ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ANNALS OF ONCOLOGY DRIVING INNOVATION IN ONCOLOGY

ORIGINAL ARTICLE

ESMO basic requirements for AI-based biomarkers in oncology (EBAI)

M. Aldea[1,2,3], M. Salto-Tellez[4,5], A. Marra[6,7], R. Umeton[8,9,10,11,12], A. Stenzinger[13], M. Koopman[14], A. Prelaj[15,16], K. L. Kehl[3], S. Gilbert[17], M.-E. Leßmann[17], J. Lipkova[18,19,20], L. Provenzano[15,16], F. Meric-Bernstam[21], S. Halabi[22,23], J. Wu[24,25], A. Pellat[26], K. P. M. Suijkerbuijk[14], B. Besse[1,2], B. Ryll[27], C. Marchió[28,29], M. Crispin-Ortuzar[30,31], R. Fehrmann[32], J. Vibert[33], D. Ferber[34], C. Pauli[35], A. Valachis[36], F. Corso[15], T. J. Brinker[37], J. Mateo[38,39], N. Harbeck[40], E. C. Winkler[41], F. Lopez-Rios[42], R. Perez-Lopez[43], G. Pentheroudakis[44], S. Delaloge[1], C. Benedikt Westphalen[45,46]* & J. N. Kather[17,34,47]*

[1]Department of Cancer Medicine, Gustave Roussy, Villejuif; [2]Faculty of Medicine, Paris-Saclay University, Kremlin Bicêtre, France; [3]Lowe Center for Thoracic Oncology, Dana-Farber Cancer Institute, Boston, USA; [4]Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast; [5]Joint Integrated Pathology Unit, The Institute of Cancer Research, The Royal Marsden NHS Foundation Trust, London, UK; [6]Division of Early Drug Development for Innovative Therapies, European Institute of Oncology IRCCS, Milan; [7]Department of Oncology and Hemato-Oncology, University of Milano, Milan, Italy; [8]Office of Data Science, St. Jude Children's Research Hospital, Memphis, USA; Departments of [9]Biological Engineering; [10]Mechanical Engineering, Massachusetts Institute of Technology, Cambridge; [11]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston; [12]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, USA; [13]Institute of Pathology, University Hospital Heidelberg and Center for Personalized Medicine (ZPM), Heidelberg, Germany; [14]Department of Medical Oncology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; [15]AI-ON-Lab, Medical Oncology Department, Fondazione IRCCS Istituto Nazionale Tumori, Milan; [16]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy; [17]Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany; Departments of [18]Pathology and Laboratory Medicine; [19]Biomedical Engineering, University of California Irvine, Irvine; [20]Chao Family Comprehensive Cancer Center, UC Health, Orange; [21]Department of Investigational Cancer Therapeutics, The University of Texas MD Anderson Cancer Center, Houston; [22]Department of Biostatistics and Bioinformatics, Duke University, Durham; [23]Duke Cancer Institute, Duke University, Durham; Departments of [24]Imaging Physics; [25]Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, USA; [26]Department of Gastroenterology, Endoscopy and Digestive Oncology Unit, Cochin Hospital AP-HP, Paris, France; [27]Melanoma Patient Network Europe, Uppsala, Sweden; [28]Department of Medical Sciences, University of Turin, Turin; [29]Division of Pathology, Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Italy; [30]Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge; [31]Department of Oncology, University of Cambridge, Cambridge, UK; [32]Department of Medical Oncology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands; [33]Drug Development Department, Gustave Roussy, Villejuif, France; [34]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany; [35]Department of Pathology and Molecular Pathology, University Hospital Zurich and Medical Faculty, University of Zurich, Zurich, Switzerland; [36]Department of Oncology, Faculty of Medicine and Health, Örebro University Hospital, Örebro University, Örebro, Sweden; [37]Division of Digital Prevention, Diagnostics and Therapy Guidance, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany; [38]Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron Barcelona Hospital Campus, Barcelona; [39]Department of Medical Oncology, Vall d'Hebron University Hospital, Barcelona, Spain; [40]Breast Centre, Department of Obstetrics & Gynaecology and Comprehensive Cancer Centre Munich, LMU University Hospital, Munich; [41]Institute for Medical and Data Ethics, Medical Faculty, Heidelberg University and Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany; [42]Pathology Department, Hospital Universitario 12 de Octubre, Universidad Complutense de Madrid, Research Institute Hospital 12 de Octubre (imas12), CIBERONC, Madrid; [43]Radiomics Group, Vall d'Hebron Institute of Oncology, (VHIO), Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain; [44]Scientific and Medical Division, European Society for Medical Oncology (ESMO), Lugano, Switzerland; [45]Comprehensive Cancer Center Munich & Department of Medicine III, University Hospital, LMU Munich, Munich; [46]German Cancer Consortium (DKTK), Partner Site Munich, German Cancer Research Center (DKFZ), Heidelberg, Germany; [47]Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK

Available online 18 November 2025

Background: Artificial intelligence (AI) is expected to introduce an increasing number of biomarkers in oncology. To bridge the gap between oncology and computer science, it is timely to define recommendations for AI-based biomarkers suitable for routine clinical use. Here, we propose the **E**SMO (European Society for Medical Oncology) **B**asic Requirements for **AI**-based Biomarkers In Oncology (EBAI).

Design: The EBAI framework was developed using a modified Delphi methodology, involving a multidisciplinary panel of 37 experts who participated in four structured consensus rounds.

Results: AI-based biomarkers were classified as 'class A' (AI quantification of established biomarkers), 'class B' (indirect measure of known biomarkers using AI-based alternative methods, to be deployed as pre-screening tests), and 'class C' (novel AI-derived biomarkers, with C1 for prognosis and C2 for prediction of treatment effect). The EBAI framework addresses AI biomarkers for clinical use. Ground truth, performance, and generalisability were considered essential;

*Correspondence to: Dr Christoph Benedikt Westphalen, ESMO Head Office – Scientific and Medical Division, Via Ginevra 4, Lugano CH-6900, Switzerland. Tel: +41-91-973-1999; Fax: +41-91-973-1902

E-mail: medicalaffairs@esmo.org (C. Benedikt Westphalen).

Prof. Dr Jakob Nikolas Kather, ESMO Head Office – Scientific and Medical Division, Via Ginevra 4, Lugano CH-6900, Switzerland. Tel: +41-91-973-1999; Fax: +41-91-973-1902

E-mail: medicalaffairs@esmo.org (J. N. Kather).

fairness was recommended. Minimal validation requirements indicate that class A requires concordance studies, class B analytical validation, class C1 high-quality retrospective real-world or clinical trial data, and class C2 additionally requires clinical validation in prospective clinical trials for the prediction of response to a new treatment. All biomarker studies should report multiple evaluation and calibration metrics, with a clearly defined primary objective. Generalisability should be demonstrated across all intended use settings, including variability in data acquisition, post-processing, and population characteristics. Biomarkers must not be applied to other cancer types or modalities without supporting evidence.

Conclusions: EBAI defines criteria for AI-based biomarker adoption in routine use, providing a common language for physicians, AI developers, and researchers.

Key words: artificial intelligence, biomarker, cancer, scale, EBAI, validation

INTRODUCTION

Artificial intelligence (AI) technologies have made rapid advancements and are approaching broad real-world use in oncology.[1] While substantial attention has focused on the automation of clinical procedures and workflows, there is another potentially transformative application emerging with clinical impact: the use of AI-based biomarkers in oncology care.[2,3] AI aims to contribute to cancer diagnostics like any other field generating new biomarkers: making the delivery of the test faster or cheaper, or delivering a new predictive test for a drug with no companion diagnostic to date. This is a shift in how we conceptualise and implement biomarkers for cancer diagnosis, prognosis, and treatment selection.[4]

A biomarker is defined by the Biomarkers Definition Working Group[5] as 'a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.' In the traditional sense, biomarkers have been physical entities—proteins, genomic alterations, or cellular characteristics measured through laboratory techniques. A biomarker, however, can also be a signal in biological measurements, such as a morphological pattern of cells in histopathology slides, a genomic signature, or a characteristic radiologic pattern in imaging scans. Fundamental concepts of biomarker development and evaluation were established for tissue-based markers.[6] AI algorithms can be trained to quantify these patterns and detect such composite biomarkers. Pragmatically, we refer to these as 'AI-based biomarkers', in line with the recent literature.[4,7-9] AI systems can function as biomarkers because they are able to analyse complex, multidimensional data to predict disease features and clinical outcomes, including treatment responses in patients with cancer. These AI systems process information and identify patterns that may even be imperceptible to human experts, effectively transforming data into actionable clinical insights.[9,10]

AI-based biomarkers have gained particular prominence in histopathology,[11] where deep learning technology can quantify features such as tumour-infiltrating lymphocytes or expression of programmed death-ligand 1 (PD-L1) and human epidermal growth factor receptor 2 (HER2) proteins from digitised tissue slides. The scope of AI-based biomarkers, however, extends well beyond pathology. AI systems utilising molecular data could potentially predict treatment response or recurrence risk based on complex molecular signatures. Similarly, AI can extract prognostic or predictive information from radiology images that outperforms conventional clinical interpretation.[12-14] Even electronic health records can be processed by AI to generate biomarker-like predictions about disease trajectory or treatment outcomes.[15,16] These diverse applications share a common thread: they aim to convert complex, medical data into clinically relevant predictions that guide treatment decisions.

These AI-based biomarkers, when implemented as software products intended for diagnostic or prediction purposes, require authorization and/or certification before clinical implementation or commercialization,[17] unless they are 'homebrew' tests.[18] Several tools adopting AI-based biomarkers have already received market authorization from regulatory bodies such as the Food and Drug Administration (FDA) in the USA and have been certified by Notified Bodies in the European Union (EU) under the EU *in vitro* diagnostic regulation (IVDR) or the EU medical devices regulation (MDR), with dozens of AI solutions currently authorized/certified for precision oncology applications.[19-21] These tools range from straightforward quantification tools to sophisticated predictive algorithms. Clinical validation data, however, were missing for >40% of the FDA approved AI devices in 2024.[22] Also, despite regulatory authorization/certification, the real-world uptake of these technologies remains disappointingly slow, inconsistent, and highly heterogenous across health care systems.

