Clinical Oncology 34 (2022) 102–113

Contents lists available at ScienceDirect

Clinical Oncology

journal homepage: www.clinicaloncologyonline.net

Evaluation of Prognostic and Predictive Models in the Oncology Clinic

M. Craddock [*], C. Crockett [†], A. McWilliam [*], G. Price [*], M. Sperrin [‡], S.N. van der Veer [‡], C. Faivre-Finn [*†]

[*] *University of Manchester, Radiotherapy Related Research Group, Division of Cancer Sciences, School of Medical Sciences, Manchester, UK*

[†] *Department of Clinical Oncology, The Christie NHS Foundation Trust, Manchester, UK*

[‡] *Centre for Health Informatics, Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK*

Abstract

Predictive and prognostic models hold great potential to support clinical decision making in oncology and could ultimately facilitate a paradigm shift to a more personalised form of treatment. While a large number of models relevant to the field of oncology have been developed, few have been translated into clinical use and assessment of clinical utility is not currently considered a routine part of model development. In this narrative review of the clinical evaluation of prediction models in oncology, we propose a high-level process diagram for the life cycle of a clinical model, encompassing model commissioning, clinical implementation and ongoing quality assurance, which aims to bridge the gap between model development and clinical implementation.

Key words: Life cycle; Model; Oncology; Personalised; Prediction; Radiotherapy

Introduction

Oncology ranks among the most complex disciplines of modern medicine. The inherent heterogeneity of cancer, patients and the ever-expanding number of treatment options make the selection of optimal treatment regimens more challenging than ever. Clinicians have to balance evidence from clinical trials with ongoing research, their own professional experience, national guidelines and patient's values to determine the 'ideal' treatment. This complex, multifactorial decision making process inevitably results in heterogeneity in practice, particularly in the management of patients from groups under-represented in traditional clinical trials, such as ethnic minorities [1], the elderly and comorbid [2]. Figure 1 describes a case study of such a patient.

New forms of evidence are required to better support clinical decision making. The wealth of routine data generated by every patient treated, allied with it now being stored in a digital and more accessible format, offers great potential to provide this new evidence through the development of prognostic and predictive models. These models are traditionally developed using established statistical methods, such as logistic or Cox regression analysis, with the use of advanced data science techniques, including machine learning and artificial intelligence now becoming increasingly common. They can provide personalised estimates of key clinical outcomes, such as the probability of local tumour control, overall survival and treatment-related toxicity. These personalised estimates offer the potential to improve treatment selection and with it, increase efficiency, patient satisfaction, quality of life and potentially even survival.

A clinical prediction model (CPM) may be defined as a mathematical combination of known clinical, biological and treatment factors that together can be used to estimate the probability of an individual experiencing a specific outcome. According to the definitions of Clark *et al.* [3], prognostic factors may be defined by association with a clinical outcome in the absence of therapy or with the application of a standard therapy, whereas predictive

Author for correspondence: M. Craddock, Radiotherapy Related Research group, University of Manchester, Dept 58 The Christie NHS Foundation Trust, Wilmslow Road, Manchester M20 4BX, UK.

E-mail address: matthew.craddock@postgrad.manchester.ac.uk (M. Craddock).

https://doi.org/10.1016/j.clon.2021.11.022

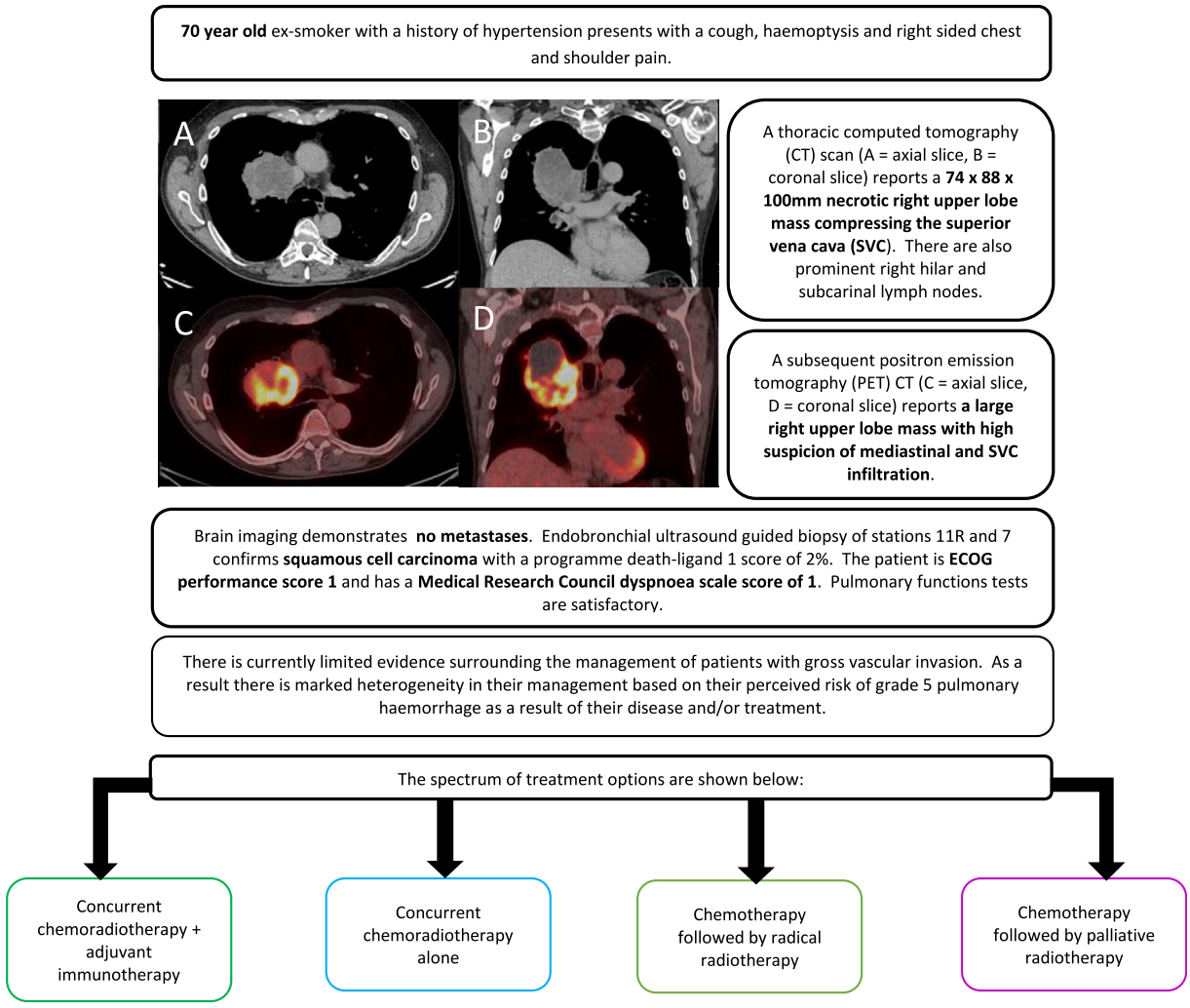

Fig 1. An exemplar case demonstrating the heterogeneity in clinical practice, where the use of a clinical prediction model may be of particular benefit.

factors are associated with the response to a particular therapy. Predictive factors imply a differential benefit of therapy that is dependent on the value of the factor. A CPM may be either prognostic or predictive, with prognostic models being more common.

