

Mediation analysis in longitudinal intervention studies with an ordinal treatment-dependent confounder

Statistical Methods in Medical Research

1–22

© The Author(s) 2026



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09622802261418211

journals.sagepub.com/home/smm

Mikko Valtanen^{1,2} , Tommi Härkänen² , Matti Uusitupa³ ,
Jaakko Tuomilehto^{2,4} , Jaana Lindström^{2,3}  and Kari Auranen^{1,5}

Abstract

In interventional health studies, causal mediation analysis can be employed to investigate mechanisms through which the intervention affects the targeted health outcome. Identifying direct and indirect effects from empirical data become complicated, however, when a confounder of the mediator-outcome association is itself affected by the treatment. Here, we investigate identification of mediational effects under such post-treatment confounding in a setting with a longitudinal mediator, time-to-event outcome and an ordinal treatment-dependent confounder. If the treatment affects the treatment-dependent confounder only in one direction (monotonicity), we show that the mediational effects are identified up to stratum-specific sensitivity parameters and derive their empirical non-parametric expressions. The feasibility of the monotonicity assumption can be assessed using empirical data, based on restrictions on the marginal distributions of counterfactuals of the treatment-dependent confounder. In an empirical analysis, we use data from the Finnish Diabetes Prevention Study to assess the extent to which the effect of a lifestyle intervention on avoiding type 2 diabetes is mediated through weight reduction in a high-risk population, with other health-related changes acting as treatment-dependent confounders. We avoid pitfalls related to post-treatment conditioning by treating the mediator as a functional entity and defining the time-to-event outcome as a restricted disease-free time.

Keywords

intervention studies, longitudinal mediator, mediation analysis, treatment-dependent confounder, type 2 diabetes

1 Introduction

Lifestyle choices play an important role in the prevention of type 2 diabetes (T2D). As T2D causes major health and economic burdens globally, lifestyle interventions to reduce its incidence in high-risk populations are an area of active research.^{1–4} Studies of such interventions typically follow a cohort of individuals over time and aim at collecting information on biomarkers and health status evaluations at several follow-up visits.^{5–7} In countries with extensive health registers, data gathered at follow-up visits can be augmented with time-to-event outcomes retrieved from the registers.

Intervention studies are often based on experimental designs where study participants are assigned to treatment and control groups and the target of inference is the causal effect of an intervention on a specific health response. A more elaborate question involves understanding the extent to which the treatment effect is mediated through an intermediate

¹Department of Mathematics and Statistics, University of Turku, Finland

²Department of Public Health, Finnish Institute for Health and Welfare, Helsinki, Finland

³Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland

⁴Department of Public Health, University of Helsinki, Helsinki, Finland

⁵Department of Clinical Medicine, University of Turku, Finland

Corresponding author:

Mikko Valtanen, Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland.

Email: mikko.valtanen@thl.fi

variable. For example, potential mediating mechanisms of liraglutide treatment on cardiovascular and chronic kidney disease through changes in clinical biomarkers have been previously investigated in people with diabetes.^{8,9}

Mediating mechanisms can be addressed within the causal mediation analysis framework, where the total effect of treatment is decomposed into direct and indirect (i.e. mediated) effects.^{10,11} The estimands of these causal effects can be defined in terms of counterfactuals, that is expected outcomes in hypothetical scenarios where the treatment and the mediator are intervened upon to set them at certain values.¹² Under non-trivial conditional independence assumptions, the estimands can be non-parametrically identified from empirical data.¹³ Studies using a lifestyle intervention as the treatment and a clinical risk factor as the mediator, however, are especially prone to violations of the identifiability assumptions, because the intervention often induces behavioural changes that affect the response through both the intended mediator and other mechanisms. The behavioural change then acts as a treatment-dependent confounder, that is a variable that confounds the mediator–outcome relationship while also lying on a causal path between the treatment and outcome. In such situations the standard independence assumptions do not suffice to identify natural mediational effects non-parametrically.¹⁴

Tchetgen Tchetgen and Vanderweele¹⁵ showed that identification can be retained even under treatment-dependent confounding by further assumptions such as monotonicity of the treatment effect on a binary treatment-dependent confounder. Here, monotonicity means that the treatment can have only a positive (or only a negative) effect on the treatment-dependent confounder. Other assumptions retaining point identification include independence of counterfactuals of the treatment-dependent confounder or absence of an additive interaction between the mediator and the treatment-dependent confounder.^{15,16} Without any additional assumptions, the mediational effects under an observed discrete treatment-dependent confounder are still partially identifiable. This means that lower and upper bounds for the estimates can be obtained.¹⁷

Further methodological issues arise when study participants are followed over time and the mediator is a longitudinal process. While causal mediation analysis has been extended to settings with longitudinal mediators,^{18–20} most literature treats the mediator as a vector-valued entity containing successive measurements of the mediating variable. Such an approach results in high dimensionality if the number of repeated measurements per individual is large and also poses challenges in handling uneven measurement intervals and missing values. Some authors have instead applied functional regression to represent the mediator trajectory as a function.^{21–23} Treating the mediator as a functional entity avoids the problem of dimensionality and enables flexible use of its full history at the time the response is evaluated. The repeated measurements in the longitudinal setting can also be used to extract information about individual-level latent properties. For example, Zheng and Liu²⁴ relaxed the assumption of no unmeasured mediator–outcome confounding under a longitudinal mediator and a time-to-event outcome by employing a joint modelling framework to estimate and control a common random effect reflecting an unobserved confounder between the two.

Time-to-event outcomes pose additional challenges in the causal inference framework.²⁵ A particular issue arises when causal estimands are defined by measures that condition on prior survival, such as the hazard function. If there exist latent variables affecting survival, conditioning opens a backdoor path from the treatment to future survival through the latent variables. This issue can be addressed by defining the response as an unconditional measure, such as the restricted mean survival time (RMST), that is the mean event-free time within a preset time period.^{23,26} If a significant portion of individuals do not experience the event during the study, RMST has the additional benefit over the mean survival time of allowing the time period to be chosen so that it remains robust to misspecification of the unobserved tail of the event-time distribution. In addition, if it is considered possible to have a zero risk for the event, the true mean survival time would be infinite, whereas RMST would remain constrained to the chosen clinically relevant time period.

In this study, we address identification of mediational causal effects in interventional studies with treatment-dependent confounding. We extend the previously presented monotonicity assumption¹⁵ to a trichotomous treatment-dependent confounder and show that this results in expressions identifiable up to a stratum-specific sensitivity parameter. Our approach is similar to the partial identification in Miles et al.¹⁷ but imposes restrictions on the unobserved joint distribution of the counterfactuals of the treatment-dependent confounder, leading to a necessary condition for their marginals. As the marginals can be estimated from observed data, the feasibility of the monotonicity assumption can be empirically assessed. As an application, we consider the effect of an intensive lifestyle intervention on T2D incidence among a high-risk population, based on the Finnish Diabetes Prevention Study (DPS).²⁷ The aim of the empirical analysis is to quantify the extent to which the effect of the intervention on T2D-free time is mediated through it inducing weight loss. We apply functional regression to represent the body mass index (BMI) trajectory as a functional entity and use a joint modelling framework to control potential latent confounding between the BMI trajectory and T2D incidence.

The paper is structured as follows. Section 2 presents the empirical problem motivating this study. Section 3 defines the targeted causal estimands, specifies conditions for their identification, and describes the methods we propose to use for their estimation. Sections 4 and 5 present the results of the empirical application and discuss the results along with further considerations of the used methodology and its possible limitations.

2 Data sources and motivation

The aim of the Finnish DPS is to assess the effectiveness of an intensive lifestyle intervention in preventing and delaying T2D onset in a high-risk population.^{27–29} The study cohort was enrolled between 1993 and 1998 and originally consisted of 522 individuals. The eligibility criteria required the participants to be overweight (BMI > 25), aged 40–64 years, and have impaired glucose tolerance at the screening visit. The study participants were randomly allocated to intervention and control groups. The active intervention lasted a maximum of six years (median 4 years), involving frequent personalised nutritional counselling and encouragement for physical activity, primarily through face-to-face sessions. The active intervention ended in 2001. The control group was given routine, non-personalised healthy lifestyle advice during the study visits. Post-intervention follow-up visits continued until 2013, with a median of ten clinical study visits per person. The participants were tested for T2D in the clinical study visits using the World Health Organization's 1985 criteria for a 2-hour oral glucose tolerance test (OGTT),³⁰ with a diagnosis requiring two OGTTs above the threshold. In addition, the Finnish Registers for Drug Reimbursements and Drug Purchases were searched for T2D-related drug purchases, extending the time-to-event follow up until the end of 2018.

A previous analysis of the DPS data showed a 40% lower hazard for T2D and greater weight loss in the intervention group during the first 13 years of follow-up and also showed greater improvements in their lifestyle compared with the control group, particularly in dietary intakes.³¹ Moreover, previous analyses have found associations between the lifestyle intervention, physical activity, nutritional components and diabetes incidence in the DPS cohort.^{32,33}

In this study, we aim to quantify the effect of the lifestyle intervention on the prevention of T2D mediated through weight loss. Since obesity is one of the most prominent risk factors for T2D, it is of interest to assess the extent to which the success of the lifestyle intervention can be attributed to its ability to reduce body weight in people with overweight or obesity. Previous literature on lifestyle intervention studies has suggested a so-called legacy effect, wherein the intervention's impact on T2D incidence persist long after the intervention ends and the obtained group differences in risk factors have diminished. As summarised by Wilding,³⁴ such results have been reported in the major T2D prevention trials, including the DPS,³¹ a study conducted in China,³⁵ and the Diabetes Prevention Program in the USA.³⁶ These findings motivated us to consider the change in the BMI during the early phase of the intervention as the effective mediator. Moreover, we apply a three-year time window because the majority of differences in BMI between the two groups occurred during this period. We will use the restricted survival time as the outcome measure, with the maximum time as 15 years, reflecting a clinically relevant time horizon. The outcome is thus interpreted as the number of healthy (i.e. T2D-free) years during the first 15 years after intervention onset.

In addition to weight reduction, the DPS intervention aimed at moderate physical activity and healthy nutritional composition measured by intakes of total fats, saturated fats and fibre.²⁸ These lifestyle factors can be assumed to influence the study participants' BMI trajectories and also T2D incidence through mechanisms other than weight loss, thus rendering them potential treatment-dependent confounders. We created a summary variable to represent individuals' lifestyle choices influenced by the intervention, combining total physical activity and the dietary intake components. The amount of total physical activity was measured by self-reports at every study visit and the components of dietary intake by three-day food diaries prior to the study visits for the first three (in addition to baseline). All variables from each study visit were standardised with respect to their baseline means and standard deviations and the lifestyle score was computed as the mean over the standardised variables across the three post-baseline study visits. The lifestyle score was then categorised into three levels based on its baseline distribution. As the distribution was nearly Gaussian, the tertiles were used as cut-points for this categorisation. Constructing the lifestyle score in this way ensures its ordinal interpretation in the sense that belonging to a higher category implies on average higher levels in each of the four lifestyle variables.

Causal mediation analysis requires controlling for any factors confounding the relationships between the treatment, mediator and outcome. We considered age, sex, smoking status and the baseline lifestyle score as potential confounding baseline variables. Age at baseline was categorised as <45, 45 to 59 and ≥60 years, while the smoking status was categorised as 'never', 'former' or 'current'.

3 Methods

In this section we present the proposed methodological framework. Sections 3.1 and 3.2 outline the assumed causal model and the estimation targets. In Section 3.3 we give assumptions sufficient to identify the causal estimands from empirical data and present the resulting expressions for the direct and indirect effects. In Sections 3.4 and 3.5 we define the parametric models and describe the strategy for their estimation.

3.1 Causal estimands

Let $A \in \{a^*, a\}$ denote the treatment group (a for intervention, a^* for control), \tilde{T} time since baseline (study onset) to T2D diagnosis, $T = \min\{\tilde{T}, t_{\max}\}$ the restricted time without a T2D diagnosis for a prespecified t_{\max} (=15 years), and $M(\cdot)$ a function of time that describes the true trajectory of BMI as a continuous mediator if remaining alive and T2D-free. We use subscripts to denote quantities under potential, possibly counterfactual scenarios: T_a refers to the restricted survival time when the treatment is set to a , whereas $T_{a, M_{a^*}(\cdot)}$ is the corresponding time when the treatment is set to a , but the mediator follows the trajectory it would take under intervention a^* .

