ELSEVIER

Correcting for Partial Verification Bias: A Comparison of Methods

JORIS A.H. DE GROOT, MSC, KRISTEL J.M. JANSSEN, PHD, AEILKO H. ZWINDERMAN, PHD, PATRICK M.M. BOSSUYT, PHD, JOHANNES B. REITSMA, PHD, AND KAREL G.M. MOONS, PHD

PURPOSE: A common problem in diagnostic research is that the reference standard has not been carried out in all patients. This partial verification may lead to biased accuracy measures of the test under study. The authors studied the performance of multiple imputation and the conventional correction method proposed by Begg and Greenes under a range of different situations of partial verification.
METHODS: In a series of simulations, using a previously published deep venous thrombosis data set ($n = 1292$), the authors set the outcome of the reference standard to missing based on various underlying mechanisms and by varying the total number of missing values. They then compared the performance of the different correction methods.
RESULTS: The results of the study show that when the mechanism of missing reference data is known, accuracy measures can easily be correctly adjusted using either the Begg and Greenes method, or multiple imputation. In situations where the mechanism of missing reference data is complex or unknown, we recommend using multiple imputation methods to correct.
CONCLUSIONS: These methods can easily apply for both continuous and categorical variables, are readily available in statistical software and give reliable estimates of the missing reference data.
Ann Epidemiol 2011;21:139–148.

KEY WORDS: Partial, Verification, Bias, Methods, Epidemiologic, Imputation.

INTRODUCTION

In studies of diagnostic accuracy, results from one or more tests under evaluation are compared with the results obtained with the reference standard. These studies are a vital step in the evaluation of new and existing diagnostic technologies. The reference standard is the best available method for identifying patients as having the disease of interest. Measures, such as sensitivity, specificity and predictive values, express how well tests under evaluation are able to identify patients as having the target disease (1).

A common problem in diagnostic research is that the reference standard has not been carried out in all patients because of ethical, practical or other reasons. Partial verification, if not accounted for, is known to lead to biased accuracy estimates, described in the literature as partial verification bias or work-up bias (2).

In clinical practice, different mechanisms can lead to partial verification (3). Sometimes it is simply unavoidable. For example, to verify results of positron emission tomography (PET) in staging esophageal cancer (4), only results of patients with PET lesions suggestive of distant metastases can be verified by histology. Histology cannot be carried out in PET negative patients. Second, incomplete verification can be prespecified in the design, for example, for efficiency reasons. This is often the case in screening test evaluation studies, where disease prevalence is low (5). In these types of studies, researchers often decide to apply the reference standard in only a random sample of the large group of patients with a negative screening test result. In other studies, partial verification is not planned, and reasons are unclear and not documented. For example, the accuracy of dobutamine atropine stress echocardiography for detecting coronary artery disease can be assessed using coronary angiography as the reference standard. In one study (6), only a small sample of the patients received this reference standard because of the practitioners' decision to refer patients to angiography or not, depending on history and other test results.

One of the methods to correct for partial verification was developed by Begg and Greenes (B&G) (7). In short, this method uses observed proportions of diseased and nondiseased among the verified patients to calculate the expected number of diseased and nondiseased among nonverified patients. The two are combined to obtain a complete

From the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands (J.A.H.d.G., K.J.M.J., K.G.M.M.); and Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center Amsterdam, Amsterdam, The Netherlands (A.H.Z., P.M.M.B., J.B.R.).
Address correspondence to: Joris A.H. de Groot, MSc, Julius Center for Health Sciences and Primary Care, UMC Utrecht, P.O. Box 85500, 3508GA Utrecht. Tel.: +31 887568050; Fax: +31 887555485. E-mail: J.degroot-17@umcutrecht.nl.
Received April 28, 2010; accepted October 6, 2010.

