Journal of Statistical Planning and Inference 167 (2015) 41–57

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Array-based schemes for group screening with test errors which incorporate a concentration effect

Y.G. Habtesllassie [a], Linda M. Haines [b,*], H.G. Mwambi [a], J.W. Odhiambo [c]

[a] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa
[b] Department of Statistical Sciences, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa
[c] Office of the Vice-Chancellor, Strathmore University, Nairobi, Kenya

ARTICLE INFO

Article history:
Received 30 August 2014
Received in revised form 26 May 2015
Accepted 27 May 2015
Available online 4 June 2015

Keywords:
Array
Group testing
Factors
False negatives
False positives
Sensitivity
Specificity

ABSTRACT

Group screening is widely used as an efficient method for identifying samples or factors from a large population that are in some sense active. The focus in the present paper is on screening blood samples for infectious diseases when errors in testing are present. Specific attention is given to the introduction of a concentration effect, that is to settings in which the error in testing a group of blood samples depends on the number of samples in that group which are infected. Four array-based group screening schemes, the Dorfman, the AND, the OR and a modification of the AND scheme, are considered and their performance appraised by deriving explicit formulae for the expected number of tests, the expected number of false negatives and the expected number of false positives. The results are illustrated by means of two examples. As an aside, relationships complementary to those derived in the context of blood screening are developed within the area of group factor screening.

1. Introduction

In 1943 Dorfman introduced an efficient procedure for the screening of a large number of blood samples for a rare disease. Specifically, the scheme involves testing pooled blood samples for the disease and then, in a second stage, testing individual samples from those pools which tested positive for the disease in the first stage. In a complementary paper in 1961, Watson examined the related problem of group factor screening in cases of effect sparsity and proposed an approach based on that of Dorfman (1943) which incorporates appropriately designed two-stage experiments and which, additionally, accommodates the errors necessarily incurred in the test procedures. Since the publication of these two seminal papers, there has been considerable interest in issues relating to the broad area of group screening and to blood screening and group factor screening in particular (Dean and Lewis, 2006). There are, in turn, two primary areas of interest within the field of blood screening, case identification and prevalence estimation, and both of these have been extensively researched. In the context of case identification attention has focused, *inter alia*, on developing multi-stage and sequential schemes that are in some sense more effective than the Dorfman procedure (Kim et al., 2007) and on relaxing the somewhat stringent assumption that the probability that a blood sample is infected is constant (Bilder et al., 2010; McMahan et al., 2012).

There are surprisingly few reported studies on the effect of errors in testing in two-stage, and more generally in multi-stage, group screening procedures for the identification of samples which are in some way defective, such as infected blood

* Corresponding author.
E-mail address: linda.haines@uct.ac.za (L.M. Haines).

http://dx.doi.org/10.1016/j.jspi.2015.05.009
0378-3758/

samples. A study based on ideas from the area of group factor screening is presented in the M.Sc. thesis of Habtesllassie (2004) and some related early work is noted in the review by Hughes-Oliver (2006). In a more recent paper, Kim et al. (2007) presented an in-depth investigation into the impact of test errors on hierarchical and two-dimensional square array schemes for case identification, in particular by evaluating the expected number of tests and selected error-based operating characteristics when the specificity and selectivity for group and individual tests are constant. Kim and Hudgens (2009) have extended this study to accommodate three-dimensional arrays. In addition, Hedt and Pagano (2008a,b) have devised an algorithm which accommodates errors in testing for square array pooling with halving at the retesting stage and have reported their findings in two working papers. In contrast, there has been a number of studies on the impact of errors in testing on the estimation of disease prevalence (Hughes-Oliver, 2006; Liu et al., 2012; Zhang et al., 2014). Interest has centered, *inter alia*, on the dilution effect, that is on settings in which false negatives which arise when a group comprising predominantly uninfected samples together with a very few infected samples tests negative. This effect was first discussed in the literature in the paper of Hwang (1976) and has been explored in more recent studies by, for example, Wein and Zenios (1996) and Hung and Swallow (1999). As noted earlier, errors in testing are necessarily embedded within the notion of group factor screening. Studies in this area are reviewed in Morris (2006) and have been restricted primarily to the Dorfman scheme and to extensions of that scheme to multi-stage and stepwise screening.

