Check for updates

Received: 12 March 2021 | Revised: 12 August 2021 | Accepted: 23 August 2021

DOI: 10.1002/sim.9190

RESEARCH ARTICLE

Statistics in Medicine WILEY

Frailty modeling under a selective sampling protocol: an application to type 1 diabetes related autoantibodies

Jaakko Nevalainen[1] | **Somnath Datta**[2] | **Jorma Toppari**[3,4] | **Jorma Ilonen**[3] | **Heikki Hyöty**[5] | **Riitta Veijola**[6] | **Mikael Knip**[7] | **Suvi M. Virtanen**[1,8,9]

[1]Health Sciences, Faculty of Social Sciences, Tampere University, Tampere, Finland

[2]Department of Biostatistics, University of Florida, Gainesville, Florida

[3]Institute of Biomedicine, University of Turku, Turku, Finland

[4]Department of Pediatrics, Turku University Hospital, Turku, Finland

[5]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

[6]Department of Pediatrics, Oulu University Hospital and University of Oulu, Oulu, Finland

[7]Children's Hospital, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

[8]Public Health and Welfare Department, Finnish Institute for Health and Welfare, Helsinki, Finland

[9]Research, Development and Innovation Centre, and Center for Child Health Research, Tampere University and University Hospital, Tampere, Finland

Correspondence
Jaakko Nevalainen, Arvo Ylpön katu 34, 33014 Tampere University, Tampere, Finland.
Email: jaakko.nevalainen@tuni.fi

Funding information
Foundation for the National Institutes of Health, Grant/Award Numbers: 1R03DE026757-01A1, 5R03DE026757-02

In studies following selective sampling protocols for secondary outcomes, conventional analyses regarding their appearance could provide misguided information. In the large type 1 diabetes prevention and prediction (DIPP) cohort study monitoring type 1 diabetes-associated autoantibodies, we propose to model their appearance via a multivariate frailty model, which incorporates a correlation component that is important for unbiased estimation of the baseline hazards under the selective sampling mechanism. As further advantages, the frailty model allows for systematic evaluation of the association and the differences in regression parameters among the autoantibodies. We demonstrate the properties of the model by a simulation study and the analysis of the autoantibodies and their association with background factors in the DIPP study, in which we found that high genetic risk is associated with the appearance of all the autoantibodies, whereas the association with sex and urban municipality was evident for IA-2A and IAA autoantibodies.

KEYWORDS
correlated data, incomplete data, multivariate survival analysis, type 1 diabetes

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

The incidence of type 1 diabetes (T1D) has increased worldwide since the 1950s[1] and in the 2000s leveled off in some high-incidence countries such as Finland with the highest incidence of T1D in the world among children younger than 15 years.[2]

Emerging clinical T1D is commonly preceded by the appearance of autoantibodies. The progression from the presence of autoantibodies to T1D has been explicitly modeled in an additive joint modeling framework.[3,4] Recent evidence suggests varying sensitivity of the autoantibodies to different exposures as well as heterogeneity of T1D as a clinical disease.[5-7] The disease process may differ according to the first-appearing autoantibody and/or by age of the child.[8,9] Autoantibody-specific associations have been shown for some dietary and infectious exposures.[5,10,11] The role of infant diet in the etiology of T1D has been studied quite extensively,[12-15] while much less is known about significance of diet later during childhood.[16-18]

The availability of the autoantibody measurements can, however, make the conduct and interpretation of autoantibody-specific analyses challenging. For example, the primary screening tool islet cell autoantibodies (ICA) in the type 1 diabetes prevention and prediction study, has been monitored systematically throughout the follow-up period up to 15 years of age during the time period 1994 to 2004. Importantly, the other three autoantibodies (insulin autoantibodies (IAA), glutamic acid decarboxylase autoantibodies (GADA) and islet antigen 2 autoantibodies (IA-2A)) were analyzed from available and subsequent samples *only if* the child seroconverted to ICA positivity. In addition, a subcohort within the study had all their autoantibodies measured (Figure 1).

The selective monitoring of other than ICA autoantibodies raises the question whether the conventional analyses regarding appearance of the secondary autoantibodies would remain statistically valid and efficient. In the motivating setting described, we would expect that a naive marginal (autoantibody-specific) analysis of the secondary autoantibodies separately on the available data would give biased estimates of the proportion of children with positivity to them and could possibly cause bias in the estimates of regression coefficients of prognostic variables.

We propose to model the appearance of autoantibodies via a multivariate frailty model, which incorporates the association parameters for the four autoantibodies. The correlation component is important for unbiased estimation of the baseline hazards under selective sampling. As further advantages, the multivariate frailty model allows for systematic evaluation of the association and the differences in regression parameters among the autoantibodies. We demonstrate the properties of the model by a simulation study and the analysis of the autoantibodies and their association with background factors in the DIPP study.

