

ORIGINAL ARTICLE

Development of practical recommendations for diagnostic accuracy studies in low-prevalence situations

Gea A. Holtman^{a,b,*}, Marjolein Y. Berger^b, Huibert Burger^b, Jonathan J. Deeks^{c,d},
Norbert Donner-Banzhoff^e, Thomas R. Fanshawe^a, Constantinos Koshiaris^a,
Mariska M. Leeftang^f, Jason L. Oke^a, Rafael Perera^a, Johannes B. Reitsma^g,
Ann Van den Bruel^{a,h}

^aNuffield Department of Primary Care Health Sciences, Radcliffe Observatory Quarter, University of Oxford, Oxford OX2 6GG, UK

^bDepartment of General Practice and Elderly Care Medicine, University Medical Centre Groningen, University of Groningen, PO Box 196, 9700 AD Groningen, the Netherlands

^cTest Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, Birmingham B15 2TT, UK

^dNIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust, University of Birmingham, Birmingham B15 2TT, UK

^eDepartment of General Practice and Family Medicine, Faculty of Medicine, Philipps University of Marburg, Karl-von-Str. 4, Marburg 35037, Germany

^fDepartment of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, the Netherlands

^gJulius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, PO Box 85500, 3508 GA Utrecht, the Netherlands

^hAcademic Centre of General Practice, University of Leuven, Kapucijnenvoer 33 blok J, Bus 7001, 3000 Leuven, Belgium

Accepted 22 May 2019; Published online 28 May 2019

Abstract

Objective: Low disease prevalence poses challenges for diagnostic accuracy studies because of the large sample sizes that are required to obtain sufficient precision. The aim is to collate and discuss designs of diagnostic accuracy studies suited for use in low-prevalence situations.

Study Design and Setting: We conducted a literature search including backward citation tracking and expert consultation. Two reviewers independently selected studies on designs for estimating diagnostic accuracy in a low-prevalence situation. During a 1-day expert meeting, all designs were discussed and recommendations were formulated.

Results: We identified six designs for diagnostic accuracy studies that are suitable in low-prevalence situations because they reduced the total sample size or the number of patients undergoing the index test or reference standard depending on which poses the highest burden. We described the advantages and limitations of these designs and evaluated efficiencies in sample sizes, risk of bias, and alignment with the clinical pathway for applicability in routine care.

Conclusion: Choosing a study design for diagnostic accuracy studies in low-prevalence situations should depend on whether the aim is to limit the number of patients undergoing the index test or reference standard, and the risk of bias associated with a particular design type. © 2019 Elsevier Inc. All rights reserved.

Keywords: Diagnostic accuracy studies; Low prevalence; Primary care

1. Introduction

A low-prevalence situation occurs when in a group of patients suspected of a particular disease or target condition, only a few will actually have the disease in question. For

example, about 1 in 170 adults who present with persistent cough in primary care will have lung cancer [1], and 1 in 125 children with an acute illness in primary care will have a serious infection [2]. This situation is quite common in primary care, but can also occur in screening situations or when diagnosing rare diseases in secondary and tertiary care.

The classic design for estimating diagnostic accuracy is a cross-sectional cohort study recruiting consecutive patients suspected of the target condition who subsequently receive the index test and the reference standard (Fig. 1A).

Conflict of interest: None.

* Corresponding author. Tel.: +31 642637125; fax: +31 (0) 503632964.

E-mail address: g.a.holtman@umcg.nl (G.A. Holtman).

What is new?**Key findings**

- We present a general guide for choosing the most suitable design in different low-disease prevalence situations.
- Suggestions for computing confidence intervals, weighted analysis, sample size calculations, binary logistic regression analysis, and population weighting are provided.
- The advantages and limitations of using health care data for diagnostic accuracy studies are discussed.

What this adds to what was known?

- An overview of studies designs and methods that could be used for diagnostic accuracy studies in low-disease prevalence situations is lacking.

What is the implication and what should change now?

- This overview provides insight on feasibility of diagnostic accuracy studies including study budget, so that important evidence on diagnostic tests in low prevalence settings can be established.

Figure 2 shows that the challenge in low-prevalence situations is that a classic design requires very large sample sizes to estimate sensitivity with sufficient precision in diseased patients and consequently to calculate other measures that involve the diseased group, for example, likelihood ratios and predictive values. For example, the prevalence of inflammatory bowel disease in children with chronic gastrointestinal symptoms presenting to primary care is around 1% [3,4], such that only 61 cases would be expected in a cohort of 6,146 children. If 49 of these 61 cases were index test positive, the estimated sensitivity of 80% would have a confidence interval ranging from 70% to 90% [5]. Such studies are not only logistically difficult to achieve, but also subject a large group of patients to both the index test and reference standard, raising ethical and financial issues depending on their level of invasiveness and costs.

Consequently, diagnostic accuracy studies in low-prevalence situations either include so few diseased patients producing imprecise estimates of diagnostic accuracy parameters or are so large that they risk being undeliverable and unaffordable. Uncertainty of the value of diagnostic tests in low-prevalence situations hampers routine practice, particularly in primary care where important conditions appear at low prevalence within the large population of health care contacts [6]. Study designs or methods that require smaller sample sizes and are therefore less

challenging, yet yield unbiased and sufficiently precise estimates of test accuracy are thus highly desirable. To the best of our knowledge, an overview listing all available study designs and methods does not exist. The aim of this article is to identify and discuss designs and methods of diagnostic accuracy studies that can be used in low-prevalence situations.

