Statistical determination of cancer biomarkers: moving forward clinically

Dr. Marika Mokou[1], Prof. Dr. Dr. Harald Mischak[1,2*], Dr. Maria Frantzi[1]

[1]Department of Biomarker Research, Mosaiques Diagnostics GmbH, Rotenburger Str. 20, D-30659 Hannover, Germany; mokou@mosaiques-diagnostics.com, mischak@mosaiques-diagnostics.com, frantzi@mosaiques-diagnostics.com
[2]Institute of Cardiovascular and Medical Science, University of Glasgow, 126 University Avenue, G12 8TA Glasgow, United Kingdom; Harald.Mischak@glasgow.ac.uk

[*]Correspondence: mischak@mosaiques-diagnostics.com, Phone: +4951155474413, Fax: +49 51155474431

Keywords: *Biomarkers, cancer, diagnostic, precision medicine, predictive, statistics*

1. Introduction

The use of biomarkers in oncology has revolutionized diagnosis, monitoring and treatment in many cancer types by promoting the concept of a personalized approach to guide treatment decision making and monitoring [1]. Cancer biomarkers can assist on risk estimation, detection of a tumor or its recurrence, prediction of response to available treatments for the specific cancer type and assessment of treatment outcome, as illustrated in **Figure 1**. The huge variability among patients with the same cancer type, as well as heterogeneity even within the same tumor specimens, necessitates a tailored cancer care according to individual patient or tumor characteristics.

2. State-of-the art in biomarker research and development

Considerable progress has been recently made regarding the implementation of biomarkers in cancer therapies, with approximately 55% of all oncology clinical trials in 2018 involving the use of biomarkers as stratification means [1]. In addition, more than 25% of patients with cancer may receive a therapy following a prior biomarker testing [1]. Some examples of biomarkers that are currently used in clinical practice to select patients that will benefit from a therapy include, among others, BRAF V600E or V600K mutations to guide treatment with vemurafenib plus cobimetinib in patients with unresectable or metastatic melanoma [2]; BRCA1/2 mutations to guide treatment with olaparib in breast cancer [3]; RAS mutational status as a negative predictor factor for the benefit of anti-EGFR monoclonal antibody to treat colorectal cancer [4]. Nevertheless, and although there are thousands of studies reporting on biomarker application in oncology, only very few of them have finally achieved success: clinical implementation. One of the main factors attributing to low cancer biomarker clinical applicability appears to be the fact that statistics are often neglected or underappreciated. This becomes evident even when searching in PubMed; of the approximately 300,000 studies reporting in the title or abstract cancer and biomarkers [keywords: (cancer) AND (biomarker*)] only the 10% mention statistics [((cancer) AND (biomarker*)) AND (statistic*)].

3. Statistical considerations

i) Study design and sample power

Several key considerations should be observed to fill the gaps of insufficient or poor use of statistics in biomarker discovery and validation. A correct study design is critical for the successful clinical application of the biomarkers and depends on, among others, the selection of the proper target population, sufficient statistical power, and consideration of the influence of possible confounding variables [5]. Power calculations are needed to ensure an adequate number of samples/events always in relation to the specific clinical context of use, the specific cancer type prevalence and the targeted performance improvement over current standards. Standard operating procedures and sampling standardization must be established prior to initiation sample collection for the study to minimize the impact of experimental/analytical variability. Multiple confounding factors (e.g. sex, age, body mass index, or comorbidities) must also be accounted for to ensure correct estimation of the biomarker value. Statistical approaches, such as inverse probability weighting or Bayesian methods can be used to reduce selection bias to the findings [6]. Biases during patient selection/ specimen collection and patient evaluation can be reduced with randomization and blinding.