We use the RMST, $\tau^{t_{\max}} = \mathbb{E}(T) = \mathbb{E}(\int_0^{t_{\max}} P(\tilde{T} > v) dv)$, as the response measure,²⁶ and define the average natural direct and indirect effects of the intervention by contrasting treatment a against a^* as

$$\begin{aligned} \text{DE} &= \mathbb{E} \left[T_{a, M_{a^*}(\cdot)} - T_{a^*, M_{a^*}(\cdot)} \right] = \tau_{a, M_{a^*}(\cdot)}^{t_{\max}} - \tau_{a^*, M_{a^*}(\cdot)}^{t_{\max}}, \\ \text{IE} &= \mathbb{E} \left[T_{a, M_a(\cdot)} - T_{a, M_{a^*}(\cdot)} \right] = \tau_{a, M_a(\cdot)}^{t_{\max}} - \tau_{a, M_{a^*}(\cdot)}^{t_{\max}} \end{aligned} \quad (1)$$

The DE describes the change in RMST if the treatment was set to a instead of a^* but the mediator follows the trajectory it would have under treatment a^* . The IE is the change in RMST if the mediator follows the trajectory it would have under treatment a instead of treatment a^* while the treatment is fixed to a (see e.g. VanderWeele³⁷). The DE and IE sum up to the average total effect, $\text{TE} = \tau_{a, M_a(\cdot)}^{t_{\max}} - \tau_{a^*, M_{a^*}(\cdot)}^{t_{\max}}$. Of note, the RMST corresponds to the area under the survival function between times zero and t_{\max} , implying that a simple nonparametric estimator of the TE can be constructed using the areas under the Kaplan–Meier curves for the two treatment groups.

If there is interaction between the treatment and the mediator, the interpretation of mediational effects depends on the choice of the levels at which the fixed variables are held in each case. The definitions (1) lead to interpreting the DE as the *pure direct effect* and the IE as the *total indirect effect*, as discussed by VanderWeele.^{38,39} For ease of notation, we hereafter denote the mediator trajectory and its realisation often simply by M and m , respectively.

3.2 Causal model

The causal model is represented graphically by the *directed acyclic graph* (DAG) shown in Figure 1, describing the assumed causal structure of the relevant variables involved. The main interest lies in the interplay between the treatment A (lifestyle intervention), the mediator trajectory M (BMI trajectory) and the restricted survival time T as the outcome. The DE comprises the paths from the treatment to the survival time that bypass the mediator M , including the portion of the effect that is mediated through the lifestyle score (path $A \rightarrow L \rightarrow T$). The IE comprises the paths that transmit the effect of the treatment to the survival time through the mediator M . The treatment-dependent confounder L represents lifestyle behaviour that, while on a path between treatment and outcome, simultaneously confounds the association between the mediator of interest (M) and the outcome (T). The variable set W consists of the confounding baseline variables that are considered to affect the mediator trajectory, the restricted survival time and the lifestyle behaviour.

We assume that individual-specific random effects (R_0, \mathbf{R}) are present and influence the outcome through the underlying mediator trajectory. In addition, R_0 is allowed to have a separate, direct effect, constituting frailty for the time-to-event outcome, and furthermore, a confounder for the causal mechanism. The random effects can be seen as some properties of the individuals which could in principle be measured and, although unobserved in reality, they can be learned through their influence on the repeated measurements of the mediator. A similar approach has previously been adopted to estimate the mediational effects of certain drug treatments on overall survival mediated by CD4 cell count.²⁴

We interpret the DAG in the framework of nonparametric structural equations models (NPSEM).¹¹ The directed arrows imply a causal ordering between the variables, and an absence of an arrow between two variables implies no direct causal relationship between them. Importantly, the absence of any bidirectional arrows implies an assumption that any randomness affecting one variable in the graph is independent of the randomness affecting any other variable.

3.3 Identification of the causal effects

Under the NPSEM framework, the following conditional independence assumptions are implied by the causal DAG of Figure 1:

1. $T_{a,l,m} \perp \{A, L, M\} \mid W, R_0$
2. $M_{a,l} \perp \{A, L\} \mid W$

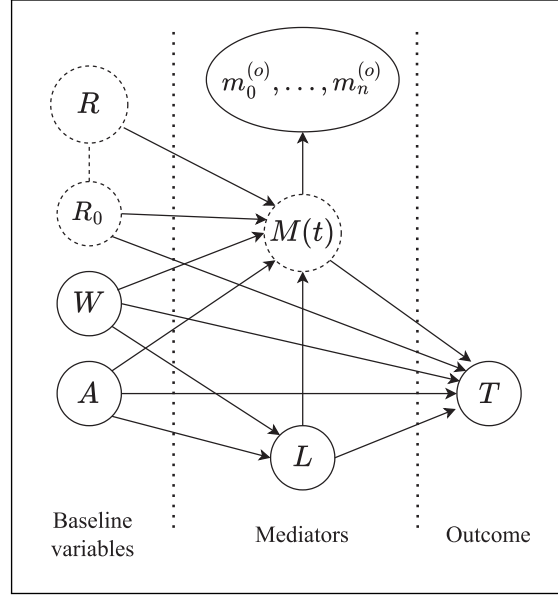


Figure 1. Directed acyclic graph describing the assumed causal mechanism within an individual. The effect of treatment A on the restricted survival time T is assumed to be mediated through the underlying mediator trajectory $M(t)$ of which n repeated measurements $m_1^{(o)}, \dots, m_n^{(o)}$ are observed over time. The random effect R_0 is shared between the mediator $M(t)$ and the time-to-event outcome T , while the random effects R affect only the mediator. The dashed line between R_0 and R implies that they are correlated. L is the treatment-dependent confounder and W consists of the baseline confounders. The dashed nodes refer to latent (unobserved) variables.

3. $L_a \perp A \mid W$
4. $M_{a,l} \perp \{L_a, L_{a^*}\} \mid W$
5. $T_{a,l,m} \perp \{L_a, L_{a^*}, M_{a^*,l}\} \mid W, R_0$.

The first three assumptions translate to no unmeasured confounding between A , L , M and T . The assumptions 4 and 5 are *cross-world independencies*, essentially requiring that there exist no further confounders between L , M and T that themselves are affected by the treatment. In addition, we take the missing data mechanism to be missing at random, that is the right censorings of the survival process and drop-outs in the longitudinal follow-ups are considered non-informative. Under these assumptions, the DE and IE (1) can be identified up to a term containing the joint distribution of (L_a, L_{a^*}) . We will show below that in the case of a trichotomous treatment-dependent confounder L and assuming the monotonicity of the treatment effect on L will further allow the identification of the estimands up to a stratum-specific sensitivity parameter.

Let m^\dagger and l^\dagger denote some reference trajectory of the mediator and a reference value of L , respectively. Following Tchetgen Tchetgen and Vanderweele,¹⁵ the expected outcome, given A , M , L , W and R_0 , can be decomposed into four terms representing the main effects of M and L , the interaction between M and L , and a reference level:

$$\begin{aligned} \beta_m(a, m, \mathbf{w}, r_0) &= \mathbb{E}(T|a, m, l^\dagger, \mathbf{w}, r_0) - \mathbb{E}(T|a, m^\dagger, l^\dagger, \mathbf{w}, r_0), \\ \beta_l(a, l, \mathbf{w}) &= \mathbb{E}(T|a, m^\dagger, l, \mathbf{w}) - \mathbb{E}(T|a, m^\dagger, l^\dagger, \mathbf{w}), \\ \beta_{m,l}(a, m, l, \mathbf{w}) &= \mathbb{E}(T|a, m, l, \mathbf{w}) - \mathbb{E}(T|a, m^\dagger, l, \mathbf{w}) \\ &\quad - \mathbb{E}(T|a, m, l^\dagger, \mathbf{w}) + \mathbb{E}(T|a, m^\dagger, l^\dagger, \mathbf{w}), \\ \bar{\beta}_{a,\mathbf{w}}(a, \mathbf{w}) &= \mathbb{E}(T|a, m^\dagger, l^\dagger, \mathbf{w}). \end{aligned}$$

The ‘no additive interaction’ assumption of Tchetgen Tchetgen and Vanderweele assumes that the terms $\beta_{m,l}(\cdot)$ are zero for all a , m , l and \mathbf{w} . However, here we will retain these additive interaction terms so that the resulting expressions for the DE and IE are (for details, see Appendix A)

$$\begin{aligned} \text{DE} &= \text{DE}^{(r)} - \Delta_{DE} + \delta, \\ \text{IE} &= \text{IE}^{(r)} + \Delta_{IE} - \delta, \end{aligned} \tag{2}$$

where

$$\begin{aligned}
DE^{(r)} &= \iiint_{m,w,r_0} [\beta_m(a, m, \mathbf{w}, r_0) - \beta_m(a^*, m, \mathbf{w}, r_0)] P_M(m|a^*, \mathbf{w}, r_0) \\
&\quad \times P_W(\mathbf{w}) P_{R_0}(r_0) dm d\mathbf{w} dr_0 \\
&\quad + \iint_{l,w} [\beta_l(a, l, \mathbf{w}) P_L(l|a, \mathbf{w}) - \beta_l(a^*, l, \mathbf{w}) P_L(l|a^*, \mathbf{w})] P_W(\mathbf{w}) dl d\mathbf{w} \\
&\quad + \int_{\mathbf{w}} [\bar{\beta}_{a,w}(a, \mathbf{w}) - \bar{\beta}_{a,w}(a^*, \mathbf{w})] P_W(\mathbf{w}) d\mathbf{w}, \\
IE^{(r)} &= \iiint_{m,w,r_0} \beta_m(a, m, \mathbf{w}, r_0) [P_M(m|a, \mathbf{w}, r_0) - P_M(m|a^*, \mathbf{w}, r_0)] \\
&\quad \times P_W(\mathbf{w}) P_{R_0}(r_0) dm d\mathbf{w} dr_0, \\
\Delta_{DE} &= \int \cdots \int_{m,l',w,r_0} \beta_{m,l}(a^*, m, l', \mathbf{w}, r_0) P_M(m|a^*, l', \mathbf{w}, r_0) P_L(l'|a^*, \mathbf{w}) \\
&\quad \times P_W(\mathbf{w}) P_{R_0}(r_0) dm dl' d\mathbf{w} dr_0, \\
\Delta_{IE} &= \int \cdots \int_{m,l,w,r_0} \beta_{m,l}(a, m, l, \mathbf{w}, r_0) P_M(m|a, l, \mathbf{w}, r_0) P_L(l|a, \mathbf{w}) \\
&\quad \times P_W(\mathbf{w}) P_{R_0}(r_0) dm dl d\mathbf{w} dr_0, \\
\delta &= \int \cdots \int_{m,l,l',w,r_0} \beta_{m,l}(a, m, l, \mathbf{w}, r_0) P_M(m|a^*, l', \mathbf{w}, r_0) P(L_a = l, L_{a^*} = l' | \mathbf{w}) \\
&\quad \times P_W(\mathbf{w}) P_{R_0}(r_0) dm dl dl' d\mathbf{w} dr_0.
\end{aligned}$$

The part of the TE that is transmitted through the additive interaction of M and L on T is given by $\Delta_{IE} - \Delta_{DE}$, which is identifiable. The term δ controls how this effect is divided into the DE and IE but is not itself identifiable as its expression relies on the joint probability of counterfactual levels of L . This also means that the DE and IE are not identifiable without further assumptions.

Tchetgen Tchetgen and VanderWeele showed that assuming no additive interaction of M and L on T , or, in the case of a binary L , assuming monotonicity of the effect of A on L , is sufficient to identify δ and, therefore, also the DE and IE from empirical data.¹⁵ The ‘no additive interaction’ assumption would imply $\Delta_{DE} = \Delta_{IE} = \delta = 0$, and subsequently $DE = DE^{(r)}$ and $IE = IE^{(r)}$. The monotonicity assumption of the effect of A on L means that an individual cannot have a worse value of L under treatment than they would have had under no treatment. If L is binary, the joint probability of L_a and L_{a^*} becomes fully determined by the marginal probabilities, which can be estimated from the observed data.