This limited adoption likely stems from a combination of factors, including a lack of confidence in AI solutions in clinical practice; structural barriers such as insufficient resources for full digitalisation and long-term data storage infrastructure; the absence of clear reimbursement models for AI-based tools; and unclear responsibility for the consequences of using these biomarkers. Confidence is further undermined by the absence of comprehensive guidance on validation criteria beyond baseline regulatory requirements. Regulatory authorization/certification typically affirms that technical functionality, safety, and performance requirements and standards have been met, but it often does not address the full spectrum of considerations

relevant to clinical implementation. These include performance benchmarks across diverse populations, generalisability across different technical set-ups and clinical settings, integration into existing workflows, cost-effectiveness, and the critical question of when an AI-based biomarker can replace rather than merely supplement conventional methods.[23,24] Furthermore, there is often uncertainty about how to evaluate the explainability of AI predictions, which affects not only clinician trust and acceptance but may also be important for physician–patient communication.[24] There is a critical need for high level consensus recommendation on minimal requirements for AI-driven biomarkers to enable their translation into clinical practice.

THE EBAI EFFORT

The European Society for Medical Oncology (ESMO) has recognised this pressing need for structured guidance in the rapidly evolving landscape of AI-based biomarkers. In response, the ESMO Precision Oncology Task Force and the ESMO Real World Data and Digital Health Task Force have collaborated to create the ESMO Basic Requirements for AI-based Biomarkers in Oncology (EBAI).

EBAI aims to provide a conceptual framework of AI-based biomarkers, as well as clear, actionable guidance for developers, physicians, regulators, and health care institutions regarding the evaluation and implementation of AI-based biomarkers in oncology. During the conceptual development of EBAI, numerous key dimensions for the evaluation of AI-based biomarkers were discussed, such as comparator validation, performance metrics, generalisability, explainability, cost considerations, turnaround time, and fairness auditing. While patient experience and provider experience were not considered as a standalone parameter, they were discussed in all dimensions. Thus, EBAI creates a shared language and consistent benchmarks to guide all parties involved. The expectation is that this guidance will help developers understand evidence requirements before investing in extensive validation studies, assist clinicians in determining when an AI-based biomarker is sufficiently validated for clinical use, and support regulatory bodies and health care institutions in setting appropriate standards for implementation.

MATERIALS AND METHODS

The development of the EBAI framework employed a modified Delphi methodology to establish consensus among experts in the field. A multidisciplinary panel of 37 experts included oncologists, pathologists, radiologists, computational scientists, bioinformaticians, biostatisticians, hybrid clinician-technologists, ethicists, a patient advocate and regulatory specialists across Europe, North America, and Asia. Panel members were selected by the ESMO Precision Oncology Task Force and the ESMO Real World Data and Digital Health Task Force, based on their expertise in AI applications in oncology, biomarker development, or regulatory affairs related to medical devices. All panellists completed ICMJE disclosures, which were reviewed by ESMO in line with its conflict of interest policies.

The process consisted of four structured Delphi rounds, combining qualitative input and quantitative assessment. The initial round was exploratory in nature, allowing experts to suggest and discuss potential criteria across various dimensions of AI-based biomarker evaluation, as well as to provide a quantitative vote on all dimensions across all classes. Building on insights from the first round, the second and third phase employed a quantitative approach with specific statements addressing a particular requirement for a biomarker class (A, B, or C) within an evaluation dimension (comparator, performance, generalisability, explainability, cost, turnaround time, or fairness). The fourth round was carried out after peer review. Consensus was defined as agreement by $\geq$75% of participants.[25] Levels of agreement were categorised as follows: strong consensus ($\geq$96%), consensus (76%-95%), majority approval (50%-75%), and no consensus ($<$50%). In addition, predefined statistical thresholds of a 7-point Likert scale with a mean score $\geq$5 and standard deviation $\leq$1.5 were applied to assess the strength and consistency of expert agreement.[26] Voting was anonymous.

The Delphi process specifically addressed criteria for AI-based biomarkers which are considered 'ready for clinical use' (Supplementary Tables S1-S4, available at https://doi.org/10.1016/j.annonc.2025.11.009). This represents AI-based biomarkers that meet all requirements for clinical implementation. While regulatory clearance (e.g. CE marking or FDA approval) is necessary, it is not sufficient; it also requires demonstrated clinical utility and integration readiness.

EBAI BIOMARKERS

Summary of classes

The EBAI framework categorises AI-based biomarkers into three distinct classes based on their functional approach, relationship to existing biomarkers, and the assumed application of current regulatory classification systems in the EU and the USA (Table 1). Class A systems (AI for biomarker quantification) automate the measurement of established biomarkers using the same input data that human experts would evaluate. Class B systems (AI as indirect measure of existing biomarkers) predict established (molecular) markers or molecular alterations using alternative, often more accessible or cost-effective methods, and are usually intended as pre-screening tests to enrich populations, rather than definitive tests. Class C systems (novel AI-based biomarkers for outcome prediction) discover entirely new patterns with prognostic or predictive value, trained directly on clinical outcome data. Class C biomarkers are further subdivided into C1 (prognostic) and C2 (predictive). C1 systems predict patients outcomes, including cancer recurrence risk or survival outcomes, such as predicting cancer recurrence from haematoxylin–eosin (H&E) pathology slides. C2 systems predict treatment

Table 1. EBAI biomarker classes and current IVDR/MDR and FDA device classification

EBAI biomarker class	Description	Potential IVDR/MDR risk class EU	FDA risk class USA	Fictitious examples
EBAI class A: AI for biomarker quantification	AI systems that support and enhance existing experimental assays by automating the interpretation of biomarker expression. These tools aim to replicate or improve upon human expert analysis. They are based on the same data which a human would use.	IVDR class C rationale: these tools are used for cancer diagnosis, staging, or monitoring, which typically falls in IVDR class 3 (rule 3h of Annex VIII to the IVDR). They support critical diagnostic decisions but do not solely determine high-risk diagnoses.	Class I (low to moderate risk): general controls or class II (moderate to high risk): general controls and special controls	1. 'HistoQuant-HER2': a software that analyses digitised slide images of breast cancer tissue to automatically count the percentage of tumour cells staining positive for HER2, providing a precise and reproducible score to aid pathologists. 2. 'LymphoCounter': an AI tool that identifies and quantifies tumour-infiltrating lymphocytes in melanoma biopsies from whole-slide images, a biomarker for immunotherapy response. 3. 'PD-L1 Professional': a deep learning algorithm that assesses the expression of PD-L1 on tumour cells from immunohistochemistry images in non-small-cell lung cancer, providing a tumour proportion score.
EBAI class B: AI as indirect measure of existing biomarkers	AI-based systems that predict established molecular markers or genetic alterations using different, often more accessible or cost-effective methods. These tools are often intended as pre-screening tests or alternatives when standard tests are unavailable.	IVDR class C or D Rationale: could be class C if used primarily for screening or initial diagnosis (rule 3h of Annex VIII to the IVDR). May be class D if used to guide critical treatment decisions, especially for targeted therapies (rule 1 of Annex VIII to the IVDR).	Class I (low to moderate risk): general controls or class II (moderate to high risk): general controls and special controls	1. 'Geno-Histo-Predictor': an algorithm that predicts the likelihood of an *EGFR* mutation in patients with lung adenocarcinoma by analysing patterns in their standard H&E stained pathology slides, flagging patients for confirmatory genetic testing. 2. 'MSI-Scanner': A CT scan analysis tool that identifies radiomic features in colorectal cancer images that are highly correlated with microsatellite instability (MSI-high) status, which is traditionally determined by genomic testing. 3. 'BRCA-Vision-X': an AI model that analyses mammogram textures to predict the probability of a patient carrying a *BRCA1* or *BRCA2* mutation, suggesting the need for genetic counselling and testing.
EBAI class C: novel AI-based biomarkers for outcome prediction	AI systems trained directly on clinical outcome data to provide prognostic information or predict treatment response. These biomarkers often integrate complex patterns not easily discernible by human observers.	IVDR class D rationale: These biomarkers directly influence high-stakes clinical decisions for life-threatening diseases. They may function as companion diagnostics, placing them in the highest risk category under the IVDR (rules 1 and 2 of Annex VIII to the IVDR). MDR class III Rationale: these biomarkers are intended to provide information which is used to take decisions with diagnosis or therapeutic purposes. Where such a decision may have an impact that may	Class III (high risk): general controls and premarket approval (PMA).	1. 'OncoScriberPro': a prognostic tool that analyses the text from a patient's electronic health record (clinical notes, lab results, and imaging reports) to generate a 'recurrence risk score' for early-stage colon cancer after surgery. 2. 'ChemoResponse-Indexer': an AI system that integrates genomic, proteomic, and digital pathology data to produce a novel score predicting a pancreatic cancer patient's likelihood of responding to a specific chemotherapy regimen.

Continued

Table 1. Continued

EBAI biomarker class	Description	Potential IVDR/MDR risk class EU	FDA risk class USA	Fictitious examples
		cause death or an irreversible deterioration of a person's state of health, the device falls under class III (rule 11 of Annex VIII to the MDR).		3. 'Rad-Immuno-Sig-DX': a predictive biomarker based on a proprietary deep learning analysis of pre-treatment PET/CT scans that identifies a unique metabolic signature to predict which patients with advanced melanoma will benefit from immune checkpoint inhibitors.

The 'fictitious examples' were generated by the authors with the use of GPT4DFCI[27] and Gemini 2.5 Pro; after using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. AI, artificial intelligence; BRCA1/2, breast cancer genes 1/2; CT, computed tomography; EBAI, ESMO Basic Requirements for AI-based Biomarkers in Oncology; EGFR, epidermal growth factor receptor; FDA, Food and Drug Administration; H&E, haematoxylin—eosin; HER2, human epidermal growth factor receptor 2; IVDR, *in vitro* diagnostic regulation; MDR, medical devices regulation; MSI, microsatellite instability; PD-L1, programmed death-ligand 1, PET, positron emission tomography.

response, such as predicting immunotherapy efficacy in lung cancer from chest computed tomography (CT) images (Figure 1A). All three classes of AI-based biomarkers (A, B, C1, and C2) will be described and defined in greater detail below.

Criteria for evaluating AI-based biomarkers

During the EBAI consensus process, experts discussed seven key dimensions for evaluating AI-based biomarkers, some of which were included in the recommendations. The 'Comparator (Ground Truth)' dimension addresses the reference standard against which the AI-based biomarker is validated, including considerations about expert validators and the appropriate gold standard for comparison. 'Performance' examines the metrics and thresholds used to evaluate an AI-based biomarker's accuracy relative to clinical gold standards. 'Generalisability' focuses on external validation and the stability of performance across new cohorts, including variations in populations, clinical environments, and technical conditions. 'Fairness', by contrast, addresses the presence and mitigation of biases within a single cohort, especially in relation to sensitive or protected attributes such as race, gender, or socioeconomic status. 'Explainability' considers the required level of interpretability and validation of clinically relevant explanatory features. 'Cost Considerations' evaluate implementation and economic sustainability associated with an AI-based biomarker. 'Turnaround Time' addresses time-to-result requirements and workflow integration and scalability (Supplementary Figure S1, available at https://doi.org/10.1016/j.annonc.2025.11.009).

