Designing deep learning studies in cancer diagnostics

Andreas Kleppe[1,2], Ole-Johan Skrede[1,2], Sepp De Raedt[1,2], Knut Liestøl[1,2], David J. Kerr[3], and Håvard E. Danielsen[1,2,3†]

[1]Institute for Cancer Genetics and Informatics, Oslo University Hospital, Oslo, Norway

[2]Department of Informatics, University of Oslo, Oslo, Norway

[3]Nuffield Division of Clinical Laboratory Sciences, University of Oxford, Oxford, United Kingdom

[†]Corresponding author: E-mail: hdaniels@ifi.uio.no

Abstract

The number of publications on deep learning for cancer diagnostics is rapidly increasing, and systems are frequently claimed to perform comparable to or better than clinicians. However, few systems have yet demonstrated real-world medical utility. In this Perspective, we discuss reasons for the moderate progress, and describe remedies designed to facilitate transition to the clinic. Recent, presumably influential deep learning studies in cancer diagnostics, of which the vast majority used images as input to the system, are reviewed to reveal the status of the field. By manipulating real data, we then exemplify that much and varied training data facilitates the generalisability of neural networks, and thus the ability to use them clinically. To reduce the risk of biased performance estimation of deep learning systems, we advocate evaluation in external cohorts, and strongly advise that the planned analyses, including a predefined primary analysis, are described in a protocol preferentially stored in an online repository. Recommended protocol items should be established for the field, and we present our suggestions.

[H1] Introduction

Deep learning **[G]** facilitates utilisation of large datasets through direct learning of correlations between raw input data and target output, providing systems that may use intricate structures in high-dimensional input data to accurately model the association with the target output[1,2]. A number of studies have reported on the applicability of deep learning in cancer diagnostics, including prediction of diagnosis, prognosis and treatment response[3-5]. While a large number of these tools are claimed to perform comparably or better than clinicians, few have yet demonstrated real-world medical utility[6]. This is partly a natural consequence of the time needed for evaluating and adapting systems affecting patient treatment. However, many studies evaluating apparently well-functioning systems are at high risk of bias[6]. Of particular concern is the frequent lack of stringent evaluation on external data[7,8] and that some systems are developed or evaluated on data that are too narrow or inappropriate for the intended medical setting[9-12]. Thus, the lack of a well-established sequence of evaluation steps for converting promising prototypes into properly evaluated medical systems clearly limits the medical utilisation of deep learning systems **[G]**.

While supervised machine learning **[G]** techniques traditionally utilised carefully selected representations of the input data to predict the target output, modern deep learning techniques use highly flexible artificial neural networks **[G]** to correlate input data directly to the target outputs[1,2,13]. The relations learned by such direct correlation will often be true but may sometimes be spurious phenomena exclusive to the data utilised for learning. In fact, the millions of adjustable parameters make deep neural networks capable of performing perfectly in training **[G]** sets even when the target outputs are randomly generated and therefore utterly meaningless[14]. Thus, the high capacity **[G]** of neural networks induces serious challenges on how to design and develop deep learning systems, and on how to validate that such a system performs adequately in the intended medical setting[15]. Adequate clinical performance will only be possible if the system has good generalisability **[G]** to subjects not included in the training data[16,17].

The design challenge involves issues related to selection of appropriate training data, such as representativeness of the target population (BOX 1), as well as modelling questions such as how the variation of training data may be artificially increased without jeopardising the relationship between input data and target outputs in the

training data[18,19]. The validation challenge includes verifying that the system generalises well, e.g. performs satisfactorily when evaluated on relevant patient populations at new locations and when input data are obtained using differing laboratory procedures or alternative equipment[15,16]. Moreover, deep learning systems are typically developed iteratively, with repeated testing and often including various selection processes that may bias results[20]. Similar selection issues have been recognised as a general concern for the medical literature for many years[21,22]. Thus, when selecting design and validation processes for diagnostic deep learning systems, one will have to focus both on the generalisation challenges and on preventing 'classical' pitfalls in data analysis. We will, however, argue that both sets of challenges may be diminished by adopting certain fairly simple principles partly borrowed from the drug clinical trial field.

In this Perspective, we first describe the validation challenges with focus on the use of external cohorts **[G]**. An evaluation of presumably influential deep learning studies is used to reveal the status of the field particularly with respect to validation procedures. We then consider generalisation issues, especially looking at the importance of both natural and artificially induced variations in training datasets. In the last part, we highlight the importance of evaluating an external cohort according to a predefined primary analysis to reduce selection bias, and we outline a suggested sequence of evaluation steps for deep learning studies in cancer diagnostics, including the use of protocols with predefined analysis plans.

[H1] External cohort evaluation

Rigorous performance evaluation is particularly important due to the inherent high complexity of deep neural networks, as seemingly well-performing deep learning systems might utilise unintentional and possibly false features[10-12] and respond unexpectedly to apparently irrelevant changes of the input data[23]. Failure to properly evaluate systems might have far-reaching consequences, including misdirection of further research, diminished credibility of research findings and, most importantly, being worthless or even harmful to patients if used to influence treatment[24,25].

[H2] The importance of an external cohort evaluation

As an initial evaluation step, the cohort used for development of a deep learning system is often partitioned randomly into three distinct subsets hereunder referred to as 'training', 'tuning' **[G]** and 'test' **[G]**, where the training subset is applied to learn candidate deep learning models **[G]**, the tuning subset to select the deep learning system that appears to perform best, and the test subset to evaluate the performance of the selected system[8]. The evaluation on the test subset may provide unbiased estimation of the performance in the development cohort **[G]**. It may also provide some information on the system's ability to perform well in other populations by considering the extent to which the system performs better on the training subset than on the test subset, as this indicates the level of overfitting **[G]** to the training data. Systems that are highly overfitted to the training data are likely not to perform well on other populations as the noise utilised to improve the performance on the training subset may negatively influence the performance on other populations. However, even a system that performs similarly in training and test subsets might perform far from acceptably on cohorts distinct from the development cohort[26,27]. As discussed below and in BOX 1, this may be caused by the system utilising data features that correlate with the target outcome only in the development cohort, which could be viewed as overfitting to the entire development cohort, or it might also be caused by important predictive features not being adequately represented in the development cohort. Thus, using a random subset of the development cohort for testing does not imply that the results have external validity, i.e. the performance of the system observed in the test subset may not generalise to patients external to the development cohort.

For example, Zech, Badgeley and colleagues[11] investigated a deep learning system for detection of pneumonia in chest X-rays, and found that it was not able to uphold the high discrimination performance achieved in the development cohort when applied to cohorts from different institutions. In this case there was a substantially higher disease prevalence in one of the training cohorts, and it appears that the poor generalisation was in part caused by utilisation of cohort-specific characteristics. In particular, the system utilised metallic tokens that radiology technicians placed on patients to indicate laterality, as these often appeared differently in different cohorts. The authors further point out that the system might not even generalise well to other patients from the same institution as the development cohort, because some correlations between input data and target outcome in the development cohort may not be present in new cohorts from the same institution. Winkler and colleagues[12] found that for their system, visible surgical skin markings present in the image were associated with higher prediction score for melanoma. Similarly, Narla and colleagues[10] reported that the presence of a ruler beside a lesion in an image was associated with a higher malignancy score. Of course, neither skin markings nor rulers

are causing the skin disease, but the apparent correlation present in the development cohort is sufficient for the deep learning system to make use of these associations. It could be argued that a more thorough quality control on the training data could mitigate this, but it is highly unlikely that one is able to detect and control for all potential confounding factors present in the training set.

Thus, unbiased performance estimation in a real-world application of a deep learning system requires external cohorts representative for a target population[22,28-30]. In an external validation **[G]**, no information from the external cohort should have influenced the design of the system or the estimation of any model parameter. Additionally, the external cohorts will implicitly define the patient population for which we have estimated the performance of the system. Thus, to know whether or not the results may be generalised to the entire target population, we need a broad validation where the cohorts may be regarded as representative of this desired target population, e.g. with respect to age, sex, ethnicity, geographical differences and disease prevalence[31,32]. Other types of evaluations may also be warranted prior to introducing the system in medical practice, including so-called domain validation to evaluate whether the system performs consistently across a range of laboratories and technical equipment (BOX 2).

Objective, non-random separation of patients from the same hospital or subjects from the same country, e.g. distinguishing between patients treated before and after a certain date, allows using one cohort for training and tuning and the other for what has been denoted 'narrow validation' (BOX 2)[22]. Such evaluation might provide unbiased performance estimation for a particular hospital. However, the two cohorts should not simply be a non-random separation of an originally larger cohort but instead be processed separately when acquiring data and ascertaining target output[33]. Narrow validation is sometimes considered a limited type of external validation[22].

[H2] Prevalence in recent studies

In order to investigate the prevalence of external cohort evaluation and other characteristics of recent studies on deep learning and cancer diagnostics, we searched PubMed on 21st of April 2020 for original research articles published in 2015 or later (Supplementary Methods). The search provided 3,578 results, and the number of

publications roughly doubled each year since 2016. To explore the use of external cohort evaluation and other characteristics in some of the most prominent and perhaps best studies, we restricted our evaluation to those with at least 20 citations per year or published in a journal with impact factor 10 or larger. Although studies satisfying either of these criteria are presumably quite influential, we acknowledge that some of the other studies might be equally good. In particular, recent studies may not have had time to accrue 20 citations even if they are currently of great interest, and such studies would only be included if published in a journal with impact factor 10 or larger. This will exclude most studies published in new journals that are expected to receive impact factors 10 or larger when this becomes available. However, we consider the selected papers to be sufficient for the purposes of this discussion, as they show that some aspects of study design could be better even in some of the presumably best studies. Only 257 (7%) of the 3,578 search results satisfied at least one of these selection criteria, and another 43 search results were excluded because the document type in Web of Science indicated that these were not original research articles. The remaining 214 studies were manually evaluated (Supplementary Table 1). We further excluded 6 studies that were not original research articles and 102 studies where deep learning was not used to predict or classify features relevant for cancer diagnosis, prognosis or treatment response, or such potential utility of the deep learning system was not evaluated. After also excluding 14 studies without human subjects or only pertaining cell biology, we ended up with 92 eligible studies[34-125], of which 85 (92%) used images as input to the deep learning system[34-57,59-64,66,67,69-93,95-99,101-121,123,125].

