The Illness-Death Model in Family Studies

Jooyoung Lee Richard J. Cook

Department of Statistics and Actuarial Science

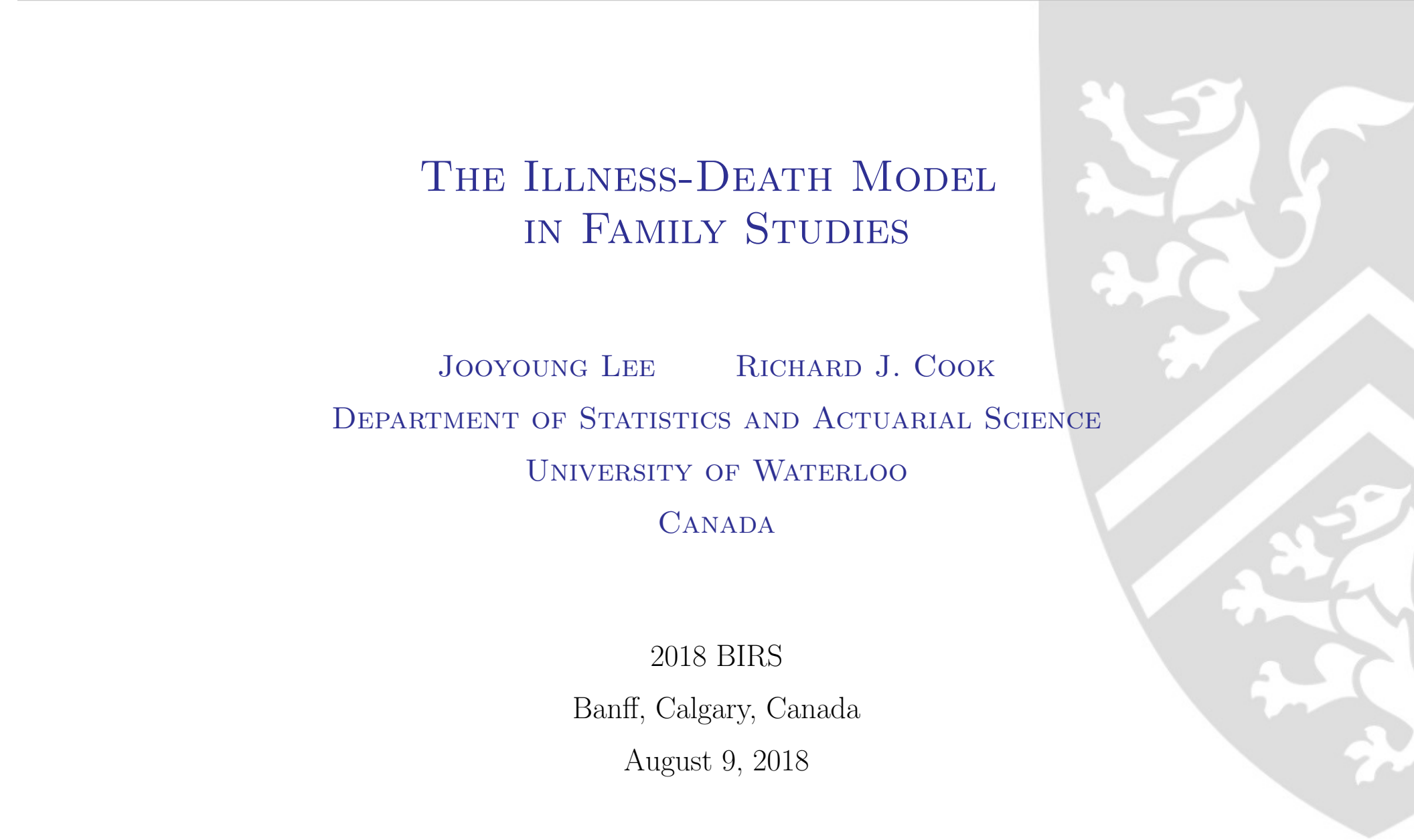
University of Waterloo

Canada

2018 BIRS

Banff, Calgary, Canada

August 9, 2018



UNIVERSITY OF WATERLOO

The Hereditary Nature of Disease

Inference regarding the hereditary nature of disease is initially based on the nature and extent of the within-family association in some feature of the disease process

Fisher (1934)

Analysis are typically based on correlated (within-families) responses on disease status

Ziegler et al. (2000)

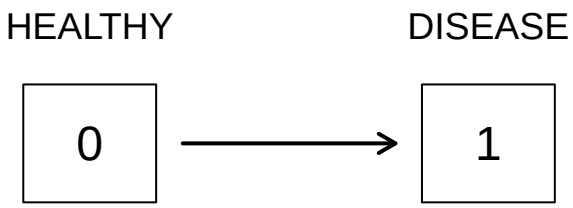
Families are typically selected through identification of an *affected individual called the proband* resulting in a *biased sampling scheme*

Much work has been carried out for the analysis of biased samples of clustered binary responses

Cannings and Thompson (1977); Burton et al. (2000)

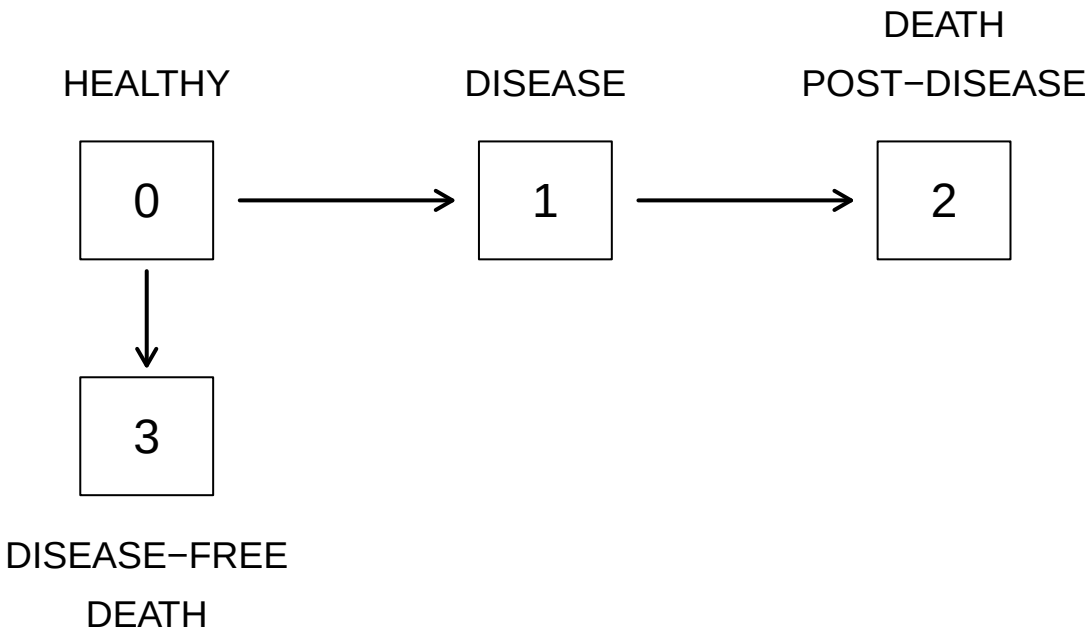
Illness-Death Models

Xu et al. (2010)



