is close to identified already, and predictors cannot be expected to make a major difference.

5 Discussion

We have presented a modular framework for effect estimation in a latent subgroup of treatable patients, identified only in one of the arms of a randomized trial. This situation can arise in several ways. A typical example is a screening trial, where patients are randomized to a screening test for a medical condition, with subsequent treatment if positive. The framework consists of a core and three plug-in models, for subgroup membership, and for outcomes of treatable and non-treatable patients. The core performs

Table 4 Analysis of death from prostate cancer in the SPCG-4 trial

Adjustment	All patients (Cox model)	Treatable patients (using framework)
	HR (95% CI)	HR (95% CI)
None	0.55 (0.41, 0.75)	0.50 (0.36, 0.70)
Outcome	0.52 (0.38, 0.70)	0.46 (0.32, 0.65)
Outcome + treatability	–	0.45 (0.32, 0.64)

The results for all patients come from ordinary Cox models. For the latent subgroup of treatable patients, the framework of this paper has been applied, with flexible spline models for the outcome. Covariate adjustment of the outcome model for treatable patients was done using age and PSA (quadratic terms), and tumor stage. The outcome model for non-treatable patients and the subgroup model were adjusted only for PSA (linear term)

maximum likelihood estimation using the EM algorithm, and computes standard errors using asymptotic closed-form expressions, while being fully ignorant about the details of the plug-in models. The latter can be of any form, and handle any type of outcome, as long as they are parametric. This structure gives the user a large flexibility, for example, predictors can easily be added. The methods have also been implemented in an R package, available at GitHub.

The framework have been validated in an extensive simulation, demonstrating asymptotic unbiasedness and correct confidence interval coverage. Including predictors of the outcome increased the power, just like for traditional randomized trials. The latent subgroup setting also allows for predictors of membership of the latent subgroup itself. If, say, men are more often treatable than women, information about sex should mean a step towards identifying the subgroup already from baseline data. Our simulations indicate that the potential power gain from such predictors can be substantial, and also that the framework is robust to misspecification of the subgroup model. The latter is in line with conventional trials, where the analysis model has been seen to be robust against misspecification [17]. For these reasons, we recommend making liberal use of both subgroup and outcome predictors.

We have also applied the framework to a prostate cancer trial, where surgery was discontinued for lymph node positive (non-treatable) patients. As expected, the treatment effect among treatable patients (biological efficacy) was seen to be stronger than the effect of intended surgery in the entire trial population (programmatic effectiveness), due to a dilution of the latter by non-treatable patients.

The EM algorithm is a convenient method to perform maximum likelihood computations in the presence of missing data, in our case treatability in the control arm. However, in the present situation it is possible to write the likelihood in a computationally tractable form, cf. Eq. (5). One could therefore consider maximizing it directly, using e.g. a quasi-Newton method [18]. Simulations suggest that this works, but that the computations are slower and more sensitive to the initial values, in line with other findings [19]. For these reasons, EM seems preferable.

In standard instrumental variable language [20], we only have two principal strata, never-takers and compliers. It can be noted that monotonicity also holds in the presence of a third stratum of always-takers, patients who receive treatment regardless of randomization. It is possible to extend the framework to such a situation, although it would complicate matters a bit. Subgroup membership is then identified for patients in both arms: always-takers in the control arm and never-takers in the intervention arm. The extension would require a third outcome model for always-takers, for a total of four plug-in models. In addition, the subgroup model needs to handle a three-level response, perhaps as a multinomial logistic regression model. The calculations of Sects. 2.4–2.5 only need minor modifications. For example, the weights w_i and $(1 - w_i)$ are to be replaced by three numbers, representing the posterior probabilities of the three principal strata.

Appendix 1. Examples of Plug-In Models

This appendix presents examples of plug-in models for the most important data types. For these cases, weighted regression is easily performed via standard software, so by Sect. 2.6 we need only derive expressions for the likelihood and its first two derivatives. In practice, it is often more convenient to do this for the log-likelihood rather than the likelihood itself, and apply logarithmic differentiation (1).