General recommendations for assessment criteria

A consensus was reached on key criteria to determine whether AI-based biomarkers are suitable for clinical use, applicable across biomarker classes A, B, and C (Figures 1B and 2) (see Supplementary Tables S1-S3, available at https://doi.org/10.1016/j.annonc.2025.11.009 for detailed DELPHI results).

Essential assessment criteria (Figure 1B)

- Ground truth, performance, and generalisability are considered essential.
- Fairness is recommended but not mandatory, as it focuses on bias within individual cohorts, a related but distinct concern from generalisability, and one that may not always be addressed in early evaluations.
- While explainability may not be mandatory, it is recommended to include targeted interpretability experiments as 'sanity checks' to guard against shortcut learning and spurious associations—for example, saliency-map randomization tests to confirm that explanations depend on learned parameters and labels; region occlusion/ablation showing that performance degrades when clinically relevant tissue is removed rather than when background is perturbed; attention-map stress tests that introduce synthetic or feature-driven confounders (e.g. stains, blur, overlaid text) to detect artefact reliance; and relevance-guided localization to verify that signal concentrates in anatomically plausible regions.[28-30]

Validation requirements (Figure 2) (before adoption)

- Multiple performance metrics must be reported. The primary endpoint of validation studies should be clearly assessed, with appropriate metrics [e.g. sensitivity, specificity, receiver operating characteristic (ROC) curves for classification tasks]. When evaluating the discrimination performance of an AI biomarker, it is important to recognise that each metric reflects a distinct dimension of predictive ability. The AUC and c-index measure how well the biomarker ranks individuals according to risk, distinguishing cases from non-cases, with the c-index extending this concept to time-to-event data while accounting for censoring. Calibration should also be assessed using the calibration slope, intercept, and observed-to-expected (O : E) ratio to ensure that predicted risks correspond well to observed outcomes. Additional metrics such as the net reclassification improvement (NRI) assess the incremental value of adding the biomarker to an existing validated model by

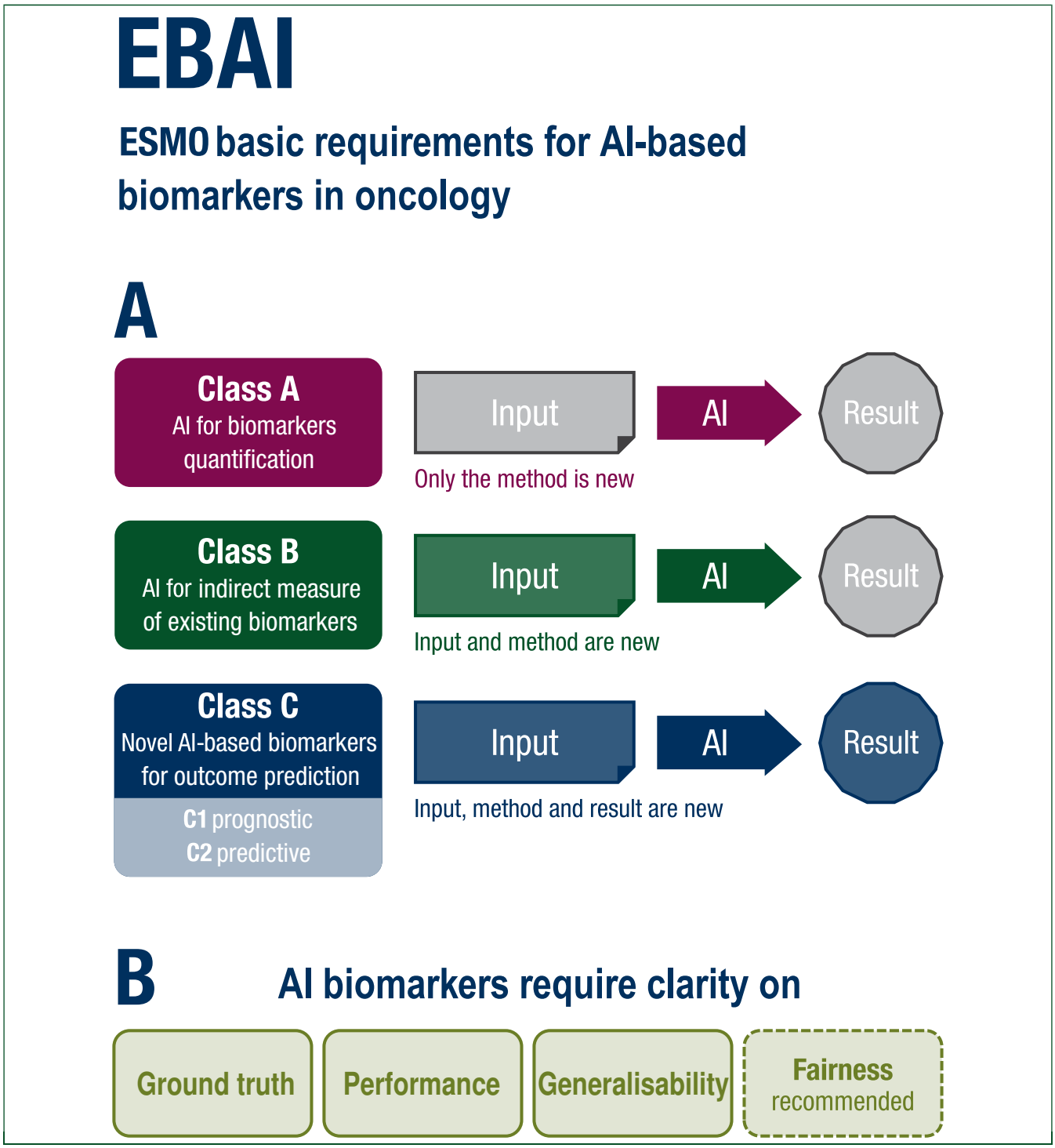

Figure 1. An overview of the EBAI framework. (A) EBAI Classes. (B) Key evaluation dimensions defined by the panel.
AI, artificial intelligence; EBAI, ESMO Basic Requirements for AI-based Biomarkers in Oncology; ESMO, European Society for Medical Oncology.

quantifying improvements in risk classification and average sensitivity/specificity. Finally, decision curve analysis (DCA) complements these measures by evaluating the clinical usefulness of the biomarker across a range of decision thresholds, determining whether its application would improve patient outcomes compared with alternative strategies. More details are available in previous work.[31-33] Examples are illustrated in Table 2.

- Sample size of the validation study(ies) should be justified through an a priori calculation[34] whenever indicated and possible, based on the study's primary objective. The sample size justification for a validation study depends on the study's primary objective. If the goal is estimation—for example, estimating the area under the ROC curve (AUC) or concordance index (c-index) of a new AI biomarker model—the sample size should

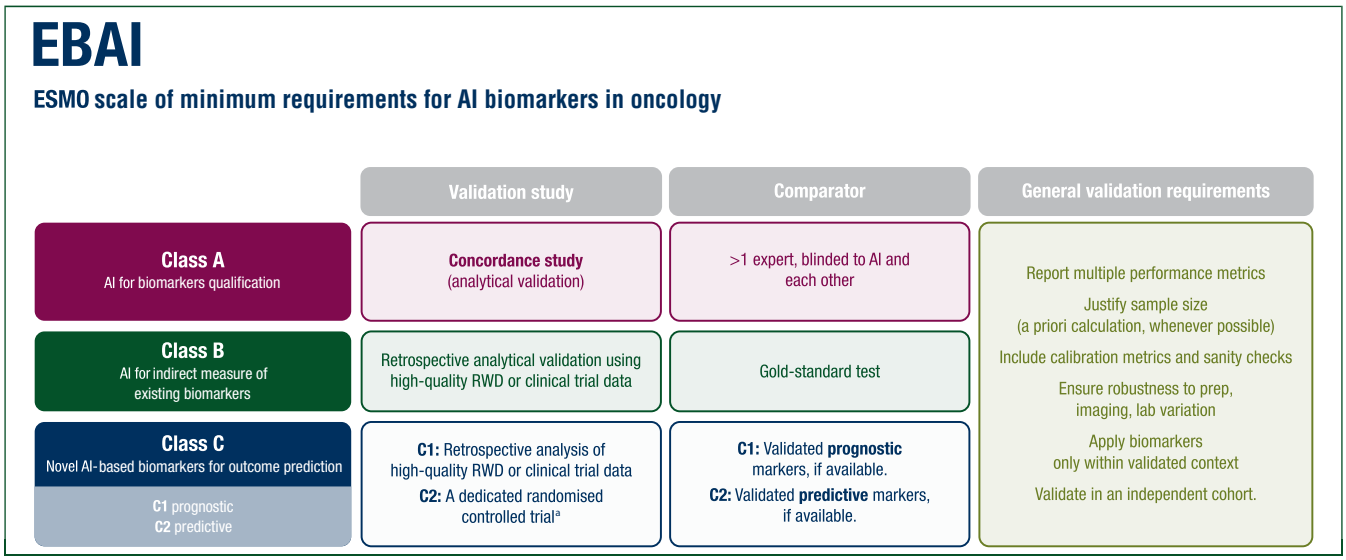

Figure 2. EBAI's general and specific basic requirements of AI-based biomarkers to be ready for clinical use.
AI, artificial intelligence; EBAI, ESMO Basic Requirements for AI-based Biomarkers in Oncology; ESMO, European Society for Medical Oncology; RWD, read-world data.
[a]For the prediction of response to a new treatment.

be based on the desired precision of the estimate. For instance, an investigator may wish to estimate the AUC of an AI-based biomarker in a prospective study with a 95% confidence interval (CI) width of ± 0.05. For a total sample size of 2750 and 550 events, the 95% CI for AUC can be estimated with CI of ± 0.047. Alternatively, if the objective is hypothesis testing—such as comparing the AUC of a new AI biomarker model against a historical or standard model—the sample size should be determined based on the expected difference in AUC, the desired power (e.g. 80%), and the significance level (e.g. 0.05). Several statistical methods are available for comparing tAUCs or c-indices, and the choice depends on whether the outcome is binary time-to-event. More details are available in.[35-40]

- Calibration, used to evaluate how well predicted probabilities align with observed outcomes in the real population, should be reported. A common approach is to plot predicted versus observed probabilities; a perfectly calibrated model will have points lying along a 45-degree line.[31]
- Results should remain consistent across variations in sample preparation, the type or model of imaging devices (e.g. scanners, microscopes), and laboratory protocols.
- Biomarkers validated in a specific cancer type or sample modality should not be applied to other contexts unless those contexts were included in the validation cohort.
- Appropriate metadata should be available and documented according to recommendations in the field.[41-45]
- An independent validation cohort is required.

