White *et al. BMC Medicine* (2023) 21:339
https://doi.org/10.1186/s12916-023-03048-6

BMC Medicine

RESEARCH ARTICLE Open Access

Evidence of questionable research practices in clinical prediction models

Nicole White[1], Rex Parsons[1], Gary Collins[2] and Adrian Barnett[1*]

Abstract

Background Clinical prediction models are widely used in health and medical research. The area under the receiver operating characteristic curve (AUC) is a frequently used estimate to describe the discriminatory ability of a clinical prediction model. The AUC is often interpreted relative to thresholds, with "good" or "excellent" models defined at 0.7, 0.8 or 0.9. These thresholds may create targets that result in "hacking", where researchers are motivated to re-analyse their data until they achieve a "good" result.

Methods We extracted AUC values from *PubMed* abstracts to look for evidence of hacking. We used histograms of the AUC values in bins of size 0.01 and compared the observed distribution to a smooth distribution from a spline.

Results The distribution of 306,888 AUC values showed clear excesses above the thresholds of 0.7, 0.8 and 0.9 and shortfalls below the thresholds.

Conclusions The AUCs for some models are over-inflated, which risks exposing patients to sub-optimal clinical decision-making. Greater modelling transparency is needed, including published protocols, and data and code sharing.

Keywords Prediction model, Area under curve, Diagnosis, Prognosis, Hacking, Statistics, Receiver operating characteristic

Background

Clinical prediction models estimate an individual's risk of being diagnosed with a disease or experiencing a future health outcome [1, 2]. A clinical prediction model uses multivariable analysis methods to estimate the risk of experiencing an outcome based on individual-level variables, for example, a model predicting a patient's risk of death after admission to intensive care using data from their medical history and test results [3].

Researchers are motivated to build clinical prediction models because of their potential to support decision-making. Clinical decisions can be based on the model's estimated probabilities or risk categories defined by probability cut-points to give qualitative interpretations, e.g. low and high risk [4]. Decisions guided by model probabilities or categories may rule out low-risk patients to reduce unnecessary treatments or identify high-risk patients for additional monitoring.

The number of published clinical prediction models has increased in recent years. A validated search strategy in MEDLINE [5] shows that an average of 4200 publications related to clinical prediction modelling are now being published weekly (searched 20 January 2023).

Despite being a popular study design, clinical prediction models are often poorly executed. Factors driving poor model quality include inadequate sample sizes,

*Correspondence:
Adrian Barnett
a.barnett@qut.edu.au
[1] Australian Centre for Health Services Innovation and Centre for Healthcare Transformation, School of Public Health and Social Work, Faculty of Health, Queensland University of Technology, Kelvin Grove, Queensland, Australia
[2] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK

inappropriate exclusions, poor handling of missing data, limited model validation, and inflated estimates of performance [6–11]. A review of prediction models for COVID-19 found only 7 of the 606 published models were potentially useful for practice [12]; other reviews have identified shortcomings in model development that introduce bias into model predictions [13, 14]. Design failings are often compounded by poor reporting, despite the availability of expert-led guidance to improve transparency [1, 15, 16].

Rigorous testing of clinical prediction models is essential before considering their use in practice [17]. A good prediction model will have a strong discrimination, which is a model's ability to separate patients based on their estimated risk. The area under the receiver operating characteristic curve (AUC) is an overall measure of model discrimination. It the probability that a model predicts a higher risk for a randomly selected patient *with* the outcome of interest than a randomly selected patient *without* the outcome of interest [18]. If the model has good discrimination and gives estimated risks for all patients with the outcome that are higher than all patients without, then the AUC will be 1. If the model discrimination is no better than a coin toss, then the AUC will be 0.5. The AUC is also known as the AUROC, c-statistic for binary outcomes, and c-index for time-to-event outcomes.

Qualitative descriptors of model performance for AUC thresholds between 0.5 and 1 have been published, for example:

- "0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding" [19].
- "The area under the ROC curve (AUC) results were considered excellent for AUC values between 0.9 and 1, good for AUC values between 0.8 and 0.9, fair for AUC values between 0.7 and 0.8, poor for AUC values between 0.6 and 0.7 and failed for AUC values between 0.5 and 0.6" [20].
- "Areas under the curve (AUCs) of 0.6 to 0.7, 0.7 to 0.8, 0.8 to 0.9 and > 0.9 were considered acceptable, fair, good and excellent for discrimination, respectively" [21].

Additional examples are in Additional file 1. These thresholds have no clear origin, but they are likely used because they transform the AUC from a number into a qualitative rating of performance. The thresholds have no scientific basis and are arbitrarily based on digit preference, often occurring at 0.7, 0.8 and 0.9 [22]. Previous research has examined the labels applied to AUC values in 58 papers and recommended that AUC values should be presented without labels [23].

Thresholds may create targets that some researchers will strive to achieve. We hypothesised that some researchers have engaged in questionable research practices or "hacking" to create models with estimated AUCs that better commonly used thresholds, including (1) re-analysing data and creating multiple models to get an AUC value over a threshold and (2) selectively reporting the best AUC value from many models [24]. Assuming the AUC has multiple thresholds (0.7, 0.8 and 0.9), we expected the distribution of AUC values would be undulating rather than smooth, with excess values just above the thresholds. We describe some ways a prediction model can be "hacked" in Table 1, but note this is not an exhaustive list.

