Cancer Epidemiology 36 (2012) 153–160

Contents lists available at ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

Estimation of disease prevalence, true positive rate, and false positive rate of two screening tests when disease verification is applied on only screen-positives: A hierarchical model using multi-center data

Eileen M. Stock [a,b,c,*], James D. Stamey [b], Rengaswamy Sankaranarayanan [d], Dean M. Young [b], Richard Muwonge [d], Marc Arbyn [e]

[a] Department of Statistical Science, P.O. Box 97140, Baylor University, Waco, TX 76798-7140, USA
[b] Department of Statistical Science, Baylor University, Waco, TX 76798, USA
[c] Center for Applied Health Research, Scott & White Healthcare System, Temple, TX 76502, USA
[d] Screening Group, International Agency for Research on Cancer, Lyon, France
[e] Unit of Cancer Epidemiology, Scientific Institute of Public Health, Brussels, Belgium

ARTICLE INFO

Article history:
Received 30 April 2011
Received in revised form 24 June 2011
Accepted 6 July 2011
Available online 19 September 2011

Keywords:
Sensitivity and specificity
Hierarchical Bayesian model
Mixed logit model
Diagnostic or screening

ABSTRACT

Objectives: A model is proposed to estimate and compare cervical cancer screening test properties for third world populations when only subjects with a positive screen receive the gold standard test. Two fallible screening tests are compared, VIA and VILI. *Methods*: We extend the model of Berry et al. [1] to the multi-site case in order to pool information across sites and form better estimates for prevalences of cervical cancer, the true positive rates (TPRs), and false positive rates (FPRs). For 10 centers in five African countries and India involving more than 52,000 women, Bayesian methods were applied when gold standard results for subjects who screened negative on both tests were treated as missing. The Bayesian methods employed suitably correct for the missing screen negative subjects. The study included gold standard verification for all cases, making it possible to validate model-based estimation of accuracy using only outcomes of women with positive VIA or VILI result (ignoring verification of double negative screening test results) with the observed full data outcomes. *Results*: Across the sites, estimates for the sensitivity of VIA ranged from 0.792 to 0.917 while for VILI sensitivities ranged from 0.929 to 0.977. False positive estimates ranged from 0.056 to 0.256 for VIA and 0.085 to 0.269 for VILI. The pooled estimates for the TPR of VIA and VILI are 0.871 and 0.968, respectively, compared to the full data values of 0.816 and 0.918. Similarly, the pooled estimates for the FPR of VIA and VILI are 0.134 and 0.146, respectively, compared to the full data values of 0.144 and 0.146. Globally, we found VILI had a statistically significant higher sensitivity but no statistical difference for the false positive rates could be determined. *Conclusion*: Hierarchical Bayesian methods provide a straight forward approach to estimate screening test properties, prevalences, and to perform comparisons for screening studies where screen negative subjects do not receive the gold standard test. The hierarchical model with random effects used to analyze the sites simultaneously resulted in improved estimates compared to the single-site analyses with improved TPR estimates and nearly identical FPR estimates to the full data outcomes. Furthermore, higher TPRs but similar FPRs were observed for VILI compared to VIA.

1. Introduction

Cervical cancer is the third most common cancer among women worldwide and is the second most common cancer among women in developing countries [2,3]. Although cervical cancer is preventable through effective screening programs, there were over 530,000 new cases diagnosed and 270,000 deaths from cervical cancer reported worldwide in 2008. Approximately 86% of these deaths occurred in developing countries, which lack the resources and infrastructure necessary to run organized screening programs [4,5].

The Pap smear's introduction, more than 40 years ago, has remarkably transformed cervical cancer from a leading killer to a rare disease in the United States and other European countries with organized screening programs by decreasing the rate of cancer by 75 percent. However, only five percent of women in low-income

* Corresponding author at: Department of Statistical Science, P.O. Box 97140, Baylor University, Waco, TX 76798-7140, USA. Tel.: +1 254 710 1699; fax: +1 254 710 4477.
E-mail addresses: Eileen_Stock@baylor.edu, estock@swmail.sw.org (E.M. Stock).

1877-7821/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.canep.2011.07.001

countries have undergone a Pap smear in the past five years [6]. Also, older and poorer women, which have a higher risk of developing cervical cancer, are least likely to be screened. In many developing countries, the majority of women have never undergone a pelvic exam [7]. Therefore new, cost-effective alternative screening methods are currently being investigated that may perform equally or possibly better than the Pap smear in developing countries [4].

Berry et al. [1] considered a single population model stratified into two groups with varying prevalences who have undergone two screening tests but with verified disease statuses for those individuals testing positive for at least one of the tests. They further allowed correlation between the screening tests and applied these methods to the detection of colorectal cancers. In this paper, we extend the model of Berry et al. [1] by presenting a multi-center hierarchical Bayesian logit model for estimating prevalence and test performance measures from a screening program that uses two imperfect diagnostic tests and a gold standard applied to all subjects testing positive for at least one of the screening tests. We further validate this model using data from a study with gold-standard measurements for all subjects. We also perform a comparison of the test parameters for the two screening procedures.

