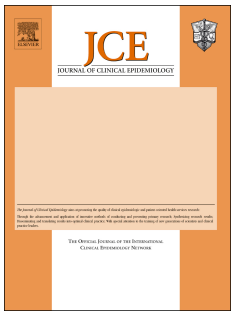
Accepted Manuscript



Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported

Rebecca Whittle, George Peat, John Belcher, Gary S. Collins, Richard D. Riley

PII: S0895-4356(18)30037-4

DOI: 10.1016/j.jclinepi.2018.05.008

Reference: JCE 9655

To appear in: *Journal of Clinical Epidemiology*

Received Date: 17 January 2018

Revised Date: 26 April 2018

Accepted Date: 14 May 2018

Please cite this article as: Whittle R, Peat G, Belcher J, Collins GS, Riley RD, Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported, *Journal of Clinical Epidemiology* (2018), doi: 10.1016/j.jclinepi.2018.05.008.

Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported

Rebecca Whittle[1], George Peat[1], John Belcher[1], Gary S. Collins[2], Richard D. Riley[1]

[1] Centre for Prognosis Research, Arthritis Research UK Primary Care Centre, Research Institute for Primary Care & Health Sciences, Keele University, Keele, Staffordshire, UK (RW: r.l.whittle@keele.ac.uk; GP: g.m.peat@keele.ac.uk; JB: j.belcher@keele.ac.uk; RR: r.riley@keele.ac.uk)

[2] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Oxford, UK (GC: gary.collins@csm.ox.ac.uk)

Corresponding author: Rebecca Whittle, Centre for Prognosis Research, Arthritis Research UK Primary Care Centre, Research Institute for Primary Care & Health Sciences, Keele University, Keele, Staffordshire, UK. Tel: 00 44 1782 734848; Email: r.l.whittle@keele.ac.uk

ABSTRACT

Objective: Measurement error in predictor variables may threaten the validity of clinical prediction models. We sought to evaluate the possible extent of the problem. A secondary objective was to examine whether predictors are measured at the intended moment of model use.

Methods: A systematic search of Medline was used to identify a sample of articles reporting the development of a clinical prediction model published in 2015. After screening according to a predefined inclusion criteria, information on predictors, strategies to control for measurement error and intended moment of model use were extracted. Susceptibility to measurement error for each predictor was classified into low and high risk.

Results: Thirty-three studies were reviewed, including 151 different predictors in the final prediction models. Fifty-one (33.7%) predictors were categorised as high risk of error, however this was not accounted for in the model development. Only 8 (24.2%) studies explicitly stated the intended moment of model use and when the predictors were measured.

Conclusion: Reporting of measurement error and intended moment of model use is poor in prediction model studies. There is a need to identify circumstances where ignoring measurement error in prediction models is consequential and whether accounting for the error will improve the predictions.

Keywords: Prediction models, prediction, prognosis, diagnosis, measurement error, error

Word count: 5416

WHAT IS NEW?

Key findings

- Many published prediction models include predictors that are susceptible to measurement error and this measurement error is not being acknowledged or accounted for in the development of the models.
- Most prediction model articles do not explicitly state the intended moment of model use, or exactly when the predictors used in the model development were measured.

What this adds to what is known

- Reporting of measurement error and intended moment of model use is poor in prediction model studies.

What is the implication, what should change now?

- There is a need to identify circumstances where ignoring measurement error in prediction models is consequential and whether accounting for the error will improve the predictions.
- Future prediction model research studies must clearly report the intended moment of use of the prediction model, and be explicit about when the predictors were measured.

BACKGROUND

Predicting a patient's future outcome risk is an important part of medical research as it guides treatment, informs clinical decision making and helps patients understand their risk. Prognosis research can be used to help predict future outcomes in patients with a particular disease or health condition by developing a prediction model [1]. The number of articles reporting clinical prediction models has been increasing steadily over time, with approximately 500 articles published in 2011 [2], and these models utilise values of multiple predictors to enable individualised risk prediction [3]. Such models are intended "to assist clinicians with their prediction of a patient's future outcome and to enhance informed decision making with the patient" [4]. Therefore, the predictions from these models should have optimal performance when being practically implemented at the "intended moment of using the model" [5].

However, when developing such models, measurement error may affect the observed predictor values, which could potentially lead to biased or incorrect estimates of predictor-outcome associations [6-9]. Measurement error is a difference between the measured values of a predictor and the true values of the predictor, or if the predictor is categorical, it is the classification to an incorrect category (misclassification). The term measurement error will be used throughout this article to refer generally to measurement error in continuous predictors and misclassification of categorical predictors. Measurement error is common within clinical studies, particularly observational studies [10], and has been found to be commonly neglected within the medical literature [11] Measurement error can occur for many different reasons such as biological variability, inaccuracy of measurement instruments, imperfect recall, cost or resource limitations, the subjective nature of measures, laboratory or measurer error and timing error. For example, measurement error in blood pressure commonly occurs due to biological variability [12]. Body mass index (BMI) is also commonly measured with error either due to the inaccuracy of measurement

instruments (i.e. the scales not being calibrated correctly), or due to imperfect recall by the patient, and this measurement error could then cause misclassification into an incorrect category.

Prognosis research is becoming increasingly more important [1], but there has been little research into the impact that measurement error in the predictors used to develop a prediction model may have, both in terms of the predictions made and model performance. It is also unclear how accounting for measurement error within the statistical modelling may improve this. A recent study demonstrated that measurement error in the predictors can dramatically reduce the c-statistic and increase the Brier score [13], and another study found that both random and systematic error in self-reported health data influences the calibration, discrimination and predicted risks [14], but in general the extent and impact of measurement error in prediction model research is often overlooked. However, the STRATOS (STRengthening Analytical Thinking for Observational Studies) initiative (www.stratos-initiative.org) have identified measurement error as a common issue in observational studies which is often ignored and for which guidance is needed. There is a vast amount of literature on the statistical effect of measurement error in general, but whether investigators consider measurement error when developing a prediction model, has not previously been evaluated. Models developed with predictors containing measurement error could therefore provide inaccurate estimates of patient risk and the model may not perform as well as expected in practice. A summary of the most commonly used methods to correct for measurement error is given by Brakenhoff et al [11] with more detailed reviews of these (and other) methods given by Caroll et al. [8] and Gustafson [9]. Several other methods that can be used to account for measurement error in the particular context of prediction research have been developed, including methods in a Bayesian framework, using an item response theory model to handle the measurement error [15] and bootstrap regression calibration [16], based on resampling techniques.

A particular aspect of measurement error in the predictors is timing error, so whether the predictors used in the model development were measured at the moment the model is intended to be used in

practice. When time-dependent predictors are not able to be measured at `baseline' this creates time-dependent bias, which has been shown to often have an impact on the estimates of key predictors and study conclusions [17]. Additionally, the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement recommends to clearly define when the predictors used in the development of the model were measured [18] and states that "all predictors should be measured before or at the study time origin and known at the intended moment the model is intended to be used" [19]. Nevertheless, for a range of practical and ethical reasons, researchers may design prognosis studies that collect time-varying predictor information after the intended moment of use, which itself may lead to errors and misleading predictions [20].

The aim of this article is to present a systematic review of recent studies developing prediction models, to ascertain how susceptible to measurement error the predictors used in the final models are and how often the measurement error was acknowledged or accounted for within the development of the models. A secondary objective is to determine whether the predictors were measured at a different time point to the intended moment of using the prediction model.