Following the notation of Sect. 2.2, we let $\boldsymbol{\vartheta}$ denote the model parameters. Thus, $\boldsymbol{\vartheta}$ may in this appendix refer either to $\boldsymbol{\vartheta}_S$, $\boldsymbol{\vartheta}_0$, or $\boldsymbol{\vartheta}_1$. We shall also consistently suppress the patient index i , understanding that the results are to be summed over patients. Moreover, p will denote the dimensionality of the covariate vector $\mathbf{x} \in \mathbf{R}^p$. As usual, \mathbf{x} may encompass an intercept, as well as interactions, non-linearities etc. Asterisks (*) in Hessians mean that the matrix is to be made symmetric.

1.1. Generalized Linear Models

In a generalized linear model, the outcome Y follows a distribution from the exponential family. Its expected value μ is related to the covariates \mathbf{x} via a link function g :

$$g(\mu) = \boldsymbol{\beta} \cdot \mathbf{x} =: \eta, \quad \boldsymbol{\beta} \in \mathbf{R}^p.$$

In addition to the regression coefficients $\boldsymbol{\beta}$, there may be additional "nuisance" parameters. Since our framework requires fully parametric models, such parameters must be included in the parameter vector $\boldsymbol{\vartheta}$, and need to be handled on a case to case basis. To begin with, we restrict ourselves to the situation of no nuisance parameters ($\boldsymbol{\vartheta} = \boldsymbol{\beta}$), which includes models for dichotomous and count data. An example of how to handle nuisance parameters is given in Sect. 1.1.1.

While the likelihood $M(\boldsymbol{\beta})$ depends on the exact nature of the outcome distribution, there is, for the exponential family, a general expression for the Jacobian of the log-likelihood [21]:

$$\nabla \log M = \frac{y - \mu}{\text{Var}(Y)} \frac{d\mu}{d\eta} \mathbf{x}.$$

Moreover, the variance can be expressed in terms of the expectation: $\text{Var}(y) = v(\mu)$ for some function v . Now, $\nabla \eta = \mathbf{x}$, whence

$$\nabla = \mathbf{x} \frac{d}{d\eta} = \mathbf{x} \frac{d\mu}{d\eta} \frac{d}{d\mu}.$$

Straightforward calculations then show that

$$\nabla \nabla \log M = -\frac{1}{v(\mu)} \left(\frac{d\mu}{d\eta} \right)^2 \mathbf{x}^{\otimes 2} + \frac{y - \mu}{v(\mu)} \left[\frac{d^2 \mu}{d\eta^2} - \frac{v'(\mu)}{v(\mu)} \left(\frac{d\mu}{d\eta} \right)^2 \right] \mathbf{x}^{\otimes 2}. \quad (11)$$

Table 5 Outcome distributions for the generalized linear model

Distribution	Likelihood M	Variance	Common links
Dichotomous (0/1)	$1 - \mu, \quad y = 0,$ $\mu, \quad y = 1$	$v(\mu) = \mu(1 - \mu),$ $v'(\mu) = 1 - 2\mu$	Logit, probit, cauchit, cloglog, log, identity
Poisson	$\mu^y e^{-\mu} / y!$	$v(\mu) = \mu, v'(\mu) = 1$	Log, identity, square root

For an outcome Y , μ is the expectation, and $v(\mu)$ the variance

To make use of this, expressions for $v(\mu)$ etc. are needed. Table 5 summarizes the likelihood and variance for dichotomous outcomes and Poisson models. In the former case, $y \in \{0, 1\}$, and μ is just the event probability. For the most popular links, the derivatives $d\mu/d\eta$ and $d^2\mu/d\eta^2$ are presented in Table 6.

1.1.1. Nuisance Parameters—The Linear Model

As an example of how to incorporate nuisance parameters, we consider the linear model $Y = \beta \cdot x + \epsilon$, where the residuals ϵ are centered normal with variance σ^2 , the nuisance parameter. The full parameter vector is then $\vartheta = (\beta, \sigma^2) \in \mathbb{R}^{p+1}$, and the derivatives will be written as block matrices accordingly.

