1 **Characterization of direct and/or indirect genetic associations for multiple traits in**
2 **longitudinal studies of disease progression**

3 Myriam Brossard[1*], Andrew D. Paterson[2,3], Osvaldo Espin-Garcia[3,4], Radu V. Craiu[5], Shelley B.
4 Bull[1,3,*]

5 [1]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada;

6 [2]Program in Genetics and Genome Biology, Hospital for Sick Children Research Institute,
7 Toronto, ON, Canada;

8 [3]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON,
9 Canada;

10 [4] Department of Biostatistics, Princess Margaret Cancer Centre, Toronto, ON, Canada;

11 [5]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada.

12 **Running title** (40/40 characters): Joint models for multiple trait genetics.

13 **Abstract (239/250)**

14 When quantitative longitudinal traits are risk factors for disease progression, endogenous, and/or
15 subject to random errors, joint model specification of multiple time-to-event and multiple
16 longitudinal traits can effectively identify direct and/or indirect genetic association of single
17 nucleotide polymorphisms (SNPs) with time-to-event traits. Here, we present a joint model that
18 integrates: *i)* a linear mixed model describing the trajectory of each longitudinal trait as a function
19 of time, SNP effects and subject-specific random effects, and *ii)* a frailty Cox survival model that
20 depends on SNPs, longitudinal trajectory effects, and a subject-specific frailty term accounting for
21 unexplained dependency between time-to-event traits. Inference is based on a two-stage
22 approach with bootstrap joint covariance estimation. We develop a hypothesis testing procedure
23 to identify direct and/or indirect SNP association with each time-to-event trait. Motivated by
24 complex genetic architecture of type 1 diabetes complications (T1DC) observed in the Diabetes
25 Control and Complications Trial (DCCT), we show by realistic simulation study that joint modelling
26 of two time-to-T1DC (retinopathy, nephropathy) and two longitudinal risk factors (HbA1c, systolic
27 blood pressure) reduces bias and improves identification of direct and/or indirect SNP
28 associations, compared to alternative methods ignoring measurement errors in intermediate risk
29 factors. Through analysis of DCCT, we identify two SNPs with indirect associations with multiple
30 time-to-T1DC traits and obtain similar conclusions using alternative formulations of time-
31 dependent HbA1c effects on T1DC. In total, joint analysis of multiple longitudinal and multiple
32 time-to-event traits provides insight into aetiology of complex traits.

33 **Key words:** joint models; longitudinal study; direct and/or indirect genetic association; pleiotropy;
34 complex genetic architecture; multiple-trait analysis; measurement error; patient's trajectory;
35 mixed model; frailty model.

36 ***Corresponding authors:** Lunenfeld-Tanenbaum Research Institute, 60 Murray Street, Box #18,
37 M5T 3L9, Toronto, ON, Canada. E-mails: bull@lunenfeld.ca, brossard@lunenfeld.ca.
38

## Introduction

Despite their known ability to improve inference in clinical and epidemiological studies, particularly in the presence of informative censoring/dropout or when longitudinal traits are measured with error[1–3], joint models for longitudinal and time-to-event outcomes have received limited attention in genetic association studies. Genome-wide association studies (GWAS) of intermediate quantitative traits (QTs) typically require follow-up studies to identify whether SNP associations detected with each of the QT(s), analyzed separately, also affect related outcomes through direct and/or indirect effects induced by those traits. Such intermediate traits may include established risk factors for related clinical outcomes and be measured with errors (e.g. biological variation). By accounting for QT measurement error and dependencies among traits, joint models can improve accuracy and efficiency of effect estimation, as well as detection of SNP associations.

The so-called shared-random-effects joint model[4,5] consists of a sub-model for a single longitudinal trait and a sub-model for a single right-censored time-to-event trait. The longitudinal sub-model describes the QT as an underlying smoothed trajectory that depends on fixed effects of time and baseline covariates, as well as subject-specific random effects. The joint model association structure is induced via the functional dependence between the hazard of an event at any time $t$ and the longitudinal trait trajectory[6,7]. This relationship can be based on prior biological knowledge of the link between that longitudinal and time-to-event traits. As elucidated by Ibrahim et al[1], this class of joint models lends itself to interpretations of direct/indirect effects because the relationship between a baseline covariate, such as a SNP genotype, and each of the longitudinal and time-to-event traits, as well as the relationship between the longitudinal and time-to-event traits can be specified via model parameters corresponding to direct, indirect and overall effects. Extensions of joint models for multiple longitudinal and/or multiple time-to-event traits can further improve inference by borrowing information among related traits. Extensions of joint models have been reviewed for multiple longitudinal traits[6,7] or multiple time-to-event traits[8]. A few extensions for both multiple longitudinal and multiple time-to-event traits have been developed ([9–11], among others), but these models are often formulated for a specific study question, and thus can lack generalizability. Such extensions raise computational challenges for maximisation of the likelihood under the multivariate random effect's distribution. Two-stage approaches[12–15] for joint model fitting appear more computationally efficient and lends itself to more flexible model formulation but inference can be mis-calibrated because parameter estimates and predictions from Stage 1 are obtained from the longitudinal model without consideration of the time-to-event outcome, and the uncertainty in Stage 1 estimates is ignored during Stage 2 estimation, a problem known as propagation of errors[16].

Motivated by the complex genetic architecture of long-term type 1 diabetes complications (T1DC), we propose joint-model extension to evaluate genetic associations with multiple time-to-event traits and multiple longitudinal risk factors. Risk of development of T1DC, including diabetic retinopathy (DR) and diabetic nephropathy (DN), is hypothesized to result from multiple genetic factors with potential direct and/or indirect effects via multiple - shared and/or specific - risk factors[17]. In addition to genetic factors, hyperglycemia (measured by Hemoglobin A1c, referred as HbA1c) represents the major QT risk factor for T1DC. Other measured longitudinal traits, that can also be influenced by genetic factors, have postulated associations with T1DC, for example,

81  association of systolic blood pressure (SBP) with DN. Because conventional analysis methods in
82  GWAS ignores dependency between related traits, SNPs can appear associated with multiple
83  traits (including SNPs with direct and/or indirect effects induced by intermediate QT(s) measured
84  or unmeasured). Accurately distinguishing between direct and/or indirect associations is crucial
85  to elucidate the biological pathways through which genetic factors operate into the etiology of
86  those complex traits, like in our motivating example of the T1DC genetic architecture. A naïve
87  approach to disentangle direct from indirect associations can be achieved by including the
88  intermediate QT as a time-dependent covariate in a Cox Proportional Hazard (PH) model.
89  However, measurement errors in intermediate QT(s), presence of informative censoring in
90  longitudinal trait, unmeasured shared risk factors between longitudinal/time-to-event traits can
91  challenge accurate classification of direct and/or indirect SNP associations.

92  The *major* contribution of our work is a general formulation for the joint analysis of multiple time-
93  to-event traits and multiple longitudinal QT risk factors in genetic association studies. We develop
94  inference methods for statistical genetic analysis based on joint model parameters estimation,
95  including a hypothesis testing procedure to classify direct and/or indirect SNP associations with
96  each time-to-event trait. A *second* contribution of the paper is the implementation of a data-
97  informed simulation algorithm to generate multiple causal SNPs with various direct effects on
98  *simulated* time-to-event traits and/or indirect effects via observed (*measured*) longitudinal QTs in
99  the Diabetes Control and Complications Trial (DCCT) study[18,19] and unobserved (*simulated*)
100  longitudinal QTs. This algorithm also provides a general approach to estimate achievable power
101  given sample size. Our simulation results show bias reduction and improved classification of direct
102  and/or indirect SNP associations in comparison with alternative approaches ignoring
103  measurement errors in longitudinal QT(s). *Thirdly*, we apply the joint model in DCCT data and
104  clarifies two SNPs as having indirect associations via the HbA1c longitudinal risk factor, with
105  consistent conclusions using alternative time-dependent association structures that account for
106  established cumulative and time-weighted effects of HbA1c on T1DC traits[20,21]. Example R codes
107  for data simulation and for application of the proposed joint model are available on GitHub.

## Methods

### Joint Modelling Approach

### *Model Formulation*

111  We assume that a set of *M* SNPs are available with *K* (*1 ≤ k ≤ K*) unordered and non-competing
112  time-to-event traits, such as multiple disease complications, and *L* (*1 ≤ l ≤ L*) longitudinal QT traits
113  (*i.e.* intermediate risk factors) possibly measured with error (*i.e.* biological variation) in *N* unrelated
114  individuals indexed by *i* (*1 ≤ i ≤ N*). For each SNP*m* (*1 ≤ m ≤ M*), with minor allele frequency (MAF)
115  $p_m$, the genotypes are coded as the number of copies of the minor allele, under an additive genetic
116  model. For each individual *i*, $Y_{i(l)} = (y_{i(l)}(t_{i1}), \dots y_{i(l)}(t_{ij}), \dots, y_{i(l)}(t_{in_{i(l)}}))'$ denotes the vector of
117  quantitative measures collected over scheduled visits at $t_{ij}$ for *j*=1,…, $n_i$ for the *l*th longitudinal trait
118  *(1≤l≤ L)* with $n_i$ represents the maximum number of visits recorded. Let $(T_{i(k)}, \delta_{i(k)})$ be the vector
119  of observed right-censored event time $T_{i(k)}$ and event indicator $\delta_{i(k)}$ for the *k*th time-to-event trait.

120    We assume $T_{i(k)} = \min(T^*_{i(k)}, C_i)$, where $T^*_{i(k)}$ is the latent (uncensored) time-to-event $k$ and $C_i$ is

121    the censoring time (e.g. administrative censoring). We define $\delta_{i(k)} = I(T^*_{i(k)} \leq C_i)$, with $\delta_{i(k)} = 1$ if

122    the event occurs during the observation period $(T^*_{i(k)} \leq C_i)$ and $\delta_{i(k)} = 0$ otherwise.

123    To characterize the genetic architecture of a system of *multiple* time-to-events and *multiple*

124    longitudinal risk factors, we formulate a shared-random-effects joint model, as specified in **Figure**

125    **1**, combining: ***(i) longitudinal*** and ***(ii) time-to-event sub-models*** connected by ***(iii) specified***

126    ***time-dependent association structures***.

127    [Insertion of Figure 1]

128    ### *(i)*    **Longitudinal sub-model**

129    The *longitudinal sub-model* is specified by a *multivariate mixed model* for the *L* longitudinal QTs

130    (or risk factors), based on the multivariate extension of the Laird and Ware linear mixed model[22].

131    It describes each vector of *observed* longitudinal QT measures $Y_{i(l)}(t_{ij})$ as noisy observations of

132    a *true* and *unobserved smoothed* subject-specific longitudinal trajectory, $Y^*_{i(l)}(t_{ij})$, with noise

133    terms $\varepsilon_{ij(l)} \sim N(0, \sigma^2_{(l)})$. Each *smooth* trajectory describes subject-specific evolution and depends

134    on time, SNP effect, individual-level random effects $b_{i(l)}$ and can also include effects of other

135    baseline covariates, for example confounding factors or ancestry-related principal components.

136    To simplify the presentation, we assume a linear trajectory, but the longitudinal sub-models can

137    be adapted for nonlinear trajectories using, for example, higher order polynomials or splines[23].

138    *Multivariate mixed model*

139    $Y_{i(l)}(t_{ij}) = Y^*_{i(l)}(t_{ij}) + \varepsilon_{ij(l)}$    (*observed* trait values, Equation 1)

140    $Y^*_{i(l)}(t_{ij}) = \beta_{0(l)} + b_{i0(l)} + (\beta_{1(l)} + b_{i1(l)})t_{ij} + \beta_{g(l)}SNP_i + \beta_{h(l)}H_{i(l)}$

141    (*smooth* trajectory, Equation 2)

142    Where:

143    •  $\beta_{0(l)}, \beta_{1(l)}, \beta_{g(l)}$ and $\beta_{h(l)}$ denote, respectively, the fixed intercept, and baseline effects of

144       time, SNP and other covariate(s) on each longitudinal trait *l*.

145    •  $b_i = ((b_{i0(l)}, b_{i1(l)})', \ldots, (b_{i0(L)}, b_{i1(L)})')'$ is the 2x*L* vector of random effects with

146       $b_i \sim N_{2L}(0, D)$.

147    •  $D = \begin{pmatrix} D_{1,1} & \cdots & D_{1,L} \\ \vdots & \ddots & \vdots \\ D_{L,1} & \cdots & D_{L,L} \end{pmatrix}$ denotes the *overall* variance-covariance matrix for random effects,

148       accounting for serial dependencies *within* longitudinal traits ($D_{l,l}$, unstructured) and cross-

149       dependencies *between* longitudinal traits ($D_{l,m}$ for two traits *l≠m*).

150     •   $\varepsilon_{i(l)} = (\varepsilon_{i1(l)}, \dots, \varepsilon_{ij(l)}, \dots, \varepsilon_{in_i(l)})'$ is the vector of error terms where $\varepsilon_{ij(l)} \sim N(0, \sigma^2_{(l)})$ is
151        assumed independent and identically distributed.
152     •   $\sigma^2_{(l)}$ is the residual variance of the QT *l*. We further assume that $\varepsilon_{ij(l)}$ and $b_{i(l)}$ are
153        independent[22].

