Computational Statistics and Data Analysis 56 (2012) 1200–1220

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

# Confidence interval construction for disease prevalence based on partial validation series

Man-Lai Tang [a], Shi-Fang Qiu [b,*], Wai-Yin Poon [c]

[a] Department of Mathematics, Hong Kong Baptist University, Hong Kong

[b] Department of Statistics, Chongqing University of Technology, Chongqing 400054, PR China

[c] Department of Statistics, The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

*Article history:*
Available online 12 February 2011

*Keywords:*
Disease prevalence
Confidence interval
Validation series
Double sampling
Approximate unconditional method
Bootstrap resampling method

ABSTRACT

It is desirable to estimate disease prevalence based on data collected by a gold standard test, but such a test is often limited due to cost and ethical considerations. Data with partial validation series thus become an alternative. The construction of confidence intervals for disease prevalence with such data is considered. A total of 12 methods, which are based on two Wald-type test statistics, score test statistic, and likelihood ratio test statistic, are developed. Both asymptotic and approximate unconditional confidence intervals are constructed. Two methods are employed to construct the unconditional confidence intervals: one involves inverting two one-sided tests and the other involves inverting one two-sided test. Moreover, the bootstrapping method is used. Two real data sets are used to illustrate the proposed methods. Empirical results suggest that the 12 methods largely produce satisfactory results, and the confidence intervals derived from the score test statistic and the Wald test statistic with nuisance parameters appropriately evaluated generally outperform the others in terms of coverage. If the interval location or the non-coverage at the two ends of the interval is also of concern, then the aforementioned interval based on the Wald test becomes the best choice.



## 1. Introduction

Disease prevalence in a population is the proportion of individuals in that population who have the disease at a given time. The estimation of disease prevalence is an important topic as it is crucial in assessing the impact of a disease or in the planning of health-care delivery. In collecting data for the estimation of disease prevalence, screening tests that are relatively inexpensive, provide results in a timely manner, and non-intrusive to test subjects are frequently used. However, such tests are more likely to misclassify subjects. In other words, the sensitivity and specificity of the tests may not be sufficiently high, where the sensitivity of a test is the probability that a diseased person is correctly identified, and the specificity is the probability that a non-diseased person is correctly specified. The use of a data set with high-level of misclassification to estimate disease prevalence can lead to biased results, as shown by Bross (1954). On the other hand, gold standard tests are those that are completely accurate and will not misclassify subjects, but are usually very costly, time-consuming to organize, or invasive when applied to subjects. As a compromise between the two, many research studies use data sets with partial validation series. Specifically, the entire data set consists of $N$ subjects that are drawn randomly and independently from the target population, out of which $N - n$ of them are classified by the fallible screening test only and $n$ of them are classified by both the fallible test and the gold standard test. The $n$ data points in the sample, which provide not only information on

* Corresponding author.
*E-mail address:* sfqiu@amss.ac.cn (S.-F. Qiu).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.02.010

**Table 1**
The hospital acquired infection data.

| | Validated series | Assessment by hospital staff | | Total |
|---|---|---|---|---|
| | | Positive | Negative | |
| Assessment by validator | Positive | 334 | 55 | 389 |
| | Negative | 78 | 4249 | 4327 |
| | Total | 412 | 4304 | 4716 |
| | | Assessment by hospital staff | | |
| | Unvalidated series | 4598 | 53105 | 57703 |
| Grand total | | 5010 | 57409 | 62419 |

**Table 2**
The aplastic anaemia patient data set.

| | Validation series | Surrogate variable (Acute GVHD) | | Total |
|---|---|---|---|---|
| | | Yes | No | |
| True variable (Chronic GVHD) | Yes | 6 | 3 | 9 |
| | No | 1 | 8 | 9 |
| | Total | 7 | 11 | 18 |
| | | Classified by surrogate variable only | | |
| | Unvalidated series | 25 | 44 | 69 |
| Grand total | | 32 | 55 | 87 |

the disease prevalence but also the accuracy of the fallible test, can be regarded as a validation series. It is worthy noting that this and other two-phase methods of data collection are sometimes referred to as the method of double sampling (see, e.g., Tenenbein, 1970, 1972). We will address this point again in the final section.

A study reported by Smyth et al. (2008) provided a typical example of this type of data. The data set was kindly recommended by Professor R.G. Newcombe. The study was on the prevalence of hospital acquired infection. A total of 62419 subjects in England and Northern Ireland were assessed by nursing staff, out of which 5010 subjects were classified as affected and 57409 were classified as unaffected. As the use of thousands of staff across hundreds of hospitals might not result in a high degree of consistency, members of the study team subsequently visited all of the hospitals involved and reviewed the records for 4716 of the patients to classify them once again as affected or unaffected. In other words, a sample of 4716 was classified by both hospital staff and the study team, and the classification results of the set of 4716 patients constituted the validation series. The classification results are shown in Table 1.

This data collection method and the data structure are frequently encountered in research designs in which surrogate variables are involved. In many medical research studies in which information on the variables of interest is difficult to obtain due to time, resource and ethical considerations, researchers may choose to use surrogate variables to mimic the variables of interest. The information provided by surrogate variables can be elicited less expensively but is less accurate, and clinical studies usually make use of all available data, including those provided by the original variables of interest and the surrogate variables. This leads to data sets with structures the same as that in Table 1. For instance, Pepe (1992) reported a study that investigated whether aplastic anaemia patients who had been given bone marrow transplants would develop graft-versus-host disease (GVHD). The data set in Table 2 was extracted from Pepe (1992) but included only those patients whose ages at the time of transplant were 20 or above. The variable of primary interest is "Chronic GVHD". However, as collecting information on "Chronic GVHD" required long-term follow-up of the patients, the variable "Acute GVHD" that could be measured instantly after the transplant was used as a surrogate. The validation set consisted of those patients who could be successfully followed-up and for whom information on "Chronic GVHD" could be ascertained.

It was suggested by Professor R.G. Newcombe that clear distinction should be drawn between three scenarios for studying disease prevalence involving a gold standard test that is completely accurate and another fallible screening test that may misclassify subjects. These are as follows. (a) *Screening a population for disease*, in which all individuals satisfying wide eligibility criteria are invited to be tested by the fallible device. For example, all women in a target population are invited to attend mammographic screening for breast cancer, and only those mammographically positive women undergo further testing using more definitive procedures. (b) *Differential diagnosis*, which is different from screening in that it applies to individuals who are identified as symptomatic or affected. Again, a sequence of tests may be used. It may be more economical to start with a fallible screening test, and only small minority of the case series will need to have the more definitive gold standard test because it may be hazardous or very expensive. One example is the differential diagnosis of porphyrias (Whatley et al., 2009). (c) *Determining prevalence using data with a set of validation series*. The application is to purely observational studies, in which there is no implication that the classification of an individual as positive by either test will alter that person's clinical management in any way. This is specifically the scenario envisaged in this paper. The data in Table 1 for the study of hospital acquired infection is a typical example.