Case–Control and Two-Gate Designs in Diagnostic Accuracy Studies

ANNE W.S. Rutjes,1* JOHANNES B. REITSCMA,1 JAN P. VANDENBROUCKE,2 AFINA S. GLS,3 and PATRICK M.M. BOSSUYT1

Background: In some diagnostic accuracy studies, the test results of a series of patients with an established diagnosis are compared with those of a control group. Such case–control designs are intuitively appealing, but they have also been criticized for leading to inflated estimates of accuracy.

Methods: We discuss similarities and differences between diagnostic and etiologic case–control studies, as well as the mechanisms that can lead to variation in estimates of diagnostic accuracy in studies with separate sampling schemes (“gates”) for diseased (cases) and nondiseased individuals (controls).

Results: Diagnostic accuracy studies are cross-sectional and descriptive in nature. Etiologic case–control studies aim to quantify the effect of potential causal exposures on disease occurrence, which inherently involves a time window between exposure and disease occurrence. Researchers and readers should be aware of spectrum effects in diagnostic case–control studies as a result of the restricted sampling of cases and/or controls, which can lead to changes in estimates of diagnostic accuracy. These spectrum effects may be advantageous in the early investigation of a new diagnostic test, but for an overall evaluation of the clinical performance of a test, case–control studies should closely mimic cross-sectional diagnostic studies.

Conclusions: As the accuracy of a test is likely to vary across subgroups of patients, researchers and clinicians might carefully consider the potential for spectrum effects in all designs and analyses, particularly in diagnostic accuracy studies with differential sampling schemes for diseased (cases) and nondiseased individuals (controls).

© 2005 American Association for Clinical Chemistry

Determining the accuracy of a test is an essential step in the overall evaluation of medical tests. Diagnostic accuracy is the ability of a test to differentiate between patients who have the condition of interest (target condition) and those who do not. The accuracy of a test is studied by comparing the results of the test under evaluation (index test) with the outcomes of a reference standard on the same series of participants. The reference standard is the best available method to establish the presence or absence of the target condition. For dichotomous test results, the findings can be summarized in a 2x2 table and expressed as the test’s sensitivity and specificity.

Diagnostic tests must be evaluated by an appropriate design and in a clinically relevant population. The observation that the accuracy of a test varies across patient subgroups complicates the issue of patient selection in diagnostic accuracy studies (1–3). The typical approach is to include those patients who would also undergo the index test in the relevant clinical situation, to perform the index test, and then to verify the results for all patients with the reference standard.

Many variations of this design can be found in the literature. It is not always clear under what circumstances these variations in study design can change the estimates of diagnostic accuracy. This uncertainty particularly applies to diagnostic case–control studies. In such studies, groups of patients with and without the target condition are identified before the index test is performed.

Strong statements have been made about the bias of diagnostic case–control studies (1, 4–6). Case–control studies have been shown to lead to 2- or 3-fold higher estimates of diagnostic accuracy compared with studies that use single series of consecutive patients to evaluate the same test (1, 5, 6). This discrepancy seems to imply that such case–control designs should be avoided. Others
have pointed out that case–control studies may have practical benefits, as they can be less expensive and easier to perform (7).

In this report, we review how estimates of sensitivity and specificity can vary across subgroups of patients and illustrate how these spectrum effects can affect diagnostic case–control studies. After discussing some potential problems and misconceptions with case–control designs in diagnostic research, we provide what we consider a more informative labeling of these studies. Our aim is to define conditions under which case–control designs can be trusted to yield valid and unbiased estimates of a test’s diagnostic accuracy. We also delineate how awareness of the effects of enrolling patients or controls with a limited disease spectrum can be turned into an advantage for specific research questions.

Spectrum Effects and Limited Challenge
Ransohoff and Feinstein (8) were among the first to report that the performance of a test in day-to-day circumstances may be misrepresented by clinical studies that include a too-narrow range of patients with the target condition or a too-narrow range of patients without target condition. They highlighted several factors that can affect diagnostic accuracy, including pathologic, clinical, and comorbid features. Three important underlying mechanisms can lead to variation: the severity of the target condition in diseased individuals, the alternative conditions in nondiseased individuals, and the presence of comorbid conditions in either diseased or nondiseased individuals.