154    ***(ii)***    **Time-to-event sub-model**

155 The *time-to-event sub-model* is specified by a Proportional Hazards (PH) mixed effects model
156 (also known as frailty PH model) where the hazard function of each time-to-event trait *k* depends
157 on the SNP effect adjusted for a function of the subject's longitudinal trajectories $W_{i(k)}(t)$. We
158 introduce a subject-specific random effect (frailty term, $u_i$) to capture unexplained dependencies
159 (e.g. due to unmeasured shared factors) among the *K* time-to-event traits. Given the common
160 frailty term $u_i$, the *K* time-to-event traits are assumed independent[24].

161 *Proportional Hazards mixed effects sub-model*

162    $\lambda_{i(k)}(t) = \lambda_{0(k)}(t) \times exp\{\gamma_{g(k)}SNP_i + W_{i(k)}(t) + \gamma_{v(k)}V_{i(k)} + u_i\}$      (Equation 3)

163 Where:

164     •   $\lambda_{0(k)}(t)$ is a (parametric or non-parametric) baseline hazard function for the time-to-event
165        trait *k*.
166     •   $\gamma_{g(k)}$ denotes the SNP effect on the time-to-event *k* adjusted for the indirect SNP effects
167        via longitudinal QT(s), taken into account by $W_{i(k)}(t)$.
168     •   $\gamma_{v(k)}$ is the effect of the baseline covariate vector(s). Those covariates can be shared or
169        trait-specific covariates. They can also be time-dependent covariates for all or a subset of
170        the *K* time-to-event traits (in this case, $V_{i(k)} = V_{i(k)}(t)$).
171     •   $u_i$ and $b_i$, random effects from the time-to-event and longitudinal sub-models respectively,
172        are assumed independent. We assume $u_i \sim \Gamma(a, b)$, with *a*, *b*>0.

173    ***(iii)***    **Time-dependent association structures**

174 The *time-dependent association structures*, $W_{i(k)}(t)$, that connect the *longitudinal* and *time-to-*
175 *event sub-models* (**Figure 1**), account for the specific temporal relationships between *each* set of
176 $L_k$ $(1 \leq L_k \leq L)$ associated QT(s) and *each* time-to-event trait *k*.

177    $W_{i(k)}(t) = \sum_{l=1}^{L_k} \alpha_{l(k)} f_{l(k)}(Y^*_{i(l)}(t))$                  (Equation 4)

178 Where:

179     •   $1 \leq L_k \leq L$ is the set of longitudinal risk factors of the time-to-event trait *k*.
180     •   $f_{l(k)}(Y^*_{i(l)}(t))$ denotes the *functional* form of the exposure effect on time-to-event trait *k*
181        corresponding to the *l*-th longitudinal risk. In the case of a *contemporaneous*
182        parametrization, the hazard of an event *k* at a time *t* depends on the longitudinal trait value

183    at the same time $t$ (i.e. $f_{l(k)}\left(Y^*_{i(l)}(t)\right) = Y^*_{i(l)}(t)$). Other parametrizations have been

184    described in the literature[6,7,25].

185    •  $\alpha_{l(k)}$ denotes the association parameter of the function of the *smooth* longitudinal

186    trajectory $Y^*_{i(l)}(t)$ with the time-to-event $k$.

187    •  $Y^*_{i(l)}(t) = \{Y^*_{i(l)}(s), with\ 0 \le s \le t, 1 \le l \le L\}$ denotes the history of the underlying

188    longitudinal trajectory for trait *l* up to time *t*.


189    *Interpretation*


190    When the joint model is *correctly* specified, the parameters accurately represent multiple

191    relationships between each SNP and the correlated traits. Note that $\beta_{g(l)}$ is the SNP effect on the

192    longitudinal risk factor *l*, $\gamma_{g(k)}$ is the *direct* SNP effect on the time-to-event trait *k* via other

193    mechanisms than the indirect SNP association induced by the observed longitudinal risk factors.

194    In this context, $\theta_{k(l)} = \mu_{g(k,l)} + \gamma_{g(k)}$ is interpreted as the *overall* SNP effect on the time-to-event

195    trait *k*, where $\mu_{g(k,l)} = \alpha_{l(k)}\beta_{g(l)}$ represents the *indirect*[1] SNP effect on the time-to-event trait *k* via

196    the longitudinal QT *l*.


197    Under the proposed joint model formulation, when a longitudinal risk factor *l* is *associated* with a

198    time-to-event trait *k* ($\alpha_{l(k)} \ne 0$), a SNP that has an effect on the longitudinal QT risk factor *l* ($\beta_{g(l)} \ne$

199    $0$), but no effect on the time-to-event trait *k* ($\gamma_{g(k)} = 0$), is interpreted as having an **indirect** SNP

200    **association** and its *overall* effect is equal to its *indirect* effect ($\theta_{k(l)} = \mu_{g(k,l)}$, with $\mu_{g(k,l)} =$

201    $\alpha_{l(k)}\beta_{g(l)}$). A SNP with an effect on the time-to-event trait *k* ($\gamma_{g(k)} \ne 0$), but no effect on the

202    longitudinal QT *l* ($\beta_{g(l)} = 0$), is interpreted as having a **direct** SNP association and its *overall*

203    effect is equal to its *direct* effect ($\theta_{k(l)} = \gamma_{g(k)}$). A SNP with an effect on the longitudinal risk factor

204    *l* ($\beta_{g(l)} \ne 0$) and an effect on the time-to-event trait *k* ($\gamma_{g(k)} \ne 0$) is interpreted as having both

205    **direct** and **indirect** **SNP associations** via distinct genetic pathways. In this case, its *overall* effect

206    is an aggregation of both *direct* and *indirect* effects ($\theta_{k(l)} = \mu_{g(k,l)} + \gamma_{g(k)}$, with $\mu_{g(k,l)} = \alpha_{l(k)}\beta_{g(l)}$).

207    It is obvious that when an associated longitudinal risk factor is omitted from a time-to-event model,

208    as occurs in separate analyses of longitudinal and time-to-event traits, the marginal SNP effect

209    on the time-to-event trait will reflect partially the overall SNP effect $\theta_{k(l)}$. This is particularly the

210    case in GWAS when intermediate risk factors are ignored in the analysis models.


211    *Effect estimation and test statistic construction*


212    To mitigate the computational complexity involved in the maximization of the joint likelihood and

213    allow more flexibility in the model formulated, we estimate the parameters using a two-stage

214    approach (see **Appendix**). We work within the framework defined by Tsiatis and Davidian[26,27]

215    which guarantees that our estimators are consistent and asymptotically normal. Specifically, in

216    Stage 1 we fit the multivariate mixed model (Equation 1) using the *mvlme*() function from the R

217    package JoineRML[28] to estimate the parameters of the longitudinal trajectories of the risk factors

218    (Equation 2), and obtain fitted values of the smoothed trajectories. In Stage 2, we fit a Cox PH

219    frailty time-to-event model (Equation 3) adjusting for functions of the smoothed trajectories as

220      time-dependent covariates using the *coxph*() function from the R survival[29,30] package. We
221      assume a Gamma distribution for the frailty term ($u_i$) and a separate baseline hazard function for
222      each time-to-event trait using the *strata* argument in *coxph*(). To account for propagation of errors,
223      due to the uncertainty in Stage 1 estimates ignored during Stage 2 parameter estimation[16] and
224      empirically estimate the joint covariance matrix of SNP and trajectories effects we apply a
225      nonparametric bootstrap[31]. For each bootstrap sample *b* ($1 \leq b \leq B$, *B* is the total number of
226      bootstraps), we generate a new dataset by randomly sampling *N* individuals with replacement
227      and refitting the joint model on each new dataset *b*. We compute the empirical joint covariance
228      matrix of the estimated parameters using the *B* bootstrap parameter estimates. Wald statistics for
229      each $\beta_{g(l)}$ are computed as $S_{\beta_{g(l)}} = \widehat{\beta_{g(l)}}/se_{\beta_{g(l)}}$ using the empirical bootstrap standard errors
230      $se_{\beta_{g(l)}}$, and similarly for each $\gamma_{g(k)}$ as $S_{\gamma_{g(k)}} = \widehat{\gamma_{g(k)}}/se_{\gamma_{g(k)}}$.

231      *Procedure to classify direct/indirect SNP associations*

232      In **Table 1**, we present a practical hypothesis testing procedure to classify SNPs as having direct
233      and/or indirect association with a time-to-event trait *k* and each associated longitudinal risk factor
234      *l*. This procedure requires two significance thresholds, $P^*_{\beta_g}$ and $P^*_{\gamma_g}$ for $\beta_{g(l)}$ and $\gamma_{g(k)}$ respectively,
235      to be specified prior to the analysis and adjusted for the number of SNPs tested. Depending on
236      the research question, we can choose different values for $P^*_\beta$ and $P^*_\gamma$, or the same one ($P^* = P^*_{\beta_g} =$
237      $P^*_{\gamma_g}$). The latter is applicable for instance, when we want to systematically classify direct/indirect
238      association among a set of *M* SNPs, and the former when assessing which SNPs, among the
239      ones reported to be associated with the longitudinal risk factor, have a direct effect on a time-to-
240      event trait.

241      [Insertion of Table 1]

242      **Motivation: The DCCT Studies**

243      The DCCT randomized-controlled trial was *pivotal* in demonstrating that intensive insulin therapy
244      prevents and delays progression of long-term T1DC[18]. Due to significant outcome differences
245      between the intensive and conventional treatment groups at interim analysis, the DCCT trial was
246      stopped early. At the closeout visit, patients were administratively censored and had a mean
247      follow-up time of 6.5 years (range 3 to 9), 99% had completed the study, and more than 95% of
248      all scheduled examinations were completed[18]. DCCT participants continued to be followed for
249      disease progression after the trial in a long-term epidemiologic cohort study and were genotyped
250      for GWAS in the DCCT Genetics Study. Because the goal of intensive therapy was to reduce
251      HbA1c into the non-diabetic range, which produced treatment differences in HbA1c values, we
252      focus on *N*=667 unrelated individuals of European descent ancestry from the Conventional
253      treatment group. Longitudinal measurements for HbA1c and SBP were collected at up to 39
254      quarterly visits during DCCT, while DR and DN events were diagnosed at annual and semi-annual
255      visits respectively. HbA1c and SBP were recorded irrespective of the occurrence of any
256      complication event(s). Although the DCCT implemented robust quality assurance procedures to
257      minimize potential sources of error during and after data collection[32,33], HbA1c and SBP are

258  subject to measurement error (e.g. individual variability). The first GWAS to analyze the DCCT
259  study phenotypes[19] identified two SNPs associated with cumulative HbA1c at genome-wide
260  significance in the Conventional treatment arm, namely rs10810632 (in *BNC2*, 9p22.2) and
261  rs1358030 (near *SORCS1*, 10q25.1), and reported consistent associations with secondary
262  outcomes of time-to-DR and/or time-to-DN. Subsequent association studies also reported
263  suggestive evidence for pleiotropy between DR and DN[34]. To preserve the inherent within-
264  individual variability, we designed simulation study evaluations of the joint modelling approach
265  that generates event times from HbA1c and SBP measurements in the DCCT Study data.

## Simulation study

266

*Design of the DCCT-data-based simulation study*

267

268  To assess the performances of the methods, we simulate the data under a *genetic causal*
269  *scenario* that imitates the complex genetic architecture of T1DC (**Figure 2**). The genetic
270  association model involves *N*=667 DCCT subjects with *K*=2 non-independent *simulated* time-to-
271  T1DC traits (DR, DN) that depend on *M*=5 causal SNPs with *direct* effects on each time-to-T1DC
272  and/or *indirect* effects via *L*=3 longitudinal QTs: two as *measured* in DCCT (HbA1c, SBP) and
273  one other *simulated* (*U*) to induce some unexplained shared dependency between the T1DC
274  traits. We also introduce an effect of sex on SBP, and effects of T1D duration (at baseline) on
275  both T1DC traits, as observed in the original DCCT data.

276  [Insertion of Figure 2]

*Algorithm for realistic data generation under a complex genetic architecture*

277

278  To generate such a data structure that combines *observed* and *simulated* traits, we formulate a
279  data generating model including: *(i)* *L*=3 linear mixed models linking each of the SNPs with an
280  indirect effect to each longitudinal risk factor, and *(ii)* *K*=2 non-independent parametric time-to-
281  event models depending on SNPs with *direct* effects and on functions of the longitudinal QT
282  trajectories. For each DCCT individual *i* with observed HbA1c ($Y_{i(1)}$), SBP ($Y_{i(2)}$), visit times ($t_i$)
283  vectors at $n_i$ time points, and covariates ($H_{i(l)}, V_{i(k)}$), we simulate longitudinal trait vector $U_i$, time-
284  to-event traits and genetic data at *M* causal SNPs as illustrated in **Figure 3** and detailed in
285  **Supplementary Information SI-1**. All SNPs are simulated under Hardy-Weinberg and linkage
286  equilibrium assumptions. Particularly, SNPs with indirect effects are simulated from the *observed*
287  (SNP1, SNP5) or *simulated* (SNP3) longitudinal QTs, while SNPs with direct effects (SNP2,
288  SNP4) on time-to-T1DC traits are simulated independently of the longitudinal QTs and are
289  included in the specified hazard function used to simulate each time-to-event trait *k* (**Figure 3**).

290  [Insertion of Figure 3]

291  For each SNP*m*, we specify MAF ($p_m$) and the SNP effects ($\beta_{g(l)}, \gamma_{g(k)}$), as well as the other
292  parameter values according to the DCCT Genetics Study and the T1DC literature (**Table 2**,

293    **Tables S1**). We assume contemporaneous association structures for HbA1c and SBP effects on
294    simulated time-to-T1DC traits. We simulate each time-to-event trait $k$ under a Weibull model, with
295    shape and scale parameters specified to generate ~54% DR events and ~25% DN events in each
296    of the $R$=1000 replicated datasets generated under the *causal* scenario from **Figure 2**. Under the
297    *global null* genetic scenario, where none of the SNPs is associated with any traits, we simulate $M$
298    SNPs independently of the traits with the same MAF as for the causal SNPs.

