Challenges in planning and conducting diagnostic studies with molecular biomarkers

Herausforderungen an die Planung und Durchführung von Diagnosestudien mit molekularen Biomarkern

Introduction

Biomarkers are currently a topic of intense research in science and medicine. However, biological markers have been used for thousands of years. Already in the ancient world, physicians used biomarkers for diagnosis. Probably the most well-known example is the examination of urine, which was described by Galen in the second century. In the Middle Ages, uroscopy was considered an almost infallible diagnostic tool for almost all diseases. Some of these indicators are still in use, such as glucose in the urine as evidence for diabetes mellitus. At the same time, many more characteristics are used as biomarkers by now, so that the development of very general definitions was necessary [1, 2].

Definition

According to Gallo et al. [2], the most common definition for a biomarker is as follows: "... a biomarker is any substance or biological structure that can be measured in the human body and may influence, explain, or predict the incidence or outcome of disease". However, it may be questioned whether the restriction of measurement in the human body is reasonable. An alternative is to define a biomarker as "any substance, structure or process that can be measured in biospecimens and may be associated with health-related outcomes" [2]. In our opinion, this definition is too general, and the definition should include a specific association with health or a clinical outcome [1]. We therefore prefer the definition by the Biomarkers Definitions Working Group [3] of the National Institutes of Health: "A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacologic responses to a therapeutic intervention." Molecular biomarkers in particular are biomarkers that can be detected using molecular technologies such as genomics or proteomics, or imaging techniques; for a comprehensive definition, see Ziegler et al. [1].

Prognostic, diagnostic, and predictive biomarkers

Biomarkers are currently applied in all patient-relevant areas, in diagnosis, prognosis, and therapy. While prognostic biomarkers predict the patients’ disease course, diagnostic biomarkers allow determining the disease. By definition, predictive biomarkers are linked to the treatment of a patient. For example, predictive biomarkers estimate the probability for the success of a specific therapy, or the probability for a specific severe adverse event of an intervention. They thus offer guidance in the selection of the best therapy for a patient.
The ACCE model

Which biomarker is a good biomarker? According to the ACCE model that was developed by the Center of Disease Control (CDC), a biomarker is evaluated using four criteria [5]:
1. Analytical validity,
2. Clinical validity,
3. Clinical utility, and
4. Ethical, legal, and social implications.

Analytical validity indicates the technical reliability of a biomarker. Here, we differentiate between accuracy, trueness, and precision according to the German norm [6]. Trueness measures whether the mean of a large series of experiments is close to the theoretically expected value, and it is therefore sometimes termed accuracy of the mean. In contrast, precision considers the variability of the measurements. Finally, accuracy is the combination of both.

Clinical validity specifies the value of the biomarker in detecting or predicting a disease. In practice, it is impossible to define a general threshold for accepting a biomarker to be clinically valid. Instead, this depends on whether alternative prognostic models are available, on the aim of biomarker testing, and on the burden of the specific disease. In general, it is difficult to justify using a diagnostic biomarker without adequate therapeutic options, regardless of its value.

Evaluating the clinical utility of a biomarker will be described in detail in the following section. The final criterion of the ACCE model considers ethical, legal and social implications that may arise in the context of a biomarker. These are detailed in the literature (e.g., [7]).

Levels of evidence and phases of diagnostic studies

To justify the use of a biomarker in practice, its clinical utility must have been shown, and this requires high analytical validity as well as high clinical validity. For this, the evaluation of the clinical validity critically depends on the quality of the clinical trials.

For diagnostic studies, the Federal Joint Committee in Germany (Gemeinsamer Bundesausschuss) has laid down the levels of evidence shown in Table 1 for its code of procedure. These levels of evidence are equally valid for screening methods. The highest level of evidence is assigned to randomized therapeutic trials and meta-analyses of randomized therapeutic trials. This raises the question how biomarkers for diagnostic purposes can be evaluated in the context of randomized therapeutic trials. As described above, the result of a diagnostic biomarker should have an effect on the subsequent intervention, i.e., the biomarker should be predictive. Accordingly, the best clinical utility is obtained for diagnostic biomarkers with proven value for application in one or more randomized therapeutic studies. Through this, a diagnostic biomarker becomes a predictive biomarker.

For purely diagnostic biomarkers with level II or lower, utility has not been investigated in the context of interventions. As a result, they have a lower level of evidence. However, the study design and the methodological quality of diagnostic studies affects the level of evidence critically. And the utilized study designs in turn depend on the phase of the biomarker study.

