SciMedCentral

Annals of Biometrics & Biostatistics

Review Article

Statistical Issues in Biomarker Validation for Use in Drug Development

Ziad Taib[1]* and Linda Akrami[2]

[1]*Department of Mathematical Sciences, Chalmers U. of Technology, Sweden*

[2]*Center of Registers in Region Vastra Gotaland, Sweden*

Abstract

Despite the fact that a large number of candidate biomarkers have been identified by biologists during the last decades, very few of these have made it all the way to clinical practice. One of the reasons for this is lack of proper validation on a level that is satisfactory to the authorities and the medical community. Biomarker validation, viewed as a confirmatory process aiming at validating a specific biomarker for a certain purpose, should ideally be based on proper statistical models and hypothesis testing methodology. In this chapter, we will consider such validation methods based on type of biomarker and discuss several associated pitfalls from a statistician's perspective.

***Corresponding author**

Ziad Taib, Department of Mathematical Sciences, Chalmers U. of Technology, 412 96 Gothenburg, Sweden, Tel: +46707655471, Email: ziad@chalmers.se

Submitted: 23 March 2020

Accepted: 17 April 2020

Published: 22 April 2020

ISSN: 2333-7109

Keywords

- Biomarker
- Prognostic biomarker
- Predictive biomarker
- Surrogate end-point
- Biomarker validation
- Interaction test
- Simulation

INTRODUCTION

Biomarkers are increasingly important in most areas of bio-medical research, especially as efficacy markers in drug development and diagnostic tools in precision medicine [1]. But to be truly, useful, scientists need to maintain a very high level of scientific rigour when selecting and validating specific biomarkers. To achieve that, the use of relevant statistical hypothesis testing methods and adequate data is warranted.

To set the scene, we start by defining a number of concepts that will be used throughout this article. A clinical endpoint is defined as any measurement that captures information on how patients would feel, function or survive. In contrast, a biomarker is defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.

Biomarkers play an important role in every aspect of clinical research and in particular in drug discovery and development. Examples of such aspects are safety, efficacy, target engagement, dose finding, patient stratification, companion diagnostics just to mention a few.

This article contains a review of statistical methods relevant for the issue of biomarker validation for use in e.g. clinical trials. These methods differ between the different types of biomarkers. There are many ways to classify biomarkers but for our purposes, we will mainly discuss three types; prognostic, predictive and surrogate biomarkers. Short good nontechnical reviews of this topic can be found in [2] and [3].

This article is organized as follows. Section 2 contains definitions of the main concepts and models used in later sections. In Section 3, we discuss how validation can be performed for various types of biomarkers while Section 4 contains illustrative examples. These examples are all related to Chronic Obstructive Pulmonary Disease. Section 5 is devoted to various sources of data that can be used for validation of biomarkers and Section 6 contains a brief discussion around relevant guidance documents. Finally, in Section 7 we include a discussion of the material presented in the whole article.

PRELIMINARIES

It is helpful to classify biomarkers as being prognostic, predictive or surrogate biomarkers. In this section, we formulate a basic statistical model against the background of which, we will provide precise definitions of the different types of biomarkers. A prognostic biomarker predicts the likely course of disease, irrespective of treatment while a predictive biomarker forecasts the likely response to treatment. A surrogate endpoint is supposed to replace a clinical endpoint while having some extra features such as being less invasive or having higher sensitivity in evaluation of the effect of a certain treatment. To formulate these in terms of statistical models, we introduce the following linear model for a continuous outcome and write

$$\mathrm{Y} = \beta_o + \beta_1 T + \beta_2 B + \beta_3 T \times B + \in$$

$$E\left[\mathrm{Y} \mid T, B\right] = \beta_o + \beta_1 T + \beta_2 B + \beta_3 T \times B$$

where

Y = Outcome

- T=Treatment
- B=Biomarker
- *e*=Random error (normally distributed with mean zero and variance σ^2)

β_o = Intercept

β_1 = Coefficients of the effect of the treatment

β_2 = Coefficients of the effect of the biomarker

β_3 = Coefficients of the interaction effect

A similar model can be used in case of time to event outcomes based on Cox regression

$$H(t) = H_0(t)\,exp(\beta_o + \beta_1 T + \beta_2 B + \beta_3 T \times B).$$

For other cases (e.g. binary data), one can define similar models based on appropriate link functions resulting in a generalized linear model e.g.

