BJR, 2025, **98**, 981–986
https://doi.org/10.1093/bjr/tqaf027
Advance access publication: 11 February 2025
Review

The current state of imaging biomarker development and evaluation

Steve Halligan, MBBS, MD, PhD, FRCR, FMedSci* **and Sue Mallett, DPhil**

Research Department of Imaging, Centre for Medical Imaging, University College London, London W1W 7TS, United Kingdom

*Corresponding author: Steve Halligan, MBBS, MD, PhD, FRCR, FMedSci, Research Department of Imaging, Centre for Medical Imaging, University College London, Charles Bell House, 43-45 Foley Street, London W1W 7TS, United Kingdom (s.halligan@ucl.ac.uk)

Abstract

With a focus on imaging, this narrative review describes why biomarker research rarely culminates in anything clinically useful. We explain the difference between diagnostic and prognostic biomarkers and the different phases of their research. We describe what factors need consideration when designing research studies to generate evidence sufficient to translate a potential imaging biomarker into clinical practice. We emphasize the need to balance the number of biomarkers investigated with the number of available patient events, so that studies are not overwhelmed by false-positive results. We explain the need for rigorous evaluation, so that clinicians can be convinced the biomarker will work in their practice and benefit their patients. In particular, we explain why novel biomarkers should be evaluated alongside factors already known to be clinically useful, so that their incremental benefit, if any, can be determined.

Keywords: biomarker; clinical markers; models; statistical; research design; methodological study.

Introduction

This article considers, "The current state of imaging biomarker development and assessment". Well, in summary, it is pretty poor, despite the huge current popularity of biomarker research: A "PubMed" search performed on September 2024 using the term "biomarker" found 277 individual article citations for the *British Journal of Radiology* alone. A large majority of these are recent—21 article citations for this journal in 2023 compared with 20 such citations for 1973 to 1999 inclusive, a period approaching 30 years. The general medical literature is replete with studies of novel high-throughput technologies like next-generation sequencing, accompanied by innumerable studies of genetic and proteomic expression signatures, all promising "personalised medicine". Why the explosion of interest? It cannot be because biomarker research usually culminates in clinically useful tools: It is well established that the vast majority of biomarker research never translates to patients. A 2014 review identified a "major disconnect between the several hundred thousand published candidate biomarkers and the less than one-hundred US FDA-approved companion diagnostics".[1] Indeed, very few biomarkers undergo any assessment beyond the articles first proposing them: A recent systematic review of multivariable models incorporating novel imaging biomarkers found that most are never evaluated by other researchers.[2]

Some believe that most biomarker research comprises "developmental studies" that lack focus on a defined clinical application and that depend on retrospective analysis of patients whose data are easily available for analysis.[3] Imaging studies are often motivated by increasingly available patient data (often via public repositories), mined by tools that interrogate and extract "information", such as radiomics. Combine this with statistical and methodological inexperience of both researchers and funders, and institutional imperatives to publish, and it is easy to appreciate why some claim that most positive research findings are simply, "noise discovery".[4,5] For example, a recent systematic review of radiomics research found an average type-I error rate (ie, false-positive results) of 76%, largely due to inadequate sample size compared to the number of variables studied.[6]

What is a biomarker?

The Oxford English Dictionary defines a biomarker as "a naturally occurring molecule, gene, or characteristic by which a particular pathological or physiological process or disease can be identified". The European Society of Radiology Imaging Biomarkers Alliance (EIBALL) states that "an imaging biomarker is a feature of an image relevant to a patient's diagnosis". It follows that a sophisticated measurement of tumour radiomics and a simple measurement of tumour diameter are both imaging biomarkers, that is, biomarkers are not restricted to cutting-edge, novel measurements, but include characteristics such as age, sex, or smoking. Many simple biomarkers are extremely effective—consider sex and breast cancer or age and colon cancer. "Personalised" medicine is a hot-topic currently and promises that data unique to an individual will lead to more appropriate and effective treatment. While recent emphasis has been on "omics", researchers should understand that age, sex, and conventional tumour stage are all personal data. The significance of such "ordinary" biomarkers has been established for decades, and it is crucial to not ignore these when researching novel biomarkers; we expand on this point later.