The use of thresholds when reporting results from statistical analysis is not new. Well-known examples include 0.05 for the statistical significance of hypothesis tests and an 80% power to justify sample size calculations [30]. Related research has examined the enormous excess of *p*-values just below the widely used 0.05 threshold, which is caused by multiple data dredging techniques, including re-analyses of data and selective reporting [26, 31–33]. Recent research has also shown the same problem for Cronbach's alpha at the "acceptable" threshold of 0.7 [34].

Methods

Data extraction

We aimed to find abstracts that included an area under the curve value or the related c-index for survival and c-statistic for binary outcomes [35]. These estimates have a variety of names, including "area under the receiver operating characteristic curve" or the acronyms "AUC" and "AUROC". We included all AUCs regardless of the study's aim and therefore included model development and validation studies. We did not consider other commonly reported metrics for evaluating clinical prediction models.

We examined abstracts published in *PubMed* because it is a large international database that includes most health and medical journals. To indicate its size, there were over 1.5 million abstracts published on *PubMed* in 2022. The National Library of Medicine make the *PubMed* data freely and easily available for research. We downloaded the entire database in XML format on 30 July 2022 from https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/.

We started with all the available *PubMed* data. Our exclusion criteria were as follows:

- Entries with an empty abstract or an abstract of 10 words or fewer

Table 1 Examples of how a clinical prediction model can be hacked to get a better AUC value that is likely to be over-inflated as the model is over-fitted. Some of these approaches create multiple results from which the best result can be selected, often without disclosing the multiple results. Some hacking may be unintentional as researchers believe they are following standard practice. Some approaches can be acceptable when combined with appropriate validation, but the number of models fitted should always be disclosed and should be pre-defined in a protocol or pre-registration [1]

- Selectively choosing data sets (from those that are available to the researcher) to build and evaluate a model
- Collecting more data until a desirable AUC value is reached [25, 26]
- Fitting multiple, potentially hundreds, of models based on subsets of potential predictors [26]
- Trialling different cut-points when dichotomising continuous predictors until a "good" AUC is achieved [27]
- Including predictors that are proxies of the outcome or that work via reverse causality, for example, using blood tests taken after the outcome
- Changing the outcome variable, for example to a proxy of the original diagnostic outcome [26]
- Trialling alternative methods for imputing missing data [26]
- Removing observations that are difficult to fit [25, 26]
- Trialling different modelling approaches, e.g. logistic regression models and classification trees [28]
- Rounding up an AUC value to pass a threshold, for example reporting 0.79 as 0.8 [25]
- Choosing the "best" random seed for split sample validation or a model's hyper-parameters [29]
- Not using internal validation, so the model performance is evaluated in the same data used to develop the model

- Pharmacokinetic studies, which often use area under the curve statistics to refer to dosages and volumes that are unrelated to prediction models
- Meta-analyses or pooled analyses, as we were interested in original research
- Tutorial papers, as these may not report original findings

Our inclusion criterion was abstracts with one or more AUC values.

We created a text-extraction algorithm to find AUC values using the team's expertise and trial and error. We validated the algorithm by randomly sampling 300 abstracts with a Medical Subject Heading (MeSH) of "Area under curve" that had an abstract available and quantifying the number of AUC values that were correctly extracted. We also examined randomly selected results from the algorithm that equalled the thresholds of 0.7, 0.8 or 1, with 300 abstracts per threshold examined. We report the validation in more detail in the results, but note here that the algorithm could not reliably extract AUC values that were exactly 1. AUC values equal to 1 were therefore excluded.

Challenges in extracting the AUC values from abstracts included the frequent use of long lists of statistics, including the sensitivity and specificity; unrelated area under the curve statistics from pharmacokinetic studies; references to AUC values as a threshold (e.g. "The AUC ranges between 0.5 and 1"); and the many different descriptors used, including "area under the curve", "receiver operating characteristic curve", and related acronyms.

AUC values reported as a percent were converted to 0 to 1. We removed any AUC values that were less than 0 or greater than or equal to 1.

We categorised each AUC value as a mean or the lower or upper limit of the confidence interval, for example, "0.704 (95% CI 0.603 to 0.806)" would be a mean, lower and upper limit, respectively.

For the specific examples from published papers in the results, we give the *PubMed* ID number (PMID) rather than citing the paper.

R version 4.2.1 was used for data extraction and analysis [36]. The code and analysis data are available online: https://github.com/agbarnett/area_under_curve [37].

Statistical analysis

Our hypothesis was that there would be an excess of AUC values just above the thresholds 0.7, 0.8 and 0.9. To examine this, we used a histogram with bins of (lower, upper], with lower thresholds of 0, 0.01 to 0.99, and an upper threshold that was + 0.01 greater. For example, the bin of (0.69, 0.70] included every AUC greater than 0.69 and less than or equal to 0.70. We excluded AUCs with a decimal place of 1 (e.g. "0.8"), as these results would create spikes in the histogram that were simply due to rounding.

We do not know what the distribution of AUC values from the health and medical literature would look like if there was no AUC-hacking. However, we are confident that it should be relatively smooth with no inflexion points. An undulating distribution, especially near the thresholds (0.7, 0.8 and 0.9), would be a strong sign of AUC-hacking, potentially caused by re-analysing the data to get a more publishable but inflated AUC.

We estimated the shape of a smooth distribution using a natural spline with 4 degrees of freedom fitted using a Poisson distribution [38]. We created residuals by subtracting the observed counts from the smooth fit. A similar approach was used to identify departures from a smooth distribution for a large sample of Cronbach's alpha statistics [34].