In general, we can distinguish four phases for diagnostic and prognostic biomarker studies (Table 2) [1, 8–12]. Phase I comprises preliminary technical and methodological studies. In these, it is investigated whether the biomarker is, in principle, suitable as a diagnostic biomarker. The typical study design for a phase I study is a case-control study with patients and healthy controls, often with extremely affected patients and unequivocally healthy or even hypernormal controls.

This phase can even be subdivided into smaller steps, because modern technologies allow for the measurement not only of a single biomarker but of possibly several million biomarkers simultaneously (phase Ia). Out of this multitude of measurements, the correct biomarker or set of biomarkers has to be selected. For the single molecule, these high-throughput technologies are typically not as accurate as other technologies that are specifically tailored to a single biomarker. Thus, this phase usually also includes the analytical validity of the biomarker, e.g., the development of a specific assay (phase Ib).

---

Table 1  Levels of evidence for diagnostic methods according to §11, para. 2, of the code of procedure of the Federal Joint Committee (Gemeinsamer Bundesausschuss), lastly modified on 2012 [65].

<table>
<thead>
<tr>
<th>Level of evidence</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>I a</td>
<td>Systematic reviews of trials with evidence level I b</td>
</tr>
<tr>
<td>I b</td>
<td>Randomized controlled trials</td>
</tr>
<tr>
<td>I c</td>
<td>Other intervention studies</td>
</tr>
<tr>
<td>II a</td>
<td>Systematic reviews of diagnostic accuracy studies with evidence level II b</td>
</tr>
<tr>
<td>II b</td>
<td>Cross sectional and cohort studies allowing to calculate all diagnostic accuracy statistics (sensitivity and specificity, likelihood ratios, positive and negative predictive value)</td>
</tr>
<tr>
<td>III</td>
<td>Other studies allowing to calculate diagnostic accuracy statistics (sensitivity and specificity, likelihood ratios)</td>
</tr>
<tr>
<td>IV</td>
<td>Observations of associations, pathophysiological considerations, descriptive reports, case reports, and the like; expert opinions with no explicit critical appraisal, reports from expert committees and consensus conferences</td>
</tr>
</tbody>
</table>

---
The sample sizes for these initial studies are generally small, because the high-throughput technologies are often rather expensive. Therefore, samples from biobanks are usually utilized to evaluate the diagnostic value of a novel biomarker (set) before conducting more cost-intensive larger validation studies (phase Ic). A combination of several biomarkers might prove to be more promising than a single biomarker measurement. The optimal combination of several biomarkers in a multimarker rule is often evaluated within the same phase using elaborate biostatistical approaches, such as machine learning algorithms.

In phase II, the validity is evaluated retrospectively in selected probands. This aims at answering the question whether the test fulfills its purpose by, e.g., detecting the disease.

Phase III is reserved for controlled diagnostic studies investigating the accuracy of the test in clinical practice. The typical study design for this phase is a cohort study in symptomatic patients. Finally, the aim of phase IV is to show efficacy of the biomarker. Thus, the first step of this phase evaluates how the test influences the clinical course. After this, cost-benefit analyses are performed. Our own practical experience has shown that a finer grading of the four phases is helpful, especially concerning the early phase I [1].

---

### Basic methodological principles for validation studies of diagnostic biomarkers

The critical feature of validation studies for biomarkers in phase III is the prospective and consecutive recruitment of symptomatic patients who represent the usual patient spectrum. Validation studies in phase IV are randomized therapeutic trials. Here, very specific study designs are often used [1, 13–19], but the most important methodological element of these studies is the randomization. However, high methodological quality and, through this, validity of a diagnostic studies is not guaranteed by randomization or prospective recruitment alone. Indeed, a number of basic methodological principles need to be considered for both study types. Specifically, for phase IV validation studies, the same basic principles as for all therapeutic studies apply [20].

The basic principles of diagnostic studies for phases I to III are compiled in Table 3 [8, 21, 22]. If these principles are not adhered to, estimates of sensitivity and specificity, i.e., of the clinical validity of the biomarker, can be substantially biased.

### Important sources of bias in validation studies of diagnostic biomarkers

There are many possibilities for errors in study designs that can lead to biases in diagnostic studies. If we consider, as in a recipe, the different ingredients of a diagnostic study, we firstly require...
study probands. Thus, one of the most important disturbance source is the bias that can arise by selective recruitment or willingness to participate.