A major limitation of prediction models is that they cannot provide any information regarding interventions or outcomes beyond those included in the training data. Informed clinical decision making requires the contextual knowledge of how a patient may respond to alternative therapies or no treatment (counterfactual outcomes). Causal predictive models show promise in being able to provide this information. Causal modelling goes beyond statistical association by combining the information contained in the training data with expert knowledge of the complex causal relationships that govern patient outcomes to allow the estimation of outcomes under hypothetical interventions [4,5]. The Predict breast tool is an example of a CPM enriched by causal reasoning, as it uses population average outcomes from clinical trials to provide survival estimates for different treatment combinations [6].

In this narrative review we evaluate the current utilisation of CPMs in the field of oncology and contrast this with the wider medical setting. We explore the challenges and opportunities presented by their widespread adoption and emphasise the necessity of assessment of clinical utility and involvement of end users throughout the model commissioning process. A high-level process diagram for the life cycle of a CPM is proposed (Figure 2), which describes the major processes of commissioning and ongoing quality assurance, with clinical utility considered from the outset.

Current Status of Models in Oncology

Despite the publication of hundreds of clinical models in the field of oncology, examples of widespread clinical implementation remain limited. In the wider healthcare

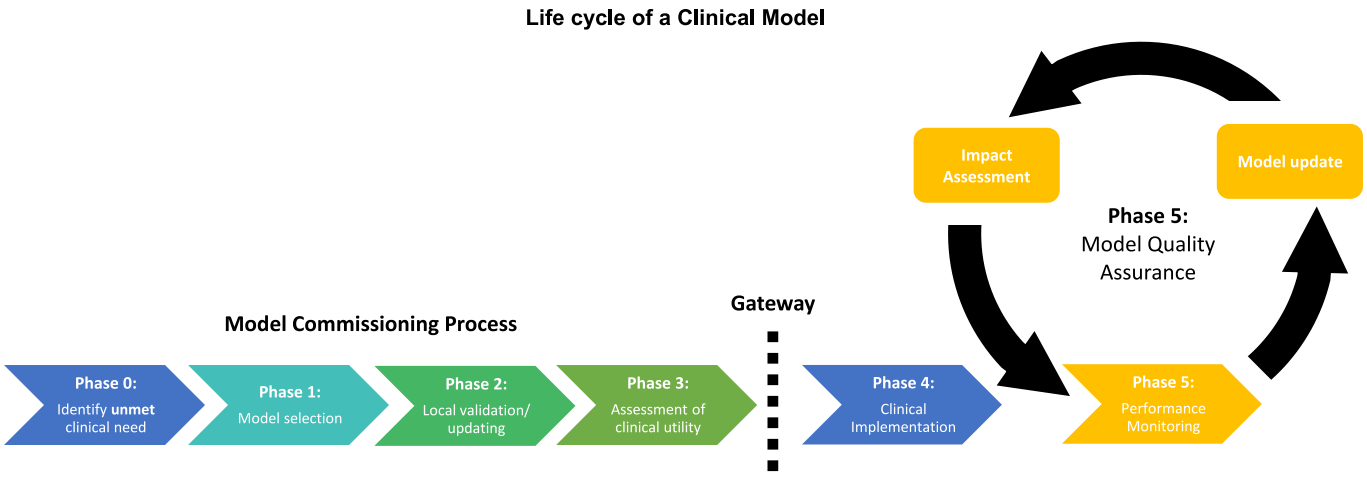

Fig 2. Process diagram for the life cycle of a clinical prediction model, encompassing the commissioning process, clinical implementation and the ongoing quality assurance cycle. In the model quality assurance cycle, 'impact assessment' refers to the processes involved in considering the clinical significance of model updates on future and historic decision making.

setting there are several commonly adopted clinical models, such as Q-RISK, Framingham Risk Score, EuroSCORE and even an artificial intelligence chatbot for patient triage [7–10]. The Predict tools for early breast cancer and, more recently, non-metastatic prostate cancer, are rare examples of commonly used prognostic models that provide personalised estimates of survival for various treatment strategies [11,12]. These models are available through an open-access online tool and are intended to be used as a consultation aid to facilitate shared decision making. Table 1 details some of the most commonly used prediction models.

Moving beyond simple predictive models, decision support systems integrate multiple sources of information to provide direct recommendations of particular therapeutic options. The IBM Watson for Oncology platform is perhaps the most complex decision support system yet developed and is currently in use in more than 100 hospitals across China, India, Thailand and South Korea [15–18]. This system utilises natural language processing and machine learning techniques to process structured and unstructured oncological data (structured referring to data stored in predefined fields, as opposed to unstructured data, which has no specific format) from literature, medical records, laboratory reports and treatment guidelines to provide treatment option recommendations tailored to local availability and supported by relevant literature. Reports of the system's performance are highly variable and it has received criticism for both a lack of transparency regarding the mechanism of evidence selection and a bias to adhere to US treatment guidelines [19].

Determining the level of clinical utilisation of the systems described in Table 1 is problematic as, with the possible exception of Watson for Oncology, there is a paucity of literature describing the clinical experience of centres that have adopted them. Nevertheless, there is clearly a gulf between the large number of published prediction models and the few actually in clinical use. This may be largely attributed to the convention of evaluating models purely on the basis of predictive performance, rather than establishing their clinical utility [20,21]. As long as clinical usefulness remains a secondary consideration in the model development process, it is unlikely that the translation of published models into clinical practice will improve.