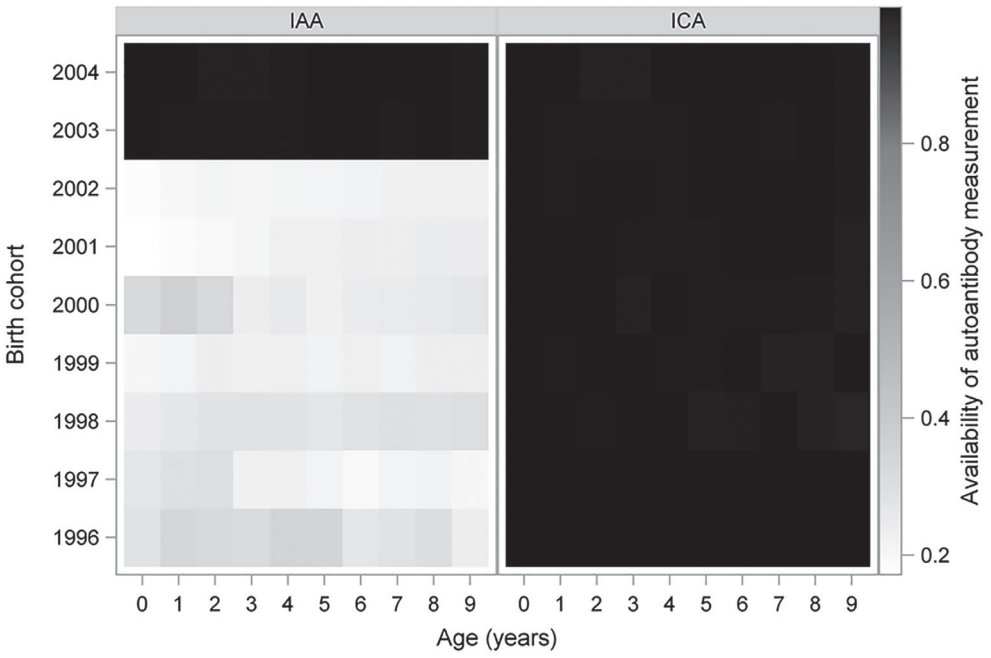

FIGURE 1 Proportion of children with at least annual IAA or ICA measurement by age and by birth cohort

2 | METHODS

2.1 | Autoantibody and background data

The Finnish prospective population-based type 1 diabetes prediction and prevention (DIPP) birth cohort study screened newborn infants born in the Tampere and Oulu University Hospitals for HLA-DQB1-conferred susceptibility to T1D using cord blood samples between September 1996 and September 2004.[19] Infants carrying increased genetic susceptibility (HLA-DQB1*02/0302 heterozygous and DQB1*0302/x-positive subjects [x stands for homozygosity or a neutral allele]) were monitored for islet autoimmunity and their blood samples are collected by venepuncture at the ages of 3, 6, 9, 12, 18, and 24 months and subsequently at 12-month intervals up to the age of 15 years. In the analyses we used actualized sampling ages reflecting some variation around the scheduled sampling times and missed visits. Islet cell autoantibodies (ICA) were used as the primary screening tool for beta cell autoimmunity. When a child seroconverted to positivity for ICA for the first time, all available retrospective and subsequent samples were analyzed for IAA, GADA, and IA-2A. From that point onwards, blood samples were collected in three-month intervals. For a subcohort, all autoantibody measurements from birth was collected. This subcohort consists of children born in 2003 to 2004. The availability of autoantibody data is illustrated for IAA and ICA in Figure 1.

HLA-DQ was genotyped using panels of sequence-specific oligonucleotide probes. Genotypes HLA-DQB1(*02/*03:02) represent "high" and HLA-DQB1*03:02/x (x not *02, *03:01, or *06:02) "moderate" risk for type 1 diabetes. INS-23 A/T (rs689) and PTPN22 1858C/T (rs2476601) were genotyped using the Sequenom platform, or using the TaqMan SNP genotyping array, with INS AA and PTPN22 TT/CT regarded as the risk genotypes for type 1 diabetes.[8]

Information on offspring sex was collected with a questionnaire after delivery. Urbanization level of home municipality was categorized according to Statistics Finland guidelines as rural, semiurban, and urban.

2.2 | Definition of the outcome and justification for the frailty model

We consider the seroconversion as a four-variate time-event-outcome; each child contributes either the timing of the first positive measurement for each of the four autoantibodies, or a right-censored observation of the autoantibody. As a child can become positive for several autoantibodies, the setting is a multivariate setting rather than a competing risks setting. To be precise, the times of the first autoantibody positivity are interval-censored.[20] Earlier, Park[21] (nonparametric approach) and Zhang et al[22] (latent variable approach) considered dependent censoring with interval-censored data. Our setting fits those approaches poorly, because the censoring times were a priori fixed by the protocol, and only the decision of whether to analyze or not to analyze the samples from those times depended on the presence of ICA autoantibody. When we assume that the monitoring algorithm is the source of incomplete data the problem can be framed as a missing value problem. Recall that ICA was analyzed throughout the follow-up period, but absence or presence of data on the three other autoantibodies was determined by the positivity of ICA. In such a missing at random setting, an analysis based on an appropriate observed data likelihood alone can provide valid inference.[23] In the next section, we propose a frailty model for the joint likelihood of the four autoantibodies to deal with the unusual study design.

2.3 | Multivariate frailty model for autoantibodies

To approximate the baseline hazard of seroconversion, we assume a piecewise linear baseline log-hazard

$$\log h_j(t) = \alpha_j + \beta_{1j}t + \beta_{2j}(t-\kappa_1)_+ + \beta_{3j}(t-\kappa_2)_+,$$

in which $j \in \mathcal{J} = \{\text{ICA}, \text{IAA}, \text{GADA}, \text{IA-2A}\}$ indicates the autoantibody to be modeled. We used two knots at $\kappa_1 = 2, \kappa_2 = 4$ in the construction of the baseline hazard, and we note that the model was not sensitive to the position of the knots. More flexible estimation would be possible by using splines,[24] but that is not our main focus here. Individual autoantibody-specific hazards were modeled with

$$h_{ij}(t) = h_j(t)\exp\left\{\theta' x_i + g_j u_i\right\}; i = 1, \ldots, n,$$

where $u_i \sim N(0, \sigma^2)$ represents individual "shared frailty" to the autoantibodies, and we assume that the autoantibodies are independent given u_i. Frailty variance can be interpreted to represent whatever the autoantibodies have in common, for example, sharing the same risk factors or being involved in the same biological process.