During data collection, we noted that many abstracts gave multiple AUC values from competing models. To examine the best model per abstract, we plotted the distribution using the highest AUC value per abstract. This subgroup analysis examined whether the best presented models were often just above the thresholds.

We used a subgroup analyses that used only AUC values from the results section of structured abstracts. This potentially increased the specificity of the extracted AUC values, as those from the introduction, methods and discussion sections were more likely to be general references to the AUC rather than results.

To investigate the role of publication bias, we used a subgroup analysis of only papers published in the journal *PLOS ONE* which welcomes "negative" results and does not select based on impact or novelty [39].

Results

The flow chart of included abstracts is shown in Fig. 1.

The number of examined abstracts was over 19 million, and 96,986 (0.5%) included at least one AUC value. The use of AUC values has become more popular in recent years (see Additional file 2: Fig. S1). The median publication year for the AUC values was 2018, with first to third quartile of 2015 to 2018.

For abstracts with at least one AUC value, the median number of AUC values was 2, with a first to third quartile of 1 to 4 (see Additional file 3: Fig. S2). There was a long tail in the distribution of AUC values, with 1.1% of abstracts reporting 20 or more AUC values. These high numbers were often from abstracts that compared multiple models. The total number of included AUC values was 306,888. There were 92,529 (31%) values reported as lower or upper confidence limits and the remainder as means.

The distribution of AUC mean values (excluding confidence intervals) and residuals from the smoothed fit to the distribution are in Fig. 2. There are clear changes in the distribution around the thresholds of 0.7, 0.8 and 0.9. There is a large excess of AUC values just above 0.7, followed by a deficit before 0.8. There is a large jump in the number of AUCs just above 0.8 compared with (0.79, 0.80]. A similar excess is observed for AUC values just above 0.9. The frequencies in the histogram and residuals are worth noting, as they indicate thousands of unexpected results. There were 2106 (1.0%) AUC values presented to 1 decimal place that were excluded from the histogram.

The distribution from the largest AUC mean value per abstract excluding confidence intervals is shown in Fig. 3. The strong changes in the distribution at the thresholds observed in Fig. 2 remain.

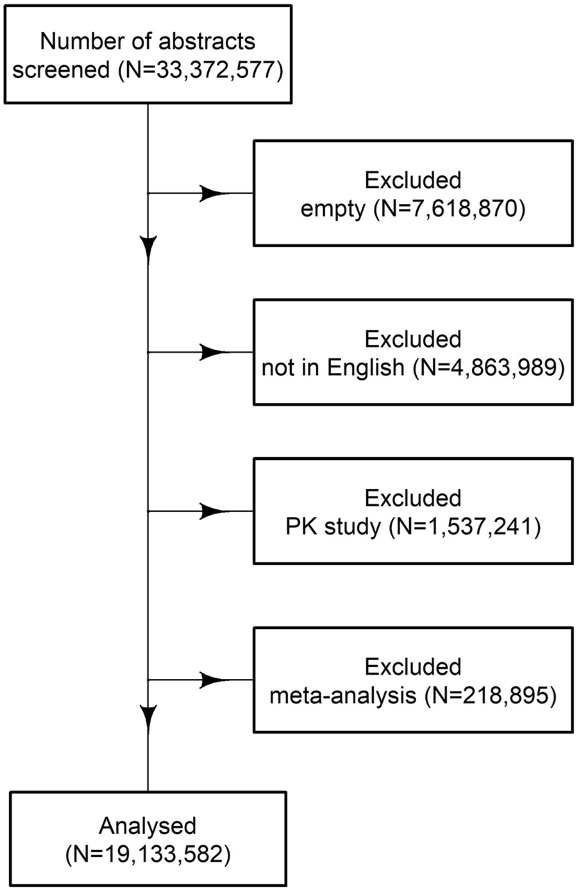
Number of abstracts screened (N=33,372,577)
Excluded empty (N=7,618,870)
Excluded not in English (N=4,863,989)
Excluded PK study (N=1,537,241)
Excluded meta-analysis (N=218,895)
Analysed (N=19,133,582)

Fig. 1 Flow chart of included abstracts. PK, pharmacokinetic

The distribution for AUC values from the results section only is in Additional file 4: Fig. S3; the shape of the distribution is similar to that using all AUC values. The distributions for the lower and upper limits of the confidence interval were generally smoother than the mean; see Additional file 5: Fig. S4. However, there was a notable excess at (0.56, 0.57] for the lower interval.

The distribution for AUC values published in *PLOS ONE* show a similar pattern to the full sample, with many more AUC values just above the 0.8 threshold (see Additional file 6: Fig. S5).