2. Methods*2.1. Search strategy*

A literature search to identify relevant methods was conducted in the Cochrane Library, Embase, MEDLINE, and Web of Science databases from inception until July 14, 2017. For the search strategies see [Appendix 1](#). In addition, invited experts in the area of diagnostic research reviewed the list of selected studies to identify any omissions and provide suggestions for other methods or study designs not covered by the literature review. Finally, we performed backward citation tracking on Scopus and Web of Science of all included studies. No language restrictions were applied.

2.2. Study selection

Two reviewers (A.V.dB. and G.A.H.) independently selected eligible studies. We included all studies describing a specific study design or method for estimating diagnostic accuracy in a low-prevalence situation. Low prevalence was defined as a maximum of 10% of the suspected population with the target condition. Studies on prognostic tests or impact studies of diagnostic tests were excluded, but screening studies included. Studies were selected on title and abstract first and on full-text articles thereafter. Disagreement between reviewers was resolved by discussion.

2.3. Evaluating methods

Data were extracted by one reviewer (G.A.H.) in a predefined extraction sheet. We invited a group of European experts in the area of diagnostic research in low-prevalence situations with varying expertise, that is, clinical, methodological, or statistical expertise. During a 1-day meeting with the experts, the identified designs and methods were discussed with regard to their advantages and limitations, and recommendations were formulated. In particular, we looked at efficiencies in sample sizes, risks of bias, and alignment with the routine clinical pathway for applicability in routine care. In addition, we discussed considerations that could help choosing the most suitable design in different low-prevalence situations: depending on whether patient selection, the burden associated with index test or reference standard, or the target condition itself poses the largest problem, one design may be more suitable than another. We made a general guide taking into account the main risks of bias according to the QUADAS-2, the quality assessment tool for diagnostic accuracy studies [7,8].

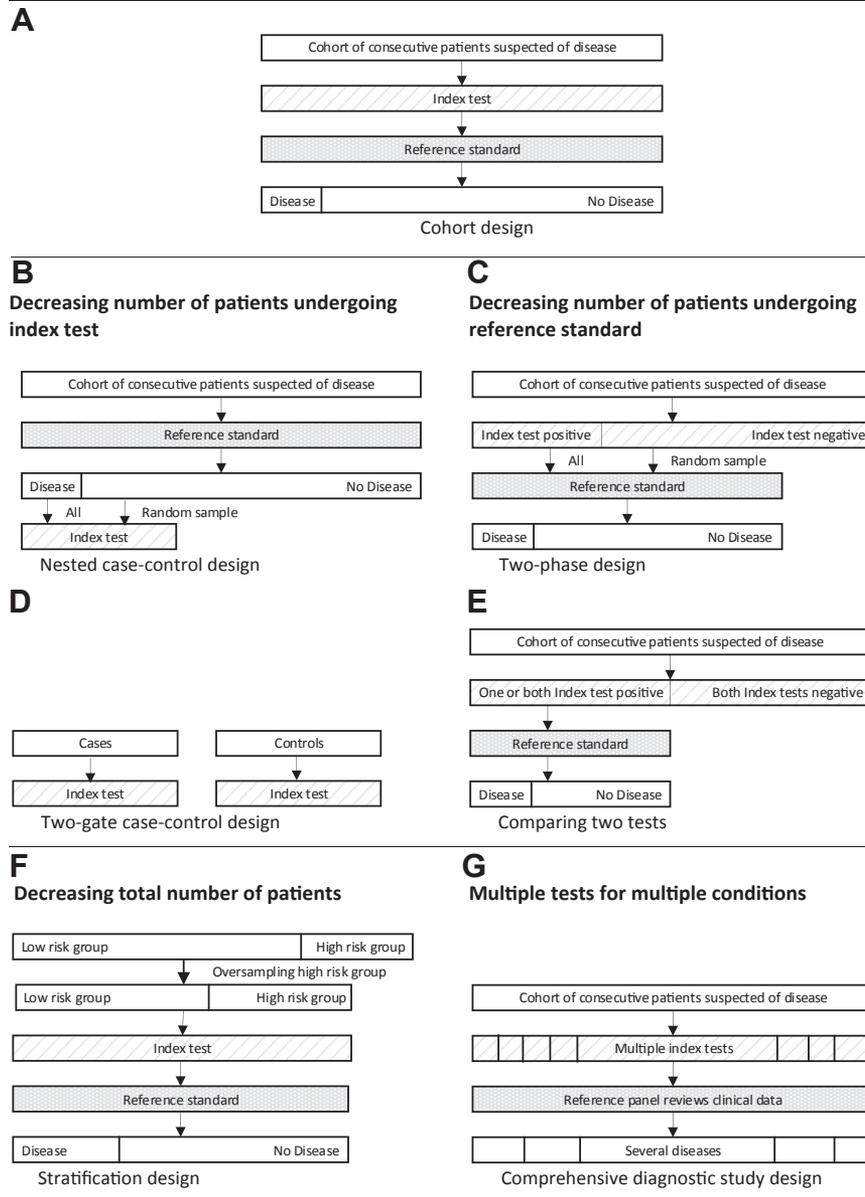


Fig. 1. Study designs for diagnostic accuracy studies.