Most imaging biomarker research concerns "diagnostic" or "prognostic" biomarkers. Diagnostic biomarkers aim to identify a current patient event. A published example would

Received: 27 October 2024; **Revised:** 8 January 2025; **Accepted:** 8 January 2025
© The Author(s) 2025. Published by Oxford University Press on behalf of the British Institute of Radiology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

be whether radiomics can diagnose pancreatic neuroendocrine tumour grade.[7] In contrast, prognostic biomarkers attempt to identify future outcomes. A published example would be whether radiomics can predict survival from renal carcinoma.[8] We published a biomarker systematic review that revealed researchers frequently confuse diagnosis and prediction (often combining both types of data within a single study), meaning that it can be exceptionally difficult to identify exactly what was done.[9] For example, the study of neuroendocrine tumours was titled, "Prediction of tumour grades by radiomics",[7] whereas "Diagnostic prediction models of tumour grades by radiomics" would be more helpful to readers. Fundamentally, diagnostic studies must gather reference data to establish whether diagnosis is correct or not, whereas prognostic studies must accrue future data that confirm or refute predicted outcomes (although such data are often collected retrospectively because this is far less burdensome). Oncologists often make a further distinction, between "prognostic" and "predictive" biomarkers, with the former predicting outcomes independent of treatment and the latter indicating whether treatment will be effective or not.

We must perform biomarker studies properly so that patients are not harmed by potentially misleading results (and so that time and money is not wasted on poor research). Good research means answering a specific clinical question, via an appropriate design within an adequately powered study, using unbiased, representative data, analysed correctly and without bias, all ultimately presented transparently and without "spin". Unfortunately, most biomarker research is probably the opposite. "Fishing trips" to identify viable biomarkers are the rule rather than the exception, often with no prior hypothesis, tested in underpowered studies, using incorrect methods. Researchers frequently claim clinical utility, but without any evaluation to confirm this. A 1994 article stated that numerous studies of new prognostic factors in oncology were "not planned prospectively like clinical trials with specification of hypotheses and sample size determination to assure adequate statistical power".[10] Despite these points being made over 30 years ago, very little has changed.

Stages of biomarker research

It is important that researchers consider and understand what stage (or phase) of biomarker research they are addressing and are explicit about this. Biomarker research can be partitioned into 4 sequential phases as follows[11,12]:

- Discovery: Is the biomarker associated with pathology? Is it biologically plausible? Can it be measured reliably?
- Translation: Does the biomarker separate diseased from normal patients (or high and low risk, or who do and do not respond to treatment, etc.)?
- Single-centre studies: Is the biomarker useful in clinical practice?
- Multi-centre studies: Is clinical utility maintained across multiple centres? Is the biomarker cost-effective?

Biomarker discovery; "predictor finding" studies

From our reading, imaging studies appear dominated by what could be termed "predictor finding", which would fall under the first 2 categories above. These studies frequently partition a retrospective dataset into patients with and without an outcome of interest, and then the proportion of each exhibiting a particular biomarker(s) is compared. Any "significant" differences in proportions are then assumed to indicate that the biomarker in some way differentiates between patients with and without the outcome. An example from this journal would be a study of patients who did (33 patients) or did not (19 patients) die following Covid-19 infection.[13] Patients were identified retrospectively, and multiple biomarkers of body composition are then compared between patients who did and did not die, with significant differences assumed to indicate predictive utility. The authors performed 27 univariable comparisons overall, finding that several were "significant" because the *P*-value was less than .05, corresponding to a probability of 5% false-positive rate. However, if there was no predictive association, we would still anticipate significant results by chance in at least 1 or 2 univariable analysis from 27 comparisons, even assuming an unbiased study. Increasing the number of hypothesis tests encourages false-positive findings (type 1 errors).[6] Predictor finding studies are valuable but probably only when investigating pre-specified hypotheses that are biologically plausible and that are powered adequately to minimise false-positives.

