BMC Medical Research
Methodology

RESEARCH Open Access

A scoping review of statistical methods in studies of biomarker-related treatment heterogeneity for breast cancer

L Sollfrank[1], SC Linn[2,3,4], M Hauptmann[1] and K Jóźwiak[1*]

Abstract

Background Many scientific papers are published each year and substantial resources are spent to develop biomarker-based tests for precision oncology. However, only a handful of tests is currently used in daily clinical practice, since development is challenging. In this situation, the application of adequate statistical methods is essential, but little is known about the scope of methods used.

Methods A PubMed search identified clinical studies among women with breast cancer comparing at least two different treatment groups, one of which chemotherapy or endocrine treatment, by levels of at least one biomarker. Studies presenting original data published in 2019 in one of 15 selected journals were eligible for this review. Clinical and statistical characteristics were extracted by three reviewers and a selection of characteristics for each study was reported.

Results Of 164 studies identified by the query, 31 were eligible. Over 70 different biomarkers were evaluated. Twenty-two studies (71%) evaluated multiplicative interaction between treatment and biomarker. Twenty-eight studies (90%) evaluated either the treatment effect in biomarker subgroups or the biomarker effect in treatment subgroups. Eight studies (26%) reported results for one predictive biomarker analysis, while the majority performed multiple evaluations, either for several biomarkers, outcomes and/or subpopulations. Twenty-one studies (68%) claimed to have found significant differences in treatment effects by biomarker level. Fourteen studies (45%) mentioned that the study was not designed to evaluate treatment effect heterogeneity.

Conclusions Most studies evaluated treatment heterogeneity via separate analyses of biomarker-specific treatment effects and/or multiplicative interaction analysis. There is a need for the application of more efficient statistical methods to evaluate treatment heterogeneity in clinical studies.

Keywords Predictive, Biomarker, Treatment heterogeneity, Interaction, Breast cancer, Review, Statistical methods

*Correspondence:
K Jóźwiak
Katarzyna.Jozwiak@mhb-fontane.de
[1] Institute of Biostatistics and Registry Research, Brandenburg Medical School Theodor Fontane, Fehrbelliner Straße 39, Neuruppin 16816, Germany
[2] Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[3] Department of Medical Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[4] Department of Pathology, University Medical Center, Utrecht, The Netherlands

Introduction

Breast cancer (BC) is the most common cancer among women worldwide [1]. In Europe, one in 11 women is diagnosed with BC at least once in her life [2]. Although 5-year survival rates after BC diagnosis increased steadily during the last decades and currently exceed 80% [2], modern treatments still fail in many patients causing significant morbidity and mortality.

At the same time, thousands of scientific papers are published annually and substantial resources are spent to develop tests for precision medicine in oncology. These tests are usually based on biomarkers, e.g., characteristics measurable in healthy or tumor tissue which influence a patient's clinical outcome. Prognostic biomarkers describe the likelihood of a future recurrence or progression of cancer, i.e., they identify patients who require additional systemic therapy besides local therapy like surgery or radiotherapy - they indicate who needs additional therapy. Predictive biomarkers identify patients who are more likely to respond to a treatment, i.e., they select the most promising treatment for a specific patient - they indicate how one should be treated [3]. Therefore, predictive biomarkers are essential for personalized medicine and are a topic of current research, e.g., mutations in the BRCA1 gene [4, 5] or tumor infiltrating lymphocytes (TILs) [6, 7]. However, only a handful of these tests are currently being used in clinical practice, for example, presence/absence of human epidermal growth factor 2 (HER2) gene amplification or estrogen receptor (ER) in BC [8, 9]. The reason is that a candidate biomarker has to pass several stages of development. Perhaps the most challenging stage is the translation from a convincing preclinical test to using the same test in patients with cancer in daily practice. To demonstrate clinical utility of a test, series of patients are required who received the treatment of interest or an alternative treatment and have positive or negative test results.

Obtaining conclusive results in such studies depends on the choice of the study design and the statistical method for data analysis. Several guidelines for designing biomarker studies are available [10–14]. A commonly used statistical technique to evaluate a predictive biomarker is a test for interaction between biomarker and treatment in a cohort of suitable patients. This approach aims to evaluate, whether a relative benefit of a specific experimental treatment compared with a control treatment differs by biomarker level. An example of such a test is a comparison of the benefit of adjuvant tamoxifen versus no tamoxifen on the risk of BC recurrence between ER positive and negative disease [15]. In this context, there is an important distinction for clinical utility between "quantitative" and "qualitative" interaction. If a new treatment benefits all patients relative to standard, but a biomarker only associates with magnitude ("quantitative" interaction) but not direction of the effect ("qualitative" interaction), then the predictive biomarker is not likely to alter therapy choice if both effect sizes are clinically meaningful, and hence the marker is not clinically useful [16]. In statistical terms, the interaction analysis evaluates departure from a multiplicative model for the joint relative effect of biomarker and treatment on the outcome. However, interaction analyses are known to require large series of patients [17, 18], which may not be available, or performing many measurements may be too expensive.

The spectrum of statistical methods commonly used in studies evaluating biomarkers for BC treatment heterogeneity is unknown. Such knowledge, however, is essential to determine whether developing or using alternative statistical methods offers an opportunity to advance precision medicine for BC. We therefore provide a methods review of a representative sample of observational and randomized studies on predictive biomarkers for BC. We focus on study designs, statistical methods and sample sizes.

Methods

The study complied with reporting recommendations for scoping reviews (PRISMA-ScR) criteria [19], although a protocol does not exist.