Among 516 original research articles on artificial intelligence for diagnostic analysis of medical images published in 2018, Kim, Jang and colleagues[7] found only 31 (6%) studies that evaluated an external cohort. In contrast, 50 (54%) of our 92 eligible studies evaluated the performance of the deep learning system on an external cohort[37,40,48,49,51,53,55,60,62,63,65,70,73-75,78-80,82-87,90,92,93,95,96,98,100-102,104-116,120,121,123,125]. This discrepancy is most likely mainly attributed to our selection of presumably influential studies, and partly attributed to the increasing usage of external cohorts (FIG. 1a); 34 (72%) of the 47 eligible studies published in 2019 and 2020 evaluated an external cohort compared to 9 (39%) of the 23 eligible studies published in 2018 and 7 (32%) of the 22 eligible studies published before 2018.

Among studies satisfying both our selection criteria, 79% (11 of 14) evaluated an external cohort, compared to 68% (25 of 37) for studies that satisfied only the impact factor criterion and 34% (14 of 41) for studies that

satisfied only the citation frequency criterion. It thus appears that journals with high impact factor have a preference for studies evaluating external cohorts. This is consistent with the call by editors of leading scientific journals for rigorous evaluation of artificial intelligence tools[126,127] and explicit prioritisation of biomarker studies that evaluate external cohorts by some journals, e.g. the Journal of Clinical Oncology (https://ascopubs.org/jco/authors/journal-policies).

[H1] Generalisability

While increased use of external cohorts is an important step towards proper validation of deep learning systems, one is still left with the challenge of ensuring that the results obtained on such a population provides a satisfactory measure of the performance within the entire intended target population. This target population may typically be patients who have a specific cancer type, and although often restricted e.g. to certain stages of the disease, the target population is normally broad. Although some studies may use more than one external cohort and some use trials with many centres distributed over several countries, it is difficult to obtain external cohorts that entirely cover the target population. Thus, successful application of a deep learning system will depend on good generalisation properties, so that good performance on one population also indicate satisfactory performance on populations differing with respect to some properties. Fortunately, exploring generalisation in deep learning is an active research area[128], and by utilising certain design principles, deep learning systems have shown remarkably good generalisation performance on a number of tasks[2-5].

One way of increasing generalisation is to control the neural network's capacity to express complex mappings, e.g. by limiting the number of adjustable parameters in the network, imposing various constraints on the network or regularising the optimisation[129,130]. Transfer learning could also increase generalisation, particularly when training data for the task at hand is scarce[131,132]. In transfer learning, the network is initialised with parameters optimised using data for a different task, typically using large datasets such as ImageNet[133,134], which may mitigate overfitting at the possible cost of introducing biases[135-137]. Making the training dataset more diverse and more representative of the target population is another way of increasing generalisation[138]. Of particular importance is to ensure adequate and unbiased representation across demographic characteristics such as sex, race and ethnicity (BOX 1). In addition to expanding the natural training dataset, i.e. the set of training data

acquired from a range of patient samples with associated target outcome, one may artificially augment the training dataset by applying smaller transformations on the inputs while maintaining their relationship to the target output[18,139]. This can reduce the network's ability to memorise details of the training data and thereby increase generalisation, especially in situations where the availability of training data is limited. The transforms can randomly change, often called 'distort', the input data by e.g. adding noise, erasing parts, shifting and scaling colours or altering the image geometry[19]. Artificially diversifying the training data may increase generalisation by enabling the resulting system to ignore vagaries of the measurement process and even become applicable to multiple data acquisition procedures, e.g. different acquisition equipment[140,141]. Other augmentation techniques include those that generate artificial input data, e.g. by mixing multiple data inputs[19]. The value of augmentation techniques has been observed in various application domains[19], including the use on images obtained in radiology[38,142-144] and histopathology[141,145].

To illustrate the importance of the amount and variation in training data, and more specifically show how data distortion may work to improve deep learning systems in cancer diagnostics, we show this type of analysis here using data from a previously published study[113]. This previous study applied deep learning to predict colorectal cancer-specific survival directly from conventional haematoxylin and eosin stained sections, with training and tuning data derived from 2,473 patients from four cohorts. The performance was evaluated on an external cohort consisting of 1,122 patients from a randomised controlled trial on a drug that was observed to not affect survival[146]. We applied the convolutional neural network called Inception-v3[147], which is a commonly used network in medical image diagnostics[8], in both the previously published analyses and the new analyses presented here.

Initially, we applied the same distortion process as in our published analyses[113]. This process artificially increased the variation of the training images by randomly distorting their colours, which is an augmentation technique that appears crucial when training deep learning systems in histopathology[145]. Initially, the maximum amount of distortion we allowed was quite modest (FIG. 2a). To illustrate the effect of reducing the number of patients while keeping the patient heterogeneity implied by having data from four cohorts, we randomly sampled 979 patients in such a manner that the data had the same number of training and tuning patients with and without cancer-specific death as in the cohort from the Gloucester Colorectal Cancer Study, UK (the largest of the four

training and tuning cohorts). The decreased performance of the resulting deep learning system when evaluated on the external cohort (FIG. 2b) exemplifies the importance of a large natural training dataset and its intrinsic variation[138]. Further reduction of the number of patients decreased the performance further; training and tuning on a quarter of the 979 patients or less (that is, less than 250 patients) provided systems that did not perform substantially better than random guessing (FIG. 2b).

We then showed that modifying the distortion process may mitigate for the performance loss observed when reducing the number of patients in training and tuning. Compared to using all 2,473 patients for training and tuning, using 979 randomly selected patients and four times the original amount of colour distortion provided similar performance on the external cohort (FIG. 2c). For this modified distortion process we allowed quite substantial colour distortions (FIG. 2d), and the results showed that artificial augmentation may in some cases compensate for limited natural training and tuning data. However, increasing the amount of colour distortion further provided worse performance (FIG. 2c), illustrating the trade-off between preventing overfitting through random distortions and occluding relevant information for the prediction task.

Randomly sampling 979 patients from all four cohorts maintained much of the variation in the natural training and tuning data. If we instead used only the Gloucester cohort, which contained the same number of training and tuning patients with and without cancer-specific death as in the random sample, we obtained worse performance on the external cohort, most clearly when including more colour distortion in training (FIG. 2e). This underlines the importance of designing studies such that the natural training data is diverse, and FIG. 2e additionally illustrates that natural and artificial variation works well together to increase generalisability.

In general, the most suitable distortion process will depend on the particular medical prediction task because the involved data will tolerate different amounts of the various types of distortions before true correlations between input and target output are occluded. For instance, deep learning systems that classify based on images of skin lesions or tumour sections are likely to benefit from being invariant to rotations, while systems aimed at supporting radiology might rely on the orientation in images of larger organ structures and thereby perform worse if forced to be rotation invariant. Thus, the distortion process needs to be fine-tuned to the particular

application, as findings about which distortion process appears most beneficial in one scenario, e.g. findings from the example presented in FIG. 2, are not necessarily directly applicable to other scenarios. However, the general principle is that including much and varied training data is important. As the importance of artificial augmentation decreases with the amount and diversity in the natural training data, prediction tasks where the true correlations between input data and target output are easily obscured by distortion warrants a more comprehensive natural training dataset.

[H1] Predefined primary analysis

In the development of a deep learning system, researchers will often evaluate different systems sequentially, each time having the possibility to learn from interpreting the previous evaluations and adapt the system to the specific data used for evaluation. Such repeated evaluations will bias the estimates, and their dependence on previous evaluations makes established statistical approaches for adjusting for multiple comparisons not applicable[148,149]. Similar re-analysis issues may arise if the initial analysis of a specific deep learning system reveals issues that are then corrected and the performance is re-evaluated. Such problems of repeated or multiple evaluations are well-known from examinations of the data analysis in various types of published medical studies, and have been identified as important contributors to biased inference and irreproducible results[20,150].

As discussed above, evaluation on an external cohort is required for unbiased performance estimation in a real-world application of the deep learning system, but it is only a prerequisite as multiple or repeated evaluations may cause bias even if evaluating an external cohort. Great caution would therefore be needed when interpreting studies that report multiple analyses without specifying which was initially planned to be the primary analysis, if any.

[H2] Prevalence of predefined primary analysis

In our evaluation of recent, presumably influential deep learning studies in cancer diagnostics, all studies performed multiple analyses of the external cohort in the form of either evaluating multiple systems, analysing multiple subpopulations or using various analysis methods. Only 3 (6%) of the 50 eligible studies that evaluated

an external cohort used one of the well-established methods for adjustment for multiple comparisons[51,62,114], e.g. Bonferroni correction. This implies that most studies should have specified which analysis was considered the primary analysis prior to evaluation of the external cohort, if such a decision was made, in order to inform the reader which analysis was not affected by selection bias and to help distinguish studies with a predefined primary analysis from those that repeatedly evaluated the external cohort and might have ended up reporting severely biased performance estimates. Although the principle of using an external dataset only once to evaluate the final hypothesis should be well-known in the machine learning community[151,152], it seems currently that there is no tradition for specifying the predefined primary analysis in deep learning publications other than those reporting on clinical trials. In our evaluation, 20 (40%) of the 50 studies evaluating an external cohort specified one or more primary performance metrics (FIG. 1b)[55,60,73,82,83,85,86,93,98,102,105,108-110,113,115,116,120,121,125], but only 8 (16%) of the 50 studies specified a predefined primary analysis (FIG. 1c)[73,83,102,105,109,113,120,121].

Prespecification of the primary analysis has previously been advocated in diagnostic and prognostic research[153,154], but this is unfortunately still not common practise despite being the only direct protection against selection bias[20]. To ensure unbiased estimation, the primary analysis should be unequivocally specified prior to all investigations that could reveal correlations between input data and target output in the external cohort. This would require the researchers to define all relevant aspects of the validation prior to analysing the cohort, including the deep learning system, target output, and patient and input data in the external cohort. Predefining the primary analysis will entail a commitment to the main analysis, which implies that the analysis should be carefully planned in advance and that researchers will be discouraged from performing creative data dredging[155].