$$g(E[Y \mid T, B]) = exp(\beta_o + \beta_1 T + \beta_2 B + \beta_3 T \times B).$$

The above model can be specialized to cover the different cases for the Biomarker:

1. Both Prognostic and Predictive Biomarker

$$E[Y \mid T, B] = \beta_o + \beta_1 T + \beta_2 B + \beta_3 T \times B.$$

2. Predictive but not prognostic Biomarker

$$E[Y \mid T, B] = \beta_o + \beta_1 T + \beta_3 T \times B.$$

3. Prognostic but not predictive biomarker

$$E[Y \mid T, B] = \beta_o + \beta_1 T + \beta_2 B.$$

4. Neither prognostic nor predictive biomarker

$$E[Y \mid T, B] = \beta_o + \beta_1 T$$

VALIDATION

For a biomarker to be used in e.g. clinical trials or in clinical practice, it needs to be somehow validated. Such validation is of course primarily biological but statistical considerations play a key role in the initial validation process.

In the remaining of this chapter, we will adopt the following strict definition of the term Biomarker Validation namely demonstration by statistical methods that the biomarker fulfills one of the following:

- is associated with a given clinical endpoint regardless of treatment (prognostic biomarkers)
- predicts the effect of a therapy on a clinical endpoint (predictive biomarkers),
- is a substitute for the clinical endpoint when assessing the effect of a certain therapy (surrogate end points).

The problem of validating biomarkers is no different from many other statistical tasks where it is required that a certain factor (Biomarker) be shown to be important to a certain outcome (Clinical outcome or treatment effect). But there are caveats and issues that renders this goal hard to achieve from a practical point of view. Such challenges will be discussed at the end of the article.

Validation of Prognostic Biomarkers

Establishing a prognostic biomarker is relatively straightforward from a statistical point of view and can be performed using many different study designs. The issue can be formulated as one of association between the biomarker and the clinical outcome. This can be formulated as a test of the null hypothesis $H_o : \beta_2 = 0 \text{ vs. } H_1 : \beta_2 \neq 0.$

Internal validation can be done using cross validation, but to become an established biomarker, the result needs to be replicated in several additional independent data sets.

Validation of predictive biomarkers

A biomarker is predictive when the baseline value, or changes in the value of the biomarker over time, predicts the response to the treatment based on the clinical outcome. Statistical validation of a predictive biomarkers requires data from large randomized trials that include patients with a wide range of values of the biomarker.

To establish a predictive biomarker based on data from a clinical study involving both a control and a treatment arm, a traditional statistical approach can be based on the model

$$E[Y \mid T, B] = \beta_o + \beta_1 T + \beta_2 B + \beta_3 T \times B.$$

The highest level of evidence is obtained when using an `interaction' design to show that the effect of the treatment T on the clinical endpoint Y depends on the biomarker B. Accordingly, one needs to test for interaction between the biomarker and the treatment effect, i.e. to test the null hypothesis that $H_o : \beta_3 = 0 \text{ vs. } H_1 : \beta_3 \neq 0.$

Although this seems to be straightforward, there are many issues that need to be dealt with. One of the most important issues is that the power to test for interaction effect is often poor since sample size calculation is based on addressing main effects as the major objective of the study as is pointed out by [4].

Determining the proper sample size, requires the specification of a testing method for the interaction effect parameter. In what follows, we discuss two natural testing approaches in the normal distribution case. For other types of data (e.g. binary or survival data) the same methods work with small changes. For more complicated cases involving other distributions or correlated data, an approach based on simulations might be the most direct approach.

Approach 1: Wald's test: The Wald type test statistic for the test of interaction in our basic model above is of the form

$$\beta_3 X_w = \beta_3^2 \Big/ \sigma_{\beta_3}^2$$

The Wald test statistic follows an approximate chi-square (x^2) distribution under large sample conditions.

Approach 2: A Likelihood Ratio test: A likelihood ratio test can be obtained to test $\beta_3 = 0$ against a certain alternative by comparing the log-likelihood for the full model (L_F) to the log-likelihood for the reduced model without the interaction term (L_R). The resulting test statistics is $X_{LR} = -2\left[\log(L_R) - \log(L_F)\right]$

Like the Wald test, this statistic is approximately (x^2) distributed under large sample conditions.