It is fundamental that researchers understand how many patients are needed for their studies to provide good evidence. The best approach is to discuss (and refine) the research question with a statistician, as this informs how many patients are needed: the "sample size". Sample size will depend on what question, precisely, the biomarker is expected to answer. Is the biomarker diagnostic or prognostic? Is the main question how the biomarker identifies different groups of patients, how the biomarker influences current tests, or how it affects patient trajectory? Sample size is influenced by study design (cohort, case-control, or randomised controlled trial?), what is being compared (eg, biomarker measurements within patients or across different groups of patients?), and how the outcome is measured (continuous, binary, time-to-event). What is the claim being made (superiority, non-inferiority), what are its characteristics (eg, mean, standard deviation, repeatability), and in which patients (normal vs diseased or severity of disease)? Missing data and loss to follow-up will also inflate the sample size.

Proving that a biomarker and outcome are related may be addressed simply by a scatterplot of continuous biomarker measurements for patients with and without the outcome of interest. The number of patients required in each group depends on the mean and standard deviation of biomarker values in each group; a discovery phase study could include as few as 10 patients per group, to obtain estimates for these values. Alternatively, sample size could be based on the precision of 95% CI estimates of group means. While researchers often believe that calculating a correlation coefficient is appropriate, a result such as "0.7" is uninformative compared to a scatterplot because correlation coefficients are interpreted differently by different researchers and the value depends on multiple assumptions that may not be evaluated or met. As we move from discovery to translation, the aim is then to determine whether the biomarker can distinguish patients with and without the outcome of interest. Several possibilities must be considered. Again, a statistician should be consulted to determine the most appropriate study design, outcome (endpoint) to be measured, and sample size options, all of which depend on the precise research question being asked.

Multivariable modelling now comprises a substantial proportion of published imaging research, and is frequently used to investigate biomarker associations.[2] Traditionally, a "rule of 10" has been helpful when understanding sample sizes needed for regression modelling,[14] although recent methods have updated this rule of thumb.[15,16] The "event rate" for multivariable modelling is the least frequent outcome of interest; in our Covid example, this would be 19 (ie, the 19 patients who died). Following the rule of 10, a study developing a multivariable model using these data has power to examine just 2 biomarkers at most. If 27 candidate variables were analysed during model building, then significant results are more likely false-positive than not. In an extreme example from this journal, a study with an event rate of just 7 (from 16 patients total) investigated PET/MR biomarkers of both glioblastoma response and survival[17]; investigating multiple different outcomes in a single study requires even more patients to achieve adequate power. Although the authors concluded that PET/MR "allows an early and reliable identification of response to treatment and predict overall survival",[17] the study had insufficient power to investigate a single biomarker. While methodological advice for powering multivariable biomarker studies is relatively immature (eg, compared to randomised controlled trials), we would nevertheless advocate the rule of 10 if the alternative is to ignore power altogether.

Most predictor finding studies use "convenience" samples, but these are very unlikely to yield reliable results. Patients are often non-consecutive and/or only available due to various biases. Most obviously, patients with more advanced, higher-stage disease are imaged more frequently and thereby frequently over-represented in convenience datasets. A biomarker identified using such a sample may not be useful in clinic, where patients of all disease stages are encountered, often alongside patients without disease.