Eligibility criteria

Eligible for our review were studies among female patients with BC comparing at least two different treatment groups, one of which chemotherapy or endocrine treatment, by levels of at least one biomarker. Reviews and other reports without original data were ineligible.

Search strategy

A query was developed for a literature search of publications written in English and available in PubMed (Additional File 1). The query was then limited to 2019 as year of publication and to the following 15 journals:

> *Annals of Oncology, Breast Cancer Research, Breast Cancer Research and Treatment, Clinical Cancer Research, International Journal of Cancer, Journal of the American Medical Association, Journal of the American Medical Association Oncology, Journal of Clinical Oncology, Journal of the National Cancer Institute, The Lancet, Lancet Oncology, Molecular Cancer Therapeutics, Nature Medicine, New England Journal of Medicine and PloS One.*

Full text versions of all identified articles were obtained. Based on a title and abstract review by at least two authors (LS, MH, KJ), ineligible articles were identified and excluded. Supplements, if any, were obtained for eligible articles.

Data extraction

For eligible articles, we abstracted the following information: first author, journal, type of study, patient inclusion criteria, endpoint definition(s), biomarker(s) analyzed, experimental and standard treatment, total number of patients included in the original study and in the analyses of treatment effects by biomarker level, number of patients and events by biomarker level and treatment, median follow-up time, details about the statistical analysis and results.

Abstraction was done independently by at least two authors (LS, MH, KJ). Discordance was resolved by consensus. Extracted data were checked by three authors (LS, MH, KJ) for completeness. Extracted data were summarized in Additional File 2: Supplementary Tables 1–3.

Since several studies reported results for various biomarkers and/or endpoints, the description in this report was a selection. We presented results for up to two biomarkers reported in the article abstracts. If the abstract mentioned more than two biomarkers, we chose the two which appeared most important based on the amount of details provided and the strengths of the results. We omitted biomarkers that were only reported in the articles' supplements. We summarized results in the total patient group and up to one subgroup mentioned in the abstract. If different types of endpoints were reported (binary and survival endpoints), we described them separately. We reported up to two time-to-event endpoints. We reported hazard ratios (HR) and odds ratios (OR) for effects of treatment, biomarker and/or interaction of both, as well as p-values for the interaction coefficients. For significant interaction tests, we reported whether the interaction was quantitative (i.e., the treatment effects differed in magnitude but not direction between marker levels or the effects of a continuous marker differed in magnitude but not direction between treatment groups) or qualitative (i.e., the treatment effects differed in direction between marker levels or the effects of a continuous marker differed in direction between treatment groups).

Results

The query performed on January 15, 2021, identified 7,243 articles of which 1,830 were published in the selected 15 journals. Of those, 164 articles were published in 2019 in 10 of the 15 journals (none published in the *Journal of the American Medical Association, The Lancet, Lancet Oncology, Nature Medicine* and *New England Journal of Medicine*). We excluded 132 of the 164 articles during title and abstract review and one during full text review (interaction described, but no results presented [20]), leaving 31 articles for description in this report (Fig. 1). These 31 articles were published in 6 of the selected journals, with 11 (35%) published in *Breast Cancer Research and Treatment.* Sample sizes ranged from 42 to 3,746 patients. One half of the studies included less than 367 patients, while one fourth included less than 175 patients. On average, these analyses included 52% of the original sample size. Median follow-up time, if reported, was mostly between 5 and 10 years. Most studies (23 of 31 (74%)) used patient data from randomized controlled trials and 20 (87%) of these studies used archived specimens. Time-to-event endpoints (progression or mortality) were evaluated in 25 of the 31 studies (81%), 7 studies (23%) evaluated binary endpoints (usually pathologic complete response, pCR). About half of the studies evaluated 1 biomarker, while 7 (23%) studies evaluated 5 or more with a maximum of 18 biomarkers in one study [21]. Similarly, while most studies evaluated treatment heterogeneity for only one endpoint, about one third used two or more (Table 1, Additional File 2: Supplementary Table 3).

Clinical characteristics

Studies were diverse in terms of types of biomarkers, treatments and patient selection. In total, over 70 different biomarkers were evaluated. Biomarkers described in at least two articles were age, ER, progesterone receptor status (PR), HER2, stromal TILs (sTILs) and PIK3CA. Biomarkers were often used in different parametrizations, i.e., continuously and categorically, or with different categorizations.

Tamoxifen treatment was investigated in 6 studies and was compared to either a treatment with an aromatase inhibitor or no/less tamoxifen. Four studies reported on the effect of adjuvant chemotherapy (ACT) in comparison to no ACT. Three studies compared trastuzumab with no trastuzumab and two studies described the AKT inhibitor and anti-HER2 tyrosine kinase inhibitors. Treatments in all other studies were unique. Patient selection was based on the tumor, node, metastasis (TNM) staging system in 23 of 31 studies. Twenty-six studies used hormone receptor status (HoR) and 6 selected patients based on age. Moreover, 9 studies used only one criterion for patient selection, while the remaining 22 studies used specific combinations of patients and tumor characteristics.