SEVERITY OF TARGET CONDITION
Most diseases and other target conditions cover a continuum, ranging from the first pathologic changes to overt clinical disease. For the majority of tests, the ability to detect the target condition will depend on the severity of the target condition (8, 9); e.g., larger tumors are more easily detected by imaging tests than smaller ones; larger myocardial infarctions produce higher concentrations of cardiac enzymes than smaller infarctions. Failure of the index test to identify the target condition in advanced cases is less frequent, yielding fewer false negatives and more true positives. This implies that in studies with a higher proportion of patients with more advanced stages of the target condition, estimates of sensitivities are likely to be more favorable.

ALTERNATIVE DIAGNOSES
The type of alternative diagnosis present in individuals without the target condition can also influence the performance of a test. Some alternative diseases may produce pathophysiologic changes similar to those induced by the target condition, leading to false-positive test results.

One example is the production of tumor markers by urinary tract infections rather than by cancer when these markers are used to identify patients with bladder cancer (10). The exclusion of all patients with fever in a diagnostic study designed to evaluate the accuracy of urinary tumor markers in diagnosing bladder cancer could lead to a lower false-positive rate and, hence, a higher specificity. The exclusion of “difficult” patients for a particular test is known as “limited challenge” (11).

COMORBID CONDITIONS
The presence of comorbid conditions can interfere with the performance of a test and can be responsible for false-positive or false-negative test results.

Individuals who do not have the target condition but who suffer from other diseases can be expected to produce false-positive results more often than otherwise healthy individuals. Advanced age can also lead to changes in body composition and metabolism that produce false-positive test results, in particular for diagnostic tests that are based on increased concentrations of substances that are naturally present in low concentrations in the human body, such as hormones and enzyme markers.

When a test is intended for patients with a broader age range, the predominant inclusion of individuals of advanced age could lead to “increased challenge”. An increase in false positives can also occur when the sampling scheme focuses on the inclusion of patients with poor general health status. Individuals who do not have the target condition but who suffer from other diseases can be expected to produce false-positive results more often than healthy individuals.

Alternatively, comorbid conditions can hinder the detection of the target condition by the index test, leading to false-negative results. ELISA tests in microbiology aim to detect specific antibodies produced in response to infection. False-negative ELISA results can occur if patients are immunocompromised or take corticosteroids and fail to produce sufficient antibodies when infected. Studies that exclude immunocompromised patients may produce more favorable sensitivities. Another example is antibiotics administered to hospitalized patients with unexplained fever when urinary cultures are used to detect urinary tract infections. The frequency of false-negative test results will be higher in patients taking antibiotics that reduce the growth of bacteria, thereby impairing detection. Studies excluding patients on antibiotics are likely to produce more favorable sensitivities compared with studies including patients on antibiotics.

The mechanisms discussed here explain why diagnostic accuracy is not a feature of the test itself but a description of how the test behaves in a particular clinical population. We will explore how these issues affect diagnostic case–control studies after introducing case–control studies in general.

Case–Control Design in Etiologic Studies
In epidemiology, case–control studies are used to answer questions about etiology. The typical way of thinking about etiology is from cause to effect; for example, to
ascertain whether smoking causes lung cancer, one might imagine a study that enrolls a large group of apparently healthy men (presumably without lung cancer), measures the extent of their exposure (smoking), and uses follow-up to determine the incidence of lung cancer. In the analysis, the extent of exposure (amount of smoking) is related to the incidence of lung cancer to quantify the effect of smoking on the incidence of lung cancer. Such a design is known as a “cohort study”. One sine qua non for causality is temporality: the necessity that the cause precedes the disease in time.

Etiologic case–control studies reverse the order of investigations and start at the end: individuals who have developed lung cancer (cases), the disease of interest, are compared with a group of individuals who are free of lung cancer (controls) and who represent source population from which the cases emerged. For both groups, past exposure (smoking behavior) is determined. In the analysis, the relative frequency of exposure among cases and controls is compared with the calculation of an odds ratio, which is a measure of the relative risk (12).