The parameters g_j are constrained with $g_{ICA} = 1$ and $(1/3)\sum_{j\in\mathcal{J}\setminus ICA} |g_j| = 1$ so that the model is identifiable. Note that the formulation allows direct and inverse associations between two autoantibodies, as well as different magnitudes of shared frailties with ICA. We chose ICA as the reference for the frailty term as it was completely observed.

Let T_{ij}^L denote the last time instance where the ith child was measured as not positive for autoantibody j, and T_{ij}^U the first time the child was positive to that autoantibody. If a child was never observed to be positive, $T_{ij}^U = \infty$, which is a right-censored observation. The likelihood contribution of an individual with interval-censored event times (T_{ij}^L, T_{ij}^U) is

$$L_i(\Theta; u_i) = \left[S_{i,ICA}(T_{i,ICA}^L) - S_{i,ICA}(T_{i,ICA}^U)\right] \prod_{j\in\mathcal{J}\setminus \text{ICA}} \left[S_{ij}(T_{ij}^L) - S_{ij}(T_{ij}^U)\right]^{\delta_i},$$

where $\delta_i = \mathbb{1}\left[\text{child } i \text{ ICA positive } \vee\, i \in \mathcal{S}\right]$, $S_{ij}(t)$ is the survival function of individual i and autoantibody j and $\mathcal{S}$ is the set of subcohort members. The full observed data likelihood,

$$L(\Theta) = \int \prod_{i=1}^{n} L_i(\Theta; u_i) dF(u_i),$$

which can be programmed and optimized in statistical software capable of numerical integration.

2.4 | Simulation setup

A simulation study was conducted to demonstrate the performance of the models and to investigate their statistical properties. For individuals $i = 1, \ldots, n$, frailty $u_i \sim N(0, 1)$ was generated, which converts to a multiplication of the marginal baseline hazard by a factor of $E(e^{u_i}) = 1.649$ compared to a hazard with no frailty. This is because if a random variable $X \sim N(\mu, \sigma^2)$, then $E[\exp(X)] = \exp(\mu + \sigma^2/2)$. Then, at each time t, we generated

$$Y_{ijt} \sim \text{Ber}\left(h_{ij}(t)\Delta t\right),$$

with a grid of 0.1 years. The chosen hazards were:

$$h_{i1}(t) = \exp\left\{-6 + 0.5t - 0.7(t-2)_+ + 0.2x_i + u_i\right\}, \tag{1}$$

$$h_{i2}(t) = \exp\left\{-6 + 0.5t - 0.2(t-2)_+ - 0.8(t-4)_+ - 0.2x_i + u_i\right\}. \tag{2}$$

The time to positivity T_{ij} was defined to be the minimum of t at which $Y_{ij} = 1$ (uncensored observation) or 10 (censored observation), whichever came first. A particular setting with two autoantibodies was studied so that the data were complete for the first autoantibody, but a proportion of the data was set to missing for the second autoantibody. To mimic the actual conditional data collection protocol of the study, these data were available when either (i) the first autoantibody was positive *or* (ii) the individual fell into a $100 \times \pi\%$ random sample. The parameter π was set at .1,.5, and 1.0, and $\theta = 0.2$ to represent the effect of x_i. The number of individuals was $n = 5000$ in the simulation and the model was fitted with maximum likelihood in the SAS/NLMIXED procedure. To improve the estimation of the frailty variance parameter, we estimated its logarithm.

The performance of two models was compared. Both had the correct model structure (1)-(2) and they were identical with the exception that the frailty term was absent from the other, that is, autoantibodies were modeled independently. For benchmarking purposes, the models were fitted on all available data and on the random sample with complete data.

3 | RESULTS

3.1 | Simulation study

Table 1 shows that the estimates from the multivariate frailty model on all data available hit their target values for all the parameters. For the first autoantibody, which was completely observed, the estimates and their standard errors remained similar regardless of the π-parameter controlling the amount of data for the second autoantibody. A slight improvement for coverage probabilities from $\pi = 0.1$ to $\pi = 1.0$ was observed. For the second autoantibody, the estimators behaved also well: they showed no meaningful bias. Standard errors increased as π decreased since less information was available for that autoantibody. Coverage probabilities remained at their nominal levels. While point estimation of the frailty variance parameter was overall without bias, interval estimation improved from 0.88 coverage to 0.95 coverage when the complete data subset was larger than 10% of the individuals.

When the frailty term was omitted from the model, a marked bias was observed in the α-parameter. For the first autoantibody, as well as for the $\pi = 1.0$ case, this is a result of a misspecification of the model rather than a result of the selective sampling. When the estimates were compared to the reference value of $\alpha^* = \alpha + \log 1.649$ from the marginal hazard $h_j^*(t) = E_U\left[h_{ij}(t)\right]$, that is, -5.5, the bias disappeared in the cases of complete data. However, the bias remained when $\pi = 0.1, 0.5$ and it was more evident with less data on the second autoantibody. Interestingly, other estimators including the estimators of the regression coefficients suffered little from the offset bias in the baseline hazard and from the omission of the frailty parameter.

Table 1 also lists the results obtained from fitting the models on the random sample only. The results are similar to identical analyses with all available data with the exception of the standard error estimates, which were notably larger. These indicate that a much higher precision can be achieved with the use of all available data.

Tables S1 and S2 show simulation results supporting these observations for a smaller data set, and for the case of four antibodies.

