Rethinking PICO in the Machine Learning Era: ML-PICO

Xinran Liu[1,2] James Anstey[1] Ron Li[3] Chethan Sarabu[4,5] Reiri Sono[2] Atul J. Butte[6]

[1] Division of Hospital Medicine, University of California, San Francisco, San Francisco, California, United States
[2] University of California, San Francisco, San Francisco, California, United States
[3] Division of Hospital Medicine, Stanford University, Stanford, California, United States
[4] doc.ai, Palo Alto, California, United States
[5] Department of Pediatrics, Stanford University, Stanford, California, United States
[6] Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, California, United States

Address for correspondence Xinran Liu, MD, MS, FAMIA, 2177 Third Street, Unit 502, San Francisco, CA 94107, United States (e-mail: xnrnliu@gmail.com).

Appl Clin Inform 2021;12:407–416.

Abstract

Background Machine learning (ML) has captured the attention of many clinicians who may not have formal training in this area but are otherwise increasingly exposed to ML literature that may be relevant to their clinical specialties. ML papers that follow an outcomes-based research format can be assessed using clinical research appraisal frameworks such as PICO (Population, Intervention, Comparison, Outcome). However, the PICO frameworks strain when applied to ML papers that create new ML models, which are akin to diagnostic tests. There is a need for a new framework to help assess such papers.

Objective We propose a new framework to help clinicians systematically read and evaluate medical ML papers whose aim is to create a new ML model: ML-PICO (Machine Learning, Population, Identification, Crosscheck, Outcomes). We describe how the ML-PICO framework can be applied toward appraising literature describing ML models for health care.

Conclusion The relevance of ML to practitioners of clinical medicine is steadily increasing with a growing body of literature. Therefore, it is increasingly important for clinicians to be familiar with how to assess and best utilize these tools. In this paper we have described a practical framework on how to read ML papers that create a new ML model (or diagnostic test): ML-PICO. We hope that this can be used by clinicians to better evaluate the quality and utility of ML papers.

Keywords
- machine learning
- electronic health record
- artificial intelligence

Background and Significance

The topic of machine learning (ML) has become increasingly prominent in clinical medicine. A search of PubMed using the Medical Subject Headings (MeSH) term "Machine Learning" returns 49 results for the year 2010, but over 7,000 results in 2019—over a hundred-fold increase. Physicians in fields such as Radiology, Neurology, Pathology, Gastroenterology, Cardiology, and others all see ML as the next enabler of innovation and advancement in their respective fields and specialty journals.[1–4] Despite the growing importance of ML in medicine, clinicians have historically received little training to systematically and

received
November 17, 2020
accepted after revision
March 24, 2021

Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/s-0041-1729752.
ISSN 1869-0327.

thoughtfully evaluate ML research papers.[5,6] Although there are publications that try to address this need,[7,8] they tend to be written from a bird's eye view, or more technical perspective, and as a result might be difficult for physicians without prior ML training to apply in practice.

This paper provides a simple and practical framework for clinicians to systematically read and evaluate medical ML papers whose aim is to create a new ML model: ML-PICO (Machine Learning, Population, Identification, Crosscheck, Outcomes). We hope that this will empower clinicians to more rigorously assess vendors, startups, and researchers who approach health systems with, at times, exaggerated promises from their ML based solutions.[9] This framework is inspired by the PICO (Population, Intervention, Comparison, Outcome) framework, which is a methodology to help formulate research questions that can also be used to digest a research paper to its core components.[10] As the formal PICO framework strains when applied to diagnostic tests and ML papers,[11,12] we have created ML-PICO to fill this void. As we discuss and demonstrate this new framework, will refer to a use case of using ML techniques to model sepsis in hospitalized patients as a guide.

Machine Learning—ML

When thinking about how to read medical ML papers, it can be helpful to imagine two separate approaches we see regularly in clinical research. The first is the study of a new diagnostic test, and the evaluation of its value when compared with an established gold standard (e.g., Wells' criteria for deep venous thrombosis, Ranson's criteria for pancreatitis). The second approach is outcomes-based research, where cohort(s) of subjects are, or have been, exposed to various treatments and controls, and are studied to better understand clinical outcomes.[13] Many papers that clinicians read follow the format of outcome-based research, so there might be a tendency to assume that the same is true for ML papers. However, the reality is that many ML papers are effectively creating new "diagnostic tests" in the form of an ML model and rarely go as far as showing that these new tests can improve clinical outcomes.[14,15]

When reading a medical ML paper, it is important to identify which of the two approaches is being presented. ML papers that describe outcome-based research (e.g., testing the impact of implementing an ML model on patient outcomes) are more ideal, and offer stronger evidence of clinical utility. These are the types of medical ML papers that readers should focus on looking for and drawing conclusions from. For such papers, applying the traditional PICO framework to assess their quality works well. Unfortunately, only a minority of ML papers do so,[16,17] and even fewer are able to show improved outcomes prospectively.