Metrology

Metrology is the science of measurement. Assessment of biomarker linearity, precision, repeatability, reproducibility, etc. is often placed first in the "biomarker pipeline" since a biomarker that cannot be measured reliably is useless. However, while it could be argued that technical assessments should precede those of clinical utility, metrology is studied less frequently. This is probably because "fishing trips" comparing multiple potential biomarkers between patients are much easier to perform (and publish) because single-centre retrospective imaging data are easy to access and analyse (assuming ethical permission has been obtained). Metrology studies usually require pre-specified hypotheses and planned analyses of data collected prospectively, for example, to investigate multiple measurements made on the same patient, by different observers and/or on different platforms. It is also certainly true that there is little motivation to assess metrology for biomarkers that have no clinical utility and, as we have said, evidence of utility is frequently weak. In contrast, guidelines to design, perform, and analyse metrology studies are well disseminated within the radiological community, certainly relative to assessments of clinical utility, possibly because radiologists and basic scientists appear more focused on technical aspects.[18]

Biomarker evaluation

Neither the Covid[13] or glioblastoma[17] study described an evaluation of their promising biomarkers. Evaluation is often termed "validation" (a term statisticians tend to avoid since it implies success) and is an assessment of biomarker diagnosis/prediction, that is, does it "work"? Researchers with limited data often develop and evaluate their biomarker using the same patients. A variety of methods (eg, bootstrapping, jack-knifing, cross-validation) examine slightly different combinations of patients, sampled from those used to develop the biomarker, and provide an "internal" evaluation of the model.[19] Clearly, assessing performance in the same patients used to develop the biomarker will encourage over-optimistic results; it is far preferable to evaluate the biomarker in different patients. "Split-sample" is a common approach (although still "internal" evaluation) and partitions patients into 2 sets (often using randomisation), investigating a range of biomarkers in the first set, and then exploring whether any relationship is maintained in the second. This still provides weak evidence since both sets originate from the same group of patients and will be similar, especially when random splitting is used. "Temporal" evaluation accrues patients from the same centre but assigns consecutive patients to the development set while allocating subsequent patients to the second. This is far preferable to randomising between development and evaluation sets, because randomisation aims to minimise differences between sets. However, temporal evaluation still employs data from similar patients from the same hospital and clinics, imaged on the same platforms. Ultimately, a viable biomarker must work in representative patients from other centres, imaged on platforms from multiple vendors, and using generally available software, etc. This reflects how the biomarker would eventually be deployed in daily clinical practice, and is termed "external" evaluation. Demonstrating accuracy in completely new patients from other centres is fundamental to proving that the biomarker is genuinely diagnostic and/or predictive. The most extreme example is "geographical" evaluation of patients from different countries.

Very unfortunately, most imaging biomarker research omits evaluation altogether, and very few authors present genuine external evaluations. A systematic review of radiological diagnostic and prognostic models found that just 15% presented an evaluation of any type.[2] Ignoring evaluation, even if just simple internal evaluation, reduces research impact considerably and will not attract the interest of other researchers because there is poor evidence that the biomarker is useful. The same systematic review found that almost no radiological models are investigated further by other researchers.[2]

Multi-centre evaluation

Clinicians will not use biomarkers unless proven to work in their patients, which demands external evaluation in multiple centres.[20] However, multi-centre biomarker studies are very uncommon, possibly because few biomarkers demonstrate enough promise to warrant such studies, and/or because multi-centre research demands considerable planning and resources. It is illuminating that all of the 20 pancreatic biomarker studies listed by EIBALL are single centre (https://www.myesr.org/research/biomarkers-inventory).

