Estimation of marriage incidence rates by combining two cross-sectional retrospective designs : Event history analysis of two dependent processes

Kulathinal, Sangita

2021-06-01

Kulathinal , S , Sääväla , M , Auranen , K & Saarela , O 2021 , ' Estimation of marriage incidence rates by combining two cross-sectional retrospective designs : Event history analysis of two dependent processes ' , Journal of Indian Statistical Association , vol. 59 , no. 1 , pp. 21-57 . <
https://www.indstatassoc.org/journal-jisa/previous-volumes/june-2021-vol-591 >

http://hdl.handle.net/10138/349626

unspecified
publishedVersion

Journal of the Indian Statistical Association
Vol. 59 No. 1, June 2021, 21-57.

Estimation of marriage incidence rates by combining two cross-sectional retrospective designs: Event history analysis of two dependent processes

Sangita Kulathinal

Department of Mathematics and Statistics, University of Helsinki, Finland

Minna Säävälä

Population Research Institute, Väestöliitto, Helsinki, Finland

Kari Auranen,

Department of Mathematics and Statistics, University of Turku, Finland

Olli Saarela,

Dalla Lana School of Public Health, University of Toronto, Canada

Abstract

The aim of this work is to develop methods for studying the determinants of marriage incidence using marriage histories collected under two different types of retrospective cross-sectional study designs. These designs are: sampling of ever married women, that is, women who have been married at least once before the cross-section, a prevalent cohort, and sampling of women irrespective of marital status, a general cross-sectional cohort. While retrospective histories from a prevalent cohort do not identify incidence rates without parametric modelling assumptions, the rates can be identified when combined with data from a general cohort. Moreover, education, a strong endogenous covariate, and marriage processes are correlated. Hence, they need to be modelled jointly in order to estimate the marriage incidence. For this purpose,

we specify a multi-state model and propose a likelihood-based estimation method. We outline the assumptions under which a likelihood expression involving only marriage incidence parameters can be derived. This is of particular interest when either retrospective education histories are not available or related parameters are not of interest. Our simulation results confirm the gain in efficiency by combining data from the two designs, while demonstrating how the parameter estimates are affected by violations of the assumptions used in deriving the simplified likelihood expressions. Two Indian National Family Health Surveys are used as motivation for the methodological development and to demonstrate the application of the methods.

Key Words : *Correlated processes, Cross-sectional surveys, Event history analysis, Incidence rate, Multi-state models, Prevalent cohort, Retrospective histories*

1 Introduction

In sociology and demography, population-based cross-sectional surveys have been used to estimate rates of events such as marriage or cohabitation especially in the absence of reliable population registers (Hayford and Morgan, 2008, Raj et al., 2009). For estimation of marriage incidence rates, retrospective marriage histories, e.g. ages at first marriage, can be collected by sampling at a cross-section. Two commonly employed sampling designs at a cross-section are; (i) sampling of ever married women, women who have been married at least once before the cross-section, a prevalent cohort, and (ii) sampling of women irrespective of marital status, a general cross-sectional cohort. Marriage histories are collected retrospectively under the two designs. We refer to studies based on these two designs as retrospective cohort studies I and II, respectively.

Similar designs are used in epidemiology to estimate incidence rate of a disease based on retrospective disease histories, with methods described in e.g. Keiding (1991), Keiding et al. (2012). Keiding (2006) gives an overview of event history analysis and the cross-section with focus on complex sam-

pling patterns. Further, Saarela et al. (2009) proposed combining retrospective event histories from individuals with prevalent disease and prospective follow-up of disease free individuals at the cross-section, incident cohort, to improve efficiency in estimating effects of time-invariant covariates on disease incidence. Gain in efficiency has also been demonstrated in estimation of survival time from disease onset to death based on combined prevalent and incident cohort data (Ning et al., 2017, Wolfson et al., 2019).

Although incidence rate estimation methods using retrospective event histories are known in epidemiology, their application in other fields are sparse. In the sociological context, retrospective event histories are typically collected under the cross-sectional retrospective designs described earlier. The quality of retrospective data on cohabitation by comparing data collected in four surveys, all having the sampling design of type II above, has been studied by Hayford and Morgan (2008). They estimated average probabilities of cohabitation under discrete-time event history logit model with fixed covariates. To estimate incidence of the outcome when the outcome of interest is correlated with an endogenous covariate process, the outcome and the covariate processes need to be modelled jointly. Moreover, the estimation method should account for the sampling. In the absence of complete covariate process histories at the cross-section, incidence rates estimation may be possible only under special assumptions or sufficient background information on the covariate processes.

The novelty of the present work is in modelling marriage and education processes jointly using a multi-state model by combining the two retrospective cohort studies. We thus extend the existing likelihood-based methods for estimation of incidence rates to simultaneously account for two different sampling patterns; two correlated processes; and two time scales. We outline the assumptions under which the likelihood expressions for the marriage incidence rates can be derived when complete retrospective histories of the education process are not available or when parameters characterising the

education process are not of interest. In a simulation study, we assess the gain in efficiency due to using the proposed method over relying on data from either of the two studies. We apply the methods to two nationally representative Indian National Family Health Surveys (NFHS) data to study the trends and determinants of marriage incidence in India. While we present results in the context of education and marriage, the results are general and can be applied to other similar settings.

The paper is organised as follows. Section 2 introduces the empirical data from the two NFHS. Section 3 outlines the model of female marriage incidence and derives the necessary likelihood expression of the model parameters to estimate them from cross-sectional data. Section 4 considers calculation of predictive probabilities based on the model. A simulation study and data analysis results are presented in sections 5 and 6. The paper concludes with a discussion.

2 The data

The motivation for this work comes from the estimation of marriage incidence rates and their determinants using two NFHS; surveys conducted in India during 1998-99 (NFHS-2) and 2005-06 (NFHS-3). The NFHS-2, an example of retrospective cohort study I, was a cross section of a nationally representative sample of 91196 households with 90265 ever-married women aged 15-49 years and gave a retrospective cohort of ever married women. The NFHS-3, an example of retrospective cohort study II, included 109041 households with 124373 women aged 15-49 years irrespective of marital status and gave a retrospective cohort, irrespective of the current status of marriage at the time of survey. The data and reports of the NFHS are available through the National Family Health Survey website (`http://rchiips.org/nfhs/`). A schematic Lexis diagram illustration of the two cohorts is presented in Figure 1. The x-axis represents calendar years and the y-axis represents age in years. Each line shows the life of a woman where the solid line indicates

life before marriage and dashed line indicates life after marriage. A dot shows the calendar time and age at the time of the first marriage. There is no fully black solid line because study I included only ever married women. The oldest birth cohort included in study I was 1949 and that in study II was 1956. Education is known to be a key determinant of marriage (Dommaraju, 2009, Goswami, 2014, Kalmijn, 1991, Ruwali, 2018). Moreover, who one marries depends on one's education more so than in the past (Cherlin, 2010). Hence, we model the joint dependency of the education and marriage processes in this context.

The data used in the current analysis include each female participant's age at the time of the survey, age at the first marriage, birth cohort, state, urban/rural residence, caste category, religion, and highest educational level completed, categorized as in Table 1. For analytical purposes, we have chosen, in addition to whole India, four Indian states , *viz.*, Kerala, Maharashtra, Punjab and Rajasthan that differ geographically, socially and economically (cf. Appendix A). *Caste* refers here to the four administrative categories of caste, *viz.*, Scheduled Caste, Scheduled Tribe, Other Backward Class and Other that are used by the Government of India to represent disadvantaged groups and to provide reservations based on the caste system. Scheduled Castes, Scheduled Tribes and Other Backward Classes are groups that have faced a varying degree of social and economic discrimination in the past (Desai and Kulkarni, 2008, Government of India, 2011). The total number of study subjects in the four states was 38052 as seen from Table 6, Appendix A.

3 Joint modelling of education and marriage processes and estimation of marriage incidence rates

As noted earlier, education is known to be a key determinant of marriage and vice versa, and hence, are highly correlated with each other. We model the

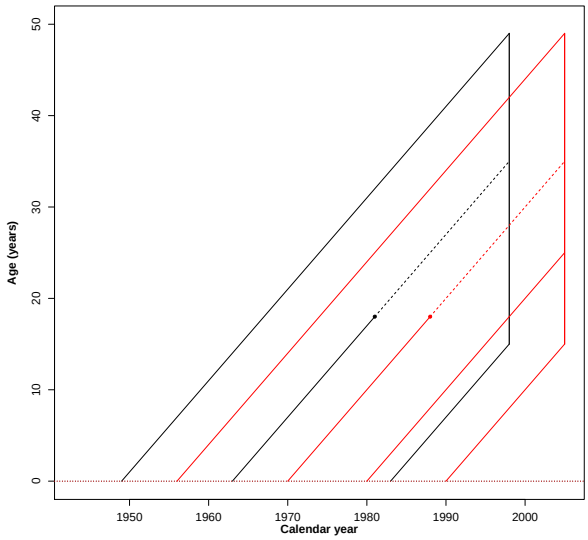

Figure 1: Lexis diagram illustration of the two cross-sectional surveys. NFHS-2 (black, retrospective cohort study I) collected retrospective marriage histories from ever married women only. NFHS-3 (red, retrospective cohort study II) included also never married women and collected retrospective marriage histories from currently married women.

two correlated processes; at-school and marriage processes, in a multi-state modelling framework. Here at-school process models continuing education starting from the primary school to the highest possible university level education, and marriage process models transition from never married state to married state. In the sequel, never married is referred as unmarried. Our interest is in the incidence of the first marriage and we thus do not model the subsequent changes to other possible states such as divorced, widowed or remarried. Each process has two states indicating respective status, and the joint process can be described using a multi-state model as depicted in Figure 2. The state space of the joint process is {*at school and unmarried, at school and married, out of school and unmarried, out of school and married, dead* }. We denote these five states as $\{1, 2, \ldots, 5\}$, respectively. Of note, the

Table 1: Covariates in the marriage incidence model. The reference categories are indicated as 'ref.'