We now extend the monotonicity assumption to a trichotomous L and show that, in each of the strata defined by the baseline covariates, the joint probability is identified up to a sensitivity parameter. For convenience, we omit denoting the stratum in what follows. Suppose that L is trichotomous, taking values $L \in \{0, 1, 2\}$, and the treatment effect on L is monotonic. The stratum-specific joint probability of (L_a, L_{a^*}) can then be represented in terms of six probability parameters:

		L_a			
		0	1	2	
L_{a^*}	2	0	0	p_{22}	Φ_6
	1	0	p_{11}	p_{12}	Φ_5
	0	p_{00}	p_{01}	p_{02}	Φ_4
		Φ_1	Φ_2	Φ_3	

where the marginal probabilities Φ can be estimated (i.e. are identifiable) from the observations. We obtain directly $p_{00} = \Phi_1 = P(L = 0|a)$ and $p_{22} = \Phi_6 = P(L = 2|a^*)$. Conditionally on the marginals, there is thus only one degree of freedom for the remaining four parameters. Considering p_{11} as the free parameter, it is constrained by the marginals to the interval

$$p_{\min} = \max\{0, 1 - \Phi_3 - \Phi_4\} \leq p_{11} \leq \min\{\Phi_2, \Phi_5\} = p_{\max}. \quad (3)$$

The choice of p_{11} can be parameterised by a sensitivity parameter $\rho \in [0, 1]$ as follows:

$$p_{11} = p_{\min} + \rho(p_{\max} - p_{\min}). \quad (4)$$

The sensitivity parameter ρ indicates the relative location of p_{11} within its admissible interval. It can also be interpreted as the assumed heterogeneity of the treatment effect on the treatment-dependent confounder. Assigning more probability mass to p_{11} (i.e. no effect) necessarily increases the probability p_{02} (i.e. large effect) and decreases the probabilities p_{01} and p_{12} (i.e. moderate effects).

For a fixed ρ , the term δ is identified and its range of possible values can be calculated by varying the sensitivity parameter ρ within the $[0, 1]$ interval. The lower and upper bounds for the DE and IE can be found by minimising and maximising δ within each stratum. Since each δ is linear in the probabilities p_{ij} , its minimum and maximum values for a given stratum are found at the opposite ends of the ρ interval. Furthermore, (3) implies that $1 - \Phi_3 - \Phi_4 \leq \min\{\Phi_2, \Phi_5\}$ is a necessary (but not sufficient) condition for the monotonicity assumption to hold. The marginal probabilities Φ can be estimated from data and thus allow one to assess based on empirical data whether the monotonicity assumption can be considered viable. Generalisation to the case with a treatment-dependent confounder with arbitrary number of levels is given in Appendix B.

If $p_{11} = 1 - \Phi_3 - \Phi_4$, then $p_{02} = 0$ and we obtain a special case, here referred to as *step monotonicity*. This assumption means that the treatment can either have no effect on L , or elevate L to one level higher. Under step monotonicity, the joint probability for L_a and L_{a^*} is identifiable by the marginals without the need of the sensitivity parameter, provided that the marginals are consistent with the step monotonicity. Of note, under step monotonicity the identifiability holds for any ordinal variable L with any number of possible values (for details, see Appendix C).

3.4 Parametric models

Under the identifying assumptions of Section 3.3, the DE and IE can be expressed in terms of the observed data. Although the effects and their corresponding empirical expressions were derived nonparametrically, the components forming the empirical expressions were estimated parametrically. In this section we describe the parametric models we used to estimate the terms in the expressions (2) for the DE and IE.

3.4.1 Mediator trajectory. We assumed a linear mixed model for the mediator trajectory. The underlying true mediator was assumed to be a smooth trajectory $M_i(t)$ from which observations are made with stochastic deviations. In particular, the observed longitudinal measurements $m_{ij}^{(o)}$ for individual i were assumed to arise from the model

$$m_{ij}^{(o)} = M_i(t_{ij}) + \epsilon_{ij}, \quad j = 1, \dots, n_i, \\ M_i(t) = (\beta_0 + R_{i0}) + \beta_1' X_i + \beta_2' W_i + \sum_{k=1}^4 (\alpha_k + \psi_k' X_i) B_k(t) + \sum_{k=1}^3 R_{ik} B_k'(t), \quad (5)$$

where $X_i = (A_i, I(L_i = 1), I(L_i = 2), A_i I(L_i = 1), A_i I(L_i = 2))'$ and W_i is a vector containing the baseline covariates (age, sex, smoking status and lifestyle score). The terms B_k are population-level basis functions for natural cubic splines with outer knots placed at the baseline and 10 years and three inner knots placed at 1, 3 and 5 years since study onset. Similarly, B_k' are basis functions corresponding to individual-level random effects, with two inner knots placed at years 1 and 5. The ϵ_{ij} are mutually independent, normally distributed error terms with variance σ^2 and the random effects $(R_{i0}, R_{i1}, R_{i2}, R_{i3})'$ have a multinormal distribution with mean zero and a full covariance matrix Σ . Note that the model includes time-independent effects of the treatment and treatment-dependent confounder on the outcome. This entails the assumptions that the treatment-dependent confounder is affected by the treatment without delay, and it in turn affects the mediator trajectory without delay.

3.4.2 Time-to-event outcome. We assumed a parametric proportional hazards model for the time-to-event outcome with separate piecewise-constant baseline hazards for the two treatment groups. Denoting the baseline hazard functions as $h_{00}(t)$ and $h_{01}(t)$, the hazard for individual i was modelled as

$$h_i(t) = h_{00}(t)^{(1-A_i)} h_{01}(t)^{A_i} \exp \{ \gamma'_1 \mathbf{L}_i + \gamma'_2 \mathbf{L}_i A_i + \gamma'_3 \mathbf{W}_i + g(M_i(\cdot), t) \mathbf{G}'_i \boldsymbol{\zeta} + \xi R_{i0} \}, \quad (6)$$

where $g(\cdot)$ is a function of the full mediator trajectory evaluated at t and determines the parametric form of the dependency between the mediator process and the hazard function. \mathbf{G}_i is the vector $(1, \mathbf{L}'_i, A_i)'$, meaning that parameter ζ_1 is interpreted as the main effect of the mediator, and the rest as interaction effects between the mediator and the levels of the treatment-dependent confounder and the treatment. The individual-level random intercept R_{i0} from the longitudinal submodel enters the hazard function as a frailty, with the parameter ξ describing the strength of its effect on the hazard, thus controlling for latent confounding of the mediator–outcome-relationship.

The choice of the functional form of $g(\cdot)$ should depend on the biological mechanism by which the mediator is assumed to affect the outcome. Here, we choose a three-year legacy parameterisation, that is $g(M_i(\cdot), t) = \int_0^3 [M_i(v) - M_i(0)] dv$ for all $t \geq 0$, implying that the hazard at any time is affected by the cumulative change of the latent trajectory over the first three years since the baseline. We return to the interpretation of this choice in the Discussion. For comparison, we also consider a current change parameterisation where the hazard at time t is affected by the change by time t of the level of the latent trajectory since the baseline, that is $g(M_i(\cdot), t) = M_i(t) - M_i(0)$.

The RMST for an individual i can be derived from the hazard function as

$$\tau_i^{t_{\max}} = \int_0^{t_{\max}} P(\tilde{T}^{(i)} > v) dv = \int_0^{t_{\max}} \exp \left(- \int_0^v h_i(v') dv' \right) dv. \quad (7)$$

The parameters of the proportional hazards model should not themselves be interpreted as causal effects. Such an interpretation would require the implausible assumption that any source of between-individual heterogeneity would be accounted for in the model, for otherwise the parameters would suffer from selection bias due to unobserved heterogeneity.⁴⁰

3.4.3 Treatment-dependent confounder. We used multinomial logistic regression to model the dependence of the trichotomous treatment-dependent confounder L on treatment A and the baseline confounders \mathbf{W} . The log-ratios of probabilities for belonging to the categories 1 or 2 compared to the reference category 0 were determined by

$$\begin{aligned} \log \left(\frac{P(L = 1|A, \mathbf{W})}{P(L = 0|A, \mathbf{W})} \right) &= \phi_0^{(1)} + \phi_1^{(1)} A + \phi_2^{(1)} \mathbf{W} =: \varphi^{(1)}(A, \mathbf{W}), \\ \log \left(\frac{P(L = 2|A, \mathbf{W})}{P(L = 0|A, \mathbf{W})} \right) &= \phi_0^{(2)} + \phi_1^{(2)} A + \phi_2^{(2)} \mathbf{W} =: \varphi^{(2)}(A, \mathbf{W}). \end{aligned} \quad (8)$$

The marginal probabilities Φ_1, \dots, Φ_6 of the joint distribution of (L_a, L_{a^*}) in a given stratum \mathbf{W} were obtained by applying the inverse-logit transform, for example, $\Phi_2 = \exp\{\varphi^{(1)}(a, \mathbf{W})\} / (1 + \exp\{\varphi^{(1)}(a, \mathbf{W})\} + \exp\{\varphi^{(2)}(a, \mathbf{W})\})$.

The above model includes only the main effects of each predictor and could thus be considered relatively inflexible. As the research question is concerned with the mediating mechanism of the treatment, we also considered models with interaction terms between the treatment and the baseline covariates to allow the treatment effect on the treatment-dependent confounder to differ among the baseline covariate strata. Model comparison was carried out to determine whether any of the more flexible models should be favoured against model (8).

3.4.4 Parametric causal effects. Since the underlying BMI trajectory as the mediator is stripped of the stochastic error terms, its distribution conditionally on the covariates and treatment is determined by the distribution of the random effects. Integrating an arbitrary functional $q(m)$ over the possible realisations of m , given A, L and \mathbf{W} , thus reduces to integrating over the joint distribution of the random effects.

With the definitions of mediational causal effects in (1), the assumed parametric models imply the following formulae for the terms determining the DE and IE:

$$\begin{aligned}
DE^{(r)} &= \int_r \sum_{\mathbf{w}, l'} [\beta_m(a, m(a^*, l', \mathbf{r}, \mathbf{w}), r_0, \mathbf{w}) - \beta_m(a^*, m(a^*, l', \mathbf{r}, \mathbf{w}), r_0, \mathbf{w})] \\
&\quad \times P_R(\mathbf{r}) P_L(l' | a^*, \mathbf{w}) P_W(\mathbf{w}) d\mathbf{r} \\
&\quad + \sum_{l, \mathbf{w}} [\beta_l(a, l, \mathbf{w}) P_L(l | a, \mathbf{w}) - \beta_l(a^*, l, \mathbf{w}) P_L(l | a^*, \mathbf{w})] P_W(\mathbf{w}) \\
&\quad + \int_{r_0} \sum_{\mathbf{w}} [\bar{\beta}_{a,c}(a, \mathbf{w}, r_0) - \bar{\beta}_{a,c}(a^*, \mathbf{w}, r_0)] P_{R_0}(r_0) P_W(\mathbf{w}) dr_0, \\
IE^{(r)} &= \int_r \sum_{\mathbf{w}, l, l'} [\beta_m(a, m(a, l, \mathbf{r}, \mathbf{w}), r_0, \mathbf{w}) P_L(l | a, \mathbf{w}) \\
&\quad - \beta_m(a, m(a^*, l', \mathbf{r}, \mathbf{w}), r_0, \mathbf{w}) P_L(l' | a^*, \mathbf{w})] P_R(\mathbf{r}) P_W(\mathbf{w}) d\mathbf{r}, \\
\Delta_{DE} &= \int_r \sum_{l', \mathbf{w}} \beta_{m,l}(a^*, m(a^*, l', \mathbf{r}, \mathbf{w}), l', \mathbf{w}) P_R(\mathbf{r}) P_L(l' | a^*, \mathbf{w}) P_W(\mathbf{w}) d\mathbf{r}, \\
\Delta_{IE} &= \int_r \sum_{l, \mathbf{w}} \beta_{m,l}(a, m(a, l, \mathbf{r}, \mathbf{w}), l, \mathbf{w}) P_R(\mathbf{r}) P_L(l | a, \mathbf{w}) P_W(\mathbf{w}) d\mathbf{r}, \\
\delta &= \int_r \sum_{l, l', \mathbf{w}} \beta_{m,l}(a, m(a^*, l', \mathbf{r}, \mathbf{w}), l, r_0, \mathbf{w}) P_R(\mathbf{r}) P(L_a = l, L_{a^*} = l' | \mathbf{w}) \\
&\quad \times P_W(\mathbf{w}) d\mathbf{r}.
\end{aligned}$$

Fully parametric expressions can then be obtained by plugging in the assumed joint distribution for the random effects, the parametric forms of $P_L(l | a, \mathbf{w})$ as implied by model (8), and $\beta(\cdot)$ as implied by (7) and (6). Lastly, under the monotonicity assumption, the parametric expression of the joint distribution $P(L_a = l, L_{a^*} = l' | \mathbf{w})$ can be obtained from model (8) by choosing the sensitivity parameter ρ for each stratum (see equation (4)).

