On the distribution of individual causal effects of binary exposures using latent variable models

Richard A.J. Post
Department of Mathematics and Computer Science, Eindhoven University of Technology
and
Edwin R. van den Heuvel
Department of Mathematics and Computer Science, Eindhoven University of Technology
Department of Preventive Medicine and Epidemiology, Boston University

November 1, 2022

Abstract

In recent years the field of causal inference from observational data has emerged rapidly. This literature has focused on (conditional) average causal effect estimation. When (remaining) variability of individual causal effects (ICEs) is considerable, average effects may be less informative, and possibly misleading for an individual. The fundamental problem of causal inference precludes the estimation of the joint distribution of potential outcomes without making assumptions, while this distribution is necessary to describe the heterogeneity of causal effects. In this paper we describe these assumptions and present a family of flexible latent variable models that can be used to study individual effect modification and estimate the ICE distribution from cross-sectional data. We will also discuss how the distribution is affected by misspecification of the error distribution or ignoring possible confounding-effect heterogeneity. How latent variable models can be applied and validated in practice is illustrated in a case study on the effect of Hepatic Steatosis on a clinical precursor to heart failure. Assuming that there is (i) no unmeasured confounding and (ii) independence of the individual effect modifier and the potential outcome under no exposure, we conclude that the individual causal effect distribution deviates from Gaussian. We estimate that the ‘treatment’ benefit rate in the population is 23.7% (95% Bayesian credible interval: 2.6%, 53.7%) despite a harming average effect.

Keywords: Causal inference, Heterogeneity of treatment effects, Precision medicine, Bayesian analysis, Linear mixed model
1 Introduction

The main result of an epidemiological study is often summarized by an average treatment effect (ATE). As a consequence, the ATE might (subconsciously) be interpreted as the causal effect for each individual while individuals may react differently to exposures. These individual causal effects (ICEs) can be highly variable and may even have opposite signs (Greenland et al. 2020, Hand 1992). Therefore, one would like to beyond mean estimation and draw inference on the distribution of the ICEs and quantify characteristics like the treatment benefit rate (TBR) and treatment harm rate (THR). Such inference involves the joint distribution of the potential outcomes of an individual, while it is well known that the pair of potential outcomes cannot be observed simultaneously (Holland 1986). This fundamental problem of causal inference makes it impossible to estimate the ICEs without making strong assumptions. However, since effect heterogeneity leads to different outcome distributions in the treated and untreated group it can be possible to learn characteristics of the distribution of the ICEs from cross-sectional data. For example, as we will explain further in this work, when the variances of exposed and unexposed individuals are different in a well-executed randomized controlled trial (RCT), effect homogeneity can be ruled out.

Nowadays, the variability in ICEs is typically studied by estimation of conditional average treatment effects (CATEs), assuming the effect is modified by certain observed features. Standard causal inference methods can typically include covariates to account for such effect modification, e.g. strata-specific marginal structural models (Robins et al. 2000), and many machine learning (ML) algorithms have been proposed for flexible CATE estimation, see Caron et al. (2022) for an elaborate review. Particularly in this ML literature a fine-granular CATE, i.e., based on many observed features, is sometimes viewed as equivalent to the ICE, see for instance Lu et al. (2018). Despite the increasing number of covariates
collected in Epidemiological studies nowadays, there may still remain (important) effect modifiers unmeasured such that the personalized CATE deviates from the personal ICE. There is less work on other characteristics, than the mean, of the (conditional) ICE distribution. Individualized $1 - \alpha$ prediction intervals for ICEs have been derived under the assumption of conditional (on measured features) independence between the potential outcomes of an individual (Lei & Candès 2021). In case of remaining dependence between individual’s potential outcomes, estimation of the (conditional) percentage of individuals that would benefit from a exposure is complicated but bounds have been obtained (Tian & Pearl 2000, Li & Pearl 2019, Mueller et al. 2022). Furthermore, assuming conditional independence of the ICE and the potential outcome under no exposure, parametric latent variable models have been proposed to estimate the TBR (Yin et al. 2018) and the (conditional) ICE distribution (Shahn & Madigan 2017).

In this paper we explain under which assumptions the (conditional) ICE distribution becomes identifiable and thus when latent variable models could suffice to estimate the ICE distribution. We will present a semi-parametric linear mixed model (LMM) to estimate the ICE distribution and discuss the complications in estimation that result from heterogeneity in the effect of confounders on the outcome. Firstly, we will illustrate how and under which assumptions the variance and distribution of the ICE can be identified from a RCT and from observational data respectively in Section 2. The semi-parametric causal mixed models as well as possible ways of fitting these models are presented in Section 3. To illustrate how the methods presented in this paper could be used in practice, a case study using the Framingham Heart Study is presented in Section 4 including detailed discussion on the modelling considerations. Finally, we reflect on the importance, limitations and future of the presented work in Section 5.
2 Identifiability of ICE heterogeneity

2.1 Notation and framework

Probability distributions of factual and counterfactual outcomes are defined in terms of the potential outcome framework [Rubin 1974, Splawa-Neyman et al. 1990]. Let $Y_i$ and $A_i$ represent the (factual) stochastic outcome and the random exposure assignment level of individual $i$. Let $Y^a_i$ equal the potential outcome (counterfactual when $A_i \neq a$) under an intervention on the exposure to level $a$. We will consider only two exposure levels, $a \in \{0, 1\}$, with 0 indicating no exposure. The individual causal effect of an arbitrary individual $i$ from the population of interest is defined as $Y^a_i - Y_i^0$ [Hernán & Robins 2020]. Throughout this work we will assume causal consistency: $Y^a_i = Y_i | A_i = a$, implying that potential outcomes are independent of the exposure levels of other individuals (no interference) [Hernán & Robins 2020]. Furthermore, we will assume that for all levels of measured features the probability of receiving exposure is bounded away from 0 and 1, known as positivity [Hernán & Robins 2020].

Potential outcomes under no exposure, $Y^0$, will differ among individuals caused by differences in individual characteristics that also result in dependence between $Y^0$ and $Y^1$. The random variable $N_Y$ will be used to represent such differences. Also the ICE can differ among individuals and we will use the random variable $U$ to refer to these differences. Studying causal inference from observational data requires us to pay attention to confounders, $L$, that cause both the exposure assignment $A$ as well as the potential outcomes $Y^a_i$. Individuals with the same level of confounders can have a different exposure assignment, accommodated by the variable $N_A$. With the introduction of these random variables, a system of cause-effect relations can be parameterized as a structural causal
model (SCM), see e.g. Peters et al. (2018). It consists of a joint probability distribution for \((N_A, N_Y, L, U)\) and a collection of structural assignments \((f_A, f_Y)\) such that

\[
\begin{align*}
A_i &:= f_A(N_{Ai}, L_i) \\
Y_i^a &:= f_Y(N_{Y_i}, L_i, U_i, a),
\end{align*}
\]

(1)

where \(Y_i^0 \perp \perp U_i \mid L_i, N_{Y_i}\). Note that the data generating mechanism is described by this SCM as \(Y_i^{A_i} = Y_i\). For data obtained from a RCT there are no confounders \(L\) as a result of the randomization and thus \(Y^a \perp \perp A\), which is often referred to as exchangeability.

Causal-effect heterogeneity is often addressed by studying the CATEs in subpopulations defined by measured characteristics \(M\) (e.g. sex, age or co-morbidity) that might be effect modifiers. These modifiers may or may not (completely) overlap with the set of confounders \(L\). Pure confounders do affect the distributions of potential outcomes and of the exposure assignment but not the ICE distribution. The SCM (1) could also be parameterized by including details on the effect modifiers \(M\) that affect the ICE distribution, i.e.

\[
\begin{align*}
A_i &:= f_A(N_{Ai}, L_i) \\
Y_i^a &:= f_Y(N_{Y_i}, L_i, M_i, U_i, a),
\end{align*}
\]

(2)

where \(Y_i^0 \perp \perp U_i \mid M_i, L_i, N_{Y_i}\). The \(U_i\) in (1) was decomposed in measured modifiers \(M_i\) and remaining unmeasured modifiers \(U_i\) in (2). In SCM (2), \(U_i\) and \(M_i\) might still be dependent, e.g. when the variability in ICE varies among strata with different levels of \(M_i\).

For observational data, causal inference requires absence of unmeasured confounding, i.e. \(N_Y, U \perp \perp N_A\), such that \(Y^a \perp \perp A \mid M, L\). The latter is referred to as conditional ex-
changeability (Hernán & Robins 2020). Then, using the nowadays available causal methods such as inverse probability of treatment weighting (IPTW) to adjust for the confounders unbiased CATE estimates can be obtained (Hernán & Robins 2020). Note that $Y^1 - Y^0 \mid M$ is independent of $L$.

