Risk and Uncertainty Communication in Deployed AI-based Clinical Decision Support Systems: A scoping review

Nicholas Gray[1,2,*], Helen Page[1,2], Iain Buchan[2], Dan W. Joyce[1,2]

[1]Institute of Population Health
[2]Civic Health Innovation Labs
University of Liverpool, Liverpool, UK
*Corresponding author: ngg@liverpool.ac.uk

Abstract

Clinical decision support systems (CDSS) employing data-driven technology such as artificial intelligence, machine- and statistical-learning are increasingly deployed in healthcare settings. These systems often provide clinicians with diagnostic, prognostic, or risk scores modelled from curated patient-level data and frequently involve iterative and non-deterministic optimisation of flexible, parameterised models. All of these data and algorithms have uncertainties associated with them that should be taken into account when used to support clinical decisions at the patient level. This scoping review aims to describe the literature on how deployed data-driven CDSSs present information about uncertainty to their intended users. We describe common clinical applications of CDSSs, characterise the decisions that are being supported, and examine how the CDSS provides outputs to end users, including uncertainty at the individual patient level, as well as indirect measures such as CDSS performance against a reference standard. We conclude with a discussion and recommendations on how CDSS development can be improved.

1 Introduction

Clinical decision support systems (CDSSs) are important tools in modern healthcare used to augment clinicians' decision-making processes [143]. With the rapid advance in data-driven technologies – notably, artificial intelligence (AI) and machine learning (ML) alongside more traditional statistical learning – their incorporation into CDSSs is frequently described as having the potential to revolutionise healthcare by augmenting human decision-making to improve diagnostic accuracy and personalised treatment.

The outputs of these systems, often probabilistic in nature, encapsulate various forms of uncertainty [76]; but there is ambiguity in how risk data and different model predictions are presented and this can be confusing for the end user. For example, QRISK3[1] [66] calculates the risk of a patient having a heart attack or stroke over the next 10 years.

[1]https://www.qrisk.org/

QRISK3 uses 21 predictors and outputs the resulting risk of an event as a percentage and with a natural frequency style expression[2]. The same natural frequency representation is presented visually as an icon array showing the risk of an event alongside a comparison to a "healthy" individual (of the same age, ethnicity and sex) delivered as a numerical relative risk or risk ratio (i.e. the ratio of the patient and the "healthy" person's QRISK score). These methods of communicating risk need to be clear to the end user of the tool as different expressions of uncertainty may alter comprehension, immediate decisions, and consequent actions.

Incomplete medical knowledge, which affects the labelling of data [32] used to develop CDSS algorithms, or the complexity of interacting and comorbid diagnoses/conditions [68], can also introduce uncertainty into an algorithm. Accurately representing and effectively communicating this uncertainty in CDSS algorithms is crucial, as it should influence clinical decision-making and may affect clinician and patient trust in AI recommendations [19, 59, 123]. There are numerous approaches that could be used to express the uncertainty of a clinical outcome all of which have potential benefits and drawbacks, although there is little consensus as to what the best approach would be [137].

The increasing accessibility of programming tools that can embed prediction models in clinical workflows has made creating data-driven clinical decision support systems relatively straightforward. The first step is to pick a clinical decision that the tool will support and then, locating and curating the relevant data and choosing a suitable algorithm to be used to fit the model to the data. Once this has been completed the model needs to be evaluated and, hopefully, it can be concluded that the clinical decision support tool is appropriate for use. The final step is the deployment of the tool, requiring constructing a user interface so that the tool can be used fluently in clinical practice. Then the model needs to be maintained as many lose performance with drifts in population health and data [65]. This review focuses on deployed systems as it is assumed that the creators have had to at least consider the communication of the tool's outputs to the end-user.

This scoping review aims to explore the representation of uncertainty in *deployed* data-driven and AI-based clinical decision support systems. By systematically examining the existing literature, we seek to identify current practices, highlight challenges, and propose directions for future research.

1.1 Defining deployed AI based Clinical decision support systems

For the purposes of this scoping review, we will make use of the following three definitions.

Definition 0. *A clinical decision support system is based on an algorithm designed to aid a medical decision or augment the decision making process where a human could not be reasonably expected to perform the calculations using the same data manually.*[3]

Definition 1. *A clinical decision support system is considered to be* ***AI-based*** *if it uses an algorithm that requires learning from some training data (e.g. Logistic regression, neural networks, etc).*[4]

[2]These are of the style: "In other words, in a crowd of 100 people with the same risk factors as you, N are likely to have a heart attack or stroke within the next 10 years."

[3]For example, a computerised version of the PHQ9 questionnaire [82] that simply reports the total score for screening would be excluded since it could be calculated manually and on it's own, a computer-based implementation of PHQ9 is not prognostic.

[4]As opposed to an algorithm that been derived from pre-existing logical rules.

Definition 2. *A clinical decision support system is considered to be* ***deployed*** *if any of the following are true:*

(a) The tool is being used within clinical practice,

(b) The tool has been validated by use in clinical practice, or

(c) The authors of the tool have made it publicly accessible.

Hereafter we will use the acronym CDSS to refer to deployed AI clinical decision support tools.

2 Methods

2.1 Protocol and registration

This scoping review's protocol has been published and is available on the Open Science Framework website [55].

2.2 Inclusion/Exclusion criteria

To be included in the scoping review, papers needed to present a deployed AI-CDSS (meeting definitions 1 and 2). Papers need to have been published in a peer reviewed journal before 31st March 2024 (with no start date) and written in English. Papers were excluded if they presented CDSSs that were clearly not related to medicine or healthcare. Tutorial, commentary, perspective, discussion and literature review papers were also excluded from the review. Where the search returned two papers that describe the same CDSS only one has been included within the review[5]. This may be due to authors publishing 'development of' and 'evaluation of' papers for the same CDSS[6] or independent groups publishing evaluations of pre-existing CDSSs[7]. In these cases the origin paper, that is the paper that first described the CDSS, was found unless it was possible to extract all the required information from the evaluation paper. Some CDSS have been iteratively refined over time, meaning that where a more recent or updated version of an original tool was located, the older paper was rejected [8].

2.3 Search

The following bibliographic databases were searched: Pubmed, Web of Science and IEEE Xplore. The final searches were carried out in April 2024. The search terms are shown the Appendix A. PubMed was accessed using the Biopython library [29]. The PubMed and IEEE results were filtered post-search to exclude papers that were published after 1st April 2024.

[5]Very often this would be the development paper even though it itself might not be in scope as it fails to meet definition 2

[6]Jauk et al. present two papers of their delirium prediction model, in [71] the model is introduced whereas Jauk et al. 70 qualitative explores the use if the algorithm within a clinical setting. As such, only [71] is included within this review.

[7]For example, [6] independently assess the performance of EuroScore II in patients with structural deterioration of aortic bioprostheses.

[8]For example, EuroSCORE II [105] was included over the original EuroSCORE model [104]

2.4 Screening

After removing duplicates, papers were subjected to title and abstract screening against the inclusion/exclusion criteria yielding a set of papers eligible for full-text review and evaluation with respect to the inclusion/exclusion criteria. During full-text review, additional papers were identified from reviewed papers' bibliographies ("snowballing" or citation tracing) and any papers that presented CDSSs that were also in another paper were removed and the origin paper prefered. In total, we located 130 papers for data extraction.

2.5 Data Extraction

Data extraction proceeded by reviewing the full-text of each paper against four key questions (the detailed criteria for each key question are given in the Appendix B):

QS0– Meta information about the paper, including year of publication, the authors' or their institution's geographical region, the clinical specialty of the CDSS and its intended use.

QS1– What algorithms / methods were used and what output was produced by the CDSS?

QS2– Is uncertainty considered and presented and if so, how is uncertainty presented?

QS3– How is the performance of the CDSS assessed?

Whilst not directly relevent to the presentation of uncertainty from deployed CDSSs, it is still important to consider how the performance of the models are assessed. For models that present high accuracy, people may derive reassurance that the system is reliable (by implication, certain) by looking at its summary performance. This may be misleading as there still can be significant uncertainty about a patients diagnosis, for instance because of the prevalence of the condition or because of the uncertainty associated with the creation of the model [124, 125]. This is similar to the *base rate fallacy* in traditional medical testing, where despite the fact that test can have excellent accuracy, there can still be significant uncertainty about whether or someone has a disease after a positive test [46]. We were also interested to explore how, for CDSSs that do present non-probabilistic uncertainty (for example "don't know' classifications), the whether performance of the uncertain outputs are explicitly assessed.

3 Results

3.1 Search Results

The search returned a total of 3897 papers across the three bibliographic databases. After removing 538 duplicates, the title and abstracts of 3359 papers were screened for inclusion. After abstract screening 306 papers were carried forward for full paper review. During the full text review 36 papers were discovered through the snow-balling process and 8 origin papers were used to extract the required data about the models. In total 350 papers where subjected to full text review, of which, 130 papers were included in this scoping review and data extraction. A PRISMA diagram is shown in figure 1. The reasons for exclusion are shown in figure 2. 108 papers were excluded as they did not

present a model that met the deployed definition in definition 2. Another 51 did not meet the definition of AI (definition 1) or CDSS (definition 0). Finally, 32 papers were deemed out of scope for other reasons and 3 were not accessible to the authors.

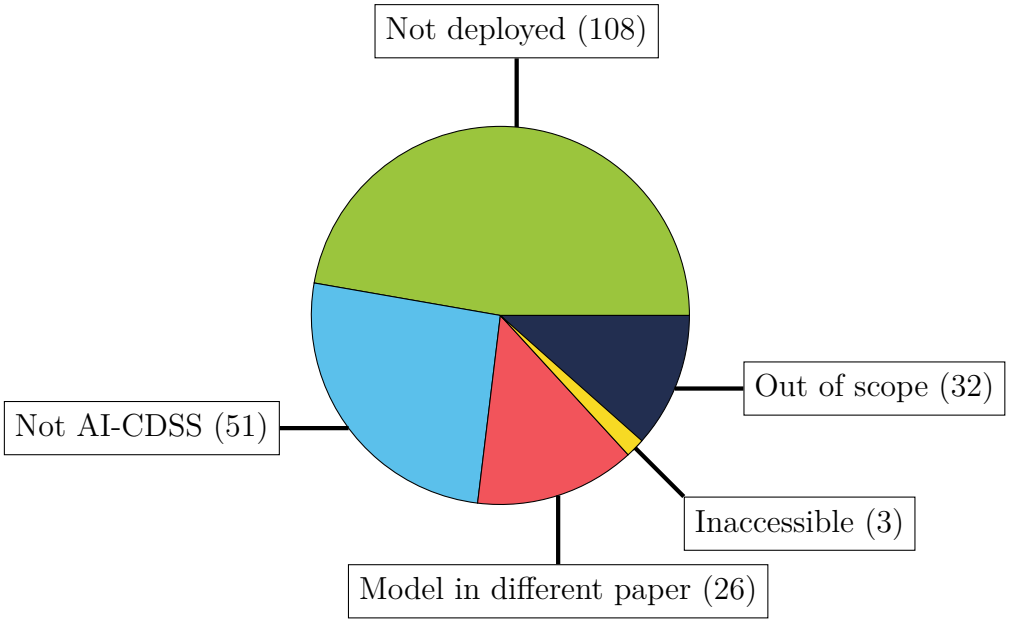

Figure 2: Reasons for exclusion

3.2 Characteristics of included studies

The number of deployed CDSSs (meeting our definition) has been increasing since the mid 2010s as shown in figure 3. The models found in the review covered a total of 44 different medical specialities, as shown in figure 4 and Table 1 in the Appendix. Eight of the eleven infectious disease CDSSs specifically concern the diagnosis/prognosis of COVID-19.

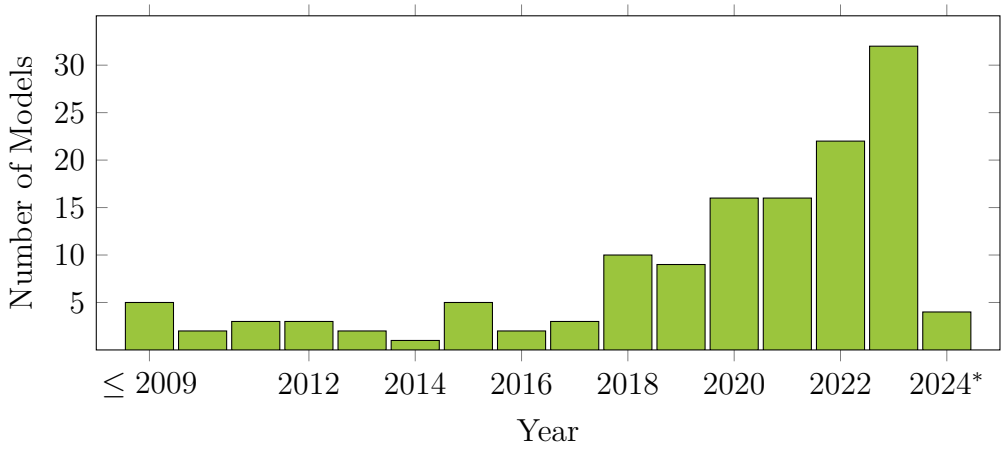

Figure 3: Number of models published by year. The 2024* label includes papers only up to April, and the ≤ 2009 label contains papers from years before 2009.

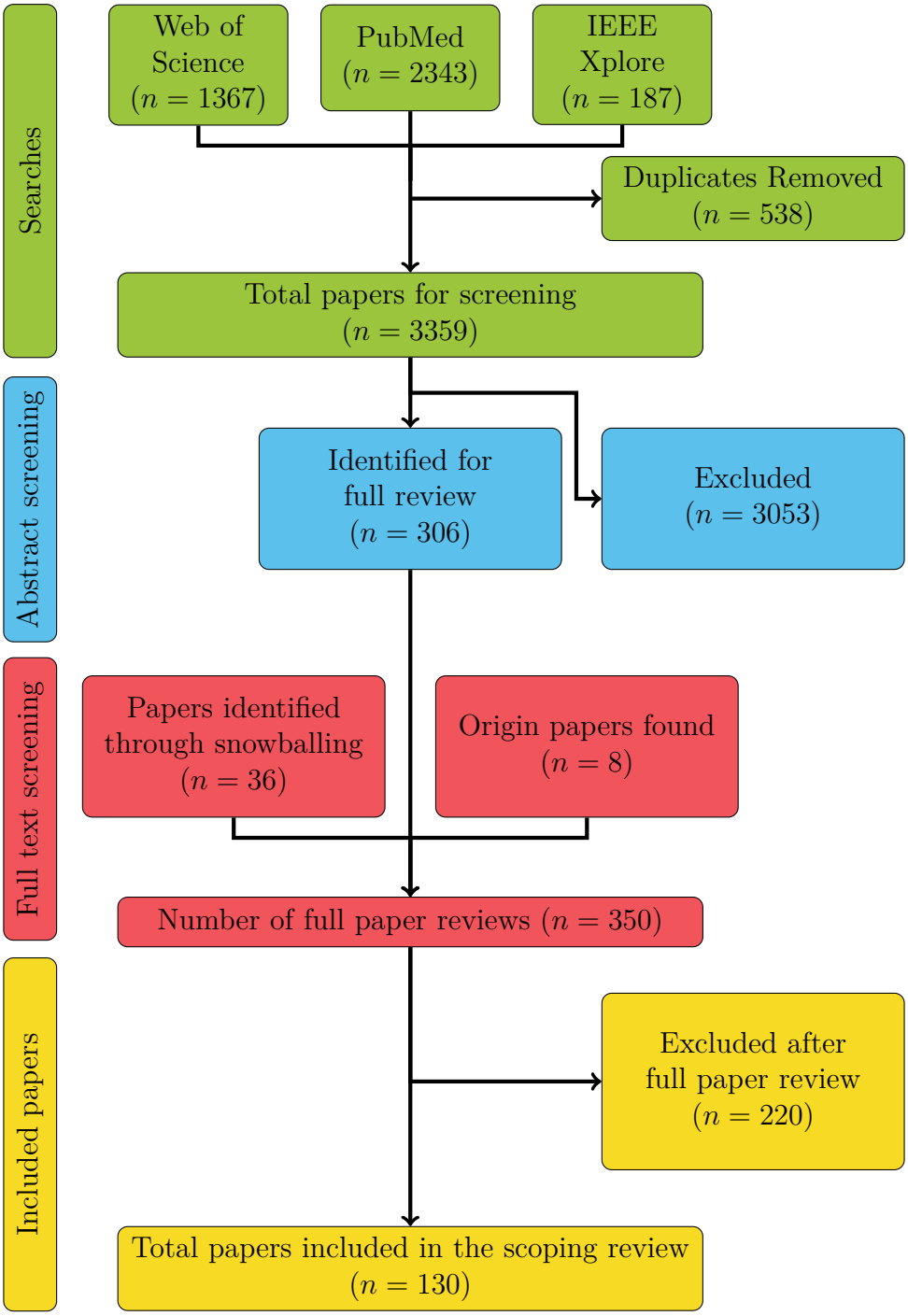

Figure 1: PRISMA diagram.