Next, we require an index test, which is the selected biomarker that is being tested. For this, random but also systematic errors can occur in the laboratory. In addition, a reference standard is needed for comparison with the novel test. Again, there are a number of sources of error, and the relevant problems will be described below. The interplay of the index test and the reference test is illustrated in the proof of citrullinated peptides (ACP) for rheumatoid arthritis [23]. The index test for ACPA is an ELISA of the third generation, the reference standard usually is the classification based on the criteria of the American College of Rheumatology.

When evaluating diagnostic accuracy, it needs to be considered that there might be an interaction between the index test and the reference standard. For example, the two tests might have been applied at time points that are different enough to allow for a change in the true condition. Also, knowing the result of one test might influence in some way the proceeding for the other test or the result, even if the proceeding remains the same. Another possibility for errors lies in the rating of the test itself.

After the study has been conducted, the data are analyzed. Here, one question is how to handle missing data or results that cannot be interpreted. Finally, the diagnostic study needs to be published comprehensively, and errors can be avoided by publishing according to the STARD (Standards for Reporting of Diagnostic Accuracy) statement [24].

Table 4 summarizes the sources of bias described so far, and a more detailed discussion can be found in the literature [21, 25–29].

### Biases on the subject level

Since a non-representative selection is the most important source of bias, the first entry in Table 4 is the spectrum composition bias. The possibilities for a selection bias are manifold [28, 30, 31], and Box 1 gives a more elaborate list of recruitment problems that can lead to this bias; the first three entries given there are summarized under the term spectrum composition bias.

### Selection bias

A number of systematic reviews have shown that selection bias is the source of bias with the most detrimental effect on the estimation of sensitivity and specificity in validation studies of biomarkers [25–27, 32]. Figure 1 summarizes the findings of the systematic reviews, and it specifically shows how the estimates of accuracy change whenever a specific methodological criterion for the study design is not fulfilled, as compared to the situation in which it is fulfilled. Accuracy is here expressed by the odds ratio, and a relative diagnostic odds ratio greater than 1 means that studies not fulfilling the criterion yield higher accuracy estimates [27].

If extremely affected cases are compared with healthy controls, the diagnostic accuracy is overestimated by about 5-fold on average. A classical example for this is the diagnosis of colorectal cancer based on carcinoembryonic antigen (CEA) [33]. In a case-control study including only patients with known advanced colorectal or rectal cancer, the CEA was found to be increased in 35 of the 36 patients, while it was considerably lower in healthy controls [34]. In subsequent studies, patients were included in earlier stages of colorectal cancer, and the accuracy of the CEA

---

**Table 4** Important sources of bias in validation studies for diagnostic biomarkers. Adapted from ref. [21].

<table>
<thead>
<tr>
<th>Bias</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrum composition bias</td>
<td>Spectrum of patients not representative of the patients who will receive the test in practice</td>
</tr>
<tr>
<td>Partial verification bias</td>
<td>Reference standard is applied only to a selected part of the patients, not all</td>
</tr>
<tr>
<td>Differential verification bias</td>
<td>Use of different reference standards, use depending on test result</td>
</tr>
<tr>
<td>Disease progression bias</td>
<td>Time period between reference standard and index test so long that target condition might have changed</td>
</tr>
<tr>
<td>Incorporation bias</td>
<td>Reference standard and index test not independent; special case: Index test part of reference standard</td>
</tr>
<tr>
<td>Test review bias</td>
<td>Index test results interpreted with knowledge of results of reference standard</td>
</tr>
<tr>
<td>Reference standard review bias</td>
<td>Reference standard results interpreted with knowledge of results of index test</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Index test interpreted in the light of clinical data that would not be available when test used in practice</td>
</tr>
</tbody>
</table>

---

**Box 1 Reasons for selection bias.**

- only cases with a restricted disease spectrum are enrolled, e.g., more severely affected cases (selection for symptoms; severe cases);
- enrollment of patients differs between the study and clinical practice; e.g., the patient spectrum differs between an emergency unit and a day hospital;
- individuals are enrolled depending on the result of the index test (referral for index test bias); this bias is not identical to verification bias; for verification bias, the reference standard is not applied in all probands of the study;
- healthy probands do not appear for the follow-up so that their data are missing (loss to follow-up bias);
- only a restricted spectrum of probands participates in the study; e.g. only patients with a confirmed diagnosis (participation bias, also self-selection bias);
- only individuals with specific previous examinations are included (limited challenge bias);
- only individuals with specific previous diagnoses are included (increased challenge bias);
- only individuals are included who are „suitable“ and „can endure“ the trial (study examination bias).
test decreased decidedly (e.g., [35]). Consequently, the CEA test was banned from clinical routine for diagnosis as well as for screening [33].

