In this issue: The Impact of Laboratory Medicine

The Journal of the International Federation of Clinical Chemistry and Laboratory Medicine

eJIFCC

Evaluating biomarkers for guiding treatment decisions

Patrick M. Bossuyt, Parvin Tajik

Department of Clinical Epidemiology, Biostatistics & Bioinformatics, Academic Medical Center, University of Amsterdam, The Netherlands

ARTICLE INFO

Corresponding author:
Patrick M. Bossuyt
Dept. Clinical Epidemiology, Biostatistics & Bioinformatics
Academic Medical Center
University of Amsterdam
Room J2-127 PO Box 22700
1100 DE Amsterdam
The Netherlands
E-mail: p.m.bossuyt@amc.uva.nl

Key words:
biomarker, medical test evaluation, evidence-based medicine, clinical effectiveness

ABSTRACT

The genetic revolution is expected to lead to improved targeting of new and existing forms of treatment. Rather than a one-size-fits-all blockbuster strategy in battling disease with drugs and other interventions, a more precise approach is becoming available, one in which treatment is only offered to those likely to benefit. The identification of those likely to benefit from treatment could be based on one or more biomarkers, but in an era where medical decisions aim to be evidence-based, the use of treatment selection markers should not just be based on hope and optimism, but on solid data from sound research. The performance of the treatment selection marker should be expressed in quantitative terms, similar to the way we express the clinical performance of diagnostic markers, or the performance of prognostic markers.

We describe recent research on this issue. First we present in intuitive terms a general, decision-theoretical framework for making treatment decisions. We then describe some measures for expressing the performance of treatment selection markers, showing that conventional measures of clinical performance, such as clinical sensitivity and specificity, are not decisive or helpful. In the last part of the paper, we provide a brief summary of study designs for evaluating treatment selection markers. Like all other forms of medical testing, potential treatment selection markers should be properly evaluated before they are implemented in routine clinical practice.

INTRODUCTION

The unraveling of the human genome has fuelled high hopes for the advancement of clinical medicine. Many believed that our improved understanding of the role of genes, the function of proteins, and the characterization of small-molecule metabolite profiles would strengthen our understanding of the origins of disease, and would help to clarify disease mechanisms. This would eventually lead to new and better forms of treatment, enabling clinicians to sustain and restore health for their patients, and to prevent premature death.

The benefits from the genetic revolution would not just come from new forms of treatment. The advances in knowledge were also expected to lead to improved targeting of new and existing forms of treatment. Rather than a one-size-fits-all blockbuster strategy in battling disease with drugs and other interventions, a more precise approach would become available, one in which treatment is only offered to those likely to benefit. The identification of those likely to benefit from treatment would be made based on one or more biomarkers. We will refer to such biomarkers as “treatment selection markers”.

In an era where medical decisions aim to be evidence-based, the use of treatment selection markers would not just be based on hope and optimism, but on solid data from sound research. It is not sufficient to expect a benefit from using a biomarker to guide treatment decisions, one should also have convincing evidence that the marker is actually able to do so. The performance of the treatment selection marker should be expressed in quantitative terms, similar to the way we express the clinical performance of diagnostic markers, or the performance of prognostic markers.

These new ambitions pose a challenge for laboratory professionals, and for researchers and methodologists in general. How does one know that a marker is fit to serve as a guide for treatment decisions? How can one express the performance of a treatment selection marker?

This paper summarizes some recent research on this issue. First we present in intuitive terms a general, decision-theoretical framework for making treatment decisions. We then present some measures for expressing the performance of treatment selection markers, showing that conventional measures of clinical performance, such as (clinical) sensitivity and specificity, are not decisive or helpful. In the last part of the paper, we provide a brief summary of study designs for evaluating treatment selection markers.

THE ANATOMY OF TREATMENT DECISIONS

In general, a treatment decision is based on balancing the positive, hoped-for effects against the negative, feared effects. The latter could be a combination of the side-effects of treatment, the burden of treatment (going to the hospital at regular intervals, or taking pills daily), and the societal costs: the resources used to develop, build and administer treatment. The positive effects are the health gains expected from treatment: restoration of health, or the prevention of worsening.

If we assume the negative effects are all known, we can re-express the treatment decision as a threshold issue. Are the positive effects large enough to offset the negative ones? Assume, for example, that the positive, hoped-for effect is an increase in 5-year survival from adjuvant chemotherapy for a cancer patient. Assume, additionally, that we have a reliable estimate of the 5-year survival for that patient. We then can present the negative effects of adjuvant chemotherapy to the patients and ask the patient how large the gain in 5-year survival have to be to justify treatment for that patient. Assume

then that a new, large RCT comes out that has estimated the survival benefit of this form of adjuvant chemotherapy for patients similar to the one facing the decision. That patient then can compare the gain in survival – in absolute terms – with the personal threshold. If the gain is larger than the threshold, adjuvant chemotherapy seems justified. Otherwise, if the gain is smaller than the threshold, this is not the case.