METHODS

Data source and search

A systematic search was carried out in Medline on 27th November 2015 to identify the 30 most recent articles reporting the development of a multivariable prediction model for either individualised diagnosis or prognosis. It was decided a priori that approximately 30 articles would be sufficient in providing qualitative saturation of whether measurement error and incorrect timings was a general concern for the prediction model field.

The search strategy used was an adaptation of a published search string for finding prognostic and diagnostic prediction studies in Medline [21]. The search string was adapted by changing the term "OR 'Multivariable'" to "AND (Multivariable OR Multivariate)" to refine the search further to studies developing multivariable prediction models for individualised prediction, which would hopefully remove other studies just examining associations between specific factors and an outcome but not developing a prediction model (see Supplementary Table 1 for the full search string).

Selection of articles

The titles and abstracts of the 1000 most recently published articles found using the search string were screened for inclusion, as we estimated this would return approximately 30 articles to be included. The full article was then obtained for any articles which were deemed to be potentially eligible or for any articles in which it was unclear from the title and abstract whether they met the eligibility criteria. These full articles were then screened for suitability and categorised into one of three groups: 'include', 'exclude' and 'unsure'. The selection of articles until this stage was undertaken by a single reviewer (RW). Articles in the 'include' and the 'unsure' groups were sent to two additional reviewers (GP & RR). Both reviewers checked all 'unsure' articles, and the 'include' articles were split between the two reviewers to check they met the eligibility criteria. Any 'unsure' articles on which an agreement could not be reached were checked by a fourth reviewer (JB) and the decision to include or exclude was based on the verdict of the fourth reviewer.

Inclusion/exclusion criteria

Articles were included if they reported the development of a clinical prediction (prognostic or diagnostic) model for individualised prediction in human participants, based on a multivariable regression model or studies updating a previously developed prediction model by adding new

153 predictors. Articles were excluded if they developed a model using non-regression based techniques,
154 validated a previously developed prediction model, created a risk score from an existing prediction
155 model, used a multivariable model to examine whether a particular predictor is associated with the
156 outcome when adjusting for other factors (prognostic factor research), estimated the prognostic
157 effect (e.g. hazard ratio) of a previously developed score, updated a previously developed model
158 without adding any new predictors to the model or investigated the optimal cut off value of a
159 previously developed model. Any articles in excess of the required 30 that met the eligibility criteria
160 were also retained for inclusion, to avoid any potential selection bias concerns when choosing which
161 articles to remove.

162

163 Data extraction

164 Data were extracted from the selected articles by a single reviewer (RW). The list of items presented
165 in Table 1, where available, were extracted from each article and were based on the CHARMS
166 checklist [5], with the addition of information related to the intended moment of using the model
167 and measurement error.

168

>> insert Table 1: Data extracted from each article<<

Measurement error

The level of susceptibility to measurement error for each predictor used in the final models of the included articles were classified into two categories:

- Low risk: Unlikely to be measured with error, or possibly/likely to be measured with error but expected to be unimportant;
- High risk: Possibly/likely to be measured with error and may be important.

For example, age and gender are both extremely unlikely to be measured with error, and any error in age recorded would be expected to be negligible. Thus, age and gender would be classed as 'low risk' with regards to important measurement error. Whereas, blood pressure could be measured with error, as error in blood pressure measurement commonly occurs because of improper techniques such as talking during measurement or wrong cuff size [22] and blood pressure is also commonly measured with error due to biological variability [12]. This error could be large, and could be important when developing a prediction model for hypertension, for example, because blood pressure is an important component of the diagnostic evaluation for hypertension. Hence, blood pressure would be classed as 'high risk' of measurement error. Another high risk example would be body mass index (BMI), which the extent of the measurement error would depend on the way in which it was measured, but there would be a high chance it would be measured with some error.

To categorise the list of predictors into the two groups of susceptibility to measurement error, first the literature was searched for any publications discussing measurement error in any of the predictors of interest. For those where no evidence could be found, the categorisations were made based on the judgement of the reviewer (RW), which was corroborated by a postdoctoral academic General Practitioner.

Timing of measurements and intended moment of model use

If the timing of predictor measurements and intended moment of using the model was not explicitly stated in the article then, wherever possible, information on where the predictor information came from and the setting they were measured in were used to establish a likely time of measurement. If the intended moment of use of the model was not stated then, again where possible, information on what the model would be used for and the predictors that would be used within the model were considered to make a decision on the most probable intended moment of use.

RESULTS

Included studies

A total of 1000 titles and abstracts were extracted and screened for inclusion. Of these, 876 were excluded based on not meeting the inclusion criteria from screening their title and abstract (Figure 1). Study eligibility was then assessed for 124 full-text articles and 33 met eligibility criteria for inclusion [23-55], hence all 33 were retained for the review. The 33 included articles consisted of 27 prognostic model studies and 6 diagnostic model studies published in 2015. The additional information extracted from the articles that is not related to measurement error or the timing of measurements is presented in Supplementary Table 3.

>>insert Figure 1: Flow chart of included studies<<

Measurement error

In the 33 articles reviewed, there was a total of 151 different predictors in the final prediction models. Many of the predictors were included in several different models, for example, age and gender were in many of the models (13 models and 5 models, respectively). Of the reported predictors included in the final models, we categorised 51 (33.8%) as high risk of being susceptible to measurement error (Table 2) and the remaining 100 (66.2%) as low risk (Supplementary Table 2).

>> insert Table 2: Predictors in final models at high risk of measurement error<<

Despite a third of the included predictors being at high risk of being susceptible to measurement error, only three studies acknowledged, or accounted for, measurement error within their model development. One study mentioned measurement error as a general limitation due to the study being from a single centre [23], but did not specify any particular predictor which may be at risk of error. Additionally, two studies used repeated measurements of a predictor within the modelling process [24, 27]. The first of the studies that used repeated measures [24] used generalised estimating equations (GEE's) to fit models accounting for the correlations among multiple biopsies that were performed on the same patients. The authors state that GEE's yield the same mean predictions as maximum likelihood, but result in inflated standard errors, wider confidence intervals and diminished statistical significance that more accurately reflect the amount of uncertainty in the data. There is no mention of measurement error within the article, and so it assumed that the authors have not made use of the repeated measures in a conscious effort to reduce measurement error, but to take advantage of all the data available (which may consequently potentially minimise measurement error). The second study using repeated measures [27] used joint modelling of longitudinal measures of CA125 with the stated purpose of estimating the time trend of CA125 rather than explicitly accounting for measurement error (although again, this may consequently account for measurement error).

Despite only two of the reviewed articles included repeated measurements in the modelling process, repeated measurements of at least one of the candidate predictors were actually reported to be available in 6 (18.2%) of the articles [24, 27, 30, 31, 37, 45]. Of these 6 articles that repeated measures were available, 4 of these had repeated measures that we categorised as being high risk [27, 30, 31, 45], one of which used the repeated measures within the modelling [27].

Predictors considered at high risk of measurement error have been grouped by key reasons for being susceptible to error in Table 2, and several examples of these predictors with more detail and related references are given in Table 3.

>> insert Table 3: Key reasons for measurement error in examples of predictors at high risk of being susceptible to error<<

One example of a predictor at high risk of error and used in several of the final prediction models is prostate specific antigen (PSA). Roehrborn et al. [69] conclude that there is significant variability between two serum PSA measurements obtained within a short time interval, which is due to chance alone. Biomarkers such as CA125, creatinine, C-reactive protein, serum albumin and other serological markers are also likely to change if a second sample was assessed, meaning they are measured with error due to biological variability causing discrepancies away from an underlying (mean) value [71]. There is also the possibility of laboratory error being present in these biomarkers, as the equipment or methods used to take the measurements within the laboratory may not be accurate.

