AI and Ethics (2025) 5:2435–2444
https://doi.org/10.1007/s43681-024-00561-z

ORIGINAL RESEARCH

Ethical considerations of use of hold-out sets in clinical prediction model management

Louis Chislett[1] · Louis J. M. Aslett[2,3] · Alisha R. Davies[3] · Catalina A. Vallejos[1,3] · James Liley[2]

Received: 8 June 2024 / Accepted: 21 August 2024 / Published online: 10 September 2024

Abstract

Clinical prediction models are statistical or machine learning models used to quantify the risk of a certain health outcome using patient data. These can then inform potential interventions on patients, causing an effect called performative prediction: predictions inform interventions which influence the outcome they were trying to predict, leading to a potential underestimation of risk in some patients if a model is updated on this data. One suggested resolution to this is the use of hold-out sets, in which a set of patients do not receive model derived risk scores, such that a model can be safely retrained. We present an overview of clinical and research ethics regarding potential implementation of hold-out sets for clinical prediction models in health settings. We focus on the ethical principles of beneficence, non-maleficence, autonomy and justice. We also discuss informed consent, clinical equipoise, and truth-telling. We present illustrative cases of potential hold-out set implementations and discuss statistical issues arising from different hold-out set sampling methods. We also discuss differences between hold-out sets and randomised control trials, in terms of ethics and statistical issues. Finally, we give practical recommendations for researchers interested in the use hold-out sets for clinical prediction models.

Keywords Performative prediction · Hold-out sets · Clinical prediction models · Ethics

1 Introduction

One typical goal of machine learning is to predict an outcome Y given a set of covariates (features) X [1]. Here we focus on binary outcomes, where Y denotes whether an event of interest occurs ($Y = 1$) or not ($Y = 0$). Risk scores are often used to estimate the probability of observing the event $Y = 1$ given the values of X. In healthcare applications, this is commonly referred to as a clinical prediction model (CPM), where X captures the patient-level information and Y would represent a health outcome being present or occurring in the future [2]. It is worth noting that models which aim to predict a probability of future outcome are referred to as "prognostic rules", while those that aim to predict the probability of an intervention being successful are called "prescriptive rules" [3]. This paper primarily deals with prognostic rules. Thus the output of a CPM in our case is a risk score, that is $Pr(Y = 1|X)$. For example, the EuroSCORE II CPM predicts whether a patient will die prior to hospital discharge after cardiac surgery using the patient's age, gender and existing medical conditions as input covariates [4].

A fundamental issue in the development and application of CPMs is their accuracy and how this evolves over time. If the distribution of (X, Y) changes (often called "drift"), then the risk scores may become biased, leading to less accurate predictions [5]. For example, one study showed that drift affecting model accuracy occurred in several machine learning methods when predicting 30-day mortality after hospital admission [6].

✉ James Liley
james.liley@durham.ac.uk

Louis Chislett
louis.chislett@ed.ac.uk

Louis J. M. Aslett
louis.aslett@durham.ac.uk

Alisha R. Davies
alisha.davies@wales.nhs.uk

Catalina A. Vallejos
catalina.vallejos@ed.ac.uk

[1] MRC Human Genetics Unit, University of Edinburgh, Crewe Rd S, Edinburgh EH4 2XU, UK

[2] Department of Mathematical Sciences, Durham University, Stockton Rd, Durham DH1 3LE, UK

[3] The Alan Turing Institute, 96 Euston Rd, London NW1 2DB, UK

Reasons for drift can be diverse and can act simultaneously [7]. These can include changes in the distribution of Y (e.g. increase prevalence of a disease), changes in relationships between X and Y (e.g. a risk factor becoming less predictive), changes in the distribution of X (e.g. an aging population, if age is a predictor), as well as "performative prediction" effects [8, 9]. Performative prediction is a potentially understudied concept in which a risk score influences the distribution of (X, Y) through interventions made based on its predictions. When used in this way, CPMs inform what are referred to as Clinical Decision Support Systems [10]. In a health context, interventions made based on risk scores distributed to patients or doctors may induce some performative effect: individuals with high risk scores may be prioritised for intervention which, if successful, may reduce their risk - thus changing the relationship between X and Y. For example with QRISK3, a model which predicts 10-year risk of cardiovascular disease, a high risk score can inform an intervention such as a prescription of statins for the patient [11]. If the model were to be retrained using post-intervention data as input, similar patients would receive risk scores which are underestimations of their true risk [12]. Performative effects can be amplified by having a more effective model or intervention, resulting in models becoming "victims of their own success" [13].

There is no consensus within the field on how to deal with a performative effect in a prediction-intervention system. The natural response to drift is to re-train (update) the model using more recent data. However, as illustrated above, this is not optimal in the presence of performative effects. One method which has been proposed is to "hold-out" a set of individuals who do not receive risk score guided interventions, on which a model can be retrained [12]. Hold-out sets are similar to control arms in clinical trials, but the goal is to mitigate the effects of performative prediction when updating a CPM rather than to measure the effect of the CPM itself (such as whether a measured decrease in patients with $Y = 1$ occurs). While hold-out sets may be trivial to implement in other contexts where the prediction does not affect a person's well-being (e.g. a social media platform which uses a model to predict whether someone will click an advertisement), this is not the case with CPMs. If hold-out sets are to be a solution to the performative prediction problem in CPMs, there are a number of ethical, practical and statistical issues which need to be considered. We consider these issues in the context of different forms of hold-out set sampling, with the view to aiding more informed decision making by researchers looking to use this method. Although the ethical considerations are applicable to other health systems, this paper takes a UK focus.