In this case we base the recommendation about treatment not on the statistical significance of the treatment effect, as estimated in the randomized trial. As is well known, such a significance test only evaluates whether the difference in survival is zero. In case of a significant result, we have rejected the null hypothesis of equality. With a two-sided test, this implies that the alternative hypothesis specifies that the survival difference is either negative or positive; with a one-sided test, the alternative hypothesis typically specifies that there is some survival gain. So conventional statistical significance tests typically do not indicate whether the health gains are large enough. We must add that, in principle, it would be perfectly possible to formulate an alternative statistical hypothesis test, one in which we test whether the treatment effect exceeds a prespecified threshold, but this is not typically done in randomized trials.

The recommendation about treatment is also not based on the target difference, as used in the sample size calculations. This target difference helps to calculate the desired precision of a study, which is typically driven by the number of included study participants. The target difference can provide reassurance that the study will be informative, in the sense that a relevant difference, if one exists, is likely to be detected with the required statistical precision (1).

Asking for a threshold for the treatment effect sounds like a complicated question to ask a patient. It is probably not an easy task to define a personal threshold, but existing research has shown that the question is indeed answerable. For adjuvant chemotherapy, for example, the actual question “what makes it worthwhile” has been asked to patients with non-small-cell lung cancer (2), to patients with early colon cancer (3), and patients with early breast cancer (4).

The threshold does not have to be same for every individual patient: for some the required gain may be fairly large, while for others extending survival is extremely important, and their threshold for accepting treatment is close to zero. This is definitely an area for personalized medicine: not in the abundant use of next-generation sequencing, but in the recognition that personal values and trade-offs differ. Despite this recognition, we will assume for now that there is one common threshold, to ease the exposition.

In itself, the threshold approach is as old as decision theory. It was introduced, or re-introduced, into medicine in the 1970s, through impressive articles written by Steve Pauker and Jerome Kassirer, which formed the start of clinical decision analysis and helped to launch economic evaluations in health care (5, 6).

Note also that the question about a large benefit is usually phrased in terms of the absolute benefit: the survival gain in percentage points at five years, for example. Although treatment effects in trials are typically expressed in relative terms, answering the question about the threshold in such relative terms is much more challenging and complicated.

TREATMENT SELECTION MARKERS

So, when can a marker act as a treatment selection marker, to guide decisions about treatment?

The threshold approach to decision-making, as just introduced, allows us a simple rule to arrive at a conclusion when evaluating a biomarker's potential to guide treatment. We assume for now that the marker is present or absent, or takes values in a well-known range. To be sufficiently general, we suppose that the marker is quantitative, be it on a dichotomous (1/0), ordinal, or interval scale.

One condition for a marker to act as a treatment selection marker is the existence of heterogeneity in the treatment effect. Keeping to the example of survival gain from adjuvant chemotherapy, this means that not everybody in the trial population is expected to benefit to the same degree from the treatment: for some the benefit is larger, for others smaller, and there may be subgroups who do not benefit from chemotherapy, but are even harmed by it: their 5-year survival is lower after treatment.

A second condition is then the existence of a reliable association between the putative treatment selection marker and treatment benefit. We can further specify this condition in terms of a classification, relative to the (common) threshold: the marker is able to identify a subgroup for which the survival gain is equal to or larger than the threshold, separating it from another subgroup where the survival gain is smaller, or even nonexistent: patients are not helped or even harmed by the treatment. The first group benefits from treatment – the gains exceed the threshold – while the second group does not.

A marker can then act as a treatment selection marker if there is a value, or a range of values, that corresponds to a group who benefits, and the remaining values correspond to a group that does not benefit.

What then if the personal treatment thresholds vary? In that case we have to generalize the second condition, over the distribution of values for the treatment threshold. The marker may be able to act as a treatment selection marker for some, but not for all. If it can act as a marker for at least one (group of) patients, in the sense we just described, then it can be qualified as a (potential) treatment selection marker.

To further facilitate presentation of concepts and performance measures we describe a clinical decision scenario with a potential treatment selection marker and discuss which measures do and which ones do not measure the performance of the marker for guiding treatment decision.

As an example, we consider using vaginal culture in women with preterm premature rupture of membranes to guide the decision for immediate delivery. In pregnant women in whom rupture of membranes occurs prematurely and before the onset of labour, a decision dilemma is whether to follow a strategy of wait-and-see or to perform immediate delivery to prevent infection and sepsis in the foetus. Bacterial infection causing neonatal sepsis is most commonly associated with the Group B *streptococcus* (GBS) from the mother's vagina. Therefore testing the vaginal GBS colonisation in mothers could potentially identify foetuses at higher risk of infection and may be a good candidate marker for guiding the decision for immediate delivery.