Statistical characteristics

The majority of articles used one of two common statistical approaches to evaluate heterogeneity of treatment outcome by biomarker level. The first approach, evaluating the significance of a multiplicative interaction term between treatment and biomarker in a regression model with individual terms for treatment and biomarker, was used in 21 studies (Table 2). In 15 of these studies (68%),

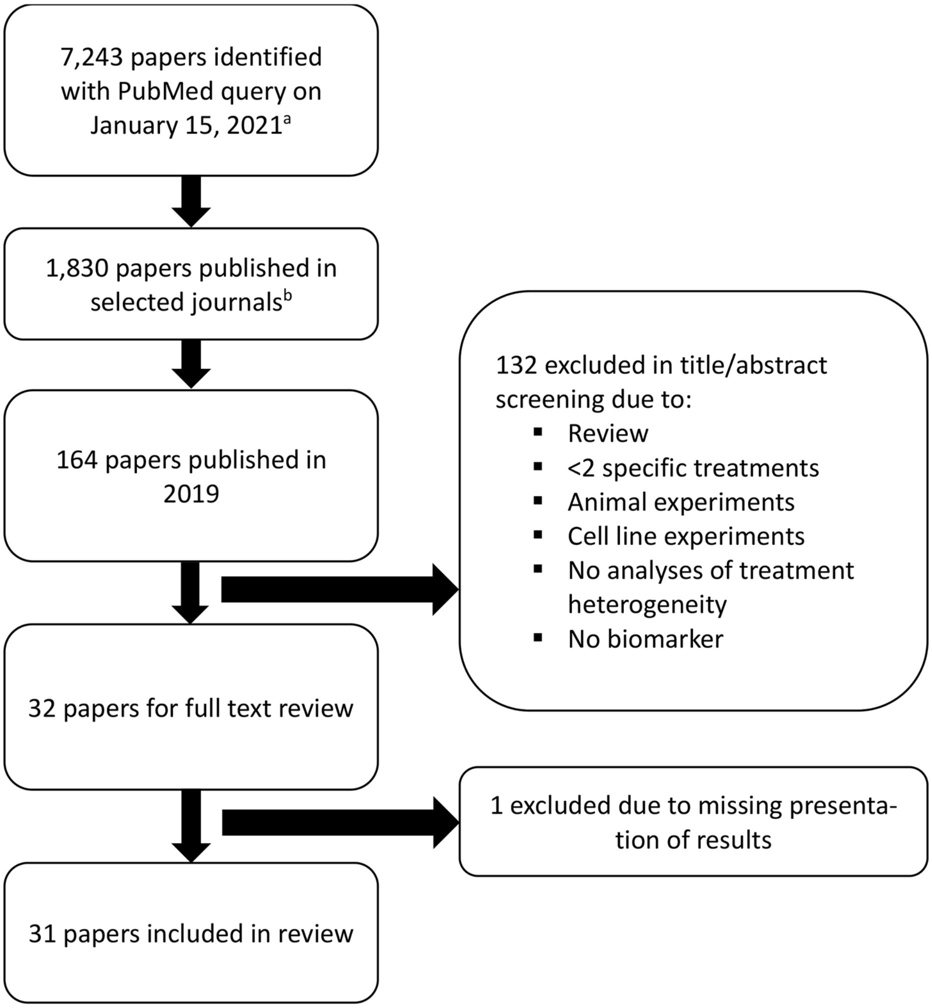
7,243 papers identified with PubMed query on January 15, 2021[a]
1,830 papers published in selected journals[b]
164 papers published in 2019
32 papers for full text review
31 papers included in review
132 excluded in title/abstract screening due to:
Review
<2 specific treatments
Animal experiments
Cell line experiments
No analyses of treatment heterogeneity
No biomarker
1 excluded due to missing presentation of results

Fig. 1 Flow chart of selection of relevant articles for the review

[a] ("Breast Neoplasms"[Majr] OR ((breast[tiab] OR mammary[tiab]) AND (neoplas*[tiab] OR cancer*[tiab] OR tumor*[tiab] OR malignan*[tiab] OR oncolog*[tiab]))) AND (heterogeneity[TIAB] OR effect[TIAB] OR predict*[TIAB] OR prognostic[TIAB] OR interaction[TIAB]) AND (marker* OR biomarker*) AND (cohort[TIAB] OR patient*[TIAB] OR female[TIAB] OR women[TIAB]) AND (endocrine OR chemotherapy OR neoadjuvant)

[b] Ann Oncol, Breast Cancer Res, Breast Cancer Res Treat, Clin Cancer Res, Int J Cancer, JAMA, JAMA Oncol, J Clin Oncol, J Natl Cancer Inst, Lancet, Lancet Oncol, Mol Can Ther, Nat Med, N Engl J Med, PloS one

only a Cox proportional hazards model for a time-to-event outcome was applied, in 5 studies (23%) only a logistic regression model for a binary outcome was used and 1 study presented results from interaction analyses with binary as well as time-to-event endpoints [21–41]. The second common approach, subgroup analysis of relative treatment effects by biomarker levels, was additionally used in 12 studies with time-to-event endpoints and 4 studies with binary endpoints. Two articles with time-to-event endpoints presented interaction analyses and relative biomarker effects by treatment subgroups. One study performed interaction analyses and both types of subgroup analyses for time-to-event as well as binary endpoints [30]. Eight studies showed only results from subgroup analyses: 3 evaluated the treatment effect by biomarker subgroups [42–44], 4 the biomarker effect by treatment subgroups [46–49] and 1 only presented Kaplan-Meier curves for recurrence-free survival for all biomarker-treatment combinations [50]. Thus, there were 20, 7 and 1 studies that performed subgroup analysis for treatment effect by biomarker subgroup, biomarker effect by treatment subgroup and all biomarker-treatment subgroups, respectively (Table 2).