3. Results

The literature search, expert input, and backward citation tracking yielded 16 studies (Appendix 2). All studies and their details are shown in Appendix 3. We categorized the identified methods in the following sections: designs, analytic considerations, and using routine health care data. We identified six designs that are particularly efficient for diagnostic test evaluation in low-prevalence situations (Figure 1). Table 1 provides an example of each design. In appendix 4, we comment on analytical considerations such as sample size calculations, computing confidence intervals, weighted analysis, binary logistic regression, and population weighting in these situations.

3.1. Designs

3.1.1. Design decreasing the total sample size

3.1.1.1. Stratification design. In the stratification design, the population of interest is divided in two subgroups by prevalence to oversample the higher prevalence subgroup (Fig. 1F). This leads to more diseased patients and fewer nondiseased patients being recruited into the study. The oversampling is subsequently accounted for in the analyses by weighting the sensitivity and specificity of the subgroups to compute estimates for the entire population of interest [9].

The design leads to lower sample sizes overall, but every participant undergoes both index test and reference

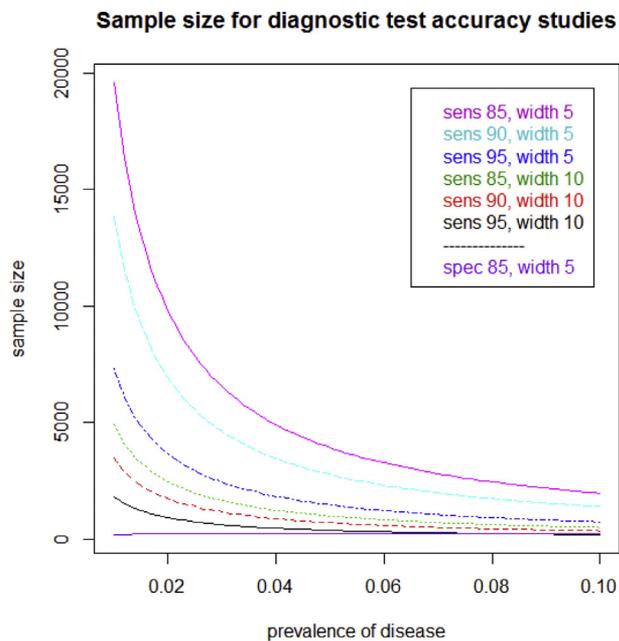


Fig. 2. Sample size for the classical design of diagnostic test accuracy studies with different disease prevalence.

standard. This means the design can be useful when either index test or the reference standard is invasive, costly, or logistically difficult to organize.

To be able to oversample one subgroup within the population of interest, the stratification design requires prior knowledge on the prevalence within each stratum, and knowledge on how to distinguish the different strata. For example, if we want to evaluate the accuracy of a new test for age-related macular degeneration, prevalence of the target condition will be higher in older patients [16]. The stratification design would require knowledge of the prevalence of age-related macular degeneration in different age groups, and the sampling fraction in each age group will be determined accordingly before starting recruitment.

If screening for patient eligibility and measuring the variables that are used for stratification are themselves invasive or costly, this design may be less efficient. In addition, it will be more efficient with greater difference in prevalence between both subgroups and a large subgroup with higher prevalence in the overall population because it results in smaller overall sample sizes for a fixed number of cases [9]. Moreover, the validity of the overall estimate relies on diagnostic accuracy across strata being similar, but it is unlikely that the data would allow formally assessing this as the stratification design leads to a low number of cases in the lowest prevalence strata. Therefore, it is important to have reasonable evidence that the characteristics of the cases detected in low- and high-prevalence cases are similar. However, this evidence is often lacking and sometimes evidence shows that tests are more sensitive when the prior probability is higher.

Considering the stratification design is prospective, it is possible to undertake the index test at the point in the diagnostic pathway where it would be used in practice, increasing applicability of the results to routine care.

3.1.2. Designs decreasing the number of patients undergoing the index test

3.1.2.1. Nested case-control design. In a nested case-control study (single-gate case-control study), cases and controls are both sampled from a previously enumerated cohort ensuring representativeness (Fig. 1B) [17]. This cohort is formed by the total population of patients initially suspected of having the target condition; the diseased are all cases occurring in the suspected population and a sample of controls can be chosen randomly.

To ascertain the presence of the target condition, all cohort members need to be subjected to the reference standard. Where disease is rare, index test results are obtained from all cases, and a random sample of the controls. The nature of the index test will depend on how this is logistically done—if the test uses samples that can be safely stored at low cost, these could be obtained from all participants at the stage in the diagnostic pathway where the test would be used in clinical practice, and later selected for testing once disease status is known. For example, venous blood taken at study entry in the GRACE cohort study, was analyzed for C-reactive protein at a later stage in 100 patients with and 100 patients without pneumonia. Considering entry criteria were similar and the prevalence of pneumonia was known, diagnostic accuracy of different C-reactive protein devices could be calculated [12]. Otherwise, participants are selected to undergo the index test at a second time point once disease status is known, and the applicability of the results would be compromised if test-related health states have changed over time. Cohorts that allow further testing with a new index test are not very common, limiting the applicability of the design.

The nested case-control design is advantageous over the cohort design when obtaining the index test results is associated with high burden or cost. An increasing number of controls results in test characteristics that are more precise. A further advantage of the nested case-control design over a two-gate case-control design, in which cases and controls are sampled separately from different source populations, is the possibility to calculate positive and negative predictive values because the sample size of the source population, that is, the size of the cohort, with the overall disease prevalence and hence the sampling fractions are known [17].