In absence of unmeasured modifiers, there is no remaining effect heterogeneity in sub-populations defined by levels of $M$, i.e., the conditional distribution of $Y^1 - Y^0 \mid M = m$ is degenerate in the CATE equal to $\theta_1(m)$. In general, $\theta_1(m)$ is informative at the individual level when there is no or limited variability remaining in the $M = m$ subpopulation. An example with remaining effect heterogeneity is presented with following the SCM,

$$A_i := \mathbb{I} \left\{ \frac{\exp(\alpha_0 + \mathbf{L}_i \mathbf{\alpha}_T^T)}{1 - \exp(\alpha_0 + \mathbf{L}_i \mathbf{\alpha}_T^T)} > N_{A_i} \right\} \quad \text{(3)}$$
$$Y_{i}^a := \theta_0 + \mathbf{L}_i \mathbf{\theta}_T^T + a \cdot (\theta_1(M_i) + U_i) + N_{Y_i},$$

where $\mathbb{E}[U_i] = 0$. The conditional distribution of $Y^1 - Y^0 \mid M = m$ is equal to the distribution of $\theta_1(m) + U$ that can still be highly variable and have different shapes, e.g.

$$Y^1 - Y^0 \mid M = m \sim N(\theta_1(m), \sigma^2),$$

or

$$Y^1 - Y^0 \mid M = m \sim LN(\mu, \sigma^2) + (\theta_1(m) - \exp(\mu + 0.5\sigma^2)),$$

or

$$Y^1 - Y^0 \mid M = m \sim \begin{cases} N(\mu_1, \sigma_1^2) & p \\ N(\mu_2, \sigma_2^2) & 1 - p \end{cases} : p \mu_1 + (1 - p) \mu_2 = \theta_1(m).$$

Individual differences within a (sub-) population might be high in variability, skewed or multimodal, examples ofICE distributions for which the CATE (or ATE in case no modifiers are measured) equals $-15$ are presented in Figure 1.
Figure 1: Examples of effect distributions where (C)ATE = −15. ICEs follow a Gaussian distribution with $\sigma = 2$ (1a) and $\sigma = 15$ (1b) respectively, a Log-Normal distribution ($\mu = 4, \sigma = \sqrt{0.5}$, 1c), or a Gaussian mixture distribution with ($p = 0.6, \mu_1 = -31, \sigma_1 = 10, \mu_2 = 9, \sigma_2 = 5$, 1d).

In the ideal case of almost homogeneous (conditional) causal effects, e.g. the distribution shown in Figure 1a, the CATE can be used to make treatment decisions for an individual from the subpopulation of interest. However, when the ICEs are still highly variable, see e.g. Figure 1b and 1c, the CATE doesn’t inform enough and can even be of opposite sign of the effect for some individuals, then the heterogeneity should be taken into account during decision making. Despite accounting for many patient characteristics, there might be other and unknown characteristics that modify the exposure effect such that some individuals in the stratum are expected to be harmed by the exposure while others to benefit from the action, see e.g. Figure 1d. It may be worthwhile to invest in research to discover the underlying effect modifier.

2.2 Variance of the ICE

In case of (remaining) effect heterogeneity, the (conditional) variance of the ICE, $\text{var}(Y^1 - Y^0 \mid M = m)$ depends on the (conditional) joint distribution of the potential outcomes and is thus not identifiable as a result of the fundamental problem of causal inference. However, we will show next that (conditional) effect homogeneity might be ruled out without additional assumptions.
2.2.1 RCT data

Let us first consider data from a well-executed RCT. The latter guarantees that $Y^1, Y^0 \perp \perp A$.

In the case of effect homogeneity within the subpopulation of individuals sharing the level of $M$,

\[
\text{var}(Y^1 \mid M = m) = \text{var}(Y^0 \mid M = m),
\]

since $Y^1_i = Y^0_i + \theta(M_i)$. Then, by causal consistency,

\[
\text{var}(Y \mid A = 1, M = m) = \text{var}(Y \mid A = 0, M = m).
\]

However, when there exists one level $m$ of the effect modifier such that

\[
\text{var}(Y \mid A = 1, M = m) > \text{var}(Y \mid A = 0, M = m),
\]

it becomes clear that there has to be effect heterogeneity as all other characteristics of individuals are exchangeable and can thus not explain the difference in variance of the outcome between exposed and unexposed individuals.

Quantification of the heterogeneity is more challenging, since

\[
\text{var}(Y^1 - Y^0 \mid M = m) = \text{var}(Y^1 \mid M = m) + \text{var}(Y^0 \mid M = m) - 2\text{cov}(Y^1, Y^0 \mid M = m),
\]

while, by the fundamental problem of causal inference, $\text{cov}(Y^1, Y^0 \mid M = m)$ is not identifiable without making additional assumptions. Nevertheless, a lower bound on the variance of the ICE follows from the Cauchy-Schwarz inequality and equals

\[
\text{var}(Y^1 - Y^0) \geq \text{var}(Y^1) + \text{var}(Y^0) - 2\sqrt{\text{var}(Y^1)\text{var}(Y^0)},
\]
where we left out $M = m$ for readability. The lower bound is presented in Figure 2 as a function of the variance of $Y^0$ and the difference of the variances of the potential outcomes. If $Y^1 - Y^0$ is independent of $Y^0$ given $M$, i.e. $Y_i^a = Y_i^0 + U_i$ and $U \perp N_Y | M$, then

$$
\text{var}(Y^1 - Y^0 | M = m) = \text{var}(Y^1 | M = m) - \text{var}(Y^0 | M = m),
$$

and thus by causal consistency,

$$
\text{var}(Y^1 - Y^0 | M = m) = \text{var}(Y | M = m, A = 1) - \text{var}(Y | M = m, A = 0).
$$

Then the (conditional) variance of the ICE can be directly estimated from the RCT data.

Figure 2: Lower bound for the variance of $Y^1 - Y^0$ (z-axis) as a function of the variance of $Y^0$ (x-axis) and the difference between the variance of $Y^1$ and $Y^0$ (y-axis).
However, when \(Y_i^a = Y_i^0 U_i\) and \(U \perp N_y \mid M\), \(Y^1 - Y^0\) does depend on \(Y^0\) given \(M\) and
\[
\text{var}(Y^1 - Y^0 \mid M = m) = \text{var}(Y^1 \mid M = m) + (1 - 2\mathbb{E}[U_1 \mid M = m])\text{var}(Y^0 \mid M = m).
\]
(10)

Such that by causal consistency the conditional variance of the ICE becomes identifiable from the observed data and equals
\[
\text{var}(Y \mid A = 1, M = m) - \left(1 - 2\frac{\mathbb{E}[Y \mid A = 1, M = m]}{\mathbb{E}[Y \mid A = 0, M = m]}\right)\text{var}(Y \mid A = 0, M = m),
\]
(11)
as \(\mathbb{E}[U \mid M = m] = \frac{\mathbb{E}[Y \mid A = 1, M = m] \mathbb{E}[U \mid M = m]}{\mathbb{E}[Y \mid A = 0, M = m]}\).

In summary, based on cross-sectional RCT data alone, we can decide whether effect heterogeneity is present but we cannot identify the variance of \(Y^1 - Y^0 \mid M\) without making the untestable assumption how \(Y^1 - Y^0 \mid M\) depends on \(Y^0 \mid M\).

2.2.2 Observational data

For observational cross-sectional data, even when we make assumptions on the dependence of \(Y^1 - Y^0\) and \(Y^0\), the variance is not identifiable without additional assumptions. Confounders, as they affect both exposure assignment and the outcome of interest, can cause a difference in both the location and the shape of the distribution of the outcome for the exposed and unexposed groups. In observational data, the inequality \([\text{10}]\) can hold in the case of effect homogeneity as a result of confounders. However, realize that
\[
\text{var}(Y^1 - Y^0 \mid M) = \text{var}(Y^1 - Y^0 \mid M, L),
\]
(12)
as by definition all measured modifiers are contained in $M$ such that the confounders $L$ cannot contain additional modifiers and thus $U \perp L \mid M = m$. Identifiability of the variability in ICE requires absence of unmeasured confounding, i.e. $Y^0, Y^1 \perp A \mid M, L$ (equivalent to $N_Y, U \perp N_A$) as is necessary for marginal causal inference (Hernán & Robins 2020). It is well known that the absence of unmeasured confounding cannot be verified using the observed data and it’s appropriateness relies on expert knowledge (Hernán & Robins 2020).

If $Y_i^1 = Y_i^0 + U_i$ and $U \perp N_Y \mid M = m$, by (12) and similar to the RCT case but now adjusted for the confounders,

$$\text{var}(Y^1 - Y^0) = \text{var}(Y^1 \mid M = m, L = \ell) - \text{var}(Y^0 \mid M = m, L = \ell),$$

(13)

for all levels of $l$. And, by causal consistency equals,

$$\text{var}(Y \mid M = m, L = \ell, A = 1) - \text{var}(Y \mid M = m, L = \ell, A = 0),$$

(14)

by causal consistency. So, $\text{var}(Y^a - Y^0 \mid M = m)$ could be computed from a subsample with individuals sharing their levels of the confounders. Similarly, when $Y_i^1 = Y_i^0 U_i$, while $U \perp N_Y \mid M = m$, $\text{var}(Y^1 - Y^0 \mid M = m)$ equals

$$\text{var}(Y \mid M = m, L = \ell, A = 1) - \left(1 - \frac{\mathbb{E}_{Y \mid M = m, L = \ell, A = 1}}{\mathbb{E}_{Y \mid M = m, L = \ell, A = 0}}\right) \text{var}(Y \mid M = m, L = \ell, A = 0),$$

(15)

for all $l$. 