Two studies showed Subpopulation Treatment Effect Pattern Plots (STEPP) in addition to results from a standard multiplicative Cox proportional hazards model with interaction terms and subgroup analysis [31, 41]. STEPP is a graphical tool which divides patients into partially overlapping subpopulations based on subsequent values of a continuous biomarker. For each subpopulation,

Table 1 Characteristics of the study design of selected studies of biomarker-based treatment heterogeneity

Characteristic	**Number (%[a])**
Journal[b]	
Ann Oncol	4 (13)
Breast Cancer Res	3 (10)
Breast Cancer Res Treat	11 (35)
Clin Cancer Res	3 (10)
Int J Cancer	5 (16)
J Clin Oncol	5 (16)
Number of patients included in analyses of treatment heterogeneity	
< 100	3 (10)
100–300	10 (32)
300–500	4 (13)
> 500	14 (45)
Minimum	42
25% quantile	175
Mean (SE)	851 (192)
Median	367
75% quantile	858
Maximum	3746
Type of study	
Randomized	23 (74)
Observational	8 (26)
Endpoints for which treatment heterogeneity was evaluated[c]	
Binary	7 (23)
pCR	6 (19)
> 50% relative decrease in 11-gene proliferation signature	1 (3)
Time to event	25 (81)
Time to progression[d]	25 (81)
Time to death[e]	9 (29)
Number of biomarkers for which treatment heterogeneity was evaluated	
1	14 (45)
2–4	10 (32)
5–10	5 (16)
> 10	2 (7)
Number of endpoints for which treatment heterogeneity was evaluated	
1	21 (68)
2	7 (23)
3	3 (10)
Median follow up time	
≤ 5 years	4 (13)
5–10 years	9 (29)
> 10 years	2 (6)
Not applicable	6 (19)
Not reported	10 (32)

[a] Percentage do not sum up to 100, if a study qualified for several categories or due to rounding

[b] Journals with no relevant papers were omitted

[c] One study analyzed binary as well as survival endpoints

[d] Time to progression includes: BCFI: invasive breast cancer-free interval; BCFS: breast cancer-free survival, DDFS: distant-disease-free survival; DFS: disease-free survival; DRFI: distant recurrence-free interval; EFS: event-free survival; iDFS: invasive disease-free survival; MRFS: metastatic recurrence-free survival; PFS: progression-free survival; RFI: recurrence-free interval; RFS: recurrence-free survival

[e] Time to death includes: BCSS: breast cancer-specific survival, OS: overall survival

Table 2 Characteristics of statistical analysis of selected studies of biomarker-based treatment heterogeneity

Characteristic	**Number (%[a])**
Type of analysis	
Cox model with biomarker-treatment interaction[b]	16 (52)
Logistic model with biomarker-treatment interaction[b]	6 (19)
Subgroup analysis[b]	
Treatment effect by biomarker subgroup	20 (65)
Biomarker effect by treatment subgroup	7 (23)
All biomarker-treatment subgroups	1 (3)
Subpopulation Treatment Effect Pattern Plot	2 (6)
Bayesian covariate-adjusted logistic model	1 (3)
Difference in cumulative incidence at one time point	1 (3)
Test of interaction	
Yes	22 (71)
No	9 (29)
At least one significant result	
Yes	21 (68)
No	10 (32)

[a] Percentages do not sum up to 100, if a study qualified for several categories or due to rounding

[b] One study presented interaction models for binary as well as time-to-event endpoints, and additionally biomarker effects by treatment subgroup and treatment effects by biomarker subgroup and is included in more than one category

results such as the treatment-related HR, the cumulative incidence or the absolute risk difference are calculated and then plotted against the midpoints of the biomarker values in corresponding subpopulations. The plot can reveal biomarker values for which the experimental treatment is superior to the alternative treatment [51]. One of the two studies performing STEPP additionally included a statistical test to evaluate whether there was a change in the HR over the different biomarker intervals [41].

One article used a Bayesian covariate-adjusted logistic model [52]. The study evaluated the pCR after treatment with oral pan-AKT inhibitorMK-2206 in subgroups of patients defined by biomarkers HER2, HoR and MammaPrint status. Patients within each biomarker combination were randomly and adaptively assigned to MK-2206 or control arms. The treatment was considered superior in a subgroup of patients with a particular biomarker combination when it had≥85% Bayesian predictive probability of success in a hypothetical phase III trial.

In addition to results from a Cox proportional hazards model with interaction term and subgroup analysis, one study presented differences in the cumulative incidence obtained with a Kaplan-Meier method at a particular time point between two treatment arms separately by biomarker subgroups to illustrate an absolute treatment effect [24].

While 8 of the 31 articles reported results for only one analysis with regard to a predictive biomarker, many studies performed multiple evaluations, mostly for several biomarkers, but also for several outcomes and subpopulations (Additional File 2: Supplementary Table 3). Five studies performed 10 or more analyses on the same patients [21, 30, 35, 40, 41]. The maximum number of analyses performed was 20 in a study with 18 different biomarkers [21]. Only 3 of the studies with more than one analysis adjusted their results for multiple testing [40, 41, 46]. The adjustment in all three studies was done by controlling the family-wise type 1 error rate via a Benjamini-Hochberg correction.