3.1.2.2. Two-gate case-control design. In a diagnostic case-control study, patients with and without the disease are selected after disease status has been established, and then undergo the index test (Fig. 1D). This means that the index test will be conducted on two different source populations, hence the term two-gate case-control design

Table 1. Clinical examples of different designs for diagnostic accuracy studies in low-prevalence situation

Design	Population	Target condition	Index test	Reference standard
Stratification design (Obuchowski 2002) [9]	Patients with and without coronary artery disease (CAD)	Myocardial infarctions	New test for silent myocardial infarctions (MI)	Magnetic resonance imaging
Two-phase design (Duvekot 2015) [10]	Children referred to mental health centers	Autism spectrum disorder	Parent and teacher-reported Social Responsiveness Scale (SRS)	Developmental, Dimensional, and Diagnostic interview (3Di) and Autism Diagnostic Observation Schedule (ADOS)
Two-gate case-control design (Ukoumunne 2017) [11]	Cases: hearing impaired children recruited from audiology services Controls: children with no previous identified impairment recruited from schools	Hearing impairment	Pure tone screen and HearCheck Screener	Pure tone audiometry
Nested case-control design (Minnaard 2015) [12]	Adult patients presenting with acute cough (the Genomics to combat Resistance against Antibiotics in Community-acquired LRTI in Europe [GRACE] Network)	Pneumonia	Five point-of-care test devices for estimating C-reactive protein (CRP) and the laboratory analyzer	Chest radiographs
Comparing two tests without verification of double negatives design (Armitage 1985) [13]	Colorectal cancer screening population	Colorectal cancer	Two fecal occult blood tests (chemical and immunological test)	Sigmoidoscopy
Comprehensive diagnostic study design (Bosner 2009 and Donner-banzhoff 2014) [14,15]	Patients presenting with chest pain in primary care	Multiple outcome categories, e.g., acute coronary syndrome, severe respiratory disorders, psychogenic causes	Symptoms and findings	Interdisciplinary reference panel reviewed clinical data of each patient

[18]. The selection of these populations can greatly affect the diagnostic accuracy measures. For example, if those without disease are healthy controls, there may be fewer false-positive results than when used in patients suspected of disease because alternative diagnoses may also lead to positive results on the index test, and hence specificity in healthy controls will be overestimated. Similarly, if the cases include patients with more advanced disease, there may be fewer false-negatives than in patients suspected of disease, and sensitivity may be overestimated [19]. Empirical evidence suggests that accuracy (expressed as a diagnostic odds ratio) may be overestimated by on average a factor of five [18]. In addition, the flow of the study is such that the index test can only be performed after the disease status of patients has been established and cannot be performed at the point in the diagnostic pathway, where it would be used in practice, making it less suited for target conditions that progress rapidly such as infectious diseases, or situations where test accuracy in incident cases differs from that of prevalent cases. Matching on patient covariates such as patient characteristics should be carefully considered in a case-control study as

it can complicate the estimation of the added value of the test [20].

The main advantage of the case-control design is the ease of implementation and the lower cost. They have a role as an early-stage study carried out to investigate whether a new test has potential and require confirmation by a different, prospective study design. Poor test accuracy in a case-control study (especially when healthy controls are used) can stop further evaluation of the test. Case-control studies using data from biobanks, such as EuroBioBank and the 100,000 Genomes Project, are valuable in the evaluation of future new markers in patients with rare diseases [21,22].

A variation of the case-control design includes sampling of patients with an alternative diagnosis (instead of healthy controls) with similar symptoms as the target condition. This can result in overestimation or underestimation of the specificity measures depending on the alternative diagnosis used. For example, if the alternative diagnosis produces positive test values similar to those seen in cases, then the false-positive rate is increased which results in an underestimation of specificity. Therefore, the calculated

specificity cannot be applied in routine care. However, this type of evaluation provides information on the likelihood of false-positive test results in specific control population(s).

3.1.3. Designs decreasing the number of patients undergoing the reference standard

3.1.3.1. Two-phase design. The objective of the two-phase design is to estimate diagnostic accuracy while not requiring that each patient in the study undergoes the reference standard (Fig. 1C). Instead, all patients receive the index test (phase 1), but only those with a positive index test result, and a random sample of those with a negative index test result, go on to receive the reference standard (phase 2) [9]. The sampling fraction of those with a negative index test result is used to reweight sensitivity and specificity estimates.

The main advantage of this design in low-prevalence scenarios is that it improves efficiency by prioritizing positive over negative index test results, reducing the number of times the reference standard needs to be performed. This may be beneficial when the reference standard is invasive, expensive, or logistically difficult to organize. An example arises in psychiatric diagnosis, when the reference standard often requires a detailed interview of the participant [10,23].

Irwig et al. provide estimates of the resulting savings in sample size for different sensitivity and specificity combinations and conclude that greater savings accrue for lower values of sensitivity [24]. Therefore, when the expected sensitivity is high, the total number of patients recruited is often increased compared with cohort design [9]. Consequently, the design is more suited to situations in which the index test is inexpensive or routinely used. In all designs, bias may occur if there is differential dropout, but this design is more exposed to this particular type of bias because the reference standard is often invasive and there could be a period between the two phases. The index test can be evaluated at a point in the diagnostic pathway where it would be used in practice.

Baker et al. discuss a similar design that they term a “partial testing design”, in which a sampling fraction for phase 2 testing (reference standard) is also used among individuals who receive an initial positive result [25]. Moreover, a third phase could be used in which patients who did not undergo the reference standard are followed up for a certain period [25].