In another example of a predictor likely measured with error, Ali et al. [68] found that the depth of myometrial invasion (DMI) was different in 29% of cases when the DMI was reassessed. The area under a patients pain curve could also be measured with error as it is a subjective measurement that may be affected by various things including how the question is asked, the setting in which the question is asked or when the question is asked. It could also be subject to recall error if the patient is asked about previous days pain levels. Another example is pulse rate, where Kobayashi [63] found that error occurred when pulse rates were objectively scored for various durations (e.g. 10, 15 or 30 seconds) rather than for a whole minute, so the error in a pulse rate could depend on how long the pulse was taken for.

A patient's primary tumour diameter is another example of a predictor which may be susceptible to being measured with error. If a histologist determined the diameter under a microscope there would be little deviation from the true value, whereas if a surgeon recorded the diameter using an endoscopy then this could be recorded with error and could have an effect on the therapy chosen to be used [72]. Another example is BMI, which again the amount of error would depend on how it was

measured. If measured by a clinician then there is unlikely to be much measurement error, but if measured by the patient and recalled this may be subject to error [73]. Other examples include duration of convulsions, duration of neck pain, duration on nervousness and duration of tingling which are all self-reported predictors which could be subject to imperfect recall by the patient.

There were also many examples of predictors that were considered to be at low risk of important error. For example, one model that aimed to identify trauma patients at high risk of pulmonary embolism included a predictor indicating if the patient arrived at the hospital by helicopter [26], and it would be unlikely this would be incorrectly classified. Other models included the patient's disease location as a predictor and again, it is unlikely that this would not be recorded correctly.

Timing of measurements and relation to intended moment of model use

Only eight of the articles explicitly stated exactly when the intended moment of using the model would be, or exactly when the predictors used in the final model were measured. However, for the majority of the 33 included articles it was possible to make a reasonable assumption about these details. If these assumptions were indeed correct, then in 30 (90.9%) of the 33 articles, the predictor measurements were all either taken at the intended moment of model use or were available prior to this. For example, one study [42] developed a model to predict survival prognosis after surgery in patients with symptomatic metastatic spinal cord compression from non-small cell lung cancer, with the aim of being able to provide optimal treatment. Although the specific timing of the predictor measurements was not stated, the predictors were specified as preoperative characteristics. The assumption was made that the model would be intended to be used at the point when a treatment decision was being made, as it was reported that those with the most favourable survival prognosis may instead be treated with more radical surgery. Therefore, it was assumed that the preoperative characteristics considered as predictors were either measured prior to or at the point that the model would be intended to be used.

In another example [40], a diagnostic model was developed to predict colorectal cancer in patients selected for colonoscopy in a primary health care setting, with the aim of identifying high risk patients to reduce the time till diagnosis and hence provide more efficient treatment strategies and success. As the model is to be used to help identify high risk patients when being considered for colonoscopy, which would happen during a GP consultation, it was assumed that the model would be intended to be used during a GP consultation when considering referral for colonoscopy. The model used predictors recorded in routine care data, which would all be available at the point of care, and although the article did not state at which time the predictors were recorded, it was assumed that only measurements recorded prior to colonoscopy referral were considered in the model development.

In all 6 of the articles in which repeated measures were available, each of the repeated measures were recorded either at or prior to the intended moment of using the prediction model.

In two (6.1%) of the articles [32, 48] it was not possible to make an assumption with regards to when the predictors were measured in relation to when the model was intended to be used. In the first article [32], a prognostic model was developed to predict the specific risk of non-sentinel node metastases in women with breast cancer with the aim of preventing unnecessary axillary lymph node dissections. The model was intended to be used after diagnosis of breast cancer, and as it is to be used to prevent unnecessary axillary lymph node dissections it could be assumed that the model would be intended to be used when deciding whether to perform an intraoperative axillary lymph node dissection. Little information was given on the predictors used in the model meaning the timing of the measurements of the predictors could not be deciphered, hence it was not possible to determine whether the predictors were measured at the intended moment of using the model or not. In the second article [48], a model was developed to predict unfavourable disease in patients with prostate cancer. The aim of the model was to avoid or postpone interventions in subjects with prostate cancer of low biological potential. The article states that the model is intended to be used

in patients after radical prostatectomy, but who were eligible for active surveillance. The predictors included were recorded from clinical evaluation, prostatic biopsy and radical prostatectomy specimens, but the timing of the clinical evaluation and prostatic biopsy was unclear and hence it was unknown whether these were before, at, or after the intended moment of using the model.

For one of the included articles [30], a classification algorithm was developed for the diagnosis of non-alcoholic fatty liver disease (NAFLD). The model was not developed to be intended to be used at a specific time but to be used to identify large scale longitudinal cohorts from electronic medical records for use in research studies.

DISCUSSION

Our review suggests that many published clinical prediction models include predictors that are susceptible to potentially important measurement error and yet this was seldom acknowledged. Of 33 articles in our review only two used methods that could potentially account for measurement error by using repeated measurements in the modelling. Though the impact of ignoring measurement error in the articles reviewed is difficult to establish, it raises an important methodological consideration for future prediction model research to address, particularly as a third of the predictors used in the prediction models were categorised as being at high risk of being susceptible to measurement error. The review also found that over three-quarters of the articles included did not explicitly state the exact timing that the model is intended to be used in clinical practice, or exactly when the predictors used in the modelling development were measured. However, a reasonable assumption could be made for the majority of the articles included and, based on this, there were no articles that obviously recorded a predictor after the time it was intended to be used.

Related research

Measurement error has been found to generally have three main effects if not accounted for in medical research: biased or inaccurate estimates of the parameters, loss of power and masking the features of the data (making it harder to spot relationships via graphical methods) [8]. The direction and magnitude of bias from measurement error depends heavily on whether the distribution of errors for one variable depends on the actual value of the variable, the actual values of other variables, or the errors in measuring other variables [7], as well as on the true strength of association, the prevalence of the predictors [74] and whether the errors are random or systematic. Hence, the direction of bias from predictor measurement error is likely to be difficult to predict. However, failing to adjust for random measurement error could potentially lead to estimates being biased towards the null [6], which could subsequently lead to an underestimate of a patients' probability of outcome if measurement error is present in the prediction model used. Conversely, failing to account for systematic errors may change the results in different directions, which could again lead to incorrect predictions of a patients' probability of future outcome.

There are currently two conflicting views about whether measurement error in prediction models is an issue or not. Firstly, Carroll and colleagues [8] state that if a predictor (*X*) is measured with error, and this measure (*W*) is used to predict a patients outcome, then if it is this same surrogate measure of *X* that will be used when applying the prediction model in practice, there is little issue with using *W* to develop the prediction model. On the other hand, a prediction model should provide the most accurate estimate possible, and if a predictor used in the development of a model is measured with error then the estimates of the predictor-outcome associations will be biased, meaning the predictions made may be untrue. Measurement error in the candidate predictors could also lead to certain predictors not being included in the final model due to the measurement error.

In etiologic research we are most interested in the (adjusted) estimate of a single predictor-outcome association and hence would want to minimise bias of this particular estimate. Whereas when

developing a prediction model, we are not predominantly interested in the individual estimates of one (or more) of the predictor-outcome associations, but in the actual absolute risk predictions calculated from the model (and the predictive performance of these risk predictions from the model). Hence, even if one (or more) of the estimates of a predictor-outcome association in a prediction model is biased due to measurement error, this may not be an issue if the model as a whole performs well in terms of the absolute risk predictions. However, measurement error in prediction models has been shown to reduce the c-statistic and increase the Brier score dramatically [13], but in that article the authors focussed on the gain in prediction performance from using error-free predictors instead of error-prone predictors, rather than the gain in prediction performance from accounting for the measurement error in the model when the true error-free values are not known. The article also only evaluated the scenario where only one error-prone predictor was included in the prediction model.