Clinical and Societal Acceptance

The views of both clinicians and patients on the use and acceptability of CPMs are mixed [22]. Reported advantages from the clinician perspective include supplementing their existing clinical knowledge, with the more accurate prediction they provide enhancing decision making confidence and potentially improving patient outcomes [22–25]. CPMs may also facilitate greater patient engagement in decision making by providing additional information to support the discussion of treatment options [22,23]. However, this must be balanced by the risk posed by over-reliance on CPMs. If clinicians feel they are bound to making decisions concordant with model predictions, regardless of patient preference, the patient's role in shared decision making is nullified. Total dependence on models would therefore represent a retrograde step from patient-centred practice to a paternalistic 'model knows best' approach to medicine [26]. A further challenge in using CPMs to support shared decision making is the complexity of communicating the risk of competing outcomes, in terms of both severity and likelihood, in addition to the uncertainty in model predictions.

Despite the potential advantages of CPMs, multiple barriers to their routine use and acceptance by clinicians and patients have been identified [22–24,27]. Clinicians and patients have reported concerns regarding CPM use threatening autonomy, in addition to the risk of becoming dependent on and overconfident in a model [22–24,27]. Many clinicians do not feel they can trust the CPM's output or that their own clinical judgement is superior, despite multiple studies demonstrating the superior predictive performance of CPMs when compared with clinicians' predictions [23–25,27,28]. The potential for them to

Table 1
Details of three widely used prediction models and decision support systems in the field of oncology

Model	Description	Input	Output	Training cohort	Validation cohort	Clinical usage
Predict Breast (V2.2) [6,11,13]	Cox proportional hazards model of breast cancer-specific mortality following surgery. Mortality from other causes modelled separately.	Age ER status HER2 status Ki-67 status Tumour size Tumour grade Detection mode Number of positive modes	Personalised estimates of survival for different treatment options up to 15 years post-surgery, contextualised with survival rates excluding deaths from breast cancer.	(For V1.0) 5694 patients from Eastern Cancer Registration & Information Centre, 1999−2003. Median follow-up: 5.6 years	(For V1.0) 5000 patients from WM Cancer Intelligence Unit, 1999−2003. Median follow-up: 4.8 years (V2.0) 45 789 patients from Scottish Cancer registry, 2001−2015	About 30,000 website visits per month
Predict Prostate (V1.1) [12,14]	Cox proportional hazards models for prostate cancer-specific and non-prostate cancer mortality. Overall survival estimates based on both models under a competing risks framework.	Age Prostate-specific antigen T-stage Hospital admission in previous 2 years BRCA Histological grade group/Gleason score Biopsy data	Personalised estimates of survival for radical and conservative treatment options, contextualised with survival rates excluding deaths from breast cancer. Non-personalised information on side-effect incidence derived from clinical trials.	(For V1.0) 7062 patients treated in East Anglia, 2000−2010. Median follow-up: 9.8 years	3000 patients from the training cohort 2546 patients treated in Singapore, 1990−2015. Median follow-up: 5.1 years	About 500 website visits per month
Watson for Oncology [15,16]	A complex multi-component system that integrates natural language processing and machine learning techniques.	Literature, medical records, imaging, laboratory and pathology reports, treatment guidelines	Recommendations of specific treatment options tailored to local availability and supported by relevant evidence from the literature.	Initial system training performed by Memorial Sloan Kettering Oncologists	Decision concordance studies performed for multiple sites in Denmark, China, India, Thailand, South Korea	>100 hospitals across China, India, Thailand, South Korea

misapply or misinterpret the CPM has also been described [22,27]. Clinicians also report concerns regarding a lack of time to utilise CPMs and a lack of organisational or peer support [22,24,27]. Strategies to overcome some of these barriers have been proposed but require further study [27].

Life Cycle of a Clinical Prediction Model

In response to perceived shortcomings in current approaches to clinical modelling, we propose a high-level process diagram for the life cycle of a CPM (Figure 2), encompassing both model commissioning (phases 0–4) and ongoing quality assurance processes (phase 5). This is intended to support clinical teams in implementing CPMs in routine practice.

Phase 0: Identify Unmet Clinical Need

A critical preliminary step in the implementation of a CPM is to explicitly consider the clinical decisions that the model is intended to support and the information required to improve on current decision making. The failure to do this is unfortunately common and results in the development of models that offer no clinical utility as they predict events of no clinical relevance, or fail to provide predictions at the time they are needed to inform decision making [29]. To avoid this, it is essential that clinicians lead this phase of model commissioning.

The process of identifying unmet clinical need involves firstly identifying shortcomings in current practice for a particular disease site or subgroup of patients due to, for example, a paucity of evidence to guide management, poor outcomes and/or marked heterogeneity in current decision making. Second, the identified clinical need must be translated into the definition of key measures by which the model's clinical impact may be judged in phase 3.

The final step of phase 0 is to draw on clinical expertise to identify the relevant predictor variables for the outcome of interest and to then determine the availability and robustness of these predictors in pre-existing datasets. This provides an early feasibility check of whether the current routinely captured patient data are sufficient to support the application of a CPM in the intended clinical context.

Phase 1: Model Selection

In selecting a model for clinical evaluation, two of the most important characteristics to consider from a pragmatic standpoint are the model's input variables and end points. The former need to be routinely recorded in a robust and consistent manner and be available at the time the model's prediction is required. Any novel variables that are not currently routinely recorded require critical evaluation in terms of their added benefit, ease of acquisition and robustness. The model's end points must be appropriate to support the intended clinical usage scenario; the information provided must be both pertinent to the clinical decision and build upon current expert knowledge.

A more subjective but no less important consideration in model selection is that of face validity. A model with face validity or sensibility may be defined as one that has an underpinning logic that is consistent with current clinical knowledge and a range of input variables that may be considered comprehensive, while excluding any obviously spurious factors [27]. Consideration of the importance of face validity must be contextualised with the potential offered by highly accurate, but uninterpretable 'black box' artificial intelligence-based systems. A recent study has reported that the public value accuracy more than explainability in healthcare applications [30]. However, clinicians may place a higher priority on face validity and consequently achieving clinical acceptance of 'black box'-type systems may prove challenging [31].

Following this, the next factors to consider are the classical measures of model accuracy – calibration and discrimination [32]. Calibration describes how closely the predicted probability matches the actual probability, whereas discrimination assesses the ability of the model to assign a higher risk to the patients who experience the outcome. The relative importance of these metrics depends on the intended application; patients are concerned by their individual risk, not their relative risk compared with another patient and so calibration is more important if the model is used for patient counselling [33]. Critically, model performance must be considered in tandem with the generalisability of the model. Models with broad external validation should offer more consistent performance when applied to external cohorts (transportability) [33], whereas narrower validation in a specific target group may offer insights that are lost in a more generalised model [34].

