Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 6582–6595
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A data-augmentation method for infectious disease incidence data from close contact groups

Yang Yang[a,*], Ira M. Longini Jr.[a,b], M. Elizabeth Halloran[a,b]

[a] *Program of Biostatistics and Biomathematics, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA*
[b] *Department of Biostatistics, University of Washington, Seattle, WA 98195, USA*

Received 5 November 2006; received in revised form 7 March 2007; accepted 8 March 2007
Available online 12 March 2007

Abstract

A broad range of studies of preventive measures in infectious diseases gives rise to incidence data from close contact groups. Parameters of common interest in such studies include transmission probabilities and efficacies of preventive or therapeutic interventions. We estimate these parameters using discrete-time likelihood models. We augment the data with unobserved pairwise transmission outcomes and fit the model using the EM algorithm. A linear model derived from the likelihood based on the augmented data and fitted with the iteratively reweighted least squares method is also discussed. Using simulations, we demonstrate the comparable accuracy and lower sensitivity to initial estimates of the proposed methods with data augmentation relative to the likelihood model based solely on the observed data. Two randomized household-based trials of zanamivir, an influenza antiviral agent, are analyzed using the proposed methods.

Keywords: Antiviral agent; Data augmentation; EM algorithm; Infectious disease; Intervention efficacy; Linear model; MLE

1. Introduction

Close contact groups, such as households, are the important places of transmission for many infectious diseases. Data collected from these contact groups provide a basis for evaluating person-to-person transmission risks and effectiveness of intervention methods such as antiviral treatments or vaccine (Halloran et al., 1997; Becker et al., 2003). Using different levels of information available in the data, various statistical methods have been developed for data analysis. If only the final infection status of participants are known, methods utilizing recursive final-size probabilities can be applied, including likelihood maximization (Longini and Koopman, 1982; Addy et al., 1991), Bayesian approaches (O'Neill and Roberts, 1999), generalized linear models (Magder and Brookmeyer, 1993), and estimating equations with martingale techniques (Becker and Hasofer, 1997). In many modern clinical trials, sequential laboratory tests and symptom diary of participants provide time-to-event data with individual-specific longitudinal exposure information. To take into account exposure and transmission dynamics at the individual level, Rampey et al. (1992) constructed discrete-time likelihoods based on assumptions about the natural history of the disease such as the distributions of the latent and infectious

* Corresponding author. Tel.: +1 206 6676987.
E-mail address: yang@scharp.org (Y. Yang).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.03.007

periods. Yang et al. (2006) extended this method to the more realistic case-ascertained design. Cauchemez et al. (2004) proposed a Bayesian model with the flexibility of estimating the natural history of the disease, but time-dependent covariates have not been accommodated.

The discrete-time likelihoods in Rampey et al. (1992) and Yang et al. (2006) are built solely upon the observed data, including symptom onset dates, laboratory test results and household structure (which individuals live in which households), and involve summing probability components over the latent period. Summations or integrals are commonly seen in likelihoods based solely on the observed data, and such complicated structure may present difficulties for standard analyses or prevent extension by other methods (O'Neill et al., 2000). More importantly, when data are sparse because of rare incidences and/or a multicovariate structure, iterative estimation procedures (e.g., the Newton–Raphson algorithm) using only the observed data may be sensitive to the initial estimates in locating the maximum likelihood estimates (MLEs). This fact can be seen in Sections 3 and 4 of this paper, and is also mentioned in Yang et al. (2006). Data augmentation is a popular technique to circumvent computational difficulties in classical likelihood methods because likelihood functions conditional on unobserved variables are often simpler (van Dyk and Meng, 2001; Paap, 2002). In a transmission model for infectious diseases, a basic element is the transmission probability given a contact between an infective person and a susceptible person. The contact may be defined in various ways, for example, one day of living in the same household. The outcome of each contact, infection or escape, is generally not observable since a person may make multiple contacts before infection. In this paper, we revise the discrete-time likelihood in Yang et al. (2006) by augmenting the observed symptom onset data with the unobserved transmission outcome for each contact. This likelihood based on the augmented data has a simpler form than the one based on only the observed data and can be maximized with the EM algorithm. To illustrate the potential use of the simple likelihood by a different method, we derive a linear model that can be fitted using the iteratively reweighted least squares (IRLS) procedure. We show via simulation studies that both the maximum likelihood (ML) and the IRLS methods using the augmented data are less sensitive to initial estimates as compared to the ML method using only the observed data in Yang et al. (2006). We use the proposed approaches to estimate the prophylactic and treatment effectiveness of an influenza antiviral agent in two household trials.

2. Methods

Suppose that the disease under investigation is influenza and the data arise from a clinical trial in which household members are randomized to either an antiviral agent or control when an index case is identified by clinical symptoms. Let us assume that the antiviral agent provides temporary protection for susceptible contacts and therapy for cases. In the discrete-time likelihood model setting, risks are evaluated for each susceptible participant in each time interval. Suppose that the time intervals are consecutive days, and define a contact as the exposure of a susceptible person to an infective person in the same household throughout a day. The pairwise transmission probability per contact between a susceptible person i with covariates x_i and an infective person j with covariates x_j in the same household is expressed as $p(x_i, x_j)$. If x_i and x_j are scalars denoting treatment status of antiviral agent (1 = yes, 0 = no), then one can define efficacy measures $\mathrm{AVE_S}=1-p(1,0)/p(0,0)$, $\mathrm{AVE_I}=1-p(0,1)/p(0,0)$ and $\mathrm{AVE_T}=1-p(1,1)/p(0,0)$, where in the epidemiological literature $\mathrm{AVE_S}$ measures the antiviral efficacy in reducing susceptibility, $\mathrm{AVE_I}$ measures the efficacy in reducing infectiousness, and $\mathrm{AVE_T}$ is called the total effectiveness (Halloran et al., 1997). Let $p=p(0,0)$ be the baseline daily pairwise transmission probability without any treatment. For notational convenience, a reparameterization leads to $p(x_i, x_j) = \theta^{x_i(1-x_j)}\phi^{(1-x_i)x_j}\eta^{x_i x_j}p$ where $\theta = 1 - \mathrm{AVE_S}$, $\phi = 1 - \mathrm{AVE_I}$ and $\eta = 1 - \mathrm{AVE_T}$. For simplicity, we assume multiplicativity of θ and ϕ such that $\eta = \theta\phi$, and thus $p(x_i, x_j) = \theta^{x_i}\phi^{x_j}p$. In Yang et al. (2006), we explored the assumption of multiplicativity for the ML method using only the observed data.

As our interest centers around estimation of transmission probabilities and treatment efficacies, we assume that (1) the latent period (time from infection to being infectious) coincides with the incubation period (time from infection to the onset of symptoms); and (2) durations of the latent and the infectious periods have known probability distributions. If the latent and the incubation periods do not coincide but are both known, the model can be adjusted for such situation.

2.1. The maximum likelihood method based on the augmented data

Suppose that the trial is conducted on a population of size N and is observed on a daily basis from day 1 to day T. Let us assume that day 1 is the first calendar day of exposure for the whole study population. The observed data for