299    *Scenarios for DCCT-based complex genetic architecture*

300    Overall, the simulated complex genetic architecture covers multiple types of SNP-trait
301    associations (**Figure 2** and **Table 2**): *direct* association with each T1DC trait (SNP2, SNP4),
302    *indirect* associations with both T1DC traits via *measured* (SNP1) and *unmeasured* (SNP3)
303    longitudinal QTs; and *direct and indirect* associations via a *measured* longitudinal QT (SNP5); all
304    longitudinal risk factors are subject to measurement error. Except for SNP3, all other SNP
305    scenarios represent GWAS discovery SNPs detectable by separate GWAS of a longitudinal risk
306    factor (SNP1, SNP5) or a time-to-event trait (SNP2, SNP4, SNP5).

307    [Insertion of Table 2]

308    SNP1, SNP3 and SNP5 have indirect effects on T1DC traits, such that their associations with the
309    T1DC traits are detectable with the marginal Cox PH time-to-event models (**Tables 2** and **S2**).
310    SNP1 corresponds roughly to rs10810632 and rs1358030 reported in the motivating DCCT
311    GWAS of HbA1c[19], while SNP5 represents a top hit detected by each GWAS (longitudinal and
312    time-to-event traits), analyzed separately. Because of their association with longitudinal and time-
313    to-event traits, SNP1 and SNP5 associations with time-to-event outcomes are conventionally
314    investigated using the Cox PH model adjusting for the *observed* longitudinal trait as a time-
315    dependent risk factor to determine if SNP association with the time-to-event trait is fully (or
316    partially) explained via the associated longitudinal QT(s).

317    **Analysis of the simulated data**

318    For the analysis of *each* SNP, done *separately*, we compare four analysis models on the
319    simulated data:

320    •   JM-cmp: a *completely* specified joint model that includes all other non-SNP variables used
321        for the data simulation.
322    •   JM-mis: includes the same variables as JM-cmp, except $U$, the unobserved longitudinal
323        QT.
324    •   JM-sep($k$): includes the same variables as JM-mis but is fitted separately for the *two* time-
325        to-event traits. JM-sep($k$) does not account for the shared dependency between the time-
326        to-event traits.
327    •   CM-obs: a Cox PH frailty model that includes the same variables as JM-mis, but adjusts
328        for the *observed* longitudinal QT values as time-dependent covariates.

329  Due to the latent nature of *U, JM-cmp* cannot be fitted in practice because *U* is unobserved, but
330  we include it as a benchmark for comparison against the *observable* models fitted without *U*. For
331  all these models, empirical covariance matrices are computed using 500 bootstrap iterations. We
332  assess the performance of the hypothesis testing procedure to classify direct/indirect association
333  for each of the 5 causal SNPs analyzed separately with each joint model (JM-cmp, JM-mis). To
334  assess the impact of longitudinal trait measurement errors on classification of SNPs, we substitute
335  test $P_{\gamma_{g(k)}}$ for $\gamma_{g(k)}$ estimated by the joint model by $P_{\gamma_{g(k)}}$ obtained with CM-obs. In our evaluations,
336  we assume the level of significance $P^*$ such that $P^* = P_\beta^* = P_\gamma^*$, with $P^*$ varying from 0.05 to 5x10$^-$
337  $^8$. Under the *causal* genetic scenario, we estimate for each SNP the proportion of replicates that
338  detect the correct direct and/or indirect association simulated (correct classification rate) and the
339  proportion of replicates where the causal SNP association is misclassified (misclassification rate).
340  Under the *global null* genetic scenario, we assess the classification test for each SNP and each
341  type of association at $P^* = 0.05$ and $P^* = 0.01$ to assess the type I error control of the hypothesis
342  testing procedure for each test of direct and/or indirect association.

343  **Availability**

344  DCCT data are available to authorized users at https://repository.niddk.nih.gov/studies/edic/
345  and https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000086.v3.p1.
346  Example R codes for DCCT-data-based simulation and analysis of the simulated data are
347  provided on GitHub (https://github.com/brossardMyriam/Joint-model-for-multiple-trait-genetics).
348  Supplemental files are available at FigShare. File S1 includes the Supplementary Information;
349  File S2 includes the Supplementary Figures and File S3 includes the Supplementary Tables.

350  **Results**

351  **Simulation results**

352  *Estimation accuracy and single-parameter tests validity*

353  When the longitudinal QT(s) are observed and measured with random *errors*, the simulation
354  results confirm that the proposed joint model reduces the bias of the parameter estimates for all
355  types of SNP association (**Figure S1** and **Tables S3-S6**), even when the analysis model is not
356  fully specified. This applies when: *(i)* two correlated time-to-event traits are analyzed *jointly* rather
357  than *separately* in the joint model (JM-cmp/JM-mis compared to JM-sep(*k*)); *(ii)* a SNP has a
358  *direct* effect on a time-to-event trait with a low event rate (SNP4/SNP5, JM-cmp/JM-mis compared
359  to CM-obs); *(iii)* a SNP has an *indirect* effect via a longitudinal QT measured with errors (for
360  example SNP5, JM-cmp/JM-mis compared to CM-obs). Particularly, for SNP5, CM-obs
361  overestimates the *direct* SNP5 effect on time-to-DN due to its inability to fully account for the
362  indirect SNP5 effect induced via SBP. Direct SNP2 and SNP4 effects on time-to-T1DC traits show
363  overall the larger relative bias reduction using joint models versus CM-obs (Ranges with JM-mis
364  0.702-0.983, with JM-cmp 0.934-0.966, see **Tables S4 and S6**).

365     When a SNP has an indirect effect on both T1DC traits fully explained via an *unmeasured*
366     longitudinal QT (SNP3), all methods that ignore the longitudinal QT (JM-mis, JM-sep($k$) or CM-
367     obs compared to JM-cmp) produce biased direct SNP3 estimates for both T1DC towards the
368     overall SNP3 effect ($\theta_{k(U)} = \alpha_{U(k)}\beta_{g(U)}$ =0.36, **Figure S1-c** and **S1-d, Table S5**). In this scenario,
369     the use of the frailty term in JM-mis does not reduce this bias (JM-mis *vs* JM-sep).

370     *Classification of direct/indirect SNP associations with time-to-event traits*

371     Under the *global null* genetic scenario of no genetic association with any of the traits, single-
372     parameter SNP test *P*-values $P_{\beta_{g(l)}}$ and $P_{\gamma_{g(k)}}$ from the joint model do not show departure from the
373     expected large sample distributions ($\chi^2$ with 1 degree of freedom (*df)*, see **Figure S2**) and the
374     type I error of each test under the null is reasonably well controlled (**Table S7**). The hypothesis
375     testing procedure applied to single-parameter SNP tests, classifies each SNP as direct or indirect
376     association with each T1DC trait with rates close to the nominal level (**Table 3** and **Figure S3**).
377     However, the classification rates for test of *direct and indirect* SNP association test appears less
378     than nominal (see **Table 3** for SNP5 with SBP/DN traits and **Figure S3** for the other SNPs with
379     other QT/time-to-event trait pairs).

380     [Insertion of Table 3]

381     Under the *causal* scenario, the hypothesis testing procedure has high correct classification rates
382     for *indirect* (SNP1) or *direct* associations (SNP2, SNP4), as shown in **Figures 4-a, 4-b** and **S4**.
383     Conclusions are similar for JM-mis, with attenuated correct classification rates compared to JM-
384     cmp for direct SNP associations but still higher than classification rates based on CM-obs
385     (particularly for SNP4). Compared to results based on CM-obs, the largest relative classification
386     improvement is found for direct SNP4 association with DN (ranges: 1.18-251.9 with JM-cmp and
387     1.18-110.98 with JM-mis, **Figure 4**).

388     [Insertion of Figure 4]

389     On the other hand, SNP3 with an *indirect* effect via the unobserved longitudinal trait *U*, is most
390     frequently detected incorrectly as a *direct* association with each T1DC trait by both JM-mis and
391     CM-obs (**Figure 4-b**). SNP5, which has both *direct* and *indirect* effects on DN induced via SBP,
392     is most frequently misclassified by JM-cmp and JM-mis as an *indirect* association at significance
393     levels varying from $P^* = 0.01$ to 5x10$^{-8}$ (**Figure 4-c** and **4-d**). SNP5 misclassification rate even
394     increases with $P^*$ decreasing to 5x10$^{-8}$, due to the direct SNP5 $P_{\gamma_{g(k)}}$. In comparison, the
395     procedure based on CM-obs results most frequently correctly classify SNP5 association in more
396     than 75% of the replicates at the same significance levels (**Figures 4-c** and **4-d**); this is explained
397     by the overestimated direct SNP5 effect, arising from measurement error in SBP as noted before.

398     In summary, our simulations show that by accounting for measurement errors in the longitudinal
399     QT risk factors and for dependencies between/within the traits, the proposed joint model improves
400     estimation accuracy and correct classification of associations of SNPs *directly* associated with

401  each time-to-event trait or *indirectly* associated via a measured longitudinal QT in comparison to
402  classification using the CM-obs approach. However, when a SNP has both *direct* and *indirect*
403  effects on a time-to-event trait, the proposed testing procedure can be conservative since it
404  requires the joint significance of the two SNP effects, $\beta_{g(l)}$ and $\gamma_{g(k)}$, where the power of each
405  test depends on effect size and the trait distribution. As a result, a SNP with a direct and an indirect
406  association can be misclassified as either a direct or an indirect association. Finally, when a SNP
407  has an indirect effect on both T1DC traits via an unmeasured QT, the testing procedure based on
408  JM-mis, that captures some of the unexplained dependency between time-to-event traits through
409  the frailty term, does not prevent misclassification as a direct association. This observation also
410  shows the importance of the correct specification of the joint model - including all the intermediate
411  QT(s) - to avoid misclassification of direct and/or indirect SNP associations.

412  **Application in the DCCT Genetics Study data**

413  We demonstrate the feasibility of the proposed approach by an application in the DCCT Genetics
414  Study data. We use time to mild DR and time to persistent microalbuminuria, for DR and DN
415  outcomes respectively, as previously defined in the motivating GWAS of HbA1c[19] (see
416  **Supplementary information SI-2** for details). Genome-wide genotyping in DCCT subjects was
417  performed using HumanCoreExome Bead Array (Illumina, San Diego, CA, USA) and standard
418  quality controls procedures were applied to individuals and genetic markers[19,35]. Ungenotyped
419  autosomal SNPs were imputed using 1000 Genomes data phase 3[36] (v5) and minimac3[37]
420  (v.1.0.13), as previously described[35].

421  Out of the 667 DCCT individuals, we analyze *N*=516 subjects with genetic data, without mild to
422  moderate non-proliferative retinopathy or without DN event at DCCT baseline. By the time of the
423  DCCT close-out visit, 297 (57.6%) experienced a DR event, 61 (11.8%) a DN event, including 47
424  subjects (9.1%) that experienced both events. After SNP filtering and pruning on linkage
425  disequilibrium (see **Supplementary Information SI-2** for details), we analyze 307 candidate
426  SNPs reported as associated with HbA1c, SBP, and multiple definitions of DR and/or DN[19,34,38–42]
427  (see **Table S8** for the full list of SNPs).

428  Compared to the DCCT-based simulated datasets, the proportion of DN events observed in the
429  application is lower. We expect that the test power for direct SNP association with DN under the
430  joint model to be reduced; this implies reduce chances to correctly classify a SNP with either a
431  *direct* or a *direct and indirect* association with DN. We also observe larger effects of HbA1c on
432  T1DC traits ($\alpha_{1(k)}$, *k*=1, 2) and a smaller effect of SBP on DN ($\alpha_{2(2)}$), as shown in **Tables 2 vs**
433  **S9**. Larger $\alpha_{l(k)}$ values increase the contribution of the indirect effect ($\mu_{g(l,k)}$) to the overall SNP
434  effect ($\theta_{k(l)}$), which can result in a direct SNP effect ($\gamma_{g(k)}$) more severely biased if the longitudinal
435  QT is not accounted properly in the joint model, while lower $\alpha_{l(k)}$ values reduce $\mu_{g(l,k)}$ and thus
436  $\gamma_{g(k)}$ is less severely biased. Overall, in the DCCT application, we expect a SNP with an *indirect*
437  effect via HbA1c on any T1DC trait to be more subject to be misclassified as having both a *direct*
438  and *indirect* association, compared to the simulation results, while a SNP with an *indirect* effect
439  via SBP on DN is less subject to this misclassification.

440    Given prior evidence for cumulative effects of HbA1c on T1DC traits[20,21], we compare joint model
441    results obtained with contemporaneous, updated cumulative mean, and time-weighted
442    cumulative HbA1c effects on T1DC (**Supplementary Information SI-2** for details). We present
443    results under the latter association structure since it has a stronger prior association with T1DC
444    and in the DCCT individuals analyzed here (**Table S9**). We obtain similar conclusions with
445    alternative association structures (**Figure S5-a**). As shown in **Figure 5-a**, rs10810632 and
446    rs1358030 are classified as indirect associations with both T1DC traits via their association with
447    HbA1c shared risk factor ($P_{\beta_{g(l)}} \leq P^*$ and $P_{\gamma_{g(k)}} > P^*$, $P^*$=1.7x10$^{-4}$ after Bonferroni correction for
448    the 289.02 effective SNPs tested[43]). These two SNPs have calculated indirect effects ($\mu_{g(1,k)}$)
449    and 95% bootstrap confidence intervals that do not include 0 (**Figure 5-b**). As noted before,
450    rs10810632 and rs1358030 were discovered previously in a GWAS of HbA1c in DCCT[19].
451    Classification of these two SNPs as direct and/or indirect associations may be affected by the
452    Winner's curse bias, due to the modest sample size in this analysis, and the same individuals
453    contributing to discovery and classification[44,45]. For the other candidate SNPs which were
454    discovered in larger and/or independent studies[19,34,38–42], we find little evidence of association with
455    the discovery traits in marginal analyses (**Table S8**).