Covariate	Category	Notation
Birth cohort	1942-62 (ref.)	$x_1 = 0$
	1962-72	$x_1 = 1$
	1972-82	$x_1 = 2$
	1982-92	$x_1 = 3$
Residence status	Urban (ref.)	$x_2 = 0$
	Rural	$x_2 = 1$
Caste	Scheduled Caste (SC, ref.)	$x_3 = 0$
	Scheduled Tribe (ST)	$x_3 = 1$
	Other Backward Class (OBC)	$x_3 = 2$
	Other	$x_3 = 3$
Religion	Hindu (ref.)	$x_4 = 0$
	Muslim	$x_4 = 1$
	Christian	$x_4 = 2$
	Sikh	$x_4 = 3$
	Other	$x_4 = 4$
Education	None (< 5 years) (ref.)	$x_5 = 0$
	Primary (5-9 years)	$x_5 = 1$
	Secondary (10-12 years)	$x_5 = 2$
	Higher (> 12 years)	$x_5 = 3$

at-school process jumps to *out of school* state when the formal education, including university level education, ends and the marriage process jumps to *married state* at the time of the first marriage. Let a_e and a_0 be the minimum age of starting basic compulsory education and the minimum marriageable age $a_0 (> a_e)$, respectively. In the Indian context, (a_e, a_0) are taken as $(6, 12)$.

We denote the calendar time corresponding to age a as $t(a) = t_0 + a$ where t_0 is the birth year and define both processes in a Lexis diagram with calendar time and age as the two time scales. We define the at-school

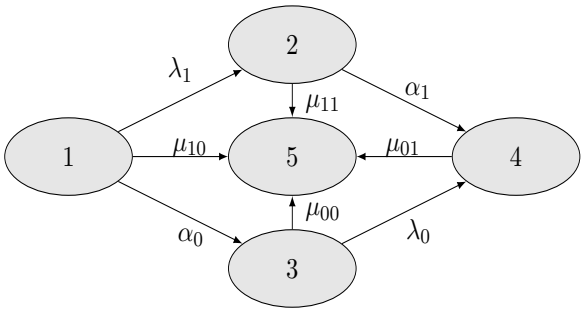

Figure 2: At-school and marriage processes as a multi-state model (states are: 1 = at school and unmarried, 2 = at school and married, 3 = out of school and unmarried, 4 = out of school and married, 5 = dead)

process $\{N_1(t,a), a \geq 0, t = t(a)\}$ as a stochastic process giving the education status with $N_1(t,a) = 1$ indicating being in school and $N_1(t,a) = 0$ having stopped formal education (out of school) by age a at time $t(a)$. Similarly, the marriage process $\{N_2(t,a), a \geq 0, t = t(a)\}$ is a stochastic process giving the marital status of a woman aged a at time $t(a)$, with $N_2(t,a) = 0$ indicating unmarried and $N_2(t,a) = 1$ married status. The corresponding histories are defined as $\mathcal{F}_r(t,a) = \{N_r(s,u), u \leq a, s(u) \leq t(a)\}$, $r = 1, 2$, respectively and the joint history as $\mathcal{F}(t,a) = \{(N_1(s,u), N_2(s,u)), u \leq a, s \leq t\}$. Note that the four states defined earlier correspond to the at-school and marriage processes taking values $(1,0), (1,1), (0,0), (0,1)$, respectively, and the state 5 corresponds to dead.

The counting process $N_1(t,a)$ remains at zero between the age 0 and a_e, that is between the birth year t_0 and the year $t(a_e)$. Because of minimum marriageable age $a_0 (> a_e)$, the process $N_2(t,a)$ is zero for all age $a < a_0$ and calendar time $t < t(a_0)$. The association between the two processes is modelled through the dependence on the joint history $\mathcal{F}(t,a)$. Because the two times grow together with the same pace we denote the history using only one time scale as $\mathcal{F}(a)$. Time invariant information or fixed covariates at birth (x) such as religion and caste are also included in this history. We

also construct a deterministic counting process giving schooling years of a woman aged a at time $t(a)$, as the accumulated history of at-school process $\{\int_{0\le u\le a} N_1(t(u),u)\,\mathrm{d}u\}$.

The intensities of making transition from at-school to out of school state given that the marriage process is in state $k=0,1$, and the history of the processses are defined as

$$\alpha_{1,k}(t,a \mid \mathcal{F}(a^-)) = \lim_{\Delta t\to 0} \frac{P(N_1(t+\Delta t, a+\Delta t)=0 \mid N(t^-,a^-)=(1,k),\mathcal{F}(a^-))}{\Delta t},$$

and similarly, the marriage intensities of making transition from unmarried to married state given that the at-school process is in state $k=0,1$, are defined as

$$\lambda_{k,0}(t,a \mid \mathcal{F}(a^-)) = \lim_{\Delta t\to 0} \frac{P(N_2(t+\Delta t, a+\Delta t)=1 \mid N(t^-,a^-)=(k,0),\mathcal{F}(a^-))}{\Delta t},$$

where $N(t,a)=(N_1(t,a),N_2(t,a))$ and $k=0,1$. Since the process $N_1(t,a)=1$ until the transition happens, we drop the subscript 1 and simplify the notation $\alpha_{1,k}(t,a \mid \mathcal{F}(a^-))$ to $\alpha_k(t,a)$. Similarly, the process $N_2(t,a)=0$ until the transition happens, and hence, we use $\lambda_k(t,a \mid \mathcal{F}(a^-))$ as a simplified notation for $\lambda_{k,0}(t,a)$. Furthermore, we define $\mu_{jk}(t,a)$ to be the intensity of moving to state 5 (dead), where $j,k\in\{0,1\}$ represent the current schooling and marriage status, respectively. Note that the transition intensities are defined in relation to the at-risk process while the transition rates describe how the process evolve over time.

Figure 3 exhibits an example sample paths of the two processes based on the retrospective information collected at the cross-sectional age of 25. In the example, the formal school ends at the age of 18 and marriage takes place at the age of 21. The at-school process remains in state 1 between the age of 6 years ($a_e=6$) and 18 years, then jumps to state 0. The marriage process

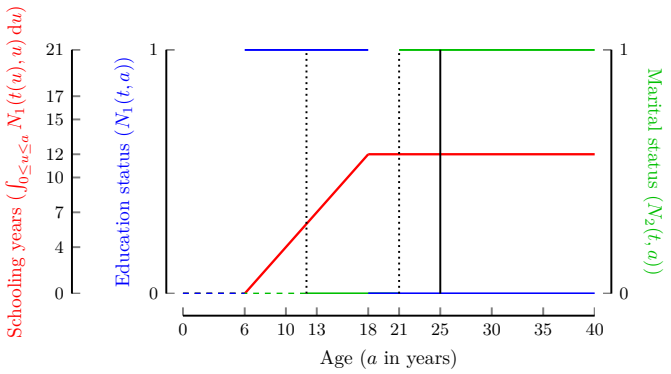

Figure 3: A sample path of at-school (blue) and marriage (green) processes. Basic education starts at the age of 6 years and marriageable age is 12 years. The inner left y-axis indicates education status 0 (= out of school) and 1 (= at school) and the outer left y-axis gives the accumulated schooling years. The right y-axis is the marital status axis, 0 (= unmarried) and 1 (= married). Dashed black lines indicate observation period relevant for marriage process and solid black line indicates the cross-section.

starts at the age of 12 years ($a_0 = 12$) in state unmarried (0) and jumps to state married (1) at the age of 21 years. In addition, the deterministic counting process giving the accumulated schooling years is also shown. The observation process stops at the first marriage and our main interest in the joint processes is in the time interval between the age a_0 and the age at the first marriage. In the observation period, the multi-state process starts in state 1 at age 12 and calendar year $(t_0 + 12)$ and moves to state 3 before transitioning to state 4 at age 21 and calendar year $(t_0 + 21)$. The schooling years at that time are 12 which are attained at the age of 18. In principle, changes in the at-school process after marriage can be inferred based on the method given below subject to the availability of data.