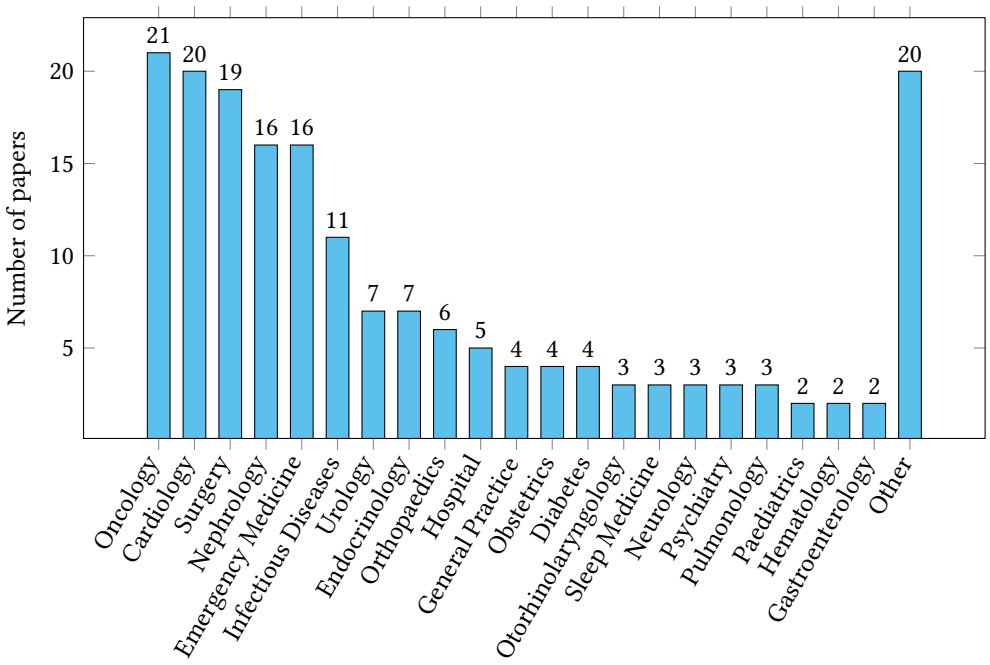

Figure 4: Medical specialities of the included models. The other category includes specialities for which only one model was found. A full list of specialities can be found in Table 1.

Most of the models (126, 93%) are expected to be used by a clinician, with 23 (17%) expecting to be used by a patient with 16 expected to be used by both. Of the remaining models, 3 were for use by administrators [148, 150, 163] and 1 was expected to be used by a carer [131].

The purposes of the CDSSs were varied but fall into several categories, see Figure 5:

- **Prediction/diagnosis of condition** – These models predict a condition. For example, Akbulut et al. [2] predicts fetal health status and Casal-Guisande et al. [18] presents a diagnosis of sleep apnea. The output of many of these CDSSs are analogous to clinical decision making, especially in the case of diagnosis.

- **Risk of condition or clinical outcome** – These models calculate the risk of some condition occurring, sometimes the risk of the condition occurring within a specified time interval. For example, QRISK3 presents the risk of a cardio- or cerebro-vascular event occurring within a 10 year period.

- **Intervention recommendation** – These models assess the likely benefit of an intervention. For example, Lau et al. [86] present a CDSS that assesses the risk/benefit of an encephalitis vaccine and Figueiredo et al. [39] predicts the expected improvement after arthroscopic hip preservation surgery.

- **Triage/Screening** –- Assisting clinicians decide on an assessment and/or treatment service or pathway for a patient. For example, using electronic heath records [37, 41], through 'lifestyle' questions [173] or through medical test results [36].

- **Prediction of outcome after intervention** – A significant number of models, especially those within the surgery domain, predict the risk of adverse outcomes after a procedure or intervention is performed.

- **Monitoring/Management** – This includes papers which predict the likelihood of patients presenting to hospital emergency departments [117, 150], Umscheid et al.'s [151] sepsis early warning systems and Bertoncelli et al.'s [12] epilepsy and seizure detection.

Only Bertoncelli et al. presents a paper that spans multiple catagories (diagnosis and monitoring). In 21 of these papers, death is the outcome that is being predicted. The full list of papers that fall into these catagories are listed in table 2 in Appendix C.2, also listed within the same table are the death-as-outcome papers.

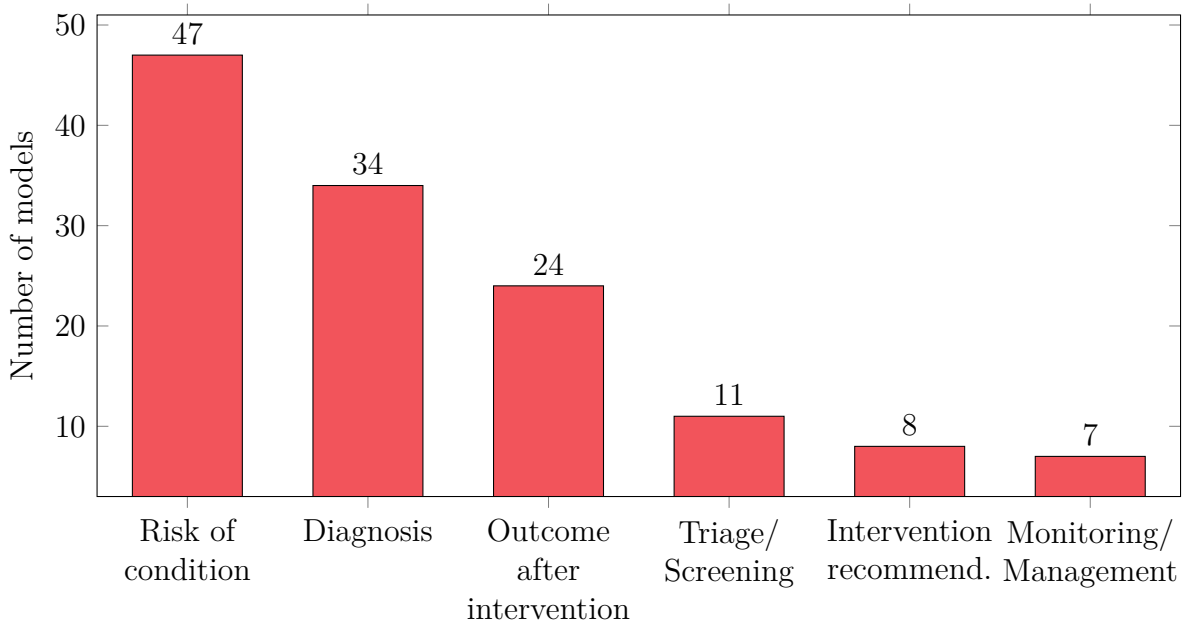

Figure 5: The use cases for the found CDSSs.

3.3 Algorithm

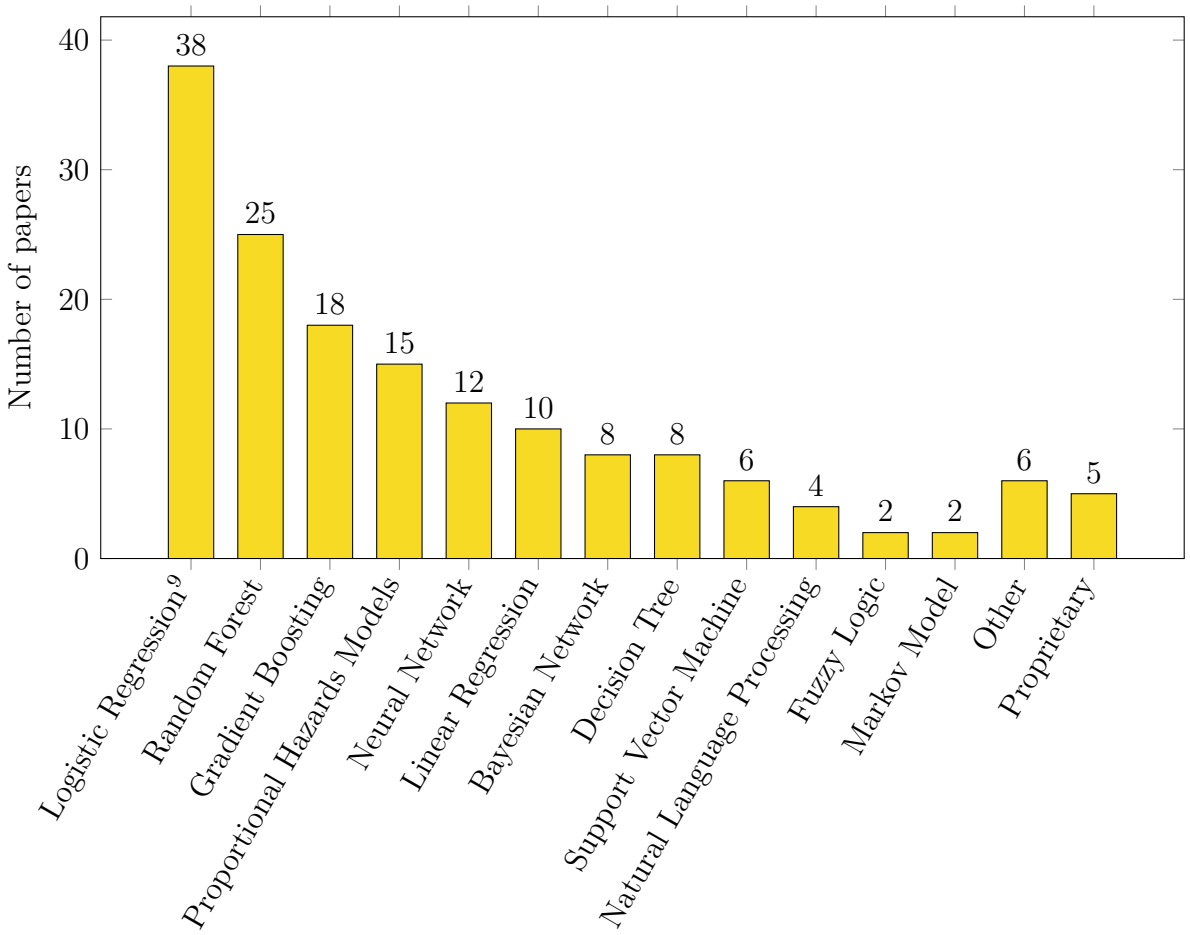

Figure 6: The algorithm used in the CDSS.

The most popular algorithm for deployed CDSSs was logistic regression, although there was a large number of different algorithms used. For 5 models [14, 20, 43, 45, 131] it was not possible toidentify the methodology uses, this was due to the fact that the CDSSs were developed commercially and details were not described in the retrieved papers. The full list of algorithms used by the CDSSs is shown in Table 3.

3.4 Outputs

The majority of the CDSSs (96) present a number as the output of the model and 64 present a classification, with 33 presenting both a number and a classification. Only one CDSS presented a different output, Anand et al. [5] presented questions for a heath care professionals to ask the patients parents (since it the CDSS is for pediatric use) to help determine risk factors (example questions, include "Is [the child] in pain today" or "Does [the child] take perscription or over-the-counter medicine"). For 2 CDSSs it was not possible to find a clear description of the output delivered to the user [163, 164].

[9]For this study, logistic regression includes related methodologies such as multivariate–, multinomial– and bayesian logistic regresion.

Of the 96 papers that presented a number, 71 presented a probability and of the 65 that presented a classification 51 classified the patient into risk levels (i.e. low risk/high risk). Of the classification models, only 35 papers explicitly describe the decision rule including how any classification thresholds were determined. Of these 35, 17 used an arbitrary threshold (e.g. Sun et al. [142] categorises the patient – on the basis of an event probability output – as low risk for $p < 0.5$, medium risk as $0.5 \leq p \leq 0.75$, and high risk as $p > 0.75$). 11 papers determined a threshold by optimising various statistics: for example Huo et al. [69] optimises the decision threshold on the basis of sensitivity and specificity, Ginestra et al. [51] used positive/negative predictive value, Sher et al. [130] used Youden's index and Patterson et al. [116] the number needed to treat. Other statistical methods were used to establish the threshold, for example, two papers [79, 85] based the threshold value of pre-existing values.

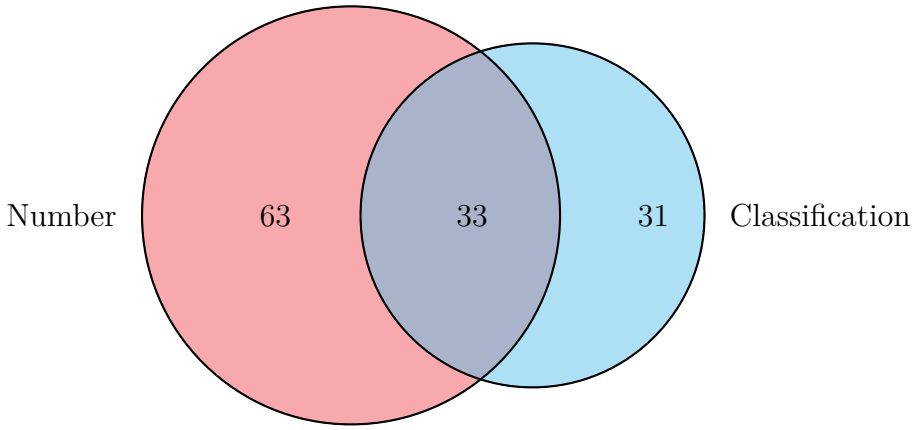

Figure 7: Outputs of the algorithms deployed in CDSSs. Note: there were 3 papers that presented 'other' outputs; Anand et al. [5] presents questions to ask the patient and for [163, 164] it was not possible to determine the output presentation.

3.5 Uncertainty

Of the 130 papers, 90 attempt to express some quantification of uncertainty in the output of the CDSS. [10] The most common representation of uncertainty (68 CDSSs) was to present a simple probability statement of the form probability of X = 0.7 or risk of X = 70%, without any further expressions of uncertainty. Other models presented probabilities using natural frequencies, i.e. 10 out of100 have X [13, 36, 66, 71, 131, 169]. More verbose natural language statements were also used as outputs, for example Xu et al. [169] a statement for the form *Today, in a group of 1000 people like me, 30 will have chlamydia and 970 people will not have chlamydia.* Some papers presented confidence intervals around the probability [42, 140, 156, 172] or continuous score [110, 129].

Some CDSSs presented uncertainty in a visual way; for example, Hippisley-Cox et al. [66] uses an icon array to present the probability of an event, Yu et al. [172] presents a graph of the risk level with error bars, Gardner et al. [45] used various visualisations (including icon arrays) to present risk information.

Many of these graphical outputs use colours to highlight risk levels. For example, Bilimoria et al. [15] presents their results with red for above average risk, yellow for average

[10]The remaining 40 includes papers for which the actual output of the model was u nclear, as such it is not appropriate to say that the do not present any uncertainty quantification.

and green for low risk. Gonçalves et al. [53] uses colours to highlight the confidence of the output, greener denotes smaller error. Some papers present bars that grow, have a moving pointer and/or change colour [See as examples 45, 64, 138, 144, 149, 172]. Another approach used by Dihge et al. [35] and Tseng et al. [149] is to show the full distribution of the esitmated risks and highlight where the patient sits within the distribution.

3.6 Performance

When assessing the performance of the CDSSs, there are two popular methods. The most popular, used by 96 CDSSs, is to use a receiver operating characteristic (ROC) curve and its associated area under the curve (AUC) metric. The second most popular method of assessing performance was to use statistics derived from a confusion matrix[11], 66 CDSSs use this method. There were a variety of different alternatives, including: comparisons to existing methods (such as X or Y), calibration plots of expected v observed risks, Brier score, Hosmer-Lemeshow test. A full list can be found in Table 4. Many of the models use multiple different performance metrics. Figure 8 shows a Venn diagram of the overlap between the (AU)ROC, CMS and the other metrics.

Of the models included in this review, only Kang et al. [78] tested how the (explicitly) outputted uncertainty, in their case a "unsure" prediction, affected the overall decision makeing process. This is acheived by checking whether uncertain predictions by the model reduced the number of false negative cases.

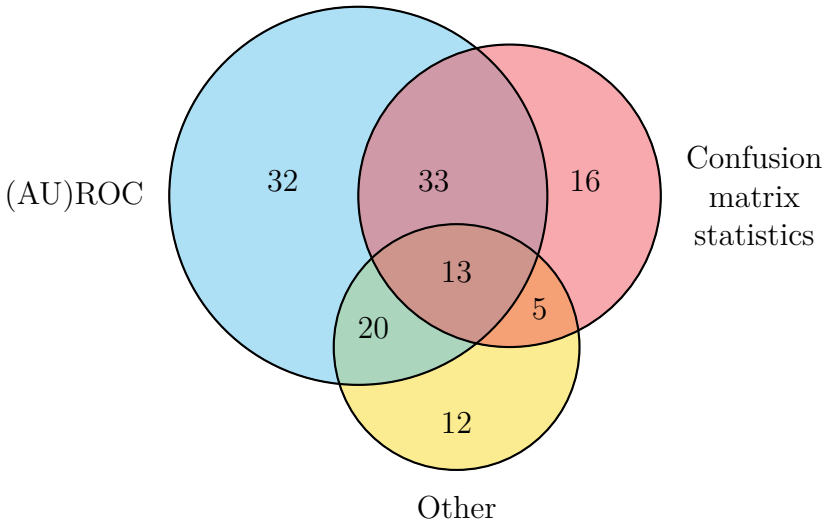

Figure 8: Venn diagram of how the performance of the models are assessed.

