Spatial and Spatio-temporal Epidemiology 16 (2016) 35–42

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

Hierarchical spatial modelling of pneumonia prevalence when response outcome has misclassification error: Applications to household data from Malawi

Lawrence N. Kazembe [a,*], Mphatso S. Kamndaya [b]

[a] Department of Statistics and Population Studies, University of Namibia, Private Bag 13301 Windhoek, 340 Mandume Ndemufayo Avenue, Pionerspark, Namibia
[b] School of Public Health, Faculty of Health Sciences, University of Witwatersrand, Johannesburg, South Africa

ARTICLE INFO

Article history:
Received 23 April 2015
Revised 22 October 2015
Accepted 4 November 2015
Available online 1 December 2015

Keywords:
Misclassification error
Pneumonia
Hierarchical spatial modelling
Malawi

ABSTRACT

Pneumonia remains a major cause of child mortality in less developed countries. However, the accuracy of its prevalence and burden remains a challenge because disease data is often based on self-reports, resulting in measurement error in a form of under- and over-reporting. We propose hierarchical disease mapping approaches that permit measurement error, through different prior distributions of sensitivity and specificity. Proposed models were used to evaluate spatial variation of risk of pneumonia in children in Malawi. Results show that the true prevalence was 0.50 (95 CI: 0.4–0.66), however, estimates were dependent on sensitivity and specificity parameters. The estimated sensitivity was 0.76 (95% CI: 0.68–0.95), whereas specificity was 0.84 (95% CI: 0.72–0.93). A lower specificity underestimated the true prevalence, while sensitivity and specificity of greater or equal to 0.75 provided reliable and stable prevalence estimates. The spatial variation in disease risk changed little; however, misclassification of areas as high risk was visible.

1. Introduction

The burden of child mortality remains persistently high in many developing countries, mostly in sub-Saharan Africa (SSA). Diseases, including pneumonia, are the main causes. According to recent estimates, pneumonia contributed to childhood deaths in the magnitudes of 0.32 million, of which 1.1 million occur in the neo-natal/post-natal periods (Campbell et al., 2013; Walker et al., 2013). Pneumonia alone, constitutes 14% of all child deaths (Walker et al., 2013).

A single source of these estimates is health facility-reported cases. However, for facility data, particularly for the data arising from developing countries, the true disease burden is unknown because only a fraction seek treatment from health facilities. Because of limitations and variability in surveillance system, health care systems, and diagnostic capacity in SSA, there is a huge under-reporting or incompleteness of cases, and the true disease burden remains unknown. This may lead to misclassification of pneumonia cases with non-pneumonia cases, because of failure to capture cases with mild symptoms that do not interact with health care systems, thereby creating errors in estimating the geographical distribution, risk and impacts.

An alternative source for estimating pneumonia prevalence, hence its burden, is to use household survey data. In the household data, the disease prevalence is determined from parental-reported status. Similar misclassification problems may arise, as care-givers may misdiagnosed the disease, rendering, household survey of limited use. Misdiagnosis, may lead, first, to over-or under-prescription of

* Corresponding author. Tel.: +264 612064515; fax: +264 612063446.
E-mail address: lkazembe@yahoo.com (L.N. Kazembe).

http://dx.doi.org/10.1016/j.sste.2015.11.002
1877-5845/

treatment, e.g. antibiotics. Second, when true prevalence is unknown, it may distort the spatial distribution of disease burden, creating false hotspots or coldspots (Berke and Waller, 2010). The resultant spatial mismatch may create difficulties in programming of interventions for reducing disease burden. The estimation of true prevalence would allow for more focussed and therefore more cost-efficient control measures (Campbell et al., 2013).

The challenges of misclassification are well known in statistical modelling. Misclassification has important implications in parameter estimates and statistical inference, leading to large biases in parameter estimates, while the validity and power of tests are diminished (Carrol et al., 2006; Gustafson, 2004). Bias and efficiency loss in regression models with misclassification have been documented. More generally, the association between exposure and outcomes under misclassification have been investigated. Rosychuk and Thompson (2004) and Neuhaus (1999) and (2002) showed that ignoring errors on response cause biased estimates on covariate effects. Gustafson (2004) provides a treatise expose of misclassification and measurement error in covariates. Paulino et al. (2005) and McInturff et al. (2004) present a Bayesian approach for binomial responses with misclassification, while Carrol et al. (2006), in Chapter 9, give a general overview of Bayesian approaches.