ii) Data missingness and false discovery rate control

To decipher the complexity of cancer, omics platforms are often applied for biomarker discovery and validation. Omics aims at the holistic/ collective characterization and simultaneous quantification of multiple molecules depicting structural, functional and dynamic status of an organism at the given timepoint. A common burden in raw omics datasets are missing data as a result of biological (feature is not present in the sample) or technical factors (detection limits of the technology). Although no golden rule exists and the best method to handle the missing data remains controversial, it seems that the optimal approach in fact is treating missing values as what they are: missing, hence the dataset cannot be used for interpretation. However, such approach obviously reduces the number of datasets (and power). Therefore, multiple efforts to develop imputation algorithms for guiding missing value imputation have been made [7], including random forest (RF), k-nearest neighbors imputation (KNN), singular value decomposition based imputation (SVD), Bayesian principal component analysis (BPCA) and others. Omics raw data complexity indicates that frequently nonparametric statistical tests, such as a Wilcoxon rank sum test, should be utilized since in most of the cases omics data do not meet the underlying assumptions for normality (a standard t-test can be applied only if the data follow a normal distribution) [6]. Moreover, given the intrinsic variances particularly for proteomics/ metabolomics, adjustment for multiple testing is required to reduce false positive identifications. Very rigorous correction as suggested by Bonferroni, although apparently ideal, preserves only little statistical power and at the same time may result in no significant findings. Therefore, methods controlling false discovery rate (FDR) like Benjamini-Hochberg correction are widely used. Multiple additional methods controlling the FDR are available [6, 8]. Data filtering, transformation or scaling may be also needed for multivariate modeling and dimensionality reduction. This can be achieved by applying low-dimensional visualizations to the processed data such as principal coordinate analysis, t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) as well as machine learning and deep learning algorithms [9].

iii) Multidimensionality and integrative models

Discovery studies based on omics datasets can result in the identification of numerous biomarker candidates (hundreds or thousands). Frequently statistical significance and fold change between cases and controls are the main criteria for biomarker candidate prioritization. However, not every feature with a change in its distribution between healthy individuals and cancer patients represents a useful cancer biomarker. For instance interleukins are highly upregulated in cancer, but similar upregulation can be found in inflammatory diseases. Consequently, interleukins are generally not used as biomarkers in oncology [10]. Biomarker prioritization depends on the application of statistical methods, machine or deep learning as well as functional enrichment analyses. The high dimensionality of omics data, as a result of the so-called p > > n problem (larger number of omic features than samples needed to detect biologically relevant attributes), is a major challenge as it can result in statistically unstable overfitted models. Machine learning techniques such as logistic regression, random forests and support vector machines can be employed to develop multivariate biomarker panels [6]. An improved performance may be achieved when using a panel of biomarkers over single markers. After demonstrating benefits in well powered trials, multi-gene biomarkers have been already endorsed by ASCO to guide decisions of adjuvant endocrine and chemotherapy in patients with early stage breast cancer [11]. Biomarkers can be also integrated with clinical variables using mathematic formulas to construct predictive models such as nomograms [12]. To develop clinical predictive nomograms logistic regression analysis can be performed to define the clinical characteristics that are correlated with e.g. patients' overall survival and those fulfilling the statistical criteria that can be next selected to develop a nomogram using multivariate COX regression model. Irrespective of the approach, validation of the performance of the model in independent datasets, ideally including patients from multiple clinical centers, must be undertaken prior to clinical implementation.