Impressive differences in the estimation of accuracy were demonstrated by Lachs et al. [36] in their investigation of the leukocyte esterase and bacterial nitrite rapid dipstick test for urinary tract infection (UTI). The study consecutively recruited 366 adult patients. Of these, 107 patients had a high prior probability for UTI based on clinical signs and symptoms. In these, the sensitivity of the dipstick test was 92\% at a specificity of 42\%. In contrast, in the remaining 259 patients with lower prior probability, sensitivity and specificity were 56\% and 78\%, respectively. Thus, the composition of the patient group led to completely different characteristics of the biomarker test.

The bias is less extreme in case-control study designs including not extremely affected patients. Still, case-control studies may give very optimistic estimates with an overestimation of the diagnostical accuracy in the order of three.

There are many further examples in the literature showing how the selection of the study sample can yield to biased results, even in the case of prospective studies instead of case-control studies. For instance, a prospective study examined fetuses at a gestational age between 11 and 14 weeks by ultrasound to evaluate the diagnostic value of a missing nasal bone as an indicator for a chromosomal anomaly [37]. The sensitivity was estimated to be 69\% (Figure 2a). However, the analysis was restricted to those fetuses with trisomy 21, which was also indicated in the title of the study. The reference standard in the study was chorionic villus sampling [37], and this as well as amnioncentesis can also detect other forms of chromosomal anomalies. Through this, 295 fetuses were excluded from analysis, of whom 124 had trisomy 18. In these, the sensitivity was only 32\%. Since the sensitivity in all fetuses was at 52\% (Figure 2b), the test for detecting chromosomal anomalies is about as good as a coin toss.

Consecutive subject recruitment

Recruiting subjects non-consecutively does not necessarily lead to bias (Figure 1). If the subjects are recruited selectively, this non-consecutive recruitment is likely to result in bias. In other scenarios, this may not be the case. However, it is important that bias cannot be excluded using a non-consecutive recruitment, and the bias can be large. Surprisingly, bias can occur even if subjects are selected at random (Figure 1), because this is not a consecutive series of probands.

Retrospective versus prospective study

As opposed to clinical trials, the terms „prospective“ and „retrospective“ are not defined coherently for biomarker studies. In general, a prospective study implies that all investigations are planned before they are conducted. This especially means that subjects are recruited only after the study has been designed. In contrast to that, many studies use data from already completed randomized controlled trials. Then, hypotheses on specific biomarkers are formulated prospectively and tested prospectively in available biospecimen. This approach has the advantage that the material can usually be attained according to stringent quality criteria and a strict protocol. But the biobank might be very restricted regarding the time point at which the material is collected as well as the specific kind of available material. Thus, the biobank might be inadequate for studies on proteins or metabolites. Since their concentrations can change very fast, a specific time point of probe extraction and a defined measurement point may be relevant; a more detailed discussion is given in the literature [17, 38]. Therefore, the biospecimen has to be available at a specific point in time, which is not a problem in prospective recruitment. However, in other studies the time point of measurement is irrelevant, for instance in the determination of genetic markers, i.e. DNA markers, because they are constant over the life span.

In contrast to prospective studies, in retrospective studies subjects have already been recruited, and the measurement of biomarkers has already been performed. This approach is often followed in genetic studies. Here, data of several so-called genome-wide association studies are often utilized to confirm novel biomarker findings. This form of retrospective study is usually termed “in-silico replication.”

As a general rule, prospective studies lead to less bias than retrospective studies. However, there are exceptions [38], and there are some scenarios in which a retrospective study can be better than a prospective study. Obviously, retrospective studies are less time and cost intensive than prospective studies. On the other hand, a prospective study typically has a higher validity, because a standardized study conduct allows for a consistent quality control of all data. However, this can also be the case for a well-maintained biobank.
On the other hand, bias can occur also in prospective studies. For instance, observational and treatment equivalence is violated if the treating or otherwise involved clinicians know the test result, and if knowledge of the test result affects their behavior. In retrospective studies, it has to be questioned whether the data are complete, assessed in a standardized manner and at correct time points, and whether selection bias might have occurred.