2 Methods

2.1 Setting

CPMs can exist in a range of settings, using different data sources, and affecting interventions in different ways. We consider the following setting:

> A CPM is trained on patient-level clinical information (e.g previous diagnoses) to predict whether a patient will go on to develop an adverse health outcome (e.g. disease) under standard medical care. Risk scores are then distributed to physicians. They may use these to inform interventions which they recommend to the patient in order to prevent or delay adverse outcomes. Clinicians do not base interventions solely on the risk scores, but also on their expert assessment of the patient (which will generally include information not used by the CPM).

We assume that patients generally benefit from more accurate CPMs, and that risk scores are interpreted as "the risk under typical clinical practice if we did not use a CPM to inform interventions". Furthermore, we assume that a sufficiently accurate CPM provides valuable information to a GP above with which they would otherwise have, allowing them to make an informed decision—this is the cause of performative effects. If drift occurs, the CPM gradually becomes less accurate [6]. To address this, CPMs need to be updated (re-trained) using recent data. If there are performative effects, then the data used to re-train the model will reflect the actions performed in response to predictions made by the original CPM. New CPMs fitted directly to these data will only estimate the risk of the outcome in the setting where the CPM was already in use. This may be a poor approximation of the "risk under typical practice" above. For instance, in the EuroSCORE2 score [4], a risk score fitted directly to population data would approximate an individual's risk of heart attacks after their doctor had already seen and possibly acted on their existing EuroSCORE2 score.

The use of a causal framework to model the effect of CPM-informed interventions has been proposed as a solution this issue [13, 14]. However, direct measurement or recording of such interventions is often impractical, particularly in most UK healthcare settings which lack the required digital infrastructure. To safely update a CPM in the presence of performative effects, we argue the need to have up-to-date data which reflects typical clinical practice without a CPM. This is tantamount to the use of a hold-out set.

2.2 Necessity of hold-out sets

The use of hold-out sets modify the setting described above is as follows:

A CPM is trained and mutually exclusive and complementary "hold-out" and "intervention" sets are sampled, which we will refer to as (X^H, Y^H) and (X^I, Y^I) respectively, where the superscript refers to hold-out (H) or intervention set (I). Risk scores are then distributed to physicians, but only for patients in the intervention set. Patients in the hold-out set have access to the same interventions as those in the intervention set, but do not have risk scores distributed to physicians which could inform decision making. When the CPM is updated, it uses data exclusively from the "hold-out" set to train the model. Figure 1 displays the causal dynamics of the system, including how hold-out sets (X^H, Y^H) are used to retrain CPMs, and how performative effects occur in the intervention set (X^I, Y^I).

Although the use of hold-out sets is not yet standard, and alternative options may be preferable in some cases, we argue toward their general necessity when monitoring and updating CPMs. Other possible methods such as causal modelling approaches have been suggested [14]. We argue that most CPMs will have a performative effect: indeed, CPMs are typically designed in order to bring about an effect on the outcome Y. Furthermore, medical systems or care pathways change with time, often deliberately, inducing non-performative drift. This can cause the accuracy of CPMs to deteriorate, resulting in the need to update. Finally, most CPMs are used to guide interventions in complex circumstances, in which it is not possible to exactly specify actions which came about as a result of the CPM (for instance, the degree to which a decision to operate was directly influenced a patient's EuroSCORE II assessment). This precludes direct measurement of the effect of the CPM on interventions and, in turn, on patient outcomes. Hence, this limits the use of causal modelling approaches at solving the problem of performative prediction when updating CPMs. By re-training a CPM on patients whose treatment is not influenced by the risk score, we continue to predict the "risk of outcome under typical clinical practice".

Hold-out sets could be used to both monitor and update a CPM. Evaluating the effects of a risk score guided intervention system is possible using a randomly sampled hold-out set, in which the only difference between hold-out and intervention set is the use of CPM derived risk scores, akin to a treatment in clinical trials. Any "treatment effect" that comes with use of the model is then easily derived, with the control group being standard medical care. However, our arguments frame the use of hold-out sets primarily as a mechanism to update CPMs in the presence of performative effects, rather than as a research tool to gain knowledge. Additionally, hold-out sets can be used to retrain a CPM when enough drift has occurred to cause a deterioration in accuracy. By retraining on the hold-out set, performative effects of the risk scores are not present in the training data, ensuring risk scores more accurately reflect standard medical care (without a CPM).

2.3 Sampling hold-out sets

We consider three sampling methods for hold-out sets; simple random sampling, cluster randomised sampling and voluntary response sampling [15]. We will explore the merits and drawbacks of each. We note that existing analysis of hold-out sets implicitly consider only simple randomised sampling [16].