Validation

We validated our algorithm against 300 manually entered abstracts. For 192 abstracts, there were no AUC values in the abstract, and the algorithm correctly identified these absences for 93%. Some errors were because the abstracts were selected using the MESH term "Area under the curve" meaning that many pharmacokinetic studies were included. For the 108 abstracts with an AUC, the

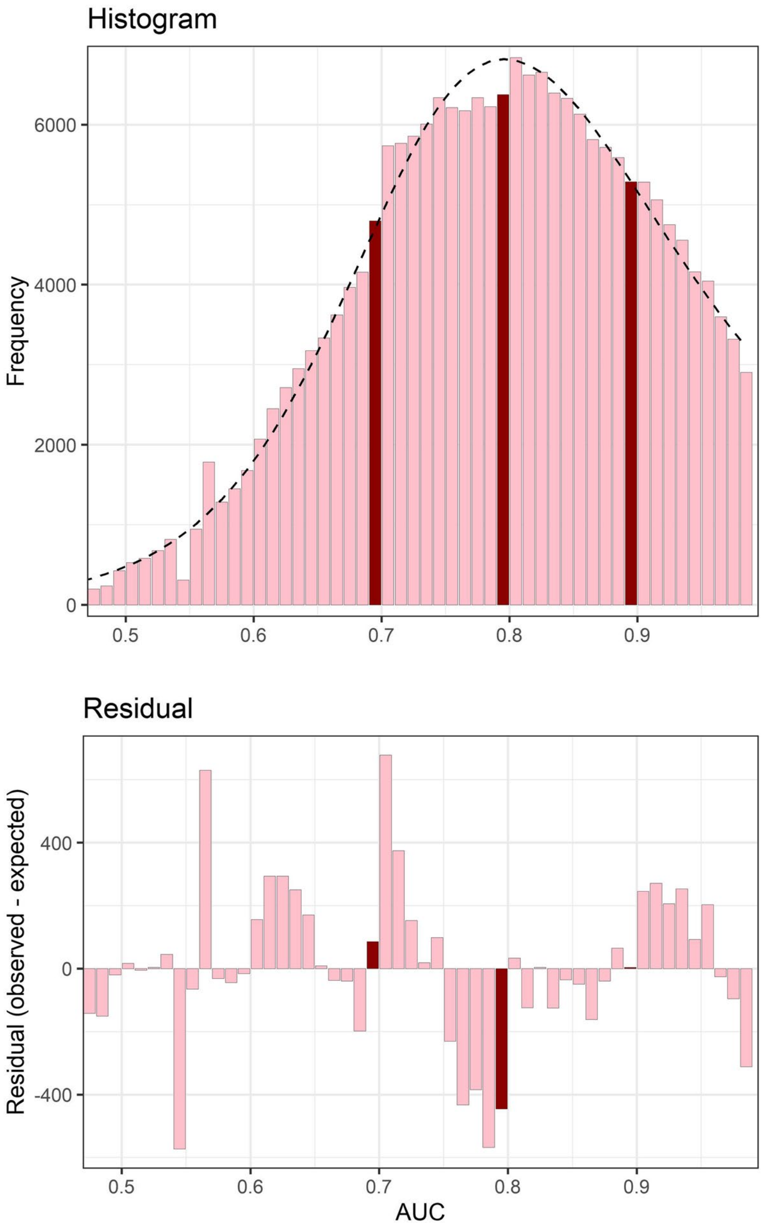

Fig. 2 Histogram of AUC mean values (top panel) and residuals from a smooth fit to the histogram (bottom panel). The dotted line in the top panel shows the smooth fit

algorithm identified 98% correctly. See Additional file 7 for details.

For abstracts where either the algorithm or manual entry found one or more AUC values, we made a Bland–Altman plot of the number of AUC values extracted (see Additional file 7: Fig. S6). The 90% limits of agreement were – 2 to 0. On average, the algorithm missed more AUC values than the manual entry, a discrepancy that was generally due to non-standard presentations. We are comfortable with this difference, as we would rather lean towards missing valid AUC values than wrongly including invalid AUC values.

We used a regression model to examine differences in the AUC values extracted by the algorithm and manual entry. AUC values that were wrongly included by the algorithm were smaller on average than the AUC values that were correctly included. This is because the values extracted were often describing other aspects of the prediction model, for example, the Brier score, sensitivity and specificity.