[H2] Choosing the primary metric

Many medical questions are categorical in nature, e.g. whether tumour or not, whether mutated or not, and whether to offer treatment or not. However, deep learning models often output continuous values reflecting the predicted probability of each possible outcome. In such cases, the predefined primary analysis should preferably evaluate a categorisation of the model output aimed at answering the medical question. The primary analysis will then be comparing predicted and target outcome in the external cohort, e.g. by measuring the so-called balanced accuracy **[G]**[156]. Measuring the performance using categorical outputs often provides more conservative

estimates[157] and avoids issues with metrics frequently applied to measure the performance using continuous outputs. For instance, the area under the receiver operating characteristic curve **[G]** (AUC)[158] and concordance index **[G]** (c-index)[159] are only affected by the ranking of the continuous outputs, not the prediction scores themselves[160]. Thus, such metrics may indicate that a deep learning system performs well even if it predicts markedly too high probabilities for all patients in a specific cohort, provided that the continuous outputs of the system rank the patients in a fairly correct order. In another cohort, the same system may similarly appear to perform well even if it predicts markedly too low probabilities for all those patients. The generalisability of such a system is poor, yet this would not be evident from the AUC and c-index of the continuous outputs, but it would be evident from the AUC and c-index of a categorisation defined irrespective of the external cohorts. The categorisation may be defined by e.g. determining suitable thresholds during tuning or selecting the outcome with highest prediction score as the predicted outcome. Defining the categorisation using the external cohort, even at predefined levels of e.g. sensitivity, adapts the categorical marker to the specific external cohort and may occlude shifts in the prediction scores as with the AUC and c-index of the continuous outputs.

In our evaluation of recent, presumably influential deep learning studies in cancer diagnostics, we found that 34 (68%) of the 50 studies evaluating an external cohort reported the estimated performance of a categorical marker on the external cohort, with a categorisation defined irrespective of the external cohort[48,49,53,55,60,62,63,65,73,75,78-80,82,85,87,90,98,100,102,104-106,108-111,113-116,120,121,125]. The proportion was lower for studies reporting on deep learning systems that used histopathology section images as input, with only 6 (40%) of 15 studies evaluating a fixed categorical marker on the external cohort[48,55,82,111,113,114], which is surprising since most histopathological evaluations provide categorical values.

For certain deep learning systems, the intended medical application directly utilises the system's continuous output, e.g. to triage patients for further examinations, and in such cases the continuous output should be evaluated in the primary analysis. This may warrant additional analyses to reveal generalisation issues that might be occluded by the selected performance metric, e.g. to consider a calibration plot in addition to the c-index when evaluating a clinical decision support system for predicting patient outcome[22,26].

[H1] From conception to application

All research with the potential to influence patient treatment should undergo careful evaluation sequences and be driven by protocols with a predefined statistical analysis plan[153]. FIG. 3 illustrates what we consider as natural and important steps in the development and evaluation of deep learning systems for medical applications.

The initial exploratory studies aim to answer whether deep learning appears suitable for the task at hand or whether further investigations based on deep learning are not warranted at this time, usually because the hypothesis seems ill-founded or the available data is not expected to provide a system with adequate performance. The performance estimates obtained in such pilot studies are frequently inflated by the use of a limited development cohort, but promising findings may motivate further investigations. After a series of explorations and possibly expansions of the development cohort, the development should conclude by deciding which system appears to perform best on the intended medical task, considering also the sensitivity to vagaries of the measurement process. Of particular importance to prevent selection of a system that performs much worse on patients outside the development cohort, the study could include sufficient amount and variation in the natural training dataset and use techniques like data distortion to increase the variation artificially.

There is a growing interest in explainable deep learning systems[161-163], including the creation of inherently more explainable systems and post-hoc explanations of existing systems[164]. For image classification tasks in particular, so-called saliency maps visualise the contribution of each pixel to the final prediction score and can be created using a number of different techniques[165-167]. By increasing the transparency, the more explained systems might have more predictable generalising abilities. This may be used to identify target populations within which the system is expected to generalise well or settings where the system is prone to fail. For example, Winkler and colleagues[12] used such a technique to support their finding that surgical skin markings unduly increased the system's prediction score for melanoma. While current explainability techniques might suggest generalisability and thereby suggest suitable target populations or influence the selection of which system to evaluate further, they will only provide indications and thus not reduce the need for proper validation.

While efficacy studies of pharmaceutical products are usually preceded by prospective trials to estimate basic features such as safety and dosing[168], deep learning systems for diagnostic purposes can to a larger extent utilise retrospective cohorts, e.g. from earlier clinical trials or medical practice. Given the risks, timeframe and costs of interventional research[168-170], we recommend rigorous, retrospective analyses to evaluate the medical validity of a deep learning system by conducting an external validation according to a predefined primary analysis. The results of such studies provide valuable information to direct further research, thus warranting publication regardless of the significance of the findings, which would also mitigate publication bias.

Rigorous, retrospective analyses of a deep learning system might warrant conducting a prospective, randomised phase III clinical trial where the system directly intervenes with the current standard of care in order to evaluate the system's medical utility in a specific real-world application, considering both benefits and harms for patients in the target population[30,171]. Systems demonstrated to have medical utility and approved by necessary governmental agencies can be applied in medical practice while monitoring the long-term benefits, harms and costs for each specific real-world medical application in phase IV clinical trials. Such surveillances might eventually indicate that the system needs to be updated because of changes in medical practice or data acquisition[172].

The levels of deep learning studies depicted in FIG. 3 and the phases of clinical trials were used to categorise recent, presumably influential deep learning studies in cancer diagnostics in relation to the reliability of the performance estimation approach and the demonstrated applicability of the system in medical practice. Although some group sizes are very small, there appears to be notable differences between research fields defined by the input to the deep learning system (FIG. 4). The proportion of studies evaluating an external cohort was lowest for the 7 studies with only non-image inputs such as omics data (29%; 2 of 7 studies), while highest for 22 studies with images other than histopathology section and radiology images as input, e.g. from gastrointestinal endoscopic examinations or dermoscopic images (64%; 14 of 22 studies). Five (23%) of the 22 studies with other images as input even had a predefined primary analysis of the external cohort[73,102,105,109,121], which included the 3 studies reporting on a randomised clinical trial, all of which evaluated a deep learning system to aid gastrointestinal examinations[102,105,121].

[H2] Recommended protocol items

When planning to evaluate the medical validity of a deep learning system through rigorous, retrospective analyses, we recommend the unequivocal specification of the predefined primary analysis to be documented in a study protocol. Relevant items in such protocols would differ from clinical trial protocols, which are the target of guidelines such as SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials)[173] and its extension to artificial intelligence[174]. Protocols should be developed before conducting the validation, and relevant items would therefore also differ from those in original research articles, which are the target of many reporting guidelines such as CONSORT (Consolidated Standards of Reporting Trials)[175] and TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis)[22] as well as their extension or anticipated adaption to machine learning[176,177]. It is therefore a need to establish guidelines dedicated to study protocols describing validations of deep learning systems. We propose a non-exhaustive list of items that we consider essential in such protocols, termed Protocol Items for External Cohort Evaluation of a deep learning System (PIECES) in cancer diagnostics.

In order to be sufficiently concrete about the predefined primary analysis, the protocol needs to describe the deep learning system and how it will be assayed, define the external cohort, including its origin, what it represents in terms of medical setting and target population, input data and target output, and clearly specify the performance evaluation. These three parts of the protocol form the basis of our PIECES recommendations together with a declaration of status (BOX 3). The status declaration should scrupulously elucidate any investigations performed before finalising the protocol that could reveal correlations between input data and target output in the external cohort, or state that no such investigations were performed.

The PIECES recommendations are designed to facilitate identification of ambiguities and disagreements between the researchers planning to conduct an external validation as well as to provide a clear description of the predefined primary analysis as reference for all readers, which may aid medical professionals in identifying well-designed studies and their applicability to their own clinical practice. The thought and work that should go into making such a protocol could also allow the researchers to make appropriate changes prior to performing the

external validation. For instance, considering what the external cohort is intended to represent and how the deep learning system is envisioned to be applied in practice, could affect the inclusion and exclusion criteria for patients and samples as well as the metric or statistical test applied in the primary analysis.

Researchers conducting an external validation would often like to perform multiple, related analyses to elucidate the performance of the deep learning system. To separate preplanned analyses from exploratory, post hoc analyses, the PIECES recommendation encourages specification of predefined secondary analyses that the researchers would like to commit themselves to report on publication of their findings. Such secondary analyses would be affected by the multiple comparisons problem but predefining and reporting all secondary analyses would provide a transparency that would substantially increase the credibility of the results. Importantly, the specification of predefined secondary analyses does not diminish the validity of the predefined primary analysis. Any analyses the researchers consider reporting, but do not wish to commit themselves to report, should not be specified as secondary analyses in the protocol and therefore should be reported as exploratory analyses, even though they might be thought of prior to analysing the external cohort.

[H2] Study registration

We recommend registration of the study protocol in an online repository before analysing the external cohort. Most major trial registries, e.g. ClinicalTrials.gov (https://www.clinicaltrials.gov) and the International Standard Randomised Controlled Trial Number (ISRCTN) registry (https://www.isrctn.com), accept registration of diagnostic accuracy studies[154]. These registries can be used to record external validation studies in deep learning, but some items will not be relevant, while some important items such as defining the deep learning system will not be encouraged. A dedicated repository to register the study protocol describing the external validation of a deep learning system is therefore warranted. We recognise that it may be undesirable to publish a detailed study protocol in an online repository prior to conclusion of the study as it would reveal novel work prior to publication of the results and perhaps in some rare cases jeopardise publication. In a dedicated repository, a submission could be partially or completely invisible to the public and the protocol encrypted until the authors choose to reveal the submission and provide the required decryption key, thus facilitating preregistration of study protocols without requiring authors to reveal novel ideas prematurely.