In a simple random sampling framework, hold-out sets are drawn as a uniform random sample of the population without the explicit informed consent of patients. This creates a hold-out set with a high degree of external validity. From a statistical perspective, this makes this form of

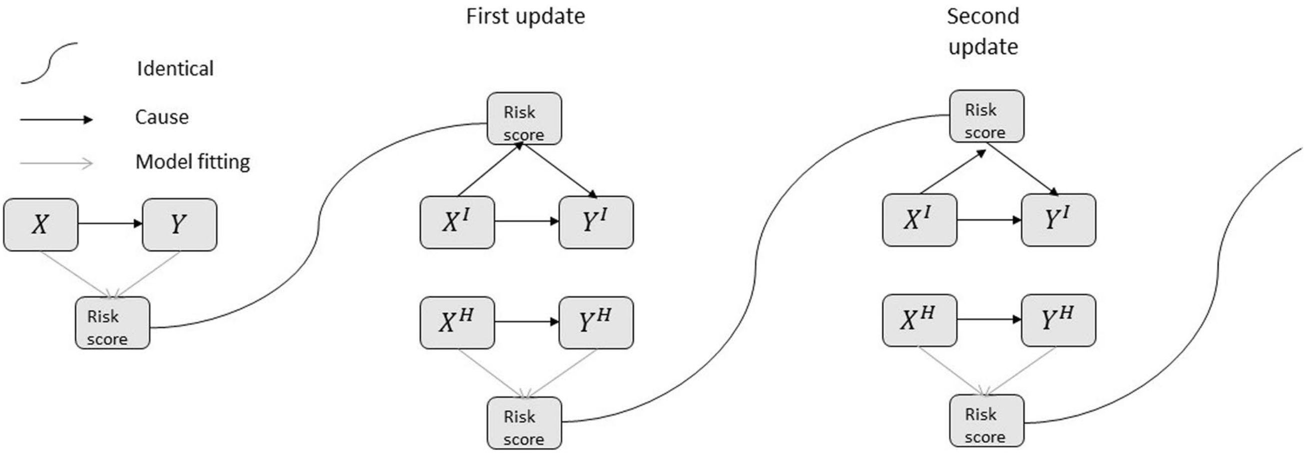

Fig. 1 Dynamics of a CPM trained once and updated twice using hold-out set methodology. Squares containing X and Y denote covariates and outcome respectively, with superscripts I and H denoting mutually exclusive intervention and hold-out sets

sampling ideal, as it does not introduce bias into the hold-out set [15].

In cluster randomised sampling, the population is split into clusters (e.g hospitals or geographic areas), with the hold-out set consisting of patients from a number of randomly selected clusters. Informed consent would not be gained from individuals. If the clusters in the hold-out set are biased in some way, then this will likely worsen the external validity of the hold-out set [15].

In a voluntary response setting, patients could explicitly consent to be included in a hold-out set. However, this would leave a number of statistical issues. Those who volunteer would likely not constitute a representative sample of the population, meaning the hold-out set would likely represent a biased sample [15]. This means that a model re-trained on the hold-out set would have poor external validity. Furthermore, it may be impractical to seek consent from sufficiently many volunteers to form an adequately sized hold-out set.

Precise comparison of statistical properties of the different types of hold-out set sampling (e.g. sample size requirements) remains an open problem which is beyond the scope of this work.

2.4 Ethical considerations

We principally consider the fundamental ethical principles of beneficence, non-maleficence, autonomy and justice [17]. We will assess hold-out set sampling methods against these ethical principles, alongside some further principles which derive from these.

2.4.1 Beneficence

The principle of beneficence requires that treatments or policies benefit the individuals in the target population [18]. This presents an apparent dilemma. The use of hold-out sets can lead to a conflict between individual welfare (for those in the hold-out set), and common welfare. On a per-patient basis, if a sufficiently accurate CPM-derived risk score is available, the principle of beneficence obliges the physician to consider the risk score in their decision making process on intervention/treatment [17]. However, this may lead to the presence of performative effects which, if not using hold-out sets, could result in a generally worse score for each individual in the population after the CPM is updated. The negative effects of inaccurate CPM-derived risk scores can accrue without bound over the population.

Generally, patients would have an expectation for physicians to act in their best interests, and thus to universally consider available CPMs, eventually meaning that the CPM becomes inaccurate (and hence we lose benefit for all patients). In simple- and cluster- randomised sampling, given that consent would not be sought, the principle of beneficence may be broken at the individual level. However, collectively (applied to each individual in the total patient population), beneficence indicates that a medical regulator should take the option which results in more accurate risk scores. This suggests use of a randomised (or cluster-randomised) hold-out set, since any other option leads to a less accurate updated CPM.

Cluster randomised sampling ensures an equivalent standard of care for patients in the same cluster, although means there are differing standards between clusters. This would ensure that at a patient level, physicians could still act in the best interests of patients using all available information to them at that time, but it would result in differences between outcomes for populations in different clusters (postcode lotteries of care).

Voluntary response sampling somewhat avoids the per-patient violation of the principle of beneficence, since patients are knowingly agreeing to turn down the benefits associated to the use of a CPM (even with informed consent gained for those in the hold-out set, a physician must still act in the best interest of a patient). However, since this approach will typically lead to a biased hold-out set and hence a less accurate model than would be attained with a randomised hold-out set, use of a voluntary response hold-out set may violate the principle of beneficence on a population scale.

