

Estimation of incidence from aggregated current status data without differential mortality

Ralph Brinks

E-Mail: ralph.brinks@ddz.de

Chair for Medical Biometry and Epidemiology
Witten/Herdecke University
Faculty of Health/School of Medicine
D-58448 Witten, Germany

Institute for Biometry and Epidemiology
German Diabetes Center
D-40225 Düsseldorf, Germany

Department for Statistics
Ludwig-Maximilians-University Munich
D-80539 München, Germany

Abstract

We use a historical data about breathlessness in British coal miners to compare two methods centered around a differential equation for deriving the age-specific incidence from aggregated current status data with age information, i.e. age-specific prevalence data. Special focus is put on estimating confidence bounds. For this, we derive a maximum likelihood (ML) estimator for estimating the age-specific incidence from the prevalence data and confidence bounds are calculated based on classical ML theory. Second, we construct a Markov-Chain-Monte-Carlo (MCMC) algorithm to estimate confidence bounds, which implements a weighted version of the differential equation into the prior of the MCMC algorithm. The confidence bounds for both methods are compared and it turns out that the MCMC estimates approach the ML estimates if the prior gives strong weight to the differential equation.

Key words: prevalence, chronic diseases, illness-death model, MCMC estimation, maximum likelihood estimation

Introduction

The question about relations between incidence and prevalence of diseases dates back at least to 1934, when Muench examined if the age-specific incidence of yellow fever in Southern America can be reconstructed from a cross-sectional sample about the age-specific prevalence [Mue34]. Muench used the phrase *catalytic curve*, which still found in recent textbooks about infectious disease epidemiology [Vyn10]. In the statistical literature, the term *current status data* is more frequently used than catalytic models. By current status data, it is assumed to have a cross-sectional sample where for each study participant, current age and current disease status are known. A literature review about current status data is found in [McK11].

In epidemiology, current status data are often aggregated in different age groups. For example, the data in Muench's article about yellow fever in Colombia is given as positive cases from a number of tests applied in five age groups: 5-9, 10-14, 15-19, 20-39 and 40+ (Table 1 in [Mue34]).

In this technical note, we use a historical data about breathlessness in British coal miners [Ela14] to compare three methods centered around a differential equation for deriving the age-specific incidence from aggregated current status data. Using the differential equation for estimating incidence from prevalence data has been described in [Bri13] and has been compared to other methods in [Lan16]. In this work, special focus is put on estimating confidence bounds based on the differential equation. The data we use here are shown in Table 1.

Age group k (in years)	Number observed (n_k)	Number with condition (c_k)	Prevalence (p_k) (%)
20-24	1952	16	0.820
25-29	1791	32	1.787
30-34	2113	73	3.455
35-39	2783	169	6.073
40-44	2274	223	9.807
45-49	2393	357	14.92
50-54	2090	521	24.93
55-59	1750	558	31.89
60-64	1136	478	42.08

Table 1: Data about breathlessness in British coal miners [Ela14].

First, we derive a novel maximum likelihood (ML) estimator for inferring the age-specific incidence from the prevalence data in Table 1. Confidence bounds are calculated from classical ML theory. In a second approach, we construct an Markov-Chain-Monte Carlo (MCMC) algorithm to estimate confidence bounds for the incidence estimates. Finally, the confidence bounds for both methods are compared.

Estimation of the age-specific incidence

Recently, we have shown that the age-specific prevalence p of a chronic condition at some time t is related to the age-specific incidence density i (synonym: incidence rate), and mortality rates m_0 , and m_1 via a partial differential equation (PDE) [Bri14]. Figure 1 shows the underlying multi-state model with the possible transitions and associated rates.

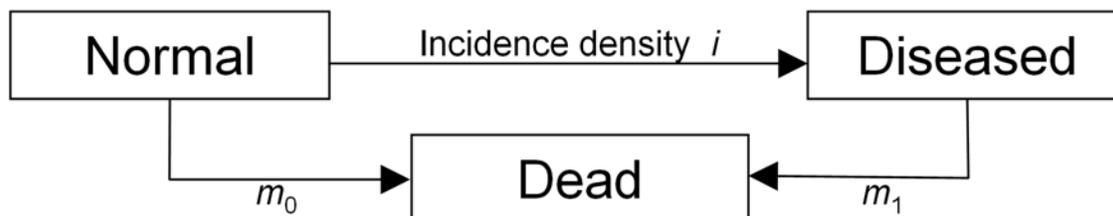


Figure 1: Illness-death model for a chronic condition ('Diseased') and associated transition densities: incidence density i , mortality without (m_0) and with the disease (m_1).