3.5 Estimation

We employed a joint modelling framework to estimate the parametric models of the mediator trajectory and the time-to-event outcome. Since these models rely on only the observed values of the treatment-dependent confounder, the model for the treatment-dependent confounder was estimated separately. Pareto smoothed importance sampling leave-one-out cross-validation⁴¹ was used to compare the treatment-dependent confounder model against its more flexible variations.

A joint model for the longitudinal and time-to-event outcomes comprises specifying submodels for both outcomes and linking them via some association structure, thus allowing incorporating any information shared between the two outcomes.^{42–45} The association structure was here induced by including a function of the longitudinal mediator trajectory into the linear predictor of the survival submodel and also assuming a random effect which is shared between the mediator trajectory and the time-to-event outcome, enabling adjustment for a latent confounder.

We adopted a Bayesian framework to estimate all parametric models. Let $\theta = (\beta, \alpha, \psi, \sigma, \Sigma, \gamma, \zeta, \xi, h_{00}, h_{01})$ denote all population-level model parameters in the joint model. The joint distribution of observed longitudinal measurements $\mathbf{m}_i^{(o)}$, event time since baseline T_i^{exit} , event (T2D) indicator d_i , and random effects \mathbf{R}_i can be factorised as

$$\begin{aligned}
p(\mathbf{m}_i^{(o)}, T_i^{\text{exit}}, d_i, \mathbf{R}_i | A_i, L_i, \mathbf{W}_i; \theta) &= p(\mathbf{m}_i^{(o)} | \mathbf{R}_i, A_i, L_i, \mathbf{W}_i; \theta) \\
&\quad \times p(T_i^{\text{exit}}, d_i | \mathbf{R}_i, A_i, L_i, \mathbf{W}_i; \theta) p(\mathbf{R}_i; \theta) \\
&=: p_i^m \times p_i^s \times p_i^r
\end{aligned}$$

The underlying mediator trajectory entering the survival submodel is determined by the random effects and the baseline covariates, making the survival observations conditionally independent of the actual measurements of the longitudinal process (see Figure 1 and Section S3 in the online Supplemental Material). The posterior distribution of the parameter

vector θ is

$$p(\theta|\text{data}) \propto \left(\prod_i p_i^m \times p_i^s \times p_i^r \right) \times p(\theta), \text{ where}$$

$$p_i^m \propto \sigma^{-n_i/2} \exp\left\{-\frac{1}{2}[\mathbf{m}_i^{(o)} - \mathbb{E}(\mathbf{m}_i^{(o)})]'[\mathbf{m}_i^{(o)} - \mathbb{E}(\mathbf{m}_i^{(o)})]\sigma^{-2}\right\},$$

$$p_i^s = h_i(t_i^{\text{exit}})^{d_i} \exp\left\{-\int_0^{t_i^{\text{exit}}} h_i(u) du\right\},$$

$$p_i^r \propto \det(\Sigma)^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{R}_i'\Sigma^{-1}\mathbf{R}_i\right\},$$

and $p(\theta)$ is the prior distribution for θ . The cumulative hazard in p_i^s does not, in general, have a convenient analytical form, as it may involve a time-dependent functional of the mediator trajectory, and numerical integration was needed for its computation.

For all regression parameters, we assumed relatively uninformative normal priors with mean zero and standard deviation 5. As the longitudinal BMI measurements were centred at $25\text{kg}/\text{m}^2$ and scaled to one fifth of the original scale, effect sizes greater than 5 would be implausible. Similarly, log hazard ratios of such magnitudes would be considered unrealistic. For the piecewise constant baseline hazards, we used $\text{Gamma}(.5, .5)$ priors for each piece. For the standard deviations of the residual terms in the longitudinal submodel and the random effects, half-Cauchy distributions were used with location parameter zero, and scale parameter 10. In addition, a Lewandowski–Kurowicka–Joe prior was assigned to the Cholesky factor of the random effects correlation matrix. The parameters ϕ in the multinomial model for the treatment-dependent confounder were given normal priors with mean 0 and standard deviation 5.

The models were run using four parallel chains with 2000 burn-in iterations and 2000 sampling iterations each, resulting in 8000 Markov chain Monte Carlo (MCMC) draws from the posterior distribution of the model parameters and the convergence was assessed by the Gelman–Rubin R -statistic,⁴⁶ which was at most 1.01 for all parameters. All computations were performed on a workstation equipped with an Intel® Xeon® w3-2435 processor. Estimating the joint model required approximately 1 hour using the legacy parameterisation and 6 hours using the current change parameterisation. To test the implementation of our framework and to assess how its performance depends on the sample size, we simulated 100 datasets with sample sizes of 100, 300, 500, or 1000, and compared the distributions of the point estimates (posterior means) with sample-based ‘true’ TE, DE and IE (Section S4 in the online Supplemental Material). With a sample size of 100, the estimates exhibited substantial variability and were slightly biased. With a sample size of 500, the estimation was already deemed satisfactory. All models were estimated using the R interface to the *Stan* software,⁴⁷ and post-processing was conducted using the R software⁴⁸ (utilising packages *survival*,⁴⁹ *statmod*,⁵⁰ *splines2*⁵¹ and *tidyverse*⁵²).

4 Application

The DPS originally followed a cohort of 522 individuals, with 265 randomised to the intervention group and 257 to the control group. After excluding 19 individuals with missing baseline covariate values, the final sample consisted of 503 individuals with 254 people in the intervention group and 249 in the control group. The median number of clinical visits per person was 11 (interquartile range 6–13) in the intervention group and 9 (IQR 4–13) in the control group. The intervention group contributed a total of 3505 person years during which 166 T2D cases were observed with a RMST of 11.2 years. The control group contributed a total of 2865 person years and 168 T2D cases with a RMST of 9.5 years.

The study endpoint was the diagnosis of T2D, either ascertained at any of the study visits or inferred from the register data. Since the clinical study visits were considered the more reliable source, we used the first diagnosis made at the study visits as the primary endpoint and considered the register data only after their last study visit for each individual. Figure 2(a) shows the cumulative incidences of T2D from both sources. The *clinical risk set* refers to the number of individuals in the risk set having not yet made their last study visit, whereas the *register risk set* is the number of individuals in the risk set being followed through the drug registers. The cumulative incidence curves based on the study visits and the drug registers grow reasonably closely in proportion to the number of individuals, implying that any bias due to uneven sensitivity of T2D detection was unlikely.

For each study participant, the follow-up started at the baseline visit and terminated at the event of T2D diagnosis, death or end of follow-up at the end of 2018. The dates and causes of deaths were obtained from the Finnish Cause of Death Register. Figure 2(b) shows the cumulative risks for the competing events of T2D diagnosis, death with a potentially T2D-related cause, that is cardiovascular and cerebrovascular complications, and death from other causes. As the proportion of potentially T2D-related deaths appears negligible, treating all deaths as uninformative right censorings was deemed

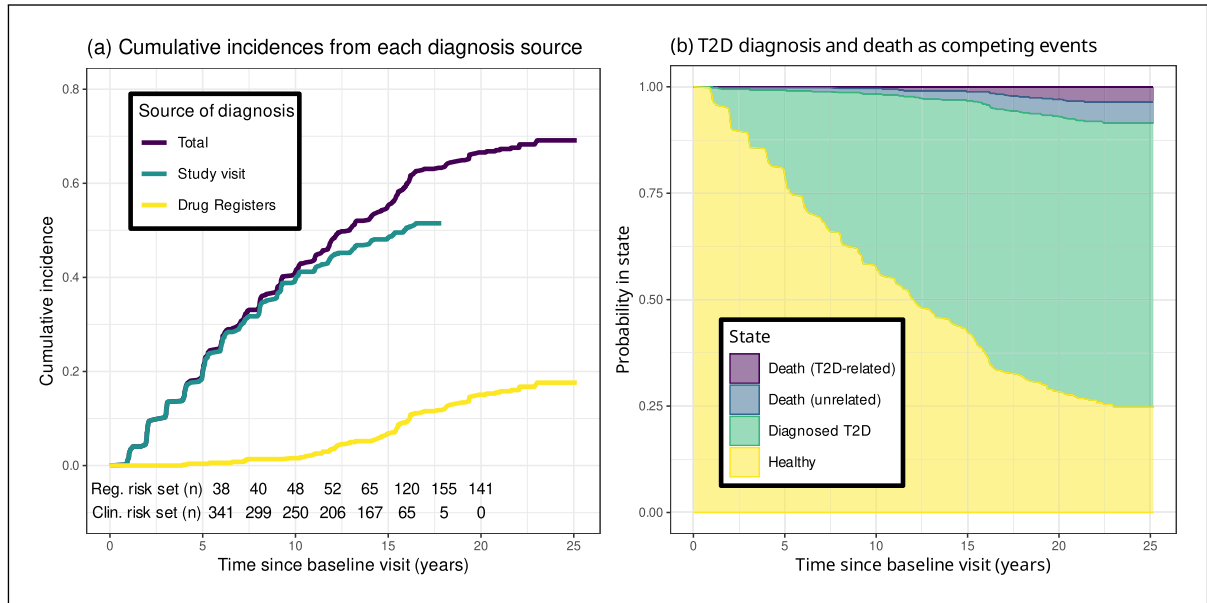


Figure 2. (a) Cumulative incidence of type 2 diabetes (T2D) diagnoses due to both sources of diagnosis. Underneath the curves, the number of individuals being followed through each source is shown. (b) Probabilities of reaching the three competing (absorbing) states as functions of time. T2D-related deaths include cases in which the cause of death is potentially associated with T2D, namely cardiovascular and cerebrovascular complications.

justified. With the availability of the register data after the clinical follow-up, deaths were the only source of censoring in the data.

4.1 Model fit

Figure 3(a) shows the observed and estimated mean BMI trajectories for the two treatment groups. Figure 3(b) to (d) display the Kaplan–Meier curves and estimated survival functions from model (6), illustrating how accurately the model reproduces the observed dependence of T2D survival on the treatment, change in BMI over the first three years, and the lifestyle score. In these plots, the cumulative change in BMI over the first three years was categorised into tertiles of the individual-level point estimates obtained from the model. The survival functions were then computed as the means of the estimated survival functions for each of the three groups. These plots are based on the model using the three-year legacy parameterisation

The mean BMI trajectory in the intervention group (Figure 3(a)) exhibits a decrease in the early years since the start of the intervention, after which the mean trajectory gradually rebounds close to the trajectory of the control group. The estimated curves replicate the observed average trajectories reasonably well, although the shape showed slight differences in the intervention group over the first two years. The Kaplan–Meier curves indicate a lower T2D risk for individuals in the intervention group (Figure 3(b)), in the higher lifestyle score groups (Figure 3(d)) or having decreased their BMI more over the first three years (Figure 3(c)). The estimated survival functions agree closely with the Kaplan–Meier curves. However, with the early BMI change, the differences in the estimated survival functions are less pronounced than those suggested by the Kaplan–Meier curves, which might indicate some lack of fit in the model with respect to the relationship between the early BMI change and T2D hazard.

4.2 Model comparison and assessment of the monotonicity assumption

In addition to model (8) of the treatment-dependent confounder, we considered more flexible models including interactions of the treatment with each baseline covariate all at the same time or each one separately. The model comparison showed no discernible difference between the model's performances and so we chose the most parsimonious one, that is, model (8) (Section S1 in the online Supplemental Material).