Illness-Death Models

Xu et al. (2010)



Offers a natural and helpful frame for joint modeling of disease onset and death

Notation and Selection Conditions

Here we consider the selection conditions for the proband in a family study

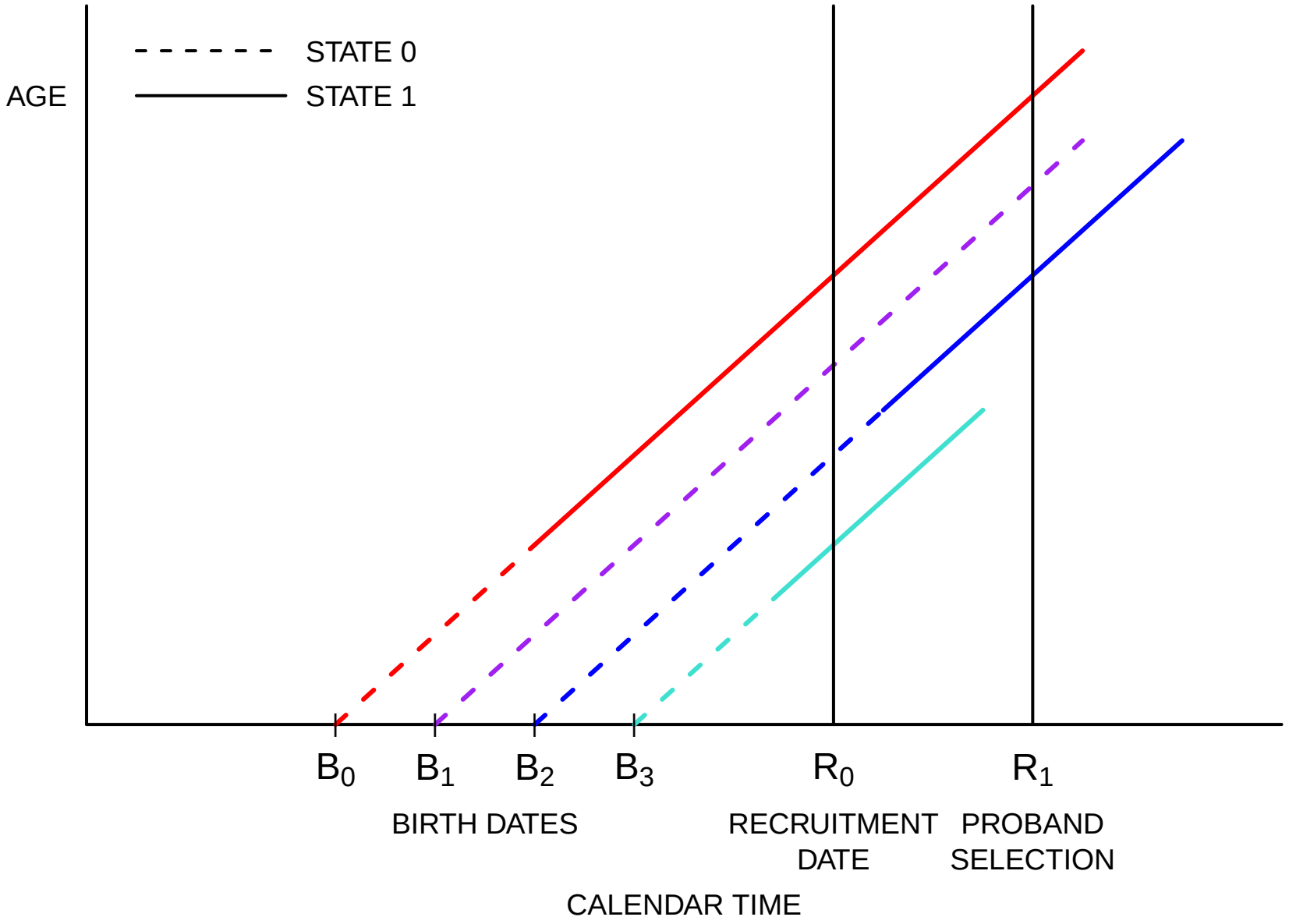
- B is calendar time of birth
- R_0 is date of screening and recruitment to a prevalent cohort study
- R_1 is date of proband sub-sampling for family study

Sampling Conditions for Proband

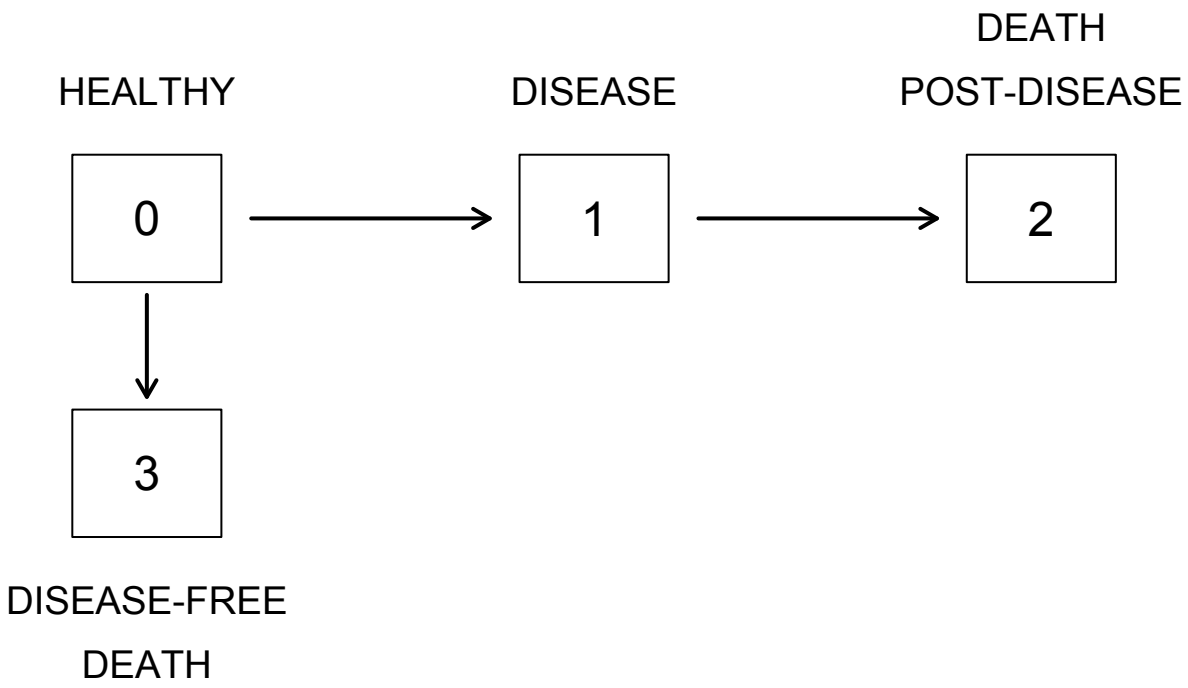
Let $C_0 = R_0 - B$ denote the age at screening for a prevalent cohort and $A_0 = R_1 - B$ denote the age at proband sampling from the prevalent cohort