iv) Performance metrics

One of the main challenges in biomarker research is to distinguish between a potential biomarker and a reliable biomarker that can guide important clinical decisions. To be clinically meaningful and provide guidance as well as a benefit over standard criteria, biomarkers must be specified for a specific context of use. During clinical validation, the performance of the biomarker can be estimated in terms of diagnostic sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), receiver operating characteristics (ROC) curve, and area under the ROC curve (AUC_{ROC}) [13]. Kaplan-Meier survival curves and log rank test can be applied to compare survival differences among the groups, a concordance index (C-index) to estimate the similarity between the true survival time and predicted risk score. For clinical application a cut-off must be defined. Among the different methods that can be chosen to define the cut-off point from a ROC curve, Youden index is highly frequently used. Although high sensitivity and specificity indicate a good biomarker, PPV and NPV represent important probabilities for the successful implementation of the biomarkers. Several blood-based biomarkers have been used in clinical practice such as prostate-specific antigen (PSA) for prostate cancer, carcinoembryonic antigen (CEA) for colorectal cancer, carbohydrate antigen 19-9 (CA19-9) for pancreatic cancer and cancer antigen 125 (CA125) for ovarian cancer [13]. However, not all of them reach the standards of sufficiently high specificity and sensitivity. For example, the application of PSA as a screening tool is being debated for more than a decade, due to low accuracy in distinguishing individuals with benign prostatic hyperplasia from those with malignant prostate cancer. High false-positive rate is also a common issue for many FDA approved urine biomarker assays (e.g. in bladder cancer), thus leading to overdiagnosis [14]. These examples further demonstrate that application of proper statistical analysis is crucial for biomarker research.

4. Conclusions

Nevertheless, even a successfully clinically validated biomarker may not reach implementation at clinical practice; as demonstrated by the large number of available biomarkers that are not yet included in clinical guidelines and/or approved by regulatory agencies. Up to date, many published biomarker studies are inconclusive or not reproducible as a result of dismissing important factors during study design and execution (including statistics). Apart from the neglected statistics, restricted access to appropriate number of specimens, limited funding options, the necessity to validate biomarkers utility in clinical trials as well as the poor communication of all parties involved represent additional challenges [15]. In fact, in many European countries availability, and reimbursement of biomarker tests is restricted, ultimately resulting in avoidable poor outcome (death) for tumor patients [1]. Dedicated strategies and methods to address these challenges have been proposed a decade ago [15]. Following these (and similar) suggestions should enable increased implementation of biomarkers in oncology, expected to significantly improve treatment outcome and reduce mortality in oncology.

References

1. Normanno N, Apostolidis K, Wolf A, et al. (2022) Access and quality of biomarker testing for precision oncology in Europe. Eur J Cancer 176:70-77
**** This article provides the latest update on the current status of biomarkers test quality and access in Europe and their implementation in precision oncology.**

2. Seth R, Messersmith H, Kaur V, et al. (2020) Systemic Therapy for Melanoma: ASCO Guideline. J Clin Oncol 38(33):3947-3970

3. Gennari A, Andre F, Barrios CH, et al. (2021) ESMO Clinical Practice Guideline for the diagnosis, staging and treatment of patients with metastatic breast cancer. Ann Oncol 32(12):1475-1495

4. Cervantes A, Adam R, Rosello S, et al. (2023) Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol 34(1):10-32

5. Ou FS, Michiels S, Shyr Y, et al. (2021) Biomarker Discovery and Validation: Statistical Considerations. J Thorac Oncol 16(4):537-545

6. Nakayasu ES, Gritsenko M, Piehowski PD, et al. (2021) Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. Nat Protoc 16(8):3737-3760
**** This tutorial describes the key considerations for proteomics data processing and statistical analyses that are required to perform meaningful interpretations and ensure the successful implementation of biomarkers.**

7. Wang S, Li W, Hu L, et al. (2020) NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. Nucleic Acids Res 48(14):e83

8. Lualdi M, Fasano M. (2019) Statistical analysis of proteomics data: A review on feature selection. J Proteomics 198:18-26

9. Diaz-Uriarte R, Gomez de Lope E, Giugno R, et al. (2022) Ten quick tips for biomarker discovery and validation analyses using machine learning. PLoS Comput Biol 18(8):e1010357
**** This educational article provides several tips that can be applied to address the current limitations and risks in biomarker discovery and validation.**

10. Briukhovetska D, Dorr J, Endres S, et al. (2021) Interleukins in cancer: from biology to therapy. Nat Rev Cancer 21(8):481-499