The monotonicity condition under the selected model was assessed from the empirical data by estimating the marginal probabilities for the two counterfactual lifestyle scores (L_a and L_{a^*}) for each of the 54 strata formed by the baseline covariates (age, sex, smoking and the baseline lifestyle score) and checking whether the marginals satisfied the necessary

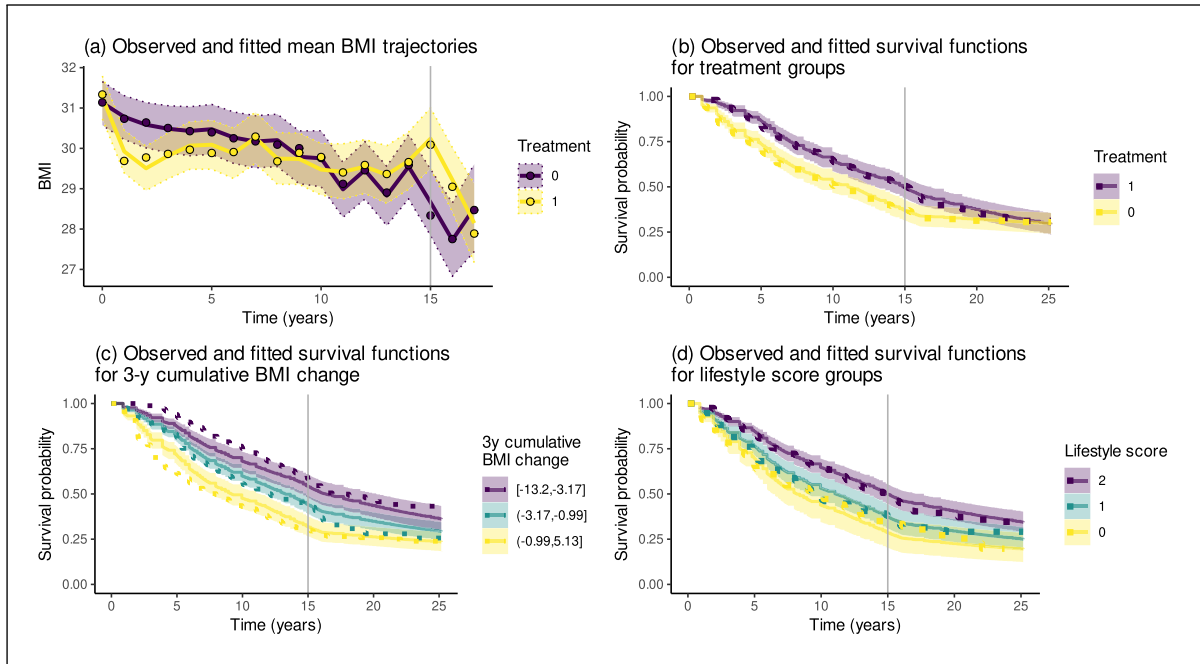


Figure 3. (a) The observed (dots) and estimated (solid lines) mean trajectories of body mass index (BMI) in the intervention (1) and control (0) groups with the 95% credible intervals. (b)–(d) Kaplan–Meier curves illustrating the effects of the treatment, early BMI change and lifestyle score on avoiding type 2 diabetes. The early BMI change is here categorised into the model-implied tertiles. The dotted lines are the Kaplan–Meier estimates, and the solid lines and the accompanying 95% credible intervals are the estimated survival functions.

condition (3) for the monotonicity assumption. Out of the 8000 MCMC samples, only 205 (2.6%) contained one or more strata for which the marginals were not consistent with the monotonicity assumption. Figure 4(a) shows the proportion of these failures for each stratum, along with the number of individuals in each stratum. The strata with the highest proportion of failures were the ones with low coverage in the data, which seems reasonable since those would be expected to be the least trustworthy. The results of the model comparison and investigating the monotonicity assumption under the alternative models are provided in Section S1 of the online Supplemental Material.

Figure 4(b) shows, for each stratum, the average boundaries within which the probability p_{11} was constrained among the monotonicity-consistent posterior draws. The interval widths are mostly around ten percent-units, suggesting that the constraints imposed by the monotonicity assumption were relatively strict and the treatment-dependent confounder should have had substantial effects on the mediator and the outcome to produce a significant change in the results when varying the sensitivity parameter.

We also investigated the agreement of the step monotonicity assumption with the data. Under the step monotonicity assumption, 3279 (41%) of the posterior draws resulted in an improper joint distribution of (L_a^*, L_a) for at least one stratum, suggesting the step monotonicity to be an overly restrictive assumption.

4.3 Total, direct and indirect effects

Table 1 shows the posterior means and 95% credible intervals (CIs) of the DE, IE and TE by choosing the stratum specific sensitivity parameters so that the mediational effects were either minimised or maximised. In addition, the value $\rho = 0.5$ was used homogeneously in all strata. The estimated effects remained virtually the same regardless of the choice. We therefore discuss the results with the choice $\rho = 0.5$.

Under the three-year legacy parameterisation, the estimated indirect treatment effect, that is the effect mediated through the change in BMI, amounts to roughly one year of additional time without T2D over the 15 years since the treatment onset (95% CI 0.13–2.1). The point estimate of the direct treatment effect was 0.60 years with a very wide credible interval extending to the negative side. Within the posterior range of the TE, the IE was almost always positive. The smaller the TE, the more prominent was the role of the IE (see Section S2 in the online Supplemental Material). The estimated total treatment effect amounts to 1.6 years of additional time remaining free of T2D (95% CI –0.08–3.0 years). For comparison,

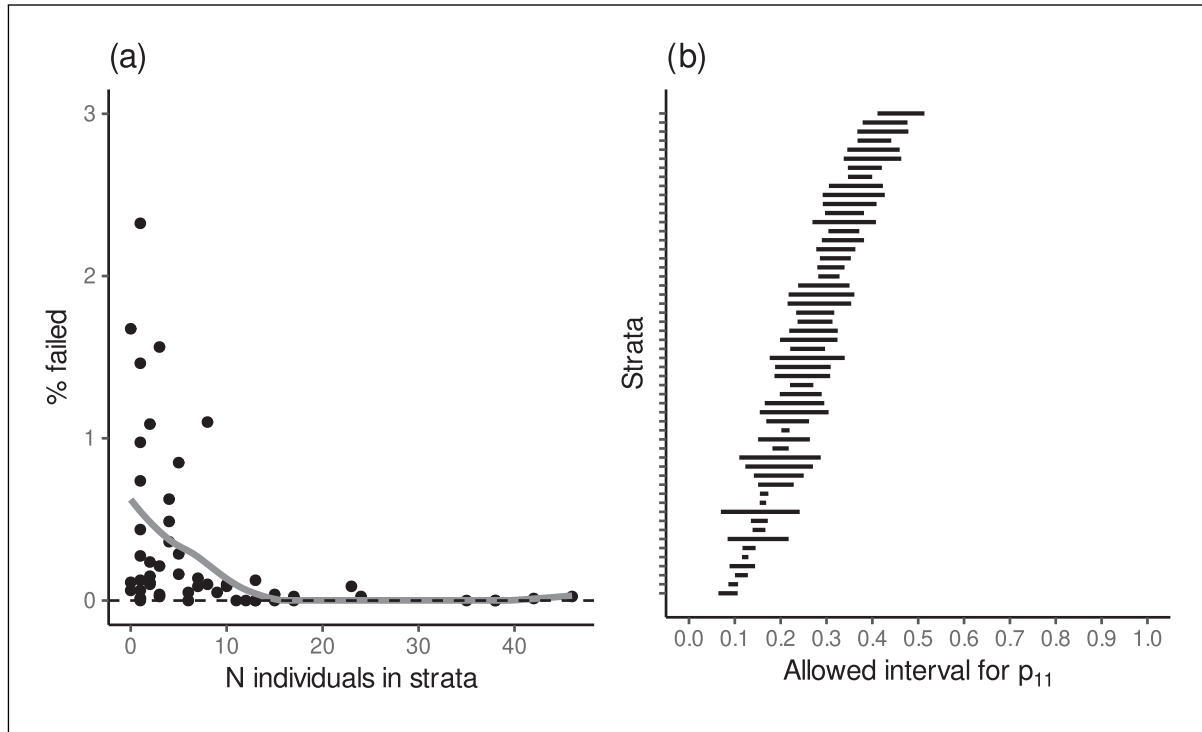


Figure 4. (a) Proportions of Markov chain Monte Carlo samples failing the monotonicity condition in the 54 strata defined by the covariates (age, sex, smoking and baseline lifestyle score) under the selected model. The largest proportions of failures occurred in strata with low numbers of individuals. (b) The average posterior boundaries within which the free parameter in the joint distribution of the two counterfactual treatment-dependent confounders (p_{11}) was constrained by the monotonicity assumption in each stratum.

Table 1. The posterior means and 95% credible intervals (in parentheses) of the DE, IE and TE under the two alternative functionals and different choices of stratum-specific sensitivity parameters ρ (see equation (4)); min/max (choosing the parameter to minimise/maximise the mediational effects separately in each covariate stratum; 0.5 (choosing $\rho = 0.5$ in all strata).

ρ	Three-year legacy			Current change		
	DE	IE	TE	DE	IE	TE
min	0.530 (-1.74, 2.20)	0.888 (0.068, 1.95)	1.57 (-0.082, 2.95)	1.24 (-0.012, 2.74)	0.243 (-0.142, 0.680)	1.61 (0.313, 3.09)
0.5	0.600 (-1.72, 2.24)	0.970 (0.133, 2.10)	1.57 (-0.082, 2.95)	1.31 (0.051, 2.81)	0.306 (-0.081, 0.748)	1.61 (0.313, 3.09)
max	0.689 (-1.57, 2.34)	1.05 (0.217, 2.13)	1.57 (-0.082, 2.95)	1.37 (0.116, 2.87)	0.369 (-0.018, 0.814)	1.61 (0.313, 3.09)

DE: direct effect; IE: indirect effect; TE: total effect.

a nonparametric estimator of the TE based on the areas under Kaplan–Meier curves of the two treatment groups yielded an estimate of 1.7 years of additional T2D-free time (95% confidence interval 0.85–2.6).

The current change parameterisation resulted in a more pronounced DE of 1.3 years of additional time free of T2D (95% CI 0.05–2.8), while the IE was 0.31 years (95% CI -0.08–0.75). The TE remained similar to that under the three-year legacy parameterisation, however, the CI became slightly narrower.

As it could be considered a plausible a priori assumption, that the effects of the intervention cannot be negative, the results are also shown by discarding posterior draws (23% of the total MCMC samples under the three-year legacy and 8% under the current change parameterisation) giving a negative estimate for the DE or the IE (Table 2). Under such a restriction, the proportion mediated, that is IE/TE, becomes a well defined quantity and describes the proportion of the TE that may be attributed to the indirect mechanism. With the three-year legacy parameterisation, the DE and the TE were larger, while the IE was slightly smaller. The estimated proportion mediated was roughly half, however, the 95% CI extended over the entire [0, 1] interval. With the current change parameterisation there was little change in results

Table 2. The posterior means and 95% credible intervals (in parentheses) of the DE, IE and TE under the two alternative functionals after discarding MCMC samples giving negative treatment effect estimates.

Model	DE	IE	TE	Proportion mediated	N discarded MCMC samples
Three-year	0.980 (0.076, 2.30)	0.831 (0.150, 1.61)	1.81 (0.860, 3.04)	0.487 (0.092, 0.940)	1776 (23%)
Current	1.33 (0.235, 2.77)	0.331 (0.032, 0.751)	1.66 (0.523, 3.11)	0.229 (0.020, 0.639)	628 (8%)

The sensitivity parameter was set at $\rho = 0.5$. MCMC: Markov chain Monte Carlo; DE: direct effect; IE: indirect effect; TE: total effect.