We derive the likelihood contributions for all possible event histories conditional on being alive in either state 1 or 3 at age a_e. Let us first consider an individual born in the calendar time t_0 and in state 1 (at school

and unmarried) at age a_e. Figure 2 shows possible transitions between the five states. We develop the model following notation of Keiding (1991), extending it to include two correlated processes. The probability density of being unmarried and at school with $w = z - a_e$ years of schooling and aged $[z, z + \mathrm{d}z)$ at time t, is proportional to $\beta_1(t - z, a_e)k_1(t, z, w)$, where $\beta_1(t - z, a_e)$ is the probability density of being born in year $t - z$ and being in state 1 at age a_e and

$$k_1(t,z,w) = \exp\left\{-\int_{a_e}^{z} [\mu_{10}(t-z+u,u) + \alpha_0(t-z+u,u)]\,\mathrm{d}u\right\} \times \exp\left\{-\int_{a_0}^{z} \lambda_1(t-z+u,u)\,\mathrm{d}u\right\}. \tag{3.1}$$

Similarly, the probability density of being unmarried and out of school with w years of schooling and alive aged $[z, z + \mathrm{d}z)$ at time t, is proportional to $\beta_1(t - z, a_0)k_0(t, z, w)\,\mathrm{d}z$ where

$$\begin{aligned} k_0(t,z,w) = {} & \exp\left\{-\int_{a_e}^{a_w} \alpha_0(t-z+u,u)\,\mathrm{d}u\right\} \\ & \times \alpha_0(t-z+a_w,a_w)^{\mathbf{1}_{\{a_e<a_w\le z\}}} \\ & \times \exp\left\{-\int_{a_e}^{z} \mu_{\mathbf{1}_{\{u<a_w\}}0}(t-z+u,u)\,\mathrm{d}u\right\} \\ & \times \exp\left\{-\int_{a_0}^{z} \lambda_{\mathbf{1}_{\{u\le a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\}, \end{aligned} \tag{3.2}$$

where $a_w = a_e + w < z$, is the age when school ended with w years of schooling attained.

Equations (3.2) and (3.1) can be combined and rewritten as follows.

$$\begin{aligned} k(t,z,w) = {} & \exp\left\{-\int_{a_e}^{\min(a_w,z)} \alpha_0(t-z+u,u)\,\mathrm{d}u\right\} \\ & \times \alpha_0(t-z+a_w,a_w)^{\mathbf{1}_{\{a_e<a_w\le z\}}} \\ & \times \exp\left\{-\int_{a_e}^{z} \mu_{\mathbf{1}_{\{u<a_w\}}0}(t-z+u,u)\,\mathrm{d}u\right\} \\ & \times \exp\left\{-\int_{a_0}^{z} \lambda_{\mathbf{1}_{\{u<a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\}. \end{aligned} \tag{3.3}$$

Similarly, the probability density of being married and having w years of schooling, alive and aged $[z, z + \mathrm{d}z)$ at time t and the first marriage at age

$[y, y + \mathrm{d}y)$ is proportional to $\beta_1(t-z, a_0)h(t, y, z, w)\,\mathrm{d}y\,\mathrm{d}z$ where $h(t, y, z, w)$ is defined as

$$\begin{aligned} h(t,y,z,w) = \exp&\left\{-\int_{a_e}^{\min(a_w,z)} \alpha_{\mathbf{1}_{\{y<u\}}}(t-z+u,u)\,\mathrm{d}u\right\} \\ &\times \alpha_{\mathbf{1}_{\{y<a_w\}}}(t-z+a_w,a_w)^{\mathbf{1}_{\{a_e<a_w\le z\}}} \\ &\times \exp\left\{-\int_{a_e}^{z} \mu_{\mathbf{1}_{\{u<a_w\}}\mathbf{1}_{\{y<u\}}}(t-z+u,u)\,\mathrm{d}u\right\} \\ &\times \exp\left\{-\int_{a_0}^{y} \lambda_{\mathbf{1}_{\{u<a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\} \\ &\times \lambda_{\mathbf{1}_{\{y<a_w\}}}(t-z+y,y). \end{aligned} \tag{3.4}$$

The likelihood contributions of individuals starting in state 3 (women who received no education) are defined similarly, but multiplied by $\beta_0(t-z, a_e)$, which is the probability density of being born in year $t-z$ and being in state 3 at age a_e, and taking $a_w = a_e$, in which case the α intensities do not appear in (3.1)-(3.4).

The probability density of the sampling event of being married, having w years of schooling, alive and aged $[z, z + \mathrm{d}z)$ at time t is

$$\int_{a_0}^{z} \beta_{\mathbf{1}_{\{a_e<a_w\}}}(t-z,a_e)h(t,y,z,w)\,\mathrm{d}y.$$

Alternatively, if we were interested in estimating intensities for both marriage and ending formal education, we could write the likelihood without conditioning on the education history. However, because our interest is in marriage intensity, we write the likelihood conditional on the education history, and consider conditions under which we can simplify the likelihood into a function of the marriage intensities alone.

The conditional likelihood contributions of individuals $i \in C_2$, in the prevalent cohort, e.g. NFHS-2, at time t_2 are

$$\begin{aligned} \prod_{i\in C_2} L_{2i}(\theta) &= \prod_{i\in C_2} \frac{\beta_{\mathbf{1}_{\{a_e<a_{w_i}\}}}(t_2-z_i,a_e)h(t_2,y_i,z_i,w_i)}{\int_{a_0}^{z_i}\beta_{\mathbf{1}_{\{a_e<a_{w_i}\}}}(t_2-z_i,a_e)h(t_2,v,z_i,w_i)\,\mathrm{d}v} \\ &= \prod_{i\in C_2} \frac{h(t_2,y_i,z_i,w_i)}{\int_{a_0}^{z_i} h(t_2,v,z_i,w_i)\,\mathrm{d}v}. \end{aligned}$$

The likelihood can be simplified under either of the following assumptions related to the counting process for number of schooling years, combined with the assumption that mortality is non-differential with respect to the marriage status, i.e. that $\mu_{j0}(t,a) = \mu_{j1}(t,a) = \mu_j(t,a)$ for $j = 0, 1$.

A1. Schooling ends always before marriage or the intensity of stopping schooling after marriage is negligible.

A2. The intensities of stopping schooling are non-differential. That is the intensities of stopping schooling are the same before and after marriage, and do not depend on the history of the marriage process $\mathcal{F}_2(t,a) = \{N_2(s,u), u \leq a, s(u) \leq t(a)\}$. In other words, this assumption states that the education process is locally independent of the marriage process; $\alpha_0(t, a \mid \mathcal{F}(a^-)) = \alpha_1(t, a \mid \mathcal{F}(a^-)) = \alpha(t, a \mid \mathcal{F}_1(a^-))$. (Cook and Lawless, 2018)

3.1 Retrospective cohort study I: likelihood under the assumptions of non-differential mortality and A2

Under the above-mentioned assumptions, equation (3.4) reduces to

$$\begin{aligned} h(t,y,z,w) = \exp&\left\{-\int_{a_e}^{\min(a_w,z)} \alpha(t-z+u,u)\,\mathrm{d}u\right\} \\ &\times \alpha(t-z+a_w,a_w)^{\mathbf{1}_{\{a_e<a_w\leq z\}}} \\ &\times \exp\left\{-\int_{a_e}^{z} \mu_{\mathbf{1}_{\{u<a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\} \\ &\times \exp\left\{-\int_{a_0}^{y} \lambda_{\mathbf{1}_{\{u<a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\} \\ &\times \lambda_{\mathbf{1}_{\{y<a_w\}}}(t-z+y,y). \end{aligned} \tag{3.5}$$

The normalising factor becomes

$$\begin{aligned}\int_{a_0}^{z} h(t,y,z,w)\,\mathrm{d}y = \exp\left\{-\int_{a_e}^{\min(a_w,z)} \alpha(t-z+u,u)\,\mathrm{d}u\right\} \\ \times \alpha(t-z+a_w,a_w)^{\mathbf{1}_{\{a_e<a_w\le z\}}} \\ \times \exp\left\{-\int_{a_e}^{z} \mu_{\mathbf{1}_{\{u<a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\} \\ \times \left(1-\exp\left\{-\int_{a_0}^{z} \lambda_{\mathbf{1}_{\{u<a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\}\right).\end{aligned} \tag{3.6}$$

Now the terms containing α and μ cancel out from the conditional likelihood, giving the likelihood contribution of an individual $i \in C_2$ in the prevalent cohort, e.g. NFHS-2, conditioned on the sampling event as

$$L_{2i}(\theta) = \frac{\exp\left\{-\int_{a_0}^{y_i} \lambda_{\mathbf{1}_{\{u<a_{w_i}\}}}(t_2-z_i+u,u)\,\mathrm{d}u\right\} \lambda_{\mathbf{1}_{\{y_i<a_{w_i}\}}}(t_2-z_i+y_i,y_i)}{1-\exp\left\{-\int_{a_0}^{z_i} \lambda_{\mathbf{1}_{\{u<a_{w_i}\}}}(t_2-z_i+u,u)\,\mathrm{d}u\right\}}. \tag{3.7}$$

If we don't include the number of schooling years in the sampling event then the denominator will have to be integrated with respect to w as well as

$$\int_{a_0}^{z}\int_{a_e}^{z} \beta_{\mathbf{1}_{\{a_e<a_w\}}}(t-z,a_0) h(t,y,z,w)\,\mathrm{d}w\,\mathrm{d}y.$$

The above expression can be simplified under the assumptions A1 or A2 but the intensities α do not cancel out, so the resulting likelihood can be used for estimating parameters characterising the education process, if these are of interest.