Assuming a promising biomarker that has been identified, can be measured reliably, and appears promising in single-centre evaluations, the fundamental question remains whether it adds anything useful to patient management in clinical practice? It follows that the biomarker must be assessed alongside established factors already known to be useful. Even if a novel biomarker has proven diagnostic or prognostic utility, this may be insufficient to displace established measurements that are easier and/or cheaper. For example, researchers developing a model of breast cancer survival found that gene expression provided no more useful information than conventional tumour stage, receptor status, and grade; indeed, "omics" only became clinically useful when standard measurements were ignored.[21] Similarly, a study of novel cardiovascular biomarkers found they contributed nothing useful over-and-above standard assessments.[22] To determine whether a biomarker contributes additional useful information to inform patient management, statisticians advocate studies that utilise multivariable models. Such models can assess the cumulative effect of multiple variables on diagnosis or prediction. A "baseline" model is developed that incorporates variables already believed to be clinically useful. Addition of information from a genuinely useful biomarker to this model will improve diagnosis/prediction significantly.[3,10,12] Studies should investigate whether the biomarker improves model "discrimination", which describes how well the model distinguishes between representative patients with and without the outcome of interest (using the c-index or area under the receiver operator characteristic curve [ROC AUC]). Evaluation should examine how the biomarker changes clinical trajectory, via the proportions of patients receiving different therapies when the biomarker is and is not included in the model. Nevertheless, researchers frequently ignore this advice. For example, the study of glioblastoma response investigated PET/MR biomarkers only, ignoring all other factors.[17] In contrast, a study of pancreatic cancer that investigated whether ADC (apparent diffusion coefficient) could predict therapeutic response also considered age, gender, and tumour stage.[23]

An example of biomarker discovery and assessment: the PROSPECT study

Many years ago, the first author and colleagues conducted a small, single-centre study that investigated whether perfusion CT of the primary colorectal tumour could identify which patients were destined to develop metastases subsequently.[24] We prospectively recruited 52 patients, measured tumour perfusion, and then followed patients up to identify those developing metastases. We found significant differences between perfusion parameters when patients with and without metastases were compared, concluding that "tumour blood flow might potentially be a useful predictor".[24] The research was "own-account", and we had performed the analysis ourselves. We then applied for external finding to further investigate our biomarker and approached a medical statistician (S. M.) for advice. She criticised our methods; in particular, we had not investigated whether our biomarker contributed anything valuable over-and-above established predictors of metastasis, especially tumour stage. She then designed a study that would provide good evidence of whether perfusion was clinically useful; improving prediction of metastatic disease in primary colorectal cancer—the PROSPECT study (ISRCTN: 95037515). The study illustrates several methodological aspects that warrant consideration when translating a biomarker to clinical practice.

Firstly, patient data were collected prospectively, because retrospective data are often biased and incomplete. For example, to eliminate selection bias, we included all consecutive patients presenting with primary colorectal cancer (irrespective of stage, etc.). To increase generalisability, data were collected from 13 hospitals, representing the range of radiologists and imaging platforms (and software) that would be used in normal clinical practice, and metrology studies were also performed.[25] We specified in advance which biomarkers we would incorporate in the model, being careful to include not only the novel perfusion measurements but also standard measurements already used widely (eg, gender, age, stage, resection margin positivity, venous invasion, etc.). Understanding the relationship between the number of events (ie, patients developing recurrent disease, which would almost certainly be smaller than disease-free patients) and the number of biomarkers studied in the model obliged us to think carefully about how many biomarkers we could investigate without damaging study power. We drafted an analysis plan that developed a multivariable model to predict recurrence following first presentation with colorectal cancer, and which then evaluated the effect on prediction when CT perfusion measurements were added to the model. Importantly, by prespecification of all biomarkers used for model development, we avoided using univariable significance within the study dataset to select variables, an approach that leads to overfitting and that is discouraged by statisticians.[26] Unfortunately, univariable significance remains the commonest method for variable selection.

We recruited 448 patients over 5 years and were able to analyse 326, 81 (25%) whose cancer recurred.[27] The best model to predict recurrence included age, sex, local stage, tumour size and location, venous invasion, and treatment details and was superior to predictions from standard staging. All of these variables are standard measurements, easily obtained in normal clinical practice. Unfortunately, prediction did not increase when CT perfusion was added to this baseline model, indicating that perfusion had no clinical utility. For example, model sensitivity and specificity at the "high" vs "medium/low" risk threshold was 0.58 (95% CI: 0.46, 0.70) and 0.75 (95% CI: 0.68, 0.81), respectively, compared with 0.57 (95% CI: 0.45, 0.68) and 0.74 (95% CI: 0.68, 0.79) without CT perfusion.[27] It is interesting to note that PROSPECT investigated a restricted number of prespecified immunohistochemical and genetic biomarkers (eg, vascular endothelial growth factor, hypoxia inducible factor, DNA mismatch repair), also assumed to have considerable prognostic potential following small retrospective studies. Like CT factors, we found these also contributed no significant predictive benefit, suggesting that poor biomarker evaluation is not limited to imaging research.