4 Discussion[12]

To frame this discussion it is useful to consider what the modal CDSS looks like and how it is may be used in practice. The tool is designed to be used by a clinician to predict the development of a health state or occurrence of a medical event in a given patient. The CDSS outputs a probability (possibly with a low/high risk classification), this probability

[11] Such as accuracy, sensitivity/specificity, precision/recall, F1 score etc

[12] The use of example papers within this section is in no way intended to denigrate particular authors or papers nor to be reflexive of the merits of the CDSS presented within them.

will be presented in the form 0.74 or 74%. The performance of the CDSS will have been validated by measuring its ability to discriminate between classes using a ROC curve and through the creation of a confusion matrix to test accuracy of the binary (high/low) risk classification.

There are two distinct types of uncertainty: the aleatory uncertainty that characterises natural variability and epistemic uncertainty that covers lack of knowledge. In a medical context, aleatory uncertainty answers the question: "*How many patients with similar symptoms/histories have the disease?*", whereas epistemic uncertainty answers *Does this patient have the disease?* When it comes to medical decision making, and for the CDSSs that present results akin to clinical reasoning, is it the latter question that needs answering. Communicating these distinct types of uncertainty via scalar probability values alone is often misunderstood, especially when describing the probability of a single patient having a disease [7, 8, 38, 48, 137, 161].

Improvements can be made by presenting information in ways people find easier to understand, such as using natural frequencies as a more intuitive way for people to understand probabilities [47, 49]. Icon arrays are another approach that can be used to communicate probabilities [44], showing natural frequency information using pictograms, however only a few models made use of these approaches despite the fact that there is a clear benefit to using them. Which approach is best is unclear, and depends on the risk, mathematical and health literacy of the user [115, 137, 157]

Another thing that is important to consider is how these probabilities should be interpreted and very often those models that present probabilities do not make it clear what exactly the probability is. For example, BASH-GN is a CDSS that assesses the risk of a patient having obstructive sleep apnea [69], that has most of the characteristics of the modal CDSS.[13] The output of the model is a naked probability alongside a low/high-risk classification (e.g. 30.9% - `low risk`[14]). It is not entirely clear from the BASH-GN's user interface what exactly this probability refers to, there are several interpretations that could be valid:

1. 31% of patients with similar symptoms suffer from sleep apnea on at least some nights,
2. 31% of nights the patient suffers from sleep apnea, or
3. 31% of the time the patient is asleep, the patient has sleep apnea episodes.

Each of these different interpretations might have a significant impact on doctors or patients decisions about further testing or treatments. The use of natural frequencies or natural language statements can help with this, for example the output could have been "*Out of 100 patients with a similar phenotype, 31 will have sleep apnea on at least some nights*" which make the output clear. However, these approaches may also be misleading as they invite a population average, whereas the key clinical decision is about what is the best decision *this* patient not what is best on average for 100 similar ones.

Many of the CDSSs present results with a high/low risk classification. As we have seen in many cases, the decision rule thresholds for classifications do not arise from a statistical methodology and most decision rules are evaluated for their performance using improper

[13] It is accessible at `https://c2ship.org/bash-gn-metric/`

[14] This result is produced with the following inputs: female, aged 54, neck circumference 24cm, weight 90kg, height 1.56m, with high blood pressure and snoring as loud as talking.

scoring rules [52]. For the modal CDSS discussed within this paper, whatever uncertain nuance is supposed to be characterised by presenting a probability with high/low risk output may be lost–especially when thresholds are set arbitrarily (i.e. without reference to clinically relevant thresholds), where thresholds have been determined from optimising classification performance (e.g. the Yourdon index) or where model-output probability calibration has not been demonstrated in advance of the imposition of a binary decision rule. A doctor might simply interpret high-risk (or a high probability) as though it implies what the correct decision should be. Such classifications are perhaps best left to epidemiologists or public health experts and not viewed as a computer science decision problem. In the framework of statistical decision theory [11, 126], the model output – a probability, or a probability distribution – should be combined with a cost/utility function for the decision to be made and the minimum-cost decision should be preferred. A difficulty with this more rigourous and formal statistical decision theoretic approach is that it is often difficult to 'design' or estimate a cost/utility function [60]. An exception is the use of decision curve analyses [154] where the cost of true-positive is fixed at unity and the relative cost of a false positive can be calculated (for mutually exclusive binary decisions) by deriving the net benefit over a range of threshold probabilities (model outputs).

Very few papers present uncertainty about the probabilities that they present, however this uncertainty certainly exists within all of the models. Very often such uncertainty is viewed as unhelpful at best, however it can be critical to the decision making process. BASH-GN (the CDSS predicting sleep apnea)outputting $\Pr = 31\%$ implies a level of confidence that may be unwarranted. If the algorithm outputted an interval probability, the output $\Pr = [29, 33]\%$ implies that the result is stable and reliable, whereas if the output was $\Pr = [5, 95]\%$, then the vacuousness of this result suggests that the algorithm should not be relied on in the decisions making process.

The most popular approaches to assessing the performance of the models are to use ROC curves or statistics derived from confusion matrices, both of these methods assess the discriminatory performance of the model. ROC curves primarily assess the trade-off between true positive rates and false positive rates across different thresholds, potentially neglecting the calibration of predicted probabilities, which is crucial for risk assessments. They also weigh errors equally and do not give information about the distribution of errors [93]. Confusion matrices, on the other hand, provide counts of true positives, true negatives, false positives, and false negatives based on a fixed threshold, which can oversimplify the performance by not capturing the uncertainty about the predictions. The important question for the user of a risk prediction model is whether an outputted probability of 31% actually implies that 31% of people have the disease. This can be done by plotting the observed vs expected risk or through a statistical means such as Hosmer-Lemeshow test or Brier score.

The use of basic binary discrimination performance metrics says nothing about clinical performance of the CDSS, especially given in most CDSSs there is significant imbalance between the classifications. For example, Cohen et al. [30] present CDSS to predict the someones suicide risk. A false negative on such a CDSS (failing to identify a suicidal individual) has much graver implications than a false positive. ROC plots cannot differentiate between the different impacts of such errors and implicitly assume that they are symmetrically consequential, whilst this is numerically convenience it is clinically nonsensical. Giving equal weighting to false positive and false negative consequences is an example of numerical convenience and clinical nonsense. Therefore, even though Cohen

et al. report AUC = 0.81 (which does indicate good discriminatory performance), the tool may still be suboptimal in clinical practice if it leads to substantial and unmitigated harms.

Often overlooked when discussing uncertainty within CDSS tools is uncertainty about the model itself. Unlike the epistemic and aleatory uncertainties discussed above, this uncertainty is artefactual, it is the result of the exact dataset used, imprecision within the data, and assumptions and decisions made within the creation of the model itself. Almost all the CDSSs we investigated did not consider this as an important source of uncertainty, instead presenting a single, middle-of-the-road, model. However, there are many different CDSSs that could have been fitted from the same data [139]. Whilst some papers do optimise the data cleaning process, the selection of algorithms and hyperparameters so that the final CDSS is optimal, it should be noted that different decisions in each of these stages can lead to models that produce significantly different results [124]. Models can be highly unstable–implying large model uncertainty–especially if there is limited If the model creation process is highly unreliable, is is undesirable if for different patients a clinical decision might be made solely as an artefact of the model creation process [125]. It also needs to be acknowledged that a patient presentation is a unique instance and that their future will be influenced by a unique set of countless environmental, physiological and psychological factors in constant interplay, ergo using a single dataset to inform decision making is flawed [106, 107].

As much of the attention of AI research (and popular discourse) has moved onto the potential of large language models (LLMs) to aid in diagnosis [134]. It is critical that LLM-based CDSSs are able to correctly handle uncertainty to ensure that they do not produce factually inaccurate or harmful statements. This can can be achievied by expressing confidence about the prediction enabling users to defer to other information sources or experts when needed. This is an artefactual uncertainty, resulting from the knowledge based, training data and model parameters of the LLM, and must be treated differently to the aleatory and epistemic uncertainty discussed above. Although LLMs do have the advantage of using natural language to be able to communicate this [90]. It is also worth remembering that the ability of LLMs to pass medical exams (See Jung et al. [75] as an example), says nothing of their ability to perform these talks in the real world.

5 Conclusion

The promise of the increasing use of AI within medicine is that better decisions will be made sooner for (and with) patients. The downside of such an ambition is the risk that these tools are used to enable doctors to do less work whilst shifting accountability onto black-box decision making processes that purport to be evidence based. Care needs to be taken to ensure that the outputs of such CDSSs are appropriately understood, especially in a risk context. When it comes to probabilistic outputs Further research needs to be conducted to establish how best to achieve this.

Reporting protocols for AI in medicine, including TRIPOD+AI [31], CLAIM [100] and DECIDE-AI [153], do not uniformly consider either the uncertainty associated with the outputs of a model nor how this is presented in a CDSS. This is bacause most of the protocols focus upon model training/development and testing/validation and not on the deployment. The informatics of end-to-end CDSS development and deployed optimisation is a more complex problem involving risk/uncertainty communication, decision theory and

human-computer interaction considerations.

The development of CDSSs needs to not be seen as a tournament of algorithms competing to be the most empirically correct, but as a part of system to improve the medical decision making process in general. Careful communication of risk and uncertainty is a key component of that process.

References

[1] Joanna Abraham, Brian Bartek, Alicia Meng, Christopher Ryan King, Bing Xue, Chenyang Lu, and Michael S. Avidan. 2023. Integrating Machine Learning Predictions for Perioperative Risk Management: Towards an Empirical Design of a Flexible-Standardized Risk Assessment Tool. *Journal of Biomedical Informatics* 137 (Jan. 2023), 104270. https://doi.org/10.1016/j.jbi.2022.104270

[2] Akhan Akbulut, Egemen Ertugrul, and Varol Topcu. 2018. Fetal Health Status Prediction Based on Maternal Clinical History Using Machine Learning Techniques. *Computer Methods and Programs in Biomedicine* 163 (Sept. 2018), 87–100. https://doi.org/10.1016/j.cmpb.2018.06.010

[3] Yagiz Aksoy, Angela Chou, Mahiar Mahjoub, Amy Sheen, Loretta Sioson, Mahsa S. Ahadi, Anthony J. Gill, and Talia L. Fuchs. 2023. A Novel Prognostic Nomogram for Predicting Survival in Diffuse Pleural Mesothelioma. *Pathology* 55, 4 (June 2023), 449–455. https://doi.org/10.1016/j.pathol.2022.11.009

[4] Salem Alkaabi, Asma Alnuaimi, Mariam Alharbi, Mohammed A Amari, Rajiv Ganapathy, Imran Iqbal, Javaid Nauman, and Abderrahim Oulhaj. 2021. A Clinical Risk Score to Predict In-Hospital Mortality in Critically Ill Patients with COVID-19: A Retrospective Cohort Study. *BMJ Open* 11, 8 (Aug. 2021), e048770. https://doi.org/10.1136/bmjopen-2021-048770

[5] Vibha Anand, Aaron E. Carroll, and Stephen M. Downs. 2012. Automated Primary Care Screening in Pediatric Waiting Rooms. *Pediatrics* 129, 5 (May 2012), e1275–e1281. https://doi.org/10.1542/peds.2011-2875

[6] Amedeo Anselmi, Vito Giovanni Ruggieri, Majid Harmouche, Sophie Mascle, Vincent Auffret, Hervé Le Breton, Xavier Beneux, and Jean-Philippe Verhoye. 2015. Is the EuroSCORE II Best Suited for Reoperative Risk Estimation in Patients with Structural Deterioration of Aortic Bioprostheses? *Medical Hypotheses* 84, 5 (May 2015), 470–473. https://doi.org/10.1016/j.mehy.2015.01.043

[7] Terje Aven. 2010. On the Need for Restricting the Probabilistic Analysis in Risk Assessments to Variability. *Risk Analysis* 30, 3 (March 2010), 354–360. https://doi.org/10.1111/j.1539-6924.2009.01314.x

[8] Terje Aven. 2023. Risk Literacy: Foundational Issues and Its Connection to Risk Science. *Risk Analysis* 44, 5 (Sept. 2023), risa.14223. https://doi.org/10.1111/risa.14223

[9] Gang Bai, Zhonglin Cai, Xiuxia Zhai, Jian Xiong, Fa Zhang, and Hongjun Li. 2021. A New Nomogram for the Prediction of Bone Metastasis in Patients with

Prostate Cancer. *Journal of International Medical Research* 49, 11 (Nov. 2021), 030006052110583. https://doi.org/10.1177/03000605211058364

[10] Sean J. Barbour, Rosanna Coppo, Hong Zhang, Zhi-Hong Liu, Yusuke Suzuki, Keiichi Matsuzaki, Ritsuko Katafuchi, Lee Er, Gabriela Espino-Hernandez, S. Joseph Kim, Heather N. Reich, John Feehally, Daniel C. Cattran, and International IgA Nephropathy Network. 2019. Evaluating a New International Risk-Prediction Tool in IgA Nephropathy. *JAMA Internal Medicine* 179, 7 (July 2019), 942. https://doi.org/10.1001/jamainternmed.2019.0600

[11] James O Berger. 2013. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

[12] Carlo M. Bertoncelli, Stefania Costantini, Fabio Persia, Domenico Bertoncelli, and Daniela D'Auria. 2023. PredictMed-epilepsy: A Multi-Agent Based System for Epilepsy Detection and Prediction in Neuropediatrics. *Computer Methods and Programs in Biomedicine* 236 (June 2023), 107548. https://doi.org/10.1016/j.cmpb.2023.107548

[13] Dimitris Bertsimas, Jack Dunn, George C. Velmahos, and Haytham M. A. Kaafarani. 2018. Surgical Risk Is Not Linear: Derivation and Validation of a Novel, User-friendly, and Machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) Calculator. *Annals of Surgery* 268, 4 (Oct. 2018), 574–583. https://doi.org/10.1097/SLA.0000000000002956

[14] Azra Bihorac, Tezcan Ozrazgat-Baslanti, Ashkan Ebadi, Amir Motaei, Mohcine Madkour, Panagote M. Pardalos, Gloria Lipori, William R. Hogan, Philip A. Efron, Frederick Moore, Lyle L. Moldawer, Daisy Zhe Wang, Charles E. Hobson, Parisa Rashidi, Xiaolin Li, and Petar Momcilovic. 2019. MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery. *Annals of Surgery* 269, 4 (April 2019), 652–662. https://doi.org/10.1097/SLA.0000000000002706

[15] Karl Y. Bilimoria, Yaoming Liu, Jennifer L. Paruch, Lynn Zhou, Thomas E. Kmiecik, Clifford Y. Ko, and Mark E. Cohen. 2013. Development and Evaluation of the Universal ACS NSQIP Surgical Risk Calculator: A Decision Aid and Informed Consent Tool for Patients and Surgeons. *Journal of the American College of Surgeons* 217, 5 (Nov. 2013), 833–842e3. https://doi.org/10.1016/j.jamcollsurg.2013.07.385

[16] Jesse Bittman, Penny Tam, Chris Little, and Nadia Khan. 2017. Who to Handover: A Case–Control Study of a Novel Scoring System to Prioritise Handover of Internal Medicine Inpatients. *Postgraduate Medical Journal* 93, 1100 (June 2017), 313–318. https://doi.org/10.1136/postgradmedj-2016-133999

[17] Jaume Canet, Lluís Gallart, Carmen Gomar, Guillem Paluzie, Jordi Vallès, Jordi Castillo, Sergi Sabaté, Valentín Mazo, Zahara Briones, Joaquín Sanchis, and on behalf of the ARISCAT Group. 2010. Prediction of Postoperative Pulmonary Complications in a Population-based Surgical Cohort. *Anesthesiology* 113, 6 (Dec. 2010), 1338–1350. https://doi.org/10.1097/ALN.0b013e3181fc6e0a

[18] Manuel Casal-Guisande, Laura Ceide-Sandoval, Mar Mosteiro-Añón, María Torres-Durán, Jorge Cerqueiro-Pequeño, José-Benito Bouza-Rodríguez, Alberto Fernández-Villar, and Alberto Comesaña-Campos. 2023. Design of an Intelligent Decision Support System Applied to the Diagnosis of Obstructive Sleep Apnea. *Diagnostics* 13, 11 (May 2023), 1854. https://doi.org/10.3390/diagnostics13111854

[19] Micah Cearns, Tim Hahn, Scott Clark, and Bernhard T Baune. 2020. Machine Learning Probability Calibration for High-Risk Clinical Decision-Making. *Australian & New Zealand Journal of Psychiatry* 54, 2 (Feb. 2020), 123–126. https://doi.org/10.1177/0004867419885448

[20] B. Chakrabarti, B. Kane, C. Barrow, J. Stonebanks, L. Reed, M. G. Pearson, L. Davies, M. Osborne, P. England, D. Litchfield, E. McKnight, and R. M. Angus. 2023. The Feasibility and Impact of Implementing a Computer-Guided Consultation to Target Health Inequality in Asthma. *npj Primary Care Respiratory Medicine* 33, 1 (Feb. 2023), 6. https://doi.org/10.1038/s41533-023-00329-8