11
2.3 ICE distribution

Understanding the variance is a first step when interested in the ICE distribution. We continue by expressing the ICE distribution in terms of the data generating distribution \( F_Y \). In absence of unmeasured confounding, i.e. \( Y^0, Y^1 \perp \perp A \mid L \), the distribution of \( Y^1 \) or \( Y^0 \) is commonly expressed in terms of \( F_Y \) using the (non-parametric) g-computation formula (Hernán & Robins 2020) which is also known as intervention formula in the field of causal graphical models (Lauritzen & Richardson 2002). Because of the independence, \( Y^a \mid L = \ell \) equals in distribution \( Y^a \mid L = \ell, A = a \) which by causal consistency is equal to the distribution of \( Y \mid L = \ell, A = a \). Thus, the conditional distribution of the potential outcome can be estimated from the observed data.

As \( \{A_i = 1\} \) and \( \{A_i = 0\} \) are mutually exclusive, the g-computation formula reasoning does not satisfy to express the joint distribution of the potential outcome in terms of \( F_Y \). It is only after conditioning on the latent shared cause, \( N_Y \) in SCM (2), that the potential outcomes of an individual become independent (Zhang et al. 2013), i.e. \( Y^1 \perp \perp Y^0 \mid L, M, N_Y \) (as \( Y^0 \mid L, M, N_Y \) is degenerate) while \( Y^1 \not\perp \perp Y^0 \mid L, M \). After conditioning on \( N_Y \), the joint distribution can be factorized in two parts, and in each part we can then condition on either \( \{A_i = 1\} \) or \( \{A_i = 0\} \) respectively. The conditional ICE distribution, \( \mathbb{P}(Y^1 - Y^0 \leq y \mid M) \) equals

\[
\int \int_{y_2 = y_1 - y} \int_{y_1 = -\infty}^{\infty} \int 1dF_{Y \mid A = 1, N_Y = n_Y, L = \ell, M}(y_1)dF_{Y \mid A = 0, N_Y = n_Y, L = \ell, M}(y_2)dF_{N_Y \mid L = \ell, M}(n_Y)dF_L(M(\ell)),
\]

where \( \int g(x)dF_X(x) \) is the Lebesgue-Stieltjes integral of \( g(X) \) with respect to probability law \( F_X \). For the derivation we refer readers to Appendix A. Recall that \( Y^1 - Y^0 \perp \perp L \mid M \), so that \( \mathbb{P}(Y^1 - Y^0 \leq y \mid M) = \mathbb{P}(Y^1 - Y^0 \leq y \mid M, L) \), but adjustment for confounding
is necessary to link the distributions of potential and observed outcomes.

Inference on the ICE distribution, thus requires inference on the conditional expectation, \[ \mathbb{E}[Y \mid A = 0, \mathbf{L}, \mathbf{M}] , \] and the conditional distribution of the unexposed observations, centralized in 0, \( F_{Y \mid \mathbf{L}, \mathbf{M}} \), which are both identifiable under positivity. Moreover, the conditional expectation, \( \mathbb{E}[Y \mid A = 1, \mathbf{L}, \mathbf{M}] \), of the exposed individuals is required and, under positivity, identifiable. Although the distribution of \( F_{Y \mid A = 1, \mathbf{L}, \mathbf{M}} \) is identifiable, the conditional distribution of the unmeasured modifiers, \( F_{U \mid N_Y, \mathbf{L}, \mathbf{M}} \), or equivalently the joint distribution \( F_{(U, N_Y) \mid \mathbf{L}, \mathbf{M}} \), is not because of the fundamental problem of causal inference. Thus, the ICE distribution can only be identified from cross-sectional data after assuming how \( Y^1 - Y^0 \) depends on \( Y^0 \) as was the case for identifiability of the variance in subsection 2.2. A sufficient assumption for identifiability of the ICE distribution (in case of no unmeasured confounding, causal consistency and positivity) is thus conditional independence of the ICE and \( Y^0 \), i.e., \( Y^0 \perp \perp Y^1 - Y^0 \mid \mathbf{M} \).

3 Causal mixed models

We continue by focusing on fitting the unknown \( F_{Y \mid A = 1, \mathbf{L}, \mathbf{M}} \), and \( F_{Y \mid A = 0, \mathbf{L}, \mathbf{M}} \) from data. First, we realize that for \( a \in \{0, 1\} \), without loss of generality,

\[
Y_i \mid \{ A_i = a, L_i = \ell, M_i = m, U_i = u \} = \mathbb{E}_{N_Y} [Y \mid A = a, L = \ell, M = m, U = u] + N_{Y_i},
\]  

(17)

where

\[
\mathbb{E}_{N_Y} [Y \mid A = a, L = \ell, M = m, U = u] = \theta_0(m, \ell) + a \cdot (\theta_1(m) + u).
\]  

(18)

This is a saturated model as the functions \( \theta_0 \) and \( \theta_1 \) are not restricted, \( N_Y \) could depend on \( \mathbf{M}, \mathbf{L} \) as well as \( U \), and the distribution of \( U \) can also depend on \( \mathbf{M} \). In the absence of
unmeasured confounding, $\int \theta_1(m)dF_M(m)$ equals the ATE, $\theta_1(m)$ equals the $m$-CATE, and $\theta_1(M_i) + U_i$ equals the ICE of individual $i$.

As the functional forms of $\theta_0$ and $\theta_1$ as well as the distribution of $(U, N_Y)$ in (17) are unknown, we propose to fit the observed data with a mixed model

$$Y_i = \beta_0(M_i, L_i) + A_i \cdot (\beta_1(M_i) + Z_{1i}) + \epsilon_i, \tag{19}$$

while assuming $Z_{1i} \perp \epsilon_i | M_i$. Note that functions $\theta_0$, $\theta_1$, and the distributions of $U$ and $N_Y$ can be directly derived from SCM (2), while the functions $\beta_0$, $\beta_1$, and the distributions of $Z$ and $\epsilon$ are part of the statistical model that will be fitted to the observed data. If, $N_Y, U \perp N_A$ (no unmeasured confounding), $N_Y \perp U | M$ (assumption on the joint distribution as discussed in Subsection 2.3), and the model (19) is well specified, then the distribution of $\beta_1(m) + Z_1$ equals the distribution of $Y^1 - Y^0 | M = m$.

For example, one could use a LMM with a random exposure effect,

$$Y_i = \beta_0 + M_i \beta_M^T + L_i \beta_L^T + A_i \cdot (M_i \beta_A^T + Z_{1i}) + \epsilon_i, \tag{20}$$

where the random effects $Z_{1i}$ and $\epsilon_i$ are independent and follow some pre-specified distribution. LMMs are often used in epidemiological studies and particularly the fitting of LMMs with (independent) Gaussian random effects ($Z_1$) and Gaussian residual ($\epsilon$) is implemented in standard software packages, e.g. PROC MIXED in SAS and the lmer package in R. It is also possible to let the variance components of $\epsilon$ and $Z_1$ depend on $M$ and $L$.

The Gaussian model might be fitted using likelihood-based methods or Bayesian methods, see e.g. Browne & Draper (2006). Note that when $U | M$ or $N_Y | M, L$ are not Gaussian distributed the Gaussian mixed model is misspecified and the conditional ICE distribution
cannot be estimated with the model. However, even in that case, a Gaussian random effects fit can be useful to study whether there is any remaining variability in ICE within a \( m \)-subpopulation. When the variance component of the random exposure effect is small one could use the CATE as an appropriate proxy for the ICE of individuals in the subpopulation. Otherwise, the TBR and THR in the subpopulation could be estimated from the LMM fit, as proposed by Yin et al. (2018). Again, these TBR and THR estimates will only be accurate when \( U | M \) and \( N_y | M, L \) are (approximately) Gaussian distributed. In the case of (only) misspecified random-effects distributions in generalized linear mixed models (GLMMs), the fixed effect parameters of the marginalized population characteristics are still valid but the standard errors of the estimates are affected (McCulloch & Neuhaus 2011).