2.4.2 Non-maleficence

Considerations of non-maleficence, the ethical principle of the avoidance of harming patients, [19, 20] largely mirror those of beneficence. Gaining informed consent from patients does not give a blanket indemnity to use hold-out sets in scenarios where direct harm could occur to individuals due to inclusion in the hold-out set. It is thus vital that risks to patients are minimised after a decision is made on whether the patient receives a risk score. Depending on the outcome that is predicted by the CPM, patients in the hold-out set may be at risk of harm if they are not intervened on. In those cases, it may not be ethically justifiable to use hold-out sets even with consenting volunteers.

However, there is also non-maleficence considerations when employing CPMs without a hold-out set. If lack of a hold-out set leads to inaccurate risk scores, particularly for risk scores which underestimate risk for high risk patients, then those patients may be at risk of harm due to misdiagnosis or lack of interventions which may have otherwise been applied.

2.4.3 Autonomy

Generally, the use of hold-out sets may be regarded as withholding valuable information with which a patient may make rational, autonomous decisions. That is to say that, although a physician can recommend an intervention, a patient has the right to make the final decision. However, withholding a risk score does not affect the set of interventions available to a patient. Moreover, patients always retain the autonomy to go ahead with any interventions agreed with a physician in any form of hold-out sets.

The principle of autonomy in our setting could cover not only what choice is available to patients from a intervention perspective, but also whether a patient has had the chance to choose whether a CPM is used in their case. In the case of cluster randomised sampling, all patients within a cluster would have the same level of autonomy as each other within that cluster, resulting in an opportunity for a CPM to inform clinical decisions at a cluster level. However, given patients within a cluster randomised hold-out set sampling frame would not have given explicit informed consent to be in the hold-out, or not, dataset, this could be seen to be an unreasonable withdrawal of patient autonomy.

Voluntary response sampling may be able to solve some issues regarding patient autonomy. In particular, patients would be volunteering to have risk scores withheld from themselves and physicians. Provided the patient maintains the ability to withdraw consent, they may still have ultimate autonomy over possible future interventions. Hence, voluntary response sampling generally may provide a greater level of autonomy to patients than randomised sampling without informed consent.

2.4.4 Justice

One key consideration in any hold-out set CPM framework is that of distributive justice—that is, the fair distribution of healthcare services and treatments to patients.

If a CPM is trained on biased data, either on initial training, or when updating using a hold-out set, it may produce a model which disproportionately benefits over-represented groups [21]. Furthermore, it is possible for models to be unfair due to other modelling aspects even when the training data itself is an unbiased sample of the population. We nevertheless argue towards a unbiased hold-out set as necessary for justice considerations.

Simple random sampling has the benefit of not systematically benefiting or harming certain protected classes such as race or gender by over/under inclusion in the hold-out set, relative to presence of these classes in the data source (as long as the general training dataset for the CPM is unbiased [22]).

Justice could be an issue in the case of cluster randomised sampling. There would need to be careful consideration given to whether any classes are being systematically selected due to availability or compromised positions. Given the geographic variance in health outcomes in the UK [23], it may also be difficult to obtain a truly representative sample, leading to a biased dataset and ultimately a less accurate CPM.

Voluntary response sampling is likely to lead to biased samples. In particular, it may lead to over-representation of certain groups in the hold-out set who are more likely to volunteer, known as volunteer bias. This occurs in traditional research cohorts such as the UK Biobank [24]. Volunteer bias will impact new risk scores upon model retraining, and may lead to worse accuracy amongst under-represented groups. Additionally, if certain groups volunteer for the hold-out set at greater rates, these groups will be disproportionately affected by any potential harms that presence in a hold-out set may bring. As discussed above, non-minimal harms to patients may not be ethically viable even with informed consent gained from patients.

Issues concerning incentives for volunteering arise when considering the principle of justice. Researchers and physicians must take care to not unduly pressure patients to volunteer to be in the hold-out set, as this could lead to systematically over-represented classes of “pressured” patients due to compromised position or availability. Correspondingly, if a group is known to be under-represented in a voluntary hold-out set, this may lead to patients volunteering in order to ensure representation of their own groups, which may be considered an unreasonable incentive.

2.4.5 Informed consent

In the UK, it is not necessarily legally required to obtain informed consent from patients whose electronic health records data are used to train a CPM, even in the absence of any use of a hold-out set [25]. Patients can, however, opt out of having their data used for research [26]. Furthermore, a patient does not need to give consent for a CPM to be used to aid decision making.

However, in a hold-out setting, there is an additional consideration. By withholding risk scores from certain individuals, we potentially deny them use of tools which could positively impact their outcomes. Withholding risk scores is not necessarily in the interests of the individual patient in the hold-out set; rather, it is in the interests of the population as a whole and potentially for the benefit of research. From an ethical perspective, informed consent may thus be necessary to withhold the use of risk scores in this way.

To guarantee an unbiased source of data which can be used to safely retrain prediction models which generalise to the full population, hold-out sets need to be randomly sampled. Seeking consent from those in the hold-out set would be highly likely to cause participation bias [15], and logistically within the NHS infrastructure it may not be feasible. However, if the goal is a more generalisable model, absence of informed consent is necessary. In any CPM-intervention setting, use of hold-out sets must be weighed against any risks to patients which arise as a result of not using a CPM. Therefore, the use of hold-out sets without informed consent must be weighed against the principle of non-maleficence as a priority, and is highly setting dependent. This argument extends to both simple and cluster randomised sampling.