- Proband must be alive and diseased at age C_0 and A_0

Selection of Family Members to Proband



1. Clustered Illness-Death Model



Notation

X_{ij1} is the *age at disease onset* for member j of family i

X_{ij2} is the age at post-disease death for member j of family i

X_{ij3} is the age at disease-free death for member j of family i

$j = 0, \ldots, m_i$, $i = 1, \ldots, n_F$

Family Data

B_{ij} is birth data of member j of family i, individual (i, j)

V_{ij} are fixed covariates

X_{ijk} is age at entry to state k for individual (i, j)

$Z_{ij}(s)$ is the state occupied for individual (i, j) at age s

$H_{ij}(a) = \{Z_{ij}(s),\ 0 < s < a,\ B_{ij},\ V_{ij}\}$

$$\lim_{\Delta a \downarrow 0} \frac{P(Z_{ij}(a + \Delta a^-) = k \mid Z_{ij}(a^-) = h,\ H_{ij}(a))}{\Delta a} = \lambda_{ijk}(t, a \mid H_{ij}(a))$$

with $t = B_{ij} + a$ and $(h, k) \in \{(0, 1),\ (0, 3),\ (1, 2)\}$

A Markov Proportional Intensity Model

Assume the disease process is Markov given $(B_{ij},\ V_{ij})$ so

$$\lambda_{ijk}(t, a \mid H_{ij}(a)) = \lambda_k(t, a \mid b_{ij}, v_{ij}), \qquad k = 1, 2, 3$$

Disease Intensity

$$\lambda_1(t, a \mid b_{ij}, v_{ij}) = \lambda_1(a) \exp(v'_{ij}\beta_1)$$

Mortality

Andersen et al. (1985)

Allow calendar time trends and set

$$\lambda_3(t, a|b_{ij}, v_{ij}) = \lambda_3(t, a|b_{ij})$$

$$\lambda_2(t, a \mid b_{ij}, v_{ij}) = \lambda_3(t, a|b_{ij})\ \nu_0(a)\ \exp(v'_{ij}\beta_2)$$

Joint Models for Age at Disease Onset

Let $\boldsymbol{B}_i = (B_{i0}, \ldots, B_{im_i})'$ and $\boldsymbol{V}_i = (V_{i0}, \ldots, V_{im_i})'$

Copula Models with Latent Variable Framework

Joe (1997)

$$P(X_{i01} > a_0, \ldots, X_{im_i1} > a_{m_i} \mid \boldsymbol{V}_i; \varphi) = \mathcal{C}(\mathcal{F}(a_0|V_{i0}; \phi_1), \ldots, \mathcal{F}(a_{m_i} \mid V_{im_i}; \phi_1); \rho)$$

where $\varphi = (\phi_1', \rho)'$

Let $\phi = (\phi_1', \phi_2')'$ where ϕ_2 is a parameter vector for the transition from the disease to the death state and $\psi = (\phi', \rho)$

The Cross Ratio for Age at Onset with Pair j and k

Oakes (1989)

$$\theta(a_j, a_k) = \frac{\lambda_1(a_k \mid X_{ij1} = a_j; \boldsymbol{V}_i, \varphi)}{\lambda_1(a_k \mid X_{ij1} > a_j; \boldsymbol{V}_i, \varphi)}$$

Cause-specific Cross Ratio for Age at Onset

Bandeen-Roche and Liang (2002)

We are really in a *semi-competing risk setting*, so make two additional assumptions:

A.1 *Independent semi-competing risks*: $X_{ij1} \perp X_{ij3} \mid V_{ij}$

A.2 $X_{ij3} \perp \{Z_{ik}(s),\ 0 < s\} \mid \boldsymbol{B}_i,\ \boldsymbol{V}_i$ for $j \neq k$

Then

$$\theta_1(a_j, a_k) = \frac{\lambda_1(a_k \mid X_{ij1} = a_j, X_{ij3} > a_j;\ \boldsymbol{B}_i,\ \boldsymbol{V}_i,\ \varphi)}{\lambda_1(a_k \mid X_{ij1} > a_j, X_{ij3} > a_j;\ \boldsymbol{B}_i,\ \boldsymbol{V}_i,\ \varphi)} = \theta(a_j, a_k)$$

Cross-Odds Ratio

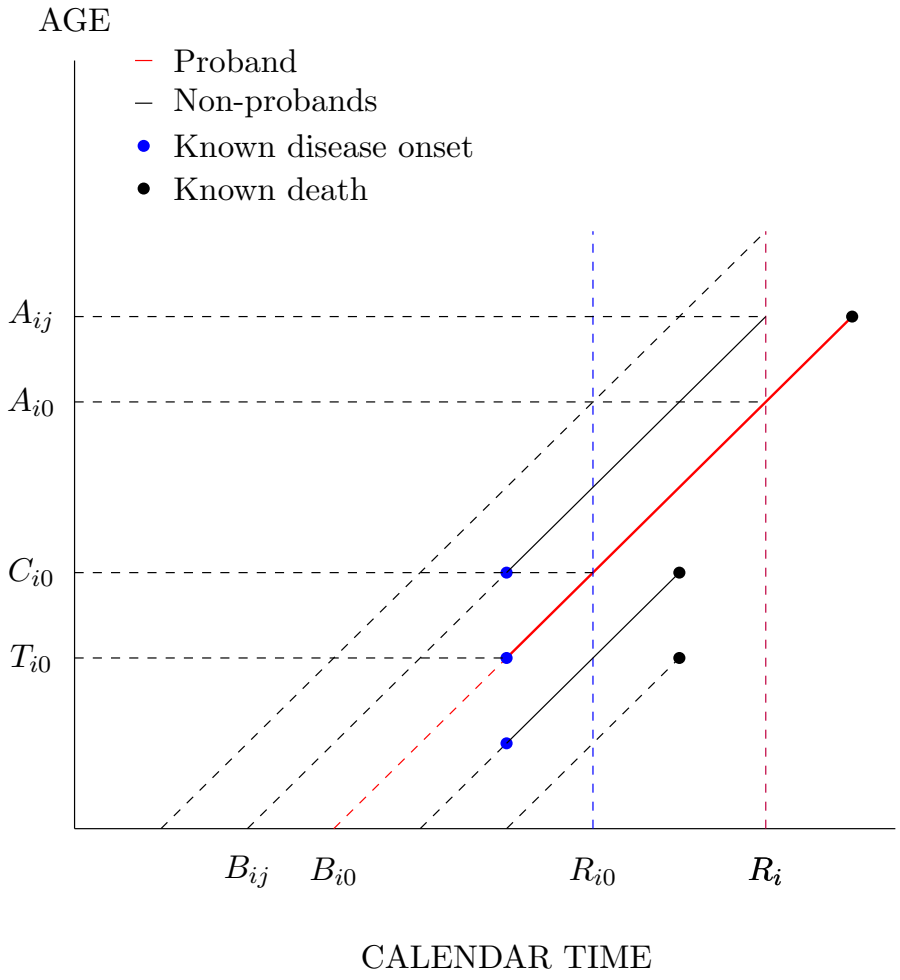
Scheike et al. (2010)