How to do better biomarker research and a cautionary tale

As we have stated, most published imaging biomarker research describes predictor finding and then progresses no further. We summarise how to do better biomarker research in Table 1: Advancing the field requires more focused identification, based on biological plausibility and potential clinical utility.

Table 1. How to do better biomarker research.

Problem:	Solution:
Is my biomarker(s) biologically plausible and is there existing evidence of clinical utility?	Consider a systematic review and meta-analysis of "promising" biomarkers to identify which deserve further research. Consult referring clinicians to help identify "promising" biomarkers.
Am I considering too many biomarkers?	To avoid "data dredging", prespecify the biomarkers to be investigated. Do not chose biomarkers based solely on univariable significance in your own dataset.
Do I have good quality data?	Obtain patient data prospectively if possible, so that data are not missing and can be collected in a way that mimics what is possible in normal clinical practice.
Is my study powered properly?	Balance the number of biomarkers investigated with the number of patient "events" available in your data.
Am I using the right study design?	Consult a qualified medical statistician or methodologist. Evaluate your biomarker(s) alongside factors already known to be useful in clinical practice (eg, within a multivariable model), so that the incremental benefit of the new biomarker (if any) can be determined.
Will my biomarker(s) be useful in normal clinical practice?	Perform evaluation in consecutive representative patients. Aim for a genuine "external" evaluation, that is, using patients from other centres, whose data were not used for biomarker development.
Is my study reported properly?	Consult the EQUATOR Network to identify the correct reporting guideline, for example, TRIPOD for studies of diagnostic and prognostic multivariable models.

Systematic review and meta-analysis can indicate the most promising biomarkers across several studies and should be used more frequently to identify which are most deserving of further research.[12,28] Studies must be properly powered, avoiding the temptation to investigate too many biomarkers compared with the number of available patient events. For example, multiple individual measurements (eg, from radiomics or CT scanning) can be combined into a single factor using statistical methods such as principal component analysis.[29] Authors should present a well-designed evaluation of performance in representative patients, avoiding patient data used for biomarker identification and/or model development. Biomarkers should be studied alongside factors already believed to be useful, so that their incremental benefit, if any, can be determined. Researchers must also understand that machine learning approaches (which may replace traditional regression-based model development for predictor finding, etc.) do not eliminate the need for well-designed, properly powered studies: A recent systematic review found no evidence that machine learning was superior to logistic regression.[30] Researchers should use the appropriate guideline so that their work is reported with sufficient methodological clarity, for example, consulting TRIPOD for diagnostic/prognostic models.[31] The EQUATOR Network will help identify the correct guideline.[32]