The likelihood of a data point y is

$$M = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \beta \cdot x)^2}{2\sigma^2}\right).$$

Taking logs and differentiating, we find

$$\log M = \text{const} - \frac{1}{2} \log \sigma^2 - \frac{(y - \beta \cdot x)^2}{2\sigma^2},$$

Table 6 Common links for the generalized linear model

Link	$g(\mu)$	$\frac{d\mu}{d\eta}$	$\frac{d^2\mu}{d\eta^2}$
Logit	$\log \frac{\mu}{1-\mu}$	$\frac{1}{4 \cosh^2(\eta/2)}$	$-\frac{\tanh(\eta/2)}{4 \cosh^2(\eta/2)}$
Probit	$\Phi^{-1}(\eta)$	$\phi(\eta)$	$-\eta\phi(\eta)$
Cauchit	$\tan\left(\pi\mu - \frac{\pi}{2}\right)$	$\frac{1}{\pi(1+\eta^2)}$	$\frac{-2\eta}{\pi(1+\eta^2)^2}$
cloglog	$\log(-\log(1 - \mu))$	$\exp(-e^\eta)e^\eta$	$\exp(-e^\eta)e^\eta(1 - e^\eta)$
Log	$\log \mu$	e^η	e^η
Identity	μ	1	0
Square root	$\sqrt{\mu}$	2η	2

The link function g connects the expected outcome μ and the regression coefficients β via $g(\mu) = \beta \cdot x = \eta$. For the probit link, ϕ and Φ denote the standard normal density and cumulative distribution functions

whence

$$\nabla \log M = \left[\frac{\nabla_{\beta} \log M}{\frac{\partial \log M}{\partial \sigma^2}} \right] = \left[\begin{array}{c} \left(\frac{y-\beta \cdot \mathbf{x}}{\sigma^2} \right) \mathbf{x} \\ -\frac{1}{2\sigma^2} + \frac{(y-\beta \cdot \mathbf{x})^2}{2\sigma^4} \end{array} \right]$$

and

$$\nabla \nabla \log M = \left[\begin{array}{c|c} \left(-\frac{1}{\sigma^2} \right) \mathbf{x}^{\otimes 2} & (*) \\ \hline \left(\frac{y-\beta \cdot \mathbf{x}}{\sigma^4} \right) \mathbf{x} & \frac{1}{2\sigma^4} - \frac{(y-\beta \cdot \mathbf{x})^2}{\sigma^6} \end{array} \right].$$

1.2. Proportional Odds Model for Ordinal Data

Next, consider an ordinal response with $K \geq 3$ levels $1, 2, \dots, K$. The proportional odds model arises by assuming the effect of a covariates, as measured by odds ratios, to be the same for all $K - 1$ possible dichotomizations along the ordinal scale. In other words,

$$P(Y \leq k) = (1 + e^{\alpha_k + \beta \cdot \mathbf{x}})^{-1}, \quad 1 \leq k \leq K - 1, \tag{12}$$

where $\alpha_1 > \alpha_2 > \dots > \alpha_{K-1}$ is a set of intercepts, and β are the usual regression coefficients. The full set of model parameters is $\boldsymbol{\vartheta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbf{R}^{K-1+p}$.

To write this as a tractable likelihood, we introduce some additional notation. First, let $\alpha_K = -\infty$, making (12) valid also for $k = K$. Defining $v_k = (\alpha_k + \beta \cdot \mathbf{x})/2$, the outcome probabilities can be written as

$$\begin{cases} P(Y = 1) = (1 + e^{2v_1})^{-1}, \\ P(Y = k) = (1 + e^{2v_k})^{-1} - (1 + e^{2v_{k-1}})^{-1}, \quad 2 \leq k \leq K. \end{cases} \tag{13}$$

This works also for $k = K$, since the first term is then unity ($\alpha_K = -\infty$).