A far more subtle, but equally important, factor to consider in model selection is bias [35]. All clinical models are developed and evaluated on existing datasets and this presents a risk of well-trained models inadvertently reflecting, and potentially exacerbating, current biases in medical decision making and inequalities in healthcare access [36]. Models developed on national datasets may perform poorly in minority groups and/or those with poor healthcare access due to data poverty; for example, applying the Framingham risk score to non-white populations has been found to result in poor performance [37]. Bias may also be inadvertently introduced by design; the Watson for Oncology system was trained by experts at the Memorial Sloan Kettering Cancer Center and as a result has been shown to preferentially reflect their own treatment guidelines over and above those determined by expert panels in other countries [19]. Sources of bias are rarely explicitly considered in published literature and although all forms of bias may not be easily identifiable, routine consideration of common sources of bias is warranted and can be carried out using frameworks such as PROBAST, which is a tool to assess the risk of bias and applicability of prediction models [38].

In the unlikely event that no suitable model exists or can be modified for the intended application, developing a new model from scratch can be considered. However, this should be a last resort to avoid exacerbating the proliferation of

unused models. Guidance on CPM development can be found in [21,39,40].

Phase 2: Local Validation/Updating of the Clinical Prediction Model

Local validation describes the process of assessing model performance in a sample of the target population it is to be applied to, as opposed to internal validation, which uses a sample of the training dataset, or external validation, which assesses general transportability to other populations. Local validation is an essential process, as predictive accuracy may be degraded in external cohorts for several reasons, as summarised in Table 2. The local population may be poorly represented by the training dataset due to differences in demographics, case-mix or environment and the effect of predictor variables on outcomes may vary between groups [41]. Changes over time in diagnosis and treatment, as well as the methods used to record and report the model's input variables may also affect model performance, even in the absence of any actual change in the population [42]. Model over-fitting, in which the model is inadvertently trained to reflect random peculiarities of the training dataset, can also lead to poor generalisability [43].

Where local validation indicates poor performance of the proposed model in the target population, model updating strategies should be applied rather than abandoning the model, along with any useful insights it may provide, and starting again. Model updating can range from a simple recalibration-in-the-large (updating the model intercept) through to re-estimating the entire model; guidance is available regarding the extent of updating required in different situations [44]. Model updating is discussed further in phase 5.

With many prediction models freely available online, there is a clear risk of clinical usage without any form of local validation. The degree of risk posed by this is difficult to assess and dependent upon the intended usage scenario, but where multiple validation studies on a diverse range of cohorts have already been carried out, it could reasonably be argued that the model's generalisability is well proven and therefore the risk of poor performance is small. The PREDICT tools are arguably an example of this, having been independently and successfully validated on multiple large multinational datasets [45,46]. However, studies have also identified that the PREDICT breast tool performs poorly in the subgroup of younger women, suggesting that such freely available tools should be used with great caution in the absence of local validation [47].

Table 2
Descriptions of the major factors that can affect model performance in external validation studies

Variation type	Description
Geographical	Changes in patient demographics, disease incidence, healthcare services, etc. by area.
Temporal	Changes over time in population characteristics, incidence rates, diagnosis, treatment, etc.
Methodological	Data collected by different methods to those used in the training dataset, e.g. computed tomography versus positron emission tomography staging
Spectrum	Patients from different care settings, such as within a clinical trial versus primary care

Phase 3: Assessment of Clinical Utility

Although the process of developing clinical models and assessing their predictive performance on retrospective data is well established, methods of determining clinical utility are in their infancy [32]. Model performance, in terms of calibration and discrimination, does not account for the clinical consequences of model predictions and as such cannot be linked to outcomes. A clear demonstration of this is that in some cases models with inferior performance metrics can actually offer superior clinical utility. As described by Cook [48], a model with perfect discrimination that predicts a clinically negligible difference in patients classified as high/low risk will not affect the decision making process, whereas a model with inferior discrimination but a clinically significant difference in predicted risk would be of greater clinical use.

An additional step is needed to determine clinical utility. A common suggestion to address this issue is to perform a randomised trial comparing standard clinical decision making with model informed decision making [20]. This approach has been successfully demonstrated by the STarT Back trial introducing stratified care for back pain driven by a prognostic screening tool [49]. The value offered by such trials lies in establishing the clinical effectiveness, cost-effectiveness and the procedures required for implementation of a CPM. However, randomised trials are not the most practical method for the evaluation of CPMs due to their high cost, complexity and long timespan. Furthermore, regular reassessment of clinical utility following model updates is required after the introduction of CPMs in routine practice. As such, although clinical trials will probably play a critical role in driving the early adoption of prediction models, much as they did with the implementation of intensity-modulated radiotherapy, alternative approaches are required to support the timely adoption of new and updated models [50].

Decision Concordance

A manifestation of the lack of established methods for assessing the clinical utility of CPMs is the inappropriate use of concordance (the percentage of model decisions that agree with current clinical decision making) as a surrogate for overall model accuracy and even utility. As reviewed by Tupasela and Di [19], the use of concordance in this way is both illogical and peculiar; the implication that higher concordance reflects a better system is inherently flawed, as a 100% concordant system would simply mirror current

clinical decision making and therefore would offer no utility beyond automation. Furthermore, percentage concordance is very much of secondary importance to where the conflict with current clinical decision making occurs and whether this may be expected within the context of the clinical need defined in phase 0. For example, discordant model predictions in typical patients, for whom there is a clear choice of optimal treatment well supported by evidence and guidelines, may represent poor model performance, whereas in more complex patients, for whom there is debate over optimal treatment, discordant predictions could actually be a result of the model improving on current decision making. As such, if two systems have similar concordance percentages, this may not translate to them performing similarly in clinical use. In spite of this, concordance percentages have been quoted in advertising material for commercial decision support systems [19].

Decision Curve Analysis

Decision curve analysis (DCA) is a method of evaluating the value of the information provided by a prediction model for informing a clinical decision that does not require any additional data beyond those used for model development [51]. The central concept of DCA is that the threshold probability of disease or event at which a clinician or patient will opt for treatment is informative of how the individual views the relative risk of under-versus over-treatment. Values of this threshold probability will be determined by the likelihood of the event occurring, the harm and benefit associated with the intervention (or lack of intervention) both in the presence and absence of the event, as well as the individual's personal preferences.

As depicted in Figure 3, decision curves plot net benefit as a function of threshold probability. Here, net benefit is defined as the difference between the number of true positives (instances where the model prediction is above the threshold and the patient experiences an event) and false positives, weighted by the associated 'harm' that is derived from the threshold probability. This provides a visual means of identifying the range of threshold probabilities across which the model offers a net benefit when compared with current decision making processes, a treat all/none strategy or alternative models. Although specific values of the highly subjective threshold do not need to be determined, the clinically relevant range of threshold values should be identified to aid meaningful interpretation of the decision curve. For a thorough introduction to the interpretation of decision curves, the reader is referred to Vickers *et al.* [52].