3.1.3.2. Comparing two tests without verification of double negatives. A design used for evaluating tests when the accuracy of two alternative index tests needs to be compared involves both index tests being used in all recruited participants, but only those found to be positive on at least one of the index tests receiving the reference standard (Fig. 1E) [26–28]. This design can be very efficient in low-prevalence situations, particularly for index tests with high specificity, as none of the many individuals who are negative on both tests undergo the reference standard investigation. This may be particularly desirable when the reference standard is invasive or expensive. While it is not possible to obtain estimates of the actual sensitivity and specificity of

either test from this design, valid estimates of the relative sensitivity and relative false-positive (1-specificity) fractions can be obtained, together with valid tests of their statistical significance [27]. If one test shows higher true-positive and lower false-positive rates clearly, dominance can be concluded; should one test show both higher true-positive and higher false-positive rates, the preferred test depends on the ratio of extra true positives to extra false negatives which can also be computed [27]. If required, estimates of sensitivity and specificity for both tests could be estimated indirectly if values for one test are known from existing studies, or additionally verifying a random sample of participants with both test negative as in a two-phase design and using a weighted analysis.

A requirement for use of this study design is that it is acceptable and ethical to use both index tests at the same time in each individual. This may not be possible if index tests are burdensome or if the quantity of specimen available for analysis is inadequate. The interpretation of each index test should also be undertaken independently.

3.1.4. Design for multiple tests and multiple conditions

3.1.4.1. Comprehensive diagnostic study design. All study designs discussed so far have in common that they evaluate the accuracy of index tests for one target condition only. Primary care physicians, however, are confronted with a wide range of conditions. The comprehensive diagnostic study provides estimates of test accuracy of multiple index tests and for more than one target condition, including more and less prevalent conditions (Fig. 1G) [14].

Patients are recruited on the basis of a particular symptom or finding. For each patient, further symptoms, physical findings, and the results of investigations recorded. Some tests are mandatory; others are left at the discretion of the clinician. Depending on the disease in question, patients are followed up for a defined period, after which a reference panel reviews clinical data to determine each patient’s status regarding the target conditions of interest [29].

As a result, the efficiency of a comprehensive diagnostic study is in providing estimates of the prevalence [30] and accuracy of selected diagnostic tests for several target conditions in a single study. Moreover, it allows the derivation and validation of complex prediction rules or strategies for the presenting problem, which may include complex algorithms for computerized decision support systems.

While a comprehensive diagnostic study closely reflects real life, especially for primary care clinicians, the amount of data required may pose logistical, ethical, and financial challenges. Moreover, there are no efficiencies in this design for the accuracy of a single test for a low-prevalence condition. However, it allows information to be gleaned about low-prevalence conditions alongside others that have higher prevalence, which potentially makes it a more efficient design than doing separate independent studies for each disease.

Table 2. General guide for designs in different low-prevalence situations

Low-prevalence situation	Stratification design	Nested case-control design	Case-control design
Population			
Different strata in patient population	Only option		
Predefined cohort		Only option	
Index test			
Costly or invasive index test	Second option	First option	Third option
Two index tests	Second option		
Multiple index tests			
Reference standard			
Costly or invasive reference standard	Third option		
Target condition			
Extremely low prevalence (<0.1%)			Only option ^a
Multiple target conditions			
Main issues QUADAS-2	Higher risk of bias if there is more difference between prevalence strata	Higher risk of bias if the selection of index test is not random and if the index test is carried out outside the care pathway	Higher risk of bias if cases or controls do not represent the same population and if index test is carried out outside the care pathway

If there were more than one design applicable, the first three options were ranked on suitability based on sample size and risk of bias. Other designs might be possible as well in specific situations.

^a Another option is to conduct a cohort study using routine healthcare data containing millions of patients.

4. Practical recommendations for design choice in low-prevalence situations

Table 2 provides a general guide of the three most suitable designs for different low-prevalence situations, taking into account their respective advantages and limitations. In addition, it is important to specify the goal for which a test is used and whether it is more important to have a high sensitivity or a particularly high specificity, and estimate these with sufficient precision. For example, a triage test requires a high sensitivity and an invasive or dangerous follow-up or treatment requires a high specificity. If sensitivity is considered more important than specificity, it is a challenge to estimate this with sufficient precision in low-prevalence settings. However, estimating specificity with sufficient precision is less of a problem in relation to low prevalence because there are more nondiseased patients than diseased.

When the index test is costly or invasive, three options are available: the nested case-control design reduces the number of index tests by testing a random sample of patients with a negative reference standard and all patients with a positive reference standard, resulting in a low risk of bias. We consider the stratification design a good alternative when it is difficult to recruit a large population in general, or when the index test cannot be performed after the selection of the cases, for example, no stored data/samples available or the target condition if progressing rapidly. Finally, a two-gate case-control design is the last option but is prone to bias.

To compare the diagnostic properties of more than one index test, one could opt for a study in which patients who have a negative test result for both tests are not verified. The design will allow ranking of the two index tests based on the number of false-positives for each true-positives detected by one test compared with the other. Two other designs could be considered: stratification design and two-phase design. The stratification design results in higher savings than the two-phase design [9].

When the reference standard is costly or invasive, two designs would apply: the two-phase design and the design comparing two tests without verification of double negatives. Although the latter results in a higher reduction in the number of patients having to undergo the reference standard, this design only allows calculating the relative true-positive and -negative rates. Therefore, the two-phase design is the first choice as it provides unbiased estimates of test characteristics. The stratification design reduces the overall sample size and therefore this design could be considered as a third option in this situation.