456    Conclusions regarding classification of direct and/or indirect associations based on the procedure
457    using $P_{\gamma_{g(k)}}$ from CM-obs are similar to those based on the joint model results for all SNPs (**Figure
458    S5-b**), although we noticed some differences in $P_{\gamma_{g(k)}}$; that could be explained by the attenuated
459    association of HbA1c and SBP with T1DC traits due to measurement errors ignored from CM-obs
460    (**Tables S10 *vs* S9**).

## Discussion

462    We present new methods for statistical genetic analysis under a joint model specification of
463    multiple time-to-event traits and multiple longitudinal risk factors designed to characterize the
464    complex genetic architecture of related traits in longitudinal studies of disease progression. The
465    proposed model accounts for dependencies within and between traits and can account for trait-
466    specific and shared covariates, measurement errors in the longitudinal traits, as well as effects of
467    unobserved baseline confounding factors between the time-to-event traits through the subject-
468    specific frailty term.

469    Evaluation by *realistic* data-informed simulation study of complex T1DC genetic architecture
470    shows that the proposed joint model improves estimation accuracy and efficiency compared to
471    the Cox PH frailty model adjusted for observed longitudinal trait values as time-dependent
472    covariates. This improvement holds particularly when the longitudinal risk factors are measured
473    with errors or the time-to-event outcome has a low event rate. Indeed, measurement error
474    attenuates the association of the longitudinal risk factor with the time-to-event trait, and as a result,
475    if a SNP has an indirect effect induced by the QT, the estimated direct SNP effect on a time-to-
476    event is biased away from zero. The magnitude of this bias tends to be worse with larger
477    measurement error[14,15]. Identification of direct and/or indirect SNP associations is also facilitated
478    under joint modelling in the presence of risk factor measurement error. When, however, a SNP
479    has an indirect effect via an unmeasured longitudinal risk factor, both the proposed and Cox PH

480 model approaches produce biased estimates of the direct SNP effects. The extent of this bias
481 depends on the magnitude of the effect sizes $\beta_{g(l)}$ and $\alpha_{l(k)}$. Although the frailty term in the joint
482 model time-to-event sub-model captures some of the unexplained shared dependency between
483 the time-to-event traits, we did not observe bias reduction of the direct SNP estimate in the case
484 of an unmeasured risk factor, which can lead to misclassification of the SNP as a direct
485 association. This may be a limitation of using a time-invariant frailty term that does not fully
486 account for the indirect SNP effect via the unmeasured longitudinal risk factor. Because joint
487 significance of both SNP effects is required to classify a SNP as having both direct and indirect
488 associations, a SNP with both direct and an indirect association can be misclassified as having
489 either a direct or an indirect association if one of the SNP effects is too small to be detected at
490 the specified level. The proposed procedure can be adapted to other research questions, for
491 example to identify SNPs that have direct association among the SNPs found associated with a
492 longitudinal risk factor. In this case, it may be desirable to apply a less stringent significance
493 threshold for the test of the direct SNP effect while controlling the overall type I error. Alternatively,
494 intersection-union tests, as based on the likelihood ratio test[46], could be used to assess joint
495 significance of $\beta_{g(l)}$ and $\gamma_{g(k)}$ but would require computation of the full joint likelihood.

496 Application of the proposed joint model approach in longitudinal studies of disease progression,
497 such as in the DCCT Genetics Study, improves classification of direct and/or indirect SNP
498 association which can help to elucidate the genetic architecture of complex traits. In the context
499 of mediation analysis, Liu et al[47] discussed various formulations and interpretations of joint
500 models, with shared-random-effects accounting for potential unmeasured baseline confounding
501 factors between *one* longitudinal and *one* time-to-event traits. Using applications in datasets from
502 two clinical trials, they illustrate interpretation of sensitivity analysis to unmeasured baseline
503 confounders. Adaptation of the joint model we propose for multiple longitudinal and multiple time-
504 to-event traits for mediation analysis requires extension of the mediation assumptions[48,49] to the
505 case of multiple mediators and multiple time-to-event traits. Specific evaluations of the proposed
506 model under these assumptions are also warranted.

507 Although our primary aim in this report is to develop statistical methods to accurately distinguish
508 among direct and/or indirect SNP associations with each time-to-event trait, the multi-trait aspect
509 of the joint model lends itself to development of multi-trait SNP association testing for SNP
510 discovery. In **Supplementary Information SI-3**, we present a joint-parameter test based on a
511 generalized Wald statistic. In application to the simulated DCCT-based complex genetic
512 architecture, we observe good type I error control under the global genetic null scenario, improved
513 power for SNP discovery when a SNP has multiple trait effects, and power maintenance in other
514 SNP association scenarios.

515 We acknowledge several features of the proposed joint model approach that warrant examination
516 in further work. *Firstly*, to reduce computational complexity and improve model flexibility, we used
517 two-stage parameter estimation. In some circumstances, this approach can produce biased
518 estimates and/or underestimated standard errors[16]. Biased estimates can result from non-random
519 censoring of the longitudinal trait values due to the occurrence of an event or from informative
520 dropout[50,51]. Our simulation results show minimal biases in the absence of informative censoring,

521 even when the joint model is mis-specified. In the DCCT application, characterized by
522 administrative censoring and a high completion rate, these biases are of minimal concern
523 because longitudinal trait values continued to be recorded regardless of the occurrence of any
524 T1DC events; we estimated the trajectories using all the available measurements. Furthermore,
525 we obtain robust bootstrap estimates of the covariance matrix, and simulation results under the
526 null did not show deviation from expected distributions. Use of the bootstrap also facilitates joint
527 testing of parameters from both stages. In the presence of informative censoring, we recommend
528 sensitivity analysis using existing implementations joint likelihood estimation. To our knowledge
529 those implementations only exist for simpler joint model formulation with either one longitudinal
530 and one time-to-event trait[52] or multiple longitudinal traits and one time-to-event outcome[28,53].
531 *Secondly*, because the joint model integrates longitudinal and time-to-event sub-models, model
532 misspecification can occur in multiple ways and lead to invalid inference[54]. Therefore, we
533 recommend a careful assessment of the model assumptions using joint model diagnostic tools[23].
534 *Thirdly*, patient visits were scheduled with high frequency in DCCT, so we ignored the modest
535 degree of interval censoring in the current implementation of the joint model; when there are
536 longer gaps between visits, extended methods are needed to account for interval censoring with
537 additional simulation studies to assess impact on joint model estimates. *Lastly*, joint models are
538 very computational demanding, particularly for genetic association studies that test millions of
539 variants. In DCCT application, it took ~60 seconds to fit the joint model for each SNP and
540 ~1090.32 more seconds (~ 18 minutes) to estimate the covariance matrix with 500 bootstraps run
541 in parallel on 4 nodes (each node with 40 CPU and 202 GB RAM). While analysis at the genome
542 level involving, up to 85 million 1000G-imputed SNPs in DCCT, seems computationally
543 unrealistic, a screening approach without bootstrap to select the SNPs based on their association
544 *P*-values, followed by the bootstrap refinement would reduce the computational burden. Recent
545 computationally efficient algorithms have been developed to improve the efficiency of the linear
546 mixed model[55] and Cox PH model[56,57] for genetic association studies, but to date, they remain to
547 be implemented for multivariate outcomes.

548 With the increasing development of national biobanks[58–60] consisting of a large collection of
549 phenotypes, environmental factors, biomarkers and other risk factors collected over time and
550 linked with genetic data, we anticipate that joint model methods can be applied to characterize
551 the genetic architecture of complex traits using these data resources[61]. In addition, the results of
552 such analysis could contribute towards the translation of human genetic findings to personalized
553 medicine by providing more efficient SNP effect estimates for increased precision in polygenic
554 risk score development[62], and causal inference using mediation and mendelian randomization
555 studies. Lastly, the joint model framework can also enable dynamic prediction beneficial for
556 dynamic risk assessment[7,63] and optimization of intervention strategies[64,65].

## Acknowledgements

## Contributions

All named authors affirm that authorship is merited based on the International Committee of Medical Journal Editors (ICMJE) authorship criteria. MB, SBB and RVC designed the study. MB conducted the simulation study and data analysis and drafted the manuscript. SBB, RVC and OEG contributed to the data analysis and interpretation of the results and reviewed/edited the manuscript. ADP contributed to the acquisition of the data and reviewed/edited the manuscript. All authors approve the final version to be published. SBB and MB are the guarantors of this work.

## Abbreviations

DCCT: Diabetes control and complications trial;

DR: diabetic retinopathy;

DN: diabetic nephropathy;

GWAS: genome-wide association study;

HbA1c: Hemoglobin A1c;

LD: linkage disequilibrium;

MAF: minor allele frequency;

PH: proportional hazards;

590    QT(s): quantitative trait(s);

591    SBP: systolic blood pressure;

592    SNP: single nucleotide polymorphism;

593    T1DC: type 1 diabetes complications.

## References

595    1. Ibrahim, J.G., Chu, H., and Chen, L.M. (2010). Basic concepts and methods for joint models of
596    longitudinal and survival data. J. Clin. Oncol. *28*, 2796–2801.

597    2. Chen, L.M., Ibrahim, J.G., and Chu, H. (2011). Sample size and power determination in joint
598    modeling of longitudinal and survival data. Stat. Med. *30*, 2295–2309.

599    3. Hogan, J.W., and Laird, N.M. (1998). Increasing efficiency from censored survival data by using
600    random effects to model longitudinal covariates. Stat. Methods Med. Res. *7*, 28–48.

601    4. Wu, L., Liu, W., Yi, G.Y., and Huang, Y. (2012). Analysis of longitudinal and survival data: joint
602    modeling, inference methods, and issues. J. Probab. Stat. *2012*, 1–17.

603    5. Asar, Ö., Ritchie, J., Kalra, P.A., and Diggle, P.J. (2015). Joint modelling of repeated
604    measurement and time-to-event data: An introductory tutorial. Int. J. Epidemiol. *44*, 334–344.

605    6. Hickey, G.L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2016). Joint modelling
606    of time-to-event and multivariate longitudinal outcomes: recent developments and issues. BMC
607    Med. Res. Methodol. *16*, 117.

608    7. Papageorgiou, G., Mauff, K., Tomer, A., and Rizopoulos, D. (2019). An overview of joint
609    modeling of time-to-event and longitudinal outcomes. Annu. Rev. Stat. Its Appl. *6*, 223–240.

610    8. Hickey, G.L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). Joint models of
611    longitudinal and time-to-event data with more than one event time outcome: a review. Int. J.
612    Biostat. *14*,.

613    9. Zhu, H., Ibrahim, J.G., Chi, Y.Y., and Tang, N. (2012). Bayesian influence measures for joint
614    models for longitudinal and survival data. Biometrics *68*, 954–964.

615    10. Tang, N.S., Tang, A.M., and Pan, D.D. (2014). Semiparametric Bayesian joint models of
616    multivariate longitudinal and survival data. Comput. Stat. Data Anal. *77*, 113–129.

617    11. Tang, A.M., and Tang, N.S. (2014). Semiparametric Bayesian inference on skew-normal joint
618    modeling of multivariate longitudinal and survival data. Stat. Med. *34*, 824–843.

619    12. Bycott, P., and Taylor, J. (1998). A comparison of smoothing techniques for CD4 data
620    measured with error in a time-dependent Cox proportional hazards model. Stat. Med. *17*, 2061–
621    2077.

622    13. Self, S., and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of

623    disease. In AIDS Epidemiology, (Boston, MA: Birkhäuser Boston), pp. 231–255.

624    14. Tsiatis, A.A., Degruttola, V., and Wulfsohn, M.S. (1995). Modeling the relationship of survival
625    to longitudinal data measured with error. Applications to survival and CD4 counts in patients with
626    AIDS. J. Am. Stat. Assoc. *90*, 27–37.

627    15. Dafni, U.G., and Tsiatis, A.A. (1998). Evaluating Surrogate Markers of Clinical Outcome When
628    Measured with Error. Biometrics *54*, 1445.

629    16. Wulfsohn, M.S., and Tsiatis, A.A. (2006). A joint model for survival and longitudinal data
630    measured with error. Biometrics *53*, 330.

631    17. Paterson, A.D., and Bull, S.B. (2012). Does familial clustering of risk factors for long-term
632    diabetic complications leave any place for genes that act independently? J. Cardiovasc. Transl.
633    Res. *5*, 388–398.

634    18. The Diabetes Control and Complications Trial Research Group (1993). The effect of intensive
635    treatment of diabetes on the development and progression of long-term complications in insulin-
636    dependent diabetes mellitus. N. Engl. J. Med. *329*, 977–986.

637    19. Paterson, A.D., Waggott, D., Boright, A.P., Hosseini, S.M., Shen, E., Sylvestre, M.P., Wong,
638    I., Bharaj, B., Cleary, P.A., Lachin, J.M., et al. (2010). A genome-wide association study identifies
639    a novel major locus for glycemic control in type 1 diabetes, as measured by both A1C and glucose.
640    Diabetes *59*, 539–549.