$$\pi(a) = \frac{ODDS(X_{ik1} \leq a, X_{ik1} < X_{ik3} | X_{ij1} \leq a, X_{ij1} < X_{ij3}; \boldsymbol{B}_i, \boldsymbol{V}_i, \varphi)}{ODDS(X_{ik1} \leq a, X_{ik1} < X_{ik3}, B_{ik}, V_{ik}, \phi_1)}$$

2. Biased Sampling and Likelihood Construction

Family studies are recruited through *affected individuals (the proband)*; subscript 0

Lexis Diagram for Family Data



B_{i0}: Date of birth for the proband

C_{i0}: Age of the proband at registry selection

T_{i0}: Age at onset of the proband

A_{i0}: Age of the proband at family selection

B_{ij}: Date of birth for non-proband

A_{ij}: Age of non-proband at family selection

$\boldsymbol{A}_i = (A_{i0}, \ldots, A_{im_i})$ for family size $m_i + 1$

Likelihood Construction

$\boldsymbol{Z}_i(\boldsymbol{s}_i) = (Z_{i0}(s_{i0}), Z_{i1}(s_{i1}), \ldots, Z_{im_i}(s_{im_i}))'$ with $\boldsymbol{s}_i = (s_{i0}, \ldots, s_{im_i})'$

$\bar{Z}_{ij}(s) = \{Z_{ij}(u), 0 < u < s; B_{ij}\}$

$\bar{\boldsymbol{Z}}_i(s_i) = \{Z_{ij}(u), 0 < u < s_{ij}, j = 0, \ldots, m_i; \boldsymbol{B}_i\}$

Biased selection (alive and diseased) of proband

Left truncated death and right-truncated disease onset time

$$L_{i0}(\phi) = P(\bar{Z}_{i0}(A_{i0}) | Z_{i0}(C_{i0}) = 1, C_{i0}, B_{i0}, V_{i0}; \phi)$$

P(proband's history I proband alive and diseased at R0)

Biased selection (alive) of non-probands

$$L_i^{II}(\psi) \propto L_{i0}(\phi) P(\bar{\boldsymbol{Z}}_i^-(\boldsymbol{A}_i^-) | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{Z}_i^-(\boldsymbol{A}_i^-) \in \{0,1\}^{m_i}, \boldsymbol{A}_i, \boldsymbol{B}_i, \boldsymbol{V}_i; \psi)$$

P(non-probands' histories I proband alive and diseased at R0)

Pairwise Conditional Composite Likelihood Construction

We wish to avoid calculating

$$P(\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{Z}_i(\boldsymbol{A}_i) \in \{0,1\}^{m_i}, \boldsymbol{A}_i, \boldsymbol{B}_i, \boldsymbol{V}_i)$$

The *composite likelihood* is Varin et al. (2011)

$$CL(\psi) \propto \prod_{i=1}^{n_F} L_{i0}(\phi) \prod_{1 \le j < l \le m_i} \left\{ L^{II}_{ijl}(\psi) \right\}^{\frac{1}{m_i - 1}} \tag{1}$$

where

$$L^{II}_{ijl}(\psi) = P(\bar{\boldsymbol{Z}}^{-}_{ijl}(\boldsymbol{A}^{-}_{ijl}) | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{Z}^{-}_{ijl}(\boldsymbol{A}^{-}_{ijl}) \in \{0,1\}^2, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)$$

- $\bar{\boldsymbol{Z}}_{ijl}(s_{ijl}) = \{Z_{ih}(u), 0 < u < s_{ih}, h = 0, j, l; \boldsymbol{B}_{ijl}\}$
- $\boldsymbol{B}_{ijl} = (B_{i0}, B_{ij}, B_{il})'$

$\boldsymbol{A}_{ijl} = (A_{i0}, A_{ij}, A_{il})'$

$\boldsymbol{V}_{ijl} = (V'_{i0}, V'_{ij}, V'_{il})'$

3. Use of Auxiliary Data to Augment likelihood

Auxiliary data are critical because

- proband simply gives a right-truncated onset time
- low incidence of disease among non-probands
- $\lambda_3(\cdot,\cdot)$ and $\lambda_2(\cdot,\cdot)$ are inestimable from the family data alone

The combination of data from different sources have been suggested

- for case-control studies — Pfeiffer et al (2008), Zheng et al (2010)
- for twin-based studies — Balliu et al (2012)

We consider — Zhong and Cook (2016, 2017)

- a registry data (with follow-up) $\longrightarrow \lambda_2()$
- a cross-sectional survey yielding current status data $\longrightarrow \lambda_1()$
- a national statistics for mortality rate $\longrightarrow \lambda_3()$

Augmented Likelihood Construction

$\mathcal{A}_1$, $\mathcal{A}_2$ is the set of individuals in the registry and the cross-sectional survey

X_{r1} is the age at onset, X_{r2} is the age at death following disease (if available)

C_r is the age at recruitment and $A_r^* = \min(C_r^*, X_{r2})$ with C_r^* the last assessment time

We multiply $CL(\psi)$ in (1) by $L_{\mathcal{A}_1} \times L_{\mathcal{A}_2}$ where

$$L_{\mathcal{A}_1} \propto \prod_r^{n_R} P(\bar{Z}_r(A_r^*)|Z_r(C_r) = 1, C_r, B_r, V_r)$$

$$L_{\mathcal{A}_2} \propto \prod_r^{n_S} P(Z_r(C_r) = 0|Z_r(C_r) \in \{0,1\}, B_r, V_r)^{I(Z_r(C_r)=0)}$$

$$\times P(Z_r(C_r) = 1|Z_r(C_r) \in \{0,1\}, B_r, V_r)^{I(Z_r(C_r)=1)}$$

Canadian National Mortality Rates over (t, a)