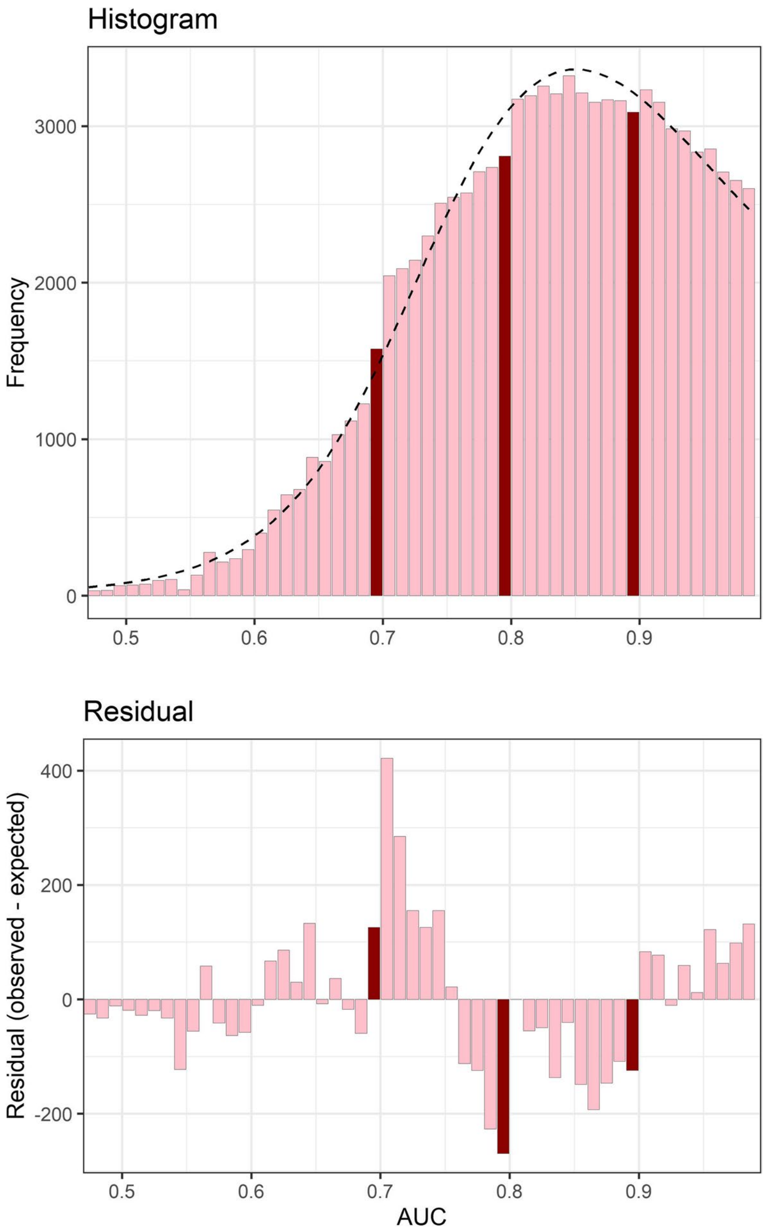

Fig. 3 Histogram of the largest AUC mean value per abstract (top panel) and residuals from a smooth fit to the histogram (bottom panel). The dotted line in the top panel shows the smooth fit

The validation helped identify MESH terms that identified pharmacokinetic studies that were excluded from our main analysis. In a second validation, we manually checked 100 randomly sampled abstracts that the algorithm identified as not having an AUC statistic and another 100 randomly sampled abstracts that the algorithm identified as having an AUC statistic. All abstracts identified as not having an AUC statistic were correctly classified (95% confidence interval for negative predictive value: 0.964 to 1.000). All but one abstract identified as having an AUC statistic was correct (95% confidence interval for positive predictive value: 0.946 to 1.000).

In a third validation, we manually checked 300 AUC values extracted by our algorithm at the thresholds 0.7, 0.8 and 1. The results led us to exclude AUC values of 1, because these could not be accurately extracted. At the thresholds of 0.7 and 0.8, there were some errors due to qualitative descriptions of the thresholds instead of actual results. For comparison with the rounded thresholds (0.7 and 0.8), we manually checked 300 AUC values at 0.81. These checks had fewer errors as they were more often AUC values and were not being used as descriptive thresholds. We do not believe that the errors undermine our main point about AUC-hacking. The greater errors at the thresholds mean that the numbers at 0.7, 0.8 and 0.9 in Fig. 2 should likely be smaller, making for a larger gap between the thresholds and values exceeding the threshold, which would be stronger evidence of poor practice.

To investigate the excess at (0.56, 0.57], we manually extracted the AUC values from 300 abstracts where our algorithm found an AUC value of 0.57 and another 300 from 0.58 as a nearby comparison with no excess. The error proportions from the algorithm were relatively low (see Additional file 7: Table S3), indicating that the excess at 0.57 was not due to errors.

Evidence of poor practice

Although documenting poor practice was not our goal, whilst reading abstracts, we encountered mistakes, poor reporting, and potential spin. Some papers had mistakes in their results, for example a 95% confidence interval for the AUC from 0 to 1 (PMID34795784) and upper limits over 1 (PMID34880677); others used excessive decimal places (PMID34456583). Some abstracts displayed a poor understanding of the AUC value, including authors declaring a "highest" AUC of 0.5 which is equivalent to a coin toss (PMID28708299); an upper confidence interval of an AUC value that was under 0.5, possibly because the disease label was mistakenly reversed (PMID28795781); the AUC being misinterpreted as a direct measure of sensitivity or specificity (PMID34674968); and the AUC being misinterpreted as a regression coefficient (PMID19410507). There were instances of potential spin [11, 40], with relatively low AUC values under 0.75 described as "excellent" (PMID35222547).

Discussion

P-hacking has been observed in large parts of the literature [31, 33, 41, 42], so it is disappointing but not surprising to see hacking in AUC values. The discontinuities in the AUC distribution around the thresholds of 0.7, 0.8 and 0.9 shown by our analysis are not as egregious as the previously observed *p*-value discontinuity at 0.05. This is likely because AUC-hacking is spread over multiple thresholds, and because the *p*-value is often the most important statistic to some authors, as demonstrated by the verbal gymnastics applied to "non-significant" *p*-values [43].

There was a surprising shortfall of AUC values at (0.54, 0.55] and excess at (0.56, 0.57]. This pattern could also be due to hacking, as estimates up to 0.55 could be viewed by some researchers as too close to 0.5, indicating model predictions that are no better than random (values under 0.55 would be presented as 0.5 if rounded to one decimal place). This would explain the distinct lack of lower confidence limits at (0.54, 0.55] (Additional file 5: Fig. S4), as some researchers would be unhappy with a confidence interval that was not statistically significant when compared with the null hypothesis at 0.5. Multiple re-analyses options are available to tweak the lower confidence limit over the threshold, including adding predictors to the model or removing attested outliers (see Table 1) [44].

The implications of hacked analyses for the literature and evidence-based medicine is that some clinical prediction models will have less utility for health systems than promised. This is potentially serious if models have been translated into practice based on "excellent" AUC values that were hacked. Decisions about patient care will be compromised, with potentially missed diseases and unnecessary interventions. Hacking likely explains some of the reduction in model performance when published prediction models are externally validated [45], with the inflated AUCs values regressing to the mean.