5. Using routine care data

The use of routine data has become increasingly popular in all types of research including diagnostic accuracy studies. Some of the most known routine data sets include the Clinical Practice Research Database [31], The Health Improvement Network [32], and Interdisciplinary

Two-phase design	Comparing two tests without verification of double negatives	Comprehensive diagnostic study design
Third option	First option	Only option
First option	Second option	Only option
Higher risk of bias if the selection of reference standard is not random	Higher risk of bias if both index tests are not blinded from each other	Higher risk of bias if there is no reference panel to determine the diagnosis

Processing of Clinical Information [33] which are all based on routine medical records, mostly from primary care.

The perceived advantages of using routine data sets is that they contain adequate numbers of patients to allow the investigation of (very) rare diseases without prolonged phases of recruitment, they provide a wealth of information including demographic information, symptoms, diagnoses, and laboratory investigations, and they allow conducting studies with long-term follow-up [34–36].

Despite these advantages, routine data sets also have weaknesses, one of the most common problems being missing data [37]. Owing to the fact that the data are not collected with research purposes in mind, there are varying degrees of completeness across patients and across time, and data are unlikely to be missing at random [38]. A second issue is that the suspicion of disease is often not recorded and therefore the suspected population cannot be created. A third issue is that it is often unclear why individuals received particular medical investigations and it is rarely recorded for which target condition a test is used. Local availability of diagnostic testing or medical tradition may have an impact on diagnostic labeling. A fourth issue is that routine data sets also rely on routinely made diagnoses, which can introduce bias because it assumes that every patient eventually receives the correct diagnostic label, which is not necessarily true. There also may be large variation in diagnostic labeling of target conditions for which there is no clear definition. Finally, routine care data do not allow the evaluation of innovative or experimental tests

that are not part of routine clinical practice. For evaluating medical tests, these issues could lead to verification bias, information bias, and applicability concerns [37]. Assumptions and clinical judgment will be required to judge the impact of these issues and select appropriate methods to deal with them.

6. Discussion

We provide an overview of existing designs and methods, each with their advantages and limitations, which can help researchers in designing a diagnostic accuracy study in low-prevalence situations. Study designs can be distinguished by their ability to decrease the number of patients undergoing the index test or reference standard, and their risks of bias. Combinations of designs could be considered as well. For example, a two-phase design with stratification in the first phase could offer more savings compared with each design separately [9]. Another example is that a nested case-control design could be performed using data from the comprehensive diagnostic study design. Moreover, adaptive designs in which the prevalence of the disease is monitored during the study could be efficient as the prevalence of the disease can vary a lot and the sampling variation is large in low-prevalence studies [39,40].

Nonetheless, situations exist in which none of these designs or methods may be possible. For example, bacterial meningitis is now estimated to occur in 0.7–0.9 per 100,000 people per year in Western countries [41].

Considering bacterial meningitis is a rapidly evolving illness (nested), case-control designs would prove problematic because any symptom or biomarker is likely to have changed after the diagnosis has been established and stored samples are unlikely to be available for an acute, mostly pediatric population. The reference standard is lumbar puncture, which is invasive and preferably avoided in patients at low risk of meningitis. This leaves the options of the two-phase design, two index tests without verification of double negatives, and the stratification design. In order for the two-phase design to result in fairly precise estimates of specificity, the index test would have to be sufficiently accurate, in particular with few false-positives because otherwise the sample size will still be quite large. Similarly, for the design in which patients are not verified when testing negative on two index tests: considering bacterial meningitis is potentially lethal, the accuracy of both the index tests needs to be sufficient to risk not verifying patients who test negative on both. Finally, for the stratification design, finding variables with which the population can be stratified safely (e.g., fever and head/neck ache) could be challenging as there is little evidence on their distribution within the population precisely because of the very low prevalence in contemporary clinical practice.

A limitation is that we could have missed studies because it is a disease independent systematic review. However, by performing an extensive search in four libraries, backward citation tracking and consultation of experts we believe it is highly likely that all relevant studies were included. We did not include designs which provide biased estimates of test accuracy, for example, the enriched design in which only positive index tests receive a reference standard. In addition, the evaluation of the diagnostic accuracy of a test is only one step in the evaluation of diagnostic tests [42]. It is desirable to evaluate whether a test actually improves the care of patients. To provide suggestions for diagnostic impact studies in low-prevalence situations was beyond the aim of this study. Moreover, the included articles do not give details of the receiver operating characteristics (ROC) curve in relation to low-prevalence situations, while the ROC curve is particularly vulnerable to low events rates, as ROC estimation requires substantial sample sizes [43].

We recommend researchers explore their options before the start of any diagnostic accuracy study, by listing all the prerequisite criteria including expected prevalence in their setting, any evidence on subgroups within the population, the index test with its associated harms and costs, the reference standard again with harms and costs, and the target condition including its rapidity of disease progression. In addition, the goal of the test will influence which accuracy parameters will need to be estimated with the most precision [44]. This will guide researchers in identifying the options available and select the best option based on their respective risks of bias.