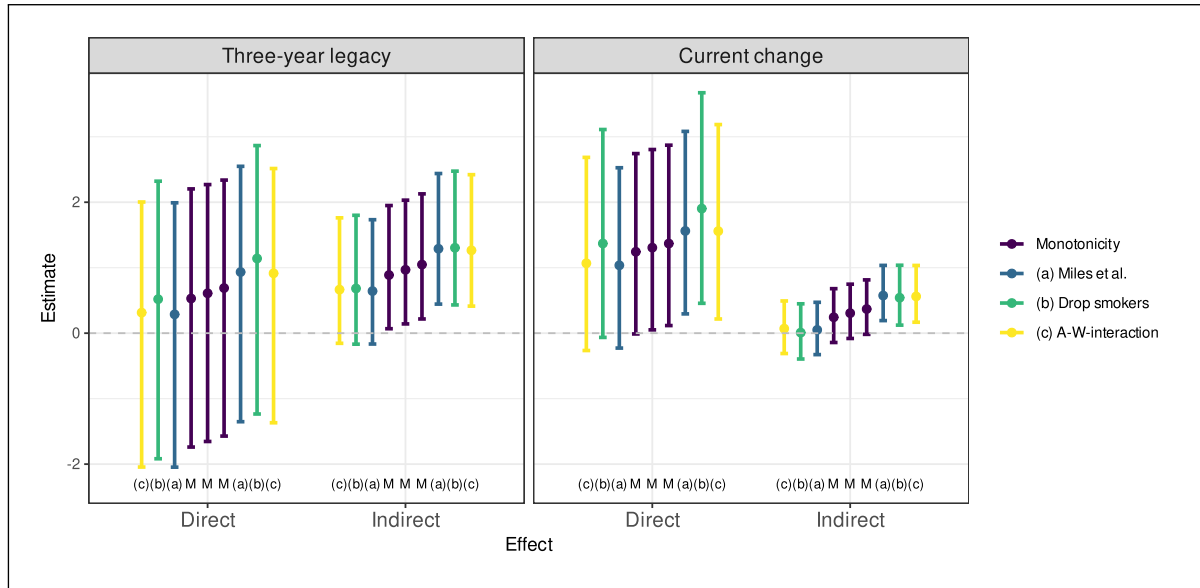


Figure 5. The lower and upper bounds for the estimates of direct and indirect effects and their 95% credible intervals under relaxing the monotonicity assumption. The estimates in the middle (with identifier ‘M’) are the results obtained under the monotonicity assumption. To the left and right are the lower and upper bounds obtained using the approach of Miles et al.¹⁷ with either (a) the full sample; (b) dropping smokers from the sample; or (c) using the treatment-dependent confounder model with all one-way interactions between the treatment and each baseline covariate.

compared with the unrestricted analysis as only 8% of the posterior draws were discarded. The estimated proportion mediated was 20% with a 95% CI from 2% to 64%.

4.4 Sensitivity analyses

For sensitivity analyses, we used the approach of Miles et al.¹⁷ to find lower and upper bounds for the estimates relaxing the monotonicity assumption, that is, optimising the expressions with respect to a joint probability matrix (L_{a^*}, L_a) without the constraints implied by the monotonicity assumption. We considered three scenarios: using either (a) the full data and the same model for L as in the primary analyses; (b) the same model for L but removing current smokers as a potential outlier group from the data; or (c) the full data and the model for L with all treatment-covariate interactions included. The lower and upper bounds for the estimates of the DE and IE and their 95% CIs are shown in Figure 5, along with the results obtained under the monotonicity assumption. As all the estimates are very similar, we conclude that the results are robust against violations of the monotonicity assumption.

Interestingly, the estimated direct effects appear to be higher when current smokers are excluded from the data while the indirect effects are largely unchanged. This suggests that the treatment might be less favourable for smokers than the others. The total effect estimates using the simple estimator based on areas under Kaplan–Meier curves resulted in a TE of 1.9 years for never smokers or former smokers and negative TE of -1.3 years for current smokers. This comparison, however, is very uncertain as smokers comprised a small group of only 30 individuals.

5 Discussion

We addressed causal mediation in the case of a longitudinal mediator and a time-to-event outcome in a randomised lifestyle intervention study, where some of the lifestyle changes acted as treatment-dependent confounders. Such situations may arise in interventional studies aimed at reducing the risk for an adverse health outcome through lifestyle changes. When the interest lies in the treatment effect mediated through changes in a biological risk factor, such as body weight in the current study, other lifestyle changes will act as treatment-dependent confounders, complicating the identification of mediational effects. The methods described here can be applied to address such situations in order to elucidate mechanisms by which treatments affect health outcomes.

Previously, Tchetgen Tchetgen and Vanderweele¹⁵ showed that if the effect of treatment on a binary-valued treatment-dependent confounder is monotonic, mediational effects can be identified from empirical data. We here extended their approach to an ordinal trichotomous case, where a single sensitivity parameter (for each baseline covariate stratum) needs to be specified to identify the mediational effects. The sensitivity parameter determines the joint probability distribution of the two counterfactuals, (L_{a^*}, L_a) , of the treatment-dependent confounder and identifies the component of the total effect that controls how the effect due to the additive interaction of the mediator and the treatment-dependent confounder is divided into the direct and indirect effects. Our approach is related to partial identification, where bounds for the effect estimates can be determined by optimising the equations (2) with respect to the unidentified joint distribution without any structural assumptions on the joint distribution of (L_{a^*}, L_a) .¹⁷ The monotonicity assumption may be seen as a structural assumption that sets some of the joint probabilities to zero. This also implies a necessary condition on the marginals of the joint distribution of (L_{a^*}, L_a) , allowing empirical assessment of the tenability of the monotonicity assumption. In Appendix A2, we generalised the necessary conditions for an ordinal treatment-dependent confounder with arbitrary many levels.

Other approaches to deal with treatment-dependent confounding have been introduced. Identification can be retained if the counterfactuals of the treatment-dependent confounder are assumed to be independent or if one is a deterministic function of the other,¹⁶ or if there is no additive interaction between the treatment and the treatment-dependent confounder.¹⁵ Another option would be to switch the estimation target from natural direct and indirect effects to their interventional analogues, which are identifiable even in the presence of treatment-dependent confounders but do not, in general, share the same interpretation as the natural effects.^{53–55}

We treated the longitudinal mediator (BMI) as a functional entity, that is a smooth function describing the underlying trajectory which the mediator follows and of which observations endowed with stochastic deviations were made over time. The contribution of the stochastic deviations to the association between the mediator and the time-to-event outcome was assumed to be negligible, hence promoting the underlying function itself as the effective mediator. An implicit assumption was that the structure given for the trajectory function is flexible enough to capture the longitudinal process to a relevant extent. By defining the mediator as a functional entity, we avoided the practical difficulties dealing with a high-dimensional mediator, as well as the conceptual challenges that might arise from having to control for post-treatment values of the mediator. Similar approaches treating longitudinally measured mediators as functional entities have been previously investigated in the mediation analysis literature.^{21–23}

We chose the restricted survival time as the time-to-event outcome to avoid issues arising from unmeasured heterogeneity under outcomes defined conditionally on previous survival.⁴⁰ For instance, if there were an unobserved genetic factor influencing T2D risk, conditioning on survival at any post-treatment time would induce a selection bias, potentially invalidating the analysis. In addition, as one third of the study participants did not develop T2D, a substantial tail proportion of the event time distribution remained unobserved. This may introduce the risk of misspecifying the parametric shape of the tail distribution. To address this, we chose the restriction time as 15 years since the baseline to represent a clinically meaningful time period providing reasonable number of follow-up visits and T2D events.

We used a joint modelling framework to estimate the parametric models. The association structure was induced by incorporating the latent mediator trajectory in the survival submodel and also specifying a shared random effects structure.^{42,45} As the mediator was considered a latent quantity, individual-specific mediator trajectories needed to be estimated and the uncertainty regarding the estimation appropriately propagated to the estimates of the causal effects. This propagation of uncertainty was straightforwardly handled in the Bayesian set-up of computations.

Incorporating the mediator into the survival submodel required selecting a functional of its trajectory to serve as a linear predictor. The indirect effect then represents the influence mediated through the property captured by the chosen functional. In our analysis, we considered two functionals, three-year legacy and current change, chosen a priori. Alternatively, one could apply some model selection procedure to identify the best-performing functional from a set of candidates. For example, within the joint modelling literature, Mauff et al.⁵⁶ proposed an approach to determine an optimal weighting function for a cumulative effect measure.

From the causal perspective, shared random effects between the longitudinal and survival submodels represent latent confounding between the mediator and the outcome. The joint modelling framework can thus be employed to account for such unmeasured confounding.²⁴ We here used the random intercept of the mediator trajectory as a shared random effect to reflect a latent property of the individual that may influence both the mediator trajectory and the time-to-event outcome. In our empirical application this might be, for example, an unmeasured metabolism-related genetic factor.

Estimating the mediator trajectories and shared random effect relied on the relatively large number of repeated measurements of the mediator. With a small number of repeated measurements, it may not be possible to estimate the mediator trajectory accurately enough to justify, for example, the current change parameterisation. In such cases, it would be necessary to use simpler functionals that can be estimated reliably from the available data. For example, the three-year legacy parameterisation could be feasible with even a few repeated measurements over the early follow-up.

The primary goal of our empirical application was to decompose the treatment effect on the restricted mean T2D-free time into the indirect (mediated through weight reduction) and direct (all other mechanisms) effects. Using the three-year legacy parameterisation, the estimated IE translated to one year of additional time without T2D over the 15 years after the start of follow-up, with a 95% CI ranging from 1.5 months to two years. The estimated TE was 1.6 years with a 95% CI from negative one month to positive three years. These results suggest that the early weight reduction may indeed constitute a major mechanism through which a lifestyle intervention affects long-term T2D-free survival, accounting for two thirds of the TE as calculated crudely from the point estimates. However, using the current change parameterisation, the IE was estimated to be considerably smaller (0.31 years, 95% CI -0.08 – 0.75 years), accounting, based on the point estimates, roughly 20% of the TE (1.6 years, 95% CI 0.31–3.1). Sensitivity analyses relaxing the monotonicity assumption yielded largely consistent results.

Assuming that the treatment effects cannot be negative had little impact on the estimates under the current change parameterisation, and the proportion mediated was estimated to be 23% with the 95% CI ranging from 2% to 64%. With the three-year legacy parameterisation, however, 23% of the numerical posterior samples were discarded and the DE estimate increased and the IE estimate decreased compared with the unrestricted case. The decrease of the IE estimate after removing samples with negative effects was due to the fact that the two mediational effects had a strong negative correlation. Since the posterior distribution of DE extended far to the negative side, a large number of samples with a very large negative DE, and consequently a very large positive IE, were not admissible.

The choice of the legacy parameterisation was motivated by previous literature demonstrating the persistence of the intervention effects on T2D risk long after the discontinuation of the active treatment and diminishing of the acquired group differences in the clinical risk factors.³⁴ We used the cumulative change at three years as a time-constant predictor in the survival submodel starting from the onset of treatment. We justify this by noting that the individual trajectories are determined by the observed baseline covariates and the random effects and can thus be interpreted to exist at the baseline, even though they may be learned only by observing the trajectory unfold over time. In this context, the weight reduction during the first three years was interpreted as a surrogate for some biological process that responds quickly to the initiation of the lifestyle intervention and the induced behavioural changes. This underlying biological process was then assumed to manifest through the subsequent weight reduction and be the true causal mechanism linking weight reduction to a decrease in T2D risk.


The DE and IE were identified up to sensitivity parameters controlling the probabilities (p_{11}) that the two counterfactual levels of the trichotomous treatment-dependent confounder in each stratum both belong to the middle category (corresponding to an absent treatment effect). The estimated causal effects remained largely unaffected by the sensitivity parameter. This was likely due to the fact that the constraint imposed by the monotonicity assumption on p_{11} (see equation (3)) was very stringent. Furthermore, increasing the probability of an absent treatment effect necessarily also increases the probability (p_{02}) of a large treatment effect, simultaneously decreasing the probabilities for intermediate treatment effects (p_{01} and p_{12}). If the treatment-dependent confounder has a monotonic effect on the mediator and the outcome, this may be seen as a self-regulating property, since the highest and lowest effects must always be up- or downweighted together.

Our empirical analysis has some potential limitations. First, as is always the case in causal analyses, the validity of inferences relies on the untestable assumption of no unmeasured confounding. We used age, sex, smoking and the baseline lifestyle score as baseline covariates and used a shared random effects structure in the joint model to include a latent confounder of the BMI trajectory and T2D incidence. Randomisation addresses confounding involving the treatment. However, it is possible that some important confounders between the treatment-dependent confounder, mediator and outcome were not accounted for. Second, the lifestyle score serving as the treatment-dependent confounder was constructed somewhat crudely by summarising four variables capturing lifestyle changes across the first three follow-up visits. Because this confounder plays a key role in the causal mechanism, any imprecision in its measurement may bias the resulting causal effect estimates. Our decision to use a three-level categorisation reflected a compromise between measurement accuracy and limiting the dimensionality of the unidentified parts of the causal model. Third, the generalisability of the results is


limited. The DPS inclusion criteria selected volunteers who were already overweight, had developed impaired glucose tolerance and were between 40 and 65 years of age at the screening visit but had not yet been diagnosed with T2D. As such, the DPS cohort represents a selected population, that is, individuals at a high risk of T2D who had managed to avoid the disease until a relatively old age.