We consider the case without migration and without differential mortality (i.e., $m_0 = m_1$), where the PDE from [Bri14] reads as

$$(\partial p / \partial t + \partial p / \partial a) = (1 - p) i. \quad (1)$$

If we assume furthermore, that the incidence density i in Equation (1) does not depend on time t , the PDE (1) becomes an ordinary differential equation (ODE)

$$dp / da = (1 - p) i, \quad (2)$$

which has the general solution

$$p(a) = 1 - (1 - p_0) \times \exp\left(-\int_{a_0}^a i(\tau) d\tau\right) \quad (3)$$

where $p_0 = p(a_0)$ is the initial condition.

Equation (2) is the basis for a straightforward estimator of the age-specific incidence density i . If we can estimate the derivative dp from the age-specific prevalence p , we obtain $i = dp / (1 - p)$.

Example 1: We fit a linear model $\text{logit}(p(a)) = \log(p(a) / \{1 - p(a)\}) = \beta_0 + \beta_1 \times a$ at the midpoints of the age groups $a = 22.5, 27.5, \dots, 62.5$ to the data from Table 1. We obtain $\beta_0 = -7.02$ and $\beta_1 = 0.110$. Inserting $p(a) = \text{expit}(\beta_0 + \beta_1 \times a)$ into $i = dp / (1 - p)$ yields $i(a) = \beta_1 p(a)$. The expit-function is the inverse of the logit-function, i.e., $\text{expit} = \exp / (1 + \exp)$, and has the derivative $\text{expit}' = \text{expit} \times (1 - \text{expit})$. The associated age-specific incidence $i(a) = \beta_1 \text{expit}(\beta_0 + \beta_1 \times a)$ with $\beta_0 = -7.02$ and $\beta_1 = 0.110$ is plotted as black line in Figure 2. For comparison, the blue line shows the estimates [Ela14] has given. Note the negative estimate at age 57.5 from [Ela14].

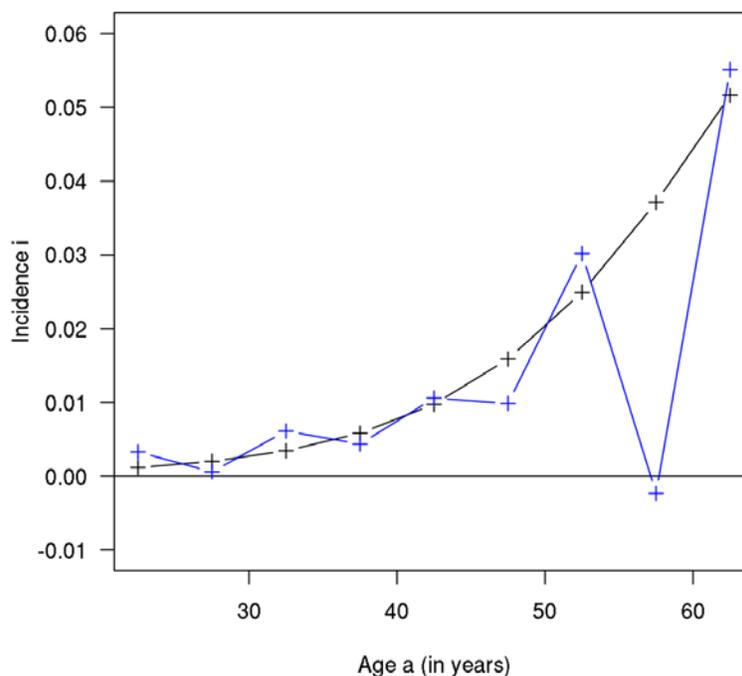


Figure 2: Estimated age-specific incidence for the data in Table 1 as calculated in Example 1.