DCA provides a valuable step beyond model assessment and selection based purely on abstract statistical measures unrelated to the clinical consequences of using the model to make a decision. The application of DCA in cancer research is gradually increasing, with published examples spanning the areas of screening, treatment stratification and comparative assessment of the utility of competing models [53–55]. However, it should be noted that the net benefit values derived by DCA are not an invariant property of the model, as they depend on the case distribution and population characteristics of the training dataset [56]. As such, the same questions regarding the generalisability of a model also apply

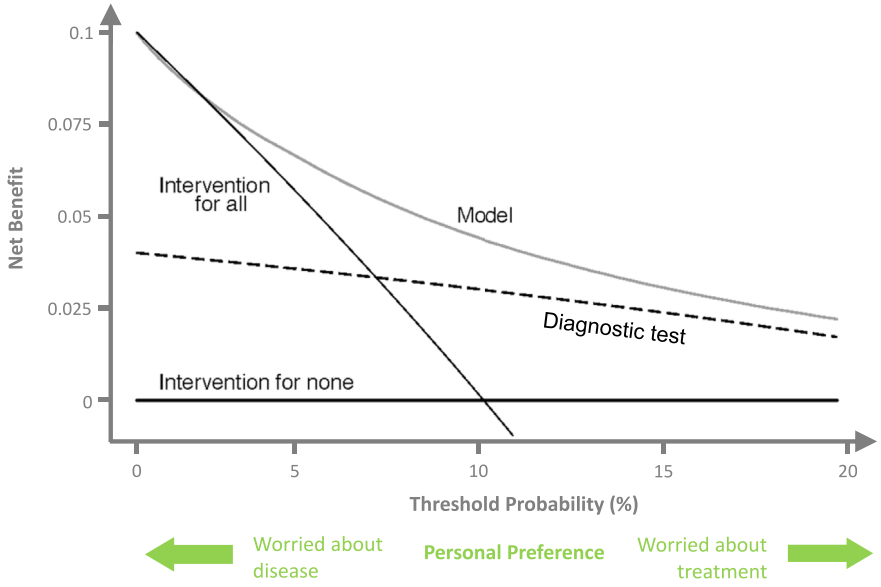

Fig 3. An example decision curve comparing the net benefit of model-based decisions with a diagnostic test and intervention for all/none strategies. The model is shown to offer value (higher net benefit) across threshold probabilities of between about 3% and 20%. The intervention for all strategy marginally outperforms the model at very low threshold probabilities, representing individuals most concerned about the presence of disease. Figure adapted from [52].

to the conclusions of DCA. Additionally, it is not clear how DCA can be applied to continuous outcome measures.

Phase 4: Clinical Implementation

As depicted by the gateway in Figure 2, based on an appropriate assessment of clinical utility a decision must be made on whether to proceed with the clinical implementation of the model. The focus of phase 4 is to integrate the model into routine clinical practice. This includes satisfying any regulatory requirements for clinical use, ensuring it is easily accessible and appropriately presented to the end user, providing training on model use and interpretation, as well as addressing barriers that may hamper its clinical acceptance and routine use.

Regulation

The prediction and prognosis of disease are included within the definition of a medical device and as such CPMs are subject to the Medical Device Regulations in Europe and the Food, Drug and Cosmetic Act in the USA [57,58]. In Europe, CPMs are required to be CE marked before they can be used clinically, with QRISK and the Predict tools being examples of models that have received this certification. Multiple national and international organisations are currently undertaking reviews of how to best approach the regulation of predictive models and mitigate the unique risks and hazards that they pose as non-static medical devices [59–63]. Key themes identified by these reviews include the need for more flexible regulation processes tailored to the risk and potential benefit of a system and the importance of ongoing post-market validation to assess performance throughout the life cycle. Critically, new regulatory frameworks must be supported by detailed guidance and dissemination programmes in order to address the current lack of knowledge, which results in potentially valuable models failing to make the leap from demonstration of utility to clinical implementation.

Presenting Model Output

For the successful integration of predictive models into clinical practice it is essential that the model's predictions are presented in a complete and user-centric manner [64]. To determine the requirements for this, the intended user groups should be surveyed and included in usability testing, following the approach adopted by Farmer *et al.* [65] in the redesign of the PREDICT breast tool interface. A major recommendation from this work was the need for flexibility in both user interaction with the model and the display of results, with multiple presentation formats desirable.

A complete presentation of CPM output should include a minimum of the following:

(i) Definition of inputs and the values used in the calculation
(ii) Model prediction and exemplar interpretation/ context
(iii) Explicit statement of the uncertainty associated with the prediction
(iv) Details of the evidence base for the model
(v) Details of the datasets used for model training and validation
(vi) Dates of model commissioning and version history

Items (i)–(iii) should always be immediately visible to the end user.

The completeness of model output presentation must be balanced with usability; the user should not be overwhelmed by the volume of information provided. As such, items (iv)–(vi), which do not vary with each calculation, could be presented separately while remaining readily accessible, to provide a less cluttered user interface.

Item (iii) refers to the multiple sources of uncertainty that influence an individual's prediction; namely missing information, bias, noise, input variable measurement uncertainty and model fit uncertainties. Ideally, every prediction should be presented with an individualised estimate of the associated uncertainty to allow the end user to consider the weight they should give to the CPM's output. The accuracy of an individual's prediction will exhibit a complex dependence on the distance between their characteristics and those of the population modelled in the training dataset. Quantifying this dependence is challenging and although methods have been proposed, they are infrequently applied [66,67].

As depicted in Figure 4, in contemporary (but incorrect) usage a nomogram is defined as a simplified graphical representation of the mathematical formulae that constitute a model. They are a very commonly used tool for presenting a prediction model in a brief, visual and accessible format that also allows for simple manual calculation of model predictions. Although nomograms offer some utility as a medium for communicating a model, they fail to provide a complete presentation of model output and in the modern era there is limited justification for manual calculation. As such, clinical usage of nomograms is considered undesirable [68].