When interested in the ICE distribution, the distribution (of the random effects) should be well-specified and thus not restricted to a Gaussian distribution. Verbeke and Lesaffre suggested to model the random effect as a mixture of normals with unknown number of components for LMM’s when Gaussian distributions are inappropriate (Verbeke & Lesaffre 1996). This model can be fitted by means of an Expectation-Maximization (EM) algorithm (Verbeke & Lesaffre 1996), while an alternative estimation has been proposed later by Proust & Jacqmin-Gadda (2005). To study features of among-individual variation, Zhang & Davidian (2001) proposed to fit a LMM by approximating the random effects density by a seminonparametric representation. For both these frequentist approaches an optimal tuning parameter is typically selected based on information criteria (Zhang & Davidian 2001, Proust & Jacqmin-Gadda 2005).

The Gaussian mixtures distribution for the random effect can also be fitted in a Bayesian Framework (Kleinman & Ibrahim 1998). As the ICE distribution is unknown, information
criteria can be used to set the optimal number of components (Gelman et al. 2021). A one-step approach using a uniform prior distribution on the number of components of the mixture distribution of the random effect has been proposed to avoid model selection (Ho & Hu 2008). Model selection is also not necessary when one does fix the number of components to a large constant $K$ while using a symmetric Dirichlet prior, with an appropriate parameter $\alpha$, for the $K$-class probabilities and rely on the natural penalization induced by Bayesian approaches (Gelman et al. 2021). Simulation studies suggest that $\alpha = (K^{-1}, ..., K^{-1})$ will suffice to empty components that are not necessary to fit a Gaussian mixture with unknown number of components (for observed outcomes) (Gelman et al. 2021). Indeed, for univariate random effects, under regularity assumptions, it has been shown that for $\alpha < 1$ the posterior distribution will have empty components when $K$ is larger than the real number of components, while this is not the case for $\alpha > 1$ (Rousseau & Mengersen 2011). Instead of approximating the distribution of the random effect with a mixture of normals the distribution of the random effect can be modeled non parametrically. A Bayesian nonparametric fit of a hierarchical model can be obtained using a Dirichlet process prior (Dunson 2009) or a truncated Dirichlet process prior (Ohlssen et al. 2007) respectively, for the distribution of the random effect. Despite the presence of all these methods, precise estimation of the distribution of a latent variable will remain challenging and is rather sensitive to the proposed model.

3.1 ICE distribution estimation

To demonstrate the use of causal mixed models we will next consider data simulated for 1000 individuals when the cause-effect relations can be parameterized as SCM (3), where $N_Y, U \perp \perp N_A, U \perp \perp N_Y$, and $N_Y$ is Gaussian distributed while $U$ follows a Gaussian
distribution, a Log-normal distribution or a mixture of two Gaussians respectively. For simplicity we will assume that there are no measured modifiers available, and there exists only one binary confounder (that is indeed not a modifier). Details on the parameters used can be found in the caption of Figure 3, and the SAS code used for the simulation can be found online at https://github.com/RAJP93/ICE-distribution.

In this paper we focus on modeling the ICE with a latent variable and not on comparing the methods to fit flexible random effects models that were introduced in the previous section. We will model the ICE distribution with a Gaussian mixture that we fit using the Bayesian method with an upper bound for the number of components, since this one-step approach can be easily implemented in both SAS and R. In line with the proposed model [20], we model the observed data as

\[ Y_i = \beta_0 + L_i \beta_L + A_i Z_{1i} + \epsilon_i, \]  

(21)

where \( Z_{1i} \) follows a Gaussian mixture distribution with mixing probabilities \( p \), component means \( \mu \) and component variances \( \tau^2 \) (i.e. \( Z_{1i} \sim GM(p, \mu, \tau^2) \)), \( \epsilon_i \sim N(0, \sigma^2) \), \( Z_{1i} \perp \perp \epsilon_i \) and \( Z_{1i}, \epsilon_i \perp \perp L_i \). Assuming, for \( 1 \leq j \leq K \), the recommended non-informative priors (Gelman 2006)

\[ \mu_j, \beta_0, \beta_L \sim N(0, 10^5) \]  

(22)

\[ \tau_j, \sigma \sim Uni[0, 100], \]  

(23)

and the weakly informative prior

\[ p \sim Dir(\alpha, \ldots, \alpha), \]  

(24)
are used. When \((21)\) is well-specified, \((16)\) simplifies to \(P(Z_1 \leq y)\). Since \(Z_1\) is an univariate random variable, \(K = 5\) should suffice to capture important characteristics, i.e. skewness and multi-modality, of the ICE distribution as also suggested by Gelman et al. (2021). As the distribution of the ICE is our main object of interest and the parameters of the mixture components are not of interest, we don’t have to worry about the non-identifiability of the mixture parameters and thus not much about the choice of \(\alpha\). As we do not have any prior knowledge about the distribution of the ICE we should be careful to not select a too small \(\alpha\) that would favor distributions closer to a normal distribution via the weakly informative prior. The prior distribution of the number of prominent components, i.e. \(p_j \in p : p_j > 0.10\), for different levels of \(\alpha\) is presented in Appendix \[E\] as Figure \[6\]. Here we use \(\alpha = 0.5\) as we then obtain Jeffreys prior for the labeling distribution (Rousseau & Mengersen 2011).

In this paper we use PROC MCMC in SAS to perform the Bayesian analysis, the script can be found online at \url{https://github.com/RAJP93/ICE-distribution}. In each Monte-Carlo iteration, for each individual it’s latent modifier is sampled from the Gaussian mixture with it’s parameters sampled for that specific iteration. Finally, the posterior \(Z_1\) distribution is presented using a kernel density estimation of all sampled \(Z_1\). The posterior distributions of \(Z_1\) are presented for 100 simulated datasets and for the different underlying ICE distribution in figs. \[3a\] to \[3c\]. In expectation (black lines) the characteristics of the ICE distribution are well recovered.
Figure 3: Estimated $Z_1$ distributions for 100 simulations for underlying ICE distributions $U \sim \mathcal{N}(-15, 10^2)$ (left), or $U \sim \text{LN}(4, \sqrt{0.5^2} - \exp(4 + 0.5 \cdot \sqrt{0.5^2}) - 15$ (middle) and a two component Gaussian mixture such that with probability $p = 0.6$ $U \sim \mathcal{N}(-31, 10^2)$ and otherwise $U \sim \mathcal{N}(9, 5^2)$ (right). The true $U$ density (yellow) and the mean of the estimated $Z_1$ densities (black) are presented. Data for individuals were simulated from SCM (3), where the mean of an unexposed individual equals $\beta_0 = 120$ and the effect of the confounder was $\beta_L = 5$. The confounder $L$ either had value $-0.3$ with probability $0.7$ and value $0.7$ with probability $0.3$ and had thus expectation $0$. The variance of the residuals was equal to $\sigma^2 = 50$ and the probability of exposure equaled $\logit(-3 + 0.7L_i)$ for individual $i$.

For illustration, for an arbitrary dataset, the kernel density estimates of the sampled $Z_1$ per MCMC iteration (grey) are presented together with the posterior distributions of $Z_1$ (colored) and the underlying $Y^1 - Y^0$ distribution (yellow) in Appendix B as Figures figs. 7a to 7c. The mean values of the posterior means and the coverage of the Bayesian credible sets are presented in Table I for the quantiles of the distribution of $Z_1$. For these datasets with 1000 individuals, for all three ICE distributions, the averages of the posterior means are close to the true levels for all characteristics of the ICE distribution considered. Furthermore, the coverage of the Bayesian credible sets are close to 0.95. Only the coverage of the 95% quantile for the heavy-tailed log-normal ICE distribution (0.77) deviates from the nominal level for 1000 individuals.
Table 1: True values of $P(Y^1 - Y^0 > 0)$ and the 5%, 25%, 50%, 75% and 95% quantiles for the different ICE distributions. For these characteristics the average of the posterior means, each estimated on 1000 individuals, and the coverage of the 95% Bayesian credible sets, based on the 100 simulations, are presented.

|                     | Truth   | Average posterior means | Coverage credible sets |
|---------------------|---------|-------------------------|------------------------|
|                     | (a)     | (b)         | (c)     | (a)     | (b) | (c)     | (a)     | (b) | (c)     |
| $P(Y^1 - Y^0 > 0)$  | 0.07    | 0.27         | 0.39    | 0.09    | 0.28 | 0.35    | 0.97    | 0.94 | 0.93    |
| $\alpha_{0.05}$    | -31.44  | -68.04       | -44.83  | -33.48  | -67.29 | -45.61  | 0.99    | 0.94 | 0.95    |
| $\alpha_{0.25}$    | -21.74  | -51.21       | -33.10  | -21.72  | -48.22 | -32.82  | 0.98    | 0.86 | 1.00    |
| $\alpha_{0.5}$     | -15     | -30.51       | -21.33  | -14.88  | -27.30 | -17.46  | 0.99    | 0.94 | 0.90    |
| $\alpha_{0.75}$    | -8.25   | 2.86         | 7.41    | -8.14   | 7.90  | 5.52    | 0.96    | 0.94 | 0.93    |
| $\alpha_{0.95}$    | 1.44    | 89.60        | 14.75   | 3.42    | 82.38 | 16.49   | 0.96    | 0.77 | 0.95    |

In practice, after verifying proper convergence of the markov chains, the model itself should be validated. Apart from the two causal assumptions ($N_Y, U \perp \perp N_A \mid L$ and $U \perp \perp N_Y \mid L$), the model can be validated using the observational data as we will elaborate on in the case study in Section 4.