Example 1 has demonstrated how the theory of differential equations can be used to estimate the age-specific incidence from aggregated current status data (prevalence data) of a chronic condition. In this work, we are interested in estimating confidence bounds for the age-specific incidence.

Maximum likelihood estimation

The binomial likelihood function L for the aggregated current status data, like the data in Table 1, is given by

$$L = \prod_k \binom{n_k}{c_k} p_k^{c_k} (1 - p_k)^{n_k - c_k} \quad (4)$$

where k is the index for the age group, n_k and c_k are the numbers of subjects and cases in an age group indexed by k , respectively. $p_k = c_k/n_k$ is the fraction of subjects in age group k with the condition under consideration (prevalence). In Table 1 we have nine age groups. Hence, for these data the product in Equation (4) consists of nine factors.

With a view to Figure 2, we have hints that the age-specific incidence density $i(a)$ grows exponentially with age. Hence, we make the approach $i(a) = \exp(\gamma_0 + \gamma_1 \times a)$ with coefficients γ_0, γ_1 . If we substitute this incidence into Equation (3) with the initial condition $a_0 = 20$ and $p(20) = p_0 = 0$, we obtain $p(a) = 1 - \exp(h(20) - (a))$ with $h(z) = \exp(\gamma_0 + \gamma_1 \times z)/\gamma_1$. Then, this p is evaluated at the age group midpoints $a_k = 22.5, \dots, 62.5$ and finally substituted into Equation (4). We end up at a likelihood function $L = L(\gamma_0, \gamma_1)$, which by optimization can be used to calculate a maximum likelihood estimator for both coefficients, γ_0 and γ_1 . If we additionally calculate the 95% confidence intervals using the inverse of the Fisher information matrix for large sample approximation of the variance-covariance matrix [Woo15], we obtain the results shown in Table 2.

	Point estimate	95% confidence bounds
γ_0	-7.8237	-8.0590 to -7.5883
γ_1	0.07559	0.07006 to 0.08111

Table 2: Maximum likelihood estimators for the coefficients γ_0 and γ_1 used for parameterization of the age-specific incidence.

MCMC approach

For data resulting from a PDE data generating process F , Xun and colleagues had the idea to separate the model fit $g = g(\theta)$ with unknown parameters θ to the observed data Y on the one hand and the model's accordance to the PDE F on the other [Xun13]. Hence, we have a functional

$$J(\theta) = f(Y - g(\theta)) + \lambda \times F(g(\theta)), \quad (5)$$

with a relaxation parameter $\lambda > 0$, which penalizes deviations from the data generating PDE model F . Xun et al. have chosen the function f for the data fit to be the square function $f(Y - g(\theta)) = \{Y - g(\theta)\}^2$. Other choices are possible, and we will use a weighted square approach below.

In our research question, the observed data Y is the age-specific prevalence p in Table 1. As in Example 1, we fit the p by an expit model: $p(a) = \text{expit}(\beta_0 + \beta_1 \times a)$. In addition, we fit the incidence i to an exponential model: $i(a) = \exp(\gamma_0 + \gamma_1 \times a)$. The function F describing the ODE (3) is defined via a vector of parameters to be estimated $\theta = (\beta_0, \beta_1, \gamma_0, \gamma_1)$:

$$F(\theta; a) = dp/da - i(1 - p).$$

Here, we have $F(\theta; a) = \{\beta_1 \text{expit}(\beta_0 + \beta_1 \times a) - \exp(\gamma_0 + \gamma_1 \times a)\} \{1 - \text{expit}(\beta_0 + \beta_1 \times a)\}$. F is squared and integrated over the whole age range, in case of the data in Table 1 from $a_0 = 20$ to $a_e = 65$ (years):

$$\mathfrak{J}(\theta) := \int_{a_0}^{a_e} [F(\theta, a)]^2 da \quad (6)$$

If we have a good agreement between the estimated parameters θ and the data generating differential equation (2), $\mathfrak{J}(\theta)$ as defined in Equation (5) is close to zero. We set up a MCMC algorithm to estimate θ and use $\mathfrak{J}(\theta)$ as a prior [Woo15]. To control the relative importance, we multiply with an penalization factor $\lambda > 0$. This leads us to following prior $\propto \exp\{-\lambda \mathfrak{J}(\theta)\}$. For the log-likelihood in the MCMC algorithm, we choose

$$l(\theta) := -\sum_k \left(\frac{p_k - \text{expit}(\beta_0 + \beta_1 a_k)}{s_k} \right)^2, \quad (7)$$

with $s_k = p_k(1 - p_k)/n_k$ which corresponds to a weighted least squares approach. Again, the a_k in Equation (7) are the midpoints of the age-groups in Table 1.