Phase 5: Ongoing Quality Assurance

Clinical practice, cancer incidence rates, treatment outcomes and patient demographics change over time and, accordingly, the performance of a clinical model should not be considered time invariant and nor should its clinical utility [42]. As such, the ongoing quality assurance of a model requires regular revalidation on more recently collected data (temporal validation) and subsequent updating, if required. There are multiple possible approaches to model updating:

- Regular: Training data replaced with most recent data
- Supplementary: Training data supplemented with most recent data
- Conditional: Training data supplemented subject to model performance criteria [70]

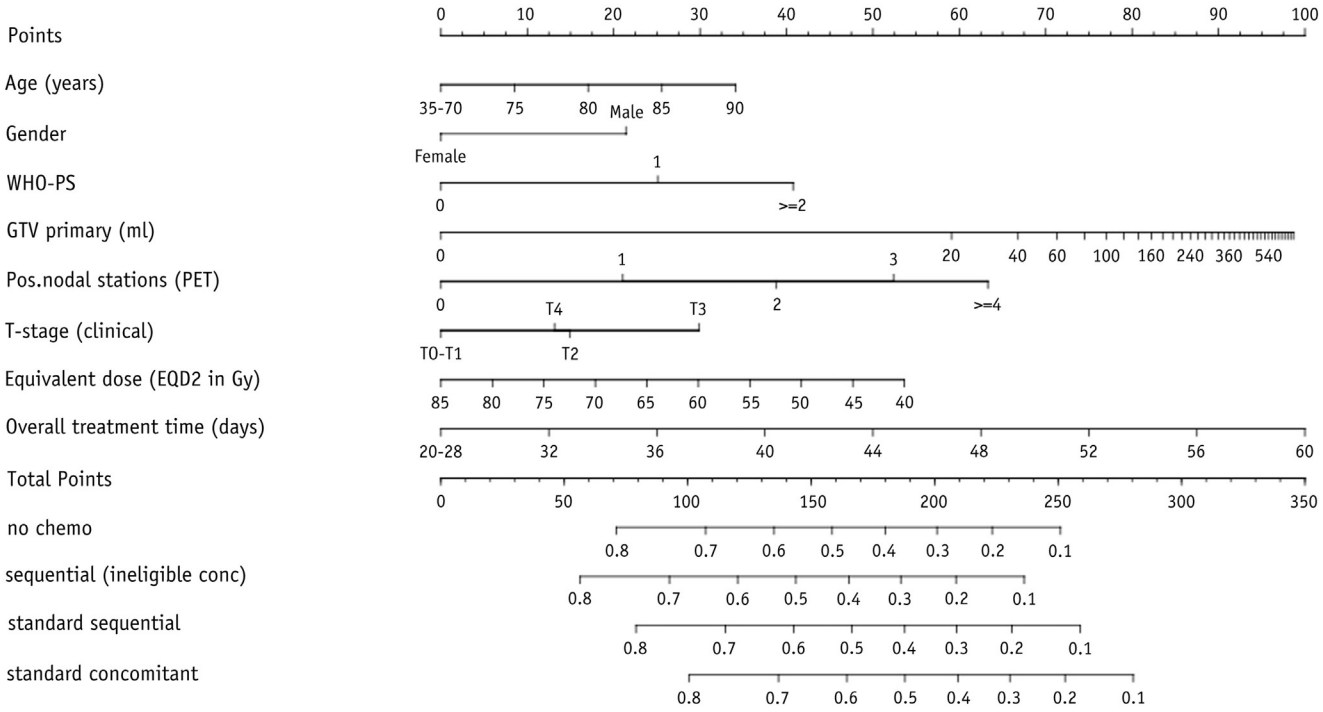

Fig 4. Exemplar nomogram for prediction of 24-month overall survival from stage III non-small cell lung cancer. Figure reproduced from Oberije et al. [69].

- Dynamic: Model continuously updates in real-time, with decreased weighting given to historic data [71]

Which of these approaches is optimal will depend upon the degree and type of change between the original and new datasets, the amount of new data available and the practicalities of applying the method in a busy clinical environment. Methods of determining the most appropriate updating method have been developed [44]. More complex approaches that preferentially weight new data should be applied with caution paid to the risk of over-fitting [72].

Conclusion

Predictive and prognostic models hold great potential to support clinical decision making in oncology and could ultimately facilitate a paradigm shift to a more personalised form of treatment. A major advantage of these models is that they can be developed using real-world data generated by patients treated in routine clinical practice, thereby providing a new form of evidence that is more inclusive of the patient groups commonly under-represented in traditional clinical trials.

Despite the publication of a large number of potentially useful models, examples of clinical implementation in oncology remain scarce. This is partly due to the challenge posed by the complexity of decision making in oncology, but the common failure to consider clinical utility both before and after model development is also a major factor. This in itself is symptomatic of a general under appreciation of the importance of the central involvement of end users, in this case oncologists and patients, throughout the model development and implementation process.

To help bridge the gap between model development and clinical implementation, we have proposed a process for the management of CPMs throughout their life cycle, encompassing model commissioning, clinical implementation and ongoing quality assurance, with the central involvement of end users embedded as a key principle. We hope this will provide a useful framework to support future efforts in translating predictive and prognostic models into clinical use.

Author Contribution

1. Guarantor of integrity of the entire study-Matthew Craddock.
2. Study concepts and design-N/A.
3. Literature research-Matthew Craddock, Cathryn Crockett.
4. Clinical studies-N/A.
5. Experimental studies/data analysis-N/A.
6. Statistical analysis-N/A.
7. Manuscript preparation-Matthew Craddock, Cathryn Crockett.
8. Manuscript editing-Matthew Craddock, Cathryn Crockett, Alan McWilliam, Gareth Price, Matthew Sperrin, Sabine N. van der Veer, Corinne Faivre-Finn.

Funding

M. Craddock, A. McWilliam, G. Price and C. Faivre-Finn are supported by Cancer Research UK via funding to the Cancer Research Manchester Centre (C147/A25254). C. Faivre-Finn was also supported by NIHR Manchester Biomedical Research Centre. G. Price is also supported by Cancer Research UK RadNet Manchester (C1994/A28701).

Conflicts of interest

The authors declare no conflict of interest.

References

[1] Hussain-Gambles M. Ethnic minority under-representation in clinical trials: whose responsibility is it anyway? *J Health Organ Manag* 2003;17:138–143. https://doi.org/10.1108/14777260310476177.

[2] Prendki V, Tau N, Avni T, Falcone M, Huttner A, Kaiser L, *et al*. A systematic review assessing the under-representation of elderly adults in COVID-19 trials. *BMC Geriatr* 2020;20:538. https://doi.org/10.1186/s12877-020-01954-5.

[3] Clark GM, Zborowski DM, Culbertson JL, Whitehead M, Savoie M, Seymour L, *et al*. Clinical utility of epidermal growth factor receptor expression for selecting patients with advanced non-small cell lung cancer for treatment with erlotinib. *J Thorac Oncol* 2006;1:837–846. https://doi.org/10.1016/s1556-0864(15)30414-7.