### 3.2 Confounding-effect heterogeneity

For causal inference from observational data it is necessary to adjust for confounders. If the interest is in the ATE, then typically only the effect of the confounder on the mean of the outcome needs to be adjusted for. However, if one is interested in the distribution of the ICE, then it is necessary to take into account the entire distribution of the effect of the confounder on the outcome of an individual. Otherwise, a difference in (the shape of the) distribution between the exposed and unexposed individuals can be caused by the non-exchangeable confounders. Then, the estimated distribution of $Z_1$ is a mixture of the heterogeneity in the effects of the confounders and in the effect of the exposure. By verifying whether the proposed distribution does fit well in strata defined by the confounders, one can decide whether accounting for the mean alone is sufficient.

The effects of a confounder on the outcome of different individuals might follow a
distribution, then there exists confounding (effect) heterogeneity. An (extreme) example of confounding-effect heterogeneity is the case where only the outcome of a proportion of individuals the is affected by the confounder (Bonvini & Kennedy 2021). In the case of confounding-effect heterogeneity the distribution of $\epsilon$ in (20) depends on the level of the confounders. Therefore, for applications, we suggest to extend (20) to

$$
\begin{align*}
Y_i &= \beta_0 + M_i \beta_M^T + L_i \beta_L^T + A_i \cdot (M_i \beta_{AM}^T + Z_{1i}) + \delta_i, \\
\delta_i &= \epsilon_i + \tilde{L}_i Z_{L,i}^T,
\end{align*}
$$

(25)

where dichotomized versions of the confounders (either $<$ or $>$ median, depending in which group the variance is larger) are presented by $\tilde{L}_i$, $Z_{1i} \sim GM(p, \mu, \tau^2)$, and $\epsilon_i$ and $Z_{L,i}$ are independently Gaussian distributed. This way one corrects for both the shift in mean as well as the increased/decreased variance in the subpopulations with different values for the confounders. Again, the fit of this distribution to the observed data should be verified. When this model is not appropriate one could decide to extend the model further by modeling part of the $Z_L$ with a mixture of Gaussians instead of the Gaussian distribution.

The distribution of $Z_1$ is also not representative for the distribution of the ICE when $\epsilon$ is not approximately Gaussian distributed. Then, the $Z_1$ distribution obtained by fitting model (20) or (25) is a mixture of the exposure effect distribution and the deviation of the distribution of $\epsilon$ from normality, such that the model will better fit the data of the exposed individuals. This problem can be verified by evaluating the fit to the data of the unexposed individuals and might be solved by assuming a Gaussian mixture distribution for $\epsilon$. 

21
4 Case study - Framingham Heart Study

In this section we consider heterogeneity in the effect of non-alcoholic fatty liver disease on cardiac structure and function studied by Chiu et al. (2020) in the Framingham Heart Study (FHS) population. We will work with the sample that consists of a subset of the FHS third generation and offspring cohorts ($n = 2352$). To illustrate our proposed method we focus on the effect of Hepatic Steatosis (fatty liver disease) on the left ventricular filling pressure (LVFP), a clinical precursor to heart failure. Hepatic Steatosis was defined as having a liver phantom ratio (measured fat attenuation in three areas of the liver divided by a phantom calibration control) exceeding 0.33 and was estimated to increase the expected LVFP by 0.46, with a 95% confidence interval equal to (0.26, 0.65), after adjustment for confounding (Chiu et al. 2020).

The sample standard deviation of the LVFP equals 2.27 for individuals exposed to Hepatic Steatosis and only 1.74 for those that were not. This difference might be the result of causal-effect heterogeneity. As the Bayesian analysis discussed in Section 3.1 is computationally intensive we started by fitting a traditional Gaussian LMM to investigate which candidate confounders affected the estimated mean or variance of the effect of Hepatic Steatosis on LVFP. Confounders accounted for in the original study were: age, sex, smoking, alcohol use, diabetes, systolic blood pressure (SBP), use of antihypertensive medication (HRX), use of lipid-lowering medication, total cholesterol, high-density-lipoprotein cholesterol, triglycerides and fasting glucose. The relevant confounders were found to consist of age, sex, diabetes, SBP and HRX (details can be found in Appendix C).

In this case study we focus on the population ICE distribution and do not condition on
any particular modifiers. To do so, we have fitted the model

\[ Y_i = \beta_0 + \mathbf{L}_i \beta_L^T + A_i Z_{1i} + \epsilon_i, \]  

where \( Z_{1i} \sim \text{GM}(\mathbf{p}, \mu, \tau^2) \), with \( K = 5 \), \( \epsilon_i \sim \text{GM}(\tilde{\mathbf{p}}, \tilde{\mu}, \tilde{\tau}^2) \) with \( K = 3 \) while restricting \( \mathbb{E}[\epsilon_i] = 0 \), and \( Z_i \perp \perp \epsilon_i \). We believe that it was appropriate to model the residual (and thus \( Y | A = 0, \mathbf{L} \)) with a more flexible distribution as we do not condition on any features other than adjusting for confounders such that the distribution is expected to be more complex than Gaussian. This is important as a misspecification of the residual distribution, will affect the estimated distribution of \( Z_1 \) as explained at the end of Section 3.2. Alternative flexible error distributions have been presented in literature, see e.g. Ghidey et al. (2010), Shahn & Madigan (2017) or Rubio & Steel (2018).

The model was fitted using the Bayesian method discussed in Section 3.1. The initial values for the parameters in the markov chain where based on the final Gaussian LMM used to select the confounders and can be found in Table 7 in Appendix C. Per chain, we have used 100000 burn-in iterations followed up by 500000 MCMC iterations. We have used a thinning rate of 100 to save computer-memory space. In total 4 chains, each contained 5000 (thinned) MCMC iterations that were saved. To investigate convergence each chain was split in two pieces and we have investigated convergence of the distribution of the \( Z_1 \) for the first unexposed individual in the dataset. Convergence of the markov chains was supported by an estimated R-hat convergence diagnostic, as defined by Vehtari et al. (2021), equal to \( \hat{R} = 1.02 \) and the estimated bulk and tail effective sample sizes equal to \( 8 \cdot 10^3 \) and \( 11 \cdot 10^3 \) respectively. The trace plot for the \( Z_1 \) of this (and another) unexposed individual can be found in Figure 13 in Appendix E. The SAS and R code used for the analysis of the case study can be found online at https://github.com/RAJP93/ICE-distribution.
To draw inference on the ICE distribution we should make sure that the model accurately fits the observed data. To validate the model we study the posterior predictive distribution of the LVFP for both the exposed and unexposed individuals \((Y \mid A = a)\) as presented in Figure 4. The distribution of the outcomes of the unexposed individuals \((Y \mid A = 0)\) are well fitted using the 3-component Gaussian mixture residual. Furthermore, the 5-component Gaussian mixture for the individual effect distribution is flexible enough to fit well the observations of the exposed individuals \((Y \mid A = 1)\).

![Figure 4: Posterior predictive distribution of the LVFP for model with a flexible residual with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for unexposed (left) and exposed individuals (right).](image)

As shown in Table 2, the standard deviation of the LVFP is higher for individuals who are older, female, have diabetes, higher blood pressure or who use antihypertensive medication for both exposure groups. In Section 3.2 we have explained how confounding-effect heterogeneity can affect the distribution of \(Z_1\).

Table 2: Sample standard deviation of the LVFP in sub-samples partitioned by the (dichotomized) confounders and the exposure.

| Hepatic Steatosis | Age>49 | Sex | Diabetes | SBP>120 | HRX |
|-------------------|--------|-----|----------|---------|-----|
| 0                 | 1.4    | 1.5 | 1.7      | 1.5     | 1.6 |
| 1                 | 1.7    | 1.8 | 2.2      | 2.0     | 1.9 |

24
The validation of the fit of the $Y \mid \mathbf{L}$ distribution for all dichotomized levels of the confounders ($>\text{median value}$) is thus a crucial part of the analysis. The posterior predictive distributions in all groups of dichotomized confounders are presented in Appendix F to save some space. Only some small differences between the model and observed distribution can be observed for specific levels of age (Figure 8), sex (Figure 9) and antihypertensive-medicine use (Figure 12). Overall, we conclude that serious (measured) confounding-effect heterogeneity is absent. In Section 4.1 we demonstrate how a model may be fitted when confounding-effect heterogeneity is present.