In conclusion, we investigated causal mediation in longitudinal intervention studies with a time-to-event outcome in the presence of an ordinal treatment-dependent confounder. Foremost, we showed that assuming monotonicity of the treatment effect on a trichotomous ordinal treatment-dependent confounder, the DE and IE can be identified up to stratum-specific scalar sensitivity parameters. The time-to-event outcome was defined as a restricted survival time to avoid issues pertaining to measures conditioning on prior survival. To overcome challenges with a high-dimensional mediator, we treated the longitudinal mediator as a functional entity and employed a joint modelling framework to control for possible unobserved confounding between the mediator and the outcome. The methodology was applied to decompose the effect of a lifestyle intervention on restricted T2D-free time into an IE through weight reduction and a DE involving other mechanisms. We found some evidence suggesting the existence of a clinically significant IE through weight reduction, however, the magnitude of the estimated IE depended considerably on the assumed effective form of the mediator. When using the weight change over the first three years as the mediator, the IE accounted for a large fraction of the TE. Conversely, when considering the current weight change since baseline as the mediator, the DE was substantially larger than the indirect one. The results remained similar in sensitivity analyses relaxing the monotonicity, indicating robustness to violations of this assumption.

ORCID iDs

Mikko Valtanen  <https://orcid.org/0009-0008-4132-7065>

Tommi Härkänen  <https://orcid.org/0000-0002-4577-1808>

Matti Uusitupa  <https://orcid.org/0000-0001-6052-2001>

Jaakko Tuomilehto  <https://orcid.org/0000-0003-1909-7407>

Jaana Lindström  <https://orcid.org/0000-0001-9255-020X>

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: MV was financially supported by the EXACTUS programme in the University of Turku Graduate School (UTUGS).

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: JT owns shares in Orion Pharma, Oriola, Aktivolabs LTD and Digostics LTD. The other authors declare no conflict of interests.

Data availability

Due to privacy laws, the data cannot be made publicly available. The code used for the data analysis is available at GitHub (<https://github.com/mv2273/R-paper-medinint>).

Supplemental materials

Supplemental materials are available online.

References

1. Tuomilehto J, Uusitupa M, Gregg EW, et al. Type 2 diabetes prevention programs—from proof-of-concept trials to national intervention and beyond. *J Clin Med* 2023; **12**: 1876.
2. Uusitupa M, Khan TA, Viguiouk E, et al. Prevention of type 2 diabetes by lifestyle changes: a systematic review and meta-analysis. *Nutrients* 2019; **11**: 2611.
3. Barry E, Roberts S, Oke J, et al. Efficacy and effectiveness of screen and treat policies in prevention of type 2 diabetes: systematic review and meta-analysis of screening tests and interventions. *BMJ* 2017; **356**: i6538.
4. Haw JS, Galaviz KI, Straus AN, et al. Long-term sustainability of diabetes prevention approaches: a systematic review and meta-analysis of randomized clinical trials. *JAMA Intern Med* 2017; **177**: 1808–1817.
5. Kivipelto M, Solomon A, Ahtiluoto S, et al. The Finnish geriatric intervention study to prevent cognitive impairment and disability (FINGER): study design and progress. *Alzheimers Dement* 2013; **9**: 657–665.

6. Knowler WC, Barrett-Connor E, Fowler SE, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002; **346**: 393–403.
7. Gunderson EP, Lewis CE, Tsai AL, et al. A 20-year prospective study of childbearing and incidence of diabetes in young women, controlling for glycemia before conception: the Coronary Artery Risk Development in Young Adults (CARDIA) study. *Diabetes* 2007; **56**: 2990–2996.
8. Buse JB, Bain SC, Mann JFE, et al. Cardiovascular risk reduction with liraglutide: an exploratory mediation analysis of the LEADER trial. *Diabetes Care* 2020; **43**: 1546–1552.
9. Mann JFE, Buse JB, Idorn T, et al. Potential kidney protection with liraglutide and semaglutide: exploratory mediation analysis. *Diabetes Obes Metab* 2021; **23**: 2058–2066.
10. Robins JM and Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992; **3**: 143–155.
11. Pearl J. Direct indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. San Francisco, CA: Morgan Kaufmann, 2001, pp. 411–420.
12. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974; **66**: 688–701.
13. Pearl J. Interpretation and identification of causal mediation. *Psychol Methods* 2014; **19**: 459–481.
14. Avin C, Shpitser I and Pearl J. Identifiability of path-specific effects. In: *Proceedings of the International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, 2005*, pp.357–363.
15. Tchetgen Tchetgen EJ and Vanderweele TJ. Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology* 2014; **25**: 282–291.
16. Robins JM and Richardson TS. Alternative graphical causal models and the identification of direct effects. In: Shrouf P (ed) *Causality and psychopathology: finding the determinants of disorders and their cures*. Oxford, England: Oxford University Press, 2010, pp. 103–158.
17. Miles C, Kanki P, Meloni S, et al. On partial identification of the natural indirect effect. *J Causal Inference* 2017; **5**: 20160004.
18. Lin SH, Young JG, Logan R, et al. Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. *Stat Med* 2017; **36**: 4153–4166.
19. VanderWeele TJ and Tchetgen Tchetgen EJ. Mediation analysis with time varying exposures and mediators. *J R Stat Soc Series B Stat Methodol* 2017; **79**: 917–938.
20. Zheng W and van der Laan M. Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *J Causal Inference* 2017; **5**(2): 20160006.
21. Lindquist MA. Functional causal mediation analysis with an application to brain connectivity. *J Am Stat Assoc* 2012; **107**: 1297–1309.
22. Coffman DL, Dziak JJ, Litson K, et al. A causal approach to functional mediation analysis with application to a smoking cessation intervention. *Multivariate Behav Res* 2023; **8**: 859–876.
23. Zeng S, Lange EC, Archie EA, et al. A causal mediation model for longitudinal mediators and survival outcomes with an application to animal behavior. *J Agric Biol Environ Stat* 2023; **28**: 197–218.
24. Zheng C and Liu L. Quantifying direct and indirect effect for longitudinal mediator and survival outcome using joint modeling approach. *Biometrics* 2022; **78**: 1233–1243.
25. Didelez V. Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime Data Anal* 2019; **25**: 593–610.
26. Royston P and Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol* 2013; **13**: 152.
27. Lindström J, Louheranta A, Mannelin M, et al. The Finnish Diabetes Prevention Study (DPS): lifestyle intervention and 3-year results on diet and physical activity. *Diabetes Care* 2003; **26**: 3230–3236.
28. Eriksson J, Lindström J, Valle T, et al. prevention of type II diabetes in subjects with impaired glucose tolerance: the diabetes Prevention study (DPS) in finland. Study design and 1-year interim report on the feasibility of the lifestyle intervention programme. *Diabetologia* 1999; **42**: 793–801.
29. Tuomilehto J, Lindström J, Eriksson JG, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med* 2001; **344**: 1343–1350.
30. World Health Organization. *Diabetes mellitus: report of a WHO study group*. Technical Report Series No 727 1985. Geneva: World Health Organization.
31. Lindström J, Peltonen M, Eriksson JG, et al. Improved lifestyle and decreased diabetes risk over 13 years: long-term follow-up of the randomised Finnish diabetes prevention study (DPS). *Diabetologia* 2013; **56**: 284–293.
32. Laaksonen DE, Lindström J, Lakka TA, et al. Physical activity in the prevention of type 2 diabetes: the Finnish diabetes prevention study. *Diabetes* 2005; **54**: 158–165.
33. Lindström J, Ilanne-Parikka P, Peltonen M, et al. Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: the follow-up results of the Finnish diabetes prevention study. *Lancet* 2006; **368**: 1673–1679.
34. Wilding JP. The importance of weight management in type 2 diabetes mellitus. *Int J Clin Pract* 2014; **68**: 682–691.

35. Li G, Zhang P, Wang J, et al. The long-term effect of lifestyle interventions to prevent diabetes in the China da qing diabetes prevention study: a 20-year follow-up study. *Lancet* 2008; **371**: 1783–1789.
36. Diabetes Prevention Program Research Group, Knowler WC, Fowler SE, et al. 10-year follow-up of diabetes incidence and weight loss in the diabetes prevention program outcomes study. *Lancet* 2009; **374**: 1677–1686.
37. VanderWeele TJ. *Explanation in causal inference: methods for mediation and interaction*. New York: Oxford University Press, 2015.
38. VanderWeele TJ. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology* 2013; **24**: 224–232.
39. VanderWeele TJ. A unification of mediation and interaction: a 4-way decomposition [published correction appears in *Epidemiology* 2016 Sep;27(5):e36. doi: 10.1097/EDE.0000000000000527]. *Epidemiology* 2014; **25**: 749–
40. Aalen OO, Cook RJ and Røysland K. Does cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Anal* 2015; **21**: 579–593.
41. Vehtari A, Gelman A and Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput* 2015; **27**: 1413–1432.
42. Gould AL, Boye ME, Crowther MJ, et al. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Stat Med* 2015; **34**: 2181–2195.
43. Ibrahim JG, Chu H and Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. *J Clin Oncol* 2010; **28**: 2796–2801.
44. Hickey GL, Philipson P, Jorgensen A, et al. Joint models of longitudinal and time-to-event data with more than one event time outcome: a review. *Int J Biostat* 2018; **14**: 20170047.
45. Papageorgiou G, Mauff K, Tomer A, et al. An overview of joint modeling of time-to-event and longitudinal outcomes. *Annu Rev Stat Appl* 2019; **6**: 223–240.
46. Gelman A and Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci* 1992; **7**: 457–472.
47. Stan Development Team. RStan: the R interface to Stan. R package version 2.32.6, <https://mc-stan.org/>.
48. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>.
49. Therneau T. A package for survival analysis in R. R package version 3.8-3, <https://CRAN.R-project.org/package=survival>.
50. Smyth GK. Numerical integration. In: Armitage P and Colton T (eds.) *Encyclopedia of biostatistics*. London: Wiley, 1998, pp. 3088–3095.
51. Wang W and Yan J. splines2: Regression spline functions and classes. R package version 0.5.4, <https://CRAN.R-project.org/package=splines2>.
52. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *J Open Source Softw* 2019; **4**: 1686.
53. Vanderweele TJ, Vansteelandt S and Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology* 2014; **25**: 300–306.
54. Nguyen TQ, Schmid I, Ogburn EL, et al. Clarifying causal mediation analysis: effect identification via three assumptions and five potential outcomes. *J Causal Inference* 2022; **10**: 246–279.
55. Miles CH. On the causal interpretation of randomised interventional indirect effects. *J R Stat Soc Ser B: Stat Methodol* 2023; **85**: 1154–1172.
56. Mauff K, Steyerberg EW, Nijpels G, et al. Extension of the association structure in joint models to include weighted cumulative effects. *Stat Med* 2017; **36**: 3746–3759.

Appendix A: Identification of causal effects

The assumed causal DAG (Figure 1 in the main text) implies the following conditional independencies:

1. $T_{a,l,m} \perp \{A, L, M\} \mid W, R_0$
2. $M_{a,l} \perp \{A, L\} \mid W$
3. $L_a \perp A \mid W$
4. $M_{a,l} \perp \{L_a, L_{a*}\} \mid W$
5. $T_{a,l,m} \perp \{L_a, L_{a*}, M_{a*,l}\} \mid W, R_0$.

In addition, we make the consistency assumption stating that the observed outcome for an individual with given treatment, mediator and treatment-dependent confounder is equivalent to the potential outcome we would have observed, had the individual been assigned those values for the treatment, mediator and intermediate confounder. This means, for example,

that $T_{a,m,l}^{(i)} = (T^{(i)}|A_i = a, M_i = m, L_i = l)$. Missing data (i.e, right censorings of the survival process and drop-outs in the mediator process) are assumed missing at random.

In what follows, we first derive the expressions for the expectation of the nested counterfactual $\mathbb{E}(T_{a,M_{a^*}})$ in terms of the observed data and the unidentified joint probability of $(L_{a^*}, L_{a'})$. Then, we proceed to show how applying the additive decomposition leads to expressions which may then be arranged into the terms $DE^{(r)}$, $IE^{(r)}$, Δ_{DE} , Δ_{IE} and δ (see equation (2) in the main text). To make the notations more succinct, we hereafter omit the baseline confounders \mathbf{W} as well as the latent confounder R_0 .