Results of individual studies

Of the 31 reviewed studies, 21 (68%) claimed to have found evidence that at least one evaluated biomarker was predictive, i.e., treatment effects differed by biomarker level (Table 2). This conclusion was based on a significant interaction test in 13 of those studies (of which 11 referred to a qualitative interaction), while 8 studies used significant p-values from subgroup analyses as evidence of treatment heterogeneity. The remaining 10 studies (32%) concluded that there was no evidence of treatment effect heterogeneity by biomarker levels. The number of significant results per study was generally low and did not strongly depend on the number of analyses conducted within the study (Spearman correlation coefficient=0.34, p=0.060; Additional File 2: Supplementary Table 3). Nearly half (14) of the articles mentioned that the study was not designed for the objective and was too small. Seven of these studies presented significant findings for at least one biomarker.

Discussion

This review shows that there were numerous studies of predictive biomarkers evaluating treatment heterogeneity among patients with BC. Most of these studies described biomarker-specific treatment effects in subgroup analyses and/or performed interaction analyses in standard multiplicative models.

The evaluation of treatment heterogeneity is usually not a primary objective of randomized clinical trials. The trials are powered to evaluate main effects of treatment, and are usually underpowered to evaluate interactions, since the sample size required for interaction analyses with adequate power can be much higher than that for main effects. For example, Brookes et al. [18] showed that detection of an interaction with adequate statistical power may require a 4-fold sample size compared with that for evaluation of a main effect of the same magnitude. Moreover, the sample size of biomarker-based analyses is often even smaller than that of the primary analysis (in our review 48% smaller, on average) due to the failure to locate tissue samples, assay failure and quality control exclusions, when archived specimen are used. Besides sample size, statistical power depends on many factors: the baseline event rate (i.e., for the group of patients with low biomarker level and standard treatment), the proportions of patients by treatment and biomarker level, and either (i) the biomarker effect in one of the treatment subgroups, the treatment effect in one of the biomarker subgroups and the interaction effect, or (ii) the marginal effects of biomarker and treatment and the interaction effect. Methods and software tools are available to support the design of treatment heterogeneity analyses prior to the study [17, 53, 54]. It is unclear what the statistical power of the treatment heterogeneity analyses in the reviewed studies was when they were planned. It may not have been determined at all at that time since treatment heterogeneity was likely not the primary objective of most studies. Calculating post-hoc power based on observed data is meaningless [55]. Nevertheless, a test of interaction is required to rigorously assess whether treatment effects are different in biomarker subgroups [56, 57]. Evaluating only treatment effects in separate subgroups may lead to erroneous conclusions [56, 58], for example, the same treatment effect is observed in both biomarker subgroups but due to a difference in sample sizes, the effect is significant only in one of the subgroups.

Half of the reviewed studies analyzed more than one biomarker and/or endpoint, but only 3 studies corrected for multiple testing [40, 41, 46] in order to reduce the probability of false positive findings. This is important since the chance of false positive conclusions is already increased for underpowered studies [59].

Several guidelines provide suggestions for the analysis of treatment heterogeneity [10–14]. The most specific guideline for this subject is the PATH statement [60, 61], which recommends evaluating variation in the treatment effect by biomarker subgroup via interaction analysis. However, the PATH statement distinguishes between (i) risk modeling, i.e., combining all available prognostic factors into a prognostic score and evaluating interaction between this score and treatment and (ii) effect modeling, i.e., evaluating interactions between treatment and single predictive biomarkers [60]. None of the studies in this review performed risk modeling. Moreover, the PATH statement emphasizes that analyses of heterogeneity of relative effects can lead to different results than comparing absolute effects, i.e., risk differences, and the latter approach is recommended [60, 61]. All studies in this review, however, used relative effect measures, i.e., HRs or ORs, to investigate treatment heterogeneity on a multiplicative scale. Only one study presented, in addition to relative effects, an absolute measure of treatment benefit, namely the difference between the reduction in risk of recurrence [24]. Using relative effect measures is convenient because it is implemented in statistical software. However, absolute effect measures, i.e., risk differences, are considered to be more relevant for clinical decision making, and the evaluation of the treatment heterogeneity should therefore be based on absolute effects [60, 61].

Our review has several limitations. We restricted the detailed evaluation of studies to those published in 2019 in a subgroup of selected scientific journals. The initial selection of journals was based on our knowledge of relevant articles published in different years, with impact factors for 2020 ranging from 4.8 to 12.5. We added journals with higher and lower impact factors, whose scope included studies on predictive biomarkers for BC. While admittedly subjective, we believe that this sample of recent research in peer-reviewed established journals allows general conclusions about the entire field of BC research. A cursory review of publications in other journals and other years confirmed our results. We also assume that results on BC are to some extent generalizable to cancers at other sites.

Another potential limitation is that we may have missed studies with rarely used statistical methods. Most of the studies we reviewed used either biomarker-specific Kaplan-Meier plots of treatment effects and/or multiplicative interaction analyses. Only 3 studies used other methods (STEPP and a Bayesian covariate-adjusted logistic model). Alternative methods have been described in the literature but were not applied in the reviewed studies, e.g., the predictiveness curve for a continuous marker [62] or the metric theta which measures a difference in the disease rate under biomarker-based treatment assignment versus the default strategy of the same treatment for all patients [63]. Therefore, our review possibly does not capture the spectrum of less commonly used methods. However, finding all applied methods was not the goal of our work.

The strengths of our review include the comprehensive search which took into account all types of biomarkers and the thorough evaluation of eligible studies with regard to methodological features. Therefore, the importance of decent study design and adequate sample size is stressed.

The results of our review illustrate an important bottleneck in the development of new predictive tests. A new candidate test has to pass several stages of development and the most difficult step is the transition from a promising preclinical test to a test which can be applied to patients with cancer in daily practice. Series of patients are required to demonstrate the clinical utility of a test, including those who received the treatment of interest or an alternative treatment, and including patients with positive and negative test results. The large series required by current statistical methods are often not available, and if they are available, limited research budgets prohibit performing the test of interest. Consequently, too small patient series are interrogated leading to inconclusive results. It is likely that many promising test are erroneously abandoned at this stage of development.