Registration of observational studies has been advocated by editors of major clinical journals[178,179], many editorial board members[180] and researchers[181,182], and the criticism it has received from epidemiologists in relation to the exploratory nature of epidemiology[183-185] does not apply to external validation studies. For diagnostic and prognostic biomarker studies in particular, the registration of a study protocol with a predefined analysis plan has been recommended by several researchers[153,154,186-188], provided that it precedes the onset of the study[189]. This would facilitate a more balanced evaluation of the proposed marker, identification and prevention of selective reporting, increased transparency, reduced proportion of false positive findings, mitigation of publication bias through identification of unpublished studies, and prevention of unnecessary duplication of research while facilitating collaboration between researchers and identification of research gaps. Consequently, widespread preregistration of detailed study protocols for deep learning systems might translate into more rapid identification of promising systems and thereby expedite progression of the research field. It would also communicate a study to peers without disclosing the findings and interpretations prior to editorial and peer review, thus providing some of the benefits of preprint archiving while allowing critical appraisal of the findings and interpretations before publication.

Amendments of clinical trial protocols are common but should be tracked and dated[173]. While clinical trials often take years to conduct due to patient recruitment and follow-up, most external validations of deep learning systems use retrospective data and the analysis part of the validation may be performed in a matter of days. Consequently, it should rarely be necessary to modify the study protocol describing the external validation of a deep learning system after initiating the validation. We therefore generally discourage protocol amendments, but if found necessary for a particular study, we recommend amendments to be included as postscripts to the study protocol, leaving the original protocol unaltered. Both the postscript and disseminations of the validation results should concretely specify what was changed as well as describe the motivation and rationale for the change.

[H1] Conclusions

Including much natural and artificial data variation when training rigorous deep learning systems appears pivotal, as analyses indicate its instrumental role in increasing the performance and generalisability of systems.

Utilising multiple sets of patients, samples and data acquisition procedures will diversify the training data, while augmentation techniques artificially enhance the variation further. The resulting systems may be capable of handling the diversity in routine medical practice and in some cases even generalise to completely new settings.

Going forward, the medical validity of a deep learning system should be evaluated according to a preregistered study protocol specifying the primary analysis and using an external cohort representative of the intended medical setting and target population. This facilitates balanced performance evaluations by reducing selection bias and increasing transparency, and helps medical professionals distinguish rigorous, retrospective validation studies from studies that repeatedly evaluated the external cohort and might end up reporting severely biased performance estimates. It would therefore assist in identifying deep learning systems that warrant prospective evaluations in randomised clinical trials and ultimately drive the development of systems that could transform current medical practice.

References

1 Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **61**, 85-117 (2015).

2 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).

3 Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500-510 (2018).

4 Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463-477 (2019).

5 Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703-715 (2019).

6 Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).

7 Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y. & Park, S. H. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. *Korean J. Radiol.* **20**, 405-410 (2019).

8 Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271-e297 (2019).

9 Ross, C. & Swetlitz, I. *IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show*. STAT. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ (2018).

10 Narla, A., Kuprel, B., Sarin, K., Novoa, R. & Ko, J. Automated Classification of Skin Lesions: From Pixels to Practice. *J. Invest. Dermatol.* **138**, 2108-2110 (2018).

11 Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).

12 Winkler, J. K. *et al.* Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatol.* **155**, 1135-1141 (2019).

13 Rueckert, D. & Schnabel, J. A. Model-Based and Data-Driven Strategies in Medical Image Computing. *Proc. IEEE* **108**, 110-124 (2020).

14 Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *Proc. Int. Conf. Learn. Represent.* (2017).

15 Liu, Y., Chen, P.-H. C., Krause, J. & Peng, L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA* **322**, 1806-1816 (2019).

16 Ransohoff, D. F. Bias as a threat to the validity of cancer molecular-marker research. *Nat. Rev. Cancer* **5**, 142-149 (2005).

17 Moons, K. G. M. *et al.* PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann. Intern. Med.* **170**, W1-W33 (2019).

18 Simard, P., Victorri, B., LeCun, Y. & Denker, J. Tangent Prop - A formalism for specifying selected invariances in an adaptive network. *Adv. Neural Inf. Process. Syst.* **4**, 895-903 (1992).

19 Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **6**, 60 (2019).

20 Ioannidis, J. P. A. What Have We (Not) Learnt from Millions of Scientific Papers with P Values? *Am. Stat.* **73**, 20-25 (2019).

21 Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med.* **2**, e124 (2005).

22 Moons, K. G. M. *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann. Intern. Med.* **162**, W1-W73 (2015).

23 Heaven, D. Why deep-learning AIs are so easy to fool. *Nature* **574**, 163-166 (2019).

24 Ioannidis, J. P. A. Evolution and translation of research findings: from bench to where? *PLoS Clin. Trials* **1**, e36-e36 (2006).

25 Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44-56 (2019).

26 Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the Generalizability of Prognostic Information. *Ann. Intern. Med.* **130**, 515-524 (1999).

27 Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21**, 345-352 (2020).

28 Ioannidis, J. P. A. & Khoury, M. J. Improving Validation Practices in "Omics" Research. *Science* **334**, 1230-1232 (2011).

29 Obermeyer, Z. & Emanuel, E. J. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *N. Engl. J. Med.* **375**, 1216-1219 (2016).

30 Keane, P. A. & Topol, E. J. With an eye to AI and autonomous diagnosis. *npj Digit. Med.* **1**, 40 (2018).

31 Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern. Med.* **178**, 1544-1547 (2018).

32 Noor, P. Can we trust AI not to further embed racial bias and prejudice? *BMJ* **368**, m363 (2020).

33 Luo, W. *et al.* Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J. Med. Internet Res.* **18**, e323 (2016).

34 Hua, K. L., Hsu, C. H., Hidayati, S. C., Cheng, W. H. & Chen, Y. J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *Onco Targets Ther.* **8**, 2015-2022 (2015).

35 Ciompi, F. *et al.* Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Med. Image Anal.* **26**, 195-202 (2015).

36 Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L. & Guevara Lopez, M. A. Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput. Methods Programs Biomed.* **127**, 248-257 (2016).

37 Setio, A. A. A. *et al.* Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Trans. Med. Imaging* **35**, 1160-1169 (2016).

38 Roth, H. R. *et al.* Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation. *IEEE Trans. Med. Imaging* **35**, 1170-1181 (2016).

39 Kallenberg, M. *et al.* Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE Trans. Med. Imaging* **35**, 1322-1331 (2016).

40 Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **6**, 26286 (2016).

41 Huynh, B. Q., Li, H. & Giger, M. L. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J. Med. Imaging* **3**, 034501 (2016).

42 Nie, K. *et al.* Rectal Cancer: Assessment of Neoadjuvant Chemoradiation Outcome based on Radiomics of Multiparametric MRI. *Clin. Cancer Res.* **22**, 5256-5264 (2016).

43 Kooi, T. *et al.* Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303-312 (2017).

44 Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118 (2017).

45 Dhungel, N., Carneiro, G. & Bradley, A. P. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Med. Image Anal.* **37**, 114-128 (2017).

46 Yu, L., Chen, H., Dou, Q., Qin, J. & Heng, P. Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. *IEEE Trans. Med. Imaging* **36**, 994-1004 (2017).

47 Sun, W., Tseng, T. B., Zhang, J. & Qian, W. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Comput. Med. Imaging Graph.* **57**, 4-9 (2017).

48 Cruz-Roa, A. *et al.* Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci. Rep.* **7**, 46450 (2017).

49 Ciompi, F. *et al.* Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci. Rep.* **7**, 46479 (2017).

50 Araújo, T. *et al.* Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS One* **12**, e0177544 (2017).

51 Becker, A. S. *et al.* Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. *Invest. Radiol.* **52**, 434-440 (2017).

52 Dou, Q., Chen, H., Yu, L., Qin, J. & Heng, P. Multilevel Contextual 3-D CNNs for False Positive Reduction in Pulmonary Nodule Detection. *IEEE Trans. Biomed. Eng.* **64**, 1558-1567 (2017).

53 Lao, J. *et al.* A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Sci. Rep.* **7**, 10353 (2017).

54 Setio, A. A. A. *et al.* Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Med. Image Anal.* **42**, 1-13 (2017).

55 Ehteshami Bejnordi, B. *et al.* Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* **318**, 2199-2210 (2017).

56 Mohamed, A. A. *et al.* A deep learning method for classifying mammographic breast density categories. *Med. Phys.* **45**, 314-321 (2018).

57 Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O. & Hajirasouliha, I. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. *EBioMedicine* **27**, 317-328 (2018).

58 Xiao, Y., Wu, J., Lin, Z. & Zhao, X. A deep learning-based multi-model ensemble method for cancer prediction. *Comput. Methods Programs Biomed.* **153**, 1-9 (2018).

59 Marchetti, M. A. *et al.* Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *J. Am. Acad. Dermatol.* **78**, 270-277.e271 (2018).

60 Chen, P.-J. *et al.* Accurate Classification of Diminutive Colorectal Polyps Using Computer-Aided Analysis. *Gastroenterology* **154**, 568-575 (2018).