[21] Eric Chalmers, Lindsey Westover, Johith Jacob, Andreas Donauer, Vicky H. Zhao, Eric C. Parent, Marc J. Moreau, James K. Mahood, Douglas M. Hedden, and Edmond H. M. Lou. 2015. Predicting Success or Failure of Brace Treatment for Adolescents with Idiopathic Scoliosis. *Medical & Biological Engineering & Computing* 53, 10 (Oct. 2015), 1001–1009. https://doi.org/10.1007/s11517-015-1306-7

[22] Zijun Chen, Tingming Li, Sheng Guo, Deli Zeng, and Kai Wang. 2023. Machine Learning-Based in-Hospital Mortality Risk Prediction Tool for Intensive Care Unit Patients with Heart Failure. *Frontiers in Cardiovascular Medicine* 10 (April 2023), 1119699. https://doi.org/10.3389/fcvm.2023.1119699

[23] Fu-Yuan Cheng, Himanshu Joshi, Pranai Tandon, Robert Freeman, David L Reich, Madhu Mazumdar, Roopa Kohli-Seth, Matthew A. Levin, Prem Timsina, and Arash Kia. 2020. Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients. *Journal of Clinical Medicine* 9, 6 (June 2020), 1668. https://doi.org/10.3390/jcm9061668

[24] Sharon Chiang, Marina Vannucci, Daniel M. Goldenholz, Robert Moss, and John M. Stern. 2018. Epilepsy as a Dynamic Disease: A Bayesian Model for Differentiating Seizure Risk from Natural Variability. *Epilepsia Open* 3, 2 (June 2018), 236–246. https://doi.org/10.1002/epi4.12112

[25] Insook Cho, Jiseon Cho, Jeong Hee Hong, Wha Suk Choe, and HyeKyeong Shin. 2023. Utilizing Standardized Nursing Terminologies in Implementing an AI-powered Fall-Prevention Tool to Improve Patient Outcomes: A Multihospital Study. *Journal of the American Medical Informatics Association* 30, 11 (Oct. 2023), 1826–1836. https://doi.org/10.1093/jamia/ocad145

[26] Lisa S Chow, Rachel Zmora, Sisi Ma, Elizabeth R Seaquist, and Pamela J Schreiner. 2018. Development of a Model to Predict 5-Year Risk of Severe Hypoglycemia in Patients with Type 2 Diabetes. *BMJ Open Diabetes Research & Care* 6, 1 (Aug. 2018), e000527. https://doi.org/10.1136/bmjdrc-2018-000527

[27] Esther H. Chung, Chaitanya R. Acharya, Benjamin S. Harris, and Kelly S. Acharya. 2021. Development of a Fertility Risk Calculator to Predict Individualized Chance of Ovarian Failure after Chemotherapy. *Journal of Assisted Reproduction and Genetics* 38, 11 (Nov. 2021), 3047–3055. https://doi.org/10.1007/s10815-021-02311-0

[28] Matthew M. Churpek, Kyle A. Carey, Dana P. Edelson, Tripti Singh, Brad C. Astor, Emily R. Gilbert, Christopher Winslow, Nirav Shah, Majid Afshar, and Jay L. Koyner. 2020. Internal and External Validation of a Machine Learning Risk Score for Acute Kidney Injury. *JAMA Network Open* 3, 8 (Aug. 2020), e2012892. https://doi.org/10.1001/jamanetworkopen.2020.12892

[29] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. De Hoon. 2009. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* 25, 11 (June 2009), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

[30] Joshua Cohen, Jennifer Wright-Berryman, Lesley Rohlfs, Douglas Trocinski, LaMonica Daniel, and Thomas W. Klatt. 2022. Integration and Validation of a Natural Language Processing Machine Learning Suicide Risk Prediction Model Based on Open-Ended Interview Language in the Emergency Department. *Frontiers in Digital Health* 4 (Feb. 2022), 818705. https://doi.org/10.3389/fdgth.2022.818705

[31] Gary S Collins, Karel G M Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten Van Smeden, Anne-Laure Boulesteix, Jennifer Catherine Camaradou, Leo Anthony Celi, Spiros Denaxas, Alastair K Denniston, Ben Glocker, Robert M Golub, Hugh Harvey, Georg Heinze, Michael M Hoffman, André Pascal Kengne, Emily Lam, Naomi Lee, Elizabeth W Loder, Lena Maier-Hein, Bilal A Mateen, Melissa D McCradden, Lauren Oakden-Rayner, Johan Ordish, Richard Parnell, Sherri Rose, Karandeep Singh, Laure Wynants, and Patricia Logullo. 2024. TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *BMJ* 385 (April 2024), e078378. https://doi.org/10.1136/bmj-2023-078378

[32] John Collins and Paul S. Albert. 2016. Estimating Diagnostic Accuracy without a Gold Standard: A Continued Controversy. *Journal of Biopharmaceutical Statistics* 26, 6 (Nov. 2016), 1078–1082. https://doi.org/10.1080/10543406.2016.1226334

[33] Ashis Kumar Das, Shiba Mishra, and Saji Saraswathy Gopalan. 2020. Predicting CoVID-19 Community Mortality Risk Using Machine Learning and Development of an Online Prognostic Tool. *PeerJ* 8 (Sept. 2020), e10083. https://doi.org/10.7717/peerj.10083

[34] Mahsa Dehghani Soufi, Taha Samad-Soltani, Samad Shams Vahdati, and Peyman Rezaei-Hachesu. 2018. Decision Support System for Triage Management: A Hybrid Approach Using Rule-Based Reasoning and Fuzzy Logic. *International Journal of Medical Informatics* 114 (June 2018), 35–44. https://doi.org/10.1016/j.ijmedinf.2018.03.008

[35] Looket Dihge, Pär-Ola Bendahl, Ida Skarping, Malin Hjärtström, Mattias Ohlsson, and Lisa Rydén. 2023. The Implementation of NILS: A Web-Based Artificial Neural Network Decision Support Tool for Noninvasive Lymph Node Staging in Breast Cancer. *Frontiers in Oncology* 13 (March 2023), 1102254. https://doi.org/10.3389/fonc.2023.1102254

[36] Yuhan Du, Anthony R. Rafferty, Fionnuala M. McAuliffe, Lan Wei, and Catherine Mooney. 2022. An Explainable Machine Learning-Based Clinical Decision Support System for Prediction of Gestational Diabetes Mellitus. *Scientific Reports* 12, 1 (Jan. 2022), 1170. https://doi.org/10.1038/s41598-022-05112-2

[37] R Scott Evans, Jose Benuzillo, Benjamin D Horne, James F Lloyd, Alejandra Bradshaw, Deborah Budge, Kismet D Rasmusson, Colleen Roberts, Jason Buckway, Norma Geer, Teresa Garrett, and Donald L Lappé. 2016. Automated Identification and Predictive Tools to Help Identify High-Risk Heart Failure Patients: Pilot Evaluation. *Journal of the American Medical Informatics Association* 23, 5 (Sept. 2016), 872–878. https://doi.org/10.1093/jamia/ocv197

[38] Scott Ferson and Lev R Ginzburg. 1996. Different Methods Are Needed to Propagate Ignorance and Variability. *Reliability Engineering & System Safety* 54, 2–3 (1996), 21.

[39] Flávio De Azevedo Figueiredo, Lucas Emanuel Ferreira Ramos, Rafael Tavares Silva, Daniela Ponce, Rafael Lima Rodrigues De Carvalho, Alexandre Vargas Schwarzbold, Amanda De Oliveira Maurílio, Ana Luiza Bahia Alves Scotton, Andresa Fontoura Garbini, Bárbara Lopes Farace, Bárbara Machado Garcia, Carla Thais Cândida Alves Da Silva, Christiane Corrêa Rodrigues Cimini, Cíntia Alcantara De Carvalho, Cristiane Dos Santos Dias, Daniel Vitório Silveira, Euler Roberto Fernandes Manenti, Evelin Paola De Almeida Cenci, Fernando Anschau, Fernando Graça Aranha, Filipe Carrilho De Aguiar, Frederico Bartolazzi, Giovanna Grunewald Vietta, Guilherme Fagundes Nascimento, Helena Carolina Noal, Helena Duani, Heloisa Reniers Vianna, Henrique Cerqueira Guimarães, Joice Coutinho De Alvarenga, José Miguel Chatkin, Júlia Drumond Parreiras De Morais, Juliana Machado-Rugolo, Karen Brasil Ruschel, Karina Paula Medeiros Prado Martins, Luanna Silva Monteiro Menezes, Luciana Siuves Ferreira Couto, Luís César De Castro, Luiz Antônio Nasi, Máderson Alvares De Souza Cabral, Maiara Anschau Floriani, Maíra Dias Souza, Maira Viana Rego Souza-Silva, Marcelo Carneiro, Mariana Frizzo De Godoy, Maria Aparecida Camargos Bicalho, Maria Clara Pontello Barbosa Lima, Márlon Juliano Romero Aliberti, Matheus Carvalho Alves Nogueira, Matheus Fernandes Lopes Martins, Milton Henriques Guimarães-Júnior, Natália Da Cunha Severino Sampaio, Neimy Ramos De Oliveira, Patricia Klarmann Ziegelmann, Pedro Guido Soares Andrade, Pedro Ledic Assaf, Petrônio José De Lima Martelli, Polianna Delfino-Pereira, Raphael Castro Martins, Rochele Mosmann Menezes, Saionara Cristina Francisco, Silvia Ferreira Araújo, Talita Fischer Oliveira, Thainara Conceição De Oliveira, Thaís Lorenna Souza Sales, Thiago Junqueira Avelino-Silva, Yuri Carlotto Ramires, Magda Carvalho Pires, and Milena Soriano Marcolino. 2022. Development and Validation of the MMCD Score to Predict Kidney Replacement Therapy in COVID-19 Patients. *BMC Medicine* 20, 1 (Sept. 2022), 324. https://doi.org/10.1186/s12916-022-02503-0

[40] Diana My Frodi, Vlad Manea, Søren Zöga Diederichsen, Jesper Hastrup Svendsen, Katarzyna Wac, and Tariq Osman Andersen. 2022. Using Consumer-Wearable Activity Trackers for Risk Prediction of Life-Threatening Heart Arrhythmia in Patients with an Implantable Cardioverter-Defibrillator: An Exploratory Observational Study. *Journal of Personalized Medicine* 12, 6 (June 2022), 942. https://doi.org/10.3390/jpm12060942

[41] Emanuele Frontoni, Luca Romeo, Michele Bernardini, Sara Moccia, Lucia Migliorelli, Marina Paolanti, Alessandro Ferri, Paolo Misericordia, Adriano Mancini, and Primo Zingaretti. 2020. A Decision Support System for Diabetes Chronic Care Models Based on General Practitioner Engagement and EHR Data Sharing. *IEEE Journal of Translational Engineering in Health and Medicine* 8 (2020), 1–12. https://doi.org/10.1109/JTEHM.2020.3031107

[42] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill. 1989. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. *JNCI Journal of the National Cancer Institute* 81, 24 (Dec. 1989), 1879–1886. https://doi.org/10.1093/jnci/81.24.1879

[43] Ajeet Gajra, Marjorie E Zettler, Kelly A Miller, Sibel Blau, Swetha S Venkateshwaran, Shreenath Sridharan, John Showalter, Amy W Valley, and John G Frownfelter. 2021. Augmented Intelligence to Predict 30-Day Mortality in Patients with Cancer. *Future Oncology* 17, 29 (Oct. 2021), 3797–3807. https://doi.org/10.2217/fon-2021-0302

[44] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. 2009. Using Icon Arrays to Communicate Medical Risks: Overcoming Low Numeracy. *Health Psychology* 28, 2 (2009), 210–216. https://doi.org/10.1037/a0014474

[45] Clarissa Gardner, Deborah Wake, Doogie Brodie, Alex Silverstein, Sophie Young, Scott Cunningham, Chris Sainsbury, Maria Ilia, Amanda Lucas, Tony Willis, and Jack Halligan. 2023. Evaluation of Prototype Risk Prediction Tools for Clinicians and People Living with Type 2 Diabetes in North West London Using the Think Aloud Method. *DIGITAL HEALTH* 9 (Jan. 2023), 205520762211286. https://doi.org/10.1177/20552076221128677

[46] Gerd Gigerenzer. 2002. *Calculated Risks: How to Know When Numbers Deceive You*. Simon and Schuster, New York, NY, USA.

[47] Gerd Gigerenzer. 2011. What Are Natural Frequencies? *BMJ (Online)* 343, 7828 (2011), 1–2. https://doi.org/10.1136/bmj.d6386

[48] Gerd Gigerenzer, Ralph Hertwig, Eva Van Den Broek, Barbara Fasolo, and Konstantinos V. Katsikopoulos. 2005. "A 30% Chance of Rain Tomorrow": How Does the Public Understand Probabilistic Weather Forecasts? *Risk Analysis* 25, 3 (June 2005), 623–629. https://doi.org/10.1111/j.1539-6924.2005.00608.x

[49] Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological Review,* 102, 4 (1995), 684–704.

[50] Mariano Cesare Giglio, Pasquale Dolce, Sezai Yilmaz, Yaman Tokat, Koray Acarli, Murat Kilic, Murat Zeytunlu, Tarkan Unek, Vincent Karam, René Adam, Wojciech Grzegorz Polak, Constantino Fondevila, Silvio Nadalin, Roberto Ivan Troisi, and for the European Liver and Intestine Transplant Association (ELITA). 2023. Development of a Model to Predict the Risk of Early Graft Failure after Adult-to-Adult Living Donor Liver Transplantation: An ELTR Study. *Liver Transplantation* (Dec. 2023). https://doi.org/10.1097/LVT.0000000000000312

[51] Jennifer C. Ginestra, Heather M. Giannini, William D. Schweickert, Laurie Meadows, Michael J. Lynch, Kimberly Pavan, Corey J. Chivers, Michael Draugelis, Patrick J. Donnelly, Barry D. Fuchs, and Craig A. Umscheid. 2019. Clinician Perception of a Machine Learning–Based Early Warning System Designed to Predict Severe Sepsis and Septic Shock*. *Critical Care Medicine* 47, 11 (Nov. 2019), 1477–1484. https://doi.org/10.1097/CCM.0000000000003803

[52] Tilmann Gneiting and Adrian E Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American statistical Association* 102, 477 (2007), 359–378.

[53] Daniel Gonçalves, Rui Henriques, Lúcio Lara Santos, and Rafael S. Costa. 2021. On the Predictability of Postoperative Complications for Cancer Patients: A Portuguese Cohort Study. *BMC Medical Informatics and Decision Making* 21, 1 (Dec. 2021), 200. https://doi.org/10.1186/s12911-021-01562-2

[54] Maya Gopalakrishnan, Suman Saurabh, Pramod Sagar, Chanaveerappa Bammigatti, and Tarun Kumar Dutta. 2022. A Simple Mortality Risk Prediction Score for Viper Envenoming in India (VENOMS): A Model Development and Validation Study. *PLOS Neglected Tropical Diseases* 16, 2 (Feb. 2022), e0010183. https://doi.org/10.1371/journal.pntd.0010183

[55] Nicholas Gray, Dan W. Joyce, and Helen Page. 2024. The Representation of Uncertainty in AI-aided Clinical Decision Support: A Scoping Review. https://doi.org/10.17605/OSF.IO/GZ4KW

[56] Seokmin Ha, Su Jung Choi, Sujin Lee, Reinatt Hansel Wijaya, Jee Hyun Kim, Eun Yeon Joo, and Jae Kyoung Kim. 2023. Predicting the Risk of Sleep Disorders Using a Machine Learning–Based Simple Questionnaire: Development and Validation Study. *Journal of Medical Internet Research* 25 (Sept. 2023), e46520. https://doi.org/10.2196/46520

[57] David E Hamilton, Jeremy Albright, Milan Seth, Ian Painter, Charles Maynard, Ravi S Hira, Devraj Sukul, and Hitinder S Gurm. 2024. Merging Machine Learning and Patient Preference: A Novel Tool for Risk Prediction of Percutaneous Coronary Interventions. *European Heart Journal* 45, 8 (Feb. 2024), 601–609. https://doi.org/10.1093/eurheartj/ehad836

[58] In Woong Han, Kyeongwon Cho, Youngju Ryu, Sang Hyun Shin, Jin Seok Heo, Dong Wook Choi, Myung Jin Chung, Oh Chul Kwon, and Baek Hwan Cho. 2020. Risk Prediction Platform for Pancreatic Fistula after Pancreatoduodenectomy Using Artificial Intelligence. *World Journal of Gastroenterology* 26, 30 (Aug. 2020), 4453–4464. https://doi.org/10.3748/wjg.v26.i30.4453

[59] Paul K. J. Han. 2013. Conceptual, Methodological, and Ethical Problems in Communicating Uncertainty in Clinical Evidence. *Medical Care Research and Review* 70, 1_suppl (Feb. 2013), 14S–36S. https://doi.org/10.1177/1077558712459361

[60] David J Hand. 2012. Assessing the Performance of Classification Methods. *International Statistical Review* 80, 3 (2012), 400–414.