Now, let $y \in \{1, 2, \dots, K\}$ be an observed outcome. We can simplify (13) further by defining the numbers ϵ_k as $\epsilon_k = 1$ if $k = y$, $\epsilon_k = -1$ if $k = y - 1$, and $\epsilon_k = 0$ otherwise. Using this notation, the likelihood can be written as

$$M = M(y) = \sum_{k=1}^K \frac{\epsilon_k}{1 + e^{2v_k}}.$$

The derivatives can also be written compactly by also letting

$$A_k = -\frac{1}{4 \cosh^2 v_k}, \quad B_k = \frac{\tanh v_k}{4 \cosh^2 v_k}, \quad 1 \leq k \leq K - 1.$$

Straightforward computations then show that the Jacobian and Hessian are given by

$$\nabla M = \begin{bmatrix} \frac{\nabla_{\alpha} M}{\nabla_{\beta} M} \end{bmatrix} = \begin{bmatrix} \epsilon_1 A_1 \\ \vdots \\ \frac{\epsilon_{K-1} A_{K-1}}{(\sum_k \epsilon_k A_k) \mathbf{x}} \end{bmatrix}$$

and

$$\nabla \nabla M = \left[\begin{array}{cc|c} \epsilon_1 B_1 & & 0 \\ & \ddots & \\ 0 & & \epsilon_{K-1} B_{K-1} & (*) \\ \hline \epsilon_1 B_1 \mathbf{x} & & \epsilon_{K-1} B_{K-1} \mathbf{x} & (\sum_k \epsilon_k B_k) \mathbf{x}^{\otimes 2} \end{array} \right],$$

where the summations run from $k = 1$ to $k = K - 1$.

1.3 Flexible Spline Model for Time to Event Data

We finally consider the case of right-censored time to event data. The Cox model is not an option for our framework, since the models need to be fully parametric. There is, however, a rich flora of parametric survival models. We shall confine ourselves to what comes perhaps closest to the Cox model, the spline-based proportional hazards model of Royston and Parmar [22].

A starting point for this model is the observation that in a Weibull model, the log cumulative hazard is a linear function of log time. A natural generalization is to replace this linear function by a spline. Thus, if $H(t)$ is the cumulative hazard, the model assumes that

$$\log H(t) = s(\log t; \boldsymbol{\gamma}) + \boldsymbol{\beta} \cdot \mathbf{x}, \tag{14}$$

where s is a natural spline with knots at $k_{\min} < k_1 < \dots < k_m < k_{\max}$ and coefficients $\boldsymbol{\gamma} \in \mathbf{R}^{m+2}$. We shall write the spline as

$$s(u; \boldsymbol{\gamma}) = \sum_{j=0}^{m+1} \gamma_j b_j(u) =: \boldsymbol{\gamma} \cdot \mathbf{b}(u), \tag{15}$$

where $\mathbf{b}(u) = (b_0(u), \dots, b_{m+1}(u))$ is the vector of spline basis functions. The latter are given by $b_0(u) = 1, b_1(u) = u$, and

$$b_{j+1}(u) = (u - k_j)_+^3 - \lambda_j (u - k_{\min})_+^3 - (1 - \lambda_j) (u - k_{\max})_+^3, \quad 1 \leq j \leq m,$$

where $u_+ = \max(u, 0)$ and $\lambda_j = (k_{\max} - k_j) / (k_{\max} - k_{\min})$. So the parameters of the Royston–Parmar model are $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}) \in \mathbf{R}^{m+2+p}$. Since (15) already contains an intercept γ_0 , there is no intercept in $\boldsymbol{\beta}$, similar to the Cox model.