Case–control studies can lead to a considerable gain in efficiency compared with prospective cohort studies. The main reason is that researchers can bypass the time- and money-consuming efforts required by long-term follow-up of every person in the cohort to determine whether the event of interest will occur. In particular, for diseases with a long latency period—a long period between first exposure and onset of disease—the savings in time and money can be substantial. In addition, case–control studies examine many fewer patients but can obtain the same results as cohort studies with only a small loss of precision. Especially when the absolute risk for disease occurrence is small, a cohort design requires a large cohort to reach adequate numbers of events with sufficient power to estimate the association between exposure and disease, whereas a case–control design obtains approximately the same confidence interval by using all available cases but only a sample of the excessive number of potential controls from the source population.

The information on past exposures of cases and controls often comes from interviews with cases and controls or from existing medical records. Erroneous estimation of exposure can occur for many reasons. When recall of exposure differs for cases and controls, “recall bias”, a form of information bias, presents a major threat for etiologic case–control studies (12, 14).

“Confounding” is an important issue in etiologic studies. A “confounder” is a variable responsible for a distorted reflection of the association between the exposure of interest and the outcome (12). Sex could be a confounder, for example, in the association between smoking and lung cancer. If women smoke less and have an inherently lower risk for lung cancer, the gender confounder may distort the calculated association between smoking and the occurrence of lung cancer.

Sampling of cases and controls is critical in etiologic case–control studies, where a distinction can be biased between population-based and non-population-based studies. In population-based studies, both cases and controls come from a well-defined source population.

**Case–Control Design in Diagnostic Studies**

The logical starting point for a prototypic diagnostic accuracy study is a consecutive series of individuals in whom the target condition is suspected. The index test is performed first in all participants, and subsequently, the presence of the target condition is determined by performing the reference standard (Fig. 1A). This design resembles the cohort design in epidemiology because individuals are enrolled before the final outcome (presence or absence of the target condition) is known.

The label “diagnostic case–control studies” has been used to refer to studies in which the disease status is already known before the index test is performed. This distinction explains the rationale for speaking about “cases” and “controls”. In this analogy, the outcome of interest has already been detected by the reference standard, and the index test is the exposure.

Unfortunately, this terminology can also lead to confusion, as there are some important differences between etiologic and diagnostic case–control studies. The fundamental difference between etiologic and diagnostic studies is that, unlike etiologic studies, diagnostic accuracy studies are cross-sectional in nature (7, 15). Their aim is to compare the result of the index test with that of the reference standard in the same participant at the same time. In this, they differ from etiologic studies, in which there is a time window between exposure and the occurrence of disease.

Etiologic studies want to eliminate confounding when assessing the effect of a potential causal exposure. In contrast, diagnostic associations between the index test and the reference standard are purely descriptive, without any causal connotation. Several important concerns in etiologic case–control studies do not transfer to diagnostic studies. An example is recall bias, a major source of information bias in case–control approaches within epidemiology, as explained above (13, 14).

Because of the cross-sectional nature of diagnostic case–control studies, some of the efficiency gains of etiologic research case–control studies do not apply in the diagnostic setting. Etiologic case–control studies can bypass the costly operation of following participants over time from exposure to occurrence of disease. These efficiency gains hardly apply in diagnostic research, where ideally, the index test and the reference standard would be performed at the same time. In diagnostic accuracy studies, case–control designs can bring other benefits, including efficiency gains, as explained below.
Types of Case–Control Designs in Diagnostics Studies

REVERSED-FLOW DESIGNS: THE IMPORTANCE OF A “SINGLE GATE”

The cross-sectional nature of accuracy studies is highlighted by considering a design in which the index test and reference standard are performed in reverse order (Fig. 1B). This design has been referred to as “retrospective sampling”, although data collection can be either prospective or retrospective (16). Often in such designs the reference test is applied only to a subsample of the participants with or without the target condition. Strictly speaking, these designs can also be labeled as case-control designs. To reduce confusion, however, we propose the label “reversed-flow design” for this setup.

The reversed-flow design indeed bears some similarities to the population-based case-control design in etio-
logic epidemiology. In a population-based case–control design, both cases and controls are sampled from a single source population. In a reversed-flow diagnostic accuracy study, cases and controls are also sampled from the same patient population.

Simply reversing the order in which the index test and reference standard are performed will not change estimates of diagnostic accuracy, such as sensitivity and specificity, as long as the same group of patients is included and all participants in the study undergo both the index test and reference standard. All patients pass through a single gate: a single set of criteria for study admission, typically defined by the clinical presentation.