Assume $\lambda_3(\cdot, \cdot)$ are given by the population mortality rates (Statistics Canada and Robert, 2017)

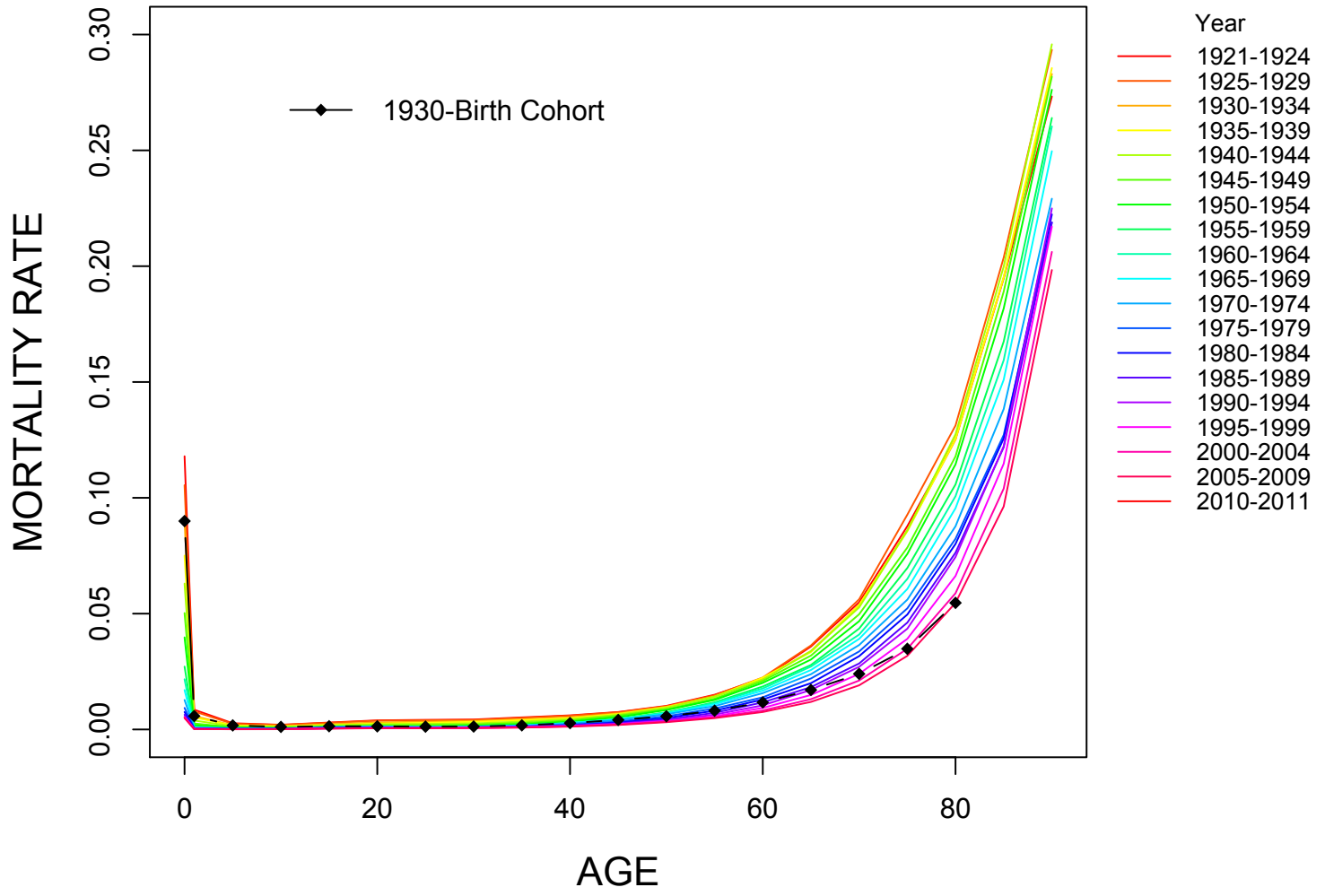


Figure 1: Age-specific population mortality rates by calender period, Canada, 1921-2011

Chatterjee et al (2006)

Genotype Variable

G_{ij} are genotype (gene carrier indicator)

$P(G_{ij} = 0) = (1-p)^2, P(G_{ij} = 1) = p^2 + 2p(1-p)$ with the allele frequency p

$\boldsymbol{G}_i = (G_{i0}, \ldots, G_{im_i})'$

$W_{ij} = (V'_{ij}, G_{ij})'$ and $\boldsymbol{W}_i = (\boldsymbol{V}'_i, \boldsymbol{G}'_i)'$

The transition intensities are given as

$$\lambda_l(t, a|b_{ij}, w_{ij}) \quad \text{for } l = 1, 2, 3$$

Assumptions	Population satisfies
1	Hardy-Weinberg Equilibrium and Mendel's law
2	$G_{ij} \perp V_{ij}$
3	$\bar{Z}_{ij}(s) \perp G_{ik} \mid G_{ij}$ for $j \neq k$
4	$\lambda_3(t, a|b_{ij}, w_{ij}) = \lambda_3(t, a|b_{ij}), \quad \lambda_2(t, a|b_{ij}, w_{ij}) = \lambda_2(t, a|b_{ij}, v_{ij})$

Missing Genotype

- *No genetic information in the current status data*
- *No available genotype data for non-probands* who died before study

With missing genotype, we model G_{ij} with allele frequency p

Pairwise Conditional Composite Likelihood

$$L_{i0}(\phi) = P(\bar{Z}_{i0}(A_{i0}), G_{i0} | Z_{i0}(C_{i0}) = 1, C_{i0}, B_{i0}, V_{i0}; \phi)$$

$$L^{II}_{ijl}(\psi) = P(\bar{\boldsymbol{Z}}^{-}_{ijl}(\boldsymbol{A}^{-}_{ijl}), \boldsymbol{G}^{-}_{ijl} | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, G_{i0}, \boldsymbol{Z}^{-}_{ijl}(\boldsymbol{A}^{-}_{ijl}) \in \{0,1\}^2, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)$$

where $\boldsymbol{G}_{ijl} = (G_{i0}, G_{ij}, G_{il})'$

Auxiliary Data

G_r denote genotype of individual r in $\mathcal{A}_1$ or $\mathcal{A}_2$

$$L_{\mathcal{A}_{1,r}} \propto P(\bar{Z}_r(A^*_r), G_r | Z_r(C_r) = 1, C_r, B_r, V_r),$$
$$L_{\mathcal{A}_{2,r}} \propto \prod_{h \in \{0,1\}} E_{G_r | Z_r(C_r) \in \{0,1\}} [P(Z_r(C_r) = h | Z_r(C_r) \in \{0,1\}, B_r, G_r, V_r)]^{I(Z_r(C_r)=h)}$$