Thus, model (26) does appropriately model the observed conditional distributions. The ICE distribution is only identifiable in the absence of other confounders and when the dependence of $Y^1 - Y^0$ and $Y^0$ given $\mathbf{L}$ is known. If we assume that $N_Y, U \perp \perp N_A \mid \mathbf{L}$ (absence of unmeasured confounding) and $U \perp \perp N_Y \mid \mathbf{L}$ (i.e., $Y^1 - Y^0 \perp \perp Y^0 \mid \mathbf{L}$), then the distribution of $Z_1$ can be used to estimate the distribution of the ICE as shown in Section 2.3. Under these assumptions, the estimated ICE distribution is shown in Figure 5.

![Figure 5:](image)

**Figure 5:** Posterior distribution of the effect of Hepatic Steatosis on the LVFP using model (26) and pointwise 95% Bayesian credible intervals. Moreover, the estimated distributions when accounting for confounding-effect heterogeneity, using a Gaussian residual and using a Gaussian LMM respectively are presented for comparison.
The ATE has a posterior mean of 0.46 with a 95% Bayesian credible interval (BCI) equal to (0.28, 0.64). This is similar to the effect found in Chiu et al. (2020) and to an ATE estimate of 0.44, with a 95% confidence interval of (0.28, 0.60), obtained using IPTW. The added value of this analysis is inference on other characteristics of the ICE distribution. Despite the on average harming effect of Hepatic Steatosis on the LVFP (ICE>0), 23.7% (BCI 2.6, 53.7) of individuals were estimated not to be harmed. The ATE should thus not be used as an individual effect measure.

4.1 Sensitivity Analysis

To demonstrate how to deal with presence of confounding heterogeneity, we performed a sensitivity analysis by extending model (26) with confounder-specific residual variances. We have fitted the model

\[
\begin{align*}
Y_i &= \beta_0 + L_i \beta_L^T + A_i Z_{1i} + \delta_i, \\
\delta_i &= \epsilon_i + \tilde{L}_i Z_{L_i}^T,
\end{align*}
\]

where again where \( Z_{1i} \sim \text{GM}(\mathbf{p}, \mu, \tau^2) \), with \( K = 5 \), \( \epsilon_i \sim \text{GM}(\mathbf{\tilde{p}}, \mathbf{\tilde{\mu}}, \mathbf{\tilde{\tau}}^2) \) with \( K = 3 \) and \( E[\epsilon_i] = 0 \), and now \( Z_{L_i} \) are Gaussian distributed, while all latent variables are independent. Dichotomized versions of the confounders (> median) are presented by \( \tilde{L}_i \). The posterior predictive checks in the different exposure and confounder strata can be found in Appendix F. The small difference observed in Appendix D disappeared. The ATE has a posterior mean of 0.42 (BCI 0.26, 0.58), and the posterior mean of the TBR equals 19.6% (BCI 4.4, 33.5). These results were (as expected) very similar to the ones obtained by fitting model (26). This holds true for the entire ICE distribution fit as is presented in Figure 5. To emphasize that it is crucial that all latent variables should be modelled with a distribution.
that is flexible enough to accurately fit the observed distributions, we have also presented the estimated ICE distribution when modelling the residual with a Gaussian distribution and when using a Gaussian LMM for the observed outcome in Figure 5.

5 Discussion

Methods for causal inference have rapidly evolve over the past decades. As a result it became more feasible to draw causal claims based on observational studies, typically relying on expert knowledge to back up the untestable assumption of no unmeasured confounding. However, such methods focus on retrieving expectations rather than distributions. The results are therefore only informative at the level of an individual when (remaining) effect heterogeneity is low. In this paper we have focused on moving from (conditional) average treatments effects to quantification of the (conditional) distribution of causal effects.

In the case of effect heterogeneity, exposure (or absence thereof) increases variability in the outcome as individuals respond differently. When we can assume (i) absence of unmeasured confounding and (ii) (conditional) independence of the ICE and the potential outcome under no exposure, inference on the ICE distribution can be drawn from cross-sectional data. To do so, the joint distribution of potential outcomes should be linked to the law of observations which can be learned from the factual data. Then, deviations from the (C)ATE could be quantified. Only when such deviations are small the (C)ATE can be used to accurately approximate the ICE. In case of large deviations, the shape of the (conditional) ICE distribution can inform on the remaining heterogeneity of treatment effect and may illustrate that there is still a serious lack in the understanding of the exposure as the effect of unmeasured modifiers is considerably. The distribution of the unmeasured modifiers can be different across populations so that estimated ICE distributions could be
helpful in understanding of differences in causal effects (Seamans et al., 2021).

In contrast to the flexible distributions of the random effects, we used a ‘simple’ linear mean model in this paper for demonstration. A misspecified mean model will also affect the fit of the random exposure effect. The conditional distributions of the data generating distribution should thus be validated, e.g. by Bayesian posterior predictive checking as illustrated in the case study. In the case that the posterior distribution for unexposed individuals is off, a more flexible mean model (involving the expected effects of modifiers and confounders) may be necessary. For cases where a lot of (rich) data are available it will be promising to investigate how flexible machine learning methods can be used to estimate the conditional means in our mixed model, see e.g. Hajjem et al. (2014) and Pellagatti et al. (2021).

While the fit of $F_{Y|A,M,L}$ can be validated, (conditional) independence of $Y^1 - Y^0$ and $Y^0$ can not be verified with cross-sectional data. As is done for the assumption of no unmeasured confounding, the validity of this assumption should thus be judged on by experts since it can seriously affect the estimated ICE distribution. In future work, we will focus on a sensitivity analysis in which models with different dependence structures between $Z_1$ and $\epsilon$ are fitted. Appropriateness of causal inference depends on the validity of assumptions that can not be tested on the data, e.g. absence of unmeasured confounding, and should resort on expert knowledge. We advocate that assumptions on the (conditional) joint distribution of potential outcomes should become part of this discussion to boost the research on heterogeneity of treatment (or exposure) effects.
Acknowledgements

We sincerely thank Michelle Long and Alison Pedley for sharing the SAS script to reproduce the sub-sample of the Framingham Heart Study as used in Section 4.

References

Bonvini, M. & Kennedy, E. H. (2021), ‘Sensitivity analysis via the proportion of unmeasured confounding’, *Journal of the American Statistical Association* 0(0), 1–11.

**URL:** [https://doi.org/10.1080/01621459.2020.1864382](https://doi.org/10.1080/01621459.2020.1864382)

Browne, W. J. & Draper, D. (2006), ‘A comparison of Bayesian and likelihood-based methods for fitting multilevel models’, *Bayesian Analysis* 1(3), 473 – 514.

**URL:** [https://doi.org/10.1214/06-BA117](https://doi.org/10.1214/06-BA117)

Caron, A., Baio, G. & Manolopoulou, I. (2022), ‘Estimating individual treatment effects using non-parametric regression models: A review’, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 185(3), 1115–1149.

**URL:** [https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12824](https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12824)

Chiu, L. S., Pedley, A., Massaro, J. M., Benjamin, E. J., Mitchell, G. F., McManus, D. D., Aragam, J., Vasan, R. S., Cheng, S. & Long, M. T. (2020), ‘The association of non-alcoholic fatty liver disease and cardiac structure and function—Framingham heart study’, *Liver International* 40(10), 2445–2454.

**URL:** [https://onlinelibrary.wiley.com/doi/abs/10.1111/liv.14600](https://onlinelibrary.wiley.com/doi/abs/10.1111/liv.14600)

Dunson, D. B. (2009), ‘Bayesian nonparametric hierarchical modeling’, *Biometrical Journal* 51(2), 273–284.

**URL:** [https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200800183](https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200800183)

Gelman, A. (2006), ‘Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)’, *Bayesian Analysis* 1(3), 515 – 534.

**URL:** [https://doi.org/10.1214/06-BA117A](https://doi.org/10.1214/06-BA117A)
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2021), *Bayesian data analysis*, third edn, CRC Press.

**URL:** http://www.stat.columbia.edu/~gelman/book/

Ghidey, W., Lesaffre, E. & Verbeke, G. (2010), ‘A comparison of methods for estimating the random effects distribution of a linear mixed model’, *Statistical Methods in Medical Research* 19(6), 575–600.

**URL:** https://doi.org/10.1177/0962280208091686

Greenland, S., Fay, M. P., Brittain, E. H., Shih, J. H., Follmann, D. A., Gabriel, E. E. & Robins, J. M. (2020), ‘On causal inferences for personalized medicine: How hidden causal assumptions led to erroneous causal claims about the d-value’, *The American Statistician* 74(3), 243–248.

**URL:** https://doi.org/10.1080/00031305.2019.1575771

Hajjem, A., Bellavance, F. & Larocque, D. (2014), ‘Mixed-effects random forest for clustered data’, *Journal of Statistical Computation and Simulation* 84(6), 1313–1328.

**URL:** https://doi.org/10.1080/00949655.2012.741599

Hand, D. J. (1992), ‘On comparing two treatments’, *The American Statistician* 46(3), 190–192.

**URL:** https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475881

Hernán, M. A. & Robins, J. M. (2020), *Causal Inference: What If.*, Boca Raton: Chapman & Hall/CRC.