Post-validation requirements (after adoption)

- AI-based biomarkers integrated into clinical practice should undergo continuous automated monitoring to ensure sustained performance, and reliability.
- Monitoring systems should trigger alerts when issues such as performance degradation, data drift, or emerging bias are detected, prompting root cause analysis.
- To increase trust, it might be helpful for retrospective studies to be conducted after validation to confirm performance across different populations, preparation protocols, and acquisition devices. Some AI tools, however, may be intended for local clinical implementation, in which case multiple retrospective studies may not be feasible. In such situations, limited prospective testing, carried out within the real clinical workflow before full deployment, may represent a feasible and safe alternative.

General recommendations for performance criteria and validation requirements of AI biomarkers

The intended clinical use of an AI-based biomarker—whether for screening, triaging, or replacing an existing diagnostic standard—depends primarily on its demonstrated performance and clinical utility. For prescreening purposes, joint human AI performance metrics should be higher than human performance alone. When a biomarker is proposed to replace a current gold standard, more stringent validation criteria must be met.

Consensus was reached on the following requirements for AI-based biomarkers intended to replace existing standards:

- 'Replacement of human assessment for class A': the AI-based biomarker must show 'very high concordance' with the clinical gold standard, defined as achieving error rates equal to or lower than the interobserver variability among trained human evaluators (e.g. for PD-L1 quantification).

Table 2. Examples of class-specific validation metrics for AI-based biomarkers (EBAI classes A, B, C1, C2)

EBAI class	Typical intended use	Primary metric(s) (pre-specified)	Secondary/reporting	Calibration
EBAI class A: AI for biomarker quantification	Replace/standardize human read of an established biomarker (e.g. PD-L1 %, HER2)	ICC/CCC versus reference; agreement at clinical cut-points (e.g. ≥1%, ≥50%) using OPA/PPA/NPA; for continuous outputs, MAE	Cohen's κ at decision thresholds; Bland–Altman; % within acceptable error; brief error-mode audit (e.g. tumour versus immune-cell confusion; artefacts)	Reliability plot across bins; slope/intercept for continuous scores
EBAI class B: AI as indirect measure of existing biomarkers	Rule-out/rule-in proxy for a molecular test (e.g. *EGFR*/MSI from H&E)	AUC and operating-point Se/Sp with PPV/NPV aligned to intended use (e.g. high NPV for rule-out)	Pre-specified non-inferiority margin versus standard assay for PPV/NPV; LR+/LR−; Brier score; stratify by specimen type/site	Reliability curves; slope/intercept
EBAI class C: novel AI-based biomarkers for outcome Prediction C1 prognostic	Risk of event independent of treatment (recurrence/OS)	Rank-based metrics, such as C-index for binary outcome, time-dependent AUC(t) for survival outcomes. IDI/NRI for comparing with prior models. DCA for clinical utility of AI based biomarkers	Risk-group KM separation	Calibration plot
C2 predictive	Predict differential treatment benefit	Treatment × Biomarker interaction (pre-specified) and effect size by subgroup (OR and 95% CI for binary outcomes and HR and 95% CI for HR for survival outcomes, ARR/ARD at fixed time)	Subgroup HRs with CIs; RMST difference	Calibration within each arm

Choose the primary metric and operating threshold to match the intended use; report calibration for any probabilistic output. AI, artificial intelligence; ARR/ARD, absolute risk reduction/difference; AUC(t), time-dependent area under the receiver operating characteristic curve; Brier score, mean squared error of probabilistic predictions; C-index, concordance index; CCC, concordance correlation coefficient; CI, confidence interval; EBAI, ESMO Basic Requirements for AI-based Biomarkers in Oncology; EGFR, epidermal growth factor receptor; H&E, haematoxylin–eosin; HER2, human epidermal growth factor receptor 2; HR, hazard ratio; ICC, intraclass correlation coefficient; IDI, integrated discrimination improvement; KM, Kaplan–Meier; LR+/LR−, positive/negative likelihood ratio; MAE, mean absolute error; MSI, microsatellite instability; NPA, negative percent agreement; NPV, negative predictive value; NRI, net reclassification improvement; OPA, overall percent agreement; OR: odds ratio; OS, overall survival; PD-L1, programmed death-ligand 1; PPA, positive percent agreement; PPV, positive predictive value; RMST, restricted mean survival time; Se/Sp, sensitivity/specificity.

- 'Replacement of molecular testing for class B': the biomarker must achieve 'error rates comparable to or better than' the variability observed across standard molecular assays. In this setting, 28.6% of experts indicated they would still recommend confirmatory testing before treatment decisions, reflecting ongoing caution toward fully autonomous AI-based diagnostics.
- 'Sample requirements': validation should be conducted on a 'sufficiently large, diverse and representative cohort', including both positive and negative cases, to ensure generalisability and clinical reliability.
- 'Independent validation': the biomarker should be tested on a dataset 'entirely independent' from the training and tuning data, that is, no overlap in data sources, patients, or collection processes. The validation dataset may come from a registry or a clinical trial, provided it reflects a contemporary patient population comparable to current clinical practice.

SPECIFIC GUIDANCE ON AI-BASED BIOMARKER CLASSES

EBAI class A AI systems

Definition of class A AI biomarkers. Class A systems automate tasks traditionally carried out by humans that are tedious and time consuming. These systems analyse the same input data as the human evaluator to automatically quantify established biomarkers, similarly to a human observer. Unlike human scoring, which is semi-quantitative and variable, AI systems can offer near-perfect reproducibility. The outputs are easily verifiable by experts, and the decision pathway follows predefined, interpretable rules. Due to their transparency and limited scope, these systems are considered low risk and can enhance efficiency, objectivity, and standardisation in routine workflows.

Examples for class A AI biomarkers. In histopathology, AI tools have been developed to quantify PD-L1 expression in non-small-cell lung cancer (NSCLC) from digitised immunohistochemistry (IHC) slides, where the output is a quantitative score of PD-L1 expression percentage that would normally be estimated by a pathologist.[46,47] Similarly, models exist for scoring HER2 expression across different tumour types, translating IHC staining into categorical outputs, a task traditionally based on subjective interpretation. Another application is the digital quantification of pathologic response on H&E-stained surgical specimens, a predictor of long-term outcomes after treatment with curative intent. In a retrospective analysis of the LCMC3 trial (NCT02927301), AI-based assessment of residual tumour after neoadjuvant therapy in early-stage NSCLC showed high concordance with manual scoring (AUC = 0.98).[48]

Beyond histopathology, in genomics, a deep learning model can be used to predict microsatellite instability (MSI) directly from genomic data. In this case, the input is the

same raw genomic sequencing data that would be used in conventional workflows, and the output is an MSI score or categorical classification (MSI-high, MSI-low, or microsatellite stable), but the intermediate analysis uses deep learning instead of traditional bioinformatics pipelines. Recent studies have shown that using an AI approach for this task is feasible and that a higher performance can be reached in cases in which only little data are available.[49] For example, if only a panel of genes is available, AI methods could potentially infer MSI status from their sequences more reliably than conventional tools.[50] These tools are class A because all maintain the same input data and output results as traditional methods while enhancing the intermediate analysis with greater consistency, speed, and potential accuracy.

Guidance. Based on the Delphi process, the following guidance is issued: For class A AI validation studies, a concordance study with analytical validation should be conducted, involving at least two experts who are blinded to AI results and each other's evaluations. A priori calculation of the sample size is recommended to ensure statistical robustness. Performance should be assessed using multiple metrics to provide a comprehensive evaluation (Table 2). Generalisability must be demonstrated through consistent results across different scanners or laboratory protocols (Figure 2).

EBAI class B AI systems

Background. In class B biomarkers, the deep learning model predicts a known biomarker in a fundamentally different way than previous methods. Several such systems are already on the market, where deep learning is used to predict biomarkers using alternative data sources. In these cases, the input data are different from what a gold standard test would use, which makes the risk and complexity of these biomarkers inherently higher than for class A. These ESMO AI class B tools may not be able to predict accurately all molecular biomarkers, but may be able to do so in a percentage of cases with sufficient sensitivity and specificity. This 'rule-out test approach' (where a negative result can reliably exclude the presence of a biomarker) is becoming increasingly common in practice, providing a potential alternative when standard testing methods are unavailable, impractical, or too costly. Importantly, EBAI class B AI systems usually aim at detecting an enriched population, rather than substituting an established genomic test.

Examples. A prominent example is an AI system that predicts MSI status from H&E-stained pathology slides of colon cancer, where the input is a standard histology slide and the output is MSI status that would traditionally require genomic sequencing with polymerase chain reaction (PCR) or next-generation sequencing (NGS).[51,52] Other applications include predicting mutation status in lung cancer [e.g. EAGLE (EGFR AI Genomic Lung Evaluation)][53] or estimating homologous recombination deficiency (HRD) from H&E images in breast and ovarian cancer.[54-57] The detection performance of molecular biomarkers remains highly variable across genes and cancer types, even when using large-scale AI models trained on massive datasets (e.g. Virchow2). This variability reflects differences in sample size in the training set, biomarker prevalence, and the strength of morphological signals, indicating that limitations are not solely attributable to data scarcity.[58]

Guidance. Based on the Delphi process, the following guidance is issued: For class B AI validation studies, an analytical study comparing AI performance with a gold-standard test is required, by using retrospective analyses of high-quality real-world data as defined per ESMO Guidance for Reporting Oncology real-World evidence (ESMO-GROW)[59] or retrospective analyses of clinical trial data. This panel agreed that, in principle, we should judge the clinical utility of AI solutions like any other biomarker in oncology. From that point of view, class B algorithms should require a strong analytical validation against other gold-standards.

As a pre-screening method for molecular alterations, class B tools may have two main consequences. (i) 'Prediction of a positive result' (presence of the biomarker) could increase the number of patients ultimately testing positive by expanding molecular testing in cases where it was previously unfeasible or not routinely carried out, thereby improving access to personalized therapies. Alternatively, it may enable rapid assays that allow faster identification of predicted alterations and shorten the time to treatment initiation. (ii) 'Prediction of a negative result' (absence of the biomarker) could help reduce the number of tests when the negative predictive value is very high. This latter approach carries substantial risk, however, as false-negative results could have important consequences for patient care. Negative results should therefore always be confirmed by the gold-standard method unless the AI-based tool has been demonstrated to be non-inferior or superior to it. In the EAGLE model for epidermal growth factor receptor (*EGFR*) prediction from H&E tissue slides, for example, patient risk was minimal because the AI system reduced the need for rapid *EGFR* testing carried out in parallel with NGS, thereby limiting tissue consumption and unnecessary duplicate testing, while patients continued to undergo systematic NGS analysis. In this case, in addition to analytical validation, the initial phase of clinical deployment included a prospective 'silent' trial. Such prospective assessments can enhance trust in AI-based pre-screening tools and demonstrate their feasibility within the genomic testing workflow.[53] Concerning the requirement of a prospective 'silent' trial for validation of class B, there was majority agreement, but not consensus among experts, suggesting that such an approach could be recommended but not strictly mandated.