3.2 | Autoantibodies in the DIPP study

Among 6081 children during a 10-year follow-up from birth, GADA positivity was observed for 285 children, IA-2A for 201, IAA for 284 and ICA for 862. Figure 2 shows their coexistence. Due to the sampling scheme, ICA was by far the most common autoantibody to appear alone, and together with other autoantibodies. For example, IAA and IA-2A autoantibodies did not appear together without either GADA or ICA being present as well. Seroconversion process was considered irreversible.

We analyzed the available data, which included all children in the frailty model, and in autoantibody-specific analyses omitting frailty terms, only the children who had at least one interval-censored or right-censored measurement on that particular autoantibody.

3.2.1 | Estimated hazards and survival functions

Autoantibody-specific estimates of the marginal log hazards, without the inclusion of covariates, were:

$$\begin{cases} -4.39 + 0.55t - 0.83(t-2)_+ - 0.05(t-4)_+, & \text{for GADA;} \\ -5.03 + 0.78t - 1.19(t-2)_+ + 0.16(t-4)_+, & \text{for IA-2A;} \\ -3.71 + 0.23t - 0.76(t-2)_+ + 0.35(t-4)_+, & \text{for IAA; and} \\ -4.32 + 0.52t - 0.74(t-2)_+ + 0.15(t-4)_+, & \text{for ICA.} \end{cases}$$

The four-variate frailty model gives the log hazards for a child with $u_i = 0$ of

$$\begin{cases} -8.51 + 1.11t - 1.18(t-2)_+ - 0.08(t-4)_+, & \text{for GADA;} \\ -11.44 + 1.66t - 1.67(t-2)_+ - 0.18(t-4)_+, & \text{for IA-2A;} \\ -8.27 + 0.95t - 1.35(t-2)_+ + 0.27(t-4)_+, & \text{for IAA; and} \\ -8.51 + 1.34t - 1.24(t-2)_+ - 0.02(t-4)_+, & \text{for ICA.} \end{cases}$$

TABLE 1 Estimation results from the simulation study with 1000 runs

Data and model used	π	Parameter	Autoantibody 1				Autoantibody 2			
			Mean	Bias	SE	Coverage	Mean	Bias	SE	Coverage
All data available	0.1	α	−5.502	0.498	0.249	0.471	−5.222	0.778	0.609	0.657
No frailty parameter		β_1	0.483	−0.017	0.159	0.943	0.531	0.031	0.371	0.959
		β_2	−0.280	−0.080	0.514	0.962				
		β_3	0.011	0.011	0.140	0.942	−0.770	0.030	0.283	0.941
		θ	0.192	−0.008	0.115	0.954	−0.156	0.044	0.239	0.939
	0.5	α	−5.502	0.498	0.249	0.471	−5.466	0.534	0.364	0.629
		β_1	0.483	−0.017	0.159	0.943	0.507	0.007	0.227	0.955
		β_2	−0.695	0.005	0.234	0.945	−0.243	−0.043	0.318	0.950
		β_3	0.011	0.011	0.140	0.942	−0.772	0.028	0.172	0.947
		θ	0.192	−0.008	0.115	0.954	−0.177	0.023	0.144	0.946
	1.0	α	−5.502	0.498	0.249	0.471	−5.521	0.479	0.277	0.549
		β_1	0.483	−0.017	0.159	0.943	0.501	0.001	0.174	0.938
		β_2	−0.695	0.005	0.234	0.945	−0.233	−0.033	0.243	0.935
		β_3	0.011	0.011	0.140	0.942	−0.776	0.024	0.131	0.933
		θ	0.192	−0.008	0.115	0.954	−0.189	0.011	0.110	0.949
All data available	0.1	α	−6.029	−0.029	0.300	0.908	−6.007	−0.007	0.761	0.926
Frailty parameter included		β_1	0.495	−0.005	0.164	0.914	0.493	−0.007	0.392	0.929
		β_2	−0.697	0.003	0.233	0.912	−0.219	−0.019	0.519	0.931
		β_3	0.020	0.020	0.163	0.909	−0.787	0.013	0.316	0.918
		θ	0.202	0.002	0.118	0.924	−0.209	−0.009	0.249	0.920
		σ^2	–	–	–	–	1.024	0.024	0.302	0.886
	0.5	α	−6.006	−0.006	0.293	0.941	−6.015	−0.015	0.390	0.945
		β_1	0.499	−0.001	0.168	0.942	0.509	0.009	0.231	0.946
		β_2	−0.694	0.006	0.246	0.936	−0.214	−0.014	0.325	0.948
		β_3	0.001	0.001	0.142	0.938	−0.793	0.007	0.174	0.942
		θ	0.201	0.001	0.120	0.958	−0.199	0.001	0.152	0.961
		σ^2	–	–	–	–	0.987	−0.013	0.221	0.948
	1.0	α	−5.985	0.015	0.274	0.937	−6.004	−0.004	0.297	0.940
		β_1	0.486	−0.014	0.158	0.945	0.508	0.008	0.176	0.935
		β_2	−0.676	0.024	0.233	0.944	−0.215	−0.015	0.246	0.932
		β_3	−0.005	−0.005	0.139	0.942	−0.791	0.009	0.132	0.939
		θ	0.201	0.001	0.120	0.958	−0.199	0.001	0.116	0.951
		σ^2	–	–	–	–	0.981	−0.019	0.206	0.952
Random sample only	0.1	α	−5.737	0.263	1.112	0.865	−5.769	0.231	1.329	0.885
No frailty parameter		β_1	0.576	0.076	0.652	0.967	0.600	0.100	0.742	0.972
		β_2	−0.790	−0.090	0.872	0.963	−0.321	−0.121	0.931	0.964
		β_3	0.001	0.001	0.463	0.943	−0.811	−0.011	0.436	0.944
		θ	0.177	−0.023	0.390	0.963	−0.197	0.003	0.355	0.946