Simulations with Genotype

Family Data

Family size: *4 or 6 members with 2 parents and 2 or 4 children*; $P(m_i + 1 = 4) = 2/3$

The *date of birth*: the uniform dist $(1920, 1950)$ if a parent or $(1950, 1980)$ otherwise

The *affected individual recruitment date*: the uniform dist (1980, 2010)

The *family sampling date* on July 1st of 2010

G_{ij} is the gene mutation with the *allele frequency p=0.06*

Generate G_i based on the family structure

$\lambda_1(a|W_{ij}) = \lambda_{01} \exp(G_{ij}\alpha)$ with $\lambda_{01} = 0.01$ and $\alpha = \log(1.5)$

$\lambda_{02}(t, a) = \nu\lambda_{03}(t, a)$ with $\nu = 1.1$

Auxiliary Data

Registry data: follow-up by July 1st of 2010 with the record of death post disease

Current status survey data: the date of birth from the uniform dist $(1930, 1980)$ and the sampling date as July 1st of 2000

Sample Size

$n_F = 1000$, $n_R = 2000$, $n_S = 1000$

Design

Not Modeling G_i:

(i) Family studies + Registry (All genotype is available)

Modeling G_i:

(ii) Family studies + Registry + Survey w/ missing genotype

Empirical Properties of Estimates

Table 1: Two sources of auxiliary data: the registry follow-up data and the current status survey data; Clayton copula with Kendall's τ=0.2, 0.4; $n_F = 1000$, $n_R = 2000$, $n_S = 1000$, and $nsim = 1000$

		Registry Data				Registry + Current Status Data			
τ	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
0.2	$\log(\lambda_{01})$	0.001	0.061	0.064	0.951	-0.000	0.042	0.042	0.941
	α	-0.003	0.070	0.071	0.953	-0.003	0.065	0.065	0.949
	$\log(\nu)$	-0.001	0.047	0.046	0.951	-0.001	0.046	0.046	0.949
	$\log(p)$	-	-	-	-	0.002	0.055	0.055	0.949
	τ	0.000	0.030	0.032	0.956	0.001	0.024	0.024	0.952
0.4	$\log(\lambda_{01})$	0.004	0.081	0.082	0.955	0.001	0.046	0.045	0.951
	α	-0.001	0.063	0.062	0.949	-0.002	0.059	0.058	0.948
	$\log(\nu)$	-0.002	0.047	0.046	0.948	-0.002	0.046	0.046	0.950
	$\log(p)$	-	-	-	-	0.002	0.053	0.053	0.942
	τ	-0.001	0.035	0.035	0.949	0.001	0.023	0.023	0.941

5. Application to Psoriatic Arthritis (PsA) Family Studies

PsA is an immune-mediated inflammatory disease occurring commonly in patients with psoriasis

The Centre for Prognosis Studies in Rheumatic Disease at the University of Toronto was established in 1976 and has been following patients since its formation

To date in April of 2017, a total of 1436 patients in the Toronto Registry include a number of 150 proband sampled for the family study

We have 168 pseudo families where two-generation families are considered with a total of 532 individuals for the family study

Patients with PsA are at higher risk for death compared to the general population of Ontario with a standardised mortality ratio of 1.36 Gladman (2008)

A total of 15307 respondents are sampled for a U.S. national survey of the National Psoriasis Foundation in 2001 Gelfand et al. (2005)