[61] Alex H. S. Harris, Alfred C. Kuo, Yingjie Weng, Amber W. Trickey, Thomas Bowe, and Nicholas J. Giori. 2019. Can Machine Learning Methods Produce Accurate and Easy-to-use Prediction Models of 30-Day Complications and Mortality After Knee or Hip Arthroplasty? *Clinical Orthopaedics & Related Research* 477, 2 (Feb. 2019), 452–460. https://doi.org/10.1097/CORR.0000000000000601

[62] Alex H. S. Harris, Amber W. Trickey, Hyrum S. Eddington, Carolyn D. Seib, Robin N. Kamal, Alfred C. Kuo, Qian Ding, and Nicholas J. Giori. 2022. A Tool to Estimate Risk of 30-Day Mortality and Complications After Hip Fracture Surgery: Accurate Enough for Some but Not All Purposes? A Study From the ACS-NSQIP Database. *Clinical Orthopaedics & Related Research* 480, 12 (Dec. 2022), 2335–2346. https://doi.org/10.1097/CORR.0000000000002294

[63] Patrick Heindel, Tanujit Dey, James J Fitzgibbon, Muhammad Mamdani, Dirk M Hentschel, Michael Belkin, Charles Keith Ozaki, and Mohamad A Hussain. 2023. Predicting Recurrent Interventions after Radiocephalic Arteriovenous Fistula Creation with Machine Learning and the PREDICT-AVF Web App. *The Journal of Vascular Access* 0, 0 (Dec. 2023), 11297298231203356. https://doi.org/10.1177/11297298231203356

[64] Lindsay N Helget, David J Dillon, Bethany Wolf, Laura P Parks, Sally E Self, Evelyn T Bruner, Evan E Oates, and Jim C Oates. 2021. Development of a Lupus Nephritis Suboptimal Response Prediction Tool Using Renal Histopathological and Clinical Laboratory Variables at the Time of Diagnosis. *Lupus Science & Medicine* 8, 1 (Aug. 2021), e000489. https://doi.org/10.1136/lupus-2021-000489

[65] G. L. Hickey, S. W. Grant, G. J. Murphy, M. Bhabra, D. Pagano, K. McAllister, I. Buchan, and B. Bridgewater. 2013. Dynamic Trends in Cardiac Surgery: Why the Logistic EuroSCORE Is No Longer Suitable for Contemporary Cardiac Surgery and Implications for Future Risk Models. *European Journal of Cardio-Thoracic Surgery* 43, 6 (June 2013), 1146–1152. https://doi.org/10.1093/ejcts/ezs584

[66] Julia Hippisley-Cox, Carol Coupland, and Peter Brindle. 2017. Development and Validation of QRISK3 Risk Prediction Algorithms to Estimate Future Risk of Cardiovascular Disease: Prospective Cohort Study. *BMJ* 357 (May 2017), j2099. https://doi.org/10.1136/bmj.j2099

[67] S.M. Hosseini Sarkhosh, M. Hemmatabadi, and A. Esteghamati. 2022. Development and Validation of a Risk Score for Diabetic Kidney Disease Prediction in Type 2 Diabetes Patients: A Machine Learning Approach. *Journal of Endocrinological Investigation* 46, 2 (Sept. 2022), 415–423. https://doi.org/10.1007/s40618-022-01919-y

[68] David M Hughes, Jose Ignacio Cuitun Coronado, Pieta Schofield, Zenas Z N Yiu, and Sizheng Steven Zhao. 2023. The Predictive Accuracy of Cardiovascular Disease Risk Prediction Tools in Inflammatory Arthritis and Psoriasis: An Observational Validation Study Using the Clinical Practice Research Datalink. *Rheumatology* 00 (Nov. 2023), kead610. https://doi.org/10.1093/rheumatology/kead610

[69] Jiayan Huo, Stuart F. Quan, Janet Roveda, and Ao Li. 2023. BASH-GN: A New Machine Learning–Derived Questionnaire for Screening Obstructive Sleep Apnea. *Sleep and Breathing* 27, 2 (May 2023), 449–457. https://doi.org/10.1007/s11325-022-02629-8

[70] Stefanie Jauk, Diether Kramer, Alexander Avian, Andrea Berghold, Werner Leodolter, and Stefan Schulz. 2021. Technology Acceptance of a Machine Learning Algorithm Predicting Delirium in a Clinical Setting: A Mixed-Methods Study. *Journal of Medical Systems* 45, 4 (April 2021), 48. https://doi.org/10.1007/s10916-021-01727-6

[71] Stefanie Jauk, Diether Kramer, Birgit Großauer, Susanne Rienmüller, Alexander Avian, Andrea Berghold, Werner Leodolter, and Stefan Schulz. 2020. Risk Prediction of Delirium in Hospitalized Patients Using Machine Learning: An Implementation and Prospective Evaluation Study. *Journal of the American Medical Informatics Association* 27, 9 (Sept. 2020), 1383–1392. https://doi.org/10.1093/jamia/ocaa113

[72] Lijing Jia, Zijian Wei, Heng Zhang, Jiaming Wang, Ruiqi Jia, Manhong Zhou, Xueyan Li, Hankun Zhang, Xuedong Chen, Zheyuan Yu, Zhaohong Wang, Xiucheng Li, Tingting Li, Xiangge Liu, Pei Liu, Wei Chen, Jing Li, and Kunlun He. 2021. An Interpretable Machine Learning Model Based on a Quick Pre-Screening System Enables Accurate Deterioration Risk Prediction for COVID-19. *Scientific Reports* 11, 1 (Nov. 2021), 23127. https://doi.org/10.1038/s41598-021-02370-4

[73] Espen Jimenez-Solem, Tonny S. Petersen, Casper Hansen, Christian Hansen, Christina Lioma, Christian Igel, Wouter Boomsma, Oswin Krause, Stephan Lorenzen, Raghavendra Selvan, Janne Petersen, Martin Erik Nyeland, Mikkel Zöllner Ankarfeldt, Gert Mehl Virenfeldt, Matilde Winther-Jensen, Allan Linneberg, Mostafa Mehdipour Ghazi, Nicki Detlefsen, Andreas David Lauritzen, Abraham George Smith, Marleen De Bruijne, Bulat Ibragimov, Jens Petersen, Martin Lillholm, Jon Middleton, Stine Hasling Mogensen, Hans-Christian Thorsen-Meyer, Anders Perner, Marie Helleberg, Benjamin Skov Kaas-Hansen, Mikkel Bonde, Alexander Bonde, Akshay Pai, Mads Nielsen, and Martin Sillesen. 2021. Developing and Validating COVID-19 Adverse Outcome Risk Prediction Models from a Bi-National European Cohort of 5594 Patients. *Scientific Reports* 11, 1 (Feb. 2021), 3246. https://doi.org/10.1038/s41598-021-81844-x

[74] Zhi-Geng Jin, Hui Zhang, Mei-Hui Tai, Ying Yang, Yuan Yao, and Yu-Tao Guo. 2023. Natural Language Processing in a Clinical Decision Support System for the Identification of Venous Thromboembolism: Algorithm Development and Validation. *Journal of Medical Internet Research* 25 (April 2023), e43153. https://doi.org/10.2196/43153

[75] Leonard B. Jung, Jonas A. Gudera, Tim L. T. Wiegand, Simeon Allmendinger, Konstantinos Dimitriadis, and Inga K. Koerte. 2023. ChatGPT Passes German State Examination in Medicine with Picture Questions Omitted. *Deutsches Ärzteblatt international* (May 2023). https://doi.org/10.3238/arztebl.m2023.0113

[76] Kerstin Kalke, Hannah Studd, and Courtney L. Scherr. 2021. The Communication of Uncertainty in Health: A Scoping Review. *Patient Education and Counseling* 104, 8 (Aug. 2021), 1945–1961. https://doi.org/10.1016/j.pec.2021.01.034

[77] Eiichiro Kanda, Bogdan Iuliu Epureanu, Taiji Adachi, and Naoki Kashihara. 2023. Machine-Learning-Based Web System for the Prediction of Chronic Kidney Disease Progression and Mortality. *PLOS Digital Health* 2, 1 (Jan. 2023), e0000188. https://doi.org/10.1371/journal.pdig.0000188

[78] Dae Y. Kang, Pamela N. DeYoung, Justin Tantiongloc, Todd P. Coleman, and Robert L. Owens. 2021. Statistical Uncertainty Quantification to Augment Clinical Decision Support: A First Implementation in Sleep Medicine. *npj Digital Medicine* 4, 1 (Sept. 2021), 142. https://doi.org/10.1038/s41746-021-00515-3

[79] Elif Kartal. 2018. Machine Learning Techniques in Cardiac Risk Assessment. *The Turkish Journal of Thoracic and Cardiovascular Surgery* 26, 3 (July 2018), 394–401. https://doi.org/10.5606/tgkdc.dergisi.2018.15559

[80] Arash Kia, Prem Timsina, Himanshu N. Joshi, Eyal Klang, Rohit R. Gupta, Robert M. Freeman, David L Reich, Max S Tomlinson, Joel T Dudley, Roopa Kohli-Seth, Madhu Mazumdar, and Matthew A Levin. 2020. MEWS++: Enhancing the Prediction of Clinical Deterioration in Admitted Patients through a Machine Learning Model. *Journal of Clinical Medicine* 9, 2 (Jan. 2020), 343. https://doi.org/10.3390/jcm9020343

[81] Ryoung-Eun Ko, Jaehyeong Cho, Min-Kyue Shin, Sung Woo Oh, Yeonchan Seong, Jeongseok Jeon, Kyeongman Jeon, Soonmyung Paik, Joon Seok Lim, Sang Joon Shin, Joong Bae Ahn, Jong Hyuck Park, Seng Chan You, and Han Sang Kim. 2023. Machine Learning-Based Mortality Prediction Model for Critically Ill Cancer Patients Admitted to the Intensive Care Unit (CanICU). *Cancers* 15, 3 (Jan. 2023), 569. https://doi.org/10.3390/cancers15030569

[82] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine* 16, 9 (Sept. 2001), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[83] Kyle N. Kunze, Austin Kaidi, Sophia Madjarova, Evan M. Polce, Anil S. Ranawat, Danyal H. Nawabi, Bryan T. Kelly, Shane J. Nho, and Benedict U. Nwachukwu. 2022. External Validation of a Machine Learning Algorithm for Predicting Clinically Meaningful Functional Improvement After Arthroscopic Hip Preservation Surgery. *The American Journal of Sports Medicine* 50, 13 (Nov. 2022), 3593–3599. https://doi.org/10.1177/03635465221124275

[84] Keun-Sang Kwon, Heejung Bang, Andrew S Bomback, Dai-Ha Koh, Jung-Ho Yum, Ju-Hyung Lee, Sik Lee, Sung K Park, Keun-Young Yoo, Sue K Park, Soung-Hoon Chang, Hyun-Sul Lim, Joong Myung Choi, and Abhijit V Kshirsagar. 2012. A Simple Prediction Score for Kidney Disease in the Korean Population. *Nephrology* 17, 3 (March 2012), 278–284. https://doi.org/10.1111/j.1440-1797.2011.01552.x

[85] Hannah Labinsky, Dubravka Ukalovic, Fabian Hartmann, Vanessa Runft, André Wichmann, Jan Jakubcik, Kira Gambel, Katharina Otani, Harriet Morf, Jule Taubmann, Filippo Fagni, Arnd Kleyer, David Simon, Georg Schett, Matthias Reichert, and Johannes Knitza. 2023. An AI-Powered Clinical Decision Support System to Predict Flares in Rheumatoid Arthritis: A Pilot Study. *Diagnostics* 13, 1 (Jan. 2023), 148. https://doi.org/10.3390/diagnostics13010148

[86] Colleen L Lau, Deborah J Mills, Helen Mayfield, Narayan Gyawali, Brian J Johnson, Hongen Lu, Kasim Allel, Philip N Britton, Weiping Ling, Tina Moghaddam, and Luis Furuya-Kanamori. 2023. A Decision Support Tool for Risk–Benefit Analysis of Japanese Encephalitis Vaccine in Travellers. *Journal of Travel Medicine* 30, 7 (Nov. 2023), taad113. https://doi.org/10.1093/jtm/taad113

[87] Seung Mi Lee, Garam Lee, Tae Kyong Kim, Trang Le, Jie Hao, Young Mi Jung, Chan-Wook Park, Joong Shin Park, Jong Kwan Jun, Hyung-Chul Lee, and Dokyoon Kim. 2022. Development and Validation of a Prediction Model for Need for Massive Transfusion During Surgery Using Intraoperative Hemodynamic Monitoring Data. *JAMA Network Open* 5, 12 (Dec. 2022), e2246637. https://doi.org/10.1001/jamanetworkopen.2022.46637

[88] Mingxing Lei, Zhencan Han, Shengjie Wang, Tao Han, Shenyun Fang, Feng Lin, and Tianlong Huang. 2023. A Machine Learning-Based Prediction Model for in-Hospital Mortality among Critically Ill Patients with Hip Fracture: An Internal and External Validated Study. *Injury* 54, 2 (Feb. 2023), 636–644. https://doi.org/10.1016/j.injury.2022.11.031

[89] Wenle Li, Qian Zhou, Wencai Liu, Chan Xu, Zhi-Ri Tang, Shengtao Dong, Haosheng Wang, Wanying Li, Kai Zhang, Rong Li, Wenshi Zhang, Zhaohui Hu, Su Shibin, Qiang Liu, Sirui Kuang, and Chengliang Yin. 2022. A Machine Learning-Based Predictive Model for Predicting Lymph Node Metastasis in Patients With Ewing's Sarcoma. *Frontiers in Medicine* 9 (April 2022), 832108. https://doi.org/10.3389/fmed.2022.832108

[90] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. https://doi.org/10.48550/ARXIV.2205.14334

[91] Jiaqi Liu, Hengqiang Zhao, Yu Zheng, Lin Dong, Sen Zhao, Yukuan Huang, Shengkai Huang, Tianyi Qian, Jiali Zou, Shu Liu, Jun Li, Zihui Yan, Yalun Li, Shuo Zhang, Xin Huang, Wenyan Wang, Yiqun Li, Jie Wang, Yue Ming, Xiaoxin Li, Zeyu Xing, Ling Qin, Zhengye Zhao, Ziqi Jia, Jiaxin Li, Gang Liu, Menglu Zhang, Kexin Feng, Jiang Wu, Jianguo Zhang, Yongxin Yang, Zhihong Wu, Zhihua Liu, Jianming Ying, Xin Wang, Jianzhong Su, Xiang Wang, and Nan Wu. 2022. DrABC: Deep Learning Accurately Predicts Germline Pathogenic Mutation

Status in Breast Cancer Patients Based on Phenotype Data. *Genome Medicine* 14, 1 (Feb. 2022), 21. https://doi.org/10.1186/s13073-022-01027-9

[92] Yexin Liu, Yan Zhang, Di Liu, Xia Tan, Xiaofang Tang, Fan Zhang, Ming Xia, Guochun Chen, Liyu He, Letian Zhou, Xuejing Zhu, and Hong Liu. 2018. Prediction of ESRD in IgA Nephropathy Patients from an Asian Cohort: A Random Forest Model. *Kidney and Blood Pressure Research* 43, 6 (2018), 1852–1864. https://doi.org/10.1159/000495818

[93] Jorge M. Lobo, Alberto Jiménez-Valverde, and Raimundo Real. 2008. AUC: A Misleading Measure of the Performance of Predictive Distribution Models. *Global Ecology and Biogeography* 17, 2 (March 2008), 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

[94] Natasha A. Loghmanpour, Marek J. Druzdzel, and James F. Antaki. 2014. Cardiac Health Risk Stratification System (CHRiSS): A Bayesian-Based Decision Support System for Left Ventricular Assist Device (LVAD) Therapy. *PLoS ONE* 9, 11 (Nov. 2014), e111264. https://doi.org/10.1371/journal.pone.0111264

[95] Wei Lu, Yulan Tong, Xiuxiu Zhao, Yue Feng, Yi Zhong, Zhaojing Fang, Chen Chen, Kaizong Huang, Yanna Si, and Jianjun Zou. 2024. Machine Learning-Based Risk Prediction of Hypoxemia for Outpatients Undergoing Sedation Colonoscopy: A Practical Clinical Tool. *Postgraduate Medicine* 136, 1 (Jan. 2024), 84–94. https://doi.org/10.1080/00325481.2024.2313448

[96] Kun Lv, Chunmei Cui, Rui Fan, Xiaojuan Zha, Pengyu Wang, Jun Zhang, Lina Zhang, Jing Ke, Dong Zhao, Qinghua Cui, and Liming Yang. 2023. Detection of Diabetic Patients in People with Normal Fasting Glucose Using Machine Learning. *BMC Medicine* 21, 1 (Sept. 2023), 342. https://doi.org/10.1186/s12916-023-03045-9

[97] Georgios Manikis, Nicholas J Simos, Konstantina Kourou, Haridimos Kondylakis, Paula Poikonen-Saksela, Ketti Mazzocco, Ruth Pat-Horenczyk, Berta Sousa, Albino J Oliveira-Maia, Johanna Mattson, Ilan Roziner, Chiara Marzorati, Kostas Marias, Mikko Nuutinen, Evangelos Karademas, and Dimitrios Fotiadis. 2023. Personalized Risk Analysis to Improve the Psychological Resilience of Women Undergoing Treatment for Breast Cancer: Development of a Machine Learning–Driven Clinical Decision Support Tool. *Journal of Medical Internet Research* 25 (June 2023), e43838. https://doi.org/10.2196/43838

[98] Carmel M. Martin, Carl Vogel, Deirdre Grady, Atieh Zarabzadeh, Lucy Hederman, John Kellett, Kevin Smith, and Brendan O' Shea. 2012. Implementation of Complex Adaptive Chronic Care: The P Atient J Ourney R Ecord System (PAJR). *Journal of Evaluation in Clinical Practice* 18, 6 (Dec. 2012), 1226–1234. https://doi.org/10.1111/j.1365-2753.2012.01880.x

[99] Tim Mathes, Carolina Pape-Köhler, Lena Moerders, Eberhard Lux, and Edmund A M Neugebauer. 2018. External Validation and Update of the RICP—A Multivariate Model to Predict Chronic Postoperative Pain. *Pain Medicine* 19, 8 (Aug. 2018), 1674–1682. https://doi.org/10.1093/pm/pnx242

[100] John Mongan, Linda Moy, and Charles E. Kahn. 2020. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiology: Artificial Intelligence* 2, 2 (March 2020), e200029. https://doi.org/10.1148/ryai.2020200029

[101] Kyoung Ja Moon, Chang-Sik Son, Jong-Ha Lee, and Mina Park. 2022. The Development of a Web-Based App Employing Machine Learning for Delirium Prevention in Long-Term Care Facilities in South Korea. *BMC Medical Informatics and Decision Making* 22, 1 (Aug. 2022), 220. https://doi.org/10.1186/s12911-022-01966-8

[102] Abu Saleh Mohammad Mosa, Md Kamruz Zaman Rana, Humayera Islam, A K M Mosharraf Hossain, and Illhoi Yoo. 2021. A Smartphone-Based Decision Support Tool for Predicting Patients at Risk of Chemotherapy-Induced Nausea and Vomiting: Retrospective Study on App Development Using Decision Tree Induction. *JMIR mHealth and uHealth* 9, 12 (Dec. 2021), e27024. https://doi.org/10.2196/27024

[103] Mayooran Namasivayam, Paul D Myers, John V Guttag, Romain Capoulade, Philippe Pibarot, Michael H Picard, Judy Hung, and Collin M Stultz. 2022. Predicting Outcomes in Patients with Aortic Stenosis Using Machine Learning: The Aortic Stenosis Risk (ASteRisk) Score. *Open Heart* 9, 1 (May 2022), e001990. https://doi.org/10.1136/openhrt-2022-001990

[104] S A M Nashef, F Roques, P Michel, E Gauducheau, S Lemeshow, and R Salamon. 1999. European System for Cardiac Operative Risk Evaluation (EuroSCORE)q. *thoracic Surgery* (1999).