(Continues)

TABLE 1 (Continued)

Data and model used	π	Parameter	Autoantibody 1 Mean	Bias	SE	Coverage	Autoantibody 2 Mean	Bias	SE	Coverage
	0.5	α	−5.517	0.483	0.350	0.677	−5.546	0.454	0.389	0.734
		β_1	0.485	−0.015	0.224	0.956	0.511	0.011	0.242	0.950
		β_2	−0.696	0.004	0.332	0.943	−0.242	−0.042	0.339	0.951
		β_3	0.010	0.010	0.195	0.942	−0.777	0.023	0.179	0.955
		θ	0.192	−0.008	0.169	0.946	−0.189	0.011	0.149	0.959
Random sample only	0.1	α	−6.263	−0.263	1.188	0.965	−6.282	−0.282	1.349	0.969
Frailty parameter included		β_1	0.592	0.092	0.655	0.969	0.612	0.112	0.733	0.973
		β_2	−0.784	−0.084	0.874	0.961	−0.301	−0.101	0.927	0.966
		β_3	−0.013	−0.013	0.464	0.943	−0.830	−0.030	0.438	0.945
		θ	0.186	−0.014	0.410	0.962	−0.208	−0.008	0.377	0.944
		σ^2	–	–	–	–	1.039	0.039	0.681	0.885
	0.5	α	−5.995	0.005	0.375	0.959	−6.023	−0.023	0.411	0.948
		β_1	0.491	−0.009	0.224	0.959	0.519	0.019	0.243	0.951
		β_2	−0.682	0.018	0.333	0.946	−0.226	−0.026	0.342	0.950
		β_3	−0.004	−0.004	0.195	0.942	−0.792	0.008	0.180	0.956
		θ	0.200	−0.000	0.176	0.948	−0.199	0.001	0.157	0.962
		σ^2	–	–	–	–	0.965	−0.035	0.287	0.952

Note: $\alpha = -6$ is the intercept parameter of the log hazard. $\beta_1, \beta_2, \beta_3$ are the slope parameters of the piecewise linear log hazard with true values at $(0.5, -0.7, 0.0)$ and, $(0.5, -0.2, -0.8)$ for the first and the second autoantibody, respectively. θ is the regression coefficient for the covariate with true value of 0.2 for the first and −0.2 for the second autoantibody.

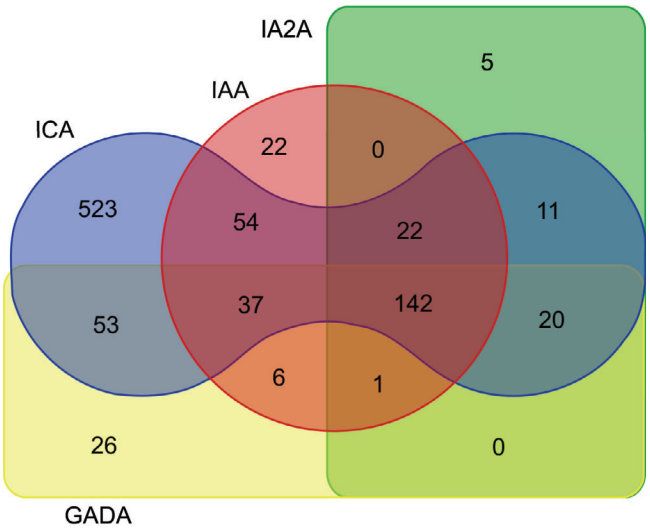

FIGURE 2 Venn diagram on the frequency of appearance of the four autoantibodies during the first 10 years of life among $n = 6081$ children with increased genetic risk of type 1 diabetes. The Venn diagram is based on the layout from http://bioinformatics.psb.ugent.be/webtools/Venn/ [Colour figure can be viewed at wileyonlinelibrary.com]

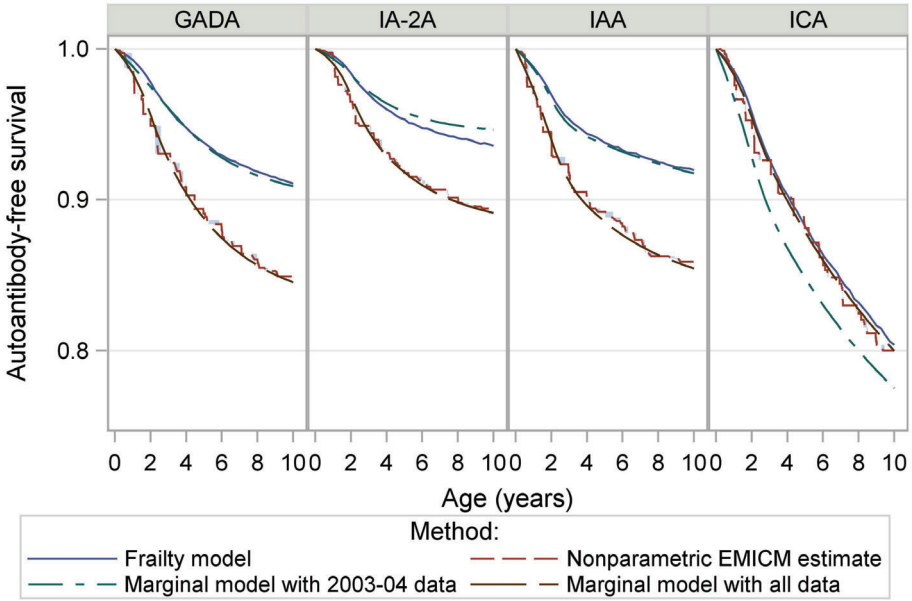

FIGURE 3 Estimates of the autoantibody-free survival functions derived nonparametrically and from two competing models. The survival function of the marginal model fitted on the subcohort data is shown as a benchmark [Colour figure can be viewed at wileyonlinelibrary.com]

These indicate that the appearance of autoantibodies accelerates during the first two years of life, but the hazard rates start to decline thereafter—all the autoantibodies follow the same pattern with some heterogeneity in the changes over time. For example, IA-2A seems to have a steeper increase in the log hazard in the first two years than IAA. Overall, the coefficients of the frailty model indicate steeper changes than those of the marginal model. Note that the intercept terms should not be compared because of the frailty variance.