In conclusion, prospective studies are almost always superior to retrospective studies. This has been confirmed in the systematic review by Rutjes et al. [27] who showed that retrospective studies overestimate the diagnostic accuracy of a test by 60%.

Selection based on the index test
If recruitment for a study deliberately depends on the result in the index test, a surprising trend for bias arises, which has been described in detail by Knotterus und Muris [38]. For example, one may preferentially include subjects with clear symptoms, or those with results from unreliable tests, or those with conflicting test results, or even those with a positive result in the index test. It is often difficult to differentiate patients with clear symptoms but possibly another differential diagnosis, so that the accuracy of the evaluation declines. Accordingly, this approach tends to yield an underestimation of accuracy (Figure 1) [27].

Biases on the test level
In many applications, it is extremely challenging to decide which procedure to use as reference standard. Obviously, the reference standard should be (nearly) perfect. However, even experienced pathologists or radiologists are not infallible. Moreover, a reference standard is not always available for some diagnostic problems as for epilepsy, or application of the reference standard was applied in only 25% of the diagnosed subjects. Thus, partial verification bias led to a drastic overestimation of the study results (Figure 1). It often occurs if the reference standard is based on an invasive procedure, such as surgery. Then, this invasive diagnostics is only applied in test positive individuals, and another reference standard, such as a clinical monitoring, is applied in subjects with negative test results.

An example for this is the study on lung ventilation/perfusion scintigraphy for the diagnosis of lung embolism [39]. Here, the reference standard is a radiological inspection of the lung artery, which is preferentially applied after a positive result from scintigraphy. Patients with a negative result are preferentially monitored only.

Similar to differential verification bias, partial verification bias can lead to substantial bias, although this phenomenon is not as widespread in practice. Detailed descriptions are given in the literature [40, 41].

Whereas real examples are given in references [40–44], Figure 3 includes a fictitious example. Assume that the reference standard is applied in only 25% subjects with a negative result in the index test, but in all subjects with a positive result in the index test. Furthermore, assume that we observe the frequencies that are shown in Figure 3 in the study. To calculate sensitivity, only subjects with positive reference standard are required. Using the observed frequencies, the sensitivity equals 80% (= 80/[80 + 20]). At the same time, the specificity requires only subjects with negative reference standard, and also equals 80% (= 40/[40 + 10]).

However, the observed frequencies need to be corrected for the calculation of sensitivity and specificity, since the reference standard was applied in only 25% of the biomarker negative, i.e. index test negative, individuals. Results from a simple extrapolation are given in Figure 3b. For this, the frequencies of test negatives from Figure 3a are increased by a factor of 3. Using these corrected numbers, sensitivity decreases to 50% (= 80/[80 + 80]) and specificity increases to 94.11% (= 160/[160 + 10]). Thus, partial verification bias led to a drastic overestimation of the sensitivity but a clear underestimation of the specificity in this fictitious study. Through this, the proportion of correctly diagnosed subjects decreases from 80% (= [80 + 40]/[80 + 20 + 40 + 20]) to 72.72% (= [80 + 160]/[80 + 10 + 80 + 160]). To correct for partial verification bias, two well-known statistical approaches have been suggested in the literature, namely the Begg-Greenes method [45] and the Diamond method [46].

Analogously, using different reference standards, i.e., differential verification bias, leads to an overestimation of diagnostic accuracy by 60% compared with studies using only one reference standard.
**Blinding**

On the test level, the most obvious error source is the lack of blinding. With no blinding, an overestimation of the diagnostic accuracy is likely, and this bias is termed review bias. Naturally, blinding is more important when using soft outcome criteria such as clinical symptoms, compared with hard endpoints such as biomarker measurements in the laboratory – although these can often be blinded easily using an adequate coding of the samples. The price to be paid for lack of blinding is an overestimation of diagnostic accuracy of about 30% on average [27].

**Gold standard versus reference standard**

The previous sections always used the term reference standard which is also used in some of the guidelines, such as the STARD Statement [24]. In common speech as well as in the literature, the term gold standard is also used. Here, the gold standard indicates the true disease state of an individual. In contrast, the reference is the best available method to indicate the disease state of the individual [47]. It is important to note that gold standard and reference standard may differ. A specific issue is that the reference standard may be flawed but might coincide better with the index test, i.e., the biomarker test. In any case, estimates of sensitivity and specificity may differ if gold standard and reference standard are not identical.