61 Bychkov, D. *et al.* Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.* **8**, 3395 (2018).

615 62 Yasaka, K., Akai, H., Abe, O. & Kiryu, S. Deep Learning with Convolutional Neural Network for
616 Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study. *Radiology*
617 **286**, 887-896 (2018).
618 63 Chang, K. *et al.* Residual Convolutional Neural Network for the Determination of IDH Status in Low-
619 and High-Grade Gliomas from MR Imaging. *Clin. Cancer Res.* **24**, 1073-1081 (2018).
620 64 Ribli, D., Horváth, A., Unger, Z., Pollner, P. & Csabai, I. Detecting and classifying lesions in
621 mammograms with Deep Learning. *Sci. Rep.* **8**, 4165 (2018).
622 65 Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep Learning–Based Multi-Omics Integration
623 Robustly Predicts Survival in Liver Cancer. *Clin. Cancer Res.* **24**, 1248-1259 (2018).
624 66 Mobadersany, P. *et al.* Predicting cancer outcomes from histology and genomics using convolutional
625 networks. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2970-E2979 (2018).
626 67 Saltz, J. *et al.* Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes
627 Using Deep Learning on Pathology Images. *Cell Rep.* **23**, 181-193.e187 (2018).
628 68 van de Goor, R., van Hooren, M., Dingemans, A.-M., Kremer, B. & Kross, K. Training and Validating
629 a Portable Electronic Nose for Lung Cancer Screening. *J. Thorac. Oncol.* **13**, 676-681 (2018).
630 69 Chang, H., Han, J., Zhong, C., Snijders, A. M. & Mao, J. Unsupervised Transfer Learning via Multi-
631 Scale Convolutional Sparse Coding for Biomedical Applications. *IEEE Trans. Pattern Anal. Mach.*
632 *Intell.* **40**, 1182-1194 (2018).
633 70 Han, S. S. *et al.* Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors
634 Using a Deep Learning Algorithm. *J. Invest. Dermatol.* **138**, 1529-1538 (2018).
635 71 Hirasawa, T. *et al.* Application of artificial intelligence using a convolutional neural network for
636 detecting gastric cancer in endoscopic images. *Gastric Cancer* **21**, 653-660 (2018).
637 72 Chang, P. *et al.* Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations
638 in Gliomas. *Am. J. Neuroradiol.* **39**, 1201-1207 (2018).
639 73 Haenssle, H. A. *et al.* Man against machine: diagnostic performance of a deep learning convolutional
640 neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.*
641 **29**, 1836-1842 (2018).
642 74 Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer
643 histopathology images using deep learning. *Nat. Med.* **24**, 1559-1567 (2018).

75 Wang, P. *et al.* Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat. Biomed. Eng.* **2**, 741-748 (2018).

76 Urban, G. *et al.* Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. *Gastroenterology* **155**, 1069-1078.e1068 (2018).

77 Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686 (2018).

78 Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).

79 Nam, J. G. *et al.* Development and Validation of Deep Learning–based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology* **290**, 218-228 (2019).

80 Byrne, M. F. *et al.* Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut* **68**, 94-100 (2019).

81 Horie, Y. *et al.* Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointest. Endosc.* **89**, 25-32 (2019).

82 Kather, J. N. *et al.* Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.* **16**, e1002730 (2019).

83 Rodríguez-Ruiz, A. *et al.* Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology* **290**, 305-314 (2019).

84 Li, X. *et al.* Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol.* **20**, 193-201 (2019).

85 Wang, S. *et al.* Predicting EGFR Mutation Status in Lung Adenocarcinoma on CT Image Using Deep Learning. *Eur. Respir. J.*, 1800986 (2019).

86 Brinker, T. J. *et al.* A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *Eur. J. Cancer* **111**, 148-154 (2019).

87 Kickingereder, P. *et al.* Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* **20**, 728-740 (2019).

88 Brinker, T. J. *et al.* Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur. J. Cancer* **113**, 47-54 (2019).

89 Choi, K. S., Choi, S. H. & Jeong, B. Prediction of IDH genotype in gliomas with dynamic susceptibility contrast perfusion MR imaging using an explainable recurrent neural network. *Neuro Oncol.* **21**, 1197-1209 (2019).

90 Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954-961 (2019).

91 Yala, A., Lehman, C., Schuster, T., Portnoi, T. & Barzilay, R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology* **292**, 60-66 (2019).

92 Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054-1056 (2019).

93 Liu, Y. *et al.* Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. *Arch. Pathol. Lab. Med.* **143**, 859-868 (2019).

94 Kehl, K. L. *et al.* Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. *JAMA Oncol.* **5**, 1421-1429 (2019).

95 Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301-1309 (2019).

96 Chen, P.-H. C. *et al.* An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453-1457 (2019).

97 Hu, L. *et al.* An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening. *J. Natl. Cancer Inst.* **111**, 923-932 (2019).

98 Rodriguez-Ruiz, A. *et al.* Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *J. Natl. Cancer Inst.* **111**, 916-922 (2019).

99 Wang, X. *et al.* Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis. *IEEE Trans. Cybern.*, 1-13 (2019).

100 Jurmeister, P. *et al.* Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci. Transl. Med.* **11**, eaaw8513 (2019).

101 Courtiol, P. *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519-1525 (2019).

102 Wang, P. *et al.* Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* **68**, 1813-1819 (2019).

103 Liao, F., Liang, M., Li, Z., Hu, X. & Song, S. Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-OR Network. *IEEE Trans. Neural Netw. Learn. Syst.* **30**, 3484-3495 (2019).

104 Luo, H. *et al.* Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. *Lancet Oncol.* **20**, 1645-1654 (2019).

105 Wu, L. *et al.* Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut* **68**, 2161-2169 (2019).

106 Shkolyar, E. *et al.* Augmented Bladder Tumor Detection Using Deep Learning. *Eur. Urol.* **76**, 714-718 (2019).

107 Yamamoto, Y. *et al.* Automated acquisition of explainable knowledge from unannotated histopathology images. *Nat. Commun.* **10**, 5642 (2019).

108 McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89-94 (2020).

109 Hollon, T. C. *et al.* Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat. Med.* **26**, 52-58 (2020).

110 Haenssle, H. A. *et al.* Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann. Oncol.* **31**, 137-143 (2020).

111 Ström, P. *et al.* Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol.* **21**, 222-232 (2020).

112 Bulten, W. *et al.* Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* **21**, 233-241 (2020).

113 Skrede, O.-J. *et al.* Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* **395**, 350-360 (2020).

114 Saillard, C. *et al.* Predicting survival after hepatocellular carcinoma resection using deep-learning on histological slides. *Hepatology* **Advance online publication**, https://doi.org/10.1002/hep.31207 (2020).

115 Jin, E. H. *et al.* Improved Accuracy in Optical Diagnosis of Colorectal Polyps Using Convolutional Neural Networks with Visual Explanations. *Gastroenterology* **158**, 2169-2179.e2168 (2020).

116 de Groof, A. J. *et al.* Deep-Learning System Detects Neoplasia in Patients With Barrett's Esophagus With Higher Accuracy Than Endoscopists in a Multistep Training and Validation Study With Benchmarking. *Gastroenterology* **158**, 915-929.e914 (2020).

117 Bangalore Yogananda, C. G. *et al.* A novel fully automated MRI-based deep-learning method for classification of IDH mutation status in brain gliomas. *Neuro Oncol.* **22**, 402-411 (2020).

118 Zheng, X. *et al.* Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. *Nat. Commun.* **11**, 1236 (2020).

119 Galateau Salle, F. *et al.* Comprehensive Molecular and Pathologic Evaluation of Transitional Mesothelioma Assisted by Deep Learning Approach: A Multi-Institutional Study of the International Mesothelioma Panel from the MESOPATH Reference Center. *J. Thorac. Oncol.* **15**, 1037-1053 (2020).

120 Baldwin, D. R. *et al.* External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax* **75**, 306-312 (2020).

121 Wang, P. *et al.* Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol. Hepatol.* **5**, 343-351 (2020).

122 Song, Q., Zheng, Y., Sheng, W. & Yang, J. Tridirectional Transfer Learning for Predicting Gastric Cancer Morbidity. *IEEE Trans. Neural Netw. Learn. Syst.*, 1-14 (2020).

123 Dong, D. *et al.* Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. *Ann. Oncol.* **31**, 912-920 (2020).

124 Shin, H. *et al.* Early-Stage Lung Cancer Diagnosis by Deep Learning-Based Spectroscopic Analysis of Circulating Exosomes. *ACS Nano* **14**, 5435-5444 (2020).

125 Kann, B. H. *et al.* Multi-Institutional Validation of Deep Learning for Pretreatment Identification of Extranodal Extension in Head and Neck Squamous Cell Carcinoma. *J. Clin. Oncol.* **38**, 1304-1311 (2020).

126 Nature. AI diagnostics need attention. *Nature* **555**, 285 (2018).

127 The Lancet. Is digital medicine different? *Lancet* **392**, 95 (2018).

128 Kawaguchi, K., Kaelbling, L. P. & Bengio, Y. Generalization in Deep Learning. Preprint at https://arxiv.org/abs/1710.05468 (2017).

129 LeCun, Y. in *Connectionism in perspective* (ed. Pfeifer, R., Schreter, Z., Fogelman, F., & Steels, L.) 143-156 (Elsevier, Zürich, Switzerland, 1989).

130 Neyshabur, B., Bhojanapalli, S., Mcallester, D. & Srebro, N. Exploring Generalization in Deep Learning. *Adv. Neural Inf. Process. Syst.* **30**, 5947-5956 (2017).

131 Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345-1359 (2010).

132 Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *J. Big Data* **3**, 9 (2016).

133 Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 248-255 (2009).

134 Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **115**, 211-252 (2015).

135 Shankar, S. *et al.* No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *NIPS Workshop Mach. Learn. Dev. World* (2017).

136 Geirhos, R. *et al.* ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *Proc. Int. Conf. Learn. Represent.* (2019).

137 Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X. & van den Oord, A. Are we done with ImageNet? Preprint at https://arxiv.org/abs/2006.07159 (2020).

138 Sun, C., Shrivastava, A., Singh, S. & Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *Proc. IEEE Int. Conf. Comput. Vis.*, 843-852 (2017).

139 Simard, P. Y., Steinkraus, D. & Platt, J. C. Best practices for convolutional neural networks applied to visual document analysis. *Proc. 7th Int. Conf. Doc. Anal. Recognit.*, 958-963 (2003).

140 Baird, H. S. Document image defect models and their uses. *Proc. 2nd Int. Conf. Doc. Anal. Recognit.*, 62-67 (1993).

141 Stacke, K., Eilertsen, G., Unger, J. & Lundstrom, C. Measuring Domain Shift for Deep Learning in Histopathology. *IEEE J. Biomed. Health Inform.* **Advance online publication**, https://doi.org/10.1109/JBHI.2020.3032060 (2020).

142 Lakhani, P. & Sundaram, B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* **284**, 574-582 (2017).

143 Hussain, Z., Gimenez, F., Yi, D. & Rubin, D. Differential Data Augmentation Techniques for Medical Imaging Classification Tasks. *AMIA Annu. Symp. Proc.* **2017**, 979-984 (2018).