3.2 Retrospective cohort study II: likelihood under the assumptions of non-differential mortality and A2

The conditional probability density of the sampling event of being alive with schooling years w and aged z at time t is the sum of the probabilities of being unmarried, alive with schooling years w and aged z at time t, and married, alive with schooling years w and aged z at time t. This is given

by $\beta_{\mathbf{1}_{\{a_e<a_w\}}}(t-z,a_e)[k(t,z,w)+\int_{a_0}^{z}h(t,y,z,w)\,\mathrm{d}y]$. Thus, the conditional likelihood contributions of individuals $i \in C_3$ in the general cohort, e.g. NFHS-3, at time t_3 are

$$L_3(\theta)=\prod_{i\in C_3}\frac{\beta_{\mathbf{1}_{\{a_e<a_{w_i}\}}}(t_3-z_i,a_0)h(t_3,y_i,z_i,w_i)^{\delta_i}k(t_3,z_i,w_i)^{1-\delta_i}}{\beta_{\mathbf{1}_{\{a_e<a_{w_i}\}}}(t_3-z_i,a_0)[k(t_3,z_i,w_i)+\int_{a_0}^{z_i}h(t_3,u,z_i,w_i)\,\mathrm{d}u]}, \tag{3.8}$$

where $\delta_i \equiv \mathbf{1}_{\{y_i\le z_i\}}$ is an indicator of marital status at time t_3. Under the assumptions A2 and non-differential mortality with respect to marriage status, as before, (3.3) reduces to

$$\begin{aligned}k(t,z,w)=&\exp\left\{-\int_{a_e}^{\min(a_w,z)}\alpha(t-z+u,u)\,\mathrm{d}u\right\}\\&\times\alpha(t-z+a_w,a_w)^{\mathbf{1}_{\{a_e<a_w\le z\}}}\\&\times\exp\left\{-\int_{a_e}^{z}\mu_{\mathbf{1}_{\{u<a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\}\\&\times\exp\left\{-\int_{a_0}^{z}\lambda_{\mathbf{1}_{\{u<a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\}\end{aligned} \tag{3.9}$$

and $h(t,u,z,w)$ to (3.6). Combining these, the normalising factor becomes

$$\begin{aligned}&k(t,z,w)+\int_{a_0}^{z}h(t,u,z,w)\,\mathrm{d}u\\&\quad=\exp\left\{-\int_{a_e}^{\min(a_w,z)}\alpha(t-z+u,u)\,\mathrm{d}u\right\}\alpha(t-z+a_w,a_w)^{\mathbf{1}_{\{a_e<a_w\le z\}}}\\&\qquad\times\exp\left\{-\int_{a_e}^{z}\mu_{\mathbf{1}_{\{u<a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\},\end{aligned} \tag{3.10}$$

which will cancel out with the similar term in the numerator of the conditional likelihood. Thus, under these assumptions, the likelihood (3.8) under the retrospective cross-sectional design II reduces to the standard likelihood for right censored survival data, given by

$$\begin{aligned}L_3(\theta)=\prod_{i\in C_3}\Bigg[&\lambda_{\mathbf{1}_{\{y_i<a_w\}}}(t_3-z_i+y_i,y_i)^{\delta_i}\\&\times\exp\left\{-\int_{a_0}^{\min(y_i,z_i)}\lambda_{\mathbf{1}_{\{u<a_w\}}}(t-z+u,u)\,\mathrm{d}u\right\}\Bigg].\end{aligned} \tag{3.11}$$

It is to be noted that the above likelihood expressions are constructed by explicitly conditioning on the calendar time of the survey, the age and schooling status of the individual at the time of the survey. This is equivalent to conditioning on the individual's birth cohort, and hence, the birth rate cancels out and the likelihood expressions simplify by assuming non-differential mortality and either A_1 or A_2 only. If the conditional likelihood were derived by conditioning only on the age range used for the sampling and not on the exact age of individuals then the probability of the conditioning event needed to be integrated over z also. The same applies for education status. In this case, stricter assumptions would be needed to carry out the estimation of the incidence rate or external information on mortality, education as well as birth rates would be required. Such information may not be available for all the stratifying groups that we will use in the real application. In the following we use a likelihood conditioned on the covariates and the sampling scheme for estimating the marriage incidence rate. The likelihood is a product of $L_2(\theta)$ and $L_3(\theta)$ from the cohorts under design I and II, respectively. We show in Appendix B that this is indeed a likelihood and hence, the maximum likelihood theory applies for estimation of θ.

4 Predictive probabilities

In societies experiencing fertility decline and manifesting imbalanced sex ratio at birth, marriage markets are affected by the marriage squeeze (Guilmoto, 2012, Neelakantan and Tertilt, 2008, Schoen, 1983). 'Marriage squeeze' refers to excess supply of marriageable females or males within the endogamy of religion, caste, language and characterised by age and education. The marriage squeeze reflects the ways in which observed males' and females' age-specific marriage rates accommodate themselves to changes in the age-sex composition and education of the population when the underlying marriage preferences remain unchanged (Schoen, 1983). Prediction of marriage squeeze has been one of the main aims of demographic models. Estimates of

marriage intensities accommodating endogenous variables including education can be used to calculate the probabilities of transition from unmarried to married state in the future. Further, such predictive probabilities could be used to evaluate the extent of marriage squeeze.

Predictive probabilities defined in this section are the probabilities of transition from unmarried to married state in the future based on existing data. The data are used to estimate the parameters characterising the marriage process, and the predictions are calculated at the estimated parameter values. Given characteristics x, we might be interested in the predictive probability of an unmarried woman aged a_1 ($\geq a_0$) at time t and schooling years w years being married before age a_2. Because education is time-dependent, generally calculation of these kinds of probabilities would involve prediction of future education also. However, for women who already reached their highest level of education (i.e. $a_w < a_1$), we can predict based on marriage intensity and mortality estimates alone. Such predictive probability for fixed schooling years is given by the cumulative incidence

$$\begin{aligned}
&\text{PredProb}(a_2 \mid t, a_1, w) \\
&\quad = \frac{\int_{a_1}^{a_2} k(t-a_1+a, a, w)\lambda_0(t-a_1+a, a; \theta)\,\mathrm{d}a}{k(t, a_1, w)} \\
&\quad = \frac{\int_{a_1}^{a_2} k_2(t-a_1+a, a, w)\lambda_0(t-a_1+a, a; \theta)\,\mathrm{d}a}{k_2(t, a_1, w)},
\end{aligned} \tag{4.1}$$

where

$$\begin{aligned}
&k_2(s, a, w) \\
&\quad = \exp\left\{-\int_{a_0}^{a} [\mu_{00}(s-a+u, u) + \lambda_0(s-a+u, u; \theta)\}]\,\mathrm{d}u\right\}.
\end{aligned}$$

Another predictive probability of interest is that of an unmarried woman, with characteristics x and aged a_1 at time t and schooling years w_1 being married before age a_2, and being alive at a_2, and is given by (for fixed

schooling years)

$$\frac{\int_{a_1}^{a_2} h(t-a_1+a_2,a,a_2,w_1)\,\mathrm{d}a}{k(t,a_1,w_1)} = \frac{\int_{a_1}^{a_2} h_2(t-a_1+a_2,a,a_2,w_1)\,\mathrm{d}a}{k_2(t,a_1,w_1)}, \tag{4.2}$$

where k_2 is defined above and h_2 is obtained from h in (3.4) by dropping terms corresponding to education process. Under the assumption of non-differential mortality with respect to both education and marriage, and possibly other covariates used to model marriage intensity, mortality rates based on official statistics can be used in the calculation.

The first predictive probability (4.1) appears to be important for population models since it gives the proportion ever getting married, which multiplied by the population count of age a_1 (with characteristics x) gives the ever-married population count. The second one (4.2) might be important for questions like: what proportion of women of age a_0 get married and live until through a typical "child-bearing age" a_1. Note that the mortality rate is needed in order to compute above probabilities. We demonstrate the former kind of predictive probabilities in Section 6.