Although the reference standard should be as perfect as possible, it is often difficult to choose a sensible reference standard in practice. Imaging studies often use surgery, pathological findings or the clinical follow-up as standard [21]. The example in reference [48] shows that the validity of the entire study is questioned if a reference standard is applied that does not conform to the usual standard [49]. Specifically, Dehdashti et al. [48] used a barium meal as reference standard for the diagnosis of gastroesophageal reflux disease, although this is not supported by the North American Society for Pediatric Gastroenterology, Hepatology and Nutrition. Instead, the current reference standard is monitoring the pH value in the esophagus. In combination with further methodological issues, the author of a letter to the editor [49] stated that “... this study has several critical methodological flaws, ...”.

**Inclusion bias: Reference standard and index test are not independent**

In some cases, the index test is part of the reference standard. Thus, the two tests are not independent from each other. The most prominent example for this has been given by Guyatt et al. [33]. In a study on screening instruments for depression in patients with terminal disease, 100% sensitivity and 100% specificity were observed. Here, the index test consisted of 9 questions and included the question: „Are you depressed?”

A second example is the study by Harvey [50] who investigated 107 patients with thyrotoxicosis. The final diagnosis was based on all available information, including the results from a thyroid function test. It was concluded that clinical disease severity was associated more strongly with concentration of free thyroxine than with any other considered index. However, the concentration of free thyroxine had been used for the primary diagnosis [51]. Therefore, all patients in the study naturally had concentrations of free thyroxine outside of the reference interval (Figure 4).

**Bias on the level of evaluating the test results**

**Missing values**

In many molecular tests, the result is not unambiguous for every proband, i.e., it is unclear, uncertain or not determined. However, values cannot simply be excluded, if not everyone can be classified as test positive or test negative, because the frequencies of the different categories are an important indicator for the usefulness of the test.

If the results are merely excluded, the estimators for sensitivity and specificity can be biased. This has been investigated methodologically in the literature, e.g. in References [52–54]. Figure 5 illustrates this phenomenon based on the previously published data by Ramos et al. [55]. Here, the value of the interferon gamma release assay (IGRA) for the diagnosis of tuberculosis was examined. In our eyes, the data were reported completely and analyzed correctly. The complete data are shown in Figure 5a. If indeterminate and invalid test results are assigned to the least favorable category (Figure 5b), sensitivity is estimated by $\frac{27}{71} = 38.00\%$, and specificity is $\frac{238}{302} = 78.81\%$. In contrast, if the missing values are ignored (Figure 5c), the estimates for sensitivity and specificity are $\frac{27}{67} = 40.30\%$ and $\frac{238}{280} = 85.00\%$, respectively. This exemplifies substantial differences in the estimates. It should be noted that in this example, only about 7% of the values are missing, whereas values of up to 40% are found in the literature [56].

**Figure 4.** Number of patients depending on free thyroxine (ng per 100 ml) for 105 patients with thyrotoxicosis according to [50]. The dotted line gives the upper reference limit of the test.

**Figure 5.** Missing data in the example by Ramos et al. [55] on the interferon gamma release assay to diagnose tuberculosis. a) complete data, b) assigning indeterminate and invalid test results to the less favorable category, c) ignoring missing values.
Two studies systematically investigated the effect of excluding test results that are not interpretable [32]. However, neither study indicates the direction and size of the possible bias [57, 58].

We finally remark that ambiguous test results might have their own diagnostic value or might hint at another disease [59].

**Post hoc definition of the threshold**

Instead of a positive or negative test result, such as mutation present or absent, many molecular biomarkers yield a quantitative test result. From this, a suspicious or unsuspicious result is defined using a threshold that is usually determined by reference or norm values. If the threshold is determined using the data of the current study, it is mostly defined to somehow optimize sensitivity and/or specificity. This generally results in an overestimation of the diagnostic accuracy of about 30% (Figure 1). Thus, it is crucial to define the threshold or, if multiple biomarkers are used, the multi-marker rule prior to the study.