641    20. Lind, M., Odén, A., Fahlén, M., and Eliasson, B. (1995). The relationship of glycemic exposure
642    (HbA(1c)) to the risk of development and progression of retinopathy in the diabetes control and
643    complications trial. Diabetes *53*, 1093–1098.

644    21. Lind, M., Odén, A., Fahlén, M., and Eliasson, B. (2010). The shape of the metabolic memory
645    of HbA1c: Re-analysing the DCCT with respect to time-dependent effects. Diabetologia *53*, 1093–
646    1098.

647    22. Laird, N.M., and Ware, J.H. (1982). Random-effects models for longitudinal data. Biometrics
648    *38*, 963–974.

649    23. Rizopoulos, D. (2012). Joint models for longitudinal and time-to-event data, with applications
650    in R (Chapman and Hall/CRC).

651    24. Hougaard, P. (1995). Frailty models for survival data. Lifetime Data Anal. *1*, 255–273.

652    25. Mauff, K., Steyerberg, E.W., Nijpels, G., van der Heijden, A.A.W.A., and Rizopoulos, D.
653    (2017). Extension of the association structure in joint models to include weighted cumulative
654    effects. Stat. Med. *36*, 3746–3759.

655    26. Tsiatis, A.A., and Davidian, M. (2001). A semiparametric estimator for the proportional
656    hazards model with longitudinal covariates measured with error. Biometrika *88*, 447–458.

657    27. Tsiatis, A.A., and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data:
658    An overview. Stat. Sin. *14*, 809–834.

659  28. Hickey, G.L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). JoineRML: A
660  joint model and software package for time-to-event and multivariate longitudinal outcomes. BMC
661  Med. Res. Methodol. *18*,.

662  29. Therneau, T.M. (2020). A Package for survival analysis in R. R package version 3.2-7, URL
663  http://CRAN.R-project.org/package=survival.

664  30. Terry M. Therneau, and Patricia M. Grambsch (2000). Modeling survival data: Extending the
665  Cox model (New York: Springer).

666  31. Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and
667  other methods. Biometrika *68*, 589.

668  32. Steffes, M., Cleary, P., Goldstein, D., Little, R., Wiedmeyer, H.-M.M., Rohlfing, C., England,
669  J., Bucksa, J., and Nowicki, M. (2005). Hemoglobin A1c measurements over nearly two decades:
670  sustaining comparable values throughout the Diabetes Control and Complications Trial and the
671  Epidemiology of Diabetes Interventions and Complications study. Clin. Chem. *51*, 753–758.

672  33. Lorenzi, G.M., Braffett, B.H., Arends, V.L., Danis, R.P., Diminick, L., Klumpp, K.A., Morrison,
673  A.D., Soliman, E.Z., Steffes, M.W., Cleary, P.A., et al. (2015). Quality control measures over 30
674  years in a multicenter clinical study: Results from the Diabetes Control and Complications Trial /
675  Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) study. PLoS One *10*,
676  e0141286.

677  34. Hosseini, S.M., Boright, A.P., Sun, L., Canty, A.J., Bull, S.B., Klein, B.E.K., Klein, R., and
678  Paterson, A.D. (2015). The association of previously reported polymorphisms for microvascular
679  complications in a meta-analysis of diabetic retinopathy. Hum. Genet. *134*, 247–257.

680  35. Roshandel, D., Gubitosi-Klug, R., Bull, S.B., Canty, A.J., Pezzolesi, M.G., King, G.L., Keenan,
681  H.A., Snell-Bergeon, J.K., Maahs, D.M., Klein, R., et al. (2018). Meta-genome-wide association
682  studies identify a locus on chromosome 1 and multiple variants in the MHC region for serum C-
683  peptide in type 1 diabetes. Diabetologia *61*, 1098–1111.

684  36. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P.,
685  Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global
686  reference for human genetic variation. Nature *526*, 68–74.

687  37. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y.,
688  Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods.
689  Nat. Genet. *48*, 1284–1287.

690  38. Wheeler, E., Leong, A., Liu, C.-T.T., Hivert, M.-F.F., Strawbridge, R.J., Podmore, C., Li, M.,
691  Yao, J., Sim, X., Hong, J., et al. (2017). Impact of common genetic determinants of Hemoglobin
692  A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic
693  genome-wide meta-analysis. PLoS Med. *14*, e1002383.

694  39. Evangelou, E., Warren, H.R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., Ntritsos,
695  G., Dimou, N., Cabrera, C.P., Karaman, I., et al. (2018). Genetic analysis of over 1 million people
696  identifies 535 new loci associated with blood pressure traits. Nat. Genet. *50*, 1412–1425.

697  40. Pollack, S., Igo, R.P., Jensen, R.A., Christiansen, M., Li, X., Cheng, C.-Y., Ng, M.C.Y., Smith,

698    A. V, Rossin, E.J., Segrè, A. V, et al. (2019). Multiethnic genome-wide association study of
699    diabetic retinopathy using liability threshold modeling of duration of diabetes and glycemic control.
700    Diabetes *68*, 441–456.

701    41. Grassi, M.A., Tikhomirov, A., Ramalingam, S., Below, J.E., Cox, N.J., and Nicolae, D.L.
702    (2011). Genome-wide meta-analysis for severe diabetic retinopathy. Hum. Mol. Genet. *20*, 2472–
703    2481.

704    42. Sandholm, N., Salem, R.M., McKnight, A.J., Brennan, E.P., Forsblom, C., Isakova, T., McKay,
705    G.J., Williams, W.W., Sadlier, D.M., Mäkinen, V.-P.P., et al. (2012). New susceptibility loci
706    associated with kidney disease in type 1 diabetes. PLoS Genet. *8*, e1002921.

707    43. Li, J., and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues
708    of a correlation matrix. Heredity (Edinb). *95*, 221–227.

709    44. Kraft, P. (2008). Curses—Winner's and Otherwise—in Genetic Epidemiology. Epidemiology
710    *19*, 649–651.

711    45. Sun, L., Dimitromanolakis, A., Faye, L.L., Paterson, A.D., Waggott, D., DCCT/EDIC Research
712    Group, and Bull, S.B. (2011). BR-squared: a practical solution to the winner's curse in genome-
713    wide scans. Hum. Genet. *129*, 545–552.

714    46. Schaid, D.J., Tong, X., Larrabee, B., Kennedy, R.B., Poland, G.A., and Sinnwell, J.P. (2016).
715    Statistical methods for testing genetic pleiotropy. Genetics *204*, 483–497.

716    47. Liu, L., Zheng, C., and Kang, J. (2018). Exploring causality mechanism in the joint analysis of
717    longitudinal and survival data. Stat. Med. *37*, 3733–3744.

718    48. Sobel, M.E. (1982). Asymptotic confidence intervals for indirect effects in structural equation
719    models. Sociol. Methodol. *13*, 290.

720    49. Mackinnon, D.P., Warsi, G., and Dwyer, J.H. (1995). A simulation study of mediated effect
721    measures. Multivariate Behav. Res. *30*, 41.

722    50. Albert, P.S., Shih, J.H., Albert, P.S., Shih, J.H., Albert1', P.S., and Shih2, J.H. (2010). On
723    estimating the relationship between longitudinal measurements and time-to-event data using a
724    simple two-stage procedure. Biometrics *66*, 983–991.

725    51. Ye, W., Lin, X., and Taylor, J.M.G. (2008). Semiparametric modeling of longitudinal
726    measurements and time-to-event data - A two-stage regression calibration approach. Biometrics
727    *64*, 1238–1246.

728    52. Rizopoulos, D. (2010). JM : An R Package for the joint modelling of longitudinal and time-to-
729    event data. J. Stat. Softw. *35*,.

730    53. Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and
731    time-to-event data using MCMC. J. Stat. Softw. *72*,.

732    54. Arisido, M.W., Antolini, L., Bernasconi, D.P., Valsecchi, M.G., and Rebora, P. (2019). Joint
733    model robustness compared with the time-varying covariate Cox model to evaluate the
734    association between a longitudinal marker and a time-to-event endpoint. BMC Med. Res.

735 Methodol. *19*, 222.

736 55. Sikorska, K., Lesaffre, E., Groenen, P.J.F., Rivadeneira, F., and Eilers, P.H.C. (2018).
737 Genome-wide analysis of large-scale longitudinal outcomes using penalization - GALLOP
738 algorithm. Sci. Rep. *8*,.

739 56. Rizvi, A.A., Karaesmen, E., Morgan, M., Preus, L., Wang, J., Sovic, M., Hahn, T., and
740 Sucheston-Campbell, L.E. (2019). gwasurvivr: an R package for genome-wide survival analysis.
741 Bioinformatics *35*, 1968–1970.

742 57. Bi, W., Fritsche, L.G., Mukherjee, B., Kim, S., and Lee, S. (2020). A fast and accurate method
743 for genome-wide time-to-event data analysis and its application to UK Biobank. J. Clean. Prod.
744 *107*, 222–233.

745 58. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic,
746 D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping
747 and genomic data. Nature *562*, 203–209.

748 59. Scholtens, S., Smidt, N., Swertz, M.A., Bakker, S.J.L., Dotinga, A., Vonk, J.M., Van Dijk, F.,
749 Van Zon, S.K.R., Wijmenga, C., Wolffenbuttel, B.H.R., et al. (2015). Cohort Profile: LifeLines, a
750 three-generation cohort study and biobank. Int. J. Epidemiol. *44*, 1172–1180.

751 60. Dummer, T.J.B., Awadalla, P., Boileau, C., Craig, C., Fortier, I., Goel, V., Hicks, J.M.T.,
752 Jacquemont, S., Knoppers, B.M., Le, N., et al. (2018). The Canadian Partnership for Tomorrow
753 Project: A pan-Canadian platform for research on chronic disease prevention. Cmaj *190*, E710–
754 E717.

755 61. Gasparini, A., Abrams, K.R., Barrett, J.K., Major, R.W., Sweeting, M.J., Brunskill, N.J., and
756 Crowther, M.J. (2020). Mixed-effects models for health care longitudinal data with an informative
757 visiting process: A Monte Carlo simulation study. Stat. Neerl. *74*, 5–23.

758 62. Young, A.I., Benonisdottir, S., Przeworski, M., and Kong, A. (2019). Deconstructing the
759 sources of genotype-phenotype associations in humans. Science *365*, 1396–1400.

760 63. Bull, L., Lunt, M., Martin, G., Hyrich, K., and Sergeant, J. (2020). Harnessing repeated
761 measurements of predictor variables for clinical risk prediction: a review of existing methods.
762 Diagnostic Progn. Res. *4*, 1–16.

763 64. Sweeting, M.J., and Thompson, S.G. (2011). Joint modelling of longitudinal and time-to-event
764 data with application to predicting abdominal aortic aneurysm growth and rupture. Biometrical J.
765 *53*, 750–763.

766 65. Yuen, H.P., Mackinnon, A., Hartmann, J., Amminger, G.P., Markulev, C., Lavoie, S., Schäfer,
767 M.R., Polari, A., Mossaheb, N., Schlögelhofer, M., et al. (2018). Dynamic prediction of transition
768 to psychosis using joint modelling. Schizophr. Res. *202*, 333–340.

769 66. Ibrahim, J.G., Chen, M.-H., and Sinha, D. (2001). Joint models for longitudinal and survival
770 data. In EMR-IBS Bi-Annual Meeting, pp. 262–289.

771

## Appendix

Inference under the joint likelihood can be based on joint maximum-likelihood estimation, Bayesian estimation of the joint likelihood parameters, or a two-stage strategy[23]. Joint maximum-likelihood and Bayesian methods enable joint parameter estimation but can be computationally challenging, especially for multiple traits. Two-stage approaches[12–15] sequentially fit the longitudinal (Stage 1) and time-to-event (Stage 2) sub-models, with trajectory estimates from Stage 1 incorporated into Stage 2; this can be relatively computationally efficient and lends itself to flexible model formulation. However, as noted in introduction, inference can be mis-calibrated because of propagation of errors[16].