Validation of surrogate biomarkers

Validation of surrogate biomarkers is a complex and technical issue that does not fit within the remit of the present work. We will however, try to include a simplified version for the sake of completeness.

At present, validation criteria for surrogate biomarkers is still an area statistical research but with signs of an emerging consensus. Some of the approaches that have been proposed as validation criteria are the following `

- Causality
- Prentice criteria
- Individual level association
- Trial level association
- Relative Effect
- Proportion explained
- Adjusted Association

In what follows, we account for one of these approaches: Prentice criteria. Much of the research in this area started with the seminal work by [5]. For a given triplet (T;B;Y), Prentice formulated the idea that for a surrogate endpoint to be validated, any test of the null hypothesis of no effect of T on B is also a valid test of the corresponding null hypothesis based on the true endpoint Y.

Despite its elegance, this idea is not practically useful for evaluating surrogate endpoints since it requires too much information. Therefore, Prentice suggested that a biomarker B is regarded as a valid surrogate for a clinical endpoint Y under a treatment T if the triplet (T;B;Y), satisfies four operational criteria:

- Treatment must have a significant effect on the biomarker.
- Treatment must have a significant effect on the clinical endpoint.
- The biomarker must have a significant effect on the clinical endpoint.
- The effect of treatment on the true endpoint vanishes when adjusted for the surrogate.

To be consistent with our basic model, we formulate a version of the above criteria in the case of parametric regression models. As before, survival- and generalized versions of the models can be formulated.

$$E[B \mid T] = \mu_{B|T} + \alpha T$$

$$E[Y \mid T] = \mu_{Y|T} + \beta T$$

$$E[Y \mid B] = \mu_{Y|B} + \gamma B$$

$$E[Y \mid T, B] = \mu_{Y|T,B} + \beta_B T + \gamma_T B$$

$$E[Y \mid T, B] = \mu_{Y|T,B} + \beta_B T + \gamma_T B + \delta TB$$

Surrogacy can now be assessed by demonstrating that the treatment has an effect on both the clinical end point and the biomarker, that the biomarker has an effect on the clinical endpoint and that in fact it (the biomarker) captures all the effect of the treatment on the clinical endpoint. This turns out to be achievable by testing null hypotheses of $\alpha = 0$, $\beta = 0$, $\gamma = 0$ and $\beta_B = \delta = 0$. In practice however, this can only be achieved based on either data from a very large clinical trial (possibly retrospectively) or meta-analysis of several trials.

ILLUSTRATIVE EXAMPLES

Even if biomarkers are mostly used in oncology, they are not less equally important in other disease areas. To demonstrate this, in what follows, we present illustrative examples of the various types of biomarkers, all from related to Chronic Obstructive Pulmonary Disease or COPD. Therefore, a short introduction to COPD is warranted. COPD, which is a common, preventable and treatable disease, is characterized by progressive airflow limitation and associated with an enhanced chronic inflammatory response in the airways and the lung. The number one risk factor for COPD is smoking. Until recently, the main primary endpoint in COPD trials was the lung function measure Forced Expiratory Volume in one second (FEV1).

While FEV1 is still used in early phase clinical trials, it has, to large extent, been replaced by Exacerbations in proof of concept- and confirmatory trials. Exacerbations are acute deteriorations triggered by e.g. bacterial and viral pathogens. They speed up disease progression and have major implications on quality of life, morbidity and mortality. Inhaled corticosteroids and long-acting β 2 agonist (ICS-LABA) combinations are commonly prescribed for COPD patients. Unlike Asthma, no advanced therapy exists for COPD but there is an expectation that in the future the use of Biomarkers will lead to more efficient clinical trials resulting in new and better therapies for treating COPD while it is in an early stage.

Prognostic biomarkers for Lung function decline

Tantucci and Molina [5] point out that lung function loss assessed as expiratory airow reduction, seems more accelerated and therefore more relevant in the initial phases of COPD. To have an impact on the natural history of COPD, it is, therefore, logical to treat COPD already in the earlier stages. Therefore, finding prognostic biomarkers capable of detecting signs of rapid decline already at an early stage would be very beneficial.

In Ostling et al., [6] results are presented from a longitudinal study conducted in Denmark with annual visits between 2005 and 2009, where lung function decline was followed in a group of

healthy smokers and ex-smokers with a history of more than 20 pack years. The overall rationale for the study was to determine the risk of smokers to develop cardiovascular disease, lung cancer and/or COPD. Based on data from this study, a number of blood plasma proteins were investigated that potentially could be used as prognostic biomarkers, capable of identifying smokers that will lose lung function more rapidly than other smokers.