**Coefficient of variation and sample size estimation**

In the planning stage of a study, a central question concerns the sample size that is required to detect differences in the biomarker means between two groups. Obviously, these depend on the precision of the biomarker measurements as well as on the difference between the groups. Specifically, the precision is expressed by the coefficient of variation (CV) \( \nu = \sigma / \mu \) which is the relative variation of the biomarker measurements. Precise tests have a CV of less than 10% = 0.1. In this context, the difference is expressed by the fold change \( f = \mu_2 / \mu_1 \) or 100 \* \( f \). This is interpreted as the factor or the percent by which the mean biomarker values in group 2 differ from the mean biomarker values in group 1. For instance, if the mean values of the biomarker in group 2 are doubled compared with the mean values in group 1, the fold change equals 2.

Assuming that the two groups are of equal size, the required sample size to detect a difference between the groups with a power of 90% at the usual significance level of 5% can be approximated by

\[
\text{n} = \frac{20 \times (1 - f)}{f^2} \quad \text{Equation (1)}
\]

per group. The appendix gives the derivation of this formula from the standard sample size formula for mean differences.

Equation (1) shows that the sample sizes increase quadratically with the CV. This means that a fourfold sample size is required if the CV is doubled due to imprecise measurements. This quadratic relationship between sample size and CV is depicted in Figure 6 for difference fold change values.

This relationship can be utilized in a number of applications in the laboratory, which is exemplified in the following.

For an experiment with gene expression chips, Figure 7 illustrates that the CV decreases considerably with increasing expression strength. For example, for expressions above the detection threshold (4.5), the CV approximates 4.5%. In contrast, for strongly expressed transcripts (normalized expression of 11 and greater), the CV is smaller by a factor of 2 to 4. Hence, if there are two transcripts with different expression levels for a validation expression threshold (4.5), the CV approximates 4.5%. In contrast, for strongly expressed transcripts (normalized expression of 11 and greater), the CV is smaller by a factor of 2 to 4. Hence, if there are two transcripts with different expression levels for a validation.

Another example is the dependence of the precision of high performance liquid chromatography (HPLC) on the time constant (Figure 8). At a time constant of 0.5 sec, the CV approximates 20%. At 2 sec, though, the CV decreases to less than 5%. In this example, the difference in the variability amounts to a factor of about 4, meaning that for a measurement at 0.5 sec, about 16 times as many probands would have to be excluded in the study compared with a measurement at 2 sec. It is certainly essential to understand the measurement technology in detail. It might well be possible that measuring at 2 sec in HPLC leads to the measurement of different analytes so that the value of the target analyte is biased. If the signals are very weak, a long measurement period also accumulates more noise than a short
measurement period. Therefore, the signal-to-noise ratio depends strongly on the measurement period, especially in weak signals.

Discussion ▼

A publication in the Journal of the American Medical Association in 1995 already stressed the importance of research on diagnostic methods [60]. In their review, the authors considered 112 publications on diagnostic tests that had been published in four important medical journals between 1978 and 1993. In total, 80% of the publications were methodologically flawed leading to relevant biases in the results [61].

More recently, in 2009, similarly sobering conclusions were drawn by Fontela et al. [25] in a study on the quality of molecular diagnostic studies for tuberculosis, HIV, and malaria. They identified 90 articles that used a commercial test kit and fulfilled their inclusion criteria. None of these publications was flawless. For instance, only 10% of the articles adequately described the reference standard, and only 16% of the studies reported a blinded follow-up.

Moreover, Fontela et al. [25] confirmed previous findings by concluding that the reporting quality of diagnostic studies was low [26, 59, 62]. However, deficits in the reporting quality could easily be avoided by completely adhering to the STARD recommendation [24], although the adoption of this STARD standard by researchers could be accelerated [63, 64].

Considering the fundamental principles for diagnostic studies in planning, performing and analyzing a study as well as subsequent publishing according to usual recommendations, such as STARD, has the potential to considerably improve the current situation.

Appendix ▼

The starting point is the following standard formula to calculate the required sample size per group if two equally sized groups are being compared:

$$n = 2 \times \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \times \sigma^2}{(\mu_2 - \mu_1)^2} \quad \text{Equation (2)}$$

Here, $\alpha$ is the significance level which is usually set to 0.05, and $1 - \beta$ is the statistical power, usually 80%, 90% or 95%. $\mu_1$ and $\mu_2$ denote the average biomarker values in the two groups, and $s$ is the variation of the biomarker measurement in a single proband. For simplicity, we set $1 - \beta = 0.9$ so that using the values of the normal distribution we obtain $(z_{1-0.025} + z_{1-0.9})^2 = 10.5074 \approx 10$. Accordingly, Equation (2) can be simplified to

$$n = 20 \times \frac{\sigma^2}{(\mu_2 - \mu_1)^2} \quad \text{Equation (3)}$$

This is the sample size that is required per group to detect a difference in average biomarker values in both groups at a significance level of 5% with a power of about 90%.