**URL:** https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Ho, R. K. & Hu, I. (2008), ‘Flexible modelling of random effects in linear mixed models—a bayesian approach’, *Computational Statistics & Data Analysis* 52(3), 1347–1361.

**URL:** https://www.sciencedirect.com/science/article/pii/S016794730700343X

Holland, P. W. (1986), ‘Statistics and causal inference’, *Journal of the American Statistical Association* 81(396), 945–960.

**URL:** http://www.jstor.org/stable/2289064

Kleinman, K. P. & Ibrahim, J. G. (1998), ‘A semiparametric bayesian approach to the random effects model’, *Biometrics* 54(3), 921–938.

**URL:** http://www.jstor.org/stable/2533846
Lauritzen, S. L. & Richardson, T. S. (2002), ‘Chain graph models and their causal interpretations’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3), 321–348.

**URL:** https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00340

Lei, L. & Candès, E. J. (2021), ‘Conformal inference of counterfactuals and individual treatment effects’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 83(5), 911–938.

**URL:** https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12445

Li, A. & Pearl, J. (2019), Unit selection based on counterfactual logic, in ‘Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19’, International Joint Conferences on Artificial Intelligence Organization, pp. 1793–1799.

**URL:** https://doi.org/10.24963/ijcai.2019/248

Lu, M., Sadiq, S., Feaster, D. J. & Ishwaran, H. (2018), ‘Estimating individual treatment effect in observational data using random forest methods’, *Journal of Computational and Graphical Statistics* 27(1), 209–219. PMID: 29706752.

**URL:** https://doi.org/10.1080/10618600.2017.1356325

McCulloch, C. E. & Neuhaus, J. M. (2011), ‘Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter’, *Statistical Science* 26(3), 388 – 402.

**URL:** https://doi.org/10.1214/11-STS361

Mueller, S., Li, A. & Pearl, J. (2022), Causes of effects: Learning individual responses from population data, in ‘Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22’, International Joint Conferences on Artificial Intelligence Organization, pp. 2712–2718.

**URL:** https://doi.org/10.24963/ijcai.2022/376

Ohlsson, D. I., Sharples, L. D. & Spiegelhalter, D. J. (2007), ‘Flexible random-effects models using bayesian semi-parametric models: applications to institutional comparisons’, *Statistics in Medicine* 26(9), 2088–2112.

**URL:** https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2666

Pellagatti, M., Masci, C., Ieva, F. & Paganoni, A. M. (2021), ‘Generalized mixed-effects random forest: A flexible approach to predict university student dropout’, *Statistical Analysis and Data Mining: The
Peters, J., Janzing, D. & Schölkopf, B. (2018), *Elements of causal inference: foundations and learning algorithms*, The MIT Press, Cambridge, Massachusetts.

URL: https://mitpress.mit.edu/9780262037310/

Proust, C. & Jacqmin-Gadda, H. (2005), ‘Estimation of linear mixed models with a mixture of distribution for the random effects’, *Computer Methods and Programs in Biomedicine* 78(2), 165–173.

URL: https://www.sciencedirect.com/science/article/pii/S0169260705000209

Robins, J. M., Hernán, M. A. & Brumback, B. (2000), ‘Marginal structural models and causal inference in epidemiology’, *Epidemiology* 11(5), 550–560.

URL: https://doi.org/10.1097/00001648-200009000-00011

Rousseau, J. & Mengersen, K. (2011), ‘Asymptotic behaviour of the posterior distribution in overfitted mixture models’, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 73(5), 689–710.

URL: http://www.jstor.org/stable/41262270

Rubin, D. B. (1974), ‘Estimating causal effects of treatments in randomized and nonrandomized studies’, *Journal of Educational Psychology* 66(5), 688–701.

URL: https://psycnet.apa.org/doi/10.1037/h0037350

Rubio, F. J. & Steel, M. F. J. (2018), ‘Flexible linear mixed models with improper priors for longitudinal and survival data’, *Electronic Journal of Statistics* 12(1), 572 – 598.

URL: https://doi.org/10.1214/18-EJS1401

Seamans, M. J., Hong, H., Ackerman, B., Schmid, I. & Stuart, E. A. (2021), ‘Generalizability of subgroup effects’, *Epidemiology* 32(3), 389–392.

URL: https://doi.org/10.1097/EDE.0000000000001329

Shahn, Z. & Madigan, D. (2017), ‘Latent Class Mixture Models of Treatment Effect Heterogeneity’, *Bayesian Analysis* 12(3), 831 – 854.

URL: https://doi.org/10.1214/16-BA1022
Splawa-Neyman, J., Dabrowska, D. M. & Speed, T. P. (1990), ‘On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9’, Statistical Science 5(4), 465 – 472.

URL: https://doi.org/10.1214/ss/1177012031

Tian, J. & Pearl, J. (2000), ‘Probabilities of causation: Bounds and identification’, Annals of Mathematics and Artificial Intelligence 28(1), 287–313.

URL: https://doi.org/10.1023/A:1018912507879

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.-C. (2021), ‘Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC (with Discussion)’, Bayesian Analysis 16(2), 667 – 718.

URL: https://doi.org/10.1214/20-BA1221

Verbeke, G. & Lesaffre, E. (1996), ‘A linear mixed-effects model with heterogeneity in the random-effects population’, Journal of the American Statistical Association 91(433), 217–221.

URL: http://www.jstor.org/stable/2291398

Yin, Y., Liu, L. & Geng, Z. (2018), ‘Assessing the treatment effect heterogeneity with a latent variable’, Statistica Sinica 28(1), 115–135.

URL: http://www.jstor.org/stable/26384235

Zhang, D. & Davidian, M. (2001), ‘Linear mixed models with flexible distributions of random effects for longitudinal data’, Biometrics 57(3), 795–802.

URL: http://www.jstor.org/stable/3068418

Zhang, Z., Wang, C., Nie, L. & Soon, G. (2013), ‘Assessing the heterogeneity of treatment effects via potential outcomes of individual patients’, Journal of the Royal Statistical Society: Series C (Applied Statistics) 62(5), 687–704.

URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12012
A Derivation ICE distribution

By the law of total probability,

\[ \mathbb{P}(Y^1 - Y^0 \leq y \mid M) = \int \mathbb{P}(Y^1 - Y^0 \leq y \mid L = \ell, M) dF_L\mid_M(l). \quad (28) \]

Again, by the law of total probability,

\[ = \int \int_{y_1 = -\infty}^{\infty} \int_{y_2 = y_1 - y}^{\infty} 1dF_{Y^1,Y^0\mid L=\ell, M}(y_1, y_2) dF_L\mid_M(l) \]
\[ = \int \int_{y_1 = -\infty}^{\infty} \int_{y_2 = y_1 - y}^{\infty} \int 1dF_{Y^1,Y^0\mid N_Y=n_Y, L=\ell, M}(y_1, y_2) dF_{N_Y\mid L=\ell, M}(n_Y) dF_L\mid_M(l). \quad (29) \]

As \( Y^0 \mid N_Y = n_Y, L = \ell, M = m \) is a degenerate random variable, \( Y^1 \perp Y^0 \mid L = \ell, M = m, N_Y = n_Y \), and thus

\[ = \int \int_{y_1 = -\infty}^{\infty} \int_{y_2 = y_1 - y}^{\infty} 1dF_{Y^1\mid N_Y=n_Y, L=\ell, M}(y_1) dF_{Y^0\mid N_Y=n_Y, L=\ell, M}(y_2) dF_{N_Y\mid L=\ell, M}(n_Y) dF_L\mid_M(l). \quad (30) \]

Since \( Y^1, Y^0 \perp\!\!\!\!\!\!\perp A \mid M, L \),

\[ = \int \int_{y_1 = -\infty}^{\infty} \int_{y_2 = y_1 - y}^{\infty} 1dF_{Y^1\mid A=1, N_Y=n_Y, L=\ell, M}(y_1) dF_{Y^0\mid A=0, N_Y=n_Y, L=\ell, M}(y_2) dF_{N_Y\mid L=\ell, M}(n_Y) dF_L\mid_M(l). \quad (31) \]

Finally, by causal consistency,

\[ = \int \int_{y_1 = -\infty}^{\infty} \int_{y_2 = y_1 - y}^{\infty} 1dF_{Y\mid A=1, N_Y=n_Y, L=\ell, M}(y_1) dF_{Y\mid A=0, N_Y=n_Y, L=\ell, M}(y_2) dF_{L\mid M}(l). \quad (32) \]
B Supplementary figures Bayesian analysis

Figure 6: The distribution of the number of \( p_j \in p : p_j > 0.10 \) when \( p \sim \text{Dir}(\alpha, \alpha, \alpha, \alpha, \alpha) \) for \( \alpha = \frac{1}{5} \) (left), \( \alpha = \frac{1}{2} \) (middle) and \( \alpha = 1 \) (right).