5 Simulation study

We conducted a simulation study to assess the efficiency gain achieved by combining data from the two retrospective cohort studies, compared to analysing each one of these separately, as well as to study the impact of various misspecification scenarios on the parameter estimates. We simulated data from a multi-state model with states similar to Figure 2. The

model was specified through the transition rate matrix

	1	2	3	4	5
1	.	$\lambda_1 = e^{m+bx+c}$	$\alpha_0 = e^{s+bx}$	0	$\mu_{10} = e^{r+bx+c}$
2	0	.	0	$\alpha_1 = e^{s+bx+d}$	$\mu_{11} = e^{r+bx+c+g}$
3	0	0	.	$\lambda_0 = e^{m+bx}$	$\mu_{00} = e^{r+bx}$
4	0	0	0	.	$\mu_{01} = e^{r+bx+g}$
5	0	0	0	0	.

where the parameters of interest are m, characterising the baseline marriage rate, b, characterising the effect of a time constant covariate x (taking values 1 or 0 with probability 0.5) on marriage rate, and c, characterising the effect of being in school on marriage rate, as well as on mortality rate. Parameters d and g characterise the effect of marriage on ending formal education and mortality, respectively. Note that $d = g = 0$ under the non-differential assumption. It is to be noted that the simulation model assumes constant rates and hence, the term transition intensity is replaced by transition rate. The initial state of the multi-state model at age $a_0 = 12$ was drawn randomly with probabilities $\text{expit}(-1 + 0.5x)$ for state 3 (unmarried, out of school) and $1 - \text{expit}(-1 + 0.5x)$ for state 1 (unmarried, at school). A cross-sectional cohort was constructed by drawing year of birth uniformly from $[1965, 1993]$, and taking 2005 as the time of the cross-sectional survey, at which time the age range was $[12, 40]$. The cohort under design I, cohort I, was constructed by simulating 2,500 event histories and including only individuals in the married states 2 and 4 at the time of the cross-section, while the one under design II, cohort II, was constructed by simulating 2,500 event histories and including individuals in states 1-4 at the cross-section. To these data, we fitted constant rate marriage models through maximizing the joint likelihood expression using data from the two cohorts, the product of (3.7) & (3.11), cohort I data only, (3.7), the likelihood expression (3.7) without the correction term in the denominator, and cohort II data only, (3.11). The data generation and model fitting were repeated 1,000 times, resulting in average sizes of the two cohorts as (1,864, 2,229), respectively

under the non-differential scenario. The likelihood expressions were maximised numerically using the R `optim` function (R Core Team, 2020), with standard errors calculated by inverting the numerically differentiated Hessian matrix at the maximum likelihood point.

The results under the non-differential scenario are given in Table 2. The results indicate that there is a clear efficiency gain (in terms of the Monte Carlo standard deviation of the point estimates) in combining the analysis of the two cohorts, as opposed to analysing each of them separately. The three types of parameters, baseline marriage intensity m, effect of a time-constant covariate b, and effect of time-dependent covariate c can be estimated without bias, with the cohort I likelihood needing the correction term in the denominator to account for the sampling mechanism. The results under the scenario of differential education process intensities are given in Table 3. These indicate that violation of the non-differential assumption for stopping school mainly causes bias in the estimated effect of ending school on marriage incidence, while the other two parameters are much less affected. The retrospective cohort likelihood under design I is more susceptible to this type of bias, but it is fairly small in all cases. Differential mortality (Table 4) on the other hand causes bias in the baseline marriage incidence estimates with both types of likelihood expression, with the covariate effect estimates affected much less. Finally, both types of non-differential assumptions are combined in the scenario of Table 5, with the two different types of biases essentially adding up. In summary, the simulation results confirm the efficiency gain in combining two types of retrospective cross-sectional cohort data, while demonstrating how the parameter estimates are affected by violations of the assumptions used in deriving the simplified likelihood expressions. Because the effect of the violations was relatively small, we proceed under the non-differential assumptions in the data analysis of Section 6.

Table 2: Results from 1000 simulation rounds under non-differential mortality and stopping school ($d = g = 0$). Mean stands for mean point estimate, MC SD for Monte Carlo standard deviation of the point estimates, Mean SE for mean estimated standard error, and Coverage for 95% confidence interval coverage probability.

Likelihood	Parameter	Truth	Mean	Bias	MC SD	Mean SE	Coverage
(3.7) & (3.11)	m	-1.500	-1.501	-0.001	0.029	0.030	0.947
	b	0.500	0.501	0.001	0.037	0.037	0.951
	c	-0.500	-0.499	0.001	0.038	0.039	0.943
(3.7)	m	-1.500	-1.501	-0.001	0.049	0.050	0.954
	b	0.500	0.501	0.001	0.059	0.060	0.942
	c	-0.500	-0.499	0.001	0.059	0.058	0.943
(3.7)	m	-1.500	-1.264	0.236	0.034	0.037	0.000
w/o	b	0.500	0.394	-0.106	0.044	0.046	0.380
correction	c	-0.500	-0.474	0.026	0.052	0.052	0.901
(3.11)	m	-1.500	-1.501	-0.001	0.038	0.037	0.941
	b	0.500	0.501	0.001	0.047	0.046	0.947
	c	-0.500	-0.500	0.000	0.052	0.052	0.947

Table 3: Results from 1000 simulation rounds under non-differential mortality ($g = 0$) and differential stopping school ($d = 1$).

Likelihood	Parameter	Truth	Mean	Bias	MC SD	Mean SE	Coverage
(3.7) & (3.11)	m	-1.500	-1.507	-0.007	0.030	0.030	0.948
	b	0.500	0.502	0.002	0.036	0.037	0.951
	c	-0.500	-0.470	0.030	0.038	0.038	0.864
(3.7)	m	-1.500	-1.516	-0.016	0.052	0.051	0.944
	b	0.500	0.501	0.001	0.060	0.060	0.953
	c	-0.500	-0.429	0.071	0.057	0.057	0.752
(3.7)	m	-1.500	-1.264	0.236	0.035	0.037	0.000
w/o	b	0.500	0.395	-0.105	0.045	0.046	0.370
correction	c	-0.500	-0.478	0.022	0.051	0.052	0.932
(3.11)	m	-1.500	-1.500	-0.000	0.038	0.037	0.935
	b	0.500	0.502	0.002	0.048	0.046	0.948
	c	-0.500	-0.505	-0.005	0.052	0.052	0.947

Table 4: Results from 1000 simulation rounds under differential mortality ($g = 1$) and non-differential stopping school ($d = 0$).

Likeli-hood	Para-meter	Truth	Mean	Bias	MC SD	Mean SE	Cover-age
(3.7) & (3.11)	m	-1.500	-1.547	-0.047	0.032	0.032	0.709
	b	0.500	0.498	-0.002	0.040	0.041	0.954
	c	-0.500	-0.497	0.003	0.042	0.042	0.945
(3.7)	m	-1.500	-1.557	-0.057	0.058	0.056	0.831
	b	0.500	0.501	0.001	0.070	0.069	0.946
	c	-0.500	-0.496	0.004	0.067	0.065	0.943
(3.7)	m	-1.500	-1.283	0.217	0.038	0.039	0.001
w/o	b	0.500	0.393	-0.107	0.050	0.051	0.443
correction	c	-0.500	-0.467	0.033	0.057	0.056	0.904
(3.11)	m	-1.500	-1.543	-0.043	0.039	0.039	0.811
	b	0.500	0.497	-0.003	0.050	0.051	0.951
	c	-0.500	-0.499	0.001	0.058	0.056	0.938

Table 5: Results from 1000 simulation rounds under differential mortality and stopping school ($g = d = 1$).

Likeli-hood	Para-meter	Truth	Mean	Bias	MC SD	Mean SE	Cover-age
(3.7) & (3.11)	m	-1.500	-1.555	-0.055	0.032	0.032	0.595
	b	0.500	0.498	-0.002	0.039	0.041	0.962
	c	-0.500	-0.469	0.031	0.043	0.042	0.874
(3.7)	m	-1.500	-1.577	-0.077	0.058	0.057	0.741
	b	0.500	0.504	0.004	0.068	0.069	0.952
	c	-0.500	-0.419	0.081	0.065	0.063	0.748
(3.7)	m	-1.500	-1.282	0.218	0.038	0.039	0.000
w/o	b	0.500	0.394	-0.106	0.049	0.051	0.472
correction	c	-0.500	-0.478	0.022	0.057	0.056	0.921
(3.11)	m	-1.500	-1.542	-0.042	0.040	0.039	0.809
	b	0.500	0.496	-0.004	0.052	0.051	0.944
	c	-0.500	-0.509	-0.009	0.057	0.056	0.947

6 Analysis of marriage incidence in India and four states based on NFHS-2 and -3 data

6.1 Model

Our main focus is on estimation of marriage incidence and hence, we apply estimation method based on likelihood expressions (3.7) and (3.11). The NFHS data did not include complete retrospective histories of education and hence, we constructed histories of education process up to the time of marriage by employing the structure of the Indian education system and assuming that everyone starts school at the same age a_e, and stays at school continuously after that until stopping.

For the purpose of illustration, we considered only four states, Kerala, Maharashtra, Punjab and Rajasthan, as described in Section 2 which are geographically spread across India and differ by way of literacy rates, women's position, and sex-ratio at birth. From the cross-sectional data from NFHS-2 and NFHS-3 surveys, the age a and calendar time t effects turned out to be strongly correlated, so instead of estimating age- and calendar period-specific marriage incidence rates, we used age as the main time scale of the analysis, and birth cohort as a covariate. Table 1 lists the covariates used in the marriage incidence model. We applied a proportional hazards model in which the covariates act multiplicatively on an age-dependent baseline rate, assumed piecewise constant over one-year age intervals except for the first and the last intervals. This results in 17 age bands $[12, 15), [15, 16), \ldots, [29, 30)$, and $[30, 50)$ years, denoted by $[a_j, a_{j+1})$, $j = 1, \ldots, 17$.