This study extends Bayesian hierarchical models to the issue of response misclassification and study the effect of misclassification on the spatial distribution of disease burden. Our study only focussed on non-differential misclassification, although differential misclassification are implied in the results thus obtained. The proposed methods are illustrated using household data from Malawi, with the objective of mapping geographical distribution of pneumonia and identify clusters of high risk of childhood pneumonia. Specifically we examine whether, different geographical clusters will emerge under varying specificity and sensitivity of diagnosis.

2. Methods

2.1. Data

We used as a case study data collected as part of the 2006 Malawi Multiple Cluster Indicators Survey (MICS). MICS was designed to provide reliable estimates of indicators of women and children health at national and regional levels, as well as allow for regional and urban–rural comparisons. Further details of the survey design are provided in the report (MICS, 2006), also available online at http://www.childinfo.org/files/MICS3_FinalReport_2006_eng.pdf. In brief, a two-stage stratified sampling design was implemented to collect the data. A total of 1040 enumeration areas (EAs) as defined in the Malawi Population and Housing Census of 1998 were selected, stratified by urban/rural status with sampling probability proportional to the population of the EA. Each EA was geo-referenced. A fixed number of households were randomly selected in each EA. All women aged 15–49 were eligible for interview. Data were realized through an interviewer – administered questionnaire. Participation in the survey was voluntary and informed consent was obtained from all participants.

Table 1
Infection status by reported symptoms of care-givers in terms of SE (η), SP (θ), p, and π.

	True disease		
Reported symptoms	Present ($D=1$)	Absent ($D=0$)	Total
Present ($Y=1$)	$\pi \times \eta$	$(1-\pi)\times(1-\theta)$	p
Absent ($Y=0$)	$\pi \times (1-\eta)$	$(1-\pi)\times\theta$	$1-p$
Total	π	$1-\pi$	1

The outcome variable was binary (had pneumonia =1 or not=0). The outcome variable was derived from parental-reported sickness status of each child as reported by the caregivers (often mothers), and as experienced within 2 weeks prior to the survey date. These were based on the following questions: “Does the child have short-breath now/Did the child have had short breath during the last 2 weeks”.

We included limited covariates to study aspects of differential misclassification. We included mother’s education level (1=none, 2=primary, 3=secondary or higher), and birth order (1=first order, 2=second or third order, and 3=fourth or higher order). Further, according to our objective, we include spatial effects modelled at 364 sub-district areas, referred to as traditional authories (TA). The TAs were identified by using geo-coordinates of the EAs, which were recorded at the time of the survey. The minimum and maximum number of TAs per district were, respectively, 2 (in Phalombe) and 17 (in Blantyre city), with a mean of 7 TAs per district.

2.2. Statistical model

Our interest is to examine how well do reported symptoms and signs of pneumonia indicate the presence of true pneumonia. This can be assessed by describing the capacity of the caregiver to report the unknown true disease status. Suppose we define infection status D: $D=1$if infected, $D=0$ if not infected, with probability, π, and $Y=1$ if reported symptoms present; $Y=0$ if reported symptoms absent, with probability p. Then the observed prevalence (p) is related to true prevalence (π) through

$$p = \pi \cdot \eta + (1-\pi)(1-\theta), \tag{1}$$

where $\eta = P(Y=1|D=1)$is the sensitivity (SE), and $\theta = P(Y=0|D=0)$is the specificity (SP). The relationship between the true prevalence (π), observed prevalence (p), sensitivity (η) and specificity (θ) are summarized in Table 1.

Using algebraic manipulation we can obtain the true prevalence adjusted for misclassification as (Rogan and Gladen, 1978),

$$\pi = \frac{p+\theta-1}{\eta+\theta-1}. \tag{2}$$

The bias in estimating the true prevalence depends on variation in three unknown quantities: p, η, and θ. If one uses the Rogan–Gladen estimator then it is negative when $SE+SP<1$. Thus an approach of incorporating available information on p, SE, $SP$is required. From a Bayesian framework, this is possible by eliciting appropriate prior distributions for η, θ, and p.