Figure 7: Kernel density estimates of the \( Z_1 \) sampled per MCMC iteration (grey) for one of the datasets for which the posterior distributions were presented in Figure 3. The corresponding posterior distributions (green/orange/blue) and the true ICE distribution (yellow) are also presented.

C Confounder selection case study

As the Bayesian analysis was computationally intensive we started by fitting a Gaussian linear mixed model (LMM) to investigate which candidate confounders did change the mean or variance of the effect of Hepatic Steatosis on the left ventricular filling pressure (LVFP). More precisely we started by fitting
\[ Y_i = \beta_0 + L_i \beta L^T + Z_{1i} A_i + \epsilon_i, \]  
(33)

where \( Z_{1i} \) and \( \epsilon_i \) are independent and Gaussian distributed and \( \mathbb{E}[\epsilon_i] = 0 \). Initially we include all candidate confounders: age, sex, smoke, drinks per week (DPW), diabetes state (DIAB), systolic blood pressure (SBP), antihypertensive-medication use (HRX), lipid lowering med use (LRX), total cholesterol (CHOL), high-density-lipoprotein cholesterol (HDL), triglyceride (TRIG) and fasting glucose (GLU).

Then, we started to remove single components of \( L \) from the model to see how \( \mathbb{E}[Z_1] \) was affected. If, the smallest change was less than 5% of \( \mathbb{E}[Z_1] \) we remove that component of \( L \) from the model. This selection procedure was continued until removing a candidate confounder from the model would result in a relative change of \( \mathbb{E}[Z_1] \) greater than 5%. The entire procedure is shown in Table 3.

**Table 3:** Change in estimate of exposure effect (\( \mathbb{E}[Z_1] \)) after removing candidate confounders from the Gaussian LMM. In each step of the procedure the covariate that had least impact is removed.

| Remove fixed effect | age  | sex  | smoke | DPW  | DIAB | SBP  | HRX  | LRX  | CHOL | HDL  | TRIG | GLU  |
|---------------------|------|------|-------|------|------|------|------|------|------|------|------|------|
| \( \mathbb{E}[Z_1] \) | 0.456| 0.398| 0.495 | 0.454| 0.449| 0.465| 0.520| 0.486| 0.452| 0.459| 0.456| 0.399|
| Relative change     | -0.126| 0.086| -0.005| -0.014| 0.021| 0.142| 0.066| -0.008| 0.006| 0.001| 0.042| 0.007|
| \( \mathbb{E}[Z_1] \) | 0.456| 0.390| 0.478 | 0.454| 0.455| 0.465| 0.522| 0.487| 0.452| 0.459| 0.456| 0.399|
| Relative change     | -0.144| 0.049| -0.004| -0.007| 0.020| 0.145| 0.067| -0.008| 0.010| 0.064| 0.007| 0.007|

At this stage age, DIAB, SBP and HRX affect \( \mathbb{E}[Z_1] \) and are therefore considered confounders. The estimates of this model can be found in Table 4.
Table 4: Gaussian LMM parameter estimates ignoring confounder effect heterogeneity.

| Fixed effects   | Variance components |
|-----------------|---------------------|
| Intercept       | 6.227               |
| Hepatic Steatosis | 0.470          |
| age             | 0.476               |
| DIAB            | 0.395               |
| SBP             | 0.260               |
| HRX             | 0.383               |
| Residual        | 2.535               |

However, candidate confounders should also be recognized as confounders when they do not seriously affect the mean $\mathbb{E}[Z_1]$ but instead affect the variance of $Z_1$. First we add a random effect for dichotomized confounders selected at this stage and fit

$$
Y_i = \beta_0 + L_i \beta_L^T + A_i Z_{1i} + \delta_i,
\delta_i = \epsilon_i + \tilde{L}_i Z_L^T
$$

(34)

where $Z_{1i}$, $\epsilon_i$ and $Z_L$ are independent and Gaussian distributed and $\mathbb{E}[\epsilon_i] = 0$. Dichotomized confounders, $\tilde{L}_i$ are obtained by comparing the level of the continuous confounder with the sample median; $>$ when the sample variance is larger in individuals with higher values of the outcome and $<$ otherwise. The estimates can be found in Table 5 and it becomes clear that the variance of $Z_1$ does seriously change after accounting for possible confounding-effect heterogeneity.

Table 5: Gaussian LMM parameter estimates for model after first selection procedure accounting for confounder effect heterogeneity.

| Fixed effects       | Variance components |
|---------------------|---------------------|
| Intercept           | 6.237               |
| Hepatic Steatosis   | 0.411               |
| age                 | 0.466               |
| DIAB                | 0.489               |
| SBP                 | 0.269               |
| HRX                 | 0.399               |
| Residual            | 1.595               |
| Hepatic Steatosis   | 0.809               |
| (age $>$ 49)        | 0.965               |
| DIAB                | 0.850               |
| (SBP $>$ 120)       | 0.476               |
| HRX                 | 1.512               |
| Residual            | 1.595               | 37
Subsequently, we did again add one of the other candidate confounders to the model and investigated how the variance of $Z_1$ changed. If for one of the covariates the relative change of the variance was more than 10%, then we added the most influential candidate confounder to the model. After adding the covariate sex to the model no other candidate confounders gave rise to a change in variance greater that 10% as shown in Table 6.

**Table 6:** Change in estimate of variance of the random effect of fatty liver (variance of $Z_1$) after adding candidate confounders to the LMM. In each step of the procedure the covariate that has most impact is added.

| Add random effect | - | sex | smoke (DPW $<$ 1.5) | 1-LRX (CHOL $>$ 189) | (HDL $>$ 52) | TRIG $>$ 98 | GLU $<$ 96 |
|-------------------|---|-----|---------------------|----------------------|-------------|--------------|------------|
| Variance $Z_1$    | 0.809 | 0.623 | 0.843 | 0.841 | 0.803 | 0.808 | 0.810 | 0.866 |
| Relative change   | -0.229 | -0.103 | 0.042 | 0.040 | -0.007 | -0.001 | 0.002 | 0.071 |
| Variance $Z_1$    | 0.623 | 0.581 | 0.630 | 0.636 | 0.623 | 0.607 | 0.582 | 0.605 |
| Relative change   | -0.067 | 0.043 | 0.053 | -0.001 | -0.026 | -0.066 | -0.029 |

With this procedure we have selected age, sex, DIAB, SBP and HRX as confounders. The parameter estimates of the final Gaussian LMM are presented in Table 7 and were used as initial values in the Bayesian procedure of the case study. The distribution of $Z_1$ was shown in Figure 5 (Gaussian LMM).

**Table 7:** Final LMM parameter estimates accounting for age, sex, diabetes, systolic blood pressure and antihypertensive-medication use.

| Fixed effects | 5.877 |
|----------------|-------|
| Hepatic Steatosis | 0.416 |
| age | 0.363 |
| sex | 0.667 |
| DIAB | 0.561 |
| SBP | 0.332 |
| HRX | 0.432 |

| Variance components | Hepatic Steatosis (age $>$ 49) | 0.623 |
|---------------------|---------------------------------|-------|
| (HDL $>$ 52) | 0.710 |
| TRIG $>$ 98 | 0.923 |
| sex | 0.747 |
| (SBP $>$ 120) | 0.723 |
| Residual | 1.374 |

Residual | 1.064 |
D  Posterior predictive distribution for confounder strata

Figure 8: Posterior predictive distribution of the LVFP for model (26), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for individuals that are older than 49 years (right) or not (left).

Figure 9: Posterior predictive distribution of the LVFP for model (26), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for males (left) and females (right).
Figure 10: Posterior predictive distribution of the LVFP for model (26), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for individuals with (right) and without (left) diabetes.

Figure 11: Posterior predictive distribution of the LVFP for model (26), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for individuals with a systolic blood pressure above 120 mmHg (right) or not (left).
Figure 12: Posterior predictive distribution of the LVFP for model (26), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for individuals with (right) and without (left) antihypertensive-medication use.

E Trace plots

Figure 13: Trace plots of simulated $Z_{1i}$ from the four thinned markov chains for two unexposed individuals while fitting model (26). The burn-in iterations are not shown.
Posterior predictive distributions for model accounting for confounding heterogeneity

Figure 14: Posterior predictive distribution of the LVFP for model (27), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for unexposed (left) and exposed individuals (right).

Figure 15: Posterior predictive distribution of the LVFP for model (27), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for individuals that are older than 49 years (right) or not (left).
Figure 16: Posterior predictive distribution of the LVFP for model (27), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for males (left) and females (right).

Figure 17: Posterior predictive distribution of the LVFP for model (27), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for individuals with (right) and without (left) diabetes.
Figure 18: Posterior predictive distribution of the LVFP for model (27), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for individuals with a systolic blood pressure above 120 mmHg (right) or not (left).

Figure 19: Posterior predictive distribution of the LVFP for model with (27), with corresponding 95% Bayesian credible intervals and the kernel density of the observed data for individuals with (right) and without (left) antihypertensive-medication use.