Given the prior and the log-likelihood, we are able to run a MCMC scheme based on the Metropolis-Hastings method [Woo15]. For each choice of λ in the range 1000 to 2,000,000, we use 500,000 iterations and discard the first 200,000 iterations as burn-in period.

The role of the relaxation parameter λ is shown in Figure 3, where the parameters γ_0 (left panel) and γ_1 (right panel) with 95% confidence bounds as estimated by the MCMC algorithm are plotted over the value λ . For comparison with the ML estimates and their confidence bounds from Table 2, the dashed blue lines and the blue areas represent the point estimates and confidence bounds for γ_0 and γ_1 , respectively.

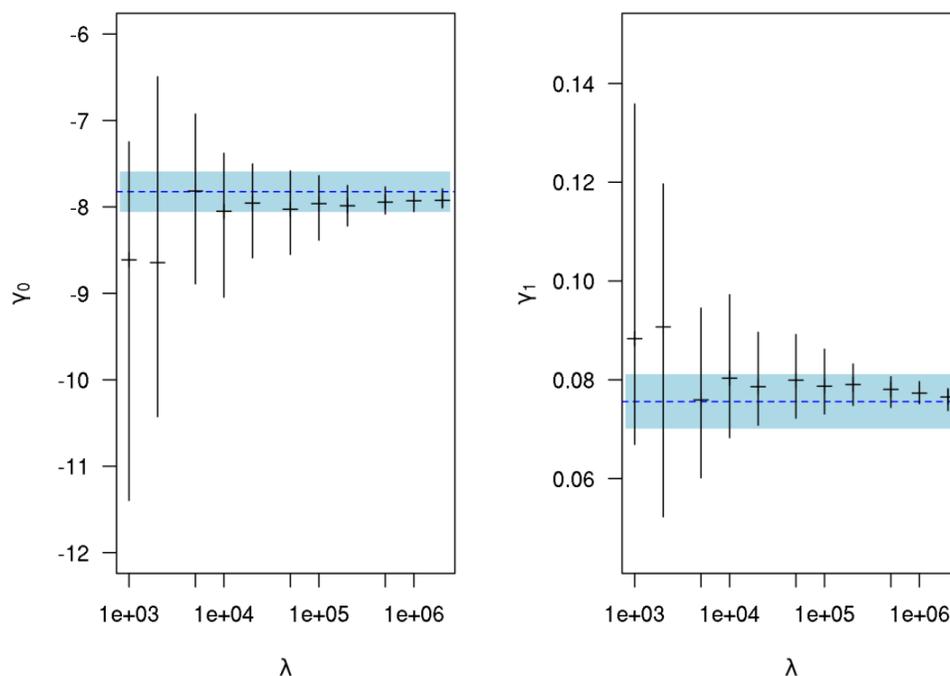


Figure 3: Incidence parameters γ_0 (left panel) and γ_1 (right) with 95% confidence bounds (vertical bars) as calculated by the MCMC algorithm with different weighting parameters λ (abscissa). The blue dashed lines and blue areas represent the ML estimates together with the 95% confidence bounds from Table 2.

Interestingly, as the parameter λ increases, the length of the confidence bounds from the MCMC decreases and finally lie inside the confidence bounds estimated from the ML estimator.

Discussion

In this work, we have described two methods about statistical inference around a differential equation, which relates the age-specific prevalence of a chronic condition with the underlying age-specific incidence.