Lexis Diagram for a family

Family: One proband with parents

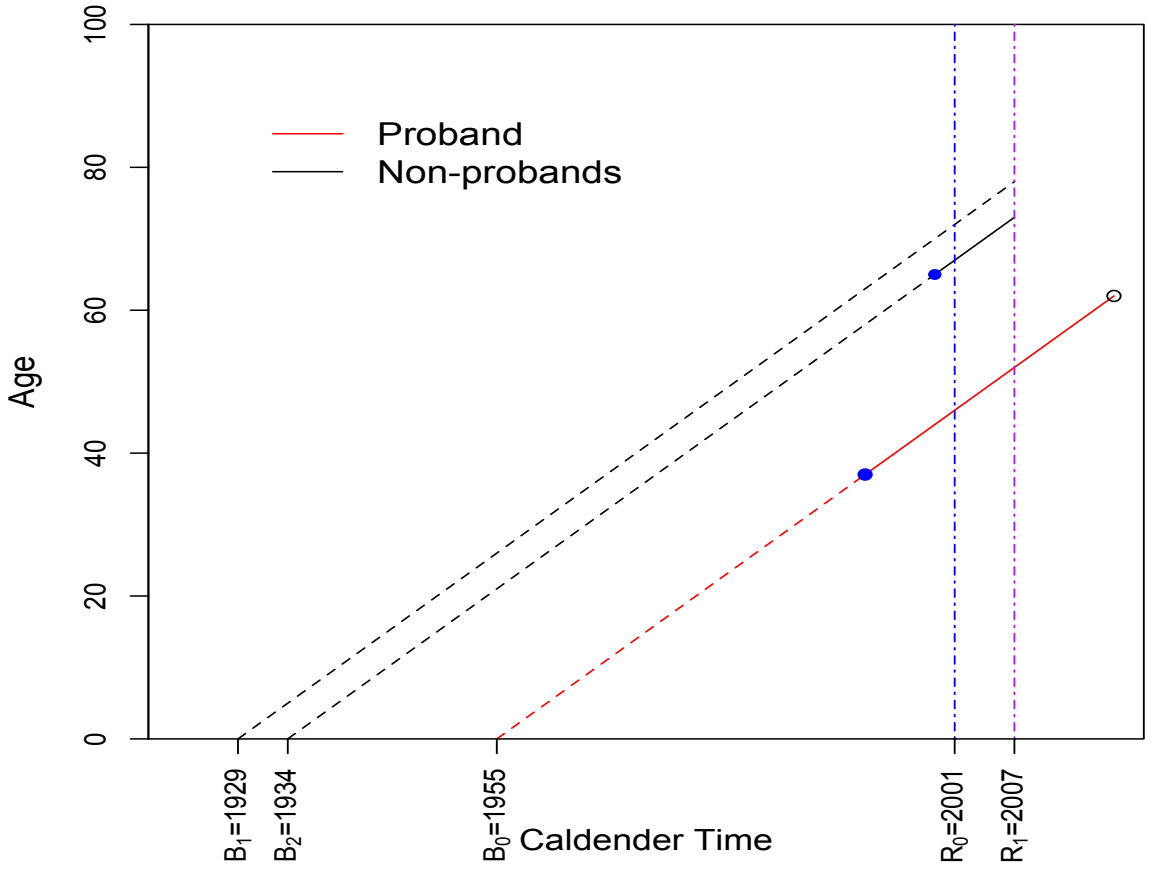


FIGURE 2: Lexis diagram for a family with 3 members; one proband and parents

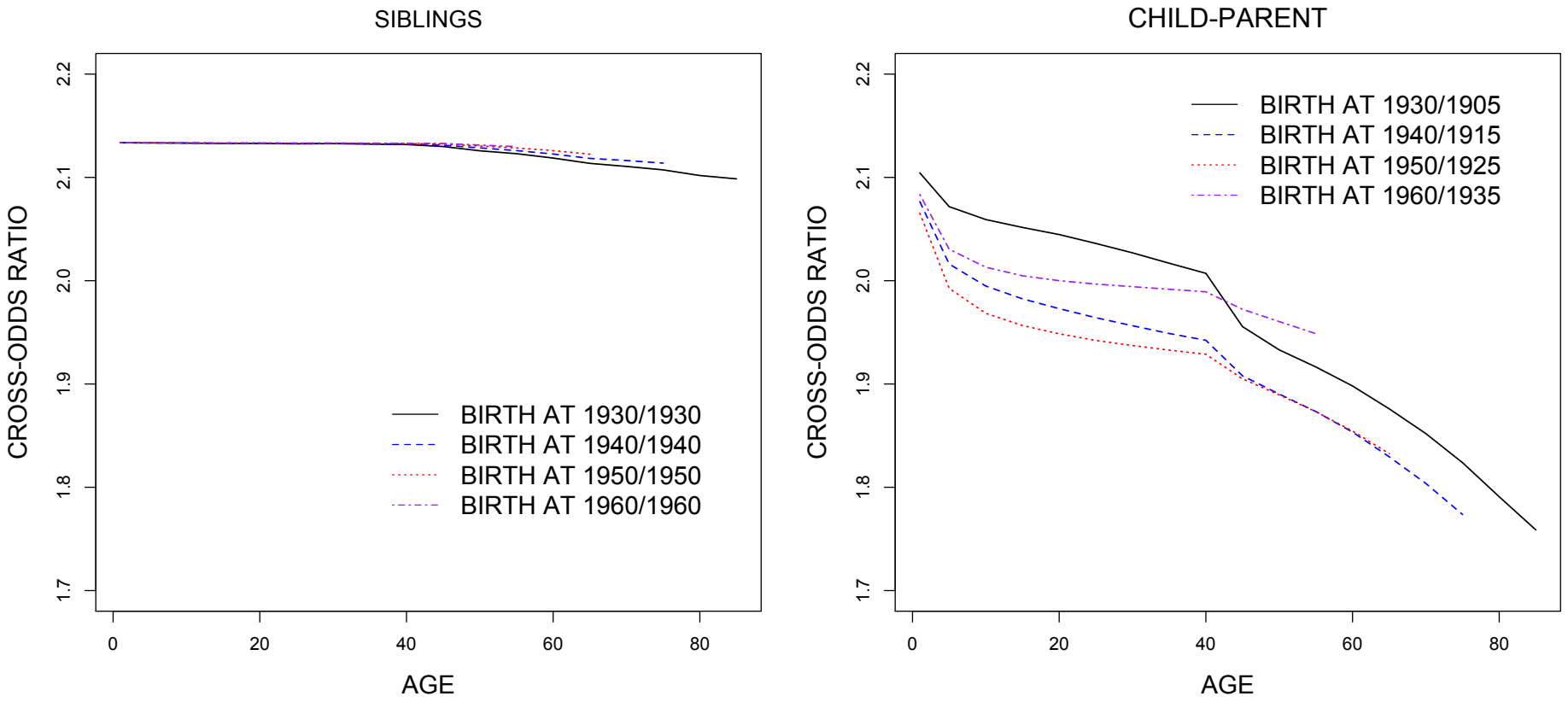
TABLE 2: Estimates of parameters based on the augmented pairwise likelihood; auxiliary data include the University of Toronto Psoriatic Arthritis Registry and the survey from Gelfand et al. (2005) without/with genotype variable under the Exponential model and piecewise constant marginal model for age at PsA onset with a cut point 40.

MARKER	α_{marker}	ν	τ	p_{marker}
-	-	1.152 (0.016)	0.362 (0.083)	-
B27	0.605 (0.239)	1.155 (0.080)	0.345 (0.085)	0.054 (0.012)
C06	0.117 (0.086)	1.155 (0.060)	0.362 (0.089)	0.115 (0.011)

- $\hat{\nu} = 1.152$; the ratio of the hazard of death post-PsA to PsA-free death is 1.152
- $\hat{\tau} = 0.362$ (95% CI: 0.199, 0.525; p value< 0.001)
- $\hat{\theta} = 2.134$ (95% CI: 1.354, 2.914; p value< 0.001)
- HLA-B27 effect on PsA onset

$$RR = 1.831;\ 95\%\ \text{CI: } 0.137,\ 1.073;\ \text{p } = 0.011$$

FIGURE 3: The cross-odds ratio for two siblings born at the same year 1930, 1940, 1950, 1960 (the right panel) and a child born at 1930, 1940, 1950, 1960 given a parent born at 1905, 1915, 1925, 1935 (the left panel) based on the fitted model with no genetic marker