Patterns of annual FEV1 decline were analyzed and FEV1 decline rate was calculated based on FEV1 values measured over the 4 years period using a linear mixed model to identify prognostic biomarkers capable of identifying rapid FEV1 decliners at early disease stages. The ratio ([ApoD/MMP9]/[E-selectin]) showed a strong linear correlation to FEV1 decline rate.

Other candidate prognostic biomarkers of lung function decline are low levels of the airway protein of Clara Cell 16 (CC16) and serum PRG4 (cf. [7,8]). CC16 is associated with severity of the disease and recent findings indicate that low levels of CC16 in serum are associated with accelerated lung function decline. The ECLIPSE study reported a significant inverse association between the rate of FEV1 decline and serum levels of CC16 in COPD patients, which was also recently confirmed in other cohorts of COPD patients. Moreover, serum PRG4 is an important biomarker for supporting the COPD diagnosis and relates to the decline in lung function in patients with COPD.

Predictive: Eosinophil count

Several recent reports indicate that the level of eosinophils in blood is predictive of the response to treatment with inhaled corticosteroids (ICS) to prevent exacerbations whereas patients with low levels constitute a group with an unmet need for better treatment. Blood eosinophils are easy and reproducible to measure, making them a good predictive biomarker that can be used for patient selection in clinical trials of novel drugs targeting exacerbations.

Based on Data from Astrazeneca studies in COPD patients (4528 in total) with a history of exacerbation [9] shows that eosinophil count determines response to Budesonide-formoterol as compared with formoterol alone in the reduction of exacerbations, improvements in lung function, and health status. Interactions were also observed between eosinophil count and the treatment effects of budesonide-formoterol over formoterol on St George's Respiratory Questionnaire (p=0.0043) and pre-bronchodilator FEV1 (linear effect p<0.0001, p=0.067). Only eosinophil count and smoking history were independent predictors of response to budesonide-formoterol in reducing exacerbations (eosinophil count, p=0013; smoking history, p=0015).

CompEx: A Novel composite surrogate endpoint for severe exacerbations

Clinical trials in COPD with severe exacerbations as the primary endpoint are lengthy and expensive due to the low rate of exacerbation events which is a serious problem in early clinical development. But it is reasonable to hypothesize that events defined based on diary data may capture less severe, but still clinically relevant, episodes of symptom worsening. Examples of such symptoms are morning PEF, evening PEF, total reliever medication use and COPD symptoms.

In [10] the aim was to establish a composite endpoint of mixed eDiary variables capturing clinically relevant disease deteriorations, to be used as a surrogate for exacerbations, predictive for effect in Ph3 trials. The resulting surroagte endpoint named CompEx resulted in 2.8 times more events than severe exacerbations while preserving the treatment effect (average HR 1.01). The increased number of events, together with the sustained treatment effect, resulted in a large net gain in power, with a 67% mean reduction in sample size.

CompEx has a potential to accelerate early clinical development of new agents for treatment of COPD by enabling shorter (3m) and smaller trials.

An attempt to validate CompEx as a surrogate endpoint for severe exacerbations can be found in a recent thesis by [11].

Diagnostics

Recent technological advances have made it possible to discover new biomarkers for the diagnosis, prognosis, therapeutic response prediction and population screening of e.g. human cancers. But despite the promise of biomarkers for use in diagnosing e.g. cancer, there are few commercialized biomarker molecular diagnostics that could help clinicians to choose between costly drug alternatives or avoid the use of toxic drugs or unnecessary interventions. We quote [12] for the need of basing such diagnostic tools on solid data and adequate statistical validation methods.

- In the post-genomics era, omics technologies offer exciting opportunities in biomarker discovery and cancer diagnostics. However, the data generated by these technologies is not reproducible or robust enough for clinical use."
- "One challenge is to validate omics findings in prospective, well-controlled clinical studies".
- "To achieve these goals, effective interdisciplinary communication and collaboration involving the fields of molecular biology, epidemiology, electronic engineering, physics, chemistry, biostatistics, computer science, mathematics with clinicians, is required, to perform successful and efficient research into biomarker discovery and molecular diagnosis."