Next to careful planning of biomarker-based studies of treatment heterogeneity, further research is necessary on statistical methods which allow evaluation of candidate predictive biomarkers with smaller numbers of patients than currently required for adequate statistical power. Case-only, hybrid designs or additive models may offer opportunities.

Conclusions

This review shows that BC studies of predictive biomarkers are usually evaluated by separately estimating treatment effect in biomarker subgroup or by testing a multiplicative interaction term between biomarker and treatment with a regression analysis. These analyses may be underpowered because the studies are designed to investigate main treatment effects and biomarker data is often not available for all patients included in the study. Therefore, there is a need for further research on more powerful statistical methods which can be applied to small studies on predictive biomarkers.

Abbreviations

ACT	adjuvant chemotherapy
BC	breast cancer
ER	estrogen receptor
HER2	human epidermal growth factor 2
HoR	hormone receptor
HR	hazard ratio
OR	odds ratio
pCR	pathological complete response
PR	progesterone receptor
STEPP	subpopulation treatment effect pattern plot
TILs	tumor infiltrating lymphocytes
TNM	tumor, node, metastasis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-023-01982-w.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

None.

Authors' contributions

LS conducted literature search, data extraction, statistical analyses and drafted the manuscript. SCL contributed to the conception of the review and edited the manuscript. MH and KJ contributed to the conception of the review, conducted data extraction, edited the manuscript and provided overall supervision and coordination of the manuscript preparation. The final version was reviewed and approved by all authors.

Funding

This work was supported by the Dutch Cancer Society, grant number KWF 10603.

Open Access funding enabled and organized by Projekt DEAL.

Data Availability

The data supporting the findings of this study are available within the reviewed articles that are publically available.

Declarations

Competing interests

None declared.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Received: 8 November 2022 / Accepted: 19 June 2023

Published online: 29 June 2023

References

1. WHO webpage [Internet]. [cited 2022 Feb 17]. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
2. European Cancer Information System [Internet]. [cited 2022 Feb 17]. Available from: https://ecis.jrc.ec.europa.eu/.
3. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Silver Spring (MD): Food and Drug Administration (US). ; 2016. Available from: https://www.ncbi.nlm.nih.gov/books/NBK326791/ Co-published by National Institutes of Health (US), Bethesda (MD).
4. Robson M, Im S-A, Senkus E, Xu B, Domchek SM, Masuda N, et al. Olaparib for metastatic breast cancer in patients with a germline *BRCA* mutation. N Engl J Med. 2017;377:523–33.
5. Tutt ANJ, Garber JE, Kaufman B, Viale G, Fumagalli D, Rastogi P, et al. Adjuvant olaparib for patients with BRCA1- or BRCA2-mutated breast cancer. N Engl J Med. 2021;384:2394–405.
6. Ignatiadis M, Van den Eynden G, Roberto S, Fornili M, Bareche Y, Desmedt C, et al. Tumor-infiltrating lymphocytes in patients receiving trastuzumab/pertuzumab-based chemotherapy: a TRYPHAENA substudy. J Natl Cancer Inst. 2019;111:69–77.
7. El Bairi K, Haynes HR, Blackley E, Fineberg S, Shear J, Turner S, et al. The tale of TILs in breast cancer: a report from the International Immuno-Oncology Biomarker Working Group. npj Breast Cancer. 2021;7:150.
8. La Thangue NB, Kerr DJ. Predictive biomarkers: a paradigm shift towards personalized cancer medicine. Nat Rev Clin Oncol. 2011;8:587–96.
9. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019;30:1194–220.
10. 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.
11. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, et al. Criteria for the use of omics-based predictors in clinical trials. Nature. 2013;502:317–20.
12. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. J Natl Cancer Inst. 2009;101:1446–52.
13. Freidlin B, Sun Z, Gray R, Korn EL. Phase III clinical trials that integrate treatment and biomarker evaluation. J Clin Oncol. 2013;31:3158–61.
14. Korn EL, McShane LM, Freidlin B. Statistical challenges in the evaluation of treatments for small patient populations. Sci Transl Med. 2013;5:178sr3.
15. Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Davies C, Godwin J, Gray R, Clarke M, Cutter D, et al. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet. 2011;378:771–84.
16. Polley M-YC, 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(22):1677–83.
17. García-Closas M, Lubin JH. Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. Am J Epidemiol. 1999;149:689–92.

18. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess. 2001;5:1–56.
19. Tricco AC, Lillie E, Zarin W, O'Brien KK, Coloquhoun H, Levac D, et al. PRISMA Extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467–73.
20. Lambertini M, Campbell C, Bines J, Korde LA, Izquierdo M, Fumagalli D, et al. Adjuvant anti-HER2 therapy, treatment-related amenorrhea, and survival in premenopausal HER2-positive early breast cancer patients. J Natl Cancer Inst. 2019;111:86–94.
21. Gluz O, Kolberg-Liedtke C, Prat A, Christgen M, Gebauer D, Kates R, et al. Efficacy of deescalated chemotherapy according to PAM50 subtypes, immune and proliferation genes in triple-negative early breast cancer: primary translational analysis of the WSG-ADAPT-TN trial. Int J Cancer. 2020;146:262–71.
22. Adamo B, Bellet M, Paré L, Pascual T, Vidal M, Pérez Fidalgo JA, et al. Oral metronomic vinorelbine combined with endocrine therapy in hormone receptor-positive HER2-negative breast cancer: SOLTI-1501 VENTANA window of opportunity trial. Breast Cancer Res. 2019;21:108.
23. Asleh K, Lyck Carstensen S, Tykjaer Jørgensen CL, Gao D, Won JR, Jensen M-B, et al. Basal biomarkers nestin and INPP4B predict gemcitabine benefit in metastatic breast cancer: samples from the phase III SBG0102 clinical trial. Int J Cancer. 2019;144:2578–86.
24. Bartlett JMS, Sgroi DC, Treuner K, Zhang Y, Ahmed I, Piper T, et al. Breast Cancer Index and prediction of benefit from extended endocrine therapy in breast cancer patients treated in the adjuvant Tamoxifen-To offer more? (aTTom) trial. Ann Oncol. 2019;30:1776–83.
25. Camp NJ, Madsen MJ, Herranz J, Rodríguez-Lescure Á, Ruiz A, Martín M, et al. Re-interpretation of PAM50 gene expression as quantitative tumor dimensions shows utility for clinical trials: application to prognosis and response to paclitaxel in breast cancer. Breast Cancer Res Treat. 2019;175:129–39.
26. Chia SKL, Martin M, Holmes FA, Ejlertsen B, Delaloge S, Moy B, et al. PIK3CA alterations and benefit with neratinib: analysis from the randomized, double-blind, placebo-controlled, phase III ExteNET trial. Breast Cancer Res. 2019;21:39.
27. Chumsri S, Li Z, Serie DJ, Mashadi-Hossein A, Colon-Otero G, Song N, et al. Incidence of late relapses in patients with HER2-positive breast cancer receiving adjuvant trastuzumab: combined analysis of NCCTG N9831 (Alliance) and NRG Oncology/NSABP B-31. J Clin Oncol. 2019;37:3425–35.
28. Dieci MV, Conte P, Bisagni G, Brandes AA, Frassoldati A, Cavanna L, et al. Association of tumor-infiltrating lymphocytes with distant disease-free survival in the ShortHER randomized adjuvant trial for patients with early HER2+ breast cancer. Ann Oncol. 2019;30:418–23.
29. Hamy A-S, Tury S, Wang X, Gao J, Pierga J-Y, Giacchetti S, et al. Celecoxib with neoadjuvant chemotherapy for breast cancer might worsen outcomes differentially by COX-2 expression and ER status: exploratory analysis of the REMAGUS02 trial. J Clin Oncol. 2019;37:624–35.
30. Janning M, Müller V, Vettorazzi E, Cubas-Cordova M, Gensch V, Ben-Batalla I, et al. Evaluation of soluble carbonic anhydrase IX as predictive marker for efficacy of bevacizumab: a biomarker analysis from the geparquinto phase III neoadjuvant breast cancer trial: evaluation of sCAIX as predictor in breast cancer. Int J Cancer. 2019;145:857–68.
31. Kensler KH, Regan MM, Heng YJ, Baker GM, Pyle ME, Schnitt SJ, et al. Prognostic and predictive value of androgen receptor expression in postmenopausal women with estrogen receptor-positive breast cancer: results from the Breast International Group Trial 1–98. Breast Cancer Res. 2019;21:30.
32. Kruger DT, Alexi X, Opdam M, Schuurman K, Voorwerk L, Sanders J, et al. IGF-1R pathway activation as putative biomarker for linsitinib therapy to revert tamoxifen resistance in ER-positive breast cancer. Int J Cancer. 2020;146:2348–59.
33. Loibl S, Untch M, Burchardi N, Huober J, Sinn BV, Blohmer J-U, et al. A randomised phase II study investigating durvalumab in addition to an anthracycline taxane-based neoadjuvant therapy in early triple-negative breast cancer: clinical results and biomarker analysis of GeparNuevo study. Ann Oncol. 2019;30:1279–88.
34. Loibl S, Treue D, Budczies J, Weber K, Stenzinger A, Schmitt WD, et al. Mutational diversity and therapy response in breast cancer: a sequencing analysis in the neoadjuvant GeparSepto trial. Clin Cancer Res. 2019;25:3986–95.
35. MYME investigators, Nanni O, Amadori D, De Censi A, Rocca A, Freschi A, et al. Metformin plus chemotherapy versus chemotherapy alone in the first-line treatment of HER2-negative metastatic breast cancer. The MYME randomized, phase 2 clinical trial. Breast Cancer Res Treat. 2019;174:433–42.
36. Nitz U, Gluz O, Clemens M, Malter W, Reimer T, Nuding B, et al. West German Study PlanB trial: adjuvant four cycles of epirubicin and cyclophosphamide plus docetaxel versus six cycles of docetaxel and cyclophosphamide in HER2-negative early breast cancer. J Clin Oncol. 2019;37:799–808.
37. Puppe J, Opdam M, Schouten PC, Jóźwiak K, Lips E, Severson T, et al. EZH2 is overexpressed in BRCA1-like breast tumors and predictive for sensitivity to high-dose platinum-based chemotherapy. Clin Cancer Res. 2019;25:4351–62.