We also need the hazard, which we get by differentiating the spline. Letting $\mathbf{b}'(u)$ be the vector of derivatives of $\mathbf{b}(u)$ (the derivative of $(u - a)_+^3$ is just $3(u - a)_+^2$), the hazard is

$$h(t) = H'(t) = H(t) \frac{\boldsymbol{\gamma} \cdot \mathbf{b}'(\log t)}{t}.$$

A time to event data point is represented by a follow-up time t , an event indicator δ , and a covariate vector \mathbf{x} . The likelihood M of such a data point is $S(t)$ if the data point is censored ($\delta = 0$), and $S(t)h(t)$ otherwise ($\delta = 1$), where $S(t) = e^{-H(t)}$ is the survival function. In other words, $M = e^{-H(t)}h(t)^\delta$ and

$$\log M = -H(t) + \delta \log h(t) = -H(t) + \delta \log H(t) + \delta \log(\boldsymbol{\gamma} \cdot \mathbf{b}'(\log t)) - \delta \log t.$$

Now, it follows from (14) and (15) that $\nabla_{\boldsymbol{\gamma}} H(t) = H(t)\mathbf{b}(\log t)$ and $\nabla_{\boldsymbol{\beta}} H(t) = H(t)\mathbf{x}$. As a consequence, the Jacobian and Hessian of the log-likelihood are given by

$$\nabla \log M = \begin{bmatrix} \nabla_{\boldsymbol{\gamma}} \log M \\ \nabla_{\boldsymbol{\beta}} \log M \end{bmatrix} = \begin{bmatrix} (-H + \delta)\mathbf{b} + \frac{\delta}{\boldsymbol{\gamma} \cdot \mathbf{b}'} \mathbf{b}' \\ (-H + \delta)\mathbf{x} \end{bmatrix}$$

and

$$\nabla \nabla \log M = \begin{bmatrix} -H\mathbf{b}^{\otimes 2} - \frac{\delta}{(\boldsymbol{\gamma} \cdot \mathbf{b}')^2} (\mathbf{b}')^{\otimes 2} & (*) \\ -H\mathbf{b} \otimes \mathbf{x} & -H\mathbf{x}^{\otimes 2} \end{bmatrix},$$

where $H = H(t)$, $\mathbf{b} = \mathbf{b}(\log t)$ etc.

Appendix 2. Example R Code

This appendix gives a simple example how to use the **latent** package, many more can be found in the package vignette. We shall follow the example surrounding Eq. (10). First, we generate data:

```
df <- data.frame(
  g = rbinom(n = 3000, size = 1, prob = 0.25), # n = 3000, pi = 25%.
  rand = rbinom(n = 3000, size = 1, prob = 0.5))
df$y <- ifelse(df$g == 0,
  rnorm(n = 3000, mean = 100, sd = 2),
  rnorm(n = 3000, mean = 100.4 - 0.4 * df$rand, sd = 2.5)) # psi = -0.4.
df$g[df$rand == 0] <- NA
```

The last statement makes the subgroup unobserved in the control arm. Next, we set up the three plug-in models:

```
modS <- latent_glm(  
  formula = (g ~ 1),  
  family_name = "binomial",  
  link = "logit")  
mod0 <- latent_linear(  
  formula = (y ~ 1))  
mod1 <- latent_linear(  
  formula = (y ~ rand))
```

The subgroup model `modS` is a logistic regression model with treatability as response, whereas the outcome models are linear models. Since there are no covariates, all model formulas are trivial, except for the outcome model of treatable patients, which depends on randomization. The analysis is then run by simply calling the main function of the package, passing data and plug-in models as parameters:

```
res <- latent_main(  
  data = df,  
  modelS = modS,  
  model0 = mod0,  
  model1 = mod1)
```

Acknowledgements The simulation was enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala partially funded by the Swedish Research Council through grant agreement no. 2018–05973. The authors also wish to thank Katja Gabrysch and Henrik Renlund for assistance with the R package, and David Robinson for providing expertise on ADT.

Author Contributions L.L. performed the main methodological development, and wrote the paper and the R package. H.G. provided expertise and data on the SPCG-4 trial. O.Ö. initiated the project, providing the main methodological ideas, and guiding the simulation. All authors reviewed the manuscript.

Funding Open access funding provided by Uppsala University. The authors state that they have no funding source for this paper.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of Interest The authors declare no competing interests.