Hacking may be lessened by using protocols, analysis plans, and registered reports [46, 47]. However, the uptake of registered reports has been modest, and protocols do not completely prevent important changes to the analysis [48–50]. Despite this poor uptake and practice, it is possible that protocols and registered reports will be an important part of future best practice, and they can help build trust in research, together with data and code sharing [51].

Our results indicate that some researchers have prioritised reporting a "good" result in their abstract that will help them publish their paper. By doing so, the wider issues of what is needed to produce a high-quality prediction model are downplayed. An AUC value alone cannot determine if a model is "acceptable" or "excellent". As a measure of model discrimination, the AUC represents just one aspect of prediction model performance. Other important aspects include the model's calibration, the costs and implications of false negatives and false positives, and whether a model is worthwhile for practice [52–56].

We found evidence that some researchers do not understand the AUC value, with errors in the presentation and interpretation of the AUC. Researchers have easy access multiple software tools to create AUC values, but may be unwilling to spend time learning the theory that underpins prediction models, leaving them with a poor understanding of a model's limitations [57].

Evidence of hacking in practice is available from recent surveys which have reported relatively high instances of researchers engaging in questionable practices and fraud. A survey of Australian researchers reported many were aware of instances where colleagues had made up data (10%), altered data (8%), selectively excluded data (27%) or trialled iterative statistical analysis until finding a model that yielded a “significant” result (45%) [58]. A survey of Dutch researchers reported 8% admitted to falsifying or manipulating data [59]. A survey of US statisticians reported that 22% had been asked in the last 5 years to remove or alter data to better support the hypothesis, and 48% had been asked to stress only the “significant” findings [60]. The widespread use of these poor practices creates a biased evidence base and is misinforming health policy.

Limitations

We did not examine other commonly reported performance metrics used to evaluate clinical prediction model performance. It is possible that values such as model sensitivity and specificity may also be influenced by “acceptable” thresholds.

We only used AUC values given in abstracts and did not examine the full text. Some papers may have only presented their best results in the abstract and given a more complete picture in the full text. However, an analysis of *p*-values found that the distribution was similarly blighted by p-hacking when using *p*-values from the abstract or full text [32], and study of spin in prediction models found its occurrence was similar in the abstract and full text [11]. It is likely that the highest AUC value presented in the abstract is also the highest in the full text, so the “best” model would be captured in the abstract, and the “best” AUC value is the one most likely to be created by hacking.

In addition to hacking, publication bias likely also plays a role in the selection of AUC values, with higher values more likely to be accepted by peer reviewers and journal editors. Our subgroup analysis of *PLOS ONE* abstracts (Additional file 6: Fig. S6) provides some evidence that the “hacking” pattern in AUC values is due to author behaviour not journal behaviour.

We used an automated algorithm that provided a large and generalisable sample but did not perfectly extract all AUC values. In particular, we were not able to reliably extract AUC values of 1, and this is an important value as it is the best possible result and could be a target for hacking. We believe that the errors and exclusions in the data are not large enough to change our key conclusion, which is that AUC-hacking has occurred.

Conclusions

Clinical prediction models are growing in popularity, likely because of increased patient data availability and accessible software tools to build models. However, many published models have serious flaws in their design and presentation. Our results show another serious issue, as the AUCs for some models have been over-inflated, and we believe this is due to hacking. Publishing overly optimistic models risks exposing patients to sub-optimal clinical decision-making. An urgent reset is needed in how clinical prediction models are built, validated and peer-reviewed. Actionable steps towards greater transparency are as follows: the wider use of protocols and registered reports, following expert reporting guidance, and increased data and code sharing.

Abbreviations

AUC, AUROC	Area under the receiver operating characteristic curve
CI	Confidence interval
PK	Pharmacokinetic
ROC	Receiver operating characteristic

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12916-023-03048-6.

Additional file 1. Examples of qualitative descriptors for AUC thresholds.

Additional file 2: Figure S1. Number and proportion of abstracts with at least one AUC value over time.

Additional file 3: Figure S2. Bar chart of the number of AUC values per abstract.

Additional file 4: Figure S3. Distribution of AUC values and residuals from a smooth fit to the distribution using only AUC values that were in the results section of the abstract.

Additional file 5: Figure S4. Histograms of AUC values that were lower or upper confidence limits and residuals from a smooth fit to the histograms.

Additional file 6: Figure S5. Subgroup analysis of AUC values from the journal *PLOS ONE*.

Additional file 7: Figure S6. Bland–Altman plot of the difference in the number of AUC values per abstract extracted manually and by the algorithm. **Figure S7.** Box-plots of AUC values grouped by whether they were extracted by the algorithm or manual-check only, or by both. **Table S1.** Estimates from a linear regression model examining the differences in AUC values extracted by the algorithm and manual checking. **Figure S8.** Proportion of correct AUC values from the algorithm for four selected AUC values. **Table S2.** Proportion of correct AUC values from the algorithm for two selected AUC values.

Acknowledgements

Thanks to the National Library of Medicine for making the *PubMed* data available for research.

Authors’ Twitter handles

Twitter handles: @nicolem_white (Nicole White); @RexParsons8 (Rex Parsons); @GSCollins (Gary Collins).