Sample size calculation of the classic cohort design is based on the following equations [5]:

Number with disease needed:

$$TP + FN = Z_{\alpha/2}^2 \frac{SN(1 - SN)}{W^2}$$

Sample size needed for sensitivity:

$$N_{sens} = \frac{TP + FN}{P}$$

Number without disease needed:

$$FP + TN = Z_{\alpha/2}^2 \frac{SP(1 - SP)}{W^2}$$

Sample size needed for specificity:

$$N_{spec} = \frac{FP + TN}{(1 - P)}$$

W = half width of the 95% CI

$Z_{\alpha/2}$ = for 2-tailed 95% CIs, $\alpha = 0.05$ and $Z_{\alpha/2} = 1.96$

P = prevalence of disease in the target population

SN = expected sensitivity of the index test

SP = expected specificity of the index test

CRediT authorship contribution statement

Gea A. Holtman: Writing - original draft, Writing - review & editing, Conceptualization, Methodology, Investigation, Data curation. **Marjolein Y. Berger:** Writing - original draft, Writing - review & editing, Conceptualization, Methodology, Data curation. **Huibert Burger:** Writing - original draft, Writing - review & editing, Conceptualization, Methodology, Data curation. **Jonathan J. Deeks:** Writing - original draft, Writing - review & editing, Data curation. **Norbert Donner-Banzhoff:** Writing - original draft, Writing - review & editing, Data curation. **Thomas R. Fanshawe:** Writing - original draft, Writing - review & editing, Data curation. **Constantinos Koshariis:** Writing - original draft, Writing - review & editing, Data curation. **Mariska M. Leeflang:** Writing - original draft, Writing - review & editing, Data curation. **Jason L. Oke:** Writing - original draft, Writing - review & editing, Data curation. **Rafael Perera:** Writing - original draft, Writing - review & editing, Data curation. **Johannes B. Reitsma:** Writing - original draft, Writing - review & editing, Data curation. **Ann Van den Bruel:** Writing - original draft, Writing - review & editing, Conceptualization, Methodology, Investigation, Data curation.

Acknowledgments

The authors are grateful to Gail Hayward and Ellie Morgan-Jones of the NIHR Community Healthcare Med-Tech and In Vitro Diagnostic Cooperative (MIC) for

facilitating the 1-day meeting. We thank Nia Robberts (Bodleian Health Care Libraries) for assisting with the literature search.

This work was supported by Ter Meulen Grant of the Royal Netherlands Academy of Arts and Sciences awarded to G.A.H. The funding source had no involvement in study design, collection or interpretation of data, writing report, and decision to submit article for publication.

T.R.F. received funding from the National Institute for Health Research (NIHR) Community Healthcare Medtech and In Vitro Diagnostics Cooperative (MIC). This paper presents independent research supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and University of Birmingham. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jclinepi.2019.05.018>.