Ethical Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this

article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Sommer, A., Zeger, S.L.: On estimating efficacy from clinical trials. *Stat. Med.* **10**(1), 45–52 (1991)
2. Burgess, S., Small, D.S., Thompson, S.G.: A review of instrumental variable estimators for Mendelian randomization. *Stat. Methods Med. Res.* **26**(5), 2333–2355 (2017)
3. Bowden, R.J., Turkington, D.A.: *Instrumental Variables*. Cambridge University Press (1990)
4. Frangakis, C.E., Rubin, D.B.: Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86**(2), 365–379 (1999)
5. Cuzick, J., Sasieni, P., Myles, J., Tyrer, J.: Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. *J. R. Stat. Soc. Ser. B Stat Methodol.* **69**(4), 565–588 (2007)
6. Altstein, L.L., Li, G., Elashoff, R.M.: A method to estimate treatment efficacy among latent subgroups of a randomized clinical trial. *Stat. Med.* **30**(7), 709–717 (2011)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat Methodol.* **39**(1), 1–22 (1977)
8. Lindhagen L.: latent: Effect estimation in latent subgroups. R package version 1.0.4. <https://github.com/lindhagen/latent>
9. Devika, S., Jeyaseelan, L., Sebastian, G.: Analysis of sparse data in logistic regression in medical research: A newer approach. *J. Postgrad. Med.* **62**(1), 26 (2016)
10. Rainey, C., McCaskey, K.: Estimating logit models with small samples. *Pol. Sci Res Methods.* **9**(3), 549–564 (2021)
11. Kahan, B.C., Jairath, V., Doré, C.J., Morris, T.P.: The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials* **15**(1), 1–7 (2014)
12. Robinson, L.D., Jewell, N.P.: Some surprising results about covariate adjustment in logistic regression models. *Int. Stat. Rev.* **59**(2), 227–240 (1991)
13. Lindhagen, L., Darkahi, B., Sandblom, G., Berglund, L.: Level-adjusted funnel plots based on predicted marginal expectations: an application to prophylactic antibiotics in gallstone surgery. *Stat. Med.* **33**(21), 3655–3675 (2014)
14. Bill-Axelsson, A., Holmberg, L., Garmo, H., Taari, K., Busch, C., Nordling, S., et al.: Radical prostatectomy or watchful waiting in prostate cancer—29-year follow-up. *N. Engl. J. Med.* **379**(24), 2319–2329 (2018)
15. Schröder, F.H., Kurth, K.H., Fosså, S.D., Hoekstra, W., Karthaus, P.P., Debois, M., et al.: Early versus delayed endocrine treatment of pN1-3 M0 prostate cancer without local treatment of the primary tumor: results of European Organisation for the Research and Treatment of Cancer 30846—a phase III study. *J. Urol.* **172**(3), 923–927 (2004)
16. Koller, M.T., Raatz, H., Steyerberg, E.W., Wolbers, M.: Competing risks and the clinical community: irrelevance or ignorance? *Stat. Med.* **31**(11–12), 1089–1097 (2012)
17. Wang, B., Ogburn, E.L., Rosenblum, M.: Analysis of covariance in randomized trials: more precision and valid confidence intervals, without model assumptions. *Biometrics* **75**(4), 1391–1400 (2019)
18. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: *Numerical Recipes: The Art of Scientific Computing*, 3rd edn. Cambridge University Press (2007)
19. Springer, T., Urban, K.: Comparison of the EM algorithm and alternatives. *Numer. Algorithms* **67**(2), 335–364 (2014)
20. Frangakis, C.E., Rubin, D.B.: Principal stratification in causal inference. *Biometrics* **58**(1), 21–29 (2002)
21. Nelder, J.A., Wedderburn, R.W.: Generalized linear models. *J. R. Stat. Soc. Ser. A Stat. Soc.* **135**(3), 370–384 (1972)

22. Royston, P., Parmar, M.K.: Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat. Med.* **21**(15), 2175–2197 (2002)