If the intended use is replacing the gold standard based on excellent performance metrics that are equivalent or superior to it, a direct application of a digital AI tool for clinical decision-making, without any measurement of

protein or nucleic acid and relying solely on image analysis but lacking clinical validation of the biomarker, represents a conceptual shift that many oncologists remain uncomfortable with in these early stages of AI-driven diagnostics. As such, the consensus in this group is to require retrospective clinical validation of the class B AI solution, in addition to the analytical validation, at least for the foreseeable future (one to two years), until this clinical familiarisation allows class B solutions to follow the rules of any other clinical biomarker (Figure 2). When the predicted AI-based biomarker lacks sufficient granularity for clinical decision-making, however, such as in cases where the specific alteration subtype determines treatment eligibility (e.g. KRAS G12C inhibitors indicated only for KRAS p.G12C, not for all *KRAS* mutations), AI-based biomarkers, given their current capability to predict the presence of an alteration but not its subtype, can only serve as pre-screening tools.

EBAI class C AI systems

Background. Class C biomarkers are entirely novel in terms of input, method, and output. In these cases, a deep learning system is not trained to predict a known biological property or biomarker, but it is trained directly on clinical endpoints, for example on survival or treatment response data. This makes it an inherently new prognostic or predictive biomarker. Class C biomarkers are further divided into two subcategories: C1 (prognostic) and C2 (predictive). C1 prognostic biomarkers are used to identify the likelihood of a clinical event, disease recurrence, or progression in patients who have the cancer of interest, independent of the treatment received. C2 predictive biomarkers identify individuals who are more likely than similar individuals without the biomarker to experience a favourable or unfavourable effect from exposure to a certain treatment. For predictive biomarkers, it is critical to compare at least two distinct groups (e.g. treated versus untreated) within defined biomarker subgroups to ensure validity.

Examples. For C1 prognostic biomarkers, examples include AI systems that predict cancer recurrence risk from H&E pathology slides or CT images, where the input is standard histology or radiology images and the output is a risk score for cancer recurrence that is not based on known biomarkers but on novel patterns recognised by the AI.[60-64] Similarly, Sybil is a deep learning model that predicts future lung cancer risk up to six years from a single low-dose CT scan, without requiring clinical data or radiologist input.[14] Other examples include an AI-based risk stratification in early-stage lung adenocarcinoma using positron emission tomography (PET)/CT habitat imaging, which identifies radiological subtypes with independent prognostic value beyond clinicopathologic factors and circulating tumour DNA (ctDNA),[13] as well as a clinical—genetic machine learning model in metastatic castration-resistant prostate cancer that integrates ctDNA aneuploidy and pathogenic alterations to refine overall survival prediction beyond clinical variables.[65] In parallel, models like DeepHRD classify 1.8- to 3.1-fold more patients with HRD than genomic tests, correlating with improved outcomes in high-grade serous ovarian and metastatic breast cancers.[55]

For C2 predictive biomarkers, the ArteraAI Prostate Test is the first AI-based predictive biomarker included in the National Comprehensive Cancer Network (NCCN) guidelines. Validated in a phase III randomised trial (NRG/RTOG 9408), it uses digitised biopsy slides and clinical data to identify patients with localised prostate cancer who benefit from short-term androgen deprivation therapy, showing a significant reduction in distant metastasis only in biomarker-positive cases.[66]

Another example is an AI system that predicts immunotherapy response in lung cancer from histopathology images[67] or from chest CT images.[12] In these cases, routine image data are used to predict treatment outcomes without relying on established biomarkers like PD-L1 expression. Also, a multistain deep learning model was developed to predict benefit from neoadjuvant chemoradiotherapy in rectal cancer, which analyses pretreatment biopsy slides stained for immune markers to identify complex tissue-level features associated with response.[68] Another example undergoing validation is the trophoblast cell surface antigen 2 (TROP2) normalised membrane ratio (QCS-NMR), a computational pathology biomarker derived from quantitative image analysis of IHC in NSCLC. In the TROPION-Lung01 trial, this biomarker predicted response to datopotamab deruxtecan (Dato-DXd), with patients having TROP2 QCS-NMR-positive tumours showing significantly longer progression-free survival and higher response rates compared with those with QCS-NMR-negative status, while no such effect was observed with docetaxel.[69] These class C AI-derived biomarkers capture subtle and complex patterns beyond conventional clinical, histological, IHC or molecular markers, offering novel, data-driven predictions of treatment benefit not linked to any previously established biomarker (Figure 2). When a stable, measurable feature underpinning a class C signal is identified, it can be formalised as an established biomarker and subsequently evaluated under class A (same input) or class B (alternative input). For instance, TROP2 QCS-NMR was discovered as a treatment response predictor, classifying it as a class C2 biomarker. Future AI-derived biomarkers that estimate a validated QCS-NMR score, however, would be categorised as class A or B (depending on the input), as they would represent novel methods for predicting an already established biomarker.

Guidance. Based on the Delphi process, the following guidance is issued. For 'class C1 AI validation studies', comparison with multiple validated prognostic markers used in routine clinical care is required. The minimum requirement should be a retrospective analysis of high-quality real-world data as defined per ESMO-GROW[59] or retrospective analysis of clinical trial data. For 'class C2 AI validation studies', comparison with multiple validated predictive markers routinely used in a defined tumour type is required, if such comparators exist. For predictive

biomarkers, validation requires comparison between at least two distinct treatment groups to ensure predictive value, or comparison with an already validated predictive biomarker. Stronger consensus supported the need for prospective validation within a clinical trial when a biomarker is used to guide stratification for a new treatment. For biomarkers predicting response or resistance to standard therapies, there was majority approval (no consensus reached) that prospective validation might be needed to predict response to standard therapies and retrospective validation using clinical trial data or high-quality real-world data to validate biomarkers of resistance, given that the effects of standard treatments are well characterized and that pre-specified, confounding-adjusted [e.g. inverse probability of treatment weighting (IPTW)/propensity] and externally validated retrospective cohorts can robustly test the treatment—biomarker interaction for clinical decision support. Also, generalisability must be tested across multiple patient cohorts from various geographic areas.

ADDITIONAL CROSS-SECTIONAL ASPECTS

Although developed in an oncology context, many aspects of this framework apply across medical specialties. Core principles are relevant to AI-based tools in other diseases and in preventive settings.

Regulatory aspects in the EU and USA

AI-based biomarkers are regulated under existing frameworks. In the EU, these include the IVDR and MDR, and in the USA, these include the FDA's medical device regulatory framework and the framework applicable to laboratory-developed tests under the Clinical Laboratory Improvement Amendments (CLIA).[70,71] While these high-level frameworks and related guidance provide, to a certain extent, adequate coverage, there remains a need for more specific guidance at the implementation level. Further, while some guidance exists on the interplay between the IVDR and the MDR and the EU AI Act,[72] the interaction between regulatory frameworks may create some procedural complexity that would benefit from further streamlining and acceleration at the EU level. In the USA there is a clear separation between CLIA (laboratory-developed test, LDT) and FDA (Companion Diagnostic-CompDX) in terms of regulatory responsibilities. In the EU all [LDT (=in-house *in vitro* diagnostic, IH-IVD, in IVDR speech) and Conformité Européenne *in vitro* diagnostic (CE-IVD) products] are regulated under IVDR and MDR. The EBAI framework itself can contribute to addressing regulatory gaps by helping define specific guidance and 'common specifications' related to IVDR/MDR requirements, providing clearer pathways for both manufacturers/developers and regulators/notified bodies to follow.

Ethical aspects

The ethical deployment of AI-based biomarkers in oncology must ensure respect for patient autonomy, transparency, and informed consent. According to the EU General Data Protection Regulation (GDPR, Article 22), individuals have the right not to be subject to decisions based solely on automated processing that significantly affect them—particularly in the context of health data. This implies that, especially for high-impact AI tools (e.g. EBAI class C or EBAI class A and B with the intent of replacing the gold standard), patients must be informed that AI plays a role in their diagnosis or treatment planning. Consent becomes ethically imperative when AI outputs directly influence clinical decisions.

The Central Ethics Committee of the German Medical Association underscores that AI systems replacing physician judgment require explicit patient consent unless a clinician performs a plausibility check, maintaining human oversight.[73] Although the Ethics Guidelines for Trustworthy AI primarily require that deployers such as clinicians are informed about an AI system's capabilities and limitations, broader legal and ethical frameworks—including the GDPR and the AI Act—support the interpretation that patients should also be informed when AI meaningfully affects decisions about their care. This obligation arises not from the system's technical robustness, but from fundamental rights to transparency, autonomy, and contestability.

Moreover, the concept of fairness in AI should not be reduced to statistical bias mitigation. It extends to include distributive, procedural, and contextual dimensions where unequal data representation can lead to disparities in diagnostic accuracy and treatment access. Hence, ethical deployment requires alignment between the AI-based biomarker's validated target population and its actual use—what might better be termed 'target group applicability'. This ensures not only accuracy but clinical appropriateness across biological and demographic subgroups.[74,75]

RELATIONSHIP TO OTHER ESMO ACTIVITIES

With EBAI, we aim to provide guidance on the development and clinical applicability of AI-based biomarkers by recommending analytical and clinical validation requirements, which should facilitate the work of clinicians, scientists, and the industry in the development of these AI-based tests. Like any other biomarker or test in oncology, AI biomarkers would need to be used in the context of ESMO's Scale for Clinical Actionability of molecular Targets[76,77] (ESCAT), which ranks biomarkers to their clinical relevance to guide treatment decisions.

FUTURE PERSPECTIVES AND CONCLUSIONS

AI-based biomarkers currently function as standalone, single-purpose tools and should be validated accordingly. While medical AI is rapidly moving toward multipurpose, large language model-based systems, biomarker validation requires specialised frameworks due to the critical treatment decisions they inform. In oncology, where biomarker results guide life-altering therapies, distinct and rigorous standards are essential.

Improving AI models to potentially replace ground-truth methods involves several strategies. A key approach is integrating AI into hospital systems across diverse regions and cancer centres to continuously feed models with varied, expert-driven data. While continuous learning is technically feasible, it raises regulatory concerns. An alternative is to send input data, such as scanned pathology slides, radiological images, or other digital inputs, to a centralised analysis facility where updated models are applied. This enables consistent performance and controlled updates across sites, while also helping to reduce the gap between highly resourced centres and those lacking adequate infrastructure. Regardless of the approach, automated monitoring systems must track key metrics against baselines and flag deviations. Regular benchmarking against gold-standard methods remains critical, even for deployed systems.