Unfortunately, good research is never easy. PROSPECT took several years, cost a considerable amount of money, and culminated in a negative result. Despite this, we argue there is considerable value in well-designed studies able to answer a clinical question with reasonable certainty, compared with small studies whose design flaws question their validity. Publishing negative results may also prove beneficial by discouraging others from pursuing fruitless biomarkers. Nevertheless, we encountered very significant difficulties when attempting to publish PROSPECT: While the design and execution was praised, journals were uninterested in publishing negative results. The current climate appears to favour small studies proposing novel biomarkers rather than larger studies of adequate power that find no benefit. Shortly before writing this article, during a meeting of senior editors from a major medical publishing group, S.H. was told that, due to vast numbers of submissions describing biomarker research, they were only interested in publishing positive results. This stance creates a perverse incentive to perform poorly designed, underpowered studies, since they are more likely to generate (false) positive findings. Indeed, publication bias towards positive results is well established and proven by systematic review.[33] We have witnessed similar behaviour in parallel disciplines: A novel genetic biomarker to identify patients destined to develop severe Crohn's disease appeared very positive in small trials, and a protocol for a large, well-designed prospective study was published.[34] However, the completed study found no biomarker efficacy and, when eventually published, emphasis switched to the positive benefits of different treatment strategies.[35] When discussing why journals publish methodologically poor biomarker research, the Editor of an indexed radiology journal told S.H. that his journal would be, "empty if I didn't". Instead, we agree with Doug Altman, who argued it is better to publish less but better research, because inaccurate results risk harming patients.[36]

Summary

We have described why there is a huge disparity between the innumerable biomarker candidates proposed in the indexed literature, and the very few that eventually enter clinical practice. We believe the best approach to avoid poor biomarker research is to circumvent the "do-it-yourself" culture and consult a medical statistician or properly trained methodologist, who will advise on proper study design and powering. While medical statisticians for imaging research appear to be in short supply,[37] as we have said, it is better to do less, but better research. A statistician rightly pointed out that a few statistics lectures at medical school doesn't make clinicians, "an expert any more than attending a first aid course makes us into clinicians".[38] If you are going to do good biomarker research, then be prepared to devote considerable time, effort, and expense, and be prepared for disappointing results. Maya Angelou, the American author and civil rights activist said, "Do the best you can until you know better. Then when you know better, do better". We hope that this article will help readers, "know better".

Funding

S.H. receives a proportion of funding from the UK National Institute for Health Research Biomedical Research Centre funding scheme. The views expressed in this article are those of the authors and not necessarily those of the UK NIHR.

Conflicts of interest

The authors declare no commercial relationship with any companies whose products or services may be related to the subject matter of the article.