In a trial, about 700 women with premature rupture of membranes were randomly assigned to immediate delivery or wait-and-see strategy (7, 8). Among the women studied, 14% had GBS-colonization and were marker-positive. Table 1 shows the association between the GBS-colonization and the outcome in the trial participants (9).

It may seem that we could quantify the performance of a treatment selection marker with the usual measures of clinical performance: why not use sensitivity or specificity here? Indeed,

Table 1 GBS - colonization and outcomes (9)

Strategy	Patients with neonatal sepsis	% of total	Patients without neonatal sepsis	Total patients
Wait-and-see				
GBS Colonization	7	*15.2%*	39	46
No GBS Colonization	8	*2.6%*	305	313
Total	15	*4.2%*	344	359
Immediate delivery				
GBS Colonization	1	*1.8%*	56	57
No GBS Colonization	9	*2.9%*	297	306
Total	10	*2.8%*	353	363

we could do so, but only if there was a straightforward clinical reference standard to identify with sufficient certainty those who benefited (sufficiently) from treatment, separating those from the rest, who did not. In that case the sensitivity of the treatment selection marker would be the proportion of those who benefited, correctly identified as such by the marker, and the specificity would be the proportion of those who did not benefit, correctly identified as such by the marker.

Unfortunately, this distinction is less easy to make on an individual basis in most treatment studies, where only the outcome under treatment is observed, or the outcome under the absence of treatment. It requires a counterfactual approach then to specify what would have happened with an alternative course of action. Below we will describe how we can use the information from the group of trial participants to evaluate the performance of a putative treatment selection marker.

PERFORMANCE OF TREATMENT SELECTION MARKERS

We have just described the necessary conditions for a marker to act as a treatment selection marker. These are absolute conditions: a marker either is or is not a (potential) treatment selection marker. Yet to make decisions about the actual use of the marker, a more quantitative estimate of its performance is required.

Janes and colleagues have explored a number of statistics to express biomarker performance, with descriptive and inferential methods to evaluate individual markers and to compare candidate markers (10, 11). They proposed useful measures for analyzing marker performance. By combining them they calculate the population benefit from using the marker as a treatment selection marker, compared to a strategy of not using the marker to decide about treatment in subgroups of patients.

We use our clinical example in Table 1 to present these measures.

Proportion of marker-positives

First we turn to the subgroup of patients who are marker-positive, in our example women with GBS colonization. GBS-positive women comprised 14% of women participating in the trial: 103 out of 722 (Table 1). So the proportion of patients in whom treatment recommendations could change following marker measurement is 0.14.

Average benefit of treatment among marker-positives

If marker-positive women receive a wait-and-see strategy, 15.2% of their neonates will develop neonatal sepsis. In contrast, when undergoing immediate delivery only 1.8% of their neonates will develop sepsis. Immediate delivery will therefore result in a reduction of 13.5% in the neonatal sepsis rate in this group: this is the average benefit of intervention in this subgroup.

Change in population event rate with marker-based treatment

This is the main composite measure of marker performance for treatment selection. It is based on the difference in overall outcome between not using the marker and using the marker for treatment decisions, aggregated over all members of the target population. Based on marker status, we will only treat marker positives, so the expected change can be calculated by multiplying the proportion of marker-positives (0.14) with the average benefit of treatment in marker-positives (13.5%): (0.14 × 13.5%) = 1.9%.

In other words, a strategy in which immediate delivery is only considered for marker positives will lead to an absolute decrease of 1.9% in the neonatal sepsis rate, compared to a wait-and-see strategy for all.

The impressive reduction in the neonatal sepsis rate in the GSB positives (minus 13.5%) may look like an adequate expression of marker performance, but it is quite clear that the prevalence of the marker positives should also be included in the evaluation.

The result is a clinically interpretable measure of performance of GBS testing for treatment selection. It evaluates the treatment selection marker in terms of its clinical effectiveness: its ability to lower the number of adverse events in the study population (12). With the same approach one can calculate the impact of application of GBS-based strategy on other outcomes such as cost of care or rate of premature birth to complete an evaluation of the costs and consequences of the marker-based strategy.

In our example we did not discuss chance variability. Janes and colleagues have described methods for statistical inference and hypothesis testing (11). They suggest that the performance measures are only estimated if a null hypothesis corresponding to no marker performance is rejected.

This approach assumed that the marker only acts as a selection mechanism, and that, in itself, it does not lead to the event one tries to prevent. It only does so by guiding treatment. We also assume that the effectiveness of the treatment itself is not affected by knowing the marker status. This could happen with some strategies, for example, through better adherence or a different way of handling side-effects. If these assumptions do not hold, the only way to evaluate the effectiveness of a marker-based strategy would be a randomized trial, allocating eligible participants, to this marker-based strategy or to an alternative: no treatment in all.