[105] S. A. M. Nashef, F. Roques, L. D. Sharples, J. Nilsson, C. Smith, A. R. Goldstone, and U. Lockowandt. 2012. EuroSCORE II. *European Journal of Cardio-Thoracic Surgery* 41, 4 (April 2012), 734–745. https://doi.org/10.1093/ejcts/ezs043

[106] Rajan Nathan and Sahil Bhandari. 2024. Risk Assessment in Clinical Practice: A Framework for Decision-Making in Real-World Complex Systems. *BJPsych Advances* 30, 1 (Jan. 2024), 53–63. https://doi.org/10.1192/bja.2022.67

[107] Rajan Nathan, Jonathon Whyler, and Peter Wilson. 2021. Risk of Harm to Others: Subjectivity and Meaning of Risk in Mental Health Practice. *Journal of Risk Research* 24, 10 (Oct. 2021), 1228–1238. https://doi.org/10.1080/13669877.2020.1819389

[108] W D Neary, B P Heather, and J J Earnshaw. 2003. The Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (POSSUM). *British Journal of Surgery* 30, 2 (2003), 157–165. https://doi.org/10.1002/bjs.4041

[109] Joseph R. Nellis, Zhifei Sun, Bora Chang, Gina Della Porta, and Christopher R. Mantyh. 2023. A Risk-Prediction Platform for Acute Kidney Injury and 30-Day Readmission After Colorectal Surgery. *Journal of Surgical Research* 292 (Dec. 2023), 91–96. https://doi.org/10.1016/j.jss.2023.07.040

[110] Oanh Kieu Nguyen, Anil N. Makam, Christopher Clark, Song Zhang, Bin Xie, Ferdinand Velasco, Ruben Amarasingham, and Ethan A. Halm. 2016. Predicting

All-cause Readmissions Using Electronic Health Record Data from the Entire Hospitalization: Model Development and Comparison. *Journal of Hospital Medicine* 11, 7 (July 2016), 473–480. https://doi.org/10.1002/jhm.2568

[111] Yoshitsugu Obi, Danh V. Nguyen, Hui Zhou, Melissa Soohoo, Lishi Zhang, Yanjun Chen, Elani Streja, John J. Sim, Miklos Z. Molnar, Connie M. Rhee, Kevin C. Abbott, Steven J. Jacobsen, Csaba P. Kovesdy, and Kamyar Kalantar-Zadeh. 2018. Development and Validation of Prediction Scores for Early Mortality at Transition to Dialysis. *Mayo Clinic Proceedings* 93, 9 (Sept. 2018), 1224–1235. https://doi.org/10.1016/j.mayocp.2018.04.017

[112] Sean M. O'Brien, Liqi Feng, Xia He, Ying Xian, Jeffrey P. Jacobs, Vinay Badhwar, Paul A. Kurlansky, Anthony P. Furnary, Joseph C. Cleveland, Kevin W. Lobdell, Christina Vassileva, Moritz C. Wyler Von Ballmoos, Vinod H. Thourani, J. Scott Rankin, James R. Edgerton, Richard S. D'Agostino, Nimesh D. Desai, Fred H. Edwards, and David M. Shahian. 2018. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 2—Statistical Methods and Results. *The Annals of Thoracic Surgery* 105, 5 (May 2018), 1419–1428. https://doi.org/10.1016/j.athoracsur.2018.03.003

[113] Koichi Ogura, Tabu Gokita, Yusuke Shinoda, Hirotaka Kawano, Tatsuya Takagi, Keisuke Ae, Akira Kawai, Rikard Wedin, and Jonathan A. Forsberg. 2017. Can A Multivariate Model for Survival Estimation in Skeletal Metastases (PATHFx) Be Externally Validated Using Japanese Patients? *Clinical Orthopaedics & Related Research* 475, 9 (Sept. 2017), 2263–2270. https://doi.org/10.1007/s11999-017-5389-3

[114] Evangelos K. Oikonomou, Marc A. Suchard, Darren K. McGuire, and Rohan Khera. 2022. Phenomapping-Derived Tool to Individualize the Effect of Canagliflozin on Cardiovascular Risk in Type 2 Diabetes. *Diabetes Care* 45, 4 (April 2022), 965–974. https://doi.org/10.2337/dc21-1765

[115] Thorsten Pachur, Ralph Hertwig, Gerd Gigerenzer, and Eduard Brandstätter. 2013. Testing Process Predictions of Models of Risky Choice: A Quantitative Model Comparison Approach. *Frontiers in Psychology* 4, 646 (2013), 1–22. https://doi.org/10.3389/fpsyg.2013.00646

[116] Brian W. Patterson, Collin J. Engstrom, Varun Sah, Maureen A. Smith, Eneida A. Mendonça, Michael S. Pulia, Michael D. Repplinger, Azita G. Hamedani, David Page, and Manish N. Shah. 2019. Training and Interpreting Machine Learning Algorithms to Evaluate Fall Risk After Emergency Department Visits. *Medical Care* 57, 7 (July 2019), 560–566. https://doi.org/10.1097/MLR.0000000000001140

[117] Christopher Pearce, Adam McLeod, Natalie Rinehart, Jon Patrick, Anna Fragkoudi, Jason Ferrigi, Elizabeth Deveny, Robin Whyte, and Marianne Shearer. 2019. POLAR Diversion: Using General Practice Data to Calculate Risk of Emergency Department Presentation at the Time of Consultation. *Applied Clinical Informatics* 10, 01 (Jan. 2019), 151–157. https://doi.org/10.1055/s-0039-1678608

[118] Zane B. Perkins, Barbaros Yet, Max Marsden, Simon Glasgow, William Marsh, Ross Davenport, Karim Brohi, and Nigel R. M. Tai. 2021. Early Identification

of Trauma-induced Coagulopathy: Development and Validation of a Multivariable Risk Prediction Model. *Annals of Surgery* 274, 6 (Dec. 2021), e1119–e1128. https://doi.org/10.1097/SLA.0000000000003771

[119] Zane B. Perkins, Barbaros Yet, Anna Sharrock, Rory Rickard, William Marsh, Todd E. Rasmussen, and Nigel R. M. Tai. 2020. Predicting the Outcome of Limb Revascularization in Patients With Lower-extremity Arterial Trauma: Development and External Validation of a Supervised Machine-learning Algorithm to Support Surgical Decisions. *Annals of Surgery* 272, 4 (Oct. 2020), 564–572. https://doi.org/10.1097/SLA.0000000000004132

[120] Wojciech Pluskiewicz, Piotr Adamczyk, Aleksandra Werner, Małgorzata Bach, and Bogna Drozdzowska. 2023. POL-RISK: An Algorithm for 10-Year Fracture Risk Prediction in the Postmenopausal Women from the RAC-OST-POL Study. *Polish Archives of Internal Medicine* 133, 3 (Jan. 2023). https://doi.org/10.20452/pamw.16395

[121] Yiming Qi, Xiaolei Lin, Wenzhi Pan, Xiaochun Zhang, Yuefan Ding, Shasha Chen, Lei Zhang, Daxin Zhou, and Junbo Ge. 2023. A Prediction Model for Permanent Pacemaker Implantation after Transcatheter Aortic Valve Replacement. *European Journal of Medical Research* 28, 1 (July 2023), 262. https://doi.org/10.1186/s40001-023-01237-w

[122] Nidan Qiao, Qilin Zhang, Li Chen, Wenqiang He, Zengyi Ma, Zhao Ye, Min He, Zhaoyun Zhang, Xiang Zhou, Ming Shen, Xuefei Shou, Xiaoyun Cao, Yongfei Wang, and Yao Zhao. 2023. Machine Learning Prediction of Venous Thromboembolism after Surgeries of Major Sellar Region Tumors. *Thrombosis Research* 226 (June 2023), 1–8. https://doi.org/10.1016/j.thromres.2023.04.007

[123] Romy Richter, Jesse Jansen, Iris Bongaerts, Olga Damman, Jany Rademakers, and Trudy Van Der Weijden. 2023. Communication of Benefits and Harms in Shared Decision Making with Patients with Limited Health Literacy: A Systematic Review of Risk Communication Strategies. *Patient Education and Counseling* 116 (Nov. 2023), 107944. https://doi.org/10.1016/j.pec.2023.107944

[124] Richard D. Riley and Gary S. Collins. 2023. Stability of Clinical Prediction Models Developed Using Statistical or Machine Learning Methods. *Biometrical Journal* 65, 8 (Dec. 2023), 2200302. https://doi.org/10.1002/bimj.202200302

[125] Richard D. Riley, Alexander Pate, Paula Dhiman, Lucinda Archer, Glen P. Martin, and Gary S. Collins. 2023. Clinical Prediction Models and the Multiverse of Madness. *BMC Medicine* 21, 1 (Dec. 2023), 502. https://doi.org/10.1186/s12916-023-03212-y

[126] Leonard J Savage. 1951. The Theory of Statistical Decision. *Journal of the American Statistical association* 46, 253 (1951), 55–67.

[127] Francesco Paolo Schena, Vito Walter Anelli, Joseph Trotta, Tommaso Di Noia, Carlo Manno, Giovanni Tripepi, Graziella D'Arrigo, Nicholas C. Chesnaye, Maria Luisa Russo, Maria Stangou, Aikaterini Papagianni, Carmine Zoccali,

Vladimir Tesar, Rosanna Coppo, V. Tesar, D. Maixnerova, S. Lundberg, L. Gesualdo, F. Emma, L. Fuiano, G. Beltrame, C. Rollino, R. Coppo, A. Amore, R. Camilla, L. Peruzzi, M. Praga, S. Feriozzi, R. Polci, G. Segoloni, L. Colla, A. Pani, A. Angioi, L. Piras, J. Feehally, G. Cancarini, S. Ravera, M. Durlik, E. Moggia, J. Ballarin, S. Di Giulio, F. Pugliese, I. Serriello, Y. Caliskan, M. Sever, I. Kilicaslan, F. Locatelli, L. Del Vecchio, J.F.M. Wetzels, H. Peters, U. Berg, F. Carvalho, A.C. Da Costa Ferreira, M. Maggio, A. Wiecek, M. Ots-Rosenberg, R. Magistroni, R. Topaloglu, Y. Bilginer, M. D'Amico, M. Stangou, F. Giacchino, D. Goumenos, M. Papasotiriou, K. Galesic, L. Toric, C. Geddes, K. Siamopoulos, O. Balafa, M. Galliani, P. Stratta, M. Quaglia, R. Bergia, R. Cravero, M. Salvadori, L. Cirami, B. Fellstrom, H. Kloster Smerud, F. Ferrario, T. Stellato, J. Egido, C. Martin, J. Floege, F. Eitner, T. Rauen, A. Lupo, P. Bernich, P. Menè, M. Morosetti, C. Van Kooten, T. Rabelink, M.E.J. Reinders, J.M. Boria Grinyo, S. Cusinato, L. Benozzi, S. Savoldi, C. Licata, M. Mizerska-Wasiak, M. Roszkowska-Blaim, G. Martina, A. Messuerotti, A. Dal Canton, C. Esposito, C. Migotto, G. Triolo, F. Mariano, C. Pozzi, R. Boero, Mazzucco, C. Giannakakis, E. Honsova, B. Sundelin, A.M. Di Palma, F. Ferrario, E. Gutiérrez, A.M. Asunis, J. Barratt, R. Tardanico, A. Perkowska-Ptasinska, J. Arce Terroba, M. Fortunato, A. Pantzaki, Y. Ozluk, E. Steenbergen, M. Soderberg, Z. Riispere, L. Furci, D. Orhan, D. Kipgen, D. Casartelli, D. GalesicLjubanovic, H. Gakiopoulou, E. Bertoni, P. Cannata Ortiz, H. Karkoszka, H.J. Groene, A. Stoppacciaro, I. Bajema, J. Bruijn, X. FulladosaOliveras, J. Maldyk, E. Ioachim, Daniela Abbrescia, Nikoleta Kouri, Maria Stangou, Aikaterini Papagianni, Francesco Scolari, Elisa Delbarba, Mario Bonomini, Luca Piscitani, Giovanni Stallone, Barbara Infante, Giulia Godeas, Desiree Madio, Luigi Biancone, Marco Campagna, Gianluigi Zaza, Isabella Squarzoni, and Concetta Cangemi. 2021. Development and Testing of an Artificial Intelligence Tool for Predicting End-Stage Kidney Disease in Patients with Immunoglobulin A Nephropathy. *Kidney International* 99, 5 (May 2021), 1179–1188. `https://doi.org/10.1016/j.kint.2020.07.046`

[128] Akash A. Shah, Sai K. Devana, Changhee Lee, Amador Bugarin, Michelle K. Hong, Alexander Upfill-Brown, Gideon Blumstein, Elizabeth L. Lord, Arya N. Shamie, Mihaela Van Der Schaar, Nelson F. SooHoo, and Don Y. Park. 2022. A Risk Calculator for the Prediction of C5 Nerve Root Palsy After Instrumented Cervical Fusion. *World Neurosurgery* 166 (Oct. 2022), e703–e710. `https://doi.org/10.1016/j.wneu.2022.07.082`

[129] Omar Shareef, James T. Kwan, Sarina Lau, Mohammad Ali Tahboub, and Hajirah N. Saeed. 2022. An Alternative Model for Assessing Mortality Risk in Stevens Johnson Syndrome/Toxic Epidermal Necrolysis Using a Random Forests Classifier: A Pilot Study. *Frontiers in Medicine* 9 (Dec. 2022), 935408. `https://doi.org/10.3389/fmed.2022.935408`

[130] Amna Sher, Sowmini Medavaram, Barbara Nemesure, Sean Clouston, and Roger Keresztes. 2020. Risk Stratification of Locally Advanced Non-Small Cell Lung Cancer (NSCLC) Patients Treated with Chemo-Radiotherapy: An Institutional Analysis. *Cancer Management and Research* Volume 12 (Aug. 2020), 7165–7171. `https://doi.org/10.2147/CMAR.S250868`

[131] Cathy Shields, Scott G Cunningham, Deborah J Wake, Evridiki Fioratou, Doogie

Brodie, Sam Philip, and Nicholas T Conway. 2022. User-Centered Design of A Novel Risk Prediction Behavior Change Tool Augmented With an Artificial Intelligence Engine (MyDiabetesIQ): A Sociotechnical Systems Approach. *JMIR Human Factors* 9, 1 (Feb. 2022), e29973. https://doi.org/10.2196/29973

[132] Roni Shouval, Myriam Labopin, Ori Bondi, Hila Mishan-Shamay, Avichai Shimoni, Fabio Ciceri, Jordi Esteve, Sebastian Giebel, Norbert C. Gorin, Christoph Schmid, Emmanuelle Polge, Mahmoud Aljurf, Nicolaus Kroger, Charles Craddock, Andrea Bacigalupo, Jan J. Cornelissen, Frederic Baron, Ron Unger, Arnon Nagler, and Mohamad Mohty. 2015. Prediction of Allogeneic Hematopoietic Stem-Cell Transplantation Mortality 100 Days After Transplantation Using a Machine Learning Algorithm: A European Group for Blood and Marrow Transplantation Acute Leukemia Working Party Retrospective Data Mining Study. *Journal of Clinical Oncology* 33, 28 (Oct. 2015), 3144–3151. https://doi.org/10.1200/JCO.2014.59.1339