Another challenge is the risk of hallucinations and erroneous predictions, particularly with complex inputs. This highlights the need for robust automated validation, especially when AI replaces human judgment or gold-standard testing. Any updated biomarker should, at minimum, be retrospectively compared with its prior version to ensure overall performance improvement.

Education is also vital. Medical training must include AI literacy to help future clinicians critically assess AI outputs and apply them as decision-support tools—not as autonomous agents—while maintaining clinical oversight.

Looking forward, AI in health care is shifting toward more autonomous, reasoning-capable agents. In this context, current biomarkers may become components within broader sophisticated AI systems. With the increased availability of validated AI-based biomarkers, we foresee the evolution of the EBAI framework to accommodate novel methods of validation, such as multiple agentic AI tools predicting the same output but trained on different datasets, or improved targeting of human review within human-in-the-loop systems, guided by AI-generated confidence scores. The acceptable performance threshold for an AI-based biomarker to progress from proof-of-concept to clinical validation will require further refinement, depending on the intended purpose of the biomarker and on a balanced assessment of its expected clinical impact relative to the resources invested and the current non-AI standard of care. While EBAI is well suited to today's landscape, where AI enhances rather than replaces conventional biomarker assessment, it may require refinement in future settings where AI-derived biomarkers themselves become standard and subject to further innovation. This evolution will require new validation frameworks, clearer accountability, and proactive governance.

ACKNOWLEDGEMENTS

This is a project initiated by the ESMO Precision Oncology Task Force and the ESMO Real World Data and Digital Health Task Force. We would like to thank ESMO leadership for their support in the development of this manuscript, as well as Svetlana Jezdic for administrative, logistical and scientific support. Furthermore, we thank the following colleagues for their critical comments on the manuscript: Joris van der Haar, Andreas Kleppe.

FUNDING

This work was supported by the European Society for Medical Oncology (no grant number).

DISCLOSURE

MA reports financial interest to the institution as a coordinating principal investigator from Amgen; financial interest from receipt of funding to the institution from AstraZeneca, Sandoz; non-financial interest from receipt of funding to the institution from Owkin; non-remunerated activity as a member and a member of the Education Committee of the International Association for the Study of Lung Cancer (IASLC). MST reports receipt of a fee to the institution for participation in advisory board from AstraZeneca, Incyte, Merck Sharp & Dohme (MSD), Roche, Sanofi; receipt of a fee to the institution as an invited speaker from Johnson and Johnson, PaigeAI; receipt of a fee to the institution for writing engagement from Eli Lilly; serving as a scientific advisor in Mindpeak, Sonrai; non-financial interest from receipt of research grants to the institution from MSD, Roche; non-financial interest to the institution for serving as a local principal investigator from Philips. AM reports receipt of a fee for participation in advisory board from Menarini/Stemline; receipt of a fee as an invited speaker from Roche; receipt of a fee for providing an expert testimony from Eli Lilly; receipt of travel support from AstraZeneca, Daiichi Sankyo, Menarini/Stemline.

AS reports receipt of a fee for participation in advisory board from Agilent, Aignostics, Amgen, Astellas, AstraZeneca, Bayer, Eli Lilly, Illumina, Janssen, Jazz Pharmaceuticals, MSD, Pfizer, Qlucore, QuIP, Sanofi, Taiho Oncology, Thermo Fisher; financial interest to the institution for participation in advisory board from BeiGene, Bristol Myers Squibb (BMS), Novartis, Takeda; receipt of a fee as an invited speaker from Incyte, Roche, Servier; non-financial interest from receipt of research grant to the institution from Bayer, BMS, Chugai, Incyte, MSD. MK reports financial interest to the institution for participation in advisory board from Bayer, MSD, Pierre Fabre, Servier; financial interest to the institution as an invited speaker from BMS, IQVIA, Merck; receipt of funding to the institution from Amgen, BMS, Merck, Nordic Pharma, Novartis, Pierre Fabre, Servier; receipt of research grants to the institution from Bayer, BMS, Merck, Personal Genomics Diagnostics, Pierre Fabre, Roche, Servier, Sirtex; financial interest to the institution for serving as a trial chair from Servier; non-remunerated activity as a principal investigator of the Dutch Prospective Colorectal Cancer Group, European Society for Medical Oncology (ESMO) faculty member for Gastrointestinal Tumours, Chair of the Real World Data and Digital Health

Working Group. APr reports receipt of a fee for participation in advisory board from Amgen, AstraZeneca, Bayer, BMS, Janssen, MSD, Pfizer; receipt of a fee as an invited speaker from AstraZeneca, Daiichi Sankyo, Gilead, IQVIA, Elli Lilly, MEDSIR, Novartis, Pfizer, Roche; receipt of a fee for training of personnel from AstraZeneca, Italfarma; receipt of travel grant from Janssen; financial interest to the institution for serving as a coordinating principal investigator from AstraZeneca, Spectrum; financial interest to the institution for serving as a local principal investigator from Bayer, BMS, Eli Lilly, MSD, Roche; non-financial interest for serving as a project lead in APOLLO 11, non-financial interest for serving as a President of the European Interdisciplinary Society of AI in Cancer Research (ESAC), non-financial interest for serving as a member of Board of Directors of the Network Italiano per la Bioterapia dei Tumori (NIBIT). KLK reports financial interest from receipt of research grant to the institution and for serving as a coordinating principal investigator from Meta; non-remunerated activities as a member of the American Association for Cancer Research (AACR) and leadership role as AACR Project GENIE Chair 2025-2027, member of the American Society of Clinical Oncology (ASCO). SG reports receipt of a fee for participation in advisory board from Prova Health; receipt of a fee for consultant activities from Ada Health GmbH, Flo Ltd, ICURA ApS, Lindus Health Ltd, Prova Health, Rock Health Inc., Thymia Ltd, Una Health GmbH; receipt of a fee for delivering training courses from FORUM Institut fur Management GmbH; receipt of a fee as a due diligence assessor from High-Tech Grunderfonds Management GmbH; receipt of a fee as an advisor/contributor to trainings from Prova Health; owning stocks/shares from Ada Health GmbH; non-financial interest for participation in advisory board and advisory role from Cansilico GmbH; non-financial interest for advisory role from DG Sante, European Commission and Advisory Group member of the EY-coordinated 'Study on Regulatory Governance and Innovation in the field of Medical Devices' conducted on behalf of the DG Sante of the European Commission. MEL reports receipt of a fee as an invited speaker from AstraZeneca, GWT-TUD. JL reports receipt of a fee to the institution as an invited speaker from British Columbia University, Cell Pharma, Dalhousie University; receipt of a fee for consultant role for Daiichi Sankyo and Mayo Clinic, non-financial interest for a person of interest position from Harvard Medical School. LP reports receipt of a fee as an invited speaker from Merck, Novartis, Pfizer. FMB reports receipt of a fee for participation in advisory board from Daiichi Sankyo, FogPharma, Harbinger Health, Incyte, Karyopharm, Mersana Therapeutics, Protai, Sanofi Pharmaceuticals, Seagen, Theratechnologies, Zentalis; receipt of a fee for providing consultancy from AstraZeneca, Becton Dickinson, Calibr, Debiopharm, EcoR1, eFFECTOR Therapeutics, Elevation Oncology, Exelixis, GT Aperion Therapeutics, Infinity Pharmaceuticals, Jazz Pharmaceuticals, LegoChem Biosciences, Lengo Therapeutics, Loxo Oncology, Menarini Group, Molecular Templates, OnCusp Therapeutics, Ribometrix, Tallac Therapeutics, Tempus, Zymeworks; receipt of a fee as an invited speaker from Dava Oncology; non-financial interest from receipt of research grants to the institution from Aileron Therapeutics, AstraZeneca, Bayer Healthcare, CytomX Therapeutics Inc., Daiichi Sankyo Co. Ltd, eFFECTOR Therapeutics, Puma Biotechnology, Repare Therapeutics, Taiho Pharmaceuticals Co., Takeda Pharmaceuticals Co.; non-financial interest to the institution for serving as a coordinating principal investigator from AstraZeneca; non-financial interest to the institution for serving as a local principal investigator from Aileron Therapeutics, AstraZeneca, Bayer Healthcare, Calithera Biosciences, Curis Inc., CytomX Therapeutics Inc., Daiichi Sankyo Co. Ltd, Debiopharm International, eFFECTOR Therapeutics, Genentech Inc., Guardant Health Inc., Jazz Pharmaceuticals, KLUS Pharma, Novartis, Taiho Pharmaceuticals Co., Zymeworks; non-financial interest to the institution for serving as a steering committee member from Genentech Inc.; receipt of honoraria from Dava Oncology; receipt of travel support from Cholangiocarcinoma Foundation, Dava Oncology, European Organisation for Research and Treatment (EORTC), ESMO. SH reports receipt of a fee for participation in advisory board from Novartis; receipt of a fee as a member of data and safety monitoring board from BeiGene, BMS, CG Oncology, Janssen, Sanofi; financial interest from receipt of research funding to the institution from ASCO; receipt of a fee as an Associate Editor from ASCO. JW reports financial interest from receipt of funding to the institution from Siemens; receipt of a fee as an invited speaker from ESMO; non-financial interest as a principal investigator from CPRIT, National Institutes of Health (NIH). APe reports receipt of a fee for participation in advisory board from Merck, Takeda; receipt of a fee as an invited speaker from Servier; receipt of a travel grant from Ipsen, Merck; non-remunerated activity as a member of the Fondation française de cancérologie digestive and Groupe des tumeurs neuroendocrines. KPMS reports receipt of a fee to the institution for participation in advisory board from AbbVie, BMS, Pierre Fabre; financial interest from receipt of research grants to the institution from BMS, Genmab, Novartis, Philips, TigaTx; receipt of a fee to the institution for participation in the safety review committee from Sairopa. BB reports receipt of a fee to the institution for participation in advisory board from AbbVie, Beijing Avistone Biotechnology, BioNTech SE, BMS, CureVac AG, PharmaMar, Regeneron, Sanofi Aventis; receipt of a fee to the institution as an invited speaker from AbbVie, AstraZeneca, BMS, Daiichi Sankyo, Eli Lilly, MSD, Ose Immunotherapeutics, Sanofi, Servier; receipt of a fee to the institution for providing an expert testimony from AbbVie, BMS, CureVac AG, Eli Lilly, Ellipse Pharma Ltd, F. Hoffmann-La Roche Ltd, Foghorn Therapeutics Inc., Genmab, Immunocore, Owkin, Sanofi; financial interest to the institution for serving as a steering committee member from Amgen, BeiGene, CureVac, Genmab, Janssen, MSD, Takeda; financial interest to the institution for serving as a coordinating principal investigator from AstraZeneca, Ose Immunotherapeutics, PharmaMar, Sanofi, Taiho; financial interest to the institution for serving as a local principal