Invoking the assumptions 1–5 along with the consistency assumption (denoted by ‘c’) and using factorisation of probabilities (denoted by ‘f’), the expression for $\mathbb{E}(T_{a,M_{a^*}})$ can be obtained as follows:

$$\begin{aligned}
\mathbb{E}(T_{a,M_{a^*}}) &= \mathbb{E}(T_{a,L_a,M_{a^*},L_{a^*}}) \\
&\stackrel{(f)}{=} \iiint_{t,l,m} tP(T_{a,L_a,M_{a^*},L_{a^*}} = t|M_{a^*},L_{a^*} = m, L_a = l) \\
&\quad \times P(M_{a^*},L_{a^*} = m, L_a = l) dt dl dm \\
&\stackrel{(c)}{=} \iiint_{t,l,m} tP(T_{a,l,m} = t|M_{a^*},L_{a^*} = m, L_a = l) \\
&\quad \times P(M_{a^*},L_{a^*} = m, L_a = l) dt dl dm \\
&\stackrel{(f)}{=} \int \dots \int_{t,l,l',m} tP(T_{a,l,m} = t|M_{a^*},L_{a^*} = m, L_a = l, L_{a^*} = l') \\
&\quad \times P(M_{a^*},L_{a^*} = m, L_a = l|L_{a^*} = l')P(L_{a^*} = l') dt dl dl' dm \\
&\stackrel{(c)}{=} \int \dots \int_{t,l,l',m} tP(T_{a,l,m} = t|M_{a^*},l' = m, L_a = l, L_{a^*} = l') \\
&\quad \times P(M_{a^*},l' = m, L_a = l|L_{a^*} = l')P(L_{a^*} = l') dt dl dl' dm \\
&\stackrel{(5)}{=} \int \dots \int_{t,l,l',m} tP(T_{a,l,m} = t)P(M_{a^*},l' = m, L_a = l, L_{a^*} = l') dt dl dl' dm \\
&\stackrel{(4)}{=} \int \dots \int_{t,l,l',m} tP(T_{a,l,m} = t)P(M_{a^*},l' = m)P(L_a = l, L_{a^*} = l') dt dl dl' dm \\
&\stackrel{(1)}{=} \int \dots \int_{t,l,l',m} tP(T_{a,l,m} = t|A = a, L = l, M = m)P(M_{a^*},l' = m) \\
&\quad \times P(L_a = l, L_{a^*} = l') dt dl dl' dm \\
&\stackrel{(2)}{=} \int \dots \int_{t,l,l',m} tP(T_{a,l,m} = t|A = a, L = l, M = m) \\
&\quad \times P(M_{a^*},l' = m|A = a^*, L = l')P(L_a = l, L_{a^*} = l') dt dl dl' dm \\
&\stackrel{(c)}{=} \int \dots \int_{t,l,l',m} tP(T = t|A = a, L = l, M = m)P_M(m|A = a^*, L = l') \\
&\quad \times P(L_a = l, L_{a^*} = l') dt dl dl' dm \\
&= \iiint_{l,l',m} \mathbb{E}(T|A = a, L = l, M = m)P_M(m|A = a^*, L = l') \\
&\quad \times P(L_a = l, L_{a^*} = l') dl dl' dm.
\end{aligned}$$

When $a = a^* \equiv a'$, the expression becomes

$$\begin{aligned}\mathbb{E}(T_{a',M_{a'}}) &= \iint_{l,m} \mathbb{E}(T|A = a', L = l, M = m)P_M(m|A = a', L = l)P(L_{a'} = l) dl dm \\ &\stackrel{(f,c,3)}{=} \iint_{l,m} \mathbb{E}(T|A = a', L = l, M = m)P_M(m|A = a', L = l)P_L(l|a') dl dm,\end{aligned}$$

since we do not need to simultaneously involve the conflicting worlds where A was set to different values.

Following the derivation given in the supplement of Tchetgen Tchetgen and Vanderweele,¹⁵ by applying the additive decomposition for $\mathbb{E}(T|a, m, l)$ we obtain

$$\begin{aligned}\mathbb{E}(T_{a,M_{a^*}}) &= \iiint_{l,l',m} [\beta_m(a, m) + \beta_l(a, l) + \beta_{m,l}(a, m, l) + \bar{\beta}_a(a)]P_M(m|a^*, l') \\ &\quad \times P(L_a = l, L_{a^*} = l') dl dl' dm \\ &= \iiint_{l,l',m} \beta_m(a, m)P_M(m|a^*, l')P(L_a = l, L_{a^*} = l') dl dl' dm \\ &\quad + \iiint_{l,l',m} \beta_l(a, l)P_M(m|a^*, l')P(L_a = l, L_{a^*} = l') dl dl' dm \\ &\quad + \iiint_{l,l',m} \bar{\beta}_a(a)P_M(m|a^*, l')P(L_a = l, L_{a^*} = l') dl dl' dm \\ &\quad + \iiint_{l,l',m} \beta_{m,l}(a, m, l)P_M(m|a^*, l')P(L_a = l, L_{a^*} = l') dl dl' dm \\ &= \iint_{l',m} \beta_m(a, m)P_M(m|a^*, l')P_L(l|a^*) dl' dm \\ &\quad + \int_l \beta_l(a, l)P_L(l|a) dl + \bar{\beta}_a(a) \\ &\quad + \iiint_{l,l',m} \beta_{m,l}(a, m, l)P_M(m|a^*, l')P(L_a = l, L_{a^*} = l') dl dl' dm.\end{aligned}$$

Applying the above expressions for $\mathbb{E}(T_{a,M_a})$ and $\mathbb{E}(T_{a,M_{a^*}})$, the definitions of the direct and indirect effects (see equation (1) in the main text) lead to the following expressions:

$$\begin{aligned}DE &= DE^{(r)} - \Delta_{DE} + \delta, \\ IE &= IE^{(r)} + \Delta_{IE} - \delta, \\ DE^{(r)} &= \int_m [\beta_m(a, m) - \beta_m(a^*, m)]P_M(m|a^*) dm \\ &\quad + \int_l [\beta_l(a, l)P_L(l|a) - \beta_l(a^*, l)P_L(l|a^*)] dl \\ &\quad + [\bar{\beta}_a(a) - \bar{\beta}_a(a^*)], \\ IE^{(r)} &= \int_m \beta_m(a, m)[P_M(m|a) - P_M(m|a^*)] dm, \\ \Delta_{DE} &= \iint_{m,l'} \beta_{m,l}(a^*, m, l')P_M(m|a^*, l')P_L(l'|a^*) dm dl', \\ \Delta_{IE} &= \iint_{m,l} \beta_{m,l}(a, m, l)P_M(m|a, l)P_L(l|a) dm dl, \\ \delta &= \iiint_{m,l,l'} \beta_{m,l}(a, m, l)P_M(m|a^*, l')P(L_a = l, L_{a^*} = l') dm dl dl' .\end{aligned}$$

Estimates of the causal effects can be obtained by estimating the components comprising the empirical formula. In the presence of (non-informative) censoring, the correct estimation of the expectations of the restricted survival time requires

that the censoring is dealt with in the estimation step. Similarly, the drop-out in the longitudinal process is required to be at most missing at random in order to enable unbiased estimation of the mediator trajectories.

Appendix B: Ordinal treatment-dependent confounder with K values

Assume (P) is a monotonic probability matrix. Denote $(P)_{jk} = P(L_{a^*} = j, L_a = k | \mathbf{W})$ and its marginals as $\mu_k = (P)_{\cdot k}$ and $\phi_j = (P)_{j \cdot}$, $j, k = 1, \dots, K$. Conditionally on the marginals, there are $\frac{1}{2}(K-1)(K-2)$ free probabilities in (P) . There are $\frac{1}{2}(K-1)(K-2)$ pairs of indices (l, r) such that $1 < l \leq r < K$. For any such pair, define $\mu^L = \{\mu_1, \dots, \mu_{l-1}\}$, $\mu^C = \{\mu_l, \dots, \mu_r\}$, $\mu^R = \{\mu_{r+1}, \dots, \mu_K\}$, and define similarly ϕ^L , ϕ^C and ϕ^R . For each (l, r) it then holds that

$$p_{\min}(l, r) = \max\{0, 1 - \|\phi^L\| - \|\mu^R\|\} \leq \|\phi^C \cap \mu^C\| \leq \min\{\|\phi^C\|, \|\mu^C\|\} = p_{\max}(l, r), \quad (9)$$

where $\|\cdot\|$ denotes the sum of the elements (probabilities) contained within the set (see the derivation below). There are thus $\frac{1}{2}(K-1)(K-2)$ constraints for the same number of independent linear combinations of the free parameters.

Let \mathbb{P} denote the full set of probabilities in (P) . The monotonicity assumption implies that $\|\mathbb{P}\| - \|\phi^L \cup \mu^R\| - \|\phi^C \cap \mu^C\| = 0$. Since $\|\phi^L \cap \mu^R\| \geq 0$, the lower limit of (9) follows from that

$$1 - (\|\phi^L\| + \|\mu^R\|) = \|\mathbb{P}\| - (\|\phi^L \cup \mu^R\| + \|\phi^L \cap \mu^R\|) \leq \|\mathbb{P}\| - \|\phi^L \cup \mu^R\| = \|\phi^C \cap \mu^C\|, \quad (10)$$

where the equality holds if $\|\phi^L \cap \mu^R\| = 0$. The upper limit follows trivially from the fact that $\phi^C \cap \mu^C$ is a subset of both ϕ^C and μ^C and that each element in the sets must be non-negative. The equality for the upper bound holds if either $\|\phi^C \cap \mu^R\| = 0$ or $\|\phi^L \cap \mu^C\| = 0$. The lower and upper limits can be estimated from data, and bounds for the causal effect estimates can be obtained by minimising and maximising the effects with respect to the joint probability matrix (P) with restrictions (9). Notice that (9) is a necessary but not sufficient condition for monotonicity. Indeed, for any marginals (ϕ, μ) , one can always construct a non-monotonic joint probability matrix as their outer product.

Appendix C: Step monotonicity

In this section, we describe a special case of the monotonicity assumption referred to as step monotonicity. We show that under step monotonicity, the joint probability of the counterfactuals of the treatment-dependent counfounder is identified and its consistency with the observed data can be empirically assessed. Let L be an ordinal variable taking values $1, \dots, K$. Assuming step monotonicity assumes that treatment can, on the individual-level, either have no effect on L or bring it one level higher. The joint distribution of (L_a, L_{a^*}) can then be expressed as a $K \times K$ matrix P , where $(P)_{jk} = P(L_a = k, L_{a^*} = j) = 0$, if $k \notin \{j, j+1\}$.

Denote the row totals corresponding to the marginal probabilities $P(L_{a^*} = j)$, as ϕ_j , $j = 1, \dots, K$, and the column totals corresponding to the marginal probabilities $P(L_a = k)$ as μ_k , $k = 1, \dots, K$. Given the marginals, the probability parameters of the joint distribution are identified, as there are $2K-1$ probability parameters and $2K-1$ restrictions imposed by the marginals. The joint probability matrix with $N = 5$, for example, would correspond to the following 5×5 table

		L_a					
		1	2	3	4	5	
L_{a^*}	5	0	0	0	0	p_{55}	ϕ_5
	4	0	0	0	p_{44}	p_{45}	ϕ_4
	3	0	0	p_{33}	p_{34}	0	ϕ_3
	2	0	p_{22}	p_{23}	0	0	ϕ_2
	1	p_{11}	p_{12}	0	0	0	ϕ_1
		μ_1	μ_2	μ_3	μ_4	μ_5	1

from which it is straightforward to see that if the marginals are known, the elements can be obtained as

$$p_{ii} = 1 - \sum_{k=1}^{i-1} \phi_k - \sum_{h=i+1}^K \mu_h, \quad p_{i(i+1)} = 1 - \sum_{k=i+1}^K \phi_k - \sum_{h=1}^i \mu_h,$$

The data are consistent with the step monotonicity assumption if the elements p_{jk} form a proper joint distribution, that is, all the elements are between 0 and 1.