By further extending this approach, Huang and colleagues define an extension of the net benefit measure: expected benefit. This measure expresses the reduction in the sum of disease and treatment cost by using the marker, based

on the comparison between a marker-based treatment-selection rule and the optimal treatment strategy without the marker information (13).

PREDICTIVE AND PROGNOSTIC MARKERS

In the oncology literature, the terms predictive and prognostic markers have increasingly been used within the context of stratified or personalized medicine, but their use has been somewhat confusing. Some have stipulated, for example, that predictive markers are associated with drug response, in contrast with prognostic markers, which are associated with disease outcome (14). We have shown that is not so much the association with outcome or drug response that counts, but the ability to separate groups who benefit – with difference in outcome compared to the threshold – from those who do not.

In this relatively young field, several other metrics and statistics have been proposed to express the performance of treatment selection markers. Some of these can be severely misleading, since they cannot provide evidence that a marker is helpful in guiding treatment decisions.

These questionable measures include expressions of the strength of the association between marker status and outcome, not benefit. In Table 1, for example, one can see that marker positives have a six-fold higher risk of neonatal sepsis under a wait-and-see strategy. With a strategy of immediate delivery, the relative risk is 0.6.

Both relative risks give information about the association between GBS colonization and outcome, but in themselves they do not reflect marker performance. Treatment decisions should not be guided by outcome in itself, but by benefit: the expected *change* in outcome produced by the treatment.

CONCLUDING COMMENTS

It is exciting to see the developments resulting from rapid progress in our understanding of molecular processes. Biomarkers and other forms of medical tests are not only used for making a diagnosis or staging a disease, but for many other purposes, including decisions about treatment. To express the performance of such treatment selection markers, and to see whether they can actually be used for this purpose, we need a different set of measures. The classical clinical performance measures, such as clinical sensitivity and specificity, can only be used in rare circumstances. Relying on familiar statistics, such as relative risks, or simple significance tests, may actually be misleading. Like all other forms of medical testing, potential treatment selection markers should be properly evaluated before they can be implemented in daily clinical practice.

REFERENCES

1. Hislop J, Adewuyi TE, Vale LD, Harrild K, Fraser C, Gurung T, et al. Methods for specifying the target difference in a randomised controlled trial: the Difference ELicitation in TriAls (DELTA) systematic review. PLoS Med 2014; 11: e1001645.

2. Blinman P, McLachlan SA, Nowak AK, Duric VM, Brown C, Wright G, et al. Lung cancer clinicians' preferences for adjuvant chemotherapy in non-small-cell lung cancer: what makes it worthwhile? Lung Cancer 2011; 72: 213-8.

3. Blinman P, Duric V, Nowak AK, Beale P, Clarke S, Briscoe K, et al. Adjuvant chemotherapy for early colon cancer: what survival benefits make it worthwhile? Eur J Cancer 2010; 46: 1800-7.

4. Duric VM, Butow PN, Sharpe L, Heritier S, Boyle F, Beith J, et al. Comparing patients' and their partners' preferences for adjuvant chemotherapy in early breast cancer. Patient Educ Couns 2008;72:239-45.

5. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980; 302: 1109-17.

6. Pauker SG, Kassirer JP. Clinical application of decision analysis: a detailed illustration. Semin Nucl Med 1978; 8: 324-35.

7. Van der Ham DP, van der Heyden JL, Opmeer BC, Mulder AL, Moonen RM, van Beek JH, et al. Management of late-preterm premature rupture of membranes: the PPROMEXIL-2 trial. Am J Obstet Gynecol 2012; 207: 276 e1-10.

8. Van der Ham DP, Vijgen SM, Nijhuis JG, van Beek JJ, Opmeer BC, Mulder AL, et al. Induction of labor versus expectant management in women with preterm prelabor rupture of membranes between 34 and 37 weeks: a randomized controlled trial. PLoS Med 2012; 9: e1001208.

9. Tajik P, van der Ham DP, Zafarmand MH, Hof MH, Morris J, Franssen MT, et al. Using vaginal Group B Streptococcus colonisation in women with preterm premature rupture of membranes to guide the decision for immediate delivery: a secondary analysis of the PPROMEXIL trials. BJOG 2014; 121: 1263-72.

10. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. Ann Intern Med 2011; 154: 253-9.

11. Janes H, Brown MD, Huang Y, Pepe MS. An approach to evaluating and comparing biomarkers for patient treatment selection. Int J Biostatistics 2014; 10: 99-121.

12. Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta 2014; 427: 49-57.

13. Huang Y, Laber EB, Janes H. Characterizing expected benefits of biomarkers in treatment selection. Biostatistics 2014 doi: 10.1093/biostatistics/kxu039

14. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 2005; 23: 2020-7.