Authors’ contributions

NW: conceptualisation, methodology, data curation, investigation, project administration, writing——review and editing. RP: conceptualisation, data curation, investigation, writing——review and editing. GC: conceptualisation, investigation, writing——review and editing. AB: conceptualisation, methodology, software, validation, data curation, formal analysis, writing——original draft, visualisation. All authors contributed to the interpretation of the results and critical revision of the manuscript. All authors read and approved the final manuscript. The corresponding author attests that all listed authors meet the authorship criteria and that no others meeting the criteria have been omitted.

Funding
GSC was supported by Cancer Research UK (programme grant: C49297/A27294).

Availability of data and materials
The code and analysis data are available online: https://github.com/agbarnett/area_under_curve, including code used to read in and wrangle the publicly available data hosted by *PubMed*. The *PubMed* data are available here https://dtd.nlm.nih.gov/ncbi/pubmed/doc/out/190101/index.html along with the attributes to the accessible data. The specific dataset used in this study can be downloaded by file transfer protocol ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline and can be accessed by a file explorer.

Declarations

Ethics approval and consent to participate
We used publicly available data that were published to be read and scrutinised by researchers and hence ethical approval was not required.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 15 June 2023 Accepted: 24 August 2023
Published online: 04 September 2023

References

1. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1–73.
2. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol. 2021;132:142–5.
3. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985;13(10):818–29. https://doi.org/10.1097/00003246-198510000-00009.
4. Wynants L, van Smeden M, McLernon DJ, Timmerman D, Steyerberg EW, Calster BV. Three myths about risk thresholds for prediction models. BMC Med. 2019;17(1). https://doi.org/10.1186/s12916-019-1425-3.
5. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS ONE. 2012;7(2):e32844.
6. Hand DJ. Classifier technology and the illusion of progress. Stat Sci. 2006;21(1). https://doi.org/10.1214/088342306000000060.
7. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14(1). https://doi.org/10.1186/1471-2288-14-40.
8. Miller E, Grobman W. Prediction with conviction: a stepwise guide toward improving prediction and clinical care. BJOG. 2016;124(3):433. https://doi.org/10.1111/1471-0528.14187.
9. Steyerberg EW, Uno H, Ioannidis JPA, van Calster B, Ukaegbu C, Dhingra T, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. J Clin Epidemiol. 2018;98:133–43. https://doi.org/10.1016/j.jclinepi.2017.11.013.
10. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020:m441. https://doi.org/10.1136/bmj.m441.
11. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models. J Clin Epidemiol. 2023;158:99–110. https://doi.org/10.1016/j.jclinepi.2023.03.024. https://www.sciencedirect.com/science/article/pii/S0895435623000756.
12. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. BMJ. 2020;369. https://doi.org/10.1136/bmj.m1328.
13. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol. 2022;22(1):1–16.
14. Meehan AJ, Lewis SJ, Fazel S, Fusar-Poli P, Steyerberg EW, Stahl D, et al. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. Mol Psychiatry. 2022;27(6):2700–8.
15. Najafabadi AHZ, Ramspek CL, Dekker FW, Heus P, Hooft L, Moons KG, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. BMJ Open. 2020;10(9):e041537.
16. Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc. 2022;29(5):983–9.
17. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.
18. Verbakel JY, Steyerberg EW, Uno H, De Cock B, Wynants L, Collins GS, et al. ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. J Clin Epidemiol. 2020;126:207–216. https://doi.org/10.1016/j.jclinepi.2020.01.028.
19. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 2010;5(9):1315–6. https://doi.org/10.1097/jto.0b013e3181ec173d.
20. Khouli RHE, Macura KJ, Barker PB, Habba MR, Jacobs MA, Bluemke DA. Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast. J Magn Reson Imaging. 2009;30(5):999–1004. https://doi.org/10.1002/jmri.21947.
21. Pitamberwale A, Mahmood T, Ansari AK, Ansari SA, Limgaokar K, Singh L, et al. Biochemical parameters as prognostic markers in severely Ill COVID-19 patients. Cureus. 2022. https://doi.org/10.7759/cureus.28594.
22. Calster BV, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med. 2023;21(1). https://doi.org/10.1186/s12916-023-02779-w.
23. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. Lancet Digit Health. 2022;4(12):e853–5. https://doi.org/10.1016/s2589-7500(22)00188-1.
24. Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F. Questionable research practices in ecology and evolution. PLoS ONE. 2018;13(7):1–16. https://doi.org/10.1371/journal.pone.0200303.
25. John LK, Loewenstein G, Prelec D. Measuring the Prevalence of questionable research practices with incentives for truth telling. Psychol Sci. 2012;23(5):524–32. https://doi.org/10.1177/0956797611430953.
26. Stefan AM, Schönbrodt FD. Big little lies: a compendium and simulation of p-hacking strategies. R Soc Open Sci. 2023;10(2):220346. https://doi.org/10.1098/rsos.220346.
27. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994;86(11):829–35. https://doi.org/10.1093/jnci/86.11.829.
28. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.
29. Picard D. Torch.manual_seed(3407) is all you need: on the influence of random seeds in deep learning architectures for computer vision. CoRR. 2021. arXiv:2109.08203.
30. White NM, Balasubramaniam T, Nayak R, Barnett AG. An observational analysis of the trope "A *p*-value of< 0.05 was considered statistically significant" and other cut-and-paste statistical methods. PLoS ONE. 2022;17(3):e0264360.
31. Masicampo EJ, Lalande DR. A peculiar prevalence of *p* values just below .05. Q J Exp Psychol (Hove). 2012;65(11):2271–2279. https://doi.org/10.1080/17470218.2012.711335.
32. Barnett AG, Wren JD. Examination of confidence intervals in health and medical journals from 1976 to 2019: an observational study. BMJ Open. 2019;9(11). https://doi.org/10.1136/bmjopen-2019-032506.