The frailty variance was estimated as $\hat{\sigma}^2 = 8.66$ with Wald 95% confidence limits from 7.54 to 9.97. Thus, autoantibodies were associated with each other and the frailty model was justified. As $\hat{g}_{IAA} = 0.97$, $\hat{g}_{GADA} = 0.88$ and $\hat{g}_{IA-2A} = 1.16$, the data suggested that the strongest association with ICA was that with IA-2A.

To illustrate how large the differences between the models are in practice, we converted the estimated hazards to marginal survival functions. For the frailty model, the estimated survival function was obtained numerically from $\hat{S}_j^*(t) = E\left\{\exp\left[\int_0^t \hat{h}_{ij}(v)dv\right]\right\}$, where the expectation was taken over the distribution of the frailty term. Figure 3 shows the different estimates. The nonparametric EMICM estimator[25] of the survival function with interval-censored data and available in the Icens package in R[26] does not rely on a model, but provides estimates which agree closely with the survival estimates of the marginal model. Both are biased as they are based on all data, which have been collected with a selective sampling protocol resulting in a particular type of incompleteness. They agree with the survival estimate of the frailty model only for ICA, which was completely measured. For other autoantibodies, they overestimate the incidence of the autoantibodies, because the individuals unlikely to be positive for them (ie, are negative for ICA) tend to have no follow-up data. On other autoantibodies, the frailty model shows remarkably lower appearance probabilities. The frailty model estimates—which are based on all data—are close to the estimate derived from the marginal model fitted on 2003 to 2004 on whom all autoantibodies were measured. The benchmark suggests that the frailty model is performing well in estimating the appearance probabilities, whereas the marginal model and the nonparametric method markedly overestimate them.

3.2.2 | Association to background factors

We then studied the association of the appearance of autoantibodies to genetic risk (high or moderate), sex (female or male), and urban (urban & semiurban or rural) surroundings. The autoantibodies were jointly modeled with autoantibody-specific hazards as

TABLE 2 Fixed parameter estimates from the frailty model

Factor		Estimate	Standard error	*Z*
High genetic risk	ICA	0.540	0.137	3.95
(vs moderate)	IAA	0.696	0.185	3.77
	IA-2A	0.650	0.226	2.87
	GADA	0.881	0.180	4.91
Female	ICA	−0.113	0.113	−1.00
(vs male)	IAA	−0.387	0.163	−2.38
	IA-2A	−0.500	0.198	−2.53
	GADA	−0.305	0.159	−1.91
Urban or semi	ICA	0.265	0.166	1.60
urban municipality	IAA	0.695	0.267	2.60
(vs rural)	IA-2A	0.894	0.339	2.64
	GADA	0.274	0.245	1.12

$$h_{ij}(t) = h_j(t) \exp\left\{\theta_j \mathbb{1}[\text{high genetic risk}_i] + \xi_j \mathbb{1}[\text{female}_i] + \eta_j \mathbb{1}[\text{urban municipality}_i] + g_j u_i\right\}.$$

The estimates of the model parameters are displayed in Table 2. Note that the standard errors of parameters related to the ICA autoantibody were always smaller than for other autoantibodies.

We observed that high genetic risk was strongly associated with appearance of all four autoantibodies. The hazard ratios ranged from 1.72 (95% confidence interval, CI: 1.31, 2.24) for ICA up to 2.41 (1.70, 3.43) for GADA. All findings were clearly significant.

Autoantibodies tended to be more frequent in boys. The strongest association was suggested for IA-2A, with a hazard ratio for of 1.65 (1.12, 2.43) for males compared to females. All the associations with sex were in the same direction but we found clear evidence of it only from IA-2A and IAA autoantibodies (Table 2); the association was nonsignificant for ICA, and borderline significant for GADA.

A similar pattern was observed for urban surroundings: especially IAA and IA-2A seem more common in urban or semiurban than in rural municipalities. Hazard ratios were 2.00 (1.19, 3.38) and 2.44 (1.26, 4.75), respectively.

Compared to the baseline model without explanatory factors, frailty variance decreased from 8.66 to 5.31, a marked 39% decline which suggests that these three factors are shared and important risk factors for seroconversion to autoantibody positivity.

When these results were compared with the estimates of regression coefficients from a simpler model, a marginal model fitted on the subcohort, we observed that the frailty model tended to result in markedly smaller standard errors (Tables 2 and S3). The estimated coefficients for high genetic risk seemed overall smaller for the subcohort. This was possibly due to the smaller precision of estimation: for example, a 95% Wald confidence interval for the coefficient for ICA from the subcohort analysis ranged from −0.049 to 0.551, and it did not exclude the point estimate of 0.540 from the analysis of the full data set. Following this procedure, possibly differing estimates between the two analyses could be the coefficients of female sex for IA-2A. In this case, the full analysis indicated a strong association with sex, but the subcohort analysis provided a nearly null result. Note, however, that there was still considerable overlap in their confidence intervals; the CI was (−0.888, −0.112) from the full analysis and (−0.492, 0.566) for the subcohort analysis. To summarize, the majority of estimates from both analyses pointed into the same direction with some variation in their estimated effect sizes, but the evidence from the frailty model was more conclusive.