A reversed-flow design can have practical advantages, as when researchers adjust the order in which they perform the index test and reference standard in response to the availability of material and human resources. Another potential benefit can be seen in situations in which the prevalence of the target condition is low, when the index test is costly, or when this test has potential side effects. In these situations, a reversed-flow design enables the researcher to balance testing costs by taking a random sample of patients with a negative result on the reference standard and performing the index test only for these patients as well as for all reference-standard–positive patients (16).

Smith et al. (17) used a reversed-flow study design to evaluate plasma B-type natriuretic peptide in detecting left ventricular systolic dysfunction in elderly patients. They screened a random sample of 817 patients from general practice with echocardiography. Random subsamples of patients with (n = 12) and without (n = 143) left ventricular systolic dysfunction were then asked to undergo venipuncture to assess the concentration of B-type natriuretic peptide, the index test under study.

In a study of second-trimester ultrasound to detect fetuses with Down syndrome, Bromley et al. (18) sampled all 53 fetuses with Down syndrome karyotypes from 4075 genetic amniocenteses. A subseries of 177 consecutive non-Down syndrome fetuses from the same set of amniocenteses served as controls. The authors then reanalyzed the previously performed ultrasound measurements in these 230 pregnancies only, rather than analyzing all 4075 images. With random sampling, the estimate of specificity is expected to be valid at the expense of a minimal loss in precision for specificity, i.e., a slightly broader confidence interval.

**TWO-GATE DESIGNS USING HEALTHY CONTROLS**

A different situation emerges when cases and controls are sampled from 2 distinct source populations (Fig. 1C). Diseased individuals, for example, are sampled from a clinical (hospital) population, whereas young, healthy controls are sampled from the general population. We refer to this as a “two-gate design using healthy controls”.

Two different sets of inclusion criteria (gates) are used: one for the diseased and another for the nondiseased participants.

For the same test, studies with two-gate design using healthy controls have been shown to produce inflated estimates of diagnostic accuracy compared with studies using a cohort of consecutive patients (single-gate study) (1, 5, 6). On average, the diagnostic odds ratio was 3-fold higher in two-gate sampling using healthy controls vs single-gate studies (5).

Spectrum effects and limited-challenge bias can explain the inflated accuracy measures in studies with two-gate sampling. Inclusion of individuals with advanced disease (the sickest of the sick) will generate fewer false-negative test results than the inclusion of more patients with limited disease. Estimates of sensitivity, therefore, are likely to be more favorable. In addition, estimates of specificity are probably higher if healthy volunteers are used as controls. Most volunteers will be without complaints and, hence, unlikely to have alternative diagnoses that generate false-positive results (the fittest of the fit).

Although the results of case–control studies with healthy volunteers may have limited applicability to clinically relevant situations, they can be useful in the early phase of the development of a test: to screen whether a test is of any use (4, 7). Disappointing results in early studies with this design can be a reason to stop further development of the test.

Healthy controls were used in the study of Che et al. (19), who evaluated a newly developed monoclonal antibody–based capture enzyme immunoassay for the detection of severe acute respiratory syndrome (SARS). The assay was tested in 13 patients with serologically confirmed SARS and in 1272 healthy blood donors. Specificity was high: 99% of the healthy volunteers had a negative test result. Because of the low prevalence of SARS worldwide (8422 total cases) at the time of the study (20), a single-gate (cohort) design would not have been feasible.

**TWO-GATE DESIGN WITH ALTERNATIVE DIAGNOSIS CONTROLS**

A different form of two-gate sampling includes only control participants diagnosed with a specific alternative condition known to produce symptoms and signs similar to those of participants with the target condition (Fig. 1D). We refer to this design as a “two-gate design with alternative diagnosis controls”.

As in any two-gate design, the selection of cases is crucial. An overrepresentation of patients with advanced disease will lead to inflated estimates of sensitivity, whereas overrepresentation of patients with mild disease will underestimate sensitivities. In a review evaluating the accuracy of urinary tumor markers in the detection of bladder cancer, Glas et al. (10) found that studies including cases with low-grade disease were associated with lower sensitivities than studies with single-gate sampling.
Depending on the type of alternative diagnosis included, specificity may be over- or underestimated. In a single-gate design with appropriate sampling, all alternative diagnoses will be represented in the group of patients with a negative reference standard outcome, with the likelihood of a false-positive test result depending on the alternative diagnosis. Sampling patients with a single alternative diagnosis may generate more or fewer false-positive results, depending on the alternative diagnosis.