investigator from Daiichi Sankyo, Eli Lilly, Ellipsis, Enliven, Nuvalent, Prelude Therapeutics; non-remunerated leadership role as a Chair of the Scientific Chairs Council in EORTC, non-remunerated advisory role as a member of the Scientific Board in the Intergroupe Francophone de Cancérologie Thoracique (IFCT). BR reports receipt of a fee for participation in advisory board from Novartis, Roche; receipt of a fee as an invited speaker from MSD SE; receipt of honoraria to the Melanoma Patient Network Europe (MPNE) for support of its activities from BMS, Pierre Fabre; non-financial interest for leadership role as a founder of MPNE. CM reports receipt of a fee for participation in advisory board from Daiichi Sankyo, Menarini, Roche; receipt of a fee as an invited speaker from Bayer, Illumina, Veracyte; non-financial interest as a member of the ESMO Precision Oncology Working Group. MCO reports full or part time employment in 52 North Health; receipt of a fee to the institution as an invited speaker from AI for Global Goals, GlaxoSmithKline (GSK); financial interest from receipt of funding to the institution from AstraZeneca, GE HealthCare. DF reports full or part time employment in Synagen; ownership interest in Synagen; holding stocks/shares in BioNTech SE, Synagen; financial interest both personal and institutional from receipt of research grant from OpenAI. CP reports receipt of a fee to the institution as an invited speaker from Roche; non-remunerated activity as a member of the International Skeletal Society (ISS), Swiss Society of Pathology (SGPATh), United States and Canadian Academy of Pathology (USCAP). AV reports non-financial interest from receipt of research grants to the institution from MSD, Roche; non-financial interest to the institution for serving as a coordinating principal investigator from AstraZeneca, BMS, Novartis. TJB reports receipt of a fee for participation in advisory board from HEINE Optotechnik, Germany; receipt of a fee as an invited speaker from Novartis, Roche; ownership interest in Smart Health Heidelberg GmbH. JM reports receipt of a fee for participation in advisory board from Amgen, Amunix Pharmaceuticals, AstraZeneca, Janssen, Pfizer, Roche; receipt of a fee to the institution for participation in advisory board from Nuage Therapeutics; receipt of a fee as an invited speaker from AstraZeneca, Guardant Health, MSD; non-financial interest from receipt of research grants to the institution from Amgen, AstraZeneca, Pfizer Oncology; non-financial interest from receiving product samples for access to drugs in early development for preclinical testing from AstraZeneca. NH reports receipt of a fee for participation in advisory board from Aptitude Health, Gilead, Pfizer, Sandoz-Hexal, Sanofi, Seagen; receipt of a fee as an invited speaker from Art Tempi, AstraZeneca, Daiichi Sankyo, Gilead, Eli Lilly, Medscape, MSD, Novartis, Onkowissen, Pierre Fabre, Roche, Sanofi, Seagen, Viatris; receipt of a fee for participation in independent data monitoring committee from Roche; spouse receipt of honoraria from WSG; ownership interest in West German Study Group; non-financial interest from receipt of funding to the institution from BMS, Daiichi Sankyo, Gilead, MSD, Roche, Seagen, TRIO, WSG; non-financial interest to the institution for serving as a coordinating principal investigator from AstraZeneca; non-financial interest to the institution for participation as a steering committee member from Eli Lilly, Pierre Fabre; non-remunerated activity as a member of the AGO Breast Committee, German AGO Breast Guideline Committee, non-remunerated activity as a member ESO/ESCO Breast Cancer Educational Programs, non-remunerated activity as a founding editor of the *Breast Care* journal. ECW reports receipt of a fee as an invited speaker from Janssen Cilag GmBH, Roche Diagnostics; full or part-time employment at Heidelberg University Hospital — National Center for Tumor Diseases (NCT) as a Director of the Institute for Medical and Data Ethics, Medical Faculty, Heidelberg University and a Senior Physician and Managing Director of the NCT Heidelberg; non-financial interest from a leadership role in 1+ Million Genomes (1+MG) initiative of European Commission for participation in the Working Group ELSI (ethical, legal, and social implications) as Appointed Member by the German Federal Ministry of Health (BMG), as a member of Board of Directors and Vice President of the German Academy for Ethics in Medicine (AEM), for advisory role and as an invited member of the interdisciplinary Working Group 'Gene technology report' of the Berlin Brandenburg Academy of Science, for leadership role in the Working Group on Medicine and Ethics in DGHO — German Society for Haematology and Medical Oncology e.V., for advisory role in the Ethics in Board Genomic Data Infrastructure (GDI) Project as Appointed Member by the German Federal Ministry of Health (BMG), for leadership role in German Ethics Council (travel fees and meeting honoraria reimbursed), as a member of Board of Directors and participation in the Working Group ELSI of the GHGA (German Human Genome-Phenome Archive), for advisory role in the Regulatory & Ethics Working Group Robert Koch Institute (RKI) in Global Alliance for Genomics & Health (GA4GH), for leadership role in ZfKD—Centre for Cancer Registry Data as Appointed Member by the German Federal Ministry of Education and Research (BMBF), as a member of Board of Directors in ZEKO (Central Ethics Committee of the German Medical Association). FLR reports receipt of a fee for participation in advisory board from AbbVie, AstraZeneca, Bayer, BMS, Daiichi Sankyo, Janssen, Eli Lilly, MSD, Pfizer, Roche, Takeda, Thermo Fisher; receipt of a fee as an invited speaker from AstraZeneca, Bayer, BMS, Janssen, Eli Lilly, MSD, Pfizer, Roche, Takeda, Thermo Fisher; financial interest from receipt of research grants to the institution from AstraZeneca, Janssen, Eli Lilly, Pfizer, Roche, Thermo Fisher. RPL reports non-financial interest from receipt of research grants to the institution from AstraZeneca, Roche. GP reports full time employment in ESMO; non-remunerated activity as a member of ASCO, Hellenic Cooperative Oncology Group (HeCOG), Hellenic Society of Medical Oncology (HeSMO). SD reports receipt of a fee to the institution for participation in advisory board from Elsan, Gilead, Novartis; receipt of a fee to the institution as an invited speaker from Gilead, MSD; financial interest to the institution for serving as a steering committee member from BMS, Roche Genentech, Sanofi, Taiho;

non-financial interest as a project lead from national funding for research project from Banque des Territoires/France 2030, non-financial interest as a principal investigator from the European Commission (H2020 funding); non-remunerated activity as a member of the Board of Directors of the Société Française de Sénologie et Pathologie Mammaire (SFSPM). CBW reports receipt of a fee for participation in advisory board from BMS, Frankfurt Institute of Clinical Cancer Research (IKF), Incyte, Johnson & Johnson, Roche; receipt of a fee as an invited speaker from Amgen, AstraZeneca, Bayer, BMS, GSK, Johnson & Johnson, Eli Lilly, Merck, MSD, Pierre Fabre, QuIP GmbH, Roche, Servier, Taiho; receipt of a fee for providing an expert testimony from Johnson & Johnson; receipt of travel support from Bayer, Johnson & Johnson, Roche, Servier, Taiho; non-financial interest from receipt of research grant both personal and to the institution from Roche; non-remunerated activity as an officer in the AIO-Arbeitsgemeinschaft Internistische Onkologie (Germany); non-financial interest for advisory role in EU Commission—DG RTD as a member of the EU Commission Mission Board for Cancer, non-financial interest for advisory role in German Federal Ministry of Research, Technology and Space, Member Strategy Board Nationale Dekade gegen Krebs. JNK reports receipt of a fee for participation in advisory board from AstraZeneca, DoMore Diagnostics, Owkin, Panakeia, Synagen GmbH; receipt of a fee as an invited speaker from AstraZeneca, Bayer, BMS, Daiichi Sankyo, Fresenius, Janssen, Merck, MSD, Pfizer, Roche; receipt of a fee for providing consultancy from Bioptimus, Mindpeak, MultiplexDx; no personal remuneration from holding shares in Ignition Lab GmbH, owning shares and part-time activities in StratifAI GmbH, owning shares in Synagen GmbH; non-financial interest to the institution for serving as a local principal investigator from GSK. All other authors have declared no conflicts of interest.