The final advantage of using the frailty model is the joint likelihood, which allowed direct testing custom hypotheses of differential effects on different autoantibodies. If $\hat{\theta}_{ICA}, \hat{\theta}_{IAA}, \hat{\theta}_{IA-2A}$ and $\hat{\theta}_{GADA}$ are the estimated regression coefficients, a quadratic form of the three-variate test statistic $(\hat{\theta}_{IAA} - \hat{\theta}_{ICA}, \hat{\theta}_{IA-2A} - \hat{\theta}_{ICA}, \hat{\theta}_{GADA} - \hat{\theta}_{ICA})$ can be used to test whether the association of background factors and autoantibodies are different from one autoantibody to another. The covariance matrix of the parameter estimates was based on the observed information matrix. Here, we observed marginal heterogeneity in the association with sex and municipality ($p = 0.10$; details not shown).

4 | CONCLUDING REMARKS

We proposed to model diabetes-related autoantibodies by a multivariate frailty model, when their monitoring was done according to a nonstandard protocol. We demonstrated by a simulation study that valid inference is obtained. Compared with a marginal (autoantibody-specific) model with these data structures, the approach provides unbiased parameter estimates throughout. In addition, it provides means to study the coexistence of the autoantibodies and their common risk factors, as well as testing for heterogeneity in regression coefficients between different autoantibodies. We believe that the proposed model has wider applicability to other disease scenarios employing selective data collection methods. However, sufficient amount of data need to be available for the estimation of the frailty variance, which was the case with our complete data on the subcohort. The subcohort also needs to be representative of the same target population as the remaining cohort, meaning for example, the same recruitment criteria and measurement standards, to ensure that the model and the frailty variance as part of it are estimated correctly.

We observed that the autoantibody-specific analysis performs similarly in this setting as the usual logistic regression on case-control data: intercept terms are estimated incorrectly, but the regression coefficients remain valid.[27,28] Therefore, a simpler approach could be adopted if a single autoantibody and its risk factors would be the only goal of analysis.

We believe that the model could be modified in terms of alternative estimates of the baseline hazard or distribution of the frailty terms. Here we used a normally distributed frailty term and a piecewise linear baseline hazard, which were sufficiently flexible choices with interval-censored data leading to reasonable computation time. The models with the largest number of parameters could be fitted in less than 2 hours computation time on a normal PC (16 GB RAM, Intel 2.70 GhZ processor on a 64-bit Windows 10 operating system).

In the analysis of the example data, we found little direct evidence of differential associations with the risk factors. The observation of a higher risk of IAA and IA-2A autoantibodies in urban or seminurban municipalities is in line with the finding of an inverse association between agricultural land cover with the risk of multiple islet autoantibodies and type 1 diabetes.[29] Our results on the association between autoantibodies and sex mostly agree with the Finnish diabetes register analysis of Turtinen et al,[30] who detected the same association on IAA and IA-2A, and found no association with ICA. We could not, however, replicate their finding of an inverse association between sex and GADA. Kukko et al[31] found no gender differences among children carrying HLA-conferred susceptibility to type 1 diabetes, but observed that genetically high-risk children were more often positive for all autoantibodies.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on a reasonable request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions. Software in the form of a SAS/NLMIXED example code is available on the GitHub repository https://github.com/jnevalainen/Frailty-models-under-selective-sampling.

ORCID

Jaakko Nevalainen https://orcid.org/0000-0001-6295-0245
Somnath Datta https://orcid.org/0000-0003-4381-1842

REFERENCES

1. Forlenza GP, Rewers M. The epidemic of type 1 diabetes. *Curr Opin Endocrinol Diabetes Obes*. 2011;18(4):248-251. https://doi.org/10.1097/med.0b013e32834872ce
2. Harjutsalo V, Sund R, Knip M, Groop PH. Incidence of type 1 diabetes in Finland. *JAMA*. 2013;310(4):427. https://doi.org/10.1001/jama.2013.8399
3. Köhler M, Beyerlein A, Vehik K, et al. Joint modeling of longitudinal autoantibody patterns and progression to type 1 diabetes: results from the TEDDY study. *Acta Diabetol*. 2017;54(11):1009-1017. https://doi.org/10.1007/s00592-017-1033-7
4. Köhler M, Umlauf N, Beyerlein A, Winkler C, Ziegler AG, Greven S. Flexible Bayesian additive joint models with an application to type 1 diabetes research. *Biom J*. 2017;59(6):1144-1165. https://doi.org/10.1002/bimj.201600224
5. Niinistö S, Takkinen HM, Erlund I, et al. Fatty acid status in infancy is associated with the risk of type 1 diabetes-associated autoimmunity. *Diabetologia*. 2017;60(7):1223-1233. https://doi.org/10.1007/s00125-017-4280-9
6. Ilonen J, Lempainen J, Hammais A, et al. Primary islet autoantibody at initial seroconversion and autoantibodies at diagnosis of type 1 diabetes as markers of disease heterogeneity. *Pediatr Diabetes*. 2017;19(2):284-292. https://doi.org/10.1111/pedi.12545
7. Ilonen J, Lempainen J, Veijola R. The heterogeneous pathogenesis of type 1 diabetes mellitus. *Nat Rev Endocrinol*. 2019;15(11):635-650. https://doi.org/10.1038/s41574-019-0254-y
8. Ilonen J, Hammais A, Laine AP, et al. Patterns of β-cell autoantibody appearance and genetic associations during the first years of life. *Diabetes*. 2013;62(10):3636-3640. https://doi.org/10.2337/db13-0300