The differential equation is not new, early approaches heading into a similar direction go back at least to the 1990ies [Bru99]. The innovative aspects in this work are twofold: 1) the link with standard maximum likelihood (ML) inference and 2) the implementation of the differential equation into the prior of an MCMC algorithm. While the maximum likelihood approach seems rather straightforward, the idea about MCMC goes back to rather recent work of Xun and colleagues [Xun13]. We could demonstrate that both methods are related, for

example by the observation that the MCMC confidence bounds tend to the ML confidence bounds if the relaxation parameter λ is chosen sufficiently large. So far this is only an observation from a computational experiment, which of course requires a deeper theoretical investigation about the relation of ML theory and MCMC simulations.

These two methods, ML and MCMC estimation, in combination with differential equations like Eq. (1) extend the methodological toolbox in epidemiology as requested about a decade ago [Chu10]. Until now, we had to employ re-sampling techniques to obtain confidence bounds in the context of differential equations. In re-sampling, a number of random samples from the reported distributions of the input parameters have been drawn to estimate how the uncertainty in the input parameters propagate through the differential equations into the outcomes. An example for the this type of re-sampling is described, for example, in [Bri15].

Appendices

Data availability

The source codes including the data sets for the analysis (and preparation of graphs) underlying this paper [Bri21a] and for running the MCMC algorithm [Bri21b] are available in the open public repository Zenodo.

Acknowledgements

The author gratefully acknowledges the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer SuperMUC at the Leibniz Supercomputing Centre (www.lrz.de).

Funding statement

The author did not receive any funding for any aspect of this work.

Competing interests

The author declares that no competing interests exists with any aspect of this work.

References

[Bri13] Brinks R, Landwehr S, Icks A, Koch M, Giani G. Deriving age-specific incidence from prevalence with an ordinary differential equation. *Stat Med.* 30;32(12):2070-8, 2013 doi: 10.1002/sim.5651.

[Bri14] Brinks R, Landwehr S: Age-and time-dependent model of the prevalence of non-communicable diseases and application to dementia in Germany. *Theoretical Population Biology*, 92:62-68, 2014

[Bri15] Brinks R, Hoyer A, Kuss O, Rathmann W: Projected Effect of Increased Active Travel in German Urban Regions on the Risk of Type 2 Diabetes. *PLOS ONE* 10(4): e0122145, 2015. <https://doi.org/10.1371/journal.pone.0122145>

[Bri21a] Brinks R: Statistical Analysis for Skinfaxi Project, Zenodo repository DOI 10.5281/zenodo.4816749 <https://doi.org/10.5281/zenodo.4816749>

[Bri21b] Brinks R: Estimation of age-specific incidence from prevalence: novel MCMC approach, Zenodo repository DOI 10.5281/zenodo.4770055 <https://doi.org/10.5281/zenodo.4770055>

[Bru99] Brunet RC, Struchiner CJ: A non-parametric method for the reconstruction of age- and time-dependent incidence from the prevalence data of irreversible diseases with differential mortality. *Theoretical Population Biology* 56(1): 76-90, 1999.

[Chu10] Chubb MC, Jacobsen KH: Mathematical modeling and the epidemiological research process. *European Journal of Epidemiology*, 25(1), 13-19, 2010

[Ela14] Elandt-Johnson RC, Johnson NL: *Survival models and data analysis*. John Wiley & Sons, 2014

[Lan16] Landwehr S, Brinks R. A comparative study of prevalence-based incidence estimation techniques with application to dementia data in Germany. *Stat Med* 28;35(5):768-81, 2016, DOI: 10.1002/sim.6736

[McK11] McKeown KM: *Topics in Current Status Data*, PhD Dissertation at the University of Berkley, 2011

[Mue34] Muench H: Derivation of Rates from Summation Data by the Catalytic Curve, *Journal of the American Statistical Association* 29 (185):25-38, 1934

[Vyn10] Vynnycky E, White RG: *An Introduction to Infectious Disease Modelling*, Oxford University Press, 2010

[Woo15] Wood SN, *Core Statistics*, Cambridge University Press, 2015

[Xun13] Xun X, Cao J, Mallick B, Maity A, Carroll RJ: Parameter estimation of partial differential equation models. *Journal of the American Statistical Association*, 108(503), 1009-1020, 2013