The literature contains numerous examples of two-gate design with alternative diagnosis controls. Hoffman et al. (21) included 21 publications in a metaanalysis of the diagnostic performance of the ratio of free to total prostate-specific antigen to detect prostate cancer. This set included 13 studies with a two-gate and 11 studies with a single-gate design. Three studies with a two-gate design used healthy controls (22–24), whereas the other 9 two-gate studies used controls with benign prostatic hyperplasia (23–31). One two-gate study reported only that the controls had a negative biopsy (32). Although further description of the control group was lacking, it is likely that this study used controls with benign prostatic hyperplasia as well.

Two-gate designs are often applied in clinical chemistry, where previously stored samples of blood and urine are used to evaluate a new test. In some studies, disease status is derived from patient charts. An adequate description of the study group is often lacking in the corresponding publications, complicating evaluation of the potential for bias (33).

In general, two-gate designs with alternative diagnosis controls can be informative because they provide data on the likelihood of false-positive results in specific subgroups. The proportion of patients with true-negative index test results, however, may not be equal to the specificity of the test. The latter equals the prevalence-weighted proportion of true negatives over all alternative diagnoses in the clinical situation in which the test is to be applied.

**TWO-GATE DESIGN WITH REPRESENTATIVE SAMPLING**

Estimates of sensitivity and specificity should be valid in a two-gate design if the group of cases is sampled in such a way that they match the group of reference-standard-positive patients in a single-gate design in terms of the spectrum of the target condition and if the group of controls matches the group of reference-standard-negative patients in terms of the relative representation of alternative conditions. We call this a “two-gate design with representative sampling”.

The difference between a two-gate design with representative sampling and the reversed-flow design is that the two-gate design still has two sets of inclusion criteria: one for cases and one for controls. Such a two-gate design with representative sampling may be difficult to realize, and we are not aware of any examples in the literature.

**Summary**

Diagnostic accuracy studies in which the presence of the target condition is known before the index test is performed are typically referred to as diagnostic case–control studies. We have highlighted some fundamental differences between diagnostic and etiologic case–control studies. Because of the cross-sectional nature of diagnostic case–control studies and the importance of timing in diagnostic research, not all efficiency gains in etiologic case–control studies transfer to the diagnostic setting.

The applicability of findings from diagnostic case–control studies is determined by spectrum effects and limited challenge. The guiding principle in all epidemiologic studies is to match patient selection to the object of study. The same principle applies to diagnostic studies. In etiologic case–control studies, a differential selection of cases and controls according to exposure history will ruin the study, as cases and controls no longer represent the same population. The resulting odds ratio, therefore, will be invalid. The situation is more complex in diagnostic studies because the object of a diagnostic accuracy study can vary, depending on the phase of test development. In an early phase of development, two-gate sampling studies with healthy controls or controls with a specific alternative diagnosis can be used to answer specific questions about a test’s potential or to study its behavior in specific subgroups of patients. These designs, however, may not provide information about a test’s specificity or sensitivity in the clinical setting in which it is to be applied. For that purpose, single-gate designs and reversed-flow designs are more appropriate.

In this report, we have focused on issues of patient selection and how they can affect measures of diagnostic accuracy in case–control designs. Several other factors can also lead to bias or variation in accuracy studies (1). These factors include the use of suboptimal reference standards, as well as incomplete and differential verification. These types of biases are not specific to particular designs, and measures to avoid them can differ among designs.

Because the accuracy of a test is likely to vary across subgroups of patients, researchers and clinicians might carefully consider the potential for spectrum effects in all designs and analyses, in particular in studies with two-gate sampling. Critical appraisal of reports on diagnostic accuracy research can help investigators decide whether the evidence about a diagnostic test is valid, clinically relevant, and applicable to specific patient groups or individuals. For that purpose, investigators need information on the inclusion and exclusion criteria, settings and locations of data collections, and methods of participant recruitment and sampling (34, 35).
This study was funded in part by a research grant from the Netherland Organisation for Scientific Research (NWO; registration no. 945-10-012).

References