References

- [1] Hamilton W, Peters TJ, Round A, Sharp D. What are the clinical features of lung cancer before the diagnosis is made? A population based case-control study. *Thorax* 2005;60:1059–65.
- [2] Van den Bruel A, Aertgeerts B, Bruyninckx R, Aerts M, Buntinx F. Signs and symptoms for diagnosis of serious infections in children: a prospective study in primary care. *Br J Gen Pract* 2007;57:538–46.
- [3] Holtman GA, Lisman-van Leeuwen Y, Kollen BJ, Norbruis OF, Escher JC, Kindermann A, et al. Diagnostic accuracy of fecal calprotectin for pediatric inflammatory bowel disease in primary care: a prospective cohort study. *Ann Fam Med* 2016;14:437–45.
- [4] Holtman GA, Lisman-van Leeuwen Y, Kollen BJ, Escher JC, Kindermann A, Rheenen PF, et al. Challenges in diagnostic accuracy studies in primary care: the fecal calprotectin example. *BMC Fam Pract* 2013;14:179.
- [5] Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Acad Emerg Med* 1996;3:895–900.
- [6] Buntinx F, Mant D, Van den Bruel A, Donner-Banzhof N, Dinant GJ. Dealing with low-incidence serious diseases in general practice. *Br J Gen Pract* 2011;61:43–6.
- [7] Wade R, Corbett M, Eastwood A. Quality assessment of comparative diagnostic accuracy studies: our experience using a modified version of the QUADAS-2 tool. *Res Synth Methods* 2013;4:280–6.
- [8] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529–36.
- [9] Obuchowski NA, Zhou XH. Prospective studies of diagnostic test accuracy when disease prevalence is low. *Biostatistics* 2002;3:477–92.
- [10] Duvekot J, van der Ende J, Verhulst FC, Greaves-Lord K. The screening accuracy of the parent and teacher-reported Social Responsiveness Scale (SRS): comparison with the 3Di and ADOS. *J Autism Dev Disord* 2015;45:1658–72.
- [11] Ukoumunne OC, Hyde C, Ozolins M, Zhelev Z, Errington S, Taylor RS, et al. A directly comparative two-gate case-control diagnostic accuracy study of the pure tone screen and HearCheck screener tests for identifying hearing impairment in school children. *BMJ Open* 2017;7:e017258.
- [12] Minnaard MC, Van De Pol, Alma C, De Groot JA, De Wit NJ, Hopstaken RM, et al. The added diagnostic value of five different C-reactive protein point-of-care test devices in detecting pneumonia in primary care: a nested case-control study. *Scand J Clin Lab Invest* 2015;75:291–5.
- [13] Armitage N, Hardcastle J, Amar S, Balfour T, Haynes J, James P. A comparison of an immunological faecal occult blood test Fecatwin sensitive/FECA EIA with Haemocult in population screening for colorectal cancer. *Br J Cancer* 1985;51:799.
- [14] Donner-Banzhoff N, Haasenritter J, Hullermeier E, Viniol A, Bosner S, Becker A. The comprehensive diagnostic study is suggested as a design to model the diagnostic process. *J Clin Epidemiol* 2014;67:124–32.
- [15] Bönsner S, Becker A, Haasenritter J, Abu Hani M, Keller H, Sönnichsen AC, et al. Chest pain in primary care: epidemiology and pre-work-up probabilities. *Eur J Gen Pract* 2009;15:141–6.
- [16] Colijn JM, Buitendijk GH, Prokofyeva E, Alves D, Cachulo ML, Khawaja AP, et al. Prevalence of age-related macular degeneration in Europe: the past and the future. *Ophthalmology* 2017;124:1753–63.
- [17] Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Advantages of the nested case-control design in diagnostic research. *BMC Med Res Methodol* 2008;8:48.
- [18] Rutjes AW, Reitsma JB, Vandembroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem* 2005;51:1335–41.
- [19] Kohn MA, Carpenter CR, Newman TB. Understanding the direction of bias in studies of diagnostic test accuracy. *Acad Emerg Med* 2013;20:1194–206.
- [20] Janes H, Pepe MS. Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. *Biometrics* 2008;64:1–9.
- [21] Mora M, Angelini C, Bignami F, Bodin A, Crimi M, Di Donato J, et al. The EuroBioBank Network: 10 years of hands-on experience of collaborative, transnational biobanking for rare diseases. *Eur J Hum Genet* 2015;23:1116.
- [22] Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, et al. The 100000 genomes project: Bringing whole genome sequencing to the NHS. *BMJ* 2018;361:k1687.
- [23] Dunn G, Pickles A, Tansella M, Vázquez-Barquero JL. Two-phase epidemiological surveys in psychiatric research. *Br J Psychiatry* 1999;174:95–100.
- [24] Irwig L, Glasziou PP, Berry G, Chock C, Mock P, Simpson JM. Efficient study designs to assess the accuracy of screening tests. *Am J Epidemiol* 1994;140:759–69.
- [25] Baker SG, Connor RJ, Kessler LG. The partial testing design: a less costly way to test equivalence for sensitivity and specificity. *Stat Med* 1998;17:2219–32.
- [26] Berry G, Smith CL, Macaskill P, Irwig L. Analytic methods for comparing two dichotomous screening or diagnostic tests applied to two populations of differing disease prevalence when individuals negative on both tests are unverified. *Stat Med* 2002;21:853–62.
- [27] Chock C, Irwig L, Berry G, Glasziou P. Comparing dichotomous screening tests when individuals negative on both tests are not verified. *J Clin Epidemiol* 1997;50:1211–7.
- [28] Schatzkin A, Connor RJ, Taylor PR, Bunnag B. Comparing new and old screening tests when a reference procedure cannot be performed on all screenees: example of automated cytometry for early detection of cervical cancer. *Am J Epidemiol* 1987;125:672–8.
- [29] Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *J Clin Epidemiol* 2003;56:1118–28.
- [30] Donner-Banzhoff N, Kunz R, Rosser W. Studies of symptoms in primary care. *Fam Pract* 2001;18:33–8.
- [31] Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staaij T, et al. Data resource profile: clinical practice research datalink (CPRD). *Int J Epidemiol* 2015;44:827–36.

- [32] Anonymous. In Practice Systems Ltd. The Health Improvement Network (THIN) 2016. <http://www.inps.co.uk/vision/health-improvement-network-thin>. Accessed July 5, 2016.
- [33] Mosseveld B. Interdisciplinary processing of clinical information. 2015. Available at <http://www.ipci.nl/Framework/Framework.php>. Accessed February 20, 2018.
- [34] Miller A, Nightingale AL, Sammon CJ, Mahtani KR, Holt TA, McHugh NJ, et al. Estimating the diagnostic accuracy of rheumatoid factor in UK primary care: a study using the Clinical Practice Research Datalink. *Rheumatology* 2015;54:1882–9.
- [35] Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2012;62:e29–37.
- [36] Rees F, Doherty M, Lanyon P, Davenport G, Riley RD, Zhang W, et al. Early clinical features in systemic lupus erythematosus: can they be used to achieve earlier diagnosis? A risk prediction model. *Arthritis Care Res (Hoboken)* 2017;69:833–41.
- [37] Oostenbrink R, Moons KG, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. *J Clin Epidemiol* 2003;56:501–6.
- [38] Sammon CJ, Miller A, Mahtani KR, Holt TA, McHugh NJ, Luqmani RA, et al. Missing laboratory test data in electronic general practice records: analysis of rheumatoid factor recording in the clinical practice research datalink. *Pharmacoepidemiol Drug Saf* 2015; 24:504–9.
- [39] Yi Q, Panzarella T, Corey P. Incorporating the sampling variation of the disease prevalence when calculating the sample size in a study to determine the diagnostic accuracy of a test. *Control Clin Trials* 2004; 25:417–27.
- [40] Brinton JT, Ringham BM, Glueck DH. An internal pilot design for prospective cancer screening trials with unknown disease prevalence. *Trials* 2015;16:458.
- [41] Brouwer MC, van de Beek D. Epidemiology of community-acquired bacterial meningitis. *Curr Opin Infect Dis* 2018;31:78–84.
- [42] Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape test evaluation. *Clin Chim Acta* 2014;427: 49–57.
- [43] Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER. Small-sample precision of ROC-related estimates. *Bioinformatics* 2010;26:822–30.
- [44] Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006; 332:1089–92.