Another article assessed the impact of random and systematic error in self-reported height and weight on the performance of a model used to predict diabetes [14]. The authors found that random error reduced the calibration and discrimination, and biased the predicted risk upwards, whereas systematic error reduced the calibration and biased the predicted risk in the direction of the bias, but had no effect on the discrimination.

Strengths and Limitations

A strength of this review was that a clearly defined search strategy which was based on a previously published search filter [21] was used. Although this review did not include a search of every prediction model published within a certain time period due to the sheer volume of prediction models published each year [2], a search of a few of the most recently published studies was deemed appropriate to enable a general overview of the current literature and provide qualitative saturation of whether there was a susceptibility for measurement error within the predictors and

whether this was considered and also the timing of the predictor measurements in relation to the intended moment of using the model.

The reviewer's judgement had to be used and assumptions were made about the timing of measurements and when the model is intended to be used. This was due to the reviewed articles not explicitly stating these details. Based on this, all of the papers here did actually measure the predictors at the intended moment of using the model (or before), in those that it was possible to decipher this information. However, it is possible that some of these assumptions made were incorrect.

Another concern within prediction models in relation to predictor timing is the relevant time window, or the length of the induction period, in which the predictor of interest is causally related to the outcome. For some prediction models, certain causal factors may need to be considered from much longer ago than others, i.e. with a longer induction period. For example, if considering asbestos exposure in relation to future lung disease, the association could span back many years, whereas recent asbestos exposure may not be related to the outcome if the induction period is only relatively short, e.g. 1-2 years. On the other hand, when predicting infectious diseases, the current and recent exposure of the patient is likely to be most important, and so a relatively short induction period would be needed. Hence, the duration of follow-up of predictors prior to the intended moment of model use should be clearly specified when developing a prediction model, however we did not assess this within this review.

When developing a prediction model, the calendar year of time in which the measurements were made is important (relative to the calendar time of the intended moment of model use), because the precision of measurements often improves when using newer measurement methods. Using a more recent, up-to-date data set that used more improved measurement techniques to develop a prediction model would potentially provide a more relevant and better performing model than if

using an older dataset. While study recruitment dates are generally reported, we did not consider this in relation to when the article was published or would be intended to be used.

Due to many of the included studies not actually stating a complete list of all of the candidate predictors considered in the model development, only the predictors included in the final models were assessed for their susceptibility to measurement error. However, measurement error in the candidate predictors could lead to the exclusion of these predictors in the model development stage and so measurement error in these predictors could be as equally as important as measurement error in the predictors in the final models.

Little information was given within the included articles about any measurement error that may be present in the predictors. Without the availability of previous research on the amount of error in certain predictors, a subjective decision on whether measurement error was likely had to be made by the reviewer, although an academic GP also reviewed the list of predictors and gave their opinion on whether they would judge the predictor to be susceptible to measurement error when using in practice. One difficulty with making a decision on whether the predictor is likely to be susceptible to measurement error was that for many of the predictors it would depend on exactly how the predictor was measured, but often this level of detail is missing from the article. Despite this subjective approach to categorising measurement error, there were several predictors included in the final models that had corresponding published research suggesting they are likely to be measured with error, and this was not considered within the development of the models.

CONCLUSIONS

It is possible that many published prediction models include predictors that are measured with error, and this is often not accounted for or even considered. Additionally, even if the authors considered the predictors to be measured without error, either because of the way they were measured, or for

some other reason, this was still not stated within the articles. This suggests a need to assess under what circumstances ignoring measurement error in prediction models is a concern and whether accounting for the error will improve the predictions made and the model performance. However, researchers should be considering how susceptible to measurement error their predictors may be when developing a model.

Although there were no clear examples within this review of a prediction model being developed using a predictor that was measured after the intended moment of using the model, it is common in prognosis studies of recurrent and long-term conditions presenting to primary care for information on predictors (e.g. pain intensity) to be ascertained by mailed self-complete questionnaires, or personal interview and examination in research clinics several days after their index consultation [75-81]. It was found in this review that the timing of the measurements and the intended moment of using the model is often not explicitly stated, which could mean that future users of the model unknowingly estimate misleading probabilities of a patients' outcome if they are using predictors measured at a different time than those used in the model development in relation to the timing of the model use. We have previously found that displacing the collection of time-varying predictors from the intended moment of use of a prediction model can result in differences in the magnitude of predictor-outcome associations and the subsequent accuracy of the model performance [20]. Hence, future prediction model research studies must clearly report the intended moment of use of the prediction model, and be explicit about whether the predictors were collected before the intended moment of use or not, and if not, justify why.

LIST OF ABBREVIATIONS

AUC=area under the curve; BMI=body mass index; DMI=depth of myometrial invasion; GP=general practitioner; NAFLD=non-alcoholic fatty liver disease; PSA=prostate specific antigen; TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis.

DECLARATIONS OF INTEREST: None.

FUNDING: RW is funded through an NIHR Research Professorship in General Practice for Christian Mallen (NIHR-RP-2014-04-026). GSC was supported by the NIHR Biomedical Research Centre, Oxford. The views expressed in this paper are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health.

AUTHORS' CONTRIBUTIONS: RW, RR, GP and JB developed the research question. RW reviewed and extracted information from the articles, with checks from RR, GP and JB. RW drafted the first version of the manuscript, with support from RR. All authors reviewed the draft manuscript, helped to revise the manuscript, and approved the final version.

ACKNOWLEDGEMENTS: The authors would like to thank Dr Lorna Clarson for assessing the list of predictors for susceptibility to measurement error.

REFERENCES

1. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, Briggs A, Udumyan R, Moons KGM, Steyerberg EW *et al*: **Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes**. *BMJ* 2013, **346**.
2. Wessler BS, Lana Lai YH, Kramer W, Cangelosi M, Raman G, Lutz JS, Kent DM: **Clinical Prediction Models for Cardiovascular Disease: The Tufts PACE CPM Database**. *Circulation Cardiovascular quality and outcomes* 2015, **8**(4):368-375.
3. Steyerberg E: **Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating**: Springer New York; 2008.
4. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG: **Prognosis Research Strategy (PROGRESS) 3: prognostic model research**. *PLoS medicine* 2013, **10**(2):e1001381.
5. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, Reitsma JB, Collins GS: **Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist**. *PLoS medicine* 2014, **11**(10):e1001744.
6. Prentice RL: **Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model**. *Biometrika* 1982, **69**(2):331-342.
7. Rothman KJ, Greenland S, Lash TL: **Modern Epidemiology**: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
8. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM: **Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition**: CRC Press; 2006.
9. Gustafson P: **Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments**: CRC Press; 2003.
10. Guolo A: **Robust techniques for measurement error correction: a review**. *Statistical methods in medical research* 2008, **17**(6):555-580.
11. Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M: **Measurement error is often neglected in medical literature: a systematic review**. *J Clin Epidemiol* 2018.
12. Grassi G, Bombelli M, Brambilla G, Trevano FQ, Dell'Oro R, Mancia G: **Total Cardiovascular Risk, Blood Pressure Variability and Adrenergic Overdrive in Hypertension: Evidence, Mechanisms and Clinical Implications**. *Current Hypertension Reports* 2012, **14**(4):333-338.
13. Khudyakov P, Gorfine M, Zucker D, Spiegelman D: **The impact of covariate measurement error on risk prediction**. *Statistics in medicine* 2015, **34**(15):2353-2367.
14. Rosella LC, Corey P, Stukel TA, Mustard C, Hux J, Manuel DG: **The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model**. *Population Health Metrics* 2012, **10**(1):20.
15. Fox J-P, Glas CAW: **Bayesian modeling of measurement error in predictor variables using item response theory**. *Psychometrika* 2003, **68**(2):169-191.
16. Li W, Mazumdar S, Arena VC, Sussman N: **A resampling approach for adjustment in prediction models for covariate measurement error**. *Computer Methods and Programs in Biomedicine* 2005, **77**(3):199-207.
17. van Walraven C, Davis D, Forster AJ, Wells GA: **Time-dependent bias was common in survival analyses published in leading clinical journals**. *J Clin Epidemiol* 2004, **57**(7):672-682.
18. Collins GS, Reitsma JB, Altman DG, Moons KGM: **Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement**. *Journal of Clinical Epidemiology*, **68**(2):112-121.
19. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS: **Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement: Explanation and Elaboration**. *Annals of Internal Medicine* 2015, **162**(1):W1-W73.