Chen et al. [8] similarly consider a Bayesian approach to breast cancer screening by estimating and obtaining credible intervals for the sensitivities, transition probabilities of the disease, and sojourn time of women in two age groups, 40–49 years and 50–59 years. Estimation of the TPR and FPR for a diagnostic or screening test requires individuals to be identified as diseased or non-diseased using a gold standard. However, often in the investigation of new screening tests, neither test is considered a gold standard. We compare the accuracy of two low-cost procedures for screening of high-grade cervical intra-epithelial neoplasia (CIN2+) using a cross-sectional multi-center study conducted by the International Agency for Research in Cancer. Similarly to Berry et al. [1], we relax the assumption of independence by allowing correlation between the diagnostic tests. Each center is stratified into two age groups for modeling the prevalence of CIN2+. The age groups are arbitrarily chosen to be (1) women < 45 years old and (2) women ≥ 45 years old. We assume the stratified groups have varying disease prevalences with unverified negatives. However, verification for these individuals has been performed so that we may compare the parameter estimates resulting from our model using the reduced data to those values obtained from the full data.

Recent developments in simulation methods and computational speed have improved the estimation of elaborate models. By specifying a model with a hierarchical structure, we reduce a complex model into a set of simpler submodels with more flexibility and features at each level resembling that of the original data. Furthermore, we obtain stable estimates through Bayesian methods. Martinez et al. [9] use a Bayesian method to estimate disease prevalence, sensitivity, and specificity of three cervical cancer screening tests in the presence of a covariate and in the absence of a gold standard. They further note that the method is advantageous in that the number of parameters to be estimated is not limited by the number of observations, as commonly encountered with frequentist approaches. We estimate the multiple parameters of interest through simulation methods based on MCMC using the software R, the WinBUGS program, and the R library R2WinBUGS.

2. Materials and methods

A research project conducted from 1999 to 2003 by the International Agency for Research in Cancer (IARC, Lyon) used a cross-sectional multi-center study involving 58,679 women, ages 25–65 years old, set up at 11 locations in five African countries and India [5]:

1. Mumbai
2. Trivandrum 1
3. Trivandrum 2
4. Calcutta 1
5. Calcutta 2
6. Bamako
7. Brazzaville
8. Conakry
9. Jaipur
10. Niamey
11. Ouagadougou

Tests involving visual inspection with 3–5% acetic acid (VIA) occured at all 11 centers and with Lugol's iodine (VILI) at 10 of the 11 centers. All screened women were verified with a gold standard, a colposcopic exploration of the cervix followed by taking biopsies from the areas that are colposcopically suspicious performed at the same time of their visit. The prevalence of cervical intra-epithelial neoplasia (CIN) was then assessed, according to four levels of disease outcome: (1) CIN I or worse (CIN1+), (2) CIN2+, (3) CIN3+, and (4) invasive cervical cancer. Different health workers, blinded to the results of the other tests, performed all tests. Since the gold standard is applied to all study subjects, the accuracy for CIN or cancer can be evaluated without verification bias. Often the gold standard is not available or is applied to only those subjects who test positive for one or more tests for time and cost purposes. We assume that the probability of selecting an individual for a clinical assessment depends on only their screening test results. Furthermore, if this individual with test results does not have a clinical assessment, then we can treat their true disease status as missing and satisfying the missing at random (MAR) assumption [10].

We will elaborate and further extend the results of this study by comparing the accuracy of the VIA and VILI screening tests for CIN2+ in a population stratified into two groups, (1) women < 45 years old and (2) women ≥ 45 years old. We propose a hierarchical Bayesian mixed logit model to examine test properties along with prevalence of CIN2+ in each of the groups. We will consider only 10 of the 11 centers since VILI testing was not performed at the Calcutta 1 center. Excluding this center, along with missing data, reduced the number of women screened to 52,779 women. Descriptive statistics for the centers, including the number of women screened, the number of women testing positive with the gold standard, and the number who tested positive for each test by age group are displayed in Table C.1. The random effects are modeled through Bayesian prior specifications reflecting heterogeneity among the centers.

Let $z_{gijk,m}$ represent the number of subjects in the g^{th} group with disease status i, Test 1 result j, and Test 2 result k for the stratified population $\Pi_{g,m}$, where

$$
\begin{aligned}
&g = 1 \ (\text{women} \geq 45),\ 2 \ (\text{women} < 45),\\
&i = 0 \ (\text{non-diseased}, \leq \text{CIN1}),\ 1\,(\text{diseased},\ \text{CIN2}+),\\
&j = 0 \ (\text{negative for VIA}),\ 1\,(\text{positive for VIA}),\\
&k = 0 \ (\text{negative for VILI}),\ 1\,(\text{positive for VILI}).
\end{aligned}
$$

at location m, m = 1 (Mumbai), ..., 10 (Ouagadougou). Considering a location m, Table C.2 shows the sampling plan for the population's stratified groups, $\Pi_{g,m}$, g = 1, 2. In Table C.2, $z_{g1\cdot1,m}$ denotes the row sum of diseased individuals who tested positive for Test 2 (VILI) in group g at location m. Similarly, $z_{g11\cdot,m}$ denotes the column sum of diseased individuals who tested positive for