[4] Lin L, Sperrin M, Jenkins DA, Martin GP, Peek N. A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagn Progn Res* 2021;5:3. https://doi.org/10.1186/s41512-021-00092-9.

[5] Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, *et al*. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell* 2020;2:369–375. https://doi.org/10.1038/s42256-020-0197-y.

[6] Candido dos Reis FJ, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, *et al*. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Res* 2017;19:1–13. https://doi.org/10.1186/s13058-017-0852-3.

[7] QRISK3, https://qrisk.org/. [Accessed 22 April 2021].

[8] Framingham Heart Study, https://framinghamheartstudy.org/. [Accessed 22 April 2021].

[9] Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, *et al*. Euroscore II. *Eur J Cardio-Thoracic Surg* 2012;41:734–745. https://doi.org/10.1093/ejcts/ezs043.

[10] Babylon Health UK - The Online Doctor and… | Babylon Health, https://www.babylonhealth.com/. [Accessed 22 April 2021].

[11] Predict Breast, https://breast.predict.nhs.uk/tool. [Accessed 27 May 2021].

[12] Predict Prostate, https://prostate.predict.nhs.uk/tool. [Accessed 27 May 2021].

[13] Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, *et al*. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res* 2010;12(R1). https://doi.org/10.1186/bcr2464.

[14] Thurtle DR, Greenberg DC, Lee LS, Huang HH, Pharoah PD, Gnanapragasam VJ, *et al*. Individual prognosis at diagnosis in nonmetastatic prostate cancer: development and external validation of the PREDICT Prostate multivariable model. *PLoS Med* 2019;16:1–19. https://doi.org/10.1371/journal.pmed.1002758.

[15] Suwanvecho S, Suwanrusme H, Jirakulaporn T, Issarachai S, Taechakraichana N, Lungchukiet P, *et al*. Comparison of an oncology clinical decision-support system's recommendations with actual treatment decisions. *J Am Med Inform Assoc* 2021;28:832–838. https://doi.org/10.1093/jamia/ocaa334.

[16] Choi YI, Chung JW, Kim KO, Kwon KA, Kim YJ, Park DK, *et al*. Concordance rate between clinicians and Watson for oncology among patients with advanced gastric cancer: early, real-world experience in Korea. *Can J Gastroenterol Hepatol* 2019;2019:8072928. https://doi.org/10.1155/2019/8072928.

[17] Yao S, Wang R, Qian K, Zhang Y. Real world study for the concordance between IBM Watson for oncology and clinical practice in advanced non-small cell lung cancer patients at a lung cancer center in China. *Thorac Cancer* 2020;11:1265–1270. https://doi.org/10.1111/1759-7714.13391.

[18] Somashekhar SPS, Kumar R, Kumar A, Patil P, Rauthan A. Validation study to assess performance of IBM cognitive computing system Watson for oncology with Manipal multidisciplinary tumour board for 1000 consecutive cases: an Indian experience. *Ann Oncol* 2016;27:ix179. https://doi.org/10.1016/S0923-7534(21)00709-2.

[19] Tupasela A, Di E. Concordance as evidence in the Watson for oncology decision – support system. *AI Soc* 2020;35:811–818. https://doi.org/10.1007/s00146-020-00945-9.

[20] Lambin P, Stiphout R, Starmans M, Rios-Velazquez E, Nalbantov G, Aerts HJWL, *et al*. Predicting outcomes in radiation oncology-multifactorial decision support systems. *Nat Rev Clin Oncol* 2012;10:27–40. https://doi.org/10.1038/nrclinonc.2012.196.

[21] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35:1925–1931. https://doi.org/10.1093/eurheartj/ehu207.

[22] Cowley LE, Farewell DM, Maguire S, Kemp AM. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagn Progn Res* 2019;3:1–23. https://doi.org/10.1186/s41512-019-0060-y.

[23] Hallen SAM, Hootsmans NAM, Blaisdell L, Gutheil CM, Han PKJ. Physicians' perceptions of the value of prognostic models: the benefits and risks of prognostic confidence. *Heal Expect* 2015;18:2266–2277. https://doi.org/10.1111/hex.12196.

[24] Saunders B, Bartlam B, Foster NE, Hill JC, Cooper V, Protheroe J. General practitioners' and patients' perceptions towards stratified care: a theory informed investigation. *BMC Fam Pract* 2016;17:1–15. https://doi.org/10.1186/s12875-016-0511-2.

[25] Vogenberg FR. Predictive and prognostic models: implications for healthcare decision-making in a modern recession. *Am Heal Drug Benefits* 2009;2:218–222.

[26] McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. *J Med Ethics* 2019;45:156–160. https://doi.org/10.1136/medethics-2018-105118.

[27] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006;144:201–209.

[28] Oberije C, Nalbantov G, Dekker A, Boersma L, Borger J, Reymen B, *et al*. A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making. *Radiother Oncol* 2014;112:37–43. https://doi.org/10.1016/j.radonc.2014.04.012.