In the case of voluntary response sampling, informed consent would be sought from all those in the hold-out set, but not necessarily those in the intervention set, which is seen as the default. This means that, ideally, patients in the hold-out set would be fully aware of any potential risks or lost benefits resulting from the withholding of risk scores. It should be noted that these patients would have access to the same interventions as those in the intervention set. However, risks to patients in the hold-out set would still need to be minimised, and in some settings this may not be acceptable even with consent. Furthermore, bias from voluntary response sampling may increase risks to individuals in the patient population through the use of less accurate risk scores.

2.4.6 Clinical equipoise

The assumption of clinical equipoise, the ethical principle that there is genuine uncertainty about the best treatment, is a key component of the ethical argument for randomisation in randomised control trials [27]. Our setting is notably different in that use of an accurate CPM is assumed to improve the expected outcome for a patient, usually due to improved prognosis of a disease or allocation of treatments. A hold-out set would be used in one of two scenarios: either risk score informed interventions have been proven to be more effective than otherwise, in which case there is not clinical equipoise; or risk score informed interventions perform no better than interventions without a risk score, in which case there is no longer an argument for use of a CPM in the first place. If clinical equipoise is assumed, this would only be the case on initial deployment of the CPM, until its effectiveness has been established.

The choice of sampling, including whether or not consent is gained, does not necessarily affect these arguments. The argument for use of a hold-out set relies on the intent that it will maintain an optimal allocation of interventions, rather than as a research tool to gain knowledge. This differs from randomised control trials, in which the primary purpose is to gain knowledge, and the medical community is assumed to be in equipoise over the effect of the potential treatment. A clinical trial may or may not improve the healthcare system over a steady baseline defined by existing treatment, whereas the corresponding baseline in the CPM updating setting deteriorates due to drift, and we seek to restore the baseline (but are not in equipoise over whether updating the CPM will do so).

2.4.7 Challenges to shared decision-making:

One key principle in modern medical ethics is that of trust and shared decision-making between patient and clinician. One aspect of this is patient autonomy, covered above [28]. However, in order to achieve this ideal of trust and shared-decision making, it is also necessary that a clinician is truthful with their patient.

When considering the principle of truth telling, it must be noted that how much information a doctor divulges to a patient may differ across cultures [29]. Furthermore, it must be balanced with the principles of beneficence and patient autonomy [30]. Nevertheless, the principle of truth telling is potentially violated by the use of hold-out sets by design, as risk scores are not generated for some patients and therefore the score is withheld from the clinical-patient discussion.

Once again, evaluation of this principle is situation dependent. It should be noted that using simple random sampling in the motivating setting, the physician would still be able to make a decision on the information present, or to evaluate risk of a future outcome without the use of the risk score for the hold-out set patients. Therefore this information may not necessarily be vital to the patient's outcome. However, a patient who is unaware that they are in the hold-out set may reasonably expect to be given a risk score, if they were aware of what this means and that it were possible to generate one for them.

Cluster randomised hold-out sets violate the principle of truth telling in the same way as simple random sampling. Whilst one could argue that patients in the hold-out set who do not receive risk scores would now be receiving the standard care of their cluster, a patient would still reasonably expect that if it were possible to generate a risk score for them, then they would want a physician to make use of this information.

In a voluntary response setting, providing a patient with the ability to withdraw consent at any time would ensure that a patient could always have access to a risk score upon request.

2.5 Case studies

In this section, some potential situation specific ethical issues are discussed through the use of two real-world models: the Scottish Patients at Risk of Readmission and Admission (SPARRA) model [31] and the Epic Sepsis Model (ESM) [32].

SPARRA: This is a model to predict the 1-year risk of emergency admission to hospital based on electronic health records in Scotland. SPARRA risk scores are calculated monthly, and individual-level scores can be accessed by GPs. GPs may choose to act on these scores with high-risk patients to lower their risk of emergency admission, such as by choosing to give an enhanced follow-up in comparison to lower-risk patients. There are no specific interventions that physicians must take in response to risk scores, as the nature of each patient's risk is unique. It has been demonstrated that preventative interventions influenced by the model have the potential to alter future risk scores, resulting in a potential underestimation of risk in high risk patients [12, 31].

ESM: The Epic Sepsis Model [32] is a prediction model which uses electronic health records to predict the risk of sepsis for patients every 15 min. It generates automatic alerts to warn clinicians that a patient may be becoming septic, at which point appropriate interventions can be taken. In the ICU setting, we would expect that interventions may occur in under 15 min, and such interventions may be life-saving; hence, a patient for whom the ESM score leads to an intervention may be observed to have a lower risk of sepsis than predicted by a well-calibrated risk score, leading to performative effects in an updated risk score.

In the case of the SPARRA model, the outcome being predicted is not imminently affecting the patient. Emergency admission could also be caused by a wide variety of factors, and interventions that may be applied are highly patient dependent. A physician and patient are likely to make decisions about interventions based on a wide variety of factors, of which the model predicted risk score is only a small factor. In this case, withholding risk scores from patients is less likely to break the principle of non-maleficence, as any harms due to withholding the risk score are likely to be non-immediate and minimal. Furthermore, the use of hold-out sets generally in this setting would provide additional value to the full patient population by way of more accurate pre-intervention risk scores.