11. Andre F, Ismaila N, Allison KH, et al. (2022) Biomarkers for Adjuvant Endocrine and Chemotherapy in Early-Stage Breast Cancer: ASCO Guideline Update. J Clin Oncol 40(16):1816-1837

12. Hendrix SB, Mogg R, Wang SJ, et al. (2021) Perspectives on statistical strategies for the regulatory biomarker qualification process. Biomark Med 15(9):669-684

13. Sarhadi VK, Armengol G. (2022) Molecular Biomarkers in Cancer. Biomolecules 12(8)

14. Ng K, Stenzl A, Sharma A, et al. (2021) Urinary biomarkers in bladder cancer: A review of the current landscape and future directions. Urol Oncol 39(1):41-51

15. Mischak H, Ioannidis JP, Argiles A, et al. (2012) Implementation of proteomic biomarkers: making it work. Eur J Clin Invest 42(9):1027-1036

**** In this perspective dedicated strategies and methods are proposed to address the limitations and challenges of biomarkers implementation in clinical practice.**

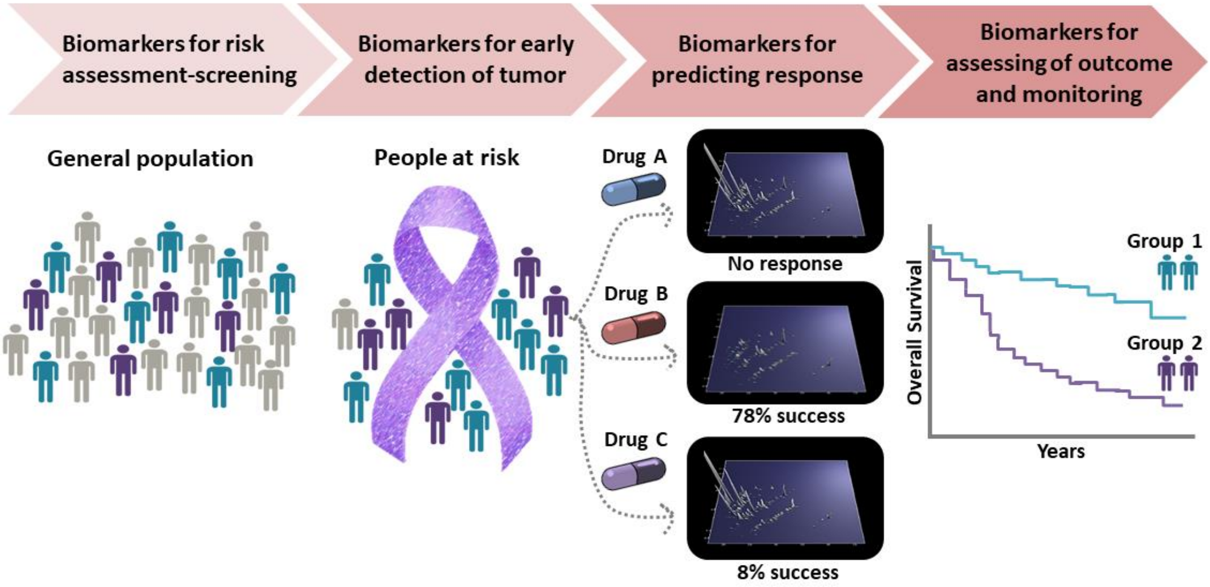
Biomarkers for risk assessment-screening
Biomarkers for early detection of tumor
Biomarkers for predicting response
Biomarkers for assessing of outcome and monitoring
General population
People at risk
Drug A
No response
Drug B
78% success
Drug C
8% success
Group 1
Group 2
Overall Survival
Years

Figure 1: Cancer biomarkers can assist on risk estimation, detection of a tumor or its recurrence, prediction of response to available treatments for the specific cancer type and assessment of treatment outcome.