Instead of using the difference in mean values $\mu_1 - \mu_2$ and the standard deviation $\sigma$, in the context of biomarkers the effect is preferably described by the fold change $f = \mu_2/\mu_1$ and the coefficient of variation (CV) $v = \sigma/\mu$. Using these parameters, Equation (3) can be re-formulated by

$$n = 20 \times \frac{v^2}{(1 - f)^2} \quad \text{Equation (4)}$$

Acknowledgments

This work has its source in presentations by the authors at the „Functional Genomics and Proteomics – Applications, Molecular Diagnostics & Next Generation Sequencing“ meeting in Frankfurt, Germany, in February Februar 2012. The authors are grateful to the Project Management Agency of the German Aerospace Center, Health Research, for the invitation to the respective session with the working title „Validierung in diagnostischen Studien“.

The work of AZ on biomarkers is financially supported by the Federal Ministry of Education and Research (01GI0916, 0315536F, 01ER0805, 01KU9098A, 01KU9098B, Cluster of Excellence Inflammation at Interfaces) and the European Union (HEALTH-2011-278913).

The authors thank Dr. Stavros Kromidas for the kind permission to use the example on high performance liquid chromatography (HPLC).

Conflict of interest: AZ is scientific advisor to Protagen AG, Dortmund, Germany, has been scientific advisor to IntegraGen SA, Evry, France, until January 14 2011, and has a cooperation contract with Affymetrix Inc., Santa Clara, USA. PSK is CSO at Protagen AG, Dortmund, Germany.

German version of this article: DOI 10.1055/s-0032-1327406
Abstract

Challenges in planning and conducting diagnostic studies with molecular biomarkers

Biomarkers are of increasing importance for personalized medicine in many areas of application, such as diagnosis, prognosis, or the selection of targeted therapies. In many molecular biomarker studies, intensity values are obtained from large scale –omics experiments. These intensity values, such as protein concentrations, are often compared between at least two groups of subjects to determine the diagnostic ability of the molecular biomarker. Various prospective or retrospective study designs are available for molecular biomarker studies, and the biomarker used may be univariate or even consist in a multimarker rule. In this work, several challenges are discussed for the planning and conduct of biomarker studies. The phases of diagnostic biomarker studies are closely related to levels of evidence in diagnosis, and they are therefore discussed upfront. Different study designs for molecular biomarker studies are discussed, and they primarily differ in the way subjects are selected. Using two systematic reviews from the literature, common sources of bias of molecular diagnostic studies are illustrated. The extreme selection of patients and controls and verification bias are specifically discussed. The pre-analytical and technical variability of biomarker measurements is usually expressed in terms of the coefficient of variation, and the effect is that the required sample size for biomarker validation quadratically increases with the coefficient of variation, and the effect is illustrated using real data from different laboratory technologies.

References

9 Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York, Oxford University Press 2003
22 Obuchowski NA. How many observers are needed in clinical studies of medical imaging? AJR Am J Roentgenol 2004; 182: 867–869
28 Sica GT. Bias in research studies. Radiology 2006; 238: 780–789
30 Blackmore CC. The challenge of clinical radiology research. AJR Am J Roentgenol 2001; 176: 327–331
35 Zielinski C. Aussagekraft der radioimmunologischen Testung. Dtsch Med Wochenschr 1995; 120: 893

Dtsch Med Wochenschr 2013; 138: e14–e24 · A. Ziegler et al., Challenges in planning ...
44 Richardson ML, Pettsavage JM. Verification bias: an under-recognized source of error in assessing the efficacy of MRI of the menisci. Acad Radiol 2011; 18: 1376–1381
45 Begg CB, Greens RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39: 207–215
57 Philbrick JT, Horwitz RI, Feinstein AR et al. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. JAMA 1982; 248: 2467–2470
60 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995; 274: 645–651
63 Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication – before-and-after study. Radiology 2008; 248: 817–823