For the ESM one may make the argument in favour of a hold-out set in this scenario, specifically that the harms to patients as a result of inaccurate risk scores would be greater in this setting, due to the nature of the outcome being predicted. However, we argue that this is not enough of a reason to use hold-out sets. In the case of the ESM, the outcome being predicted poses an immediate risk to patients. Furthermore, it is a model used on patients who are already hospitalised, and therefore highly vulnerable. Gaining consent to be included in a hold-out set from such patients would clearly be unethical, as they are in a compromised state. Furthermore, under (or over) estimation of risk presents a much greater risk to patients in the case of ESM in comparison to SPARRA. Withholding risk scores from some patients for the ESM may increase the risk of misdiagnosis, particularly false negatives. This presents severe possible harms to patients in the hold-out set, and breaks the principle of non-maleficence. This can then be viewed as a scenario where the use of hold-out sets, with or without consent, would be a gross violation of widely accepted clinical research ethics.

The two models have been chosen specifically due to the stark contrast in the nature of the predicted outcomes, as well as the risk profiles of patients. This highlights the wide variety of different CPMs that exist in the real world. Any ethical questions which arise due to the potential use of hold-out sets must be contextualised to the model in question, and treated uniquely.

2.6 Hold-out sets to measure the effectiveness of a CPM

So far in this paper we have mainly considered hold-out sets as a tool for updating CPMs, and subsequent ethical arguments have followed from that. There is however, the possibility of using hold-out sets to measure the effectiveness of a CPM compared to standard medical care.

Typically, CPMs are validated in-silico, for example using discrimination or calibration metrics and cross-validation [33]. However, this fails to capture the overall effectiveness of a CPM. Indeed, the latter depends on a variety of factors that go beyond the accuracy of the predictions. Among others, this includes: how well is the CPM integrated into the healthcare system (can clinicians or patient easily access the scores?), how well is the CPM integrated into the decision-making process (if a CPM is available to them, do clinicians and patients take it into account when deciding the best course of action?) and the effectiveness of the possible interventions themselves (does the intervention actually reduce the risk of an adverse event?).

If a hold-out set is a random sample from the population, then generally you may be able to say that any difference in downstream case numbers between the hold-out and intervention set is due to the use of the CPM. This allows for analysis to be done in a similar way to a randomised control trial. For example, one study utilised a randomised stepped wedge trial design to assess the effects (both in terms of patient outcomes and resource utilisation) associated to the introduction of a CPM (to predict the risk of emergency hospital admissions) into the Welsh primary care system [34]. It should be noted that simple random sampling would then offer the most appropriate sampling method for this use-case. Likewise, cluster randomised sampling would have potential to also offer unbiased results, provided that clusters were chosen appropriately. However, voluntary response sampling, as previously discussed, would lead to biased samples. If the intent is to measure the overall effectiveness of a CPM, voluntary response would make this very difficult to interpret.

2.7 Summary and recommendations

In general we expect that the greater the risk to the patient from having no risk score (and hence the more unappealing the hold-out set approach appears), the worse the cost from having an inaccurate risk score (so the greater the incentive for a randomised hold-out set).

The necessity of use of a hold-out set at all is dependent on drift and performative effects. The presence of drift is largely independent of the severity of the outcome predicted by the CPM. Unfortunately, in settings where ignoring a CPM has a severe consequence, performative effects are necessarily present. Recommendations can not be made at a general level, as it depends on the severity of such consequences. However, considerations of the environment in which a CPM operates can help inform choice of sampling method for a hold-out set, or whether to use a hold-out set at all. In particular, the nature of the outcome being predicted by the CPM is an important factor. If the outcome being predicted is one which will not cause immediate serious harm to patients, and the patients themselves are not in a compromised position, then a setting specific argument can be made for use of a hold-out set without consent. Use of cluster randomisation for hold-out sets may offer additional ethical safeguards for patients without sacrificing statistical validity in terms of bias, provided that appropriate clusters are chosen. Specifically, cluster randomisation ensures that patients always receive the same care as others in their cluster. However, care must be taken to select clusters which do not over-sample certain groups, otherwise issues of justice and fairness arise.

Voluntary response hold-out sets, whilst avoiding some of the ethical issues discussed in this paper, offer the potential to indirectly harm patients on the whole through inaccurate risk scores trained on a biased hold-out set. It may be possible to ensure that a voluntary response hold-out set is balanced across some protected classes or important characteristics. However, this does not correct for different rates of volunteering across unobserved patient characteristics.

Other approaches have been suggested for mitigating against the effects of performative prediction in the presence of external drift. In particular, causal modelling approaches have been suggested [14]. Whilst this may be possible, it relies on a more mature data collection system than currently exists in any area of the health care system, in particular the recording of all interventions, and the collection of all variables which exist in the causal structure. Thus, in the absence of this, hold-out sets may be the only viable method of updating risk scores in the presence of performative effects.