38. Sestak I, Martín M, Dubsky P, Kronenwett R, Rojo F, Cuzick J, et al. Prediction of chemotherapy benefit by EndoPredict in patients with breast cancer who received adjuvant endocrine therapy plus chemotherapy or endocrine therapy alone. Breast Cancer Res Treat. 2019;176:377–86.
39. Swain SM, Tang G, Lucas PC, Robidoux A, Goerlitz D, Harris BT, et al. Pathologic complete response and outcomes by intrinsic subtypes in NSABP B-41, a randomized neoadjuvant trial of chemotherapy with trastuzumab, lapatinib, or the combination. Breast Cancer Res Treat. 2019;178:389–99.
40. Szijgyarto Z, Flach KD, Opdam M, Palmieri C, Linn SC, Wesseling J, et al. Dissecting the predictive value of MAPK/AKT/estrogen-receptor phosphorylation axis in primary breast cancer to treatment response for tamoxifen over exemestane: a translational report of the Intergroup Exemestane Study (IES)-PathIES. Breast Cancer Res Treat. 2019;175:149–63.
41. Turner NC, Liu Y, Zhu Z, Loi S, Colleoni M, Loibl S, et al. Cyclin E1 expression and palbociclib efficacy in previously treated hormone receptor-positive metastatic breast cancer. J Clin Oncol. 2019;37:1169–78.
42. Kubo M, Kawai M, Kumamaru H, Miyata H, Tamura K, Yoshida M, et al. A population-based recurrence risk management study of patients with pT1 node-negative HER2+ breast cancer: a National Clinical Database study. Breast Cancer Res Treat. 2019;178:647–56.
43. Mori H, Kubo M, Kai M, Yamada M, Kurata K, Kawaji H, et al. T-bet+ lymphocytes infiltration as an independent better prognostic indicator for triple-negative breast cancer. Breast Cancer Res Treat. 2019;176:569–77.
44. Pizzuti L, Krasniqi E, Barchiesi G, Della Giulia M, Izzo F, Sanguineti G, et al. Distinct HR expression patterns significantly affect the clinical behavior of metastatic HER2+ breast cancer and degree of benefit from novel anti-HER2 agents in the real world setting. Int J Cancer. 2020;146:1917–29.
45. Chumsri S, Serie DJ, Li Z, Pogue-Geile KL, Soyano-Muller AE, Mashadi-Hossein A, et al. Effects of age and immune landscape on outcome in HER2-positive breast cancer in the NCCTG N9831 (Alliance) and NSABP B-31 (NRG) trials. Clin Cancer Res. 2019;25:4422–30.
46. Alfarsi LH, Elansari R, Toss MS, Diez-Rodriguez M, Nolan CC, Ellis IO, et al. Kinesin family member-18A (KIF18A) is a predictive biomarker of poor benefit from endocrine therapy in early ER+ breast cancer. Breast Cancer Res Treat. 2019;173:93–102.
47. Craze ML, El-Ansari R, Aleskandarany MA, Cheng KW, Alfarsi L, Masisi B, et al. Glutamate dehydrogenase (GLUD1) expression in breast cancer. Breast Cancer Res Treat. 2019;174:79–91.
48. Hrebien S, Citi V, Garcia-Murillas I, Cutts R, Fenwick K, Kozarewa I, et al. Early ctDNA dynamics as a surrogate for progression-free survival in advanced breast cancer in the BEECH trial. Ann Oncol. 2019;30:945–52.
49. Ignatov T, Claus M, Nass N, Haybaeck J, Seifert B, Kalinski T, et al. G-protein-coupled estrogen receptor GPER-1 expression in hormone receptor-positive breast cancer is associated with poor benefit of tamoxifen. Breast Cancer Res Treat. 2019;174:121–7.
50. Siddappa CM, Pillai SG, Snider J, Alldredge P, Trinkaus K, Watson MA, et al. Gene expression analysis to detect disseminated tumor cells in the bone marrow of triple-negative breast cancer patients predicts metastatic relapse. Breast Cancer Res Treat. 2019;178:317–25.
51. Bonetti M, Gelber RD. A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. Stat Med. 2000;19:2595–609.
52. Chien AJ, Tripathy D, Albain KS, Symmans WF, Rugo HS, Melisko ME, et al. MK-2206 and standard neoadjuvant chemotherapy improves response in patients with human epidermal growth factor receptor 2-positive and/or hormone receptor-negative breast cancers in the I-SPY 2 trial. J Clin Oncol. 2020;38:1059–69.
53. Lubin JH, Gail MH. On power and sample size for studying features of the relative odds of disease. Am J Epidemiol. 1990;131:552–66.
54. Dobbin KK, McShane LM. Sample size methods for evaluation of predictive biomarkers. Stat Med. 2022;41(16):3199–210.
55. Zhang Y, Hedo R, Rivera A, Rull R, Richardson S, Tu XM. Post hoc power analysis: is it an informative and meaningful analysis? Gen Psychiatry. 2019;32(4):e100069.

56. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses: power and sample size for the interaction test. J Clin Epidemiol. 2004;57(3):229–36.
57. Ou F-S, Michiels S, Shyr Y, Adjei AA, Oberg AL. Biomarker discovery and validation: statistical considerations. J Thorac Oncol. 2021;16(4):537–45.
58. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. BMC Med. 2012;10:51.
59. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2:e124.
60. Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) Statement. Ann Intern Med. 2020;172:35–45.
61. Kent DM, van Klaveren D, Paulus JK, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) Statement: explanation and elaboration. Ann Intern Med. 2020;172:W1–25.
62. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. Biometrics. 2007;63:1181–8.
63. Janes H, Brown MD, Pepe MS, Huang Y. An approach to evaluating and comparing biomarkers for patient treatment selection. Int J Biostat. 2014;10(1):99–121.

Publisher's Note