360 Park Avenue South, New York, NY 10010
1047-2797/$ - see front matter
doi:10.1016/j.annepidem.2010.10.004

AEP Vol. 21, No. 2
February 2011: 139–148

Selected Abbreviations and Acronyms

MI = multiple imputation
B&G = Begg and Greenes correction method
95% CI = 95% confidence interval
DVT = deep venous thrombosis

two-by-two table, as if all patients had received the reference standard. (for details see Appendix 1) This correction method requires knowledge about the reasons responsible for partial verification. It is disputable whether this correction method also leads to valid results when the reasons for partial verification are less clear-cut.

Recently Harel and Zhou (8) have shown that partial verification can be considered as a missing data problem and that multiple imputation (MI) methods, the practice of "filling in" missing data with plausible values, can be used to correct for this bias. Their conclusion that multiple imputation is generally better than the existing methods with regard to alleviating the bias and correcting 95% confidence interval (CI) width has been debated (9, 10). Hanley et al. (9) stated that the numerical differences between the B&G method and MI found by Harel and Zhou (8) were highly unlikely. De Groot et al. (10) concluded that these differences were due to a computational error and therefore led to spurious conclusions.

We will compare the performance of multiple imputation and the correction method of B&G under a range of situations of partial verifications using a simulation study and examine under which circumstances they produce similar results and when their results differ. Based on our findings we will propose guidance for researchers designing and analyzing diagnostic accuracy studies with partial verification.

METHODS

We have used a previously published data set, in which all patients had been verified by the reference standard. In a series of simulations, we deliberately set the outcome of the reference standard to missing based on various underlying mechanisms and by varying the total number of missing values, generating different partial verification patterns. We then compared the performance of different correction methods in each of these patterns of verification, in particular their ability to reduce the bias in estimates of accuracy by comparing it with the true values in the complete data set.

Empirical Data set with Complete Verification

Data of a large study among adults with suspected deep venous thrombosis (DVT) were used. For specific details of the study we refer to the literature (11). In brief, 1292 consecutive patients with suspected DVT were included. DVT suspicion was primarily based on the presence of swelling, redness, or pain in one of the legs. After informed consent, the physician systematically documented the patient's history and the results of a physical examination. Subsequently, venous blood was drawn to measure D-dimer level. All patients were then referred to a hospital to undergo repeated compression ultrasonography of the lower extremities, which was used as the reference standard to determine the presence or absence of DVT. Repeated compression ultrasonography revealed DVT in 251 (19%) patients, of which 225 (90%) had a positive D-dimer test result.

In our series of simulations we used the complete data of 1292 research subjects (Table 1), to which we will refer as the original study group. D-dimer test results were dichotomized, labeling results as positive if they exceeded the 1000 ng/ml threshold. The reference test used in this study (repeated compression ultrasonography) was assumed to be 100% sensitive and 100% specific. The sensitivity, specificity, and predictive values of the D-dimer test in the original study group were then calculated using standard methods (1, 12). These accuracy measures in the original study group will be referred to as the "true" sensitivity, specificity, and predictive values. Ninety-five percent confidence intervals were calculated using the Wilson "score" method (13, 14).

Patterns of Partial Verification

We selected a range of situations in which partial verification could typically arise in practice (Fig. 1).

In the first pattern of missing values, the outcome of the reference standard was set to missing in a random subset of patients with a negative D-dimer test result. This reflects a common practical situation where the practitioner thinks it unnecessary to refer all subjects with a negative

TABLE 1. Univariate association of each significant diagnostic variable with the presence or absence of DVT

Diagnostic variables	DVT present (%) (n = 251)	DVT absent (%) (n = 1041)
Patient history		
Gender + OC use:		
Males	47.4	36.2
Females using OC	12.0	9.2
Females not using OC	40.6	54.6
Absence of leg trauma	88.0	83.3
Presence of malignancy	6.8	3.3
Recent surgery	12.7	10.1
Physical examination		
Vein distention	19.5	15.5
Calf difference ⩾3 cm	62.9	34.3
Additional testing		
D-dimer abnormal ⩾1000 ng/ml	89.6	39.5

DVT = deep venous thrombosis; OC = oral contraceptive.
Values are percentages.