Any implementation of hold-out sets will need to take into account the views of key stakeholders, clinicians and patients themselves. In particular, use of hold-out sets without informed consent may risk damaging trust in the health system and lead to unintended consequences, therefore much more research is needed.

3 Discussion

Our work principally suggests that the ethical viability of hold-out sets is setting dependent. Furthermore, certain implementations of hold-out sets may be more appropriate than others. The key advantage of hold-out sets is that they offer the possibility to distribute more accurate pre-intervention risk scores to those in the intervention set. On an aggregate level this may be more optimal than distributing risk scores to all patients, given some objective function to be optimised such as lowering the number of patients with $Y = 1$ over the life-span of the CPM. However from an ethical perspective, this is may not be enough of a reason to use hold-out sets, particularly without informed consent. Withholding risk scores must not cause an unacceptable risk of harm to patients, and must respect principles of patient autonomy and justice. Ultimately, any implementation of hold-out sets in a health setting in the absence of informed consent must respect patients, putting their health and well-being first. Cumulative benefits to the group must be balanced against individual considerations at a patient level, including risk of harm, lack of autonomy or potential justice issues.

It is critical to realise that a choice must be made between not updating a CPM, updating a CPM without a hold-out set, and updating with a hold-out set. All have drawbacks at an individual or population level. Each option violates one or more ethical principles, as is typically the case in research ethics, and violation of principles thus cannot preclude use of a method: risks must be weighed against each other. We do note however that in the absence of recorded interventions in response to risk scores and therefore explicit causal modelling, hold-out sets may be the only viable way of updating in the presence of performative effects.

This paper does not seek to prescribe hold-out sets as a catch-all solution to performative prediction in CPMs, nor which implementations of sampling are universally ethically viable. Rather, it seeks to provide an initial discussion of some of the general issues in this field in the absence of existing literature. In any possible implementation of hold-out sets, a sampling framework should be implemented which offers the lowest possible degree of bias in the hold-out set given that ethical issues have been considered. This paper further highlights the need for robust study protocols and wide involvement from scientists from a range of disciplines, health practitioners and patients themselves. Furthermore, this paper does not seek to analyse the potential for hold-out sets to be practical within the NHS (or other health systems) infrastructure, although this will be a necessary consideration. Without these considerations, hold-out sets as a tool have the potential to lead to gross violations of widely accepted research ethics principles. Context dependent ethical discussion is necessary before use of a hold-out set in a CPM can be considered.

Funding Louis Chislett acknowledges the receipt of studentship awards from the Health Data Research UK-The Alan Turing Institute Wellcome PhD Programme in Health Data Science (Grant Ref: 218529/Z/19/Z). This work was supported by EPSRC Grant EP/X03870X/1 & The Alan Turing Institute'

Data availability Not applicable.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest

References

1. Hastie, T., Friedman, J., Tibshirani, R.: The Elements of Statistical Learning. Springer, New York, NY (2001)
2. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nat. Publ. Group (2019). https://doi.org/10.1038/s41591-018-0300-7
3. Cowley, L.E., Farewell, D.M., Maguire, S., Kemp, A.M.: Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. Diagn. Progn. Res. **3**(1), 16 (2019). https://doi.org/10.1186/s41512-019-0060-y
4. Nashef, S.A.M., Roques, F., Sharples, L.D., Nilsson, J., Smith, C., Goldstone, A.R., Lockowandt, U.: Euroscore II. Eur. J. Cardiothorac. Surg. **41**(4), 734–745 (2012). https://doi.org/10.1093/ejcts/ezs043
5. Žliobaitė, I.: Learning under Concept Drift: An Overview. arXiv preprint (2010). arXiv:1010.4784 [cs.AI]
6. Davis, S.E., Greevy, R.A., Lasko, T.A., Walsh, C.G., Matheny, M.E.: Detection of calibration drift in clinical prediction models to inform model updating. J. Biomed. Inform. **112**, 103611 (2020). https://doi.org/10.1016/j.jbi.2020.103611
7. Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S., Saria, S.: The clinician and dataset shift in artificial intelligence. N. Engl. J. Med. **385**(3), 283–286 (2021). https://doi.org/10.1056/nejmc2104626
8. Perdomo, J.C., Zrnic, T., Mendler-Dünner, C., Hardt, M.: Performative prediction. In: International Conference on Machine Learning (2020)
9. Toll, D.B., Janssen, K.J.M., Vergouwe, Y., Moons, K.G.M.: Validation, updating and impact of clinical prediction rules: a review. J. Clin. Epidemiol. **61**(11), 1085–1094 (2008). https://doi.org/10.1016/j.jclinepi.2008.04.008
10. Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N., Kroeker, K.I.: An overview of clinical decision support systems: benefits, risks, and strategies for success. Nat. Res. (2020). https://doi.org/10.1038/s41746-020-0221-y
11. Hippisley-Cox, J., Coupland, C., Brindle, P.: Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ (2017). https://doi.org/10.1136/bmj.j2099
12. Liley, J., Emerson, S.R., Mateen, B.A., Vallejos, C.A., Aslett, L., Vollmer, S.J.: Model updating after interventions paradoxically introduces bias. In: International Conference on Artificial Intelligence and Statistics, vol. 130 (2021).https://www.who.int/news-room/
13. Lenert, M.C., Matheny, M.E., Walsh, C.G.: Prognostic Models will be Victims of their Own Success, Unless. Oxford University Press, Oxford (2019)
14. Sperrin, M., Jenkins, D., Martin, G.P., Peek, N.: Explicit Causal Reasoning is Needed to Prevent Prognostic Models Being Victims of their Own Success. Oxford University Press, Oxford (2019)
15. Berndt, A.E.: Sampling methods. J. Hum. Lact. **36**(2), 224–226 (2020). https://doi.org/10.1177/0890334420906850
16. Haidar-Wehbe, S., Emerson, S.R., Aslett, L.J.M., Liley, J.: Optimal Sizing of a Holdout Set for Safe Predictive Model Updating. arXiv preprint (2022) https://doi.org/10.48550/arXiv.2202.06374
17. Varkey, B.: Principles of Clinical Ethics and Their Application to Practice. S. Karger AG, Germany (2021)
18. Coughlin, S.S.: How many principles for public health ethics? Open Public Health J. **1**(1), 8–16 (2008). https://doi.org/10.2174/1874944500801010008
19. Summers, J., Morrison, E.: Principles of healthcare ethics. In: Health Care Ethics, 2nd edn., pp. 41–58. Jones and Bartlett Publishers, USA (2009)
20. Guraya, S.Y., London, N.J.M., Guraya, S.S.: Ethics in medical research. J. Microsc. Ultrastruct. **2**(3), 121 (2014). https://doi.org/10.1016/j.jmau.2014.03.003
21. Chen, R.J., Chen, T.Y., Lipkova, J., Wang, J.J., Williamson, D.F.K., Lu, M.Y., Sahai, S., Mahmood, F.: Algorithm Fairness in AI for Medicine and Healthcare. arXiv preprint (2021). https://doi.org/10.48550/arXiv.2110.00603
22. Verheij, R.A., Curcin, V., Delaney, B.C., McGilchrist, M.M.: Possible sources of bias in primary care electronic health record data use and reuse. J. Med. Internet Res. (2018). https://doi.org/10.2196/JMIR.9134
23. Walsh, D., Bendel, N., Jones, R., Hanlon, P.: It's not 'just deprivation': why do equally deprived UK cities experience different health outcomes? Public Health **124**(9), 487–495 (2010). https://doi.org/10.1016/j.puhe.2010.02.006
24. Swanson, J.M.: The UK Biobank and Selection Bias. Elsevier B.V, Amsterdam (2012)
25. Taylor, R.M., Fern, L.A., Aslam, N., Whelan, J.S.: Direct access to potential research participants for a cohort study using a confidentiality waiver included in UK National Health Service legal statutes. BMJ Open (2016). https://doi.org/10.1136/bmjopen-2016-011847
26. NHS: Protecting patient data (2022). https://digital.nhs.uk/services/national-data-opt-out/understanding-the-national-data-opt-out/protecting-patient-data
27. Cook, C., Sheets, C.: Clinical equipoise and personal equipoise: two necessary ingredients for reducing bias in manual therapy

trials. J Man Manip Ther (2011). https://doi.org/10.1179/106698111X12899036752014
28. Gillon, R.: Defending the four principles approach as a good basis for good medical practice and therefore for good medical ethics. Technical Report 1 (2015). https://doi.org/10.1136/medethics-2014-102282
29. Tuckett, A.G.: Truth-telling in clinical practice and the arguments for and against: a review of the literature. Nurs Ethics **11**, 500–513 (2004). https://doi.org/10.1191/0969733004ne728oa
30. Sullivan, R.J., Menapace, L.W., White, R.M.: Truth-telling and patient diagnoses. J. Med. Ethics **27**(3), 192–197 (2001). https://doi.org/10.1136/jme.27.3.192
31. Liley, J., Bohner, G., Emerson, S.R., Mateen, B.A., Borland, K., Carr, D., Heald, S., Oduro, S.D., Ireland, J., Moffat, K., Porteous, R., Riddell, S., Cunningham, N., Holmes, C., Payne, K., Vollmer, S.J., Vallejos, C.A., Aslett, L.J.M.: Development and assessment of a machine learning tool for predicting emergency admission in Scotland. medRxiv (2023). https://doi.org/10.1101/2021.08.06.21261593
32. Wong, A., Otles, E., Donnelly, J.P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., Singh, K.: External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern. Med. **181**(8), 1065–1070 (2021). https://doi.org/10.1001/jamainternmed.2021.2626
33. Staffa, S.J., Zurakowski, D.: Statistical development and validation of clinical prediction models. Anesthesiology **135**(3), 396–405 (2021). https://doi.org/10.1097/ALN.0000000000003871
34. Snooks, H., Bailey-Jones, K., Burge-Jones, D., Dale, J., Davies, J., Evans, B.A., Farr, A., Fitzsimmons, D., Heaven, M., Howson, H., Hutchings, H., John, G., Kingston, M., Lewis, L., Phillips, C., Porter, A., Sewell, B., Warm, D., Watkins, A., Whitman, S., Williams, V., Russell, I.: Effects and costs of implementing predictive risk stratification in primary care: a randomised stepped wedge trial. BMJ Qual. Saf **28**(9), 697–705 (2019). https://doi.org/10.1136/bmjqs-2018-007976

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.