[133] Marleine Mefeugue Siga, Michel Ducher, Nans Florens, Hubert Roth, Nadir Mahloul, Denis Fouque, and Jean-Pierre Fauvel. 2020. Prediction of All-Cause Mortality in Haemodialysis Patients Using a Bayesian Network. *Nephrology Dialysis Transplantation* 35, 8 (Aug. 2020), 1420–1425. https://doi.org/10.1093/ndt/gfz295

[134] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large Language Models Encode Clinical Knowledge. *Nature* 620, 7972 (Aug. 2023), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[135] Sebastian Spänig, Agnes Emberger-Klein, Jan-Peter Sowa, Ali Canbay, Klaus Menrad, and Dominik Heider. 2019. The Virtual Doctor: An Interactive Clinical-Decision-Support System Based on Deep Learning for Non-Invasive Prediction of Diabetes. *Artificial Intelligence in Medicine* 100 (Sept. 2019), 101706. https://doi.org/10.1016/j.artmed.2019.101706

[136] John Sperger, Kushal S Shah, Minxin Lu, Xian Zhang, Ryan C Ungaro, Erica J Brenner, Manasi Agrawal, Jean-Frédéric Colombel, Michael D Kappelman, and Michael R Kosorok. 2021. Development and Validation of Multivariable Prediction Models for Adverse COVID-19 Outcomes in Patients with IBD. *BMJ Open* 11, 11 (Nov. 2021), e049740. https://doi.org/10.1136/bmjopen-2021-049740

[137] David Spiegelhalter. 2017. Risk and Uncertainty Communication. *Annual Review of Statistics and Its Application* 4, 1 (March 2017), 31–60. https://doi.org/10.1146/annurev-statistics-010814-020148

[138] John Stacy, Rachel Kim, Christopher Barrett, Balaviknesh Sekar, Steven Simon, Farnoush Banaei-Kashani, and Michael A Rosenberg. 2022. Qualitative Evaluation of an Artificial Intelligence–Based Clinical Decision Support System to Guide

Rhythm Management of Atrial Fibrillation: Survey Study. *JMIR Formative Research* 6, 8 (Aug. 2022), e36443. https://doi.org/10.2196/36443

[139] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11, 5 (Sept. 2016), 702–712. https://doi.org/10.1177/1745691616658637

[140] Miroslav Stojadinovic, Milorad Stojadinovic, and Damjan Pantic. 2019. Decision Tree Analysis for Prostate Cancer Prediction. *Srpski arhiv za celokupno lekarstvo* 147, 1-2 (2019), 52–58. https://doi.org/10.2298/SARH181127039S

[141] Andreas N. Strobl, Andrew J. Vickers, Ben Van Calster, Ewout Steyerberg, Robin J. Leach, Ian M. Thompson, and Donna P. Ankerst. 2015. Improving Patient Prostate Cancer Risk Assessment: Moving from Static, Globally-Applied to Dynamic, Practice-Specific Risk Calculators. *Journal of Biomedical Informatics* 56 (Aug. 2015), 87–93. https://doi.org/10.1016/j.jbi.2015.05.001

[142] Hong Sun, Kristof Depraetere, Laurent Meesseman, Jos De Roo, Martijn Vanbiervliet, Jos De Baerdemaeker, Herman Muys, Vera Von Dossow, Nikolai Hulde, and Ralph Szymanowsky. 2021. A Scalable Approach for Developing Clinical Risk Prediction Applications in Different Hospitals. *Journal of Biomedical Informatics* 118 (June 2021), 103783. https://doi.org/10.1016/j.jbi.2021.103783

[143] Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. 2020. An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success. *npj Digital Medicine* 3, 1 (Feb. 2020), 17. https://doi.org/10.1038/s41746-020-0221-y

[144] Mohammadamin Tajgardoon, Malarkodi J. Samayamuthu, Luca Calzoni, and Shyam Visweswaran. 2019. Patient-Specific Explanations for Predictions of Clinical Outcomes. *ACI Open* 03, 02 (July 2019), e88–e97. https://doi.org/10.1055/s-0039-1697907

[145] Navdeep Tangri. 2011. A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure. *JAMA* 305, 15 (April 2011), 1553. https://doi.org/10.1001/jama.2011.451

[146] Ammarin Thakkinstian, Atiporn Ingsathit, Amnart Chaiprasert, Sasivimol Rattanasiri, Pornpen Sangthawan, Pongsathorn Gojaseni, Kriwiporn Kiattisunthorn, Leena Ongaiyooth, and Prapaipim Thirakhupt. 2011. A Simplified Clinical Prediction Score of Chronic Kidney Disease: A Cross-Sectional-Survey Study. *BMC Nephrology* 12, 1 (Dec. 2011), 45. https://doi.org/10.1186/1471-2369-12-45

[147] Prem Timsina, Himanshu N. Joshi, Fu-Yuan Cheng, Ilana Kersch, Sara Wilson, Claudia Colgan, Robert Freeman, David L. Reich, Jeffrey Mechanick, Madhu Mazumdar, Matthew A. Levin, and Arash Kia. 2021. MUST-Plus: A Machine Learning Classifier That Improves Malnutrition Screening in Acute Care Facilities. *Journal of the American College of Nutrition* 40, 1 (Jan. 2021), 3–12. https://doi.org/10.1080/07315724.2020.1774821

[148] Takeshi Tohyama, Tomomi Ide, Masataka Ikeda, Hidetaka Kaku, Nobuyuki Enzan, Shouji Matsushima, Kouta Funakoshi, Junji Kishimoto, Koji Todaka, and Hiroyuki Tsutsui. 2021. Machine Learning-based Model for Predicting 1 Year Mortality of Hospitalized Patients with Heart Failure. *ESC Heart Failure* 8, 5 (Oct. 2021), 4077–4085. https://doi.org/10.1002/ehf2.13556

[149] Yi-Ju Tseng, Yi-Cheng Wang, Pei-Chun Hsueh, and Chih-Ching Wu. 2022. Development and Validation of Machine Learning-Based Risk Prediction Models of Oral Squamous Cell Carcinoma Using Salivary Autoantibody Biomarkers. *BMC Oral Health* 22, 1 (Nov. 2022), 534. https://doi.org/10.1186/s12903-022-02607-2

[150] Eva Tsui, Sy Au, Cp Wong, Alan Cheung, and Peggo Lam. 2015. Development of an Automated Model to Predict the Risk of Elderly Emergency Medical Admissions within a Month Following an Index Hospital Visit: A Hong Kong Experience. *Health Informatics Journal* 21, 1 (March 2015), 46–56. https://doi.org/10.1177/1460458213501095

[151] Craig A. Umscheid, Joel Betesh, Christine VanZandbergen, Asaf Hanish, Gordon Tait, Mark E. Mikkelsen, Benjamin French, and Barry D. Fuchs. 2015. Development, Implementation, and Impact of an Automated Early Warning and Response System for Sepsis: EWRS for Sepsis. *Journal of Hospital Medicine* 10, 1 (Jan. 2015), 26–31. https://doi.org/10.1002/jhm.2259

[152] Rutger R. Van De Leur, Remco De Brouwer, Hidde Bleijendaal, Tom E. Verstraelen, Belend Mahmoud, Ana Perez-Matos, Cathelijne Dickhoff, Bas A. Schoonderwoerd, Tjeerd Germans, Arjan Houweling, Paul A. Van Der Zwaag, Moniek G.P.J. Cox, J. Peter Van Tintelen, Anneline S.J.M. Te Riele, Maarten P. Van Den Berg, Arthur A.M. Wilde, Pieter A. Doevendans, Rudolf A. De Boer, and René Van Es. 2024. ECG-only Explainable Deep Learning Algorithm Predicts the Risk for Malignant Ventricular Arrhythmia in Phospholamban Cardiomyopathy. *Heart Rhythm* 21, 7 (Feb. 2024), S1547527124002108. https://doi.org/10.1016/j.hrthm.2024.02.038

[153] Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A. Clifton, Gary S. Collins, Spiros Denaxas, Alastair K. Denniston, Livia Faes, Bart Geerts, Mudathir Ibrahim, Xiaoxuan Liu, Bilal A. Mateen, Piyush Mathur, Melissa D. McCradden, Lauren Morgan, Johan Ordish, Campbell Rogers, Suchi Saria, Daniel S. W. Ting, Peter Watkinson, Wim Weber, Peter Wheatstone, Peter McCulloch, the DECIDE-AI expert group, Aaron Y. Lee, Alan G. Fraser, Ali Connell, Alykhan Vira, Andre Esteva, Andrew D. Althouse, Andrew L. Beam, Anne De Hond, Anne-Laure Boulesteix, Anthony Bradlow, Ari Ercole, Arsenio Paez, Athanasios Tsanas, Barry Kirby, Ben Glocker, Carmelo Velardo, Chang Min Park, Charisma Hehakaya, Chris Baber, Chris Paton, Christian Johner, Christopher J. Kelly, Christopher J. Vincent, Christopher Yau, Clare McGenity, Constantine Gatsonis, Corinne Faivre-Finn, Crispin Simon, Danielle Sent, Danilo Bzdok, Darren Treanor, David C. Wong, David F. Steiner, David Higgins, Dawn Benson, Declan P. O'Regan, Dinesh V. Gunasekaran, Dominic Danks, Emanuele Neri, Evangelia Kyrimi, Falk Schwendicke, Farah Magrabi, Frances Ives, Frank E. Rademakers, George E. Fowler, Giuseppe Frau, H. D. Jeffry Hogg, Hani J. Marcus, Heang-Ping Chan, Henry Xiang, Hugh F. McIntyre, Hugh Harvey, Hyungjin Kim, Ibrahim Habli, James C.

Fackler, James Shaw, Janet Higham, Jared M. Wohlgemut, Jaron Chong, Jean-Emmanuel Bibault, Jérémie F. Cohen, Jesper Kers, Jessica Morley, Joachim Krois, Joao Monteiro, Joel Horovitz, John Fletcher, Jonathan Taylor, Jung Hyun Yoon, Karandeep Singh, Karel G. M. Moons, Kassandra Karpathakis, Ken Catchpole, Kerenza Hood, Konstantinos Balaskas, Konstantinos Kamnitsas, Laura Militello, Laure Wynants, Lauren Oakden-Rayner, Laurence B. Lovat, Luc J. M. Smits, Ludwig C. Hinske, M. Khair ElZarrad, Maarten Van Smeden, Mara Giavina-Bianchi, Mark Daley, Mark P. Sendak, Mark Sujan, Maroeska Rovers, Matthew DeCamp, Matthew Woodward, Matthieu Komorowski, Max Marsden, Maxine Mackintosh, Michael D. Abramoff, Miguel Ángel Armengol De La Hoz, Neale Hambidge, Neil Daly, Niels Peek, Oliver Redfern, Omer F. Ahmad, Patrick M. Bossuyt, Pearse A. Keane, Pedro N. P. Ferreira, Petra Schnell-Inderst, Pietro Mascagni, Prokar Dasgupta, Pujun Guan, Rachel Barnett, Rawen Kader, Reena Chopra, Ritse M. Mann, Rupa Sarkar, Saana M. Mäenpää, Samuel G. Finlayson, Sarah Vollam, Sebastian J. Vollmer, Seong Ho Park, Shakir Laher, Shalmali Joshi, Siri L. Van Der Meijden, Susan C. Shelmerdine, Tien-En Tan, Tom J. W. Stocker, Valentina Giannini, Vince I. Madai, Virginia Newcombe, Wei Yan Ng, Wendy A. Rogers, William Ogallo, Yoonyoung Park, and Zane B. Perkins. 2022. Reporting Guideline for the Early-Stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI. *Nature Medicine* 28, 5 (May 2022), 924–933. `https://doi.org/10.1038/s41591-022-01772-9`

[154] Andrew J Vickers and Elena B Elkin. 2006. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making* 26, 6 (2006), 565–574.

[155] Andrew J. Vickers, Daniel D. Sjoberg, Donna P. Ankerst, Catherine M. Tangen, Phyllis J. Goodman, and Ian M. Thompson. 2013. The Prostate Cancer Prevention Trial Risk Calculator and the Relationship between Prostate-specific Antigen and Biopsy Outcome. *Cancer* 119, 16 (Aug. 2013), 3007–3011. `https://doi.org/10.1002/cncr.28114`

[156] C. Villodre, L. Taccogna, P. Zapater, M. Cantó, L. Mena, J. M. Ramia, F. Lluís, N. Afonso, V. Aguilella, J. Aguiló, J. C. Alados, M. Alberich, A. B. Apio, R. Balongo, E. Bra, A. Bravo-Gutiérrez, F. J. Briceño, J. Cabañas, G. Cánovas, I. Caravaca, S. Carbonell, E. Carrera-Dacosta, E. E Castro, C. Caula, E. Choolani-Bhojwani, A. Codina, S. Corral, C. Cuenca, Y. Curbelo-Peña, M. M. Delgado-Morales, L. Delgado-Plasencia, E. Doménech, A. M. Estévez, A. M. Feria, M. A. Gascón-Domínguez, R. Gianchandani, C. González, R. J. Hevia, M. A. González, J. M. Hidalgo, M. Lainez, N. Lluís, F. López, J. López-Fernández, J. A. López-Ruíz, P. Lora-Cumplido, Z. Madrazo, J. Marchena, Marenco B. De La Cuadra, S. Martín, Martínez I. Casas, P. Martínez, A. Mena-Mateos, D. Morales-García, C. Mulas, E. Muñoz-Forner, A. Naranjo, A. Navarro-Sánchez, I. Oliver, I. Ortega, R. Ortega-Higueruelo, S. Ortega-Ruiz, J. Osorio, M. H. Padín, J. J. Pamies, M. Paredes, F. Pareja-Ciuró, J. Parra, C. V. Pérez-Guarinós, B. Pérez-Saborido, J. Pintor-Tortolero, K. Plua-Muñiz, M. Rey, I. Rodríguez, C. Ruiz, R. Ruíz, S. Ruiz, A. Sánchez, D. Sánchez, R. Sánchez, F. Sánchez-Cabezudo, R. Sánchez-Santos, J. Santos, M. P. Serrano-Paz, V. Soria-Aledo, L. Tallón-Aguilar, J. H. Valdivia-Risco, H. Vallverdú-Cartié, C. Varela, J. Villar-del-Moral, and N. Zam-

budio. 2022. Simplified Risk-Prediction for Benchmarking and Quality Improvement in Emergency General Surgery. Prospective, Multicenter, Observational Cohort Study. *International Journal of Surgery* 97 (Jan. 2022), 106168. https://doi.org/10.1016/j.ijsu.2021.106168

[157] Vivianne H. M. Visschers, Ree M. Meertens, Wim W. F. Passchier, and Nanne N. K. De Vries. 2009. Probability Information in Risk Communication: A Review of the Research Literature. *Risk Analysis* 29, 2 (Feb. 2009), 267–287. https://doi.org/10.1111/j.1539-6924.2008.01137.x

[158] Han-Zhang Wang, Su-Wei Chen, Yong-Liang Zhong, Yi-Peng Ge, Zhi-Yu Qiao, Cheng-Nan Li, Ru-Tao Guo, Zhe Zhang, Chen-Hui Qiao, and Jun-Ming Zhu. 2023. Anzhen Risk Evaluation System for Acute Aortic Syndrome (AZSCORE-AAS): Protocol for a Multicentre Prospective Cohort Study in Northern China. *BMJ Open* 13, 6 (June 2023), e067469. https://doi.org/10.1136/bmjopen-2022-067469

[159] Shiqi Wang, Jinwan Wang, Mark Xuefang Zhu, and Qian Tan. 2022. Machine Learning for the Prediction of Minor Amputation in University of Texas Grade 3 Diabetic Foot Ulcers. *PLOS ONE* 17, 12 (Dec. 2022), e0278445. https://doi.org/10.1371/journal.pone.0278445

[160] Wenyi Wang, Sining Chen, Kieran A. Brune, Ralph H. Hruban, Giovanni Parmigiani, and Alison P. Klein. 2007. PancPRO: Risk Assessment for Individuals With a Family History of Pancreatic Cancer. *Journal of Clinical Oncology* 25, 11 (April 2007), 1417–1422. https://doi.org/10.1200/JCO.2006.09.2452

[161] Odette Wegwarth and Gerd Gigerenzer. 2011. *Statistical Illiteracy in Doctors*. MIT Press, Cambridge, MA, USA, 16.