### *Joint likelihood function*

Let $\Omega$ be the full parameter vector containing all fixed parameters from the longitudinal and time-to-event sub-models respectively. We assume:

(A1) $b_i \sim MVN(0, D)$

(A2) $\varepsilon_{ij(l)} \sim N\left(0, \sigma_{(l)}^2\right)$ and $\varepsilon_{ij(l)} \perp \varepsilon_{ip(l)}$ between visits $j \neq p$

(A3) $b_{i(l)} \perp \varepsilon_{ij(l)}$

(A4) $u_i \perp b_i$

Given (A1)-(A4), it is appropriate to assume that the random effects $b_i$ account for: the association between the *L* longitudinal traits; the association between both the longitudinal and time-to-event outcomes[66]; and the serial correlation between the repeated measurements in each longitudinal process[22]; and that given (A4), the frailty term accounts for the residual dependency between the time-to-event traits[24]. Under these conditional independence assumptions, the joint likelihood function of the model parameters given the observed data is:

$$L(\Omega|Y_i, T_i, \delta_i, SNP_i, H_i, V_i) = \prod_{i=1}^{N} \int f_1(Y_i|\, b_i, \Omega) \times f_2(T_i, \delta_i \mid b_i, u_i, \Omega) \times f_3(b_i|\Omega) \times f_4(u_i|\Omega)\, du_i db_i$$

where:

- $(T_i, \delta_i) = ((T_{i(1)}, \delta_{i(1)})', \ldots, (T_{i(k)}, \delta_{i(k)})', \ldots, (T_{i(K)}, \delta_{i(K)})')'$
- $Y_i = (Y_{i(1)}', \ldots, Y_{i(l)}', \ldots, Y_{i(L)}')'$
- $f_1(Y_i|\, b_i, \Omega) = \prod_{l=1}^{L} \left(\frac{1}{\sqrt{2\pi\sigma_{(l)}^2}}\right)^{n_i} exp\left[-\frac{1}{2\sigma_{(l)}^2} \prod_{j=1}^{n_i} (Y_{ij(l)}(t_{ij}) - Y_{ij(l)}^*(t_{ij}))^2\right]$
- $f_2(T_i, \delta_i \mid b_i, u_i, \Omega) = \prod_{k=1}^{K} \left[\lambda_{i(k)}\left(T_{i(k)}|\, b_i, u_i, \Omega\right)\right]^{\delta_{ik}} S_{i(k)}(T_{i(k)}|\, b_i, u_i, \Omega)$
- $S_{i(k)}\left(T_{i(k)}|\, b_i, u_i, \Omega\right)$

$$801 \qquad = exp\left[-\int_0^{T_{i(k)}} \lambda_{i0}(s)exp\{\gamma_{g(k)}SNP_i + W_{i(k)}(s) + \gamma_{v(k)}V_{i(k)} + u_i\}ds\right]$$

802 $\quad \bullet \quad f_3(b_i|\Omega) = \left(\frac{1}{\sqrt{2\pi}}\right)^{-q/2}|D|^{-1/2}exp\left\{-\frac{1}{2}b_i^T D^{-1}b_i\right\}$, where $q$ is the dimension of the $D$ matrix

803 $\quad \bullet \quad f_4(u_i|\Omega) = \frac{u_i^{a-1}exp\,(-u_i/b)}{\Gamma(a)b^a}$, i.e. we assume $u_i \sim \Gamma(a,b)$, with $a$, $b$>0.

804 Calculation of this full likelihood requires multivariate integration with respect to the random
805 effect's distribution, which can lead to demanding computation. When the random effect vector
806 $b_i$, has a small dimension, say less than 3, the integral can be evaluated via Gaussian quadrature
807 which approximates the integral by a weighted sum of the target function evaluated at pre-
808 specified sample points. However, when the dimension is larger, it is demanding to calculate the
809 integrals with satisfactory approximation accuracy. Although a full likelihood specification enables
810 rigorous study of asymptotic properties, its large sample approximation may not be accurate when
811 sample size is small. In comparison, the Bayesian paradigm does not require asymptotic
812 approximations, but the design of an efficient sampling algorithm to study the posterior distribution
813 is challenging.

814 ***Likelihood functions under the two-stage approximation***

815 Let $\Omega$ and $\Gamma$ be the vectors containing all fixed parameters from the longitudinal and time-to-event
816 sub-models respectively.

817 *Stage1:* Multivariate mixed model

$$818 \qquad L(\Omega|Y_i, X_i, C_i) = \prod_{i=1}^N \int_{b_i} f_1(Y_i| b_i, \Omega) \times f_3(b_i|\Omega) \, db_i$$

819 $\quad$ *Stage2:* Multivariate Cox PH model adjusted for fitted values of the trajectories $\widehat{Y_i^*(T_i)}$

$$820 \qquad L(\Gamma|T_i, \delta_i, C_i) = \prod_{i=1}^N \int_{u_i} f_2\big(T_i, \delta_i \,|\widehat{Y_{l(l)}^*}(T_i), u_i, \Gamma\big) \times f_4(u_i|\Gamma) \times du_i$$

821 Where:

822 $\quad \bullet \quad f_2\big(T_i, \delta_i \,|\widehat{Y_i}(T_i), u_i, \Gamma\big) = \prod_{k=1}^K \big[\lambda_{i(k)}\big(T_{i(k)}|\widehat{Y_{l(l)}^*}(T_{i(k)}), u_i, \Gamma\big)\big]^{\delta_{ik}} S_{ik}\big(T_{i(k)}|\widehat{Y_{l(l)}^*}(T_{i(k)}), u_i, \Gamma\big)$

823 $\quad \bullet \quad S_{i(k)}\big(T_{i(k)}|\widehat{Y_{l(l)}}(T_{i(k)}), u_i, \Gamma\big) = exp\left[-\int_0^{T_{i(k)}} \lambda_{i0}(s)exp\{\gamma_{g(k)}SNP_i + W_{i(k)}(s) + \gamma_{v(k)}V_{i(k)} + \right.$
824 $\quad u_i\}ds\Big]$

825 $\quad \bullet \quad$ With $W_{i(k)}(s) = \sum_{l=1}^{L_k} \alpha_{l(k)}f_{l(k)}(\widehat{Y_{l(l)}^*}(s))$

826    Unlike the *joint likelihood function*, where the shared random effects $b_i$ account for the
827    dependencies between the longitudinal and the time-to-event traits, the two-stage approach
828    accounts for the dependencies between the longitudinal and time-to-event traits via the fitted
829    values of the longitudinal trajectories. As mentioned previously, this approximation can produce
830    biased estimates and/or underestimated standard errors[16], particularly in presence of non-random
831    censoring of the longitudinal trait values due to the occurrence of an event or from informative
832    dropout[50,51] and/or because of propagation errors of Stage 1 parameters in Stage2[16]. Indeed,
833    informative missingness/dropouts lead in differential follow-up between patients with and without
834    an event, the time-to-event processes are related to the length of the follow-up, and thus the
835    random effects $b_{i(l)}$ can depend on the event times (e.g. patients who have an event early are
836    more likely to have positive random slopes). However, in absence of informative
837    dropouts/missingness, as we showed in our simulation studies, this approach has low bias and is
838    computationally feasible for genetic association studies.

## Multivariate mixed-effects sub-model for $L$ longitudinal QT(s) (and/or risk factors)

**Measured longitudinal traits:**

$$Y_{i(l)}(t_{ij}) = Y^*_{i(l)}(t_{ij}) + \varepsilon_{ij(l)} \sim N(0, \sigma^2_{(l)})$$

- *intermittently* (ie scheduled visits)
- measurement *errors* (ie biological variation)
- potentially *informatively* missing values

**Longitudinal trajectories:**

$$Y^*_{i(l)}(t_{ij}) = \beta_{0(l)} + b_{i0(l)} + (\beta_{1(l)} + b_{i1(l)})t_{ij} + \beta_{g(l)}SNP_i + \beta_{h(l)}H_{i(l)}$$

- *latent* and continuous processes
- various *shapes* (linear/non-linear)
- *smoothed*, account for *measurement* errors
- account for *shared* and/or trait-*specific* covariates
- $b_i \sim N_{2L}(0, D)$: subject-specific deviation from *average* trajectories
- $D$: dependencies *within/between* $L$ traits

## Time-dependent association structures

$$W_{i(k)}(t) = \sum_{l=1}^{L_k} \alpha_{l(k)} f_{l(k)}(Y^*_{i(l)}(t))$$

- account for the *indirect* SNP effects induced by each QT $l$
- $\alpha_{l(k)}$ association between each pair of traits $l$ & $k$
- $f_{l(k)}(Y^*_{i(l)}(t))$, *functional dependence* between traits $l$ & $k$, based on prior knowledge

## Proportional Hazards frailty sub-model for $K$ time-to-event traits

$$\lambda_{i(k)}(t) = \lambda_{0(k)}(t) \times exp\{\gamma_{g(k)}SNP_i + W_{i(k)}(t) + \gamma_{v(k)}V_{i(k)} + u_i\}$$

- possibly *informatively* censored
- $\gamma_{g(k)}$: *direct* SNP effect (other than via $W_{i(k)}(t)$)
- accounts for *shared* and/or trait-*specific* covariates
- $u_i \perp b_i$, subject-specific, accounts for *unexplained* dependencies

---

SNP associations with each time-to-event trait $k$ and each longitudinal trait $l$ ($\alpha_{l(k)} \neq 0$):

a. Indirect: $\beta_{g(l)} \neq 0$ & $\gamma_{g(k)} = 0$
b. Direct: $\beta_{g(l)} = 0$ & $\gamma_{g(k)} \neq 0$
c. Direct & Indirect: $\beta_{g(l)} \neq 0$ & $\gamma_{g(k)} \neq 0$
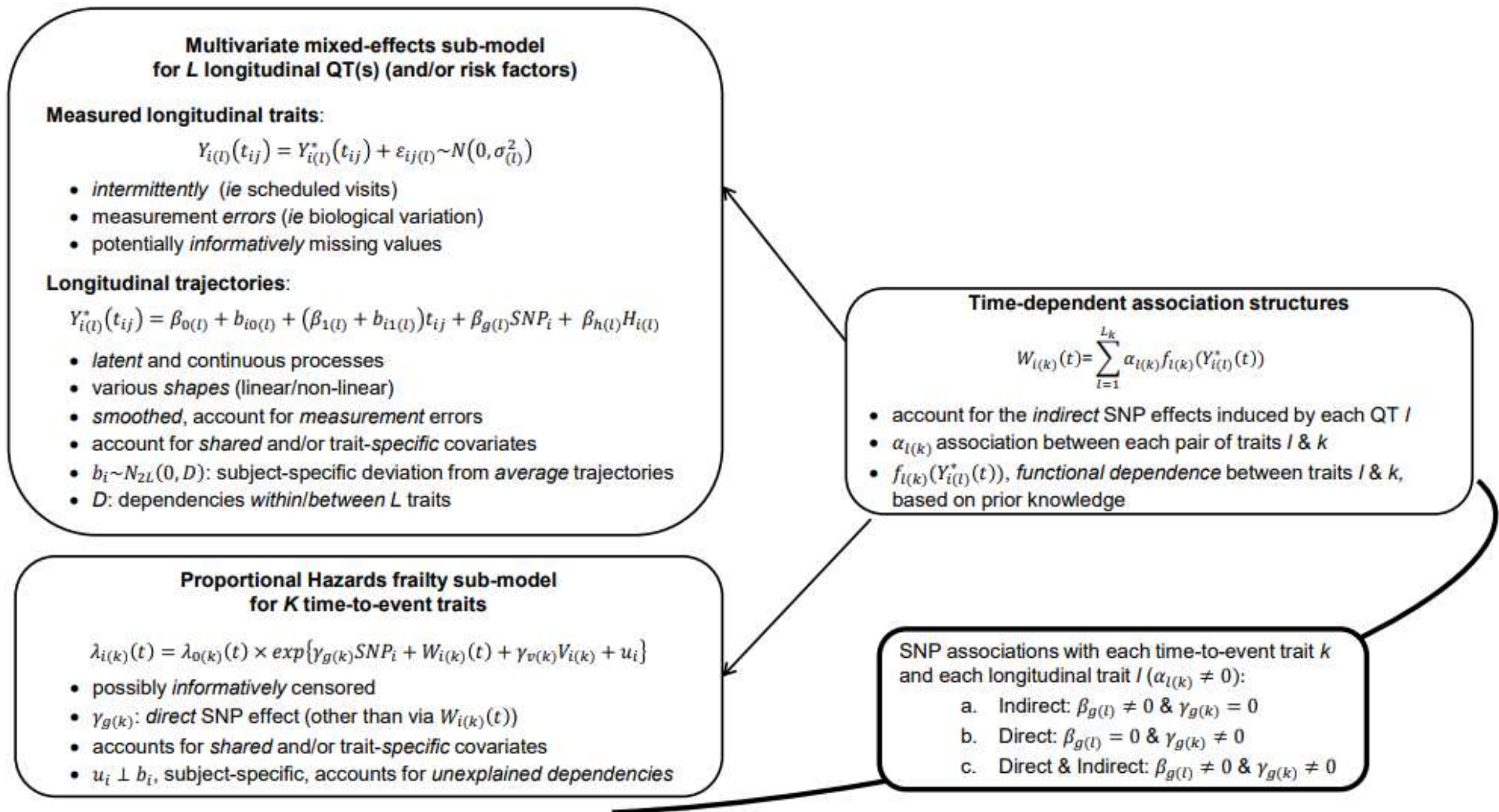
**Figure 1**. Proposed joint modelling approach for characterization of complex genetic architecture of multiple disease progression
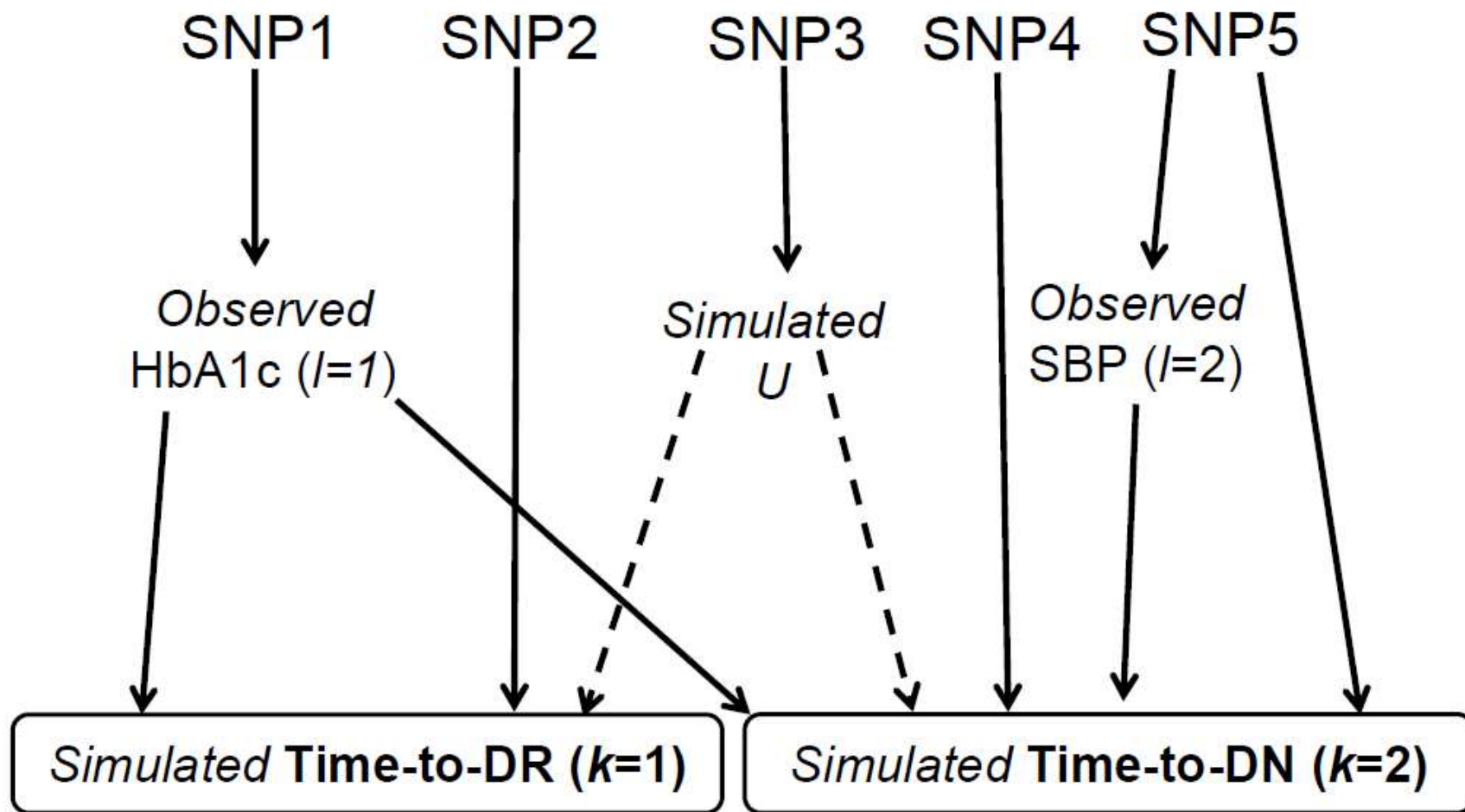
**Figure 2. Realistic DCCT-data-based *causal* genetic scenario**