33. Zwet EW, Cator EA. The significance filter, the winner's curse and the need to shrink. Stat Neerl. 2021;75(4):437–52. https://doi.org/10.1111/stan.12241.
34. Hussey I, Alsalti T, Bosco F, Elson M, Arslan RC. An aberrant abundance of Cronbach's alpha values at .70. 2023. https://doi.org/10.31234/osf.io/dm8xn.
35. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer Series in Statistics. Springer; 2013.
36. R Core Team. R: a language and environment for statistical computing. Vienna; 2023. https://www.R-project.org/.
37. Barnett AG. Code and data for our analysis of area under the curve values extracted from PubMed abstracts. 2023. https://doi.org/10.5281/zenodo.8275064.
38. Ruppert D, Wand MP, Carroll RJ. Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press; 2003. https://doi.org/10.1017/CBO9780511755453.
39. PLOS Collections. Positively Negative: A New PLOS ONE Collection focusing on Negative, Null and Inconclusive Results. 2015. https://everyone.plos.org/2015/02/25/positively-negative-new-plos-one-collection-focusing-negative-null-inconclusive-results/.
40. Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: a methodological systematic review. PLoS Biol. 2017;15(9):e2002173. https://doi.org/10.1371/journal.pbio.2002173.
41. Brodeur A, Cook N, Heyes A. Methods matter: p-hacking and publication bias in causal analysis in economics. Am Econ Rev. 2020;110(11):3634–60. https://doi.org/10.1257/aer.20190687.
42. Adda J, Decker C, Ottaviani M. P-hacking in clinical trials and how incentives shape the distribution of results across phases. Proc Natl Acad Sci U S A. 2020;117(24):13386–92. https://doi.org/10.1073/pnas.1919906117.
43. Otte WM, Vinkers CH, Habets PC, van IJzendoorn DGP, Tijdink JK. Analysis of 567,758 randomized controlled trials published over 30 years reveals trends in phrases used to discuss results that do not reach statistical significance. PLoS Biol. 2022;20(2):e3001562. https://doi.org/10.1371/journal.pbio.3001562.
44. Rohrer JM, Tierney W, Uhlmann EL, DeBruine LM, Heyman T, Jones B, et al. Putting the self in self-correction: findings from the loss-of-confidence project. Perspect Psychol Sci. 2021;16(6):1255–69. https://doi.org/10.1177/1745691620964106.
45. Moons KGM, Donders ART, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. J Clin Epidemiol. 2004;57(12):1262–70. https://doi.org/10.1016/j.jclinepi.2004.01.020.
46. Chambers CD, Tzavella L. The past, present and future of Registered Reports. Nat Hum Behav. 2021;6(1):29–42. https://doi.org/10.1038/s41562-021-01193-7.
47. Penders B. Process and bureaucracy: scientific reform as civilisation. Bull Sci Technol Soc. 2022;42(4):107–16. https://doi.org/10.1177/02704676221126388.
48. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials. JAMA. 2004;291(20):2457. https://doi.org/10.1001/jama.291.20.2457.
49. Mathieu S. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA. 2009;302(9):977. https://doi.org/10.1001/jama.2009.1242.
50. Goldacre B, Drysdale H, Powell-Smith A, Dale A, Milosevic I, Slade E, et al. The COMPare Trials Project. 2016. www.COMPare-trials.org. Accessed 10 June 2023.
51. Schwab S, Janiaud P, Dayan M, Amrhein V, Panczak R, Palagi PM, et al. Ten simple rules for good research practice. PLoS Comput Biol. 2022;18(6):1–14. https://doi.org/10.1371/journal.pcbi.1010139.
52. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):1–7.
53. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010;21(1):128.
54. Vickers AJ, Calster BV, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016:i6. https://doi.org/10.1136/bmj.i6.
55. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KG. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagn Progn Res. 2018;2(1):1–11.
56. Parsons R, Blythe RD, Barnett AG, Cramb SM, McPhail SM. predictNMB: an R package to estimate if or when a clinical prediction model is worthwhile. J Open Source Softw. 2023;8(84):5328. https://doi.org/10.21105/joss.05328.
57. Stark PB, Saltelli A. Cargo-cult statistics and scientific crisis. Significance. 2018;15(4):40–3. https://doi.org/10.1111/j.1740-9713.2018.01174.x.
58. Christian K, ann Larkins J, Doran MR. We must improve conditions and options for Australian ECRs. Nat Hum Behav. 2023. https://doi.org/10.1038/s41562-023-01621-w.
59. Gopalakrishna G, ter Riet G, Vink G, Stoop I, Wicherts JM, Bouter LM. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: a survey among academic researchers in The Netherlands. PLoS ONE. 2022;17(2):1–16. https://doi.org/10.1371/journal.pone.0263023.
60. Wang MQ, Yan AF, Katz RV. Researcher requests for inappropriate analysis and reporting: a U.S. survey of consulting biostatisticians. Ann Intern Med. 2018;169(8):554. https://doi.org/10.7326/m18-1230.

Publisher's Note

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