As one of the first success stories of a treatment and diagnostic tool, we mention the case of HER2 and Trastuzumab (Herceptin). In normal quantities, Human epidermal growth factor (HER2) promotes cell growth. However, in presence of certain mutations, it can cause certain breast cancer cells to multiply in an uncontrolled manner due to overexpression.

HER2 protein overexpression was shown to be a prognostic biomarker associated with increase relapse and mortality as well as a predictive biomarker for response to certain therapies. Understanding the role of HER2 protein as a biomarker lead to the development of Herceptin as a first line therapy. The drug was subject to fast track approval because it could help patients not being responsive to conventional treatment and,

more importantly, a diagnostic test available for identifying the right responder patient group. Its effect has been evaluated in randomized clinical trials and meta analyses.

DATA SOURCES

To establish prognostic biomarkers, observational data from retrospective studies can often be enough. However, establishing a prognostic biomarker based on one data set should include some cross validation using e.g. resampling techniques. Of course, the Gold standard, that is not always practically available, is randomized, prospective evidence in randomized clinical trials.

For predictive biomarkers, prospective randomized clinical trials produce the best type of data. In practice, however, it might not be feasible to perform large clinical trials for the sole purpose of validating a biomarker, so one is limited to existing observational data. In such cases, the retrospective analyses should be planned carefully thus mimicking as many aspects as possible of randomized clinical trials. Alternatively, it might turn out that an exploratory trial is performed but that lacks statistical power due to low sample size. Other issues are discussed in [13].

Ideally for validating surrogate endpoints, you need either data from a very large trial or data from multiple sources, for instance if several clinical trials have been performed on the same therapy, it will be possible to estimate the treatment effects upon the marker and upon the clinical endpoint in each of these trials and assess the association between these.

In an effort to validate a biomarker in the absence of perfect data, one should look at multiple sources of data. An example of such sources include but is not limited to the following:

- Randomized controlled trial
- single-arm/historical control trial
- Cohort studies
- Case-control studies (including nested)
- Cross-sectional studies
- Case series or case reports
- Registry information
- Meta-analyses

One should be aware that relying on observational data implies risk of confounding, whereas the need to match groups using e.g. propensity score.

Guidelines

One of the most relevant guidelines is Biomarker Qualification: Evidentiary Framework: Guidance for Industry and FDA Staff (cf. [Buyse et al., [2]) which is a draft guidance containing some interesting facts regarding validation of biomarkers, especially Section V on Statistical Considerations. In that document, it is pointed out that the International Conference on Harmonization (ICH) Guideline E9, Statistical Principles for Clinical Trials [10] should also be used for this purpose, even if its primary objective is interventional studies. Additional relevant information can of course also be found in various disease area Guidance documents.

For surrogate endpoints, the most important criteria for valid surrogates are summarized in the International Conference on Harmonization (ICH) Guideline E9, Statistical Principles for Clinical Trials. These mainly define the relationship between the surrogate endpoint (high blood pressure, for example) and the "hard" clinical endpoint (such as stroke) actually relevant when treating the condition. To show this relationship, the following must be

demonstrated:

- Biological plausibiliy.
- Statistical relationship in epidemiological studies.
- Evidence from clinical studies that treatment effects on the surrogate correspond to the clinical outcome.

SOME RESULTS BASED ON A SIMULATION EXPERIMENT

In this section, we discuss various issues related to statistical validation of biomarkers and present some learnings from a simulation experiment. The simulations were based on the basic model as a starting point but with modifications as in (1)-(3) in section 2. In this way, we cover the cases where the biomarker is both prognostic and predictive, when it is only predictive and when it is only prognostic. Data from these different scenarios were analyzed using PROC GLM in SAS to study what happens when the analysis is consistent with the simulated data and when the analysis assumes a different model than the true one.

Moreover, we studied the effect of powering a study for the main effect only but using the data to assess the biomarker. We also investigated the claim that the power to detect an interaction effect can be increased by increasing the Type I error.