The effect of gender on SBP (noted as $H_{i(2)}$), and effects of T1D duration at baseline (noted as $V_{i(k)}$) on both time-to-T1DC ($k=1, 2$) traits are not represented in this figure, but are included in the data generating model, see **Supplementary Information SI-1** for details. We simulated $R=1000$ replicates of $N=667$ DCCT individuals with $M=5$ causal variants and $K=2$ time-to-event traits simulated under this causal genetic scenario and $R=1000$ replicates of $M=5$ SNPs (with same MAF as the causal ones) under a global null genetic scenario where none of the SNPs is associated with any traits.
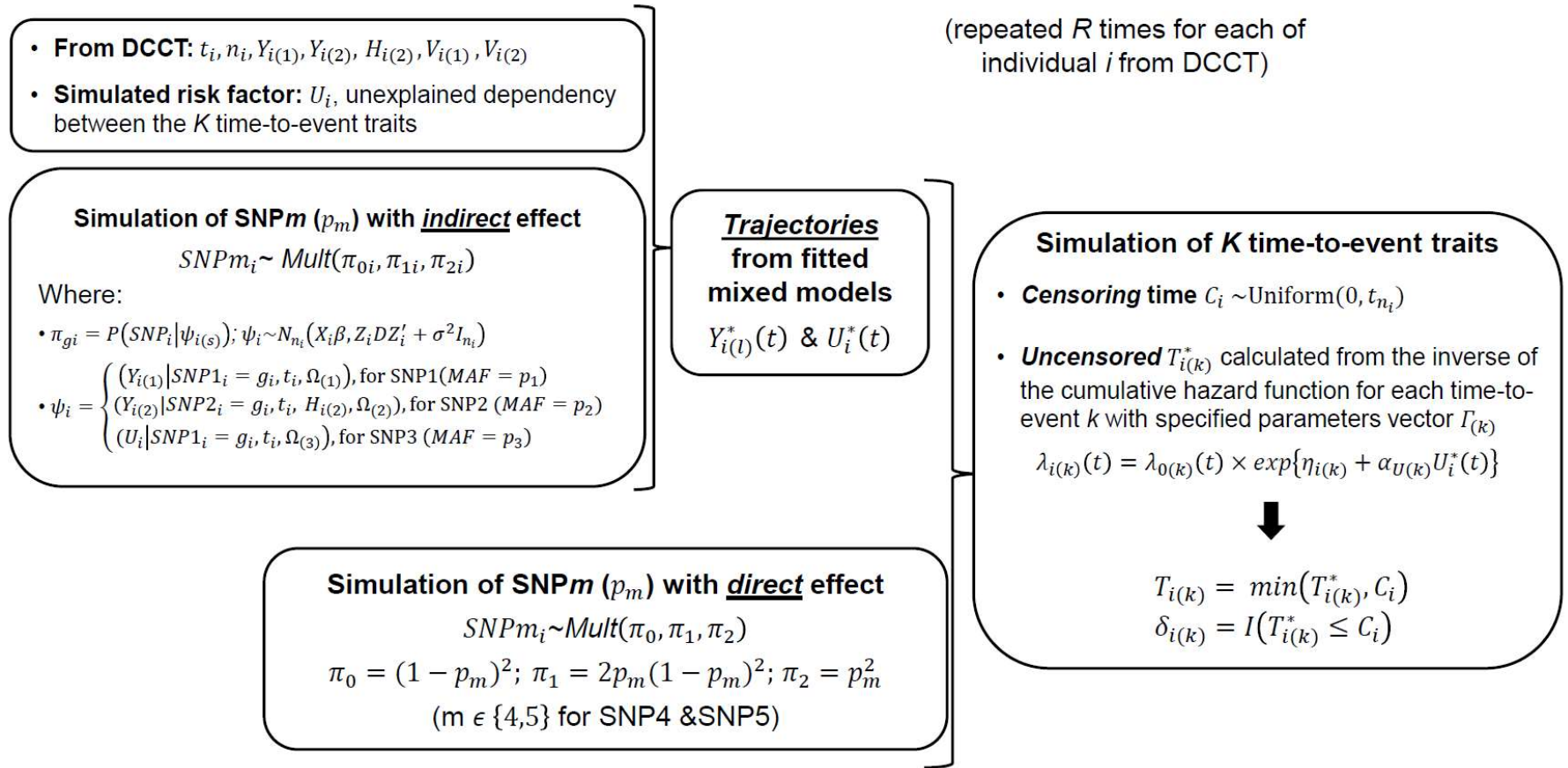
**Figure 3.** Illustration of the procedure developed for DCCT-based simulation study under the scenario from Figure 2

$\Omega_{(1)}$, $\Omega_{(2)}$ and $\Omega_{(U)}$ are specified parameter values for the longitudinal trait models, and $\Gamma_{(1)}$, $\Gamma_{(2)}$ for the time-to-event trait models, $p$ is the specified vector of minor allele frequencies vector of the $M$ SNPs. Parameters for the causal genetic scenario in **Figure 2** are summarized in **Table 2** and in **Table S1**. The DCCT-data-based simulation algorithm uses longitudinal trait/baseline covariates from DCCT and simulated genetic data and time-to-event traits. SNPs with indirect effects are simulated from a multinomial distribution with conditional genotype probabilities $(\pi_{0i}, \pi_{1i}, \pi_{2i})$ using longitudinal trait values for individual $i$. Here, $X_i$ and $Z_i$ denote the design matrices of the fixed and random effects of the mixed model assumed for each longitudinal trait and $D$ is the specified covariance matrix for the random effects. SNPs with direct effects are simulated from the population probabilities, that depend on the MAF. Each time-to-event trait $k$ is simulated by calculating the inverse of the cumulative specified hazard function using the Brent univariate root-finding method[34,35]. For DR, $\eta_{i(1)} = \gamma_{g(1)} SNP2_i + \alpha_{1(1)} Y^*_{i(1)}(t) + \gamma_{v(1)} V_{i(1)}$ and for DN, $\eta_{i(2)} = \gamma_{g(2)} SNP4_i + \gamma'_{g(2)} SNP5_i + \sum_{l=1}^{2} \alpha_{l(1)} Y^*_{i(l)}(t) + \gamma_{v(2)} V_{i(2)}$. We further added the effect of the longitudinal trajectory $U^*_i(t)$ to each $\eta_{i(k)}$ in the hazard function of each trait $k$ to induce some unexplained dependencies between the simulated time-to-event traits. Details of the DCCT-based simulation procedure are presented in **Supplementary Information SI-1**.
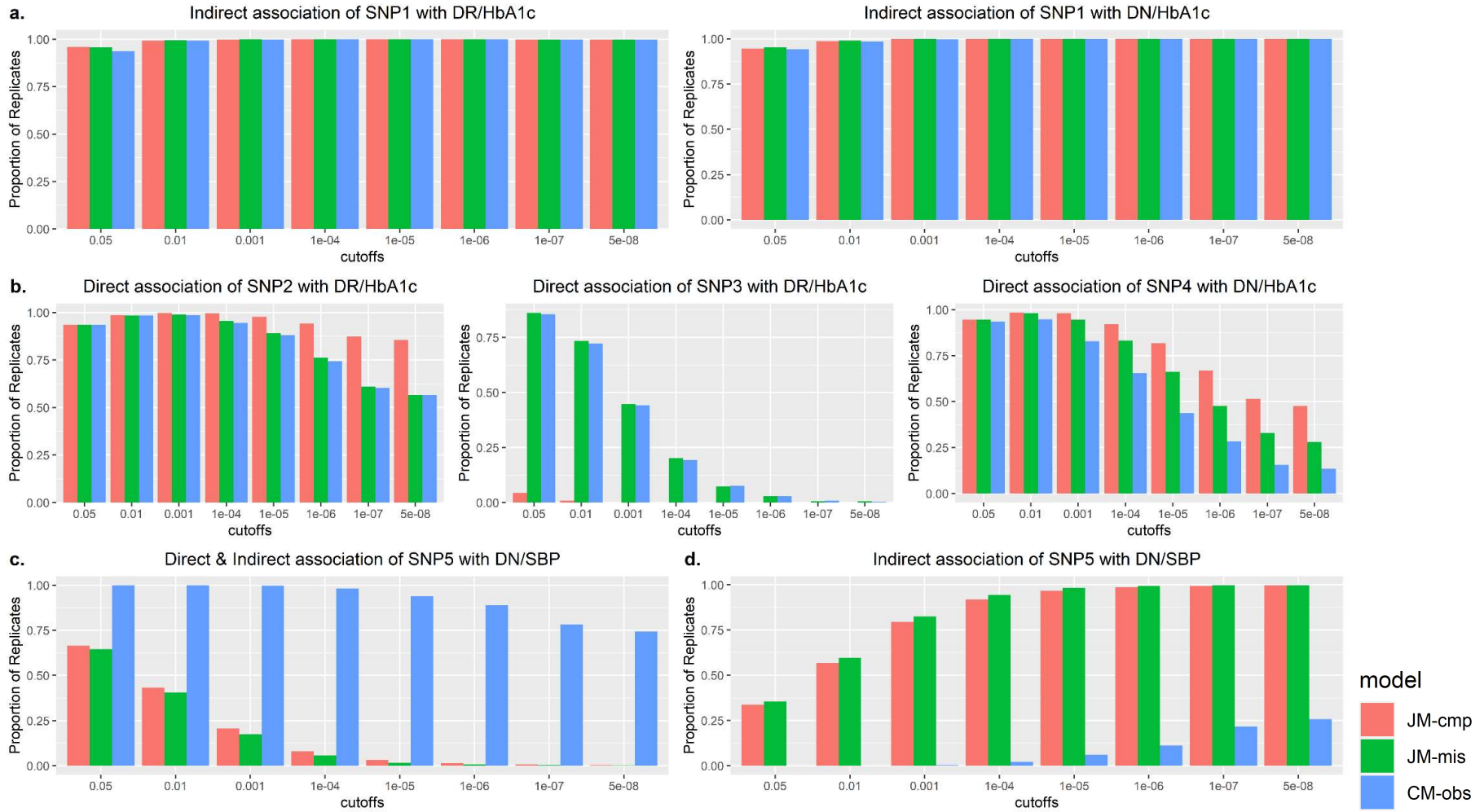
**Figure 4**. **Performance of the testing procedure to recover underlying generating *causal* relationships from Figure 2, assessed using *R*=1000 replicates of *N*=667 DCCT subjects.**