The effect of education was modeled by defining the marriage incidence as a function of the current highest education level being attempted, defining the education level x_{5j} at age band j as a time-dependent covariate by modifying the woman's highest attained level x_5, recorded at the time of survey, so that $x_{5j} = \min(x_5, 1)$ when in age band $j = 1$, $x_{5j} = \min(x_5, 2)$ when $j \in \{2, 3, 4\}$, and $x_{5j} = x_5$ otherwise. For example, consider a woman

aged 25 years at the time of survey, married at the age of 21 years, has reported education level *Secondary* ($x_5 = 2$, cf. Figure 3 and Table 1). In the analysis, her contribution to the education variable will be *Primary* in the age band $[12, 15)$ and *Secondary* in the bands $[15, 16)$, $[16, 17)$, $\cdots$, $[20, 21)$ and $[21, 22)$ to span the age range from 12 years to the age at her marriage.

To sum up, the model for the marriage incidence rate $\lambda(a; x, \theta) = \lambda_j(x, \theta)$, $a \in [a_j, a_{j+1})$, $j = 1, \ldots, 17$, conditional on covariate values x, Table 1, was specified as

$$\log\{\lambda_j(x,\theta)\} = \alpha_j + \sum_{i=1}^{3} \beta_{1i} \mathbf{1}_{\{x_1=i\}} + \beta_2 \mathbf{1}_{\{x_2=1\}} + \sum_{i=1}^{3} \beta_{3i} \mathbf{1}_{\{x_3=i\}} + \sum_{i=1}^{4} \beta_{4i} \mathbf{1}_{\{x_4=i\}} + \sum_{i=1}^{3} \beta_{5i} \mathbf{1}_{\{x_{5j}=i\}}, \tag{6.1}$$

with 31 parameters, including 17 log-baseline rates and 14 covariate effects, log-rate ratios. The same model was fitted for each state separately, and in addition to all-India data (all 29 states) to assess how the state-specific patterns differ from the national pattern, by maximising the product of likelihood expressions of the form (3.7) and (3.11), but because the age at marriage was only reported at the precision of one year, expressing the numerator contribution for married women as

$$\int_{\lfloor y \rfloor}^{\lceil y \rceil} \lambda(a; x, \theta) \exp\left\{ - \int_{a_0}^{a} \lambda(u; x, \theta)\, \mathrm{d}u \right\} \mathrm{d}a = \exp\left\{ - \int_{a_0}^{\lfloor y \rfloor} \lambda(u; x, \theta)\, \mathrm{d}u \right\} \left[1 - \exp\left\{ - \int_{\lfloor y \rfloor}^{\lceil y \rceil} \lambda(u; x, \theta)\, \mathrm{d}u \right\} \right]$$

where $\lfloor y \rfloor$ and $\lceil y \rceil = \lfloor y \rfloor + 1$ denote the floor and ceiling of the exact age y at which the marriage took place. The joint likelihood expression was maximised with respect to the parameter vector θ using the `optim` function of the R statistical environment (R Core Team, 2020). The standard errors were evaluated by inverting the numerically differentiated observed information matrix at the maximum likelihood point. The results were presented as

point estimates and 95% confidence intervals. Of note, by letting the marriage rate depend on the birth cohort, the third possible time scale (calendar time) can be omitted.

6.2 Results

Figure 4 presents the estimated age-specific baseline marriage rates in the four Indian states and in all India. Although the hazard of first marriage after age 30 has remained low in each state, different patterns emerge otherwise. The rate is generally lowest in Kerala, in particular in comparison to Maharashtra and Punjab. In Rajasthan, the rate starts increasing earliest in age.

Figure 5 shows the estimated covariate effects on the marriage rates. The rate decreases by birth cohort x_1, except for Punjab where the rate is the highest for the 1972-1982 cohort ($x_1 = 2$). By the last cohort (1982-1992, $x_1 = 3$) in this analysis, the rates have declined considerably in all four states. Since this birth cohort, being 6-16 years of age at the time of survey, was underrepresented in NFHS-2, we repeated the analysis by using only the NFHS-3 data and the estimates of marriage rates were essentially unchanged (results not shown).

Unsurprisingly, women in rural areas ($x_2 = 1$) have a larger rate of marriage (all India incidence rate ratio of 1.19) compared to urban areas ($x_2 = 0$, ref.), except in Punjab where the reverse is true. The higher rate in rural areas is particularly striking in Kerala and Maharashtra. At the India level, the marriage rates are similar for OBC ($x_3 = 2$) and SC ($x_3 = 0$, ref.) while ST ($x_3 = 1$) and Other caste ($x_3 = 3$) have lower marriage rates. However, this pattern is not evident in all of the four the state-level results. In Punjab, the confidence interval for ST is wide because this caste is rare (Table 6).

There are clear differences in the marriage rate across religions (x_4). At the India level, the marriage incidence rates are clearly smaller in Christian

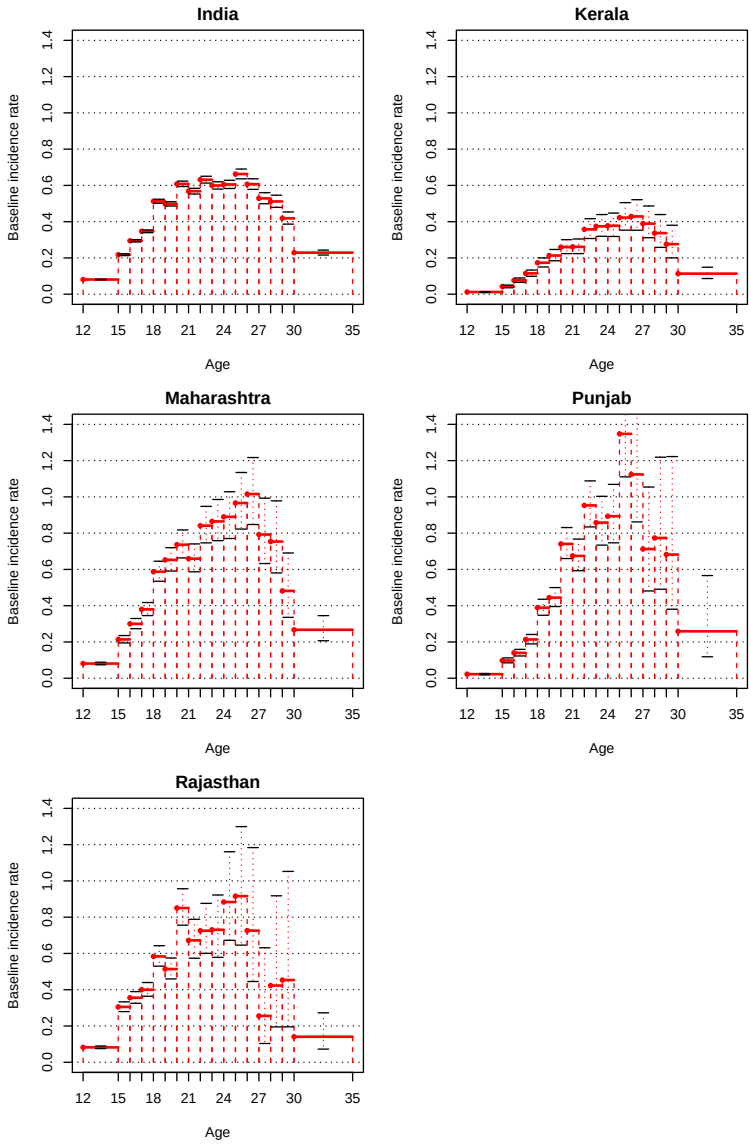

Figure 4: Age-specific baseline rates for a woman to marry in India and the four selected states. The horizontal lines show the maximum likelihood estimates of parameters $\exp\{\alpha_j\}$ in (6.1), and their corresponding 95% confidence intervals.

($x_4 = 2$), Sikh ($x_4 = 3$) and other religions ($x_4 = 4$) as compared to Hindu ($x_4 = 0$, ref.). The same pattern emerges in the state-level analysis, except for Muslims ($x_4 = 1$) in Kerala. Again, to interpret the state specific results we note that not all religions were sufficiently represented in each state (Table 6).

The effect of education (x_5) is evident. There is a clear decrease in the incidence rate when moving from no education ($x_5 = 0$, ref.) to higher education levels ($x_5 = 3$) in India and in all the four states. In the all India analysis, the incidence rate for a woman with primary education ($x_5 = 1$) to marry at any given age is about half that for a woman with no education. The corresponding rates are 31% and 28% of the uneducated rate for a woman with secondary ($x_5 = 2$) and higher education. The same patterns shows up in all four states although the effect of education level is relatively smaller in Kerala.