[29] Wyatt JC, Altman DG. Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ* 1995;311: 1539–1591. https://doi.org/10.1136/bmj.311.7019.1539.
[30] Van Der Veer SN, Riste L, Cheraghi-sohi S, Phipps DL, Tully MP, Bozentko K, *et al.* Trading off accuracy and explainability in AI decision making: findings from two citizens juries. *J Am Med Inform Assoc* 2021;28:2128–2138.
[31] London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep* 2019;49:15–21. https://doi.org/10.1002/hast.973.
[32] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2.
[33] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130: 515–524. https://doi.org/10.7326/0003-4819-130-6-199903160-00016.
[34] Wessler BS, Nelson J, Park JG, McGinnes H, Gulati G, Brazil R, *et al.* External validations of cardiovascular clinical prediction models: a large-scale review of the literature. *Circ Cardiovasc Qual Outcome.* 2021;14:e007858.
[35] Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *Npj Digit Med* 2020;3:1–8. https://doi.org/10.1038/s41746-020-0304-9.
[36] Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019;51:584–591. https://doi.org/10.1038/s41588-019-0379-x.
[37] Gijsberts CM, Groenewegen KA, Hoefer IE, Eijkemans MJC, Asselbergs FW, Anderson TJ, *et al.* Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One* 2015;10:1–13. https://doi.org/10.1371/journal.pone.0132321.
[38] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, *et al.* PROBAST: A tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 2019;170:W1–W33. https://doi.org/10.7326/M18-1377.
[39] Steyerberg EW. *Clinical Prediction Models*, 2nd edition. Cham: Springer International Publishing; 2019.
[40] Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, *et al.* Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; 10:e1001381. https://doi.org/10.1371/journal.pmed.1001381.
[41] Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010; 172:971–980. https://doi.org/10.1093/aje/kwq223.
[42] Nakatsugawa M, Cheng Z, Kiess A, Choflet A, Bowers M, Utsunomiya K, *et al.* The needs and benefits of continuous model updates on the accuracy of RT-induced toxicity prediction models within a learning health system. *Int J Radiat Oncol Biol Phys* 2019;103:460–467. https://doi.org/10.1016/j.ijrobp.2018.09.038.
[43] Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004;66:411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9.
[44] Vergouwe Y, Nieboer D, Oostenbrink R, Debray TPA, Murray GD, Kattan MW, *et al.* A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med* 2017;36: 4529–4539. https://doi.org/10.1002/sim.7179.
[45] Thurtle D, Bratt O, Stattin P, Pharoah P, Gnanapragasam V. Comparative performance and external validation of the multivariable PREDICT prostate tool for non-metastatic prostate cancer: a study in 69,206 men from Prostate Cancer data Base Sweden (PCBaSe). *BMC Med* 2020;18:1–8. https://doi.org/10.1186/s12916-020-01606-w.
[46] Wishart GC, Bajdik CD, Azzato EM, Dicks E, Greenberg DC, Racshbass J, *et al.* A population-based validation of the prognostic model PREDICT for early breast cancer. *Eur J Surg Oncol* 2011;37:411–417. https://doi.org/10.1016/j.ejso.2011.02.001.
[47] Maishman T, Copson E, Stanton L, Gerty S, Dicks E, Durcan L, *et al.* An evaluation of the prognostic model PREDICT using the POSH cohort of women aged ≤40 years at breast cancer diagnosis. *Br J Cancer* 2015;112:983–991. https://doi.org/10.1038/bjc.2015.57.
[48] Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115:928–935. https://doi.org/10.1161/CIRCULATIONAHA.106.672402.
[49] Hill JC, Whitehurst DGT, Lewis M, Bryan S, Dunn KM, Foster NE, *et al.* Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet* 2011; 378:1560–1571. https://doi.org/10.1016/S0140-6736(11)60937-9.
[50] Wallace E, Smith SM, Perera-Salazar R, Vaucher P, McCowan C, Collins G, *et al.* Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). *BMC Med Inform Decis Mak* 2011;11:62. https://doi.org/10.1186/1472-6947-11-62.
[51] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak* 2006;26: 565–574. https://doi.org/10.1177/0272989X06295361.
[52] Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;3:1–8. https://doi.org/10.1186/s41512-019-0064-7.
[53] Raji OY, Duffy SW, Agbaje OF, Baker SG, Christiani DC, Cassidy A, *et al.* Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer. *Ann Intern Med* 2012; 157:242. https://doi.org/10.7326/0003-4819-157-4-201208210-00004.
[54] Wynants L, Timmerman D, Verbakel JY, Testa A, Savelli L, Fischerova D, *et al.* Clinical utility of risk models to refer patients with adnexal masses to specialized oncology care: multicenter external validation using decision curve analysis. *Clin Cancer Res* 2017;23:5082–5091. https://doi.org/10.1158/1078-0432.CCR-16-3248.
[55] Hahn C, Eulitz J, Peters N, Wohlfahrt P, Enghardt W, Richter C, *et al.* Impact of range uncertainty on clinical distributions of linear energy transfer and biological effectiveness in proton therapy. *Med Phys* 2020;47:6151–6162. https://doi.org/10.1002/mp.14560.
[56] Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *J Clin Oncol* 2016;34:2534–2540. https://doi.org/10.1200/JCO.2015.65.5654.
[57] Medical Device Regulations. Chapter I article 2 - definitions. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017R0745&from=EN#d1e1258-1-1. [Accessed 15 November 2021].
[58] IMDRF SaMD Working Group. Software as a medical device (SaMD): key definitions 2013. Available at: http://www.imdrf.

org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf. [Accessed 15 November 2021].
[59] Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. *Science* 2019;363:810–812. https://doi.org/10.1126/science.aaw0029.
[60] MHRA - Software and AI as a Medical Device Change Programme - GOV.UK. Available at: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme. [Accessed 15 November 2021].
[61] Baird P, Hoefer E, Lewelling J, Turpin R. Machine learning in medical devices - adapting regulatory frameworks and standards to ensure safety and performance 2020. Available at: https://www.ethos.co.im/wp-content/uploads/2020/11/MACHINE-LEARNING-AI-IN-MEDICAL-DEVICES-ADAPTING-REGULATORY-FRAMEWORKS-AND-STANDARDS-TO-ENSURE-SAFETY-AND-PERFORMANCE-2020-AAMI-and-BSI.pdf. [Accessed 15 November 2021].
[62] ICMRA informal innovation network horizon scanning assessment report - artificial intelligence. Available at: https://doi.org/10.1038/s41573-019-0024-5 2021; 2021.
[63] U.S. Food & Drug Administration, Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-Based software as a medical device (SaMD). Available at: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf, accessed: 15 November 2021.
[64] Norman DA, Draper SW. *User centered system design: new perspectives on human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1986.
[65] Farmer GD, Pearson M, Skylark WJ, Freeman ALJ, Spiegelhalter DJ. Redevelopment of the Predict: breast cancer website and recommendations for developing interfaces to support decision-making. *Cancer Med* 2021;10:5141–5153. https://doi.org/10.1002/cam4.4072.
[66] Ovadia Y., Fertig E., Ren J., Nado Z., Sculley D., Nowozin S., et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. arXiv:1906.02530v2 [stat.ML]; 2019 .
[67] Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *Npj Digit Med* 2021;4:4. https://doi.org/10.1038/s41746-020-00367-3.
[68] Grimes DA. The nomogram epidemic: resurgence of a medical relic. *Ann Intern Med* 2008;149:273–275.
[69] Oberije C, De Ruysscher D, Houben R, van de Heuvel M, Uyterlinde W, Deasy JO, *et al*. A Validated Prediction Model for Overall Survival From Stage III Non-Small Cell Lung Cancer: Toward Survival Prediction for Individual Patients. *Int J Radiat Oncol* 2015;92:935–944, https://doi.org/10.1016/j.ijrobp.2015.02.048.
[70] Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. *J Biomed Inform* 2020;112:103611. https://doi.org/10.1016/j.jbi.2020.103611.
[71] Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TPA, Collins GS, *et al*. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res* 2021;5:1–7. https://doi.org/10.1186/s41512-020-00090-3.
[72] Su TL, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res* 2018;27:185–197. https://doi.org/10.1177/0962280215626466.