The *Y* axis represents the proportion of replicates that accurately classify each causal SNP as: **(a.)** indirect; **(b.)** direct and **(c.)** direct and indirect causal associations for SNP1, SNP2, SNP4 and SNP5. SNP3 is tested as a direct association because the longitudinal risk factor *U* is assumed unmeasured. Results obtained for direct association of SNP3 and SNP4 with DN/SBP are similar to those presented for DN/HbA1c (**Figure S4**) and are not shown here. Plot **(d.)** shows that SNP5 is misclassified as an indirect association in a larger number of replicates compared to **(c.)** by the joint models compared to CM-obs. This misclassification rate tends to increase with decreasing $P^*$. Classification rates of all these SNPs with other time-to-event/longitudinal trait pairs are presented in **Figure S4**.
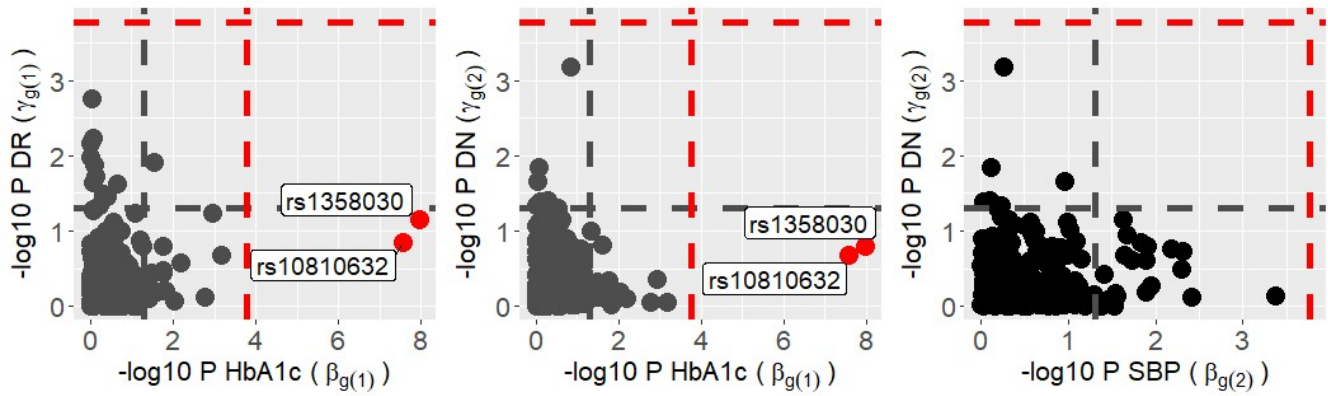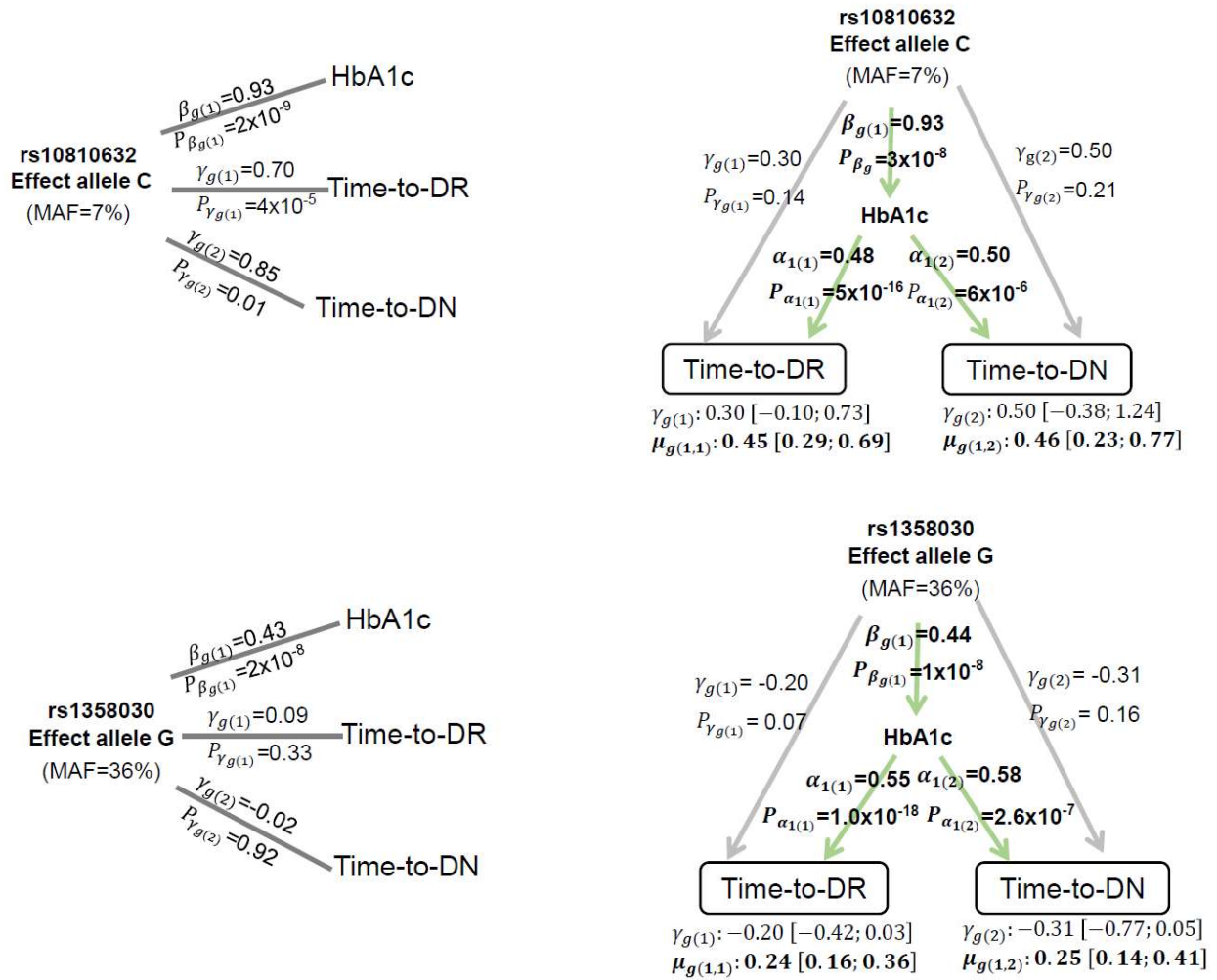
**Figure 5. Classification of direct and/or indirect SNP associations in the DCCT Genetics Study data.**

**(a.)** Scatter plots of the P-values (-log10) for tests of $\beta_{g(l)}$ ($X$ axis) and $\gamma_{g(k)}$ ($Y$ axis) for DR/HbA1c, DN/HbA1c and DN/ SBP trait pairs. Significance levels $P^* = 1.7\text{x}10^{-4}$ and $P^* = 0.05$ are indicated by red and grey horizontal and vertical dashed lines. **(b.)** Association results for the two SNPs that exibit indirect association with DR/DN outcomes at $P^* = 1.7\text{x}10^{-4}$. Left panel present results from separate analysis of each trait; and right panel presents results from the joint model with bootstrap 95% confidence intervals for the direct and indirect SNP effects. Results are presented using the time-weighted cumulative HbA1c effects on T1DC traits.

|  | | SNP association with the longitudinal risk factor $l$ | |
|  | | $P_{\beta_{g(l)}} \leq P^*_{\beta_g}$ | $P_{\beta_{g(l)}} > P^*_{\beta_g}$ |
| SNP association with the time-to-event trait $k$ | $P_{\gamma_{g(k)}} \leq P^*_{\gamma_g}$ | Direct & Indirect | Direct |
|  | $P_{\gamma_{g(k)}} > P^*_{\gamma_g}$ | Indirect | No association |

$P_{\beta_{g(l)}}$ and $P_{\gamma_{g(k)}}$ are the *P-values* from Wald test (1df) for SNP effects $\beta_{g(l)}$ and $\gamma_{g(k)}$.

$P^*_{\beta_g}$ and $P^*_{\gamma_g}$ are the significance thresholds (see methods for details). Note that thorough the paper, we use $P^* = P^*_{\beta_g} = P^*_{\gamma_g}$, but different levels can be specified for $P^*_{\beta_g}$ and $P^*_{\gamma_g}$.

**Table 1**. **Hypothesis testing procedure to classify a SNP as having a direct and/or an indirect association with a time-to-event trait *k* and an associated longitudinal risk factor *l*.**

| Causal SNPs | Type of SNP association | Specified parameters[1] | Empirical power (separate analysis of each trait)[2] | | | | |
|---|---|---|---|---|---|---|---|
| | | | Trait (*Tested parameter*) | $P^* =0.05$ | $P^* =0.001$ | $P^* =10^{-5}$ | $P^* =5\times10^{-8}$ |
| SNP1 (MAF=0.30) | *Indirect* association with both T1DC traits via HbA1c | $\beta_{g(1)} = 0.70$ | HbA1c ($\beta_{g(1)}$) | 100% | 100% | 100% | 99.8% |
| | | $\alpha_{1(1)} = 0.20$ | DR ($\gamma_{g(1)}$) | 20.4% | 1.2% | $<10^{-4}$ | $<10^{-4}$ |
| | | $\mu_{g(1,k)} = 0.14$ | DN ($\gamma_{g(2)}$) | 8.70% | 0.3% | $<10^{-4}$ | $<10^{-4}$ |
| SNP2 (MAF=0.10) | *Direct* association with DR | $\gamma_{g(1)} = 0.80$ | DR ($\gamma_{g(1)}$) | 100% | 98.3% | 87.4% | 58.2% |
| SNP3 (MAF=0.40) | *Indirect* association with both T1DC traits via *U* | $\beta_{g(U)} = 0.80$ | DR ($\gamma_{g(1)}$) | 87.5% | 42.8% | 7.6% | 0.9% |
| | | $\alpha_{1(U)} = 0.40$ $\mu_{g(1,k)} = 0.36$ | DN ($\gamma_{g(2)}$) | 32.6% | 2.6% | $<10^{-3}$ | $<10^{-3}$ |
| SNP4 (MAF=0.30) | *Direct* association with DN | $\gamma_{g(2)} = 0.70$ | DN ($\gamma_{g(2)}$) | 94.1% | 56.5% | 17.3% | 4.0% |
| SNP5 (MAF=0.20) | *Direct* and *indirect* association with DN via SBP | $\beta_{g(2)} = 7.00$ | SBP ($\beta_{g(2)}$) | 100% | 100% | 100% | 100% |
| | | $\alpha_{2(2)} = 0.20$ $\mu_{g(2,2)} = 1.40$ $\gamma_{g(2)} = 0.40$ | DN ($\gamma_{g(2)}$) | 100% | 100% | 99.8% | 98.0% |

[1]$\beta_{g(l)}$ is the SNP effect on the longitudinal trait *l*, $\gamma_{g(k)}$ is the direct SNP effect on the time-to-event trait *k*; $\alpha_{l(k)}$ is the effect of each longitudinal trajectory for trait *l* on the time-to-event trait *k*. $\mu_{g(l,k)}$ is the *indirect* SNP effect, calculated as $\mu_{g(l,k)} = \beta_{g(l)}\alpha_{l(k)}$.

[2]For each scenario of direct/indirect SNP association, we present the empirical power to detect each SNP using *separate* analysis of each trait (linear mixed model for each longitudinal trait, Cox PH model for each time-to-event trait); as typically used for SNP discovery. Models are adjusted for all non-SNP baseline covariates, except *U*, as specified to simulate the data. The empirical power is calculated as the proportion of replicates in the causal scenario from Figure 2 with a significant SNP association detected at specified levels of $P^*$. As expected, SNP1, SNP3 and SNP5 appear associated with the T1DC outcomes due to their indirect effects induced by longitudinal risk factors absents from the time-to-event models. For comparisons, empirical power of the same tests based on parameters from the joint model are presented in **Table S2**.

**Table 2**. **Simulated scenarios of direct and/or indirect SNP associations, assessed using *R*=1000 replicates of *N*=667 DCCT subjects simulated under the *causal* genetic scenario.** This table presents parameters specified for SNPs and trajectory effects on time-to-event traits for the DCCT-based simulation study and illustrates how each type of SNP would be detected in GWAS discovery analysis based on separate analysis of each trait.

| SNPs[1] | Trait pairs | $P^* = 0.05$ | | | $P^* = 0.01$ | | |
|---|---|---|---|---|---|---|---|
| | | JM-cmp | JM-mis | CM-obs | JM-cmp | JM-mis | CM-obs |
| **Indirect association** | | | | | | | |
| SNP1 | DR/HbA1c | 0.049 | 0.049 | 0.500 | 0.015 | 0.015 | 0.015 |
| | DN/HbA1c | 0.049 | 0.048 | 0.048 | 0.015 | 0.015 | 0.015 |
| **Direct association** | | | | | | | |
| SNP2 | DR/HbA1c | 0.058 | 0.045 | 0.045 | 0.009 | 0.009 | 0.007 |
| SNP3 | DR/HbA1c | 0.041 | 0.044 | 0.041 | 0.008 | 0.006 | 0.008 |
| | DN/HbA1c | 0.041 | 0.037 | 0.042 | 0.012 | 0.008 | 0.008 |
| | DR/HbA1c | 0.042 | 0.034 | 0.036 | 0.012 | 0.008 | 0.009 |
| SNP4 | DN/HbA1c | 0.046 | 0.044 | 0.048 | 0.007 | 0.009 | 0.005 |
| | DN/SBP | 0.047 | 0.044 | 0.042 | 0.008 | 0.009 | 0.005 |
| **Direct and Indirect association** | | | | | | | |
| SNP5 | DN/SBP | <0.001 | 0.002 | 0.003 | <0.001 | <0.001 | <0.001 |

[1]Each SNP is tested with the same direct/indirect association tests as applied to the causal SNPs, except SNP3 that is tested as a direct association because the longitudinal risk factor $U$ is assumed unmeasured. Classification rates for all other possible QT/time-to-event trait pairs are presented in **Figure S3**.

[2] JM-cmp denotes the complete joint model fitted for each SNP separately, including all non-genetic covariates used for the data simulation (including $U$).

**Table 3**. Classification rates of the procedure for each SNP under the *global null* genetic scenario, assessed using $R$=1000 replicates of $N$=667 DCCT subjects.