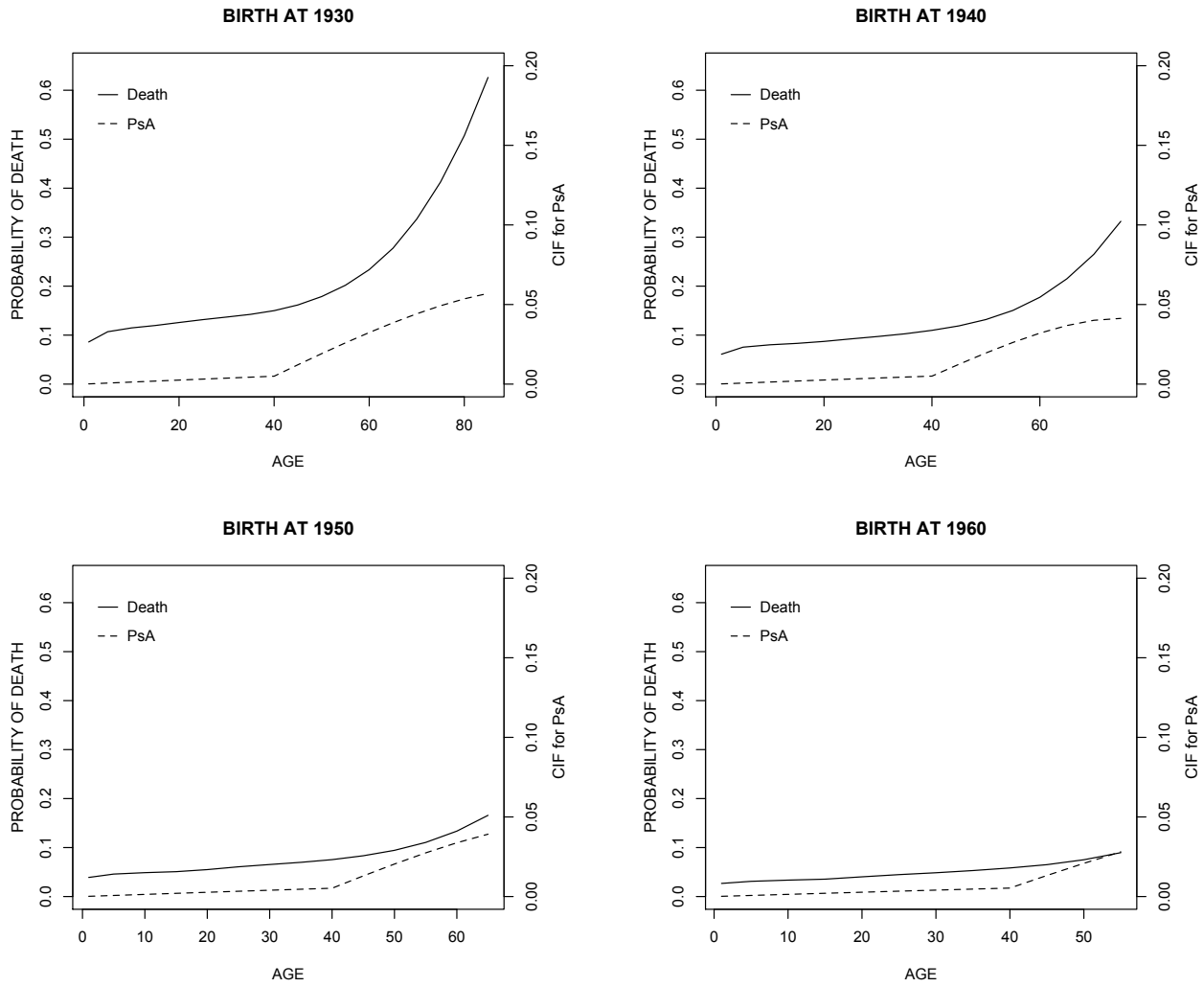


FIGURE 4: The marginal probability of death and the cumulative incidence function of PsA by the year of birth at 1930, 1940, 1950, 1960 based on the fitted model with no genetic marker

Concluding Remarks

Within family dependence in disease process must be modeled to account for selection bias

Less well-studied is the survival bias from requiring individuals (probands and non-probands) to live until the conduct of the family study if they must be examined for disease status

Modeling dependencies in the context of the illness death model is challenging

Identifiability issues require auxiliary data which enable one to fit appropriate models

Tests for genetic associations with disease onset perform well

Score tests are being developed for genetic associations

References

Aalen, O. O. (2012). Armitage lecture 2010: understanding treatment effects: the value of integrating longitudinal data and survival analysis. Statistics in medicine 31(18), 1903-1917

Andersen, P. K, Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, Philip, Keiding, Niels and Kreiner, Svend. (1985). A cox regression model for the relative mortality and its application to diabetes mellitus survival data. Biometrics, 921-932

Balliu, B., Tsonaka, R., van der Woude, D., Boehringer, S. and Houwing- Duistermaat, J. J. (2012). Combining family and twin data in association studies to estimate the noninherited maternal antigens effect. Genetic Epidemiology 36(8), 811-819

Bandeen-Roche, K. and Liang, K. (2002). Modelling multivariate failure time associations in the presence of a competing risk. Biometrika 89(2), 299-314

Chatterjee, N., Kalaylioglu, Z., Shih, J. H. and H Gail, M. (2006). Case-control and case-only designs with genotype and family history data: Estimating relative risk, residual familial aggregation, and cumulative risk. Biometrics 62(1), 36-48

Gladman, D. D. (2008). Mortality in psoriatic arthritis. Clinical & Experimental Rheumatol- ogy 26(5), S62.

Gelfand, J. M., Gladman, D. D., Mease, P. J., Smith, N., Margolis, D. J., Nijsten, T., Stern, R. S., Feldman, S. R. and Rolstad, T. (2005). Epidemiology of psoriatic arthritis in the population of the United States. Journal of the American Academy of Derma- tology 53(4), 573–e1.

Hsu, L., Chen, L., Gorfine, M., Malone, K. (2004). Semiparametric estimation of marginal hazard function from case–control family studies. Biometrics 60(4), 936-944

Hsu, L. and Gorfine, M. (2005). Multivariate survival analysis for case–control family data. Biostatistics 7(3), 387–398.

Jiang, F. and Haneuse, S. (2017). A semi-parametric transformation frailty model for semi- competing risks survival data. Scandinavian Journal of Statistics 44(1), 112–129.

Joe, H. (1997). Multivariate Models and Multivariate Dependence Concepts. CRC Press

Lakhal-Chaieb, L., Cook, R. J. and Y., Zhong. (2018). Testing the heritability and parent- of-origin hypotheses for ages at onset of psoriatic arthritis under biased sampling. Biometrics under revision.

Oakes, D. (1989). Bivariate survival models induced by frailties. Journal of the American Statistical Association 84(406), 487-493

Pfeiffer, R. M., Pee, D. and Landi, M. T. (2008). On combining family and case-control studies. Genetic Epidemiology 32(7), 638-646

Shih, J. H. and Albert, P. S. (2010). Modeling familial association of ages at onset of disease in the presence of competing risk. Biometrics 66(4), 1012–1023.

Shih, J. H. and Chatterjee, N. (2002). Analysis of survival data from case–control family studies. Biometrics 58(3), 502–509.

Scheike, T. H., Sun, Y., Zhang, M. and Jensen, T. K. (2010). A semiparametric random effects model for multivariate competing risks data. Biometrika 97(1), 133–145.

Statistics Canada and Robert, B. (2017). Mortality data for Canada. https://www. mortality.org/cgi-bin/hmd/country.php?cntr=CAN& level=1.

Varin, Cristiano, Reid, Nancy and Firth, David. (2011). An overview of composite like- lihood methods. Statistica Sinica, 5–42.

Xu, J., Kalbfleisch, J. D. and Tai, B. (2010). Statistical analysis of illness–death processes and semicompeting risks data. Biometrics 66(3), 716–725.

Zheng, Y., Heagerty, P. J., Hsu, L. and Newcomb, P. A. (2010). On combining familybased and population-based case-control data in association studies. Biometrics 66(4), 1024-1033

Zhong, Y. and Cook, R.J. (2016) Augmented composite likelihood for copula modeling in family studies under biased sampling. Biostatistics 17 (3), 437-452

Zhong, Y. and Cook, R.J. (2017) Second-order estimating equations for clustered current status data from family studies using response-dependent sampling. *Statistics in Biosciences* (online). DOI: 10.1007/s12561-017-9201-4