The simulations

Simulations were used based on the basic full model (1) and the reduced models (2)-(3) with and without covariates. The simulated data were analysed using proc GLM in SAS for a variety of models reflecting (1)-(3). The numerical values for the model parameters were as follows. alpha varied between 0.05 and 0.25, Power=1- β varied between 0.30 and 0.90. β_0 =10, β_1 =5, β_2 =1.25, β_3 =1. Moreover, MB = 2:5; SD_B = 1; M_C = 1:5; SD_C =1; SD_{Error} = 5; Corr (B;C) = 0:6

Low statistical power

Biomarker data used to assess the validity of biomarkers originate often from studies with the primary goal of investigating the effect of a treatment. Therefore, such studies are often under-powered to detect the interaction effect between a predictive biomarker and the treatment. In our simulations, this was quite clear whereas the sample size to detect the interaction effect was nearly three times the sample size needed for the main effect. For a prognostic biomarker, however, the sample size to detect the biomarker effect was about the same as the sample size needed for the main effect.

To handle the low power issue, Polley et al., [7] proposed fitting a model with interaction, but without the main effect term.

$$E\left[\mathrm{Y} \mid T, B\right] = \beta_o + \beta_2 B + \beta_3 T \times B.$$

Although we could see this effect in our simulations, there was also a decrease in power in detecting the prognostic aspect of the biomarker. Therefore, although this approach seems promising, we think that more research is needed in order to understand under what circumstances such an approach would lead to higher power.

Along the same line, some propose raising the Type I error rate, thereby increasing power, when testing interactions. However, Marshall, [8] points out that this can be a poor analysis strategy. Such an increase in power is obvious based on a theoretical ground and of course, we could clearly see such an effect in our simulations. For instance, in our simulations the power to detect the interaction effect went from 70% to nearly 80% when we increased the type I error from 0.05 to 0.10. The question, however, is whether this approach is useful from a practical point of view. To demonstrate this point, Marshall [8] quantified the gain in power for testing interactions when the Type I error rate is raised, for a variety of study sizes and types of interaction based on several test methods and different types of interaction. The conclusion is that relaxing the Type I error rate did not usefully improve the power for tests of interaction in many of the scenarios studied.

The gain in power obtained by raising the Type I error needs to be seen against the disadvantage of increased "false positives" rate. In most situations, false positives are more troublesome that false negatives. Based on our simulations, we agree with Marshal's conclusion that increasing the Type I error rate when assessing tests of interaction is not recommended.

Prognostic vs. predictive

As was already pointed out, the same biomarker can be prognostic and predictive, whereas one needs to separate these effects. Sechidis, et al., [14] proposed an information theoretic approach. We simulated data from models (1)-(3) in section 2 and analysed these in many different ways reflecting the biomarker is prognostic, predictive and both prognostic. In general, we could see that analyzing data based on a different model than the one used to simulate the data leads to inaccurate conclusions. As an example if the biomarker is predictive but not prognostic, trying to estimate the prognostic effect gives very small (negligible) values. However, not including a prognostic effect when it should have been included leads to loss of power.

Covariates

The effect of adding covariates (risk functions) to models like the ones discussed in this work has been studied using simulations by Haller et al., [15,16]. Our simulations confirm the finding in that reference that the power to detect interaction term is not affected by the addition of a prognostic risk factor or covariate, unless is correlated with biomarker. Moreover, the interaction estimate is biased when relevant prognostic factors are not considered.

Of course, adding too many covariates to the model is, in general, not recommended.

DISCUSSION AND FURTHER COMMENTS

In this section, we discuss various issues related to statistical validation of biomarkers.

Cut points

A common albeit overemphasized practice is the use of optimal cut-points or thresholds of biomarker values. Such cut-points can be needed to provide guidelines for clinical decision making, but lead to inevitable and unfortunate loss of potentially useful information if such dichotomization is performed already at the validation stage. It is therefore, good practice to work with the continuous values for as long as possible till the final stage when the model is built based on all the biomarkers values.

Pragmatism

The main conclusion of the article is that there is a divide between statistical methodology for biomarker validation and what is practically and possible. This results in current statistical methods being somewhat too strict to be useful. Therefore, scientists try to use pragmatic approaches based on whatever data is available. There is however, no consensus on how a more pragmatic approach should look like. In the meanwhile, there are things that can be done to handle certain shortcomings of the current methods. Some of these are discussed in what remains of this section.

Consensus

Another issue on which there is no consensus is the question of "how strong is strong?" i.e. what level of relationship is required to conclude that the biomarker is valid? There are no commonly accepted criteria for quantifying minimum strength in the same way as 0.05 and 0.20 in type I and type II errors respectively in hypothesis testing or confidence degree 95% for confidence intervals. Consideration of specificity and sensitivity could add some confidence but again there are no consensus values for these. It all depends on the so called Context of Use (COU) and especially on the biological knowledge/confidence.