[162] Peter W. F. Wilson, Ralph B. D'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, and William B. Kannel. 1998. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* 97, 18 (May 1998), 1837–1847. https://doi.org/10.1161/01.CIR.97.18.1837

[163] Danny J. N. Wong, Steve Harris, Arun Sahni, James R. Bedford, Laura Cortes, Richard Shawyer, Andrew M. Wilson, Helen A. Lindsay, Doug Campbell, Scott Popham, Lisa M. Barneto, Paul S. Myles, SNAP-2: EPICCS collaborators, and S. Ramani Moonesinghe. 2020. Developing and Validating Subjective and Objective Risk-Assessment Measures for Predicting Mortality after Major Surgery: An International Prospective Cohort Study. *PLOS Medicine* 17, 10 (Oct. 2020), e1003253. https://doi.org/10.1371/journal.pmed.1003253

[164] Joshua Guoxian Wong, Aung-Hein Aung, Weixiang Lian, David Chien Lye, Chee-Kheong Ooi, and Angela Chow. 2020. Risk Prediction Models to Guide Antibiotic Prescribing: A Study on Adult Patients with Uncomplicated Upper Respiratory Tract Infections in an Emergency Department. *Antimicrobial Resistance & Infection Control* 9, 1 (Dec. 2020), 171. https://doi.org/10.1186/s13756-020-00825-3

[165] Myung Woo, Brooke Alhanti, Sam Lusk, Felicia Dunston, Stephen Blackwelder, Kay S. Lytle, Benjamin A. Goldstein, and Armando Bedoya. 2020. Evaluation of ML-Based Clinical Decision Support Tool to Replace an Existing Tool in an

Academic Health System: Lessons Learned. *Journal of Personalized Medicine* 10, 3 (Aug. 2020), 104. https://doi.org/10.3390/jpm10030104

[166] L. K. Woolery and J. Grzymala-Busse. 1994. Machine Learning for an Expert System to Predict Preterm Birth Risk. *Journal of the American Medical Informatics Association* 1, 6 (Nov. 1994), 439–446. https://doi.org/10.1136/jamia.1994.95153433

[167] Guangyao Wu, Pei Yang, Yuanliang Xie, Henry C. Woodruff, Xiangang Rao, Julien Guiot, Anne-Noelle Frix, Renaud Louis, Michel Moutschen, Jiawei Li, Jing Li, Chenggong Yan, Dan Du, Shengchao Zhao, Yi Ding, Bin Liu, Wenwu Sun, Fabrizio Albarello, Alessandra D'Abramo, Vincenzo Schininà, Emanuele Nicastri, Mariaelena Occhipinti, Giovanni Barisione, Emanuela Barisione, Iva Halilaj, Pierre Lovinfosse, Xiang Wang, Jianlin Wu, and Philippe Lambin. 2020. Development of a Clinical Decision Support System for Severity Risk Prediction and Triage of COVID-19 Patients at Hospital Admission: An International Multicenter Study. *European Respiratory Journal* 56, 2001104 (July 2020), 2001104. https://doi.org/10.1183/13993003.01104-2020

[168] J. Xing, K. Dong, X. Liu, J. Ma, E. Yuan, L. Zhang, and Y. Fang. 2024. Enhancing Gestational Diabetes Mellitus Risk Assessment and Treatment through GDMPredictor: A Machine Learning Approach. *Journal of Endocrinological Investigation* (March 2024). https://doi.org/10.1007/s40618-024-02328-z

[169] Xianglong Xu, Zhen Yu, Zongyuan Ge, Eric P F Chow, Yining Bao, Jason J Ong, Wei Li, Jinrong Wu, Christopher K Fairley, and Lei Zhang. 2022. Web-Based Risk Prediction Tool for an Individual's Risk of HIV and Sexually Transmitted Infections Using Machine Learning Algorithms: Development and External Validation Study. *Journal of Medical Internet Research* 24, 8 (Aug. 2022), e37850. https://doi.org/10.2196/37850

[170] Nadir Yalçın, Merve Kaşıkcı, Hasan Tolga Çelik, Kutay Demirkan, Şule Yiğit, and Murat Yurdakök. 2023. Development and Validation of Machine Learning-Based Clinical Decision Support Tool for Identifying Malnutrition in NICU Patients. *Scientific Reports* 13, 1 (March 2023), 5227. https://doi.org/10.1038/s41598-023-32570-z

[171] Hwai-I Yang, Morris Sherman, Jun Su, Pei-Jer Chen, Yun-Fan Liaw, Uchenna H. Iloeje, and Chien-Jen Chen. 2010. Nomograms for Risk of Hepatocellular Carcinoma in Patients With Chronic Hepatitis B Virus Infection. *Journal of Clinical Oncology* 28, 14 (May 2010), 2437–2444. https://doi.org/10.1200/JCO.2009.27.4456

[172] Chengdong Yu, Xiaolan Ren, Ze Cui, Li Pan, Hongjun Zhao, Jixin Sun, Ye Wang, Lijun Chang, Yajing Cao, Huijing He, Jin'en Xi, Ling Zhang, and Guangliang Shan. 2023. A Diagnostic Prediction Model for Hypertension in Han and Yugur Population from the China National Health Survey (CNHS). *Chinese Medical Journal* 136, 9 (May 2023), 1057–1066. https://doi.org/10.1097/CM9.0000000000001989

[173] Xia Zhu, Jun Lv, Meng Zhu, Caiwang Yan, Bin Deng, Canqing Yu, Yu Guo, Jing Ni, Qiang She, Tianpei Wang, Jiayu Wang, Yue Jiang, Jiaping Chen, Dong Hang, Ci Song, Xuefeng Gao, Jian Wu, Juncheng Dai, Hongxia Ma, Ling Yang,

Yiping Chen, Mingyang Song, Qingyi Wei, Zhengming Chen, Zhibin Hu, Hongbing Shen, Yanbing Ding, Liming Li, and Guangfu Jin. 2023. Development, Validation, and Evaluation of a Risk Assessment Tool for Personalized Screening of Gastric Cancer in Chinese Populations. *BMC Medicine* 21, 1 (April 2023), 159. `https://doi.org/10.1186/s12916-023-02864-0`

A Search terms

A.1 PubMed Search Term

”ML”[Title/Abstract] OR ”Artificial intelligence”[Title/Abstract] OR ”AI”[Title/Abstract] OR ”machine intelligence”[Title/Abstract] OR ”machine learning”[Title/Abstract] OR ”intelligent *”[Title/Abstract] OR ”expert system”[Title/Abstract] OR ”neural network”[Title/Abstract] OR ”natural language processing”[Title/Abstract] OR ”generative AI”[Title/Abstract] OR ”deep learning”[Title/Abstract] OR ”bayesian”[Title/Abstract] OR ”fuzzy logic”[Title/Abstract] OR ”Artificial Intelligence”[MeSH Terms] AND (”predictive modelling”[Title/Abstract] OR ”prediction modelling”[Title/Abstract] OR ”prognostic modelling”[Title/Abstract] OR ”Decision Support Tool”[Title/Abstract] OR ”decision support system”[Title/Abstract] OR ”risk prediction”[Title/Abstract] OR ”decision support systems, clinical”[MeSH Terms]) AND (”Journal Article”[Publication Type]) AND (”uncertainty”[Title/Abstract] OR ”risk”[Title/Abstract]) AND (”English”[Language]) NOT ”imaging”[Title/Abstract] NOT ”image”[Title/Abstract] NOT ”vision”[Title/Abstract] NOT ”literature review”[Title/Abstract] NOT ”scoping review”[Title/Abstract] NOT ”systematic review”[Title/Abstract] NOT ”environment*”[Title/Abstract] NOT ”veterinary”[Title/Abstract] NOT ”Organisims”[Title/Abstract] NOT ”Drug”[Title/Abstract]

A.2 Web of Science search term

The search can be accessed at: `https://www.webofscience.com/wos/woscc/summary/056d9f82-4755-4801-b331-50c6f9b65847-e17ef8f8/relevance/1`

A.3 IEEE Xplore Search Term

((”ML” OR ”Artificial intelligence” OR ”AI” OR ”machine intelligence” OR ”machine learning” OR ”expert system” OR ”neural network” OR ”natural language processing” OR ”generative AI” OR ”deep learning” OR ”bayesian” OR ”fuzzy logic” OR ”Artificial Intelligence”) AND (”predictive modelling” OR ”prediction modelling” OR ”prognostic modelling” OR ”Decision Support Tool” OR ”decision support system” OR ”risk prediction”) AND (”uncertainty” OR ”risk”) NOT (”imaging” OR ”image” OR ”vision” OR ”literature review” OR ”scoping review” OR ”systematic review” OR ”environment*” OR ”veterinary” OR ”Organisms” OR ”Drug”))

B Questions

The use of square brackets indicates that the answer was expected to be a categorical answer. In all cases *Other* answers are followed up by establishing what the other is.

QS0 – About the CDSSs

Q0.1 – *What is the area of medicine?*

Q0.2 – *Who is the intended user of the CDSS? [Clinician/Patient/Both/Other]*

Q0.3 – *What 'decision' is the CDSS 'supporting'?*

Q0.4 – *Is the CDSS trying to be analogous to clinical reasoning? [Yes/No]*

QS1 – What is the algorithm and its output?

The first set of questions is to consider what exactly in the underlying machine learning algorithm is within the model and what the results look like.

Q1.1 – *What is the algorithm?*

Neural Network, Logistic Regression, Linear Regression, SVM, random forest, etc.

Q1.2 – *What is the output of the algorithm? [Number/Classification/Both/Other]*

If *Number* then Q1.3 is asked.
If *classification* Q1.4 is asked.
If *Both* Q1.3 and Q1.4 is asked.

Q1.3 – *What is the number? [Probability/Score/Other]*

Q1.4 – *Is it classifying into risk levels? [Yes/No]*

Q1.5 – *Does there exist a threshold value that transforms a numeric value into a classification? [Yes/No]*

If the answer is *Yes* then Q1.6 is asked

Q1.6 – *How has that threshold been established?*

QS2 – Presentation of uncertainty

Q2.1 – *Is any uncertainty presented alongside the output? [Yes/No]*

If the answer is *No*, then no more questions in this question are asked.

Q2.2 – *What is the uncertain object?*

Q2.3 – *How is the uncertainty presented?*

QS3 – Assessment of performance

Q3.1 – *How is the performance of the CDSS assessed?*

C Data Tables

C.1 Medical specialities

Table 1: Medical Specialities of the included models

Specialty	Count	Reference	Specialty	Count	Reference
Oncology	21	9, 35, 89, 91, 97, 140, 42, 149, 53, 3, 43, 81, 130, 160, 132, 141, 27, 102, 155, 171, 173	Gastroenterology	2	95, 58
Surgery	19	17, 53, 99, 13, 105, 1, 14, 15, 108, 112, 58, 61, 128, 62, 87, 159, 119, 118, 156, 163	Hematology	2	74, 122
Cardiology	19	40, 74, 94, 172, 37, 162, 138, 22, 148, 105, 112, 122, 121, 79, 152, 103, 66, 158, 57	Nursing	1	25
Emergency Medicine	14	98, 16, 54, 129, 13, 117, 144, 119, 118, 150, 156, 71, 164, 34	Gynecology	1	27
Nephrology	12	10, 63, 92, 145, 77, 64, 111, 133, 84, 109, 28, 146	Transplant	1	50
Covid	11	73, 33, 23, 39, 4, 72, 136, 167, 51, 151, 171	Hepatology	1	50
Urology	8	12, 9, 39, 140, 77, 111, 133, 141, 155, 24, 71	Tropical Medicine	1	54
Diabeties	7	96, 127, 131, 159, 135, 67, 36	Sports Medicine	1	83
Orthopaedics	6	21, 88, 113, 61, 120, 62	Orthopedics	1	83
Diabetes	5	39, 114, 168, 26, 45	Rheumatology	1	85

Continued on next page

Specialty	Count	Reference	Specialty	Count	Reference
Hospital	5	80, 142, 116, 110, 165	Travel Medicine	1	86
Obstetrics	4	2, 168, 166, 36	Psycology	1	97
General Practice	4	98, 41, 117, 45	Social Care	1	98
Neurology	3	12, 24, 71	Neurosurgery	1	128
Otorhinolaryngology	3	18, 69, 56	Outpatients	1	144
Pulmonology	3	20, 74, 122	Malnutrition	1	147
Psychiatry	3	101, 30, 71	Oral Health	1	149
Sleep Medicine	3	78, 69, 56	Geriatric Care	1	150
Paediatrics	2	170, 5	Sexual Health	1	169

C.2 Use case of the CDSSs

Table 2: The use cases of the CDSSs.

Use	Count	Models
Risk of condition	47	3, 4, 22, 24, 25, 26, 27, 28, 30, 33, 42, 43, 45, 50, 54, 56, 57, 66, 67, 71, 72, 73, 77, 79, 80, 81, 84, 88, 101, 103, 113, 118, 120, 129, 133, 135, 136, 141, 142, 144, 148, 152, 162, 166, 169, 170, 171
Prediction/diagnosis of condition	34	2, 9, 10, 12, 16, 18, 21, 23, 35, 40, 51, 63, 64, 74, 78, 85, 89, 92, 91, 94, 96, 97, 114, 116, 127, 131, 140, 145, 146, 149, 155, 158, 168, 172
Outcome after intervention	24	13, 14, 15, 17, 53, 58, 61, 62, 95, 99, 102, 105, 108, 109, 111, 112, 121, 122, 128, 130, 132, 156, 163
Triage/Screening	11	5, 20, 34, 36, 37, 41, 69, 147, 160, 167, 173
Intervention recommendation	8	39, 83, 86, 87, 110, 119, 159, 164
Monitoring/Management	7	12, 98, 117, 138, 150, 151, 165
Mortality prediction	21	13 22 33 43 54 62 80 81 88 105 108 111 112 113 129 130 133 148 163

C.3 Algorithms used within the CDSSs

Table 3: Algorithm used by CDSSs

Algorithm	Count	Paper Reference
Logistic Regression	32	9, 12, 16, 17, 21, 33, 39, 40, 42, 53, 54, 64, 67, 69, 72, 83, 84, 99, 105, 108, 110, 112, 117, 120, 121, 122, 136, 141, 144, 146, 150, 151, 155, 156, 158, 164, 163
Random Forest	25	2, 23, 27, 41, 51, 63, 64, 71, 73, 77, 80, 81, 87, 89, 92, 97, 101, 128, 129, 147, 148, 167, 168, 169, 170
Gradient Boosting	18	1, 22, 28, 56, 57, 72, 81, 87, 88, 95, 96, 109, 114, 116, 128, 148, 149, 159
Proportional Hazards Models	15	3, 10, 26, 63, 66, 85, 111, 121, 130, 145, 151, 152, 162, 171, 173
Neural Network	12	12, 35, 50, 58, 64, 91, 101, 127, 135, 142, 148, 165
Regression	10	4, 15, 39, 61, 62, 103, 148, 170, 164, 172
Decision Tree	8	79, 98, 102, 132, 140, 164
Bayesian Network	8	86, 94, 113, 119, 118, 133, 160
Support Vector Machine	6	30, 36, 64, 81, 101, 148
Proprietary	5	14, 20, 43, 45, 131
Natural Language Processing	4	30, 37, 74, 98
Markov Model	2	24, 78
Fuzzy Logic	2	18, 34
Principle Component Analysis	1	12
Reinforcement Learning	1	138
Voting Classifier	1	148
Ensemble Prediction Model	1	128
Learning From Examples Using Rough Sets (Lers)	1	166
Arden Syntax Medical Logic Modules	1	5

C.4 Performance Metric

Table 4: Metrics used to assess the performance of the CDSSs

Metric	Count	Models
(AU)ROC	97	3, 4, 9, 10, 13, 14, 15, 16, 17, 18, 23, 22, 25, 27, 30, 33, 35, 36, 37, 39, 40, 43, 42, 45, 50, 51, 53, 54, 56, 57, 58, 61, 62, 63, 64, 66, 67, 69, 71, 72, 73, 74, 77, 79, 80, 81, 83, 85, 87, 88, 89, 92, 91, 94, 95, 96, 97, 98, 99, 101, 103, 105, 108, 109, 110, 111, 112, 113, 119, 118, 116, 117, 120, 121, 122, 127, 129, 128, 130, 132, 133, 135, 136, 138, 140, 141, 142, 144, 145, 146, 147, 148, 149, 150, 152, 155, 156, 160, 159, 158, 162, 165, 166, 164, 163, 167, 168, 169, 170, 171, 172, 173
Confusion Matrix Statistics	66	1, 2, 4, 5, 12, 13, 14, 16, 18, 20, 21, 23, 24, 25, 27, 28, 30, 34, 37, 41, 43, 50, 51, 53, 54, 56, 69, 71, 72, 74, 78, 79, 80, 84, 85, 87, 88, 91, 94, 96, 97, 101, 102, 108, 109, 116, 117, 122, 128, 129, 133, 140, 144, 146, 149, 150, 152, 160, 159, 158, 165, 164, 166, 167, 169, 170
Comparison To Existing Practice	11	1, 20, 25, 34, 37, 40, 51, 86, 98, 151, 155
Calibration Plot	10	15, 42, 57, 63, 83, 88, 112, 130, 141, 148, 173
Brier Score	8	4, 15, 57, 119, 118, 141, 148, 149
Hosmer-Lemeshow	7	15, 67, 105, 119, 118, 141, 148
(AU)PRC	7	56, 80, 87, 96, 147, 152, 168
DCA	4	9, 50, 163, 172
χ^2 Test	3	26, 91, 145
User Views	2	131,138
Wilcoxon Rank-Sum Test	1	9
Hazard Ratio	1	114
Mann-Whitney U Test	1	83
D-Statistic	1	66
r^2 Test	1	66
Calibration Plots	1	57
NNT	1	116
Unclear	1	45