References

1. Poste G, Compton CC, Barker AD. The national biomarker development alliance: confronting the poor productivity of biomarker research and development. *Expert Rev Mol Diagn*. 2015; 15:211-218.
2. Hameed M, Yeung J, Boone D, Mallett S, Halligan S. Meta-research: how many diagnostic or prognostic models published in radiological journals are evaluated externally? *Eur Radiol*. 2024; 34:2524-2533.
3. Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Per Med*. 2010; 7:33-47.
4. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124.
5. Ioannidis JP. Microarrays and molecular research: noise discovery? *Lancet*. 2005;365:454-455.
6. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. *PLoS One*. 2015;10:e0124165.
7. Dong Y, Yang D-H, Tian X-F, et al. Pancreatic neuroendocrine tumor: prediction of tumor grades by radiomics models based on ultrasound images. *Br J Radiol*. 2023;96:20220783.
8. Zhang T, Ming Y, Xu J, et al. Radiomics and Ki-67 index predict survival in clear cell renal cell carcinoma. *Br J Radiol*. 2023; 96:20230187.
9. Halligan S, Boone D, Bhatnagar G, et al. Prognostic biomarkers to identify patients destined to develop severe Crohn's disease who may benefit from early biological therapy: protocol for a systematic review, meta-analysis and external validation. *Syst Rev*. 2016; 5:206.
10. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer*. 1994;69:979-985.
11. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst*. 2001; 93:1054-1061.
12. Halligan S, Menu Y, Mallett S. Why did European radiology reject my radiomic biomarker paper? How to correctly evaluate imaging biomarkers in a clinical setting. *Eur Radiol*. 2021;31:9361-9368.
13. Surov A, Thormann M, Kardas H, et al. Visceral to subcutaneous fat ratio predicts short-term mortality in patients with Covid 19. A multicenter study. *Br J Radiol*. 2023;96:20220869.
14. van Smeden M, de Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol*. 2016;16:163.
15. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II—binary and time-to-event outcomes. *Stat Med*. 2019;38:1276-1296.
16. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I—continuous outcomes. *Stat Med*. 2019;38:1262-1275.
17. Lombardi G, Spimpolo A, Berti S, et al. PET/MR in recurrent glioblastoma patients treated with regorafenib: [(18)F]FET and DWI-ADC for response assessment and survival prediction. *Br J Radiol*. 2022;95:20211018.
18. Sullivan DC, Obuchowski NA, Kessler LG, et al. RSNA-QIBA Metrology Working Group. Metrology standards for quantitative imaging biomarkers. *Radiology*. 2015;277:813-825.
19. Lopez E, Etxebarria-Elezgarai J, Amigo JM, Seifert A. The importance of choosing a proper validation strategy in predictive models. A tutorial with real examples. *Anal Chim Acta*. 2023; 1275:341532.
20. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statist Med*. 2000;19:453-473.
21. Volkmann A, De Bin R, Sauerbrei W, Boulesteix AL. A plea for taking all available clinical information into account when assessing the predictive value of omics data. *BMC Med Res Methodol*. 2019;19:162.
22. Melander O, Newton-Cheh C, Almgren P, et al. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. *JAMA*. 2009;302:49-57.
23. Niwa T, Ueno M, Ohkawa S, et al. Advanced pancreatic cancer: the use of the apparent diffusion coefficient to predict response to chemotherapy. *Br J Radiol*. 2009;82:28-34.
24. Goh V, Halligan S, Wellsted DM, Bartram CI. Can perfusion CT assessment of primary colorectal adenocarcinoma blood flow at staging predict for subsequent metastatic disease? A pilot study. *Eur Radiol*. 2009;19:79-89.
25. Lewis M, Goh V, Beggs S, et al. PROSPeCT Investigators. Quality control within the multicentre perfusion CT study of primary colorectal cancer (PROSPeCT): results of an iodine density phantom study. *Eur Radiol*. 2014;24:2309-2318.
26. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol*. 1996;49:907-916.
27. Goh V, Mallett S, Boulter V, et al. For PROSPECT Investigators Multivariable prognostic modelling to improve prediction of colorectal cancer recurrence: the PROSPeCT trial. *Eur Radiol*. 2024; 34:6992-7001.
28. Riley RD, Moons KGM, Snell KIE, et al. A guide to systematic review and meta-analysis of prognostic factor studies. *BMJ*. 2019;364:k4597.
29. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci*. 2016; 374:20150202.
30. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22.
31. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162:55-63.
32. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. *Lancet*. 2008;371:1149-1150.
33. Dwan K, Altman DG, Arnaiz JA, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*. 2008;3:e3081.
34. Parkes M, Noor NM, Dowling F, et al. PRedicting Outcomes For Crohn's disease using a moLecular biomarkEr (PROFILE): protocol for a multicentre, randomised, biomarker-stratified trial. *BMJ Open*. 2018;8:e026767.
35. Noor NM, Lee JC, Bond S, et al. PROFILE Study Group. A biomarker-stratified comparison of top-down versus accelerated step-up treatment strategies for patients with newly diagnosed Crohn's disease (PROFILE): a multicentre, open-label randomised controlled trial. *Lancet Gastroenterol Hepatol*. 2024;9:415-427.
36. Altman DG. The scandal of poor medical research. *BMJ*. 1994; 308:283-284.
37. Campbell SZ, Kuah JY, Hall-Craggs M, Parry T, Mallett S, Halligan S. Access to statistical support for medical imaging research: questionnaire survey of UK radiology trainees. *Clin Radiol*. 2022;77:920-924.
38. Stratton IM, Thorne KI. The statistician will see you now. *Diabet Med*. 2021;38:e14437.

© The Author(s) 2025. Published by Oxford University Press on behalf of the British Institute of Radiology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
British Journal of Radiology, 2025, 98, 981–986
https://doi.org/10.1093/bjr/tqaf027
Review