Caveats

Many other general statistical issues that we have not mentioned explicitly in this work could lead to inadequate analyses, some of which are discussed in the literature. Here are a few examples

- Data involving repeated measures introduce correlations, whereas the need to use mixed models where the correlation structure should be modeled carefully.
- Statistical multiplicity issues can arise for various reasons e.g. handling multiple biomarkers. Notice that including multiple biomarkers in a model also leads to colinearity problems. Multiple endpoints, also lead to multiple testing problems that need to be addressed properly using e.g. gate-keeping, sequential testing etc.
- Selection bias, Meta analyses based on e.g. published data often suffers from selection bias issues since positive results are published with higher probability.
- Relying on observational data implies risk of confounding,

whereas the need to match groups using e.g. propensity scores.

REFERENCES

1. Guest PC. The Importance of Biomarkers: The Required Tools of the Trade. In: Biomarkers and Mental Illness. Copernicus, Springer, 2017; 31-41.
2. Buyse M, Michiels S, Sargent DJ, Grothey A, Matheson A, De Gramont A. Integrating biomarkers in clinical trials. Expert Rev Mol Diagn. 2011; 11: 171-182.
3. Buyse M, Molenberghs G, Paoletti X, Oba K, Alonso A, Van der Elst W, et al. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. Biom J. 2016; 58: 104-132.
4. James J C, Tzu-Pin L, Yu-Chuan C, Wei-Jiun L. Predictive biomarkers for treatment selection: statistical considerations. Biomark Med. 2015; 9: 1121-1135.
5. Tantucci C, Modina D. Lung function decline in COPD. Int J Chron Obstruct Pulmon Dis. 2012; 7: 95-99.
6. Ostling1 J, Taib Z, Van Geest M, Bruijnzeel1 P. Blood Biomarkers Predict Rapid Fev1 Decline In COPD. Am J Respir Crit Care Med. 2014; 189; A5879.
7. Polley M, Freidlin B, Korn EL, Conley BA, Abrams JS, McShane LM. Statistical and practical considerations for clinical evaluation of predictive biomarkers. J Natl Cancer Inst. 2013; 105: 1677-1683.
8. Marshall SW. Power for tests of interaction: effect of raising the Type I error rate. Epidemiol Perspect Innov. 2007; 4: 4.
9. Bafadhel M, Peterson S, De Blas MA, Calverley PM, Rennard SI, Richter K, et al. Predictors of exacerbation risk and response to budesonide in patients with chronic obstructive pulmonary disease: a post-hoc analysis of three randomised trials. Lancet Respir Med. 2018; 6: 117-126.
10. Da Silva CA, Bengtsson T, Peterson S, Karlsson N, Fageras M, Jauhiainen A. COPD CompEx: A Novel Composite Endpoint to Accelerate Early Clinical De- velopment of New Agents for Treatment of COPD. Am J Respir Crit Care Med. 2018; 197: A7428.
11. Zolghadr M. Evaluation of Surrogate Endpoints with Applications in Respiratory Clinical Trials. Thesis for the PhD degree in pure and applied mathematics. Department of Mathematical Sciences, Politecnico di Torino and Universita degli studi di Torino. Italy, 2018.
12. Xuewu Zhang, Lin Li, Dong Wei, Yeeleng Yap, Feng Chen. Moving cancer diagnostics from bench to bedside. Trends in Biotechnology. 2007; 25.
13. Sargent DJ, Mandrekar SJ. Statistical issues in the validation of prognostic, predictive, and surrogate biomarkers. Clin Trials. 2013; 10: 647-652.
14. Sechidis K, Papangelou K, Metcalfe PD, Svensson D, Weatherall J, Brown G. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. Bioinformatics. 2018; 34: 3365-3376.
15. Haller B, Ulm K. A simulation study on estimating biomarker-treatment interaction effects in randomized trials with prognostic variables. Trials. 2018; 19: 128.
16. Haller B, Ulm K, Hapfelmeier A. A Simulation Study Comparing Different Statistical Approaches for the Identification of Predictive Biomarkers. Computational and Mathematical Methods in Medicine. 2019.

Cite this article

Taib Z, Akrami L (2020) Statistical Issues in Biomarker Validation for Use in Drug Development. Ann Biom Biostat 5(1): 1032.