REFERENCES

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38.
2. Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat Cancer*. 2022;3(9):1026-1038.
3. Marra A, Morganti S, Pareja F, et al. Artificial intelligence entering the pathology arena in oncology: current applications and future perspectives. *Ann Oncol*. 2025;36(7):712-725.
4. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer*. 2021;124(4):686-696.
5. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*. 2001;69(3):89-95.
6. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst*. 2009;101(21):1446-1452.
7. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol*. 2021;18(7):465-478.
8. Zeng Q, Klein C, Caruso S, et al. Artificial intelligence-based pathology as a biomarker of sensitivity to atezolizumab-bevacizumab in patients with hepatocellular carcinoma: a multicentre retrospective study. *Lancet Oncol*. 2023;24(12):1411-1422.
9. Lipkova J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. 2022;40(10):1095-1110.
10. Ligero M, El Nahhas OSM, Aldea M, Kather JN. Artificial intelligence-based biomarkers for treatment decisions in oncology. *Trends Cancer*. 2025;11(3):232-244.
11. Salto-Tellez M, Maxwell P, Hamilton P. Artificial intelligence-the third revolution in pathology. *Histopathology*. 2019;74(3):372-376.
12. Saad MB, Hong L, Aminu M, et al. Predicting benefit from immune checkpoint inhibitors in patients with non-small-cell lung cancer by CT-based ensemble deep learning: a retrospective study. *Lancet Digit Health*. 2023;5(7):e404-e420.
13. Sujit SJ, Aminu M, Karpinets TV, et al. Enhancing NSCLC recurrence prediction with PET/CT habitat imaging, ctDNA, and integrative radiogenomics-blood insights. *Nat Commun*. 2024;15(1):3152.
14. Mikhael PG, Wohlwend J, Yala A, et al. Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *J Clin Oncol*. 2023;41(12):2191-2200.
15. Yoo S-K, Fitzgerald CW, Cho BA, et al. Prediction of checkpoint inhibitor immunotherapy efficacy for cancer using routine blood tests and clinical data. *Nat Med*. 2025;31(3):869-880.
16. Jee J, Fong C, Pichotta K, et al. Automated real-world data integration improves cancer outcome prediction. *Nature*. 2024;636(8043):728-736.
17. Derraz B, Breda G, Kaempf C, et al. New regulatory thinking is needed for AI-based personalised drug and cell therapies in precision oncology. *NPJ Precis Oncol*. 2024;8(1):23.
18. Calderaro J, Morement H, Penault-Llorca F, Gilbert S, Kather JN. The case for homebrew AI in diagnostic pathology. *J Pathol*. 2025;266(4-5):390-394.
19. Geaney A, O'Reilly P, Maxwell P, James JA, McArt D, Salto-Tellez M. Translation of tissue-based artificial intelligence into clinical practice: from discovery to adoption. *Oncogene*. 2023;42(48):3545-3555.
20. Center for Devices, Radiological Health. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. U.S. Food and Drug Administration. Available at https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Accessed July 4, 2025.
21. EUDAMED. Available at https://ec.europa.eu/tools/eudamed/#/screen/home. Accessed July 4, 2025.
22. Chouffani El Fassi S, Abdullah A, Fang Y, et al. Not all AI health tools with regulatory authorization are clinically validated. *Nat Med*. 2024;30(10):2718-2720.
23. Hanna MG, Olson NH, Zarella M, et al. Recommendations for performance evaluation of machine learning in pathology: a concept paper from the College of American Pathologists. *Arch Pathol Lab Med*. 2024;148(10):e335-e361.
24. Gu Q, Patel A, Hanna MG, et al. Bridging the clinical-computational transparency gap in digital pathology. *Arch Pathol Lab Med*. 2025;149(3):276-287.
25. Diamond IR, Grant RC, Feldman BM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. *J Clin Epidemiol*. 2014;67(4):401-409.
26. Franc JM, Hung KKC, Pirisi A, Weinstein ES. Analysis of Delphi study 7-point linear scale data by parametric methods: use of the mean and standard deviation. *Method Innov*. 2023;16(2):226-233.
27. Umeton R, Kwok A, Maurya R, et al. GPT-4 in a cancer center — institute-wide deployment challenges and lessons learned. *NEJM AI*. 2024;1(4).
28. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. [Preprint.] arXiv Advance Access published on November 6, 2020, doi: https://arxiv.org/abs/1810.03292
29. Mahmood U, Shrestha R, Bates DDB, et al. Detecting spurious correlations with sanity tests for artificial intelligence guided radiology systems. *Front Digit Health*. 2021;3:671015.
30. Šimić I, Veas E, Sabol V. A comprehensive analysis of perturbation methods in explainable AI feature attribution validation for neural time series classifiers. *Sci Rep*. 2025;15(1):26607.

31. Halabi S, Li C, Luo S. Developing and validating risk assessment models of clinical outcomes in modern oncology. *JCO Precis Oncol*. 2019;3:PO.19.00068.
32. Uno H, Tian L, Cai T, Kohane IS, Wei LJ. A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Stat Med*. 2013;32(14):2430-2442.
33. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak*. 2008;8:53.
34. Riley RD, Ensor J, Snell KIE, et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. *Lancet Digit Health*. 2025;7(6):100857.
35. Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Stat Med*. 1997;16(13):1529-1542.
36. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Stat Med*. 2015;34(4):685-703.
37. Pavlou M, Qu C, Omar RZ, et al. Estimation of required sample size for external validation of risk models for binary outcomes. *Stat Methods Med Res*. 2021;30(10):2187-2206.
38. Riley RD, Collins GS, Ensor J, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat Med*. 2022;41(7):1280-1295.
39. Kattan MW, Hess KR, Amin MB, et al. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. *CA Cancer J Clin*. 2016;66(5):370-374.
40. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744.
41. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762.
42. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al. METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. *Insights Imaging*. 2024;15(1):8.
43. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14(1):75.
44. Lambin P, Woodruff HC, Mali SA, et al. Radiomics Quality Score 2.0: towards radiomics readiness levels and clinical translation for personalized medicine. *Nat Rev Clin Oncol*. 2025;22(11):831-846.
45. Sarkans U, Chiu W, Collinson L, et al. REMBI: Recommended Metadata for Biological Images-enabling reuse of microscopy data in biology. *Nat Methods*. 2021;18(12):1418-1422.
46. Molero A, Hernandez S, Alonso M, et al. Assessment of PD-L1 expression and tumour infiltrating lymphocytes in early-stage non-small cell lung carcinoma with artificial intelligence algorithms. *J Clin Pathol*. 2025;78(7):456-464.
47. Herbst RS, Prizant H, Ruderman D, et al. Digital versus manual PD-L1 scoring in advanced NSCLC from the IMpower110 and IMpower150 trials. *J Thorac Oncol*. 2025. https://doi.org/10.1016/j.jtho.2025.07.131.
48. Dacic S, Travis WD, Giltnane JM, et al. Artificial intelligence-powered assessment of pathologic response to neoadjuvant atezolizumab in patients with NSCLC: results from the LCMC3 study. *J Thorac Oncol*. 2024;19(5):719-731.
49. Anaya J, Sidhom J-W, Mahmood F, Baras AS. Multiple-instance learning of somatic mutations for the classification of tumour type and the prediction of microsatellite status. *Nat Biomed Eng*. 2023;8:57-67.
50. Unger M, Loeffler CML, Žigutytė L, et al. Deep learning for biomarker discovery in cancer genomes. [Preprint.] *bioRxiv*. Advance Access published on Jan 8, 2025, https://www.biorxiv.org/content/10.1101/2025.01.06.631471v2
51. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*. 2019;25(7):1054-1056.
52. Saillard C, Dubois R, Tchita O, et al. Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides. *Nat Commun*. 2023;14(1):6695.
53. Campanella G, Kumar N, Nanda S, et al. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. *Nat Med*. 2025;31(9):3002-3010.
54. Luan H, Hu T, Hu J, et al. Breast cancer homologous recombination deficiency prediction from pathological images with a sufficient and representative Transformer. *NPJ Precis Oncol*. 2025;9(1):160.
55. Bergstrom EN, Abbasi A, Díaz-Gay M, et al. Deep learning artificial intelligence predicts homologous recombination deficiency and platinum response from histologic slides. *J Clin Oncol*. 2024;42(30):3550-3560.
56. El Nahhas OSM, Loeffler CML, Carrero ZI, et al. Regression-based Deep-Learning predicts molecular biomarkers from pathology slides. *Nat Commun*. 2024;15(1):1253.
57. Lazard T, Bataillon G, Naylor P, et al. Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images. *Cell Rep Med*. 2022;3(12):100872.
58. Wang YK, Tydlitatova L, Kunz JD, et al. Screen them all: high-throughput pan-cancer genetic and phenotypic biomarker screening from H&E whole slide images. [Preprint.] *arXiv* Advance Access published on July 14, 2025, doi: https://arxiv.org/abs/2408.09554v4
59. Castelo-Branco L, Pellat A, Martins-Branco D, et al. ESMO Guidance for Reporting Oncology real-World evidence (GROW). *ESMO Real World Data Digit Oncol*. 2023;1:100003.
60. Skrede O-J, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet*. 2020;395(10221):350-360.
61. Kleppe A, Skrede O-J, De Raedt S, et al. A clinical decision support system optimising adjuvant chemotherapy for colorectal cancers by integrating deep learning and pathological staging markers: a development and validation study. *Lancet Oncol*. 2022;23(9):1221-1232.
62. Jiang X, Hoffmeister M, Brenner H, et al. End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study. *Lancet Digit Health*. 2024;6(1):e33-e43.
63. Volinsky-Fremond S, Horeweg N, Andani S, et al. Prediction of recurrence risk in endometrial cancer with multimodal deep learning. *Nat Med*. 2024;30(7):1962-1973.
64. Huang YQ, Chen XB, Cui YF, et al. Enhanced risk stratification for stage II colorectal cancer using deep learning-based CT classifier and pathological markers to optimize adjuvant therapy decision. *Ann Oncol*. 2025;36(10):1178-1189.
65. Halabi S, Guo S, Luo B, et al. Pathogenic genomic alterations in circulating tumor DNA predict overall survival in men with metastatic castrate-resistant prostate cancer. *Eur Urol*. 2025. https://doi.org/10.1016/j.eururo.2025.07.011.
66. Spratt DE, Tang S, Sun Y, et al. Artificial intelligence predictive model for hormone therapy use in prostate cancer. *NEJM Evid*. 2023;2(8):EVIDoa2300023.
67. Rakaee M, Tafavvoghi M, Ricciuti B, et al. Deep learning model for predicting immunotherapy response in advanced non-small cell lung cancer. *JAMA Oncol*. 2025;11(2):109-118.
68. Foersch S, Glasner C, Woerl A-C, et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat Med*. 2023;29(2):430-439.
69. Garassino MC, Sands J, Paz-Ares L, et al. PL02.11 normalized membrane ratio of TROP2 by quantitative continuous scoring is predictive of clinical outcomes in TROPION-lung 01. *J Thorac Oncol*. 2024;19(10):S2-S3.
70. van Genderen ME, Kant IMJ, Tacchetti C, Jovinge S. Moving toward implementation of responsible artificial intelligence in health care: the European TRAIN initiative. *JAMA*. 2025;333(17):1483-1484.

71. Furtado LV, Ikemura K, Benkli CY, et al. General applicability of existing College of American Pathologists accreditation requirements to clinical implementation of machine learning-based methods in molecular oncology testing. *Arch Pathol Lab Med*. 2025;149(4):319-327.
72. AIB 2025-1 MDCG 2025-6 interplay between the medical devices Regulation (MDR) & in vitro diagnostic medical devices Regulation (IVDR) and the artificial intelligence act (AIA). Available at https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4_en?filename=mdcg_2025-6_en.pdf. Accessed November 5, 2025.
73. Stellungnahme der Zentralen Kommission zur Wahrung ethischer Grundsätze in der Medizin und ihren Grenzgebieten (Zentrale Ethikkommission) bei der Bundesärztekammer. Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz. *Deutsches Ärzteblatt Online*. Available at http://daebl.de/UB11; 2021. Accessed June 4, 2025.
74. Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz. Zentrale Ethikkommission. Available at https://www.zentrale-ethikkommission.de/stellungnahmen/kuenstliche-intelligenz-2021. Accessed June 4, 2025.
75. Regulation - 2016/679 - EN - gdpr - EUR-Lex. Available at https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng. Accessed June 4, 2025.
76. Mateo J, Chakravarty D, Dienstmann R, et al. A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). *Ann Oncol*. 2018;29(9):1895-1902.
77. Mosele MF, Westphalen CB, Stenzinger A, et al. Recommendations for the use of next-generation sequencing (NGS) for patients with advanced cancer in 2024: a report from the ESMO Precision Medicine Working Group. *Ann Oncol*. 2024;35(7):588-606.