Predictive probabilities of type (4.1) for marrying by age a_1 were calculated as discussed in Section 4, with $a_{\min} = a_0 = 12$, using 2010 mortality rates based on census data, and marriage incidence rates corresponding to different calendar periods (Figure 6). The covariate values were set to the reference categories (urban area, scheduled caste, Hindu religion, and uneducated). Clearly, the women's absolute probability of marrying by late twenties has remained consistently high, but in Maharashtra there has been a clear shift towards marrying at a comparatively higher age. The patterns in Kerala and Rajasthan are more difficult to interpret, as the high estimated marriage rates in late twenties in the later calendar periods actually results also in higher projected absolute probabilities in late twenties. However, this projection does not reflect all the changes in the background population, since the overall education level has increased over time, bringing the population marriage incidence rates down, while in this projection education was fixed to the reference level. In Punjab, any changes over time have been comparatively small.

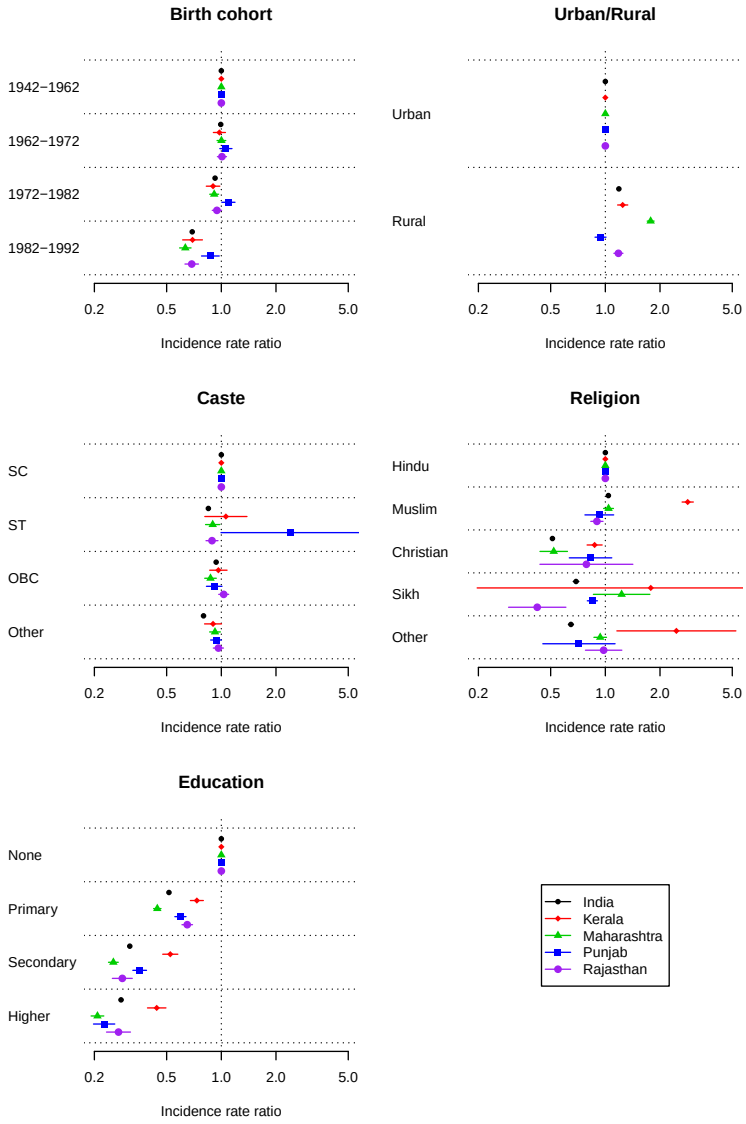

Figure 5: Forest plots of the estimated covariate effects on marriage incidence rates of women for India and the four selected states. The horizontal lines correspond to the rate ratio estimate, and 95% confidence interval.

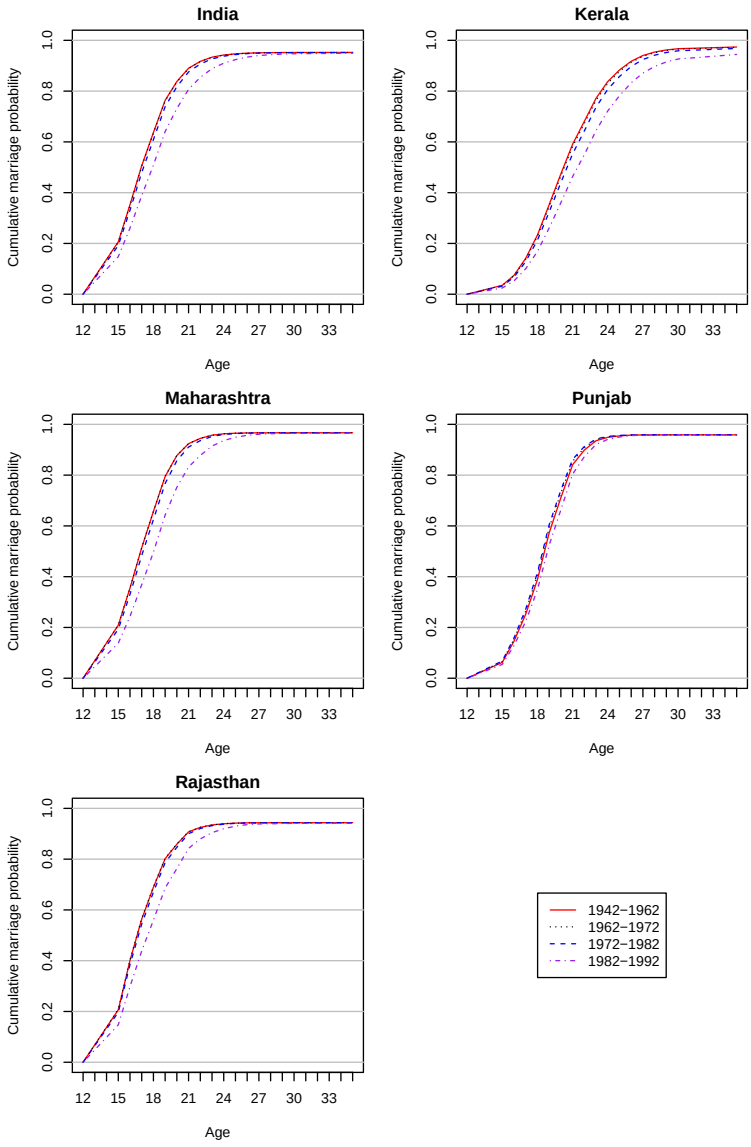

Figure 6: Predictive probabilities for women to be married by age a by birth cohort, calculated by combining 2010 mortality rates with the marriage incidence model. The other covariates were set to the reference levels.

7 Discussion

In this article we formulated a multi-state model for modeling an outcome and a covariate process jointly in two types of retrospective cross-sectional

cohort studies. Our methodological contributions can be summarised as follows. (i) Combined analysis of retrospective histories from two types of cross-sectional cohorts; (ii) multi-state modeling of retrospective histories of two correlated processes in two time scales; (iii) assessment of the performance of the method based on combining the two retrospective cross-sectional cohort designs against using either of the two; and (iv) illustration through an application to the estimation of marriage incidence rates. We have used explicitly the structure of the Indian education system in building the joint model and also extracting retrospective information from the cross-section. We also assume that everyone adheres to that. When retrospective history on schooling is available, in addition to the cross-section, this assumption can be relaxed.

Statistical methods have been developed and applied for the estimation of incidence rates from cross-sectional cohorts, with or without subsequent prospective follow-up (Keiding, 1991, 2006, Keiding et al., 2012, Saarela et al., 2009). The incidence rate, in general, is not identifiable from data under retrospective cross-section desing I only without supplementary information, e.g., data from design II. The estimation is simplified under assumptions such as time homogeneity and non-differential mortality before and after the incident event (Keiding, 1991). Much of the existing literature has focused on nonparametric estimation of cumulative incidence and survival functions through appropriately weighting the risk sets. Herein our main focus was in factors that modify the incidence rates, and therefore we applied likelihood-based methods for piecewise constant hazard models. For this purpose, we needed to combine likelihood functions arising from two different sampling plans, namely the cross-sectional cohort setting of NFHS-2, and the setting of NFHS-3. To combine information collected under the different sampling plans, the likelihood contributions from the individual surveys are conditioned on the specific sampling plan employed in the survey, with the overall likelihood expression obtained simply as the product of

these.

This result was applied in the estimation of the marriage incidence rates in four Indian states as functions of age and birth cohort, as well as demographic characteristics. Unlike previous approaches (Kashyap et al., 2015) to estimate marriage rates, the proposed method allows combining information from more than one survey and modelling education and marriage jointly. This brings several advantages. First, the increased sample size leads to more powerful analyses of age at marriage data at the sub-population level (e.g. Indian states). Second, it also allows learning of calendar time trends in the strength of association of many factors affecting marriage rates.