The aim of the present study is to extend the work reported in the literature on blood screening for the identification of infected samples in the presence of test errors by borrowing strength from the area of group factor screening and by taking cognizance of studies on prevalence estimation. More specifically, following Burns and Mauro (1987) and Habtesllassie (2004), the aim is to investigate blood screening schemes in which there is a concentration effect, that is one in which the errors in testing groups of blood samples depend on the number of infected individuals within the group. The paper is structured as follows. In Section 2 issues relating to the screening of blood samples are discussed. In particular four two-stage group screening schemes of interest are introduced and the expected number of tests, the expected number of false negatives and the expected number of false positives in the presence of test errors are evaluated explicitly for each scheme. The results are illustrated by means of examples in Section 3. In Section 4 the same issues as those discussed for blood screening are briefly revisited within the context of group factor screening and the differences which emerge, particularly in relation to the number of false negatives, are highlighted. Finally some broad conclusions and pointers for future research are given in Section 5. Note that, for ease of exposition, the setting in which blood samples are tested for an infectious disease is adopted here. However the discussion holds for a large number of group screening applications for case identification, including tests for defective items and in drug discovery. Note also that the terms concentration effect in the context of case identification and dilution effect in the context of prevalence estimation can be construed as being equivalent but that the emphasis in their interpretation is somewhat different. Thus, following Burns and Mauro (1987), the term concentration effect is used in the present study.

2. Blood screening

2.1. Preliminaries

Suppose that blood samples are arranged at random in a 2-dimensional array of cells. Four two-stage screening schemes based on this array are considered to be of interest here and are introduced as follows. The Dorfman scheme (Dorfman, 1943) involves the pooling and testing of samples in each row of the array and then the testing of individual samples in rows that test positive for the disease. The AND and the OR schemes, which were introduced in the paper by Phatarfod and Sudbury (1994) and the technical report by Langfeldt et al. (1997) respectively, are the same in the first stage and involve the pooling and testing of samples in each row and, independently, the pooling and testing of samples in each column. In the second stage of the AND scheme only individual samples lying at the intersection of rows and columns which tested positive in the first stage are tested. In contrast, the second stage of the OR scheme involves testing all samples in the rows and columns which tested positive in the first stage. Finally the scheme "square array without master pool testing" devised by Kim et al. (2007) is introduced and extended to a rectangular array. In this latter scheme all rows and columns are tested independently in the first stage. In the second stage, if at least one row and one column test positive, all cells at the intersection of positive rows and positive columns are tested; otherwise, if only rows test positive then all cells in the positive rows are tested and similarly for columns. For all four schemes, if no rows or columns test positive in the first stage then the procedure stops.

The probability that an individual blood sample is infected, that is the prevalence of the disease, is taken to be a constant p, independent of the status of all other samples. In addition, this probability is assumed to remain unchanged on dilution, that is on pooling. Most importantly here, following notions for factor screening developed in the seminal paper of Watson (1961) and for group screening for defective items by Burns and Mauro (1987), errors in testing in stage one of each of the screening schemes of interest are taken to depend on the number of infected cells in the group or pool. A concentration effect is therefore introduced into the setting for blood screening. The errors in testing in the first stage can be quantified by introducing the generic probability

$$\pi_1^\star(s, k) = P(\text{a group of } k \text{ cells tests positive given that } s \text{ out of the } k \text{ cells are infected})$$

where $s = 0, \ldots, k$. Thus the sensitivity of the test is taken to depend on the number of infected cells in the group and is given by $\pi_1^\star(s, k)$ for $s = 1, 2, \ldots, k$ and the specificity of the test is given by $1 - \pi_1^\star(0, k)$. The errors in testing an individual