20. Whittle R, Royle K-L, Jordan KP, Riley RD, Mallen CD, Peat G: **Prognosis research ideally should measure time-varying predictors at their intended moment of use**. *Diagnostic and Prognostic Research* 2017, **1**(1):1.
21. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG: **Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews**. *PloS one* 2012, **7**(2):e32844.
22. Handler J: **The Importance of Accurate Blood Pressure Measurement**. *The Permanente Journal* 2009, **13**(3):51-54.
23. Angioli R, Capriglione S, Aloisi A, Ricciardi R, Scaletta G, Lopez S, Miranda A, Di Pinto A, Terranova C, Plotti F: **A Predictive Score for Secondary Cytoreductive Surgery in Recurrent Ovarian Cancer (SeC-Score): A Single-Centre, Controlled Study for Preoperative Patient Selection**. *Annals Of Surgical Oncology* 2015, **22**(13):4217-4223.
24. Ankerst DP, Xia J, Thompson IM, Jr., Hoefler J, Newcomb LF, Brooks JD, Carroll PR, Ellis WJ, Gleave ME, Lance RS *et al*: **Precision Medicine in Active Surveillance for Prostate Cancer: Development of the Canary-Early Detection Research Network Active Surveillance Biopsy Risk Calculator**. *European Urology* 2015, **68**(6):1083-1088.
25. Bendifallah S, Canlorbe G, Laas E, Huguet F, Coutant C, Hudry D, Graesslin O, Raimond E, Touboul C, Collinet P *et al*: **A Predictive Model Using Histopathologic Characteristics of Early-Stage Type 1 Endometrial Cancer to Identify Patients at High Risk for Lymph Node Metastasis**. *Annals Of Surgical Oncology* 2015, **22**(13):4224-4232.
26. Black SR, Howard JT, Chin PC, Starr AJ: **Toward a More Robust Prediction of Pulmonary Embolism in Trauma Patients. A Risk Assessment Model Based Upon 38,000 Patients**. *Journal Of Orthopaedic Trauma* 2015.
27. Chang C, Chiang AJ, Chen W-A, Chang H-W, Chen J: **A joint model based on longitudinal CA125 in ovarian cancer to predict recurrence**. *Biomarkers In Medicine* 2015.
28. Chen J-Y, Chen J-J, Xue J-Y, Chen Y, Liu G-Y, Han Q-X, Yang W-T, Shen Z-Z, Shao Z-M, Wu J: **Predicting Non-sentinel Lymph Node Metastasis in a Chinese Breast Cancer Population with 1-2 Positive Sentinel Nodes: Development and Assessment of a New Predictive Nomogram**. *World Journal Of Surgery* 2015, **39**(12):2919-2927.
29. Cohen MH, Hotton AL, Hershow RC, Levine A, Bacchetti P, Golub ET, Anastos K, Young M, Gustafson D, Weber KM: **Gender-Related Risk Factors Improve Mortality Predictive Ability of VACS Index Among HIV-Infected Women**. *Journal Of Acquired Immune Deficiency Syndromes (1999)* 2015, **70**(5):538-544.
30. Corey KE, Kartoun U, Zheng H, Shaw SY: **Development and Validation of an Algorithm to Identify Nonalcoholic Fatty Liver Disease in the Electronic Medical Record**. *Digestive Diseases And Sciences* 2015.
31. Coté GA, Lynch S, Easler JJ, Keen A, Vassell PA, Sherman S, Hui S, Xu H: **Development and Validation of a Prediction Model for Admission After Endoscopic Retrograde Cholangiopancreatography**. *Clinical Gastroenterology And Hepatology: The Official Clinical Practice Journal Of The American Gastroenterological Association* 2015, **13**(13):2323-2332.e2329.
32. Di Filippo F, Giannarelli D, Bouteille C, Bernet L, Cano R, Cunnick G, Sapino A: **Elaboration of a nomogram to predict non sentinel node status in breast cancer patients with positive sentinel node, intra-operatively assessed with one step nucleic acid amplification method**. *Journal Of Experimental & Clinical Cancer Research: CR* 2015, **34**(1):136-136.
33. Du X-J, Tang L-L, Chen L, Mao Y-P, Guo R, Liu X, Sun Y, Zeng M-S, Kang T-B, Shao J-Y *et al*: **Neoadjuvant chemotherapy in locally advanced nasopharyngeal carcinoma: Defining high-risk patients who may benefit before concurrent chemotherapy combined with intensity-modulated radiotherapy**. *Scientific Reports* 2015, **5**:16664-16664.