144 Sajjad, M. *et al.* Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *J. Comput. Sci.* **30**, 174-182 (2019).

145 Tellez, D. *et al.* Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* **58**, 101544 (2019).

146 Kerr, R. S. *et al.* Adjuvant capecitabine plus bevacizumab versus capecitabine alone in patients with colorectal cancer (QUASAR 2): an open-label, randomised phase 3 trial. *Lancet Oncol.* **17**, 1543-1557 (2016).

147 Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2818-2826 (2016).

148 Miller, R. G. J. *Simultaneous Statistical Inference*, 2nd edn. (Springer New York, 1981).

149 Hochberg, Y. & Tamhane, A. C. *Multiple Comparison Procedures*. (Wiley, 2009).

150 Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488-492 (2005).

151 Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*, 2nd edn. (Springer-Verlag New York, 2009).

152 Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd edn. (Prentice Hall, 2010).

153 Hemingway, H., Riley, R. D. & Altman, D. G. Ten steps towards improving prognosis research. *BMJ* **339**, b4184 (2009).

154 Korevaar, D. A. *et al.* Facilitating Prospective Registration of Diagnostic Accuracy Studies: A STARD Initiative. *Clin. Chem.* **63**, 1331-1341 (2017).

155 Ioannidis, J. P. A. The Importance of Predefined Rules and Prespecified Statistical Analyses: Do Not Abandon Significance. *JAMA* **321**, 2067-2068 (2019).

156 Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. *Proc. 20th Int. Conf. Pattern Recognit.*, 3121-3124 (2010).

157 van den Hout, W. B. The Area under an ROC Curve with Limited Information. *Med. Decis. Making* **23**, 160-166 (2003).

158 Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861-874 (2006).

159 Harrell, F. E., Jr, Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *J. Am. Med. Assoc.* **247**, 2543-2546 (1982).

160 Lobo, J. M., Jiménez-Valverde, A. & Real, R. AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**, 145-151 (2008).

161 Voosen, P. *How AI detectives are cracking open the black box of deep learning*. Science. https://www.sciencemag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning (2017).

162 Adadi, A. & Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **6**, 52138-52160 (2018).

163 Barredo Arrieta, A. *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82-115 (2020).

164 Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **73**, 1-15 (2018).

165 Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *Proc. Int. Conf. Learn. Represent.* (2014).

166 Bach, S. *et al.* On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One* **10**, e0130140 (2015).

167 Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. *Proc. 34th Int. Conf. Mach. Learn.* **70**, 3319-3328 (2017).

168 Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M. & Granger, C. B. *Fundamentals of Clinical Trials*, 5th edn. (Springer, 2015).

169 van Luijn, H. E. M., Musschenga, A. W., Keus, R. B., Robinson, W. M. & Aaronson, N. K. Assessment of the risk/benefit ratio of phase II cancer clinical trials by Institutional Review Board (IRB) members. *Ann. Oncol.* **13**, 1307-1313 (2002).

170 Martin, L., Hutchens, M., Hawkins, C. & Radnov, A. How much do clinical trials cost? *Nat. Rev. Drug Discov.* **16**, 381-382 (2017).

171 Teutsch, S. M. *et al.* The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative: methods of the EGAPP Working Group. *Genet. Med.* **11**, 3-14 (2009).

172 Vollmer, S. *et al.* Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* **368**, l6927 (2020).

173 Chan, A.-W. *et al.* SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ* **346**, e7586 (2013).

174 Cruz Rivera, S. *et al.* Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* **26**, 1351-1363 (2020).

175 Moher, D. *et al.* CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**, c869 (2010).

176 Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *Lancet* **393**, 1577-1579 (2019).

177 Liu, X. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364-1374 (2020).

178 The Lancet. Should protocols for observational research be registered? *Lancet* **375**, 348 (2010).

179 Loder, E., Groves, T. & MacAuley, D. Registration of observational studies. *BMJ* **340**, c950 (2010).

180 Chambers, C. & Munafo, M. *Trust in science would be improved by study pre-registration*. The Guardian. https://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration (2013).

181 Williams, R. J., Tse, T., Harlan, W. R. & Zarin, D. A. Registration of observational studies: Is it time? *Can. Med. Assoc. J.* **182**, 1638-1642 (2010).

182 Gill, J. & Prasad, V. Improving observational studies in the era of big data. *Lancet* **392**, 716-717 (2018).

183 Sørensen, H. T. & Rothman, K. J. The prognosis for research. *BMJ* **340**, c703 (2010).

184 Vandenbroucke, J. P. Registering observational research: second thoughts. *Lancet* **375**, 982-983 (2010).

185 Epidemiology. The Registration of Observational Studies—When Metaphors Go Bad. *Epidemiology* **21**, 607-609 (2010).

186 Andre, F. *et al.* Biomarker studies: a call for a comprehensive biomarker study registry. *Nat. Rev. Clin. Oncol.* **8**, 171-176 (2011).

187 Hooft, L. & Bossuyt, P. M. Prospective Registration of Marker Evaluation Studies: Time to Act. *Clin. Chem.* **57**, 1684-1686 (2011).

188 Altman, D. G. The Time Has Come to Register Diagnostic and Prognostic Research. *Clin. Chem.* **60**, 580-582 (2014).

189 Rifai, N. *et al.* Registering Diagnostic and Prognostic Trials of Tests: Is It the Right Thing to Do? *Clin. Chem.* **60**, 1146-1152 (2014).

190 Rajkomar, A., Dean, J. & Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* **380**, 1347-1358 (2019).

191 Zou, J. & Schiebinger, L. AI can be sexist and racist - it's time to make it fair. *Nature* **559**, 324-326 (2018).

192 Adamson, A. S. & Smith, A. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol.* **154**, 1247-1248 (2018).

193 Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *N. Engl. J. Med.* **383**, 874-882 (2020).

194 Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447-453 (2019).

195 Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann. Intern. Med.* **169**, 866-872 (2018).

196 Owens, K. & Walker, A. Those designing healthcare algorithms must become actively anti-racist. *Nat. Med.* **26**, 1327-1328 (2020).

197 Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691-698 (2012).

Acknowledgements

We thank Marian Seiergren for assembling all figures, Tarjei Sveinsgjerd Hveem for discussions, Trond Ystanes, Haakon André Inderhaug and Bjørn Morten Sannes for setting up and maintaining our computer network and computational infrastructure, and the authors of Inception-v3 for making their code freely available under an open source licence (Apache License, Version 2.0). We acknowledge funding from the Research Council of Norway through its IKTPLUSS Lighthouse program (project number 259204).

Author contributions

H.E.D and D.J.K initiated the project. All authors researched data for the article. A.K., O.J.S. and K.L. assessed papers in the review of recent, presumably influential deep learning studies in cancer diagnostics. S.D.R. executed the training, tuning and evaluation of Inception-v3 systems. A.K. drafted the manuscript, and all authors contributed to reviewing and editing the manuscript.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Cancer thanks J. Kather and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Related links

Journal Policies in the Journal of Clinical Oncology: https://ascopubs.org/jco/authors/journal-policies

ClinicalTrials.gov registry: https://www.clinicaltrials.gov

International Standard Randomised Controlled Trial Number (ISRCTN) registry: https://www.isrctn.com

Supplementary information

Supplementary information is available for this paper at https://doi.org/10.1038/s415XX-XXX-XXXX-X

Glossary

Artificial neural networks

Mathematical functions mapping input data to output representations, structured as a directed graph of nodes and edges.

Deep learning

A class of machine learning methods that make use of successively more abstract representations of the input data to perform a specific task, typically implemented using artificial neural networks. They also consist of an objective function that compares the final output with a target output as well as an optimisation method that is used to optimise the objective function.

Deep learning models

Computational models obtained by training deep neural networks. Note that a single training of a neural network produces a sequence of models since each new optimisation iteration produces a model slightly different from the previous one. A tuning dataset may be used to select among these models.

Deep learning systems

Systems utilising one or more deep learning models to make predictions. A system's output may be a function of the outputs of the models, e.g. by averaging and thresholding the model outputs.

Supervised machine learning

A methodology in which learning occurs by mimicking the mapping of input data to target output labels. In contrast, the input data are not associated with any output labels in unsupervised learning.

Capacity

The ability of a model class, e.g. a particular network architecture, to express complicated correlations between input data and target output. Model classes with high capacity have the potential to produce models that are able to map training data to target outputs with a high degree of accuracy, but are also more prone to overfitting.

Development cohort

A cohort used for training and sometimes tuning and internal validation of a system.

External cohorts

Also known as independent cohorts, these differ non-randomly from the development cohort. In cancer diagnostics, the external cohorts will often contain patients suspected of having the same disease or disease attribute, at risk of developing the same event or suspected to respond to the same treatment as patients in the development cohort. However, external cohorts may be intentionally more different from the development cohort.

Training

Optimisation of model parameters based on data.

Tuning

Informed selection of hyperparameter values (parameters not optimised during training) based on data. Examples include network architecture, optimisation method and threshold for a model's continuous output. The nomenclature in machine learning is to use 'validation' instead of 'tuning'.

Test

While frequently used by the machine learning community to refer to an evaluation of a system's performance, we use 'test' to refer to evaluations other than external validations, e.g. internal validations.

External validation

An evaluation of a system's performance on an external cohort that did not influence the development of the system.

Overfitting

Utilising noise or features in the training data that are not generally relevant for the prediction task but cause the system to perform better on the training sample.

Generalisability

The ability of a system to perform similarly on subjects not included in training as on those included in the training. Poor generalisability can be caused by overfitting to the training data or by the lack of generally relevant features in the training data.

Balanced accuracy

A classification performance metric calculated by averaging the proportion of true predicted outcomes across all possible outcomes. For dichotomous outcomes, this reduces to the average between the sensitivity and specificity.

Area under the receiver operating characteristic curve (AUC)

A performance metric measuring the concordance between a dichotomous outcome and the ranking of subjects provided by a continuous or categorical marker. An AUC of 50% indicates random guessing and 100% indicates perfect prediction. For dichotomous markers, AUC and balanced accuracy are equivalent.