9. Krischer JP, Liu X, Lernmark Å, et al. The influence of type 1 diabetes genetic susceptibility regions, age, sex, and family history on the progression from multiple autoantibodies to type 1 diabetes: a TEDDY study report. *Diabetes*. 2017;66(12):3122-3129. https://doi.org/10.2337/db17-0261
10. Lynch KF, Lee HS, Törn C, et al. Gestational respiratory infections interacting with offspring HLA and CTLA-4 modifies incident β-cell autoantibodies. *J Autoimmun*. 2018;86:93-103. https://doi.org/10.1016/j.jaut.2017.09.005
11. Mattila M, , Erlund I, Lee HS, et al. Plasma ascorbic acid and the risk of islet autoimmunity and type 1 diabetes: the TEDDY study. *Diabetologia* 2019;63(2):278–286. https://doi.org/10.1007/s00125-019-05028-z
12. Writing Group for the TRIGR Study Group, Knip M, Åkerblom HK, Al Taji E, et al. Effect of hydrolyzed infant formula vs conventional formula on risk of type 1 diabetes. *JAMA*. 2018;319(1):38. https://doi.org/10.1001/jama.2017.19826
13. Virtanen SM. Dietary factors in the development of type 1 diabetes. *Pediatr Diabetes*. 2016;17:49-55. https://doi.org/10.1111/pedi.12341
14. Uusitalo U, Lee HS, Aronsson CA, et al. Early infant diet and islet autoimmunity in the TEDDY study. *Diabetes Care*. 2018;41(3):522-530. https://doi.org/10.2337/dc17-1983
15. Frederiksen B, Kroehl M, Lamb MM, et al. Infant exposures and development of type 1 diabetes mellitus. *JAMA Pediatrics*. 2013;167(9):808. https://doi.org/10.1001/jamapediatrics.2013.317
16. Hakola L, Miettinen ME, Syrjälä E, et al. Association of cereal, gluten, and dietary fiber intake with islet autoimmunity and type 1 diabetes. *JAMA Pediatrics*. 2019;173(10):953-960. https://doi.org/10.1001/jamapediatrics.2019.2564
17. Lund-Blix NA, Dong F, Mårild K, et al. Gluten intake and risk of islet autoimmunity and progression to type 1 diabetes in children at increased risk of the disease: the diabetes autoimmunity study in the young (DAISY). *Diabetes Care*. 2019;42(5):789-796. https://doi.org/10.2337/dc18-2315
18. Virtanen SM, Nevalainen J, Kronberg-Kippilä C, et al. Food consumption and advanced β cell autoimmunity in young children with HLA-conferred susceptibility to type 1 diabetes: a nested case-control design. *Am J Clin Nutr*. 2012;95(2):471-478. https://doi.org/10.3945/ajcn.111.018879
19. Kupila A, Muona P, Simell T, et al. Feasibility of genetic and immunological prediction of type I diabetes in a population-based birth cohort. *Diabetologia*. 2001;44(3):290-297. https://doi.org/10.1007/s001250051616
20. Zhang Z, Sun J. Interval censoring. *Stat Methods Med Res*. 2010;19(1):53-70. https://doi.org/10.1177/0962280209105023
21. Park Y. One- and two-sample nonparametric inference procedures in the presence of a mixture of independent and dependent censoring. *Biostatistics*. 2005;7(2):252-267. https://doi.org/10.1093/biostatistics/kxj005
22. Zhang Z, Sun L, Sun J, Finkelstein DM. Regression analysis of failure time data with informative interval censoring. *Stat Med*. 2007;26(12):2533-2546. https://doi.org/10.1002/sim.2721
23. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Stat Methods Med Res*. 2007;16:199-218.
24. Wang L, McMahan CS, Hudgens MG, Qureshi ZP. A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics*. 2016;72(1):222-231. https://doi.org/10.1111/biom.12389
25. Wellner JA, Zhan Y. A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *JASA*. 1997;92(439):945-959. https://doi.org/10.1080/01621459.1997.10474049
26. Gentleman R, Vandal A. *Icens: NPMLE for censored and truncated data*. 2018. R package version 1.54.0.
27. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979;66(3):403-411. https://doi.org/10.1093/biomet/66.3.403
28. Carroll RJ, Wang S, Wang CY. Prospective analysis of logistic case-control studies. *JASA*. 1995;90(429):157-169. https://doi.org/10.1080/01621459.1995.10476498
29. Nurminen N, Cerrone D, Lehtonen J, et al. Land cover of early life environment modulates the risk of type 1 diabetes. *Diabetes Care*. 2021;44:1-9. https://doi.org/10.2337/dc20-1719
30. Turtinen M, Härkönen T, Parkkola A, Ilonen J, Knip M, Finnish Pediatric Diabetes Register. Sex as a determinant of type 1 diabetes at diagnosis. *Pediatr Diabetes*. 2018;19(7):1221-1228. https://doi.org/10.1111/pedi.12697
31. Kukko M, Kimpimäki T, Korhonen S, et al. Dynamics of diabetes-associated autoantibodies in young children with human leukocyte antigen-conferred risk of type 1 diabetes recruited from the general population. *J Clin Endocrinol Metab*. 2005;90(5):2712-2717.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Nevalainen J, Datta S, Toppari J, et al. Frailty modeling under a selective sampling protocol: an application to type 1 diabetes related autoantibodies. *Statistics in Medicine*. 2021;1–11. https://doi.org/10.1002/sim.9190