34. Dua A, Algodi MM, Furlough C, Ray H, Desai SS: **Development of a scoring system to estimate mortality in abdominal aortic aneurysms management**. *Vascular* 2015, **23**(6):586-591.
35. Englum BR, Pura J, Reed SD, Roman SA, Sosa JA, Scheri RP: **A Bedside Risk Calculator to Preoperatively Distinguish Follicular Thyroid Carcinoma from Follicular Variant of Papillary Thyroid Carcinoma**. *World Journal Of Surgery* 2015, **39**(12):2928-2934.
36. Faget C, Taourel P, Charbit J, Ruyer A, Alili C, Molinari N, Millet I: **Value of CT to predict surgically important bowel and/or mesenteric injury in blunt trauma: performance of a preliminary scoring system**. *European Radiology* 2015, **25**(12):3620-3628.
37. Horn SD, Barrett RS, Fife CE, Thomson B: **A Predictive Model for Pressure Ulcer Outcome: The Wound Healing Index**. *Advances In Skin & Wound Care* 2015, **28**(12):560-572.
38. Kaymakcalan MD, Xie W, Albiges L, North SA, Kollmannsberger CK, Smoragiewicz M, Kroeger N, Wells JC, Rha S-Y, Lee JL *et al*: **Risk factors and model for predicting toxicity-related treatment discontinuation in patients with metastatic renal cell carcinoma treated with vascular endothelial growth factor-targeted therapy: Results from the International Metastatic Renal Cell Carcinoma Database Consortium**. *Cancer* 2015.
39. Koller L, Rothgerber D-J, Sulzgruber P, El-Hamid F, Förster S, Wojta J, Goliasch G, Maurer G, Niessner A: **History of previous bleeding and C-reactive protein improve assessment of bleeding risk in elderly patients (≥ 80 years) with myocardial infarction**. *Thrombosis And Haemostasis* 2015, **114**(5):1085-1091.
40. Koning NR, Moons LMG, Büchner FL, Helsper CW, Ten Teije A, Numans ME: **Identification of patients at risk for colorectal cancer in primary care: an explorative study with routine healthcare data**. *European Journal Of Gastroenterology & Hepatology* 2015, **27**(12):1443-1448.
41. Kusamura S, Torres Mesa PA, Cabras A, Baratti D, Deraco M: **The Role of Ki-67 and Pre-cytoreduction Parameters in Selecting Diffuse Malignant Peritoneal Mesothelioma (DMPM) Patients for Cytoreductive Surgery (CRS) and Hyperthermic Intraperitoneal Chemotherapy (HIPEC)**. *Annals Of Surgical Oncology* 2015.
42. Lei M, Liu Y, Tang C, Yang S, Liu S, Zhou S: **Prediction of survival prognosis after surgery in patients with symptomatic metastatic spinal cord compression from non-small cell lung cancer**. *BMC Cancer* 2015, **15**(1):853-853.
43. Matsuo K, Jung CE, Hom MS, Gualtieri MR, Randazzo SC, Kanao H, Yessaian AA, Roman LD: **Predictive Factor of Conversion to Laparotomy in Minimally Invasive Surgical Staging for Endometrial Cancer**. *International Journal Of Gynecological Cancer: Official Journal Of The International Gynecological Cancer Society* 2015.
44. Nykanen DG, Forbes TJ, Du W, Divekar AA, Reeves JH, Hagler DJ, Fagan TE, Pedra CA, Fleming GA, Khan DM *et al*: **CRISP: Catheterization RISk score for pediatrics: A Report from the Congenital Cardiac Interventional Study Consortium (CCISC)**. *Catheterization And Cardiovascular Interventions: Official Journal Of The Society For Cardiac Angiography & Interventions* 2015.
45. Olmedilla L, Lisbona CJ, Pérez-Peña JM, López-Baena JA, Garutti I, Salcedo M, Sanz J, Tisner M, Asencio JM, Fernández-Quero L *et al*: **Early Measurement of Indocyanine Green Clearance Accurately Predicts Short-Term Outcomes After Liver Transplantation**. *Transplantation* 2015.
46. Resch JE, Brown CN, Macciocchi SN, Cullum CM, Blueitt D, Ferrara MS: **A Preliminary Formula to Predict Timing of Symptom Resolution for Collegiate Athletes Diagnosed With Sport Concussion**. *Journal Of Athletic Training* 2015.
47. Rosenkrantz AB, Ream JM, Nolan P, Rusinek H, Deng F-M, Taneja SS: **Prostate Cancer: Utility of Whole-Lesion Apparent Diffusion Coefficient Metrics for Prediction of Biochemical Recurrence After Radical Prostatectomy**. *AJR American Journal Of Roentgenology* 2015, **205**(6):1208-1214.

48. Russo GI, Castelli T, Favilla V, Reale G, Urzì D, Privitera S, Fragalà E, Cimino S, Morgia G: **Performance of biopsy factors in predicting unfavorable disease in patients eligible for active surveillance according to the PRIAS criteria**. *Prostate Cancer And Prostatic Diseases* 2015, **18**(4):338-342.
49. Shaikh AY, Esa N, Martin-Doyle W, Kinno M, Nieto I, Floyd KC, Browning C, Ennis C, Donahue JK, Rosenthal LS *et al*: **Addition of B-Type Natriuretic Peptide to Existing Clinical Risk Scores Enhances Identification of Patients at Risk for Atrial Fibrillation Recurrence After Pulmonary Vein Isolation**. *Critical Pathways In Cardiology* 2015, **14**(4):157-165.
50. Siegel CA, Horton H, Siegel LS, Thompson KD, Mackenzie T, Stewart SK, Rice PW, Stempak JM, Dezfoli S, Haritunians T *et al*: **A validated web-based tool to display individualised Crohn's disease predicted outcomes based on clinical, serologic and genetic variables**. *Alimentary Pharmacology & Therapeutics* 2015.
51. Spolverato G, Vitale A, Cucchetti A, Popescu I, Marques HP, Aldrighetti L, Gamblin TC, Maithel SK, Sandroussi C, Bauer TW *et al*: **Can hepatic resection provide a long-term cure for patients with intrahepatic cholangiocarcinoma?** *Cancer* 2015, **121**(22):3998-4006.
52. Suh YJ, Hong YJ, Lee H-J, Hur J, Kim YJ, Lee HS, Hong SR, Im DJ, Kim YJ, Park CH *et al*: **Prognostic value of SYNTAX score based on coronary computed tomography angiography**. *International Journal Of Cardiology* 2015, **199**:460-466.
53. Tada H, Takanashi J-I, Okuno H, Kubota M, Yamagata T, Kawano G, Shiihara T, Hamano S-I, Hirose S, Hayashi T *et al*: **Predictive score for early diagnosis of acute encephalopathy with biphasic seizures and late reduced diffusion (AESD)**. *Journal Of The Neurological Sciences* 2015, **358**(1-2):62-65.
54. Takahashi N, Leng S, Kitajima K, Gomez-Cardona D, Thapa P, Carter RE, Leibovich BC, Sasiwimonphan K, Sasaguri K, Kawashima A: **Small (< 4 cm) Renal Masses: Differentiation of Angiomyolipoma Without Visible Fat From Renal Cell Carcinoma Using Unenhanced and Contrast-Enhanced CT**. *AJR American Journal Of Roentgenology* 2015, **205**(6):1194-1202.
55. Zhou J, Zhou Y, Cao S, Li S, Wang H, Niu Z, Chen D, Wang D, Lv L, Zhang J *et al*: **Multivariate logistic regression analysis of postoperative complications and risk model establishment of gastrectomy for gastric cancer: A single-center cohort report**. *Scandinavian Journal Of Gastroenterology* 2016, **51**(1):8-15.
56. Scott J, Huskisson EC: **Accuracy of subjective measurements made with or without previous scores: an important source of error in serial measurement of subjective states**. *Annals of the Rheumatic Diseases* 1979, **38**(6):558-559.
57. Daoust R, Sirois MJ, Lee JS, Perry JJ, Griffith LE, Worster A, Lang E, Paquet J, Chauny JM, Emond M: **Painful Memories: Reliability of Pain Intensity Recall at 3 Months in Senior Patients**. *Pain research & management* 2017, **2017**:5983721.
58. Tso E, Elson P, Vanlente F, Markman M: **The "real-life" variability of CA-125 in ovarian cancer patients**. *Gynecologic oncology* 2006, **103**(1):141-144.
59. Tuxen MK, Soletormos G, Petersen PH, Schioler V, Dombernowsky P: **Assessment of biological variation and analytical imprecision of CA 125, CEA, and TPA in relation to monitoring of ovarian cancer**. *Gynecologic oncology* 1999, **74**(1):12-22.
60. Peake M, Whiting M: **Measurement of Serum Creatinine – Current Status and Future Goals**. *Clinical Biochemist Reviews* 2006, **27**(4):173-184.
61. Reinhard M, Erlandsen EJ, Randers E: **Biological variation of cystatin C and creatinine**. *Scandinavian journal of clinical and laboratory investigation* 2009, **69**(8):831-836.
62. Macy EM, Hayes TE, Tracy RP: **Variability in the measurement of C-reactive protein in healthy subjects: implications for reference intervals and epidemiological applications**. *Clinical Chemistry* 1997, **43**(1):52-58.
63. Kobayashi H: **Effect of measurement duration on accuracy of pulse-counting**. *Ergonomics* 2013, **56**(12):1940-1944.