Concordance index (c-index)

A performance metric measuring the concordance between a target outcome, usually defined by time-to-event data, and the ranking of subjects provided by a continuous or categorical marker. A c-index of 50% indicates random guessing and 100% indicates perfect prediction. For dichotomous outcomes, c-index and AUC are equivalent.

983 **Figure legends**

984

985 Fig. 1 | **Characteristics of recent, presumably influential deep learning studies in cancer diagnostics. a** |
986 Percentage of studies reporting on the evaluation of a broad or narrow cohort (BOX 2) by year of publication, for
987 all 92 eligible studies. **b** | Percentage of studies specifying one, multiple or no primary performance metrics in
988 the analysis of the external cohort, for the 50 eligible studies that reported on the evaluation of an external
989 cohort. **c** | Percentage of studies specifying a predefined analysis of the external cohort, for the 50 eligible studies
990 that reported on the evaluation of an external cohort. Studies that specified predefined analyses of external
991 cohorts without defining which one was the primary, if any, were categorised as 'Predefined analyses'. Studies
992 with a predefined primary analysis were categorised according to whether the primary analysis was prespecified
993 in a protocol or not.

994

995 Fig. 2 | **Effect of data variation when training deep learning systems.** For each analysis setup, 20 individual
996 deep learning systems were trained and tuned for prediction of colorectal cancer-specific survival using images
997 of haematoxylin and eosin stained sections acquired by both Aperio AT2 (Leica Biosystems, Germany) and
998 NanoZoomer XR (Hamamatsu Photonics, Japan), as in the previously published analyses[113]. The individual
999 systems were applied to evaluate the external cohort using NanoZoomer XR slide images, and the c-index of the
1000 system's binary output was computed. Standard box plots were made using Stata/SE 16.1 (StataCorp, TX). The
1001 matched random subset contained the same number of training and tuning patients with and without cancer-
1002 specific death as in the Gloucester cohort, in total 979 patients. **a** | An example image from the training dataset
1003 and the results of applying the maximum possible amount of colour distortion at each step in the random
1004 distortion process used in the published Inception-v3 analyses[113]. Generally, the distortion process applies
1005 random colour distortions to an image by: 1) converting the image to HSV colour space, 2) adding a random
1006 value between -0.05 and 0.05 to the hue, 3) scaling the saturation by a random value between 1/1.1 and 1.1, 4)
1007 adding a random value between -0.1 and 0.1 to the saturation, 5) scaling the brightness (or technically the value
1008 channel in the HSV colour space) by a random value between 1/1.1 and 1.1, 6) adding a random value between -
1009 0.1 and 0.1 to the brightness, and 7) converting back to RGB colour space. Intuitively, the leftmost and rightmost
1010 images represent the range of the random colour distortion, i.e. the minimum and maximum possible amount of
1011 colour distortion for the applied distortion process, where the minimum is no colour distortion. Scale bar, 100

µm. **b** | Effect of changing the number of patients in training and tuning when using the original amount of colour distortion, as depicted in figure part **a**. **c** | Effect of changing the amount of colour distortion when training and tuning using the matched random subset. Label '0' on the horizontal axis identifies deep learning systems trained without any colour distortion, label '1' identifies systems trained with the colour distortion process depicted in figure part **a**, and label '4' identifies systems trained with the colour distortion process depicted in figure part **d**. **d** | Similar to figure part **a**, but four times the amount of colour distortion was used at each step in the distortion process. Scale bar, 100 µm. **e** | Effect of changing the amount of colour distortion and the number of patients and cohorts in training and tuning.

Fig. 3 | **Development and evaluation of deep learning systems.** A deep learning project often begins with testing a conceptual idea using a pilot software based on a related open source implementation and data easily available to the researchers. Successful level I studies will typically evolve into explorative testing of different modelling options that might be more suitable for the particular task. The system that appears to perform best should be determined in a level II study that includes sufficient amount and variation in the natural training dataset. Although performance estimates obtained in such studies are often inflated by the use of a subset that closely resembles the training subset, level II is an important step in the evaluation sequence that could motivate investigators to pursue evaluation on external cohorts and attract collaborators. As the lack of predefined primary analysis often entails post hoc adjustments influenced by the performance in the external cohort, we distinguish between studies without (level III) and with (level IV) a primary analysis unequivocally specified prior to all investigations that could reveal correlations between input data and target output in the external cohort. If the medical validity of a deep learning system is established in level IV studies, the indicated medical utility should be prospectively evaluated in randomised phase III clinical trials where the system directly intervenes with the current standard of care. If medical utility is demonstrated and necessary governmental agencies approve routine medical application, the system can be applied in medical practice while monitoring the long-term benefits, harms and costs of its application.

Fig. 4 | **Reliability of performance estimations in recent, presumably influential deep learning studies in cancer diagnostics.** Percentage of studies categorised in the different levels of deep learning studies or phases of clinical trials depicted in FIG. 3, for all 92 eligible studies separated by type of input to the neural network. The

input was histopathology section images in 23 (25%) of the studies **(a)**, radiology images in 40 (43%) of the studies **(b)**, other images in 22 (24%) of the studies **(c)** and other types of input in 7 (8%) of the studies **(d)**.

Boxes

Box 1 | **Representation and biases in training data**

As deep learning systems are developed by learning correlations between input data and target outcome directly from the training data, it is essential that the training data adequately represents the target population[31,190]. Otherwise, the system might learn correlations exclusive to the subpopulation represented in the training data and consequently perform worse on those not represented in the training data to a sufficient extent. Despite this, systems are often trained on datasets with prominent biases in demographic characteristics such as sex, race or ethnicity, with the consequence that many systems exhibit substantial discriminatory biases[32,191,192]. Restricting the target population to a particular sex, race or ethnicity would be appalling, and the medical application of any such deep learning system would systematically increase health care disparities. It is therefore pivotal to utilise truly representative and unbiased data for training deep learning systems in cancer diagnostics. This extends beyond ensuring representative distributions of relevant demographics in the training dataset. Racial bias may also be encoded into systems if the target outcome used in the training is affected by histories of unequal treatment of patients based on race or ethnicity[193] or is a proxy such as health care cost instead of health needs, which has been shown to be the reason why a widely used health care prediction algorithm exhibited significant racial bias[194]. Researchers should strive to identify and compensate for any such biases in their datasets, as failure to do so might reinforce health inequities if the deep learning systems are applied in clinical practice[195,196]. Deficient deep learning systems might be identified through rigorous evaluations in external datasets truly representative of the target population, or representative of minority populations, as well as through comprehensive analyses of system explainability across different demographic characteristics.

Box 2 | **Approaches for evaluating a deep learning system**

Different approaches for estimating the performance of a deep learning system provide indications of the system's ability to make accurate predictions in different scenarios. Even if successful, internal and narrow validations do not indicate a general medical validity in themselves. Successful broad or domain validations might warrant assessment of the system's medical utility in prospective, randomised phase III clinical trials.

[bH1] Internal validation

Internal validation is evaluation of a deep learning system's performance in the development cohort. This can be done by evaluating the performance in a randomly sampled subset of the development cohort disjoint from the training and tuning subsets, or by using resampling techniques such as cross-validation or bootstrapping[22].

[bH1] Narrow validation

Narrow validation is evaluation of a deep learning system's performance based on a cohort that is similar but differs non-randomly from the development cohort, e.g. on a cohort from the same hospital as the development cohort, but sampled in a time interval disjoint from the time interval where the development cohort was sampled. No information from the narrow cohort should have influenced the development of the system, including that it should be collected and handled separately from the development cohort.

[bH1] Broad validation

Broad validation is evaluation of a deep learning system's performance based on a cohort geographically separate from the development cohort, e.g. from a different hospital or country[22]. No information from the broad cohort should have influenced the development of the system.

[bH1] Domain validation

Domain validation is evaluation of a deep learning system's performance in a setting that is very different from the one where the system was developed[197]. This includes validation in a cohort with characteristics not represented by the development cohort, e.g. developing a method on one type and stage of cancer and validating it on another type or stage of cancer. Other examples are when the validation data are obtained by equipment not used in the development such as imaging systems from different vendors, or by sample preparation procedures intentionally different from the ones used for the development cohort. Domain validations should also be narrow or broad validations, and are typically performed after successful narrow or broad validations.

1093

Box 3 | **Recommended Protocol Items for External Cohort Evaluation of a deep learning System (PIECES) in cancer diagnostics**

[bH1] Status

[b1] Specify the date the protocol was last modified.

[b1] Scrupulously elucidate any investigations performed before finalising the protocol that could reveal correlations between input data and target output in the external cohort, or state that no such investigations were performed.

[bH1] System

[b1] Describe the development of the deep learning system, including utilised cohorts, network architecture, hyperparameters and any categorisation of the neural network model's output.

[b1] Unequivocally specify how to assay the deep learning system in a blinded fashion for a single, new subject, including what the system receives as input and what it directly outputs.

[bH1] External cohort

[b1] Describe the origin of the cohort, and explain why it should be regarded as external to the development cohort.

[b1] Precisely define criteria for inclusion and exclusion of subjects and samples, preferably starting from a consecutive series of subjects.

[b1] Clearly state the medical setting and target population that the cohort represents.

[b1] Specify the acquisition of input data, including whether it was acquired blinded to the deep learning system and target output. Note the expertise of any humans involved in the process, e.g. that a pathologist annotated the regions of interest in slide images.

[b1] Specify the ascertainment of target output, including whether it was ascertained blinded to the deep learning system.

[b1] If multiple external cohorts are planned to be analysed as a pooled cohort, then the preceding five protocol items should be completed for the pooled cohort and differences between the individual cohorts should be stated.

If multiple external cohorts are to be analysed independently, the five preceding protocol items should be completed for each cohort, as well as subsequent protocol items if the predefined analyses differ between cohorts.

[bH1] Analyses

[b1] Unequivocally specify the primary analysis, including the target output and the performance metric and/or statistical test with interpretation.