The analysis goes beyond simply describing the age- and sex-based marriage rates and puts forward a model which takes into account the well-recognised factors driving the marriages in India. The marriage incidence rates differ regionally (or state-wise) and hence the rates obtained using the India-level data may not bring out the real marriage squeeze problem existent in social strata defined by caste, religion and education. Although the caste effect on the marriage incidence rates did not differ much by state, those of education and religion did. Our analysis provides strong evidence towards religion, education and urban/rural area as the main factors affecting the marriage pattern among women in India. Education levels or qualifications seem to be replacing the earlier role of caste in shaping the marriage market in India. The effects of women's educational expansion on marriage incidence have been studied worldwide and found to have some impact. However, a considerable portion of the reduction in early marriage is not explained by changes in levels of education (Mensch et al., 2005). To predict the real magnitude of the marriage squeeze problem in India, predictions of married and unmarried populations in different age and social strata defined by state, caste, religion, urban/rural, and education are needed. The model proposed here will have a direct application for such predictions.

Appendices

Appendix A: Data selection and description

The NFHS reports clearly bring out differences between the states with respect to education (`http://rchiips.org/nfhs/`). All four states considered here show increasing trends in the proportion of women attaining higher education but differ by education attainment. There is a decreasing trend in the proportion of primary and no education, and increasing trend in the secondary and higher education level. Rajasthan stands out when looking the education levels of women, with the highest proportion of women with no education.

Punjab has suffered from an imbalanced child sex ratio, starting already in the 1980's (908 girls per 1000 boys in 1981) when the child sex ratios were still normal in most other states in India. Rajasthan has remained as a state with a relatively high total fertility unlike the other states examined (TFR 4.1 in 1998). Kerala has enjoyed replacement level fertility since the early 1990's. Maharashtra has come to suffer from imbalance in child sex ratio during the last two decades, combined with replacement level fertility since the 2000's.

Appendix B: Likelihood conditioning on the sampling pattern

To see that the likelihood obtained by multiplying (3.7) and (3.11) is still a conditional probability (less multiplicative terms), and thus a conditional likelihood, we partition the data collected under survey j as $(v_j, w_j) \equiv \{(v_{ij}, w_{ij}) : i \in C_j\},\ j = 2, 3$, where (v_{ij}) represents the conditioning event or sampling pattern. Further, (w_{ij}) denote the retrospective marriage histories recorded through the survey. Let $\Theta = (\theta, \beta(t), \mu(t, a))$ denote the parameters of interest θ as well as birth and mortality rates $(\beta(t), \mu(t, a))$. The parametrised joint distribution of all observed data $p(v_j, w_j, j = 2, 3 \mid \theta)$

Table 6: Observed proportions (in %) of women by state: categorical variables used are birth cohort, urban/rural, caste, religion, and education. (Source: NFHS-2 and -3 data)

	Kerala	Maharashtra	Punjab	Rajasthan	India
N	6450	14424	6477	10701	214638
Birth cohort					
1942-1962	22	15	19	18	16
1962-1972	33	28	30	30	28
1972-1982	28	33	30	36	33
1982-1992	17	24	21	16	23
Urban	33	66	36	28	40
Caste					
SC	10	15	30	18	17
ST	1	8	0.1	14	13
OBC	38	25	12	31	30
Other	51	52	58	38	40
Religion					
Hindu	55	75	41	89	75
Muslim	30	13	3	10	13
Christian	15	2	1	0.1	7
Sikh	0	0.3	55	0.5	2
Other	0.1	10	0.4	1	3
Education					
None	16	34	35	73	48
Primary	39	34	30	18	28
Secondary	31	20	26	6	16
Higher	14	12	10	4	8

may now be decomposed as

$$\begin{aligned} p(v_2, w_2, v_3, w_3 \mid \Theta) &= p(w_2, w_3 \mid v_2, v_3; \Theta) p(v_2, v_3 \mid \Theta) \\ &= \prod_{j=2}^{3} p(w_j \mid v_j; \theta) p(v_j \mid \Theta) \\ &= \prod_{j=2}^{3} \prod_{i \in C_j} p(w_{ij} \mid v_{ij}; \theta) p(v_{ij} \mid \Theta) \\ &\overset{\theta}{\propto} \prod_{j=2}^{3} L_j(\theta) \prod_{j=2}^{3} p(v_j \mid \Theta), \end{aligned}$$

where conditioning on the sampling plan (ignoring $\prod_{j=2}^{3} p(v_j \mid \Theta)$) may result in some loss of information on θ, but results in valid inferences.

Acknowledgments

The first two authors were partly supported by the project 'Precarious family formation' financed by the Kone foundation. The first author's work was also supported by the research mobility grant (No. 325990) awarded by the Academy of Finland. The authors would like to thank reviewers for their comments that helped to improve the manuscript.

References

[1] **Cherlin A.** (2010). Demographic Trends in the United States: A Review of Research in the 2000s. *J Marriage Fam.*, 72(3):403-419. doi:10.1111/j.1741-3737.2010.00710.x

[2] **Cook, R.J. and Lawless J.F.** (2018). *Multistate Models for the Analysis of Life History Data.* Monographs on Statistics and Applied Probability 158, CRC Press.

[3] **Desai S, Kulkarni V.** (2008). Changing educational inequalities in India in the context of affirmative action. *Demography*, 45(2):245-270. doi:10.1353/dem.0.0001

[4] **Dommaraju P.** (2009). Female schooling and marriage change in India. *Population-E*, 64(4):667-684.

[5] **Goswami B.** (2014). *Marriage patterns in India. Chapter IV Determinants of Marriage Change in India.* Thesis. http://shodhganga.inflibnet.ac.in:8080/jspui/handle/10603/49981

[6] **Government of India** (2011). Socio Economic and Caste Census 2011. Overview. `https://secc.gov.in/reportlistContent#`

[7] **Guilmoto, C.Z.** (2012). Skewed sex ratios at birth and future marriage squeeze in China and India, 2005-2100. *Demography*, 49:77–100.

[8] **Hayford, S.R. and Morgan, S.P.** (2008) The quality of retrospective data on cohabitation. *Demography*, 45(1):129-141.

[9] **Kalmijn M.** (1991). Shifting boundaries: Trends in religious and educational homogamy. *American Sociological Review.*, 56:786–800.

[10] **Kashyap, R., Esteve, A., and Garcia-Roman, J.** (2015). Potential (mis)match? marriage markets amidst sociodemographic change in India, 2005-2050. *Demography*, 52(1):183–208.

[11] **Keiding, N.** (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society, Series A*, 154:371–412.

[12] **Keiding, N.** (2006). Event history analysis and the cross-section. *Statistics in Medicine*, 25:2343–2364.

[13] **Keiding, N., Hansen, O. K. H., Sørensen, D. N., and Slama, R.** (2012). The current duration approach to estimating time to pregnancy. *Scandinavian Journal of Statistics*, 39:185–204.

[14] **Mensch, B. S., Singh, S., Casterline, J. B.** (2005). *Trends in the timing of first marriage among men and women in the developing world.* The Population Council, Inc.

[15] **Neelakantan, U., Tertilt, M.** (2008). A note on marriage market clearing. *Econ Lett.*, 101:103–105.

[16] **Ning, J., Hong, C., Li, L., Xuelin Huang, X. and Yu Shen, Y.** (2017). Estimating treatment effects in observational studies with

both prevalent and incident cohorts *The Canadian Journal of Statistics*, 45:202–219.

[17] **R Core Team** (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[18] **Raj A., Saggurti N., Balaiah D., Silverman J. G.** (2009). Prevalence of child marriage and its effect on fertility and fertility-control outcomes of young women in India: a cross-sectional, observational study. *Lancet*, 373(9678):1883-1889. doi:10.1016/S0140-6736(09)60246-4

[19] **Ruwali P. N.** (2018). The changing scenario of marriage in India : A sociological analysis. *Journal of Acharya Narendra Dev Research Institute*, https://andjournalin.com/2018/11/09/the-changing-scenario-of-marriage -in-india-a-sociological-analysis/

[20] **Saarela, O., Kulathinal, S., and Karvanen, J.** (2009). Joint analysis of prevalence and incidence data using conditional likelihood. *Biostatistics*, 10(3):575–587.

[21] **Schoen, R.** (1983). Measuring the tightness of a marriage squeeze. *Demography*, 20:61–78.

[22] **Wolfson, D.B., Best,A.F., Addona, V., Wolfson, J. and Gadalla, S.M.** (2019). Benefits of combining prevalent and incident cohorts: An application to myotonic dystrophy *Statistical Methods in Medical Research* 28:3333-3345.

Sangita Kulathinal
Department of Mathematics and Statistics, University of Helsinki, Finland
E-mail: sangita.kulathinal@helsinki.fi

Minna Säävälä
Population Research Institute, Väestöliitto, Helsinki, Finland
E-mail: Minna.Saavala@vaestoliitto.fi

Kari Auranen
Department of Mathematics and Statistics, University of Turku, Finland
E-mail: kajuaur@utu.fi

Olli Saarela
Dalla Lana School of Public Health, University of Toronto, Canada
E-mail: olli.saarela@utoronto.ca