64. Sawers L: **Measuring and modelling concurrency**. *Journal of the International AIDS Society* 2013, **16**(1):17431.
65. Delanaye P, Schaeffner E, Ebert N, Cavalier E, Mariat C, Krzesinski J-M, Moranne O: **Normal reference values for glomerular filtration rate: what do we really know?** *Nephrology Dialysis Transplantation* 2012, **27**(7):2664-2672.
66. Braga F, Ferraro S, Mozzi R, Panteghini M: **The importance of individual biology in the clinical use of serum biomarkers for ovarian cancer**. In: *Clinical Chemistry and Laboratory Medicine (CCLM).* vol. 52; 2014: 1625.
67. Polley M-YC, Leung SCY, McShane LM, Gao D, Hugh JC, Mastropasqua MG, Viale G, Zabaglo LA, Penault-Llorca F, Bartlett JMS *et al*: **An International Ki67 Reproducibility Study**. *JNCI Journal of the National Cancer Institute* 2013, **105**(24):1897-1906.
68. Ali A, Black D, Soslow RA: **Difficulties in assessing the depth of myometrial invasion in endometrial carcinoma**. *International journal of gynecological pathology : official journal of the International Society of Gynecological Pathologists* 2007, **26**(2):115-123.
69. Roehrborn CG, Pickens GJ, Carmody T, 3rd: **Variability of repeated serum prostate-specific antigen (PSA) measurements within less than 90 days in a well-defined patient population**. *Urology* 1996, **47**(1):59-66.
70. Winkel P, Statland BE, Bokelund H: **Factors contributing to intra-individual variation of serum constituents: 5. Short-term day-to-day and within-hour variation of serum constituents in healthy subjects**. *Clin Chem* 1974, **20**(12):1520-1527.
71. Braga F, Panteghini M: **Generation of data on within-subject biological variation in laboratory medicine: An update**. *Critical Reviews in Clinical Laboratory Sciences* 2016, **53**(5):313-325.
72. Mori H, Kobara H, Tsushimi T, Nishiyama N, Fujihara S, Masaki T: **Unavoidable Human Errors of Tumor Size Measurement during Specimen Attachment after Endoscopic Resection: A Clinical Prospective Study**. *PloS one* 2015, **10**(4):e0121798.
73. Hill A, Roberts J: **Body mass index: a comparison between self-reported and measured height and weight**. *Journal of public health medicine* 1998, **20**(2):206-210.
74. Jurek AM, Greenland S, Maldonado G, Church TR: **Proper interpretation of non-differential misclassification effects: expectations vs observations**. *International journal of epidemiology* 2005, **34**(3):680-687.
75. Wardenaar KJ, Conradi HJ, de Jonge P: **Data-driven course trajectories in primary care patients with major depressive disorder**. *Depression and anxiety* 2014, **31**(9):778-786.
76. Von Korff M, Deyo RA, Cherkin D, Barlow W: **Back pain in primary care. Outcomes at 1 year**. *Spine* 1993, **18**(7):855-862.
77. Scheele J, Luijsterburg PA, Ferreira ML, Maher CG, Pereira L, Peul WC, van Tulder MW, Bohnen AM, Berger MY, Bierma-Zeinstra SM *et al*: **Back complaints in the elders (BACE); design of cohort studies in primary care: an international consortium**. *BMC musculoskeletal disorders* 2011, **12**:193.
78. Radanov BP, di Stefano G, Schnidrig A, Ballinari P: **Role of psychosocial stress in recovery from common whiplash [see comment]**. *Lancet (London, England)* 1991, **338**(8769):712-715.
79. Licht-Strunk E, Beekman AT, de Haan M, van Marwijk HW: **The prognosis of undetected depression in older general practice patients. A one year follow-up study**. *Journal of affective disorders* 2009, **114**(1-3):310-315.
80. Hermsen LA, Leone SS, van der Windt DA, Smalbrugge M, Dekker J, van der Horst HE: **Functional outcome in older adults with joint pain and comorbidity: design of a prospective cohort study**. *BMC musculoskeletal disorders* 2011, **12**:241.
81. Diehm C, Darius H, Pittrow D, Schwertfeger M, Tepohl G, Haberl RL, Allenberg JR, Burghaus I, Trampisch HJ: **Prognostic value of a low post-exercise ankle brachial index as assessed by primary care physicians**. *Atherosclerosis* 2011, **214**(2):364-372.

742

1 *Table 1: Data extracted from each article*

Design and aim	• Prognostic versus diagnostic prediction model
	• Intended scope of the review
	  – Clinical area
	  – Aim of prediction model (e.g. inform therapeutic decision making, inform referral or withholding from invasive diagnostic testing, inform patients of probability of event)
	• Source of data (e.g. cohort, case-control, randomised trial or registry data)

Outcomes to be predicted	• Definition and method for measurement of outcome • Type of outcome (e.g. single or combined endpoints; binary or time to event)
Candidate predictors	• Number and type of predictors (e.g. demographics, patient history, physical examination, additional testing, disease characteristics) • Definition and method for measurement of candidate predictors • Timing of predictor measurement • Handling of predictors in the modelling (e.g. continuous, linear, non-linear transformations or categorised)
Sample size	• Number of participants • Number of outcomes/events • Number of outcomes/events in relation to the number of candidate predictors (events per variable)
Missing data	• How much missing data • Handling of missing data (e.g. complete-case analysis, imputation, or other methods)
Model development	• Modelling method (e.g. logistic or survival) • Method for selection of predictors for inclusion in multivariable modelling • Method for selection of predictors during multivariable modelling
Intended moment of using the model & timing of predictor measurements	• Intended moment of use • Timing of the measurement of predictors included in the final model, and whether it matched the intended moment of using the model
Measurement error of predictors	• Susceptibility to measurement error for the predictors included in the final model • Whether measurement error was accounted for and, if so, how
Model performance	• Calibration (e.g. calibration slope, calibration plot, Hosmer-Lemeshow test) • Discrimination (e.g. c-statistic, D-statistic, log-rank) • Classification measures (e.g. sensitivity, specificity, predictive values, net reclassification improvement)
Model evaluation	• Method used for testing model performance: internal (e.g. random split of data, resampling methods, none) or external (e.g. temporal, geographical, different setting, different investigators) • In case of poor validation, whether the model was adjusted or updated (e.g. intercept recalibrated, predictor effects adjusted, new predictors added)