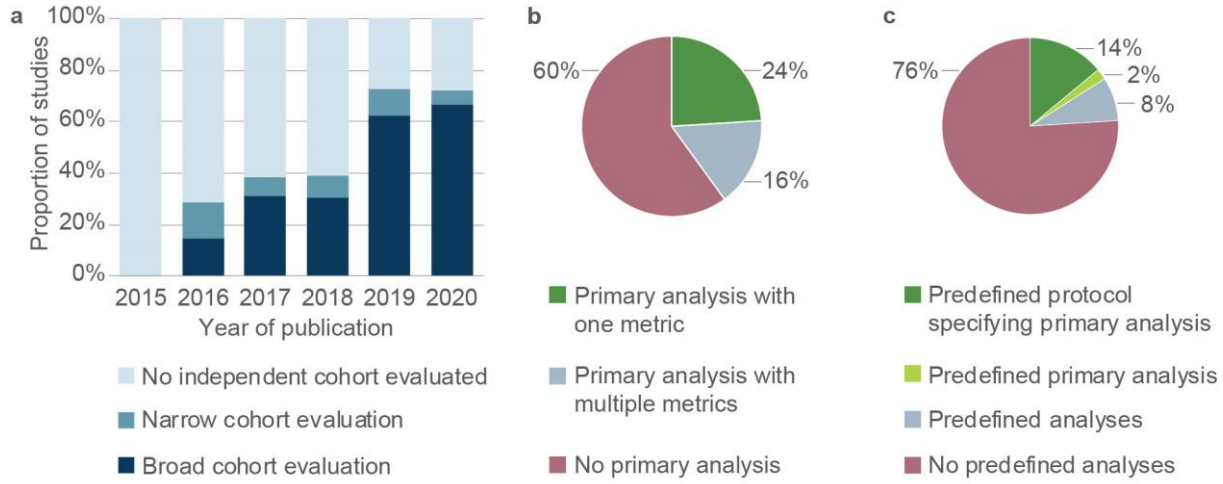
[b1] If the chosen metric or statistical test depends on other markers, describe how these markers were assayed and whether done blinded to the deep learning system and target output, and specify how missing values will be handled.

[b1] If the deep learning system was designed to evolve upon usage, e.g. by learning from unlabelled data or adapting to a cohort, specify that this will not be done when evaluating the external cohort. The system's prediction should thus not depend on the order in which a set of patients is evaluated and also be identical if the same patient is evaluated multiple times.

[b1] If additional analyses will be performed and reported in disseminations, e.g. of other deep learning systems, target outputs, metrics or statistical tests or in specific patient subgroups, specify these analyses in the same manner as the primary analysis and identify them as secondary analyses.

Table of Contents Summary

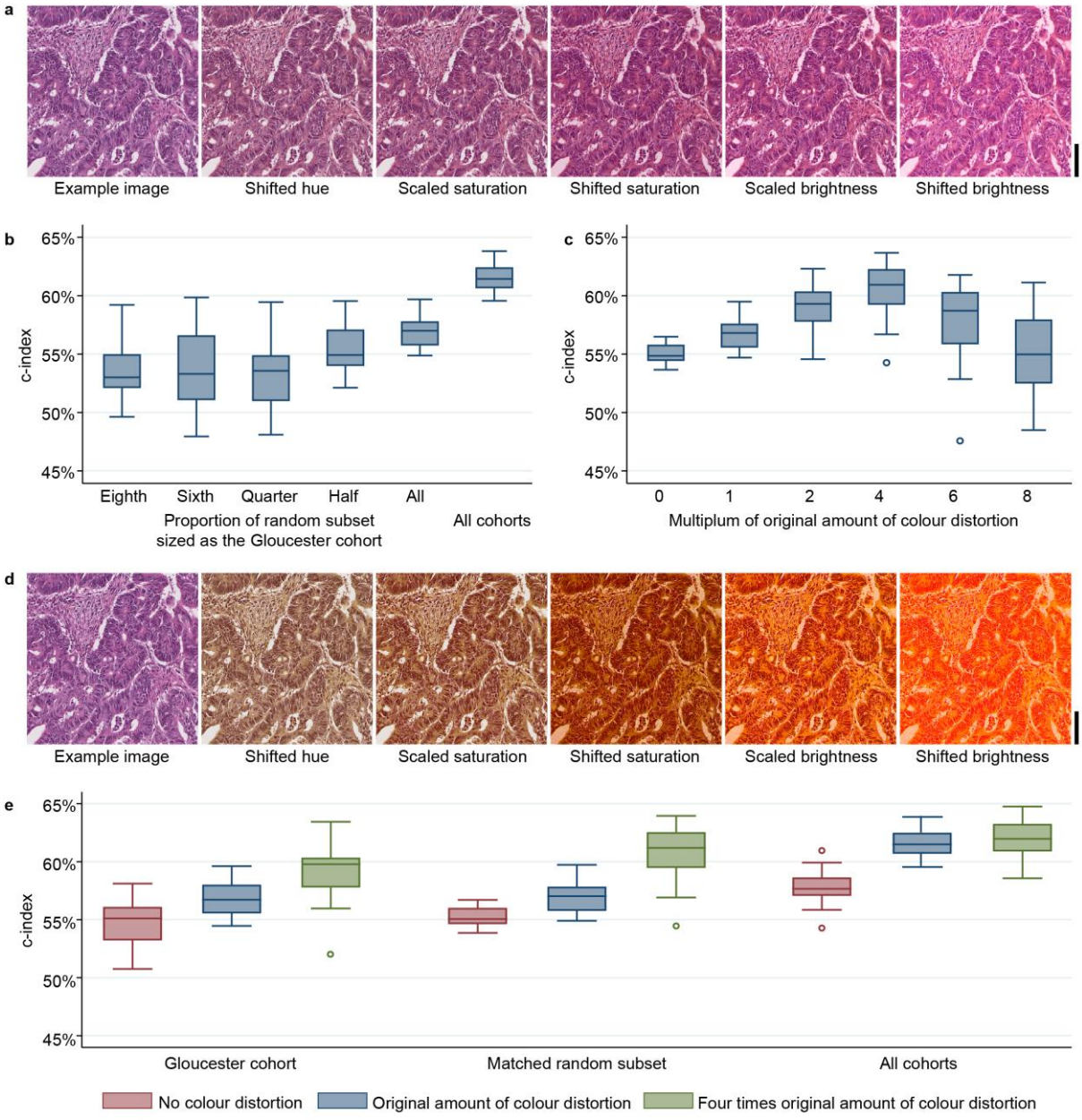
The number of publications on deep learning for cancer diagnostics is rapidly increasing, but clinical translation is slow. This Perspective advocates performance estimation in external cohorts, and strongly advises that a primary analysis is predefined in a standardised protocol preferentially stored in an online repository.



1140

1141 Figure 1.

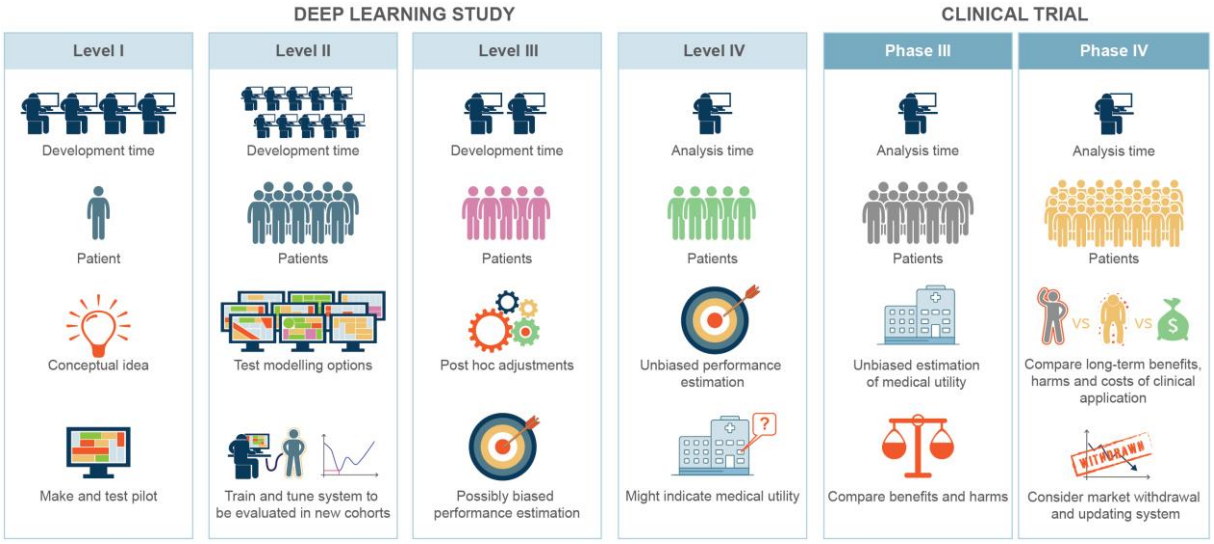
1142



1143

1144 Figure 2.

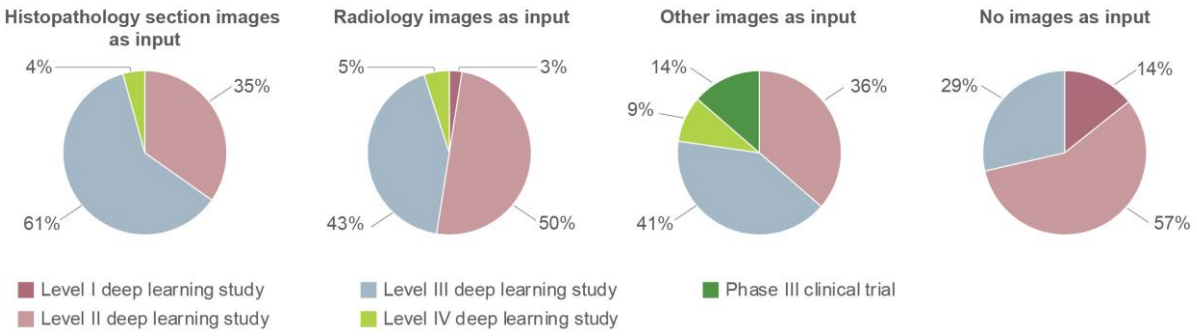
1145



1146

1147 Figure 3.

1148



1149

1150 Figure 4.