2

3 *Table 2: Predictors in final models at high risk of measurement error*

Key reasons for being at high risk of error	**Predictors included in final models**
Fluctuations in human samples/ biological variability	Serum albumin, Serologic markers, Prostate Specific Antigen (PSA) density, Prostate Specific Antigen (PSA), Ki-67, Human epididymis protein 4 (HE4), Glomerular filtration rate, Emergency room pulse rate, CRUSADE score, C-reactive protein, Creatinine on admission, CA125, Ascites
Inaccuracy of measurement instruments	Body Mass Index (BMI), Myometrial invasion depth, Emergency room pulse rate, Creatinine on admission, Weight, Ascites, International normalised ratio (INR1) , Infection/bioburden
Imperfect recall	Body Mass Index (BMI), Duration of convulsions, Duration of drowsiness, Duration of neck pain, Duration of nervousness, Duration of tingling, History of transactional sex, Area under pain curve, Congestive heart failure, Weight, Previous bleeding, Endoscopic retrograde cholangiopancreatography (ERCP) time, Time developing motor deficits, ImPACT total symptom score, Eastern Cooperative Oncology Group (ECOG) performance status, Depression, Number of non-major comorbidities, Systemic illness/organ failure
Subjective nature of measures	Abdominal pain, Tumour stage, Suboptimal pelvic examination or enlarged uterus during preoperative evaluation, Area under pain curve, Hypertension, Clinical stage, Malnutrition, Obesity, Procedure risk category, Pressure ulcer stage, ImPACT total symptom score, Eastern Cooperative Oncology Group (ECOG) performance status, Depression, Pre-catheterisation diagnosis
Laboratory or measurer error	Tumour stage, Suboptimal pelvic examination or enlarged uterus during preoperative evaluation, Myometrial invasion depth, CRUSADE score, CA125, Histologic grade, Primary tumour diameter, Clinical stage, Residual tumour, Endoscopic Retrograde Cholangiopancreatography (ERCP) Time, Tumour size, Pressure ulcer stage, Ascites, International normalised ratio (INR1) , Peritoneal Cancer Index, Infection/bioburden, Operating time and age, Wound (ulcer) age at first encounter

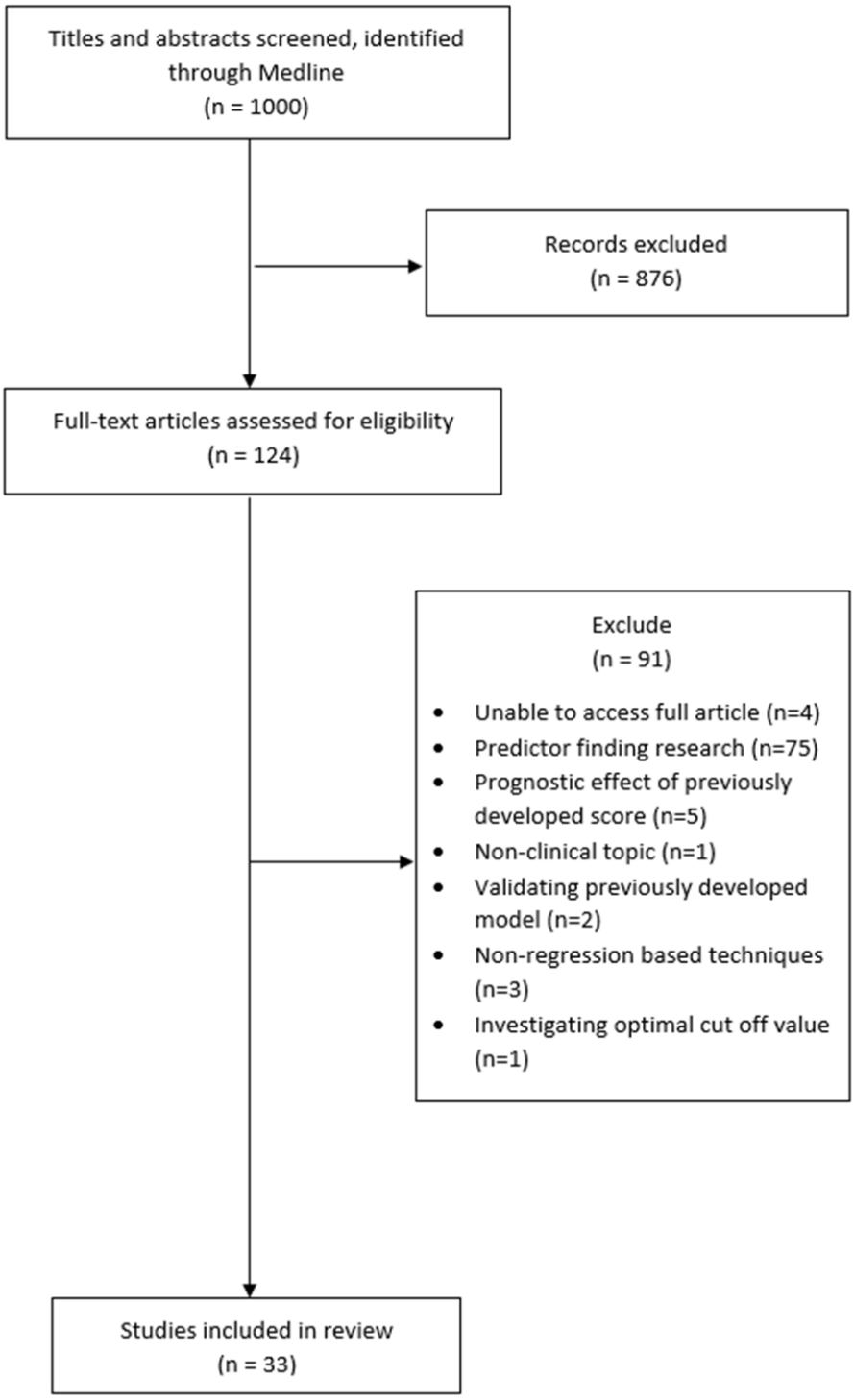
4

Table 3: Key reasons for measurement error in examples of predictors at high risk of being susceptible to error

Predictor	**Key Reasons**	**Explanation**
Area under pain curve	Subjective/subject to recall	Requires patient to report pain, which is a subjective measure and could report the same pain differently at a different time/by a different method or if previous scores were not provided [56], and recall incorrectly [57]
CA125	Biological variability/ laboratory error	Assay imprecision can contribute considerably to result variations in a conventional laboratory setting [58] and changes can occur due to normal biological variation [59]
Creatinine on admission	Biological variability/ inaccuracy of measurement method	Bias and imprecision may occur by use of different measurement methods [60] and changes can occur due to normal biological variation [61]
C-reactive protein	Biological variability	Within-individual variability exists, so a second sample may produce different results [62]
CRUSADE score	Biological variability/measurer error	May get a different value if calculated again shortly afterwards as includes measures that vary and may be affected by measurer error such as of blood pressure [22]
Emergency room pulse rate	Biological variability/inaccuracy of measurement method	May change if measured a couple of minutes later and there may be error depending on how long the measurer counted for [63]
History of transactional sex	Imperfect recall	Patient may not be truthful about history [64]
Glomerular filtration rate	Biological variability	A second sample may produce different results due to biological variation [65]
Human epididymis protein 4 (HE4)	Biological variability	A second sample may produce different results [66]
Ki-67	Biological variability	A second sample may produce different results and differences may be present from different laboratories [67]
Myometrial invasion depth	Measurer error/inaccuracy of measurement method	Results may be different when reassessed [68]

Prostate specific antigen (PSA)	Biological variability	A second sample may produce different results [69]
Serum albumin	Biological variability	A second sample may produce different results [70]
Tumour stage	Subjective/measurer error	May get a different result from different assessors dependent on experience level or areas of speciality

Figure 1: Flow chart of included studies



WHAT IS NEW?

Key findings

- Many published prediction models include predictors that are susceptible to measurement error and this measurement error is not being acknowledged or accounted for in the development of the models.
- Most prediction model articles do not explicitly state the intended moment of model use, or exactly when the predictors used in the model development were measured.

What this adds to what is known

- Reporting of measurement error and intended moment of model use is poor in prediction model studies.

What is the implication, what should change now?

- There is a need to identify circumstances where ignoring measurement error in prediction models is consequential and whether accounting for the error will improve the predictions.
- Future prediction model research studies must clearly report the intended moment of use of the prediction model, and be explicit about when the predictors were measured.