In contrast, ML papers that use retrospective data to create ML models (new diagnostic tests) have value but do not in isolation tell us if the ML models have clinical utility.[18] Despite a lack of clear utility, such papers tend to grab attention with news headlines often in the form of: Artificial Intelligence better than physicians at diagnosing X, Y, or Z.[14,19] To know if those claims can be justified requires, as with any new clinical test/drug, external validation. Simple validation can be testing the model on a different retrospectively collected dataset.[16] More ideally though, the model should be tested prospectively in an interventional trial format to see if it can truly improve clinical outcomes.[20] For ML papers that aim to create a new diagnostic test in the form of an ML model, the traditional PICO framework does not fit or perform well to assess their quality. Instead, we propose using ML-PICO.

Population—P

When evaluating clinical research, it is important to know the characteristics of the population studied. Often summarized in "Table 1," these include variables such as age, gender, race, relevant comorbidities, and more.[21] This information is important because it helps clinicians make decisions on whether or not the results of the study are applicable to the patient sitting in front of them.

Similarly, ML models are created from a population of data, which means that it is important to know the characteristics of that data, and whether or not a model built on that data might be applicable to different settings[18] (i.e., Table 1 for the data). Referring to our case example, numerous models to predict sepsis have been created using the Medical Information Mart for Intensive Care (MIMIC) dataset,[22–25] which is an open, de-identified clinical dataset from the intensive care unit (ICU) at the Beth Israel Deaconess Medical Center in Boston, Massachusetts, United States.[25] Patients in the ICU are not only sicker than those who are in non-ICU settings, but also experience different treatments (e.g., vasopressors, sedation, intubation), and generate different frequencies of data (e.g., many sets of vital signs and laboratory measurements per day vs. a few per day). For these reasons, an ML model created using the MIMIC dataset might not generalize well to different clinical settings such as the general medicine wards, or to a nonacademic hospital setting. This is referred to as *dataset shift*,[26] in which a model fails to generalize due to differences between the data used to create the model, and the data seen during deployment. This can lead to significantly decreased model performance when the model is applied to new settings.[18,26]

Another important question to consider about the data is its quality. While quality can be subjective, there are some generally agreed upon measures of data quality, which include: (1) data inaccuracy, (2) data missingness, and (3) selective measurements.

Data inaccuracy would seem to be straightforward, being either accurate or inaccurate. However, there is a less intuitive cause of data inaccuracy, which is the variable completeness and standardization of data from the electronic health record (EHR).[9,27] For example, one hospital may have a workflow to ensure that a patient's problem list is up to date at all times, while another might not use the problem list much at all. One hospital might flag patients with sepsis using the systemic inflammatory response syndrome criteria, while another might use the sepsis-3 criteria. There are significant efforts under way to standardize data definitions (e.g., Fast Healthcare Interoperability Resources), but these standards still depend

on accurate data mapping, which is not always a straightforward process.

Next is data missingness. From the perspective of computer science, all health care data has some degree of missingness.[28] For example, vital signs are generally obtained every 4 to 8 hours, meaning that if a predictive model looked for values every hour, the majority of 1-hour intervals will have no values. There are methods to handle this type of missing data, called imputation.[29] However, an easy to overlook type of missing data is the data that exists and should be included, but is not.[9] If incomplete data are used to create an ML model, the model could perform poorly as a result.[30] For example, a patient might see multiple physicians in multiple health systems, and important information might not be available because his or her data at another health system cannot be accessed.[31]

The last measure of data quality is selective measurement.[9] A good performing model should be built on data that is as complete and as relevant to a specific question as possible. For each patient encounter, a patient might have numerous types of data generated including insurance claims, EHR structured data, images, notes, etc.[32] Despite a plethora of data sources, it is not uncommon to see models built using only one source of data, such as claims data. When reading an ML paper, it is important to appreciate whether the clinical question the ML model is built to answer can be answered with the type(s) of data being used.[33]

Identification—I

In medicine, the majority of ML models (including deep learning ones) are of the supervised learning variety,[34] in which "gold standard" labels are required to be identified for the data being used.[7] For example, an ML model to identify cases of sepsis from EHR data would require structured data elements from the EHR, as well as clear labels for which patients in the dataset had sepsis, and when. Once the data and labels for the data are obtained, the ML model is able to mathematically find patterns and relationships in the data that best correlate with the labeled outcome that was provided. This process is how an ML model is "trained." Afterward, it can be fed new data to get its prediction(s). Its predictions can then be compared with the actual outcome to determine its effectiveness. Different ML algorithms use different mathematical and statistical methods to do so, but generally work using similar concepts[8,35] (►Fig. 1).

Previously, the importance of the quality of data was emphasized, so it should not be surprising that how the data are labeled is just as important.[34] Going back to our sepsis example, how should cases of sepsis from EHR data be labeled? Should the sepsis-2 or sepsis-3 definition be used? International Classification of Diseases, Tenth Revision, (ICD-10) diagnosis codes could be used to identify positive cases of sepsis,[22,36,37] but diagnosis codes are generally unreliable,[38] with the sensitivity of explicit sepsis codes only between 30 and 50%.[39,40] A better gold standard for sepsis than ICD-10 diagnosis codes might make the generated model more relevant.

One could thus decide not to use ICD-10 diagnosis codes and instead use "clinical criteria." However, if there were a hundred thousand encounters in the dataset, it might be infeasible to manually review all these charts and assign individual labels. In addition, what level of experience is needed for someone to be qualified to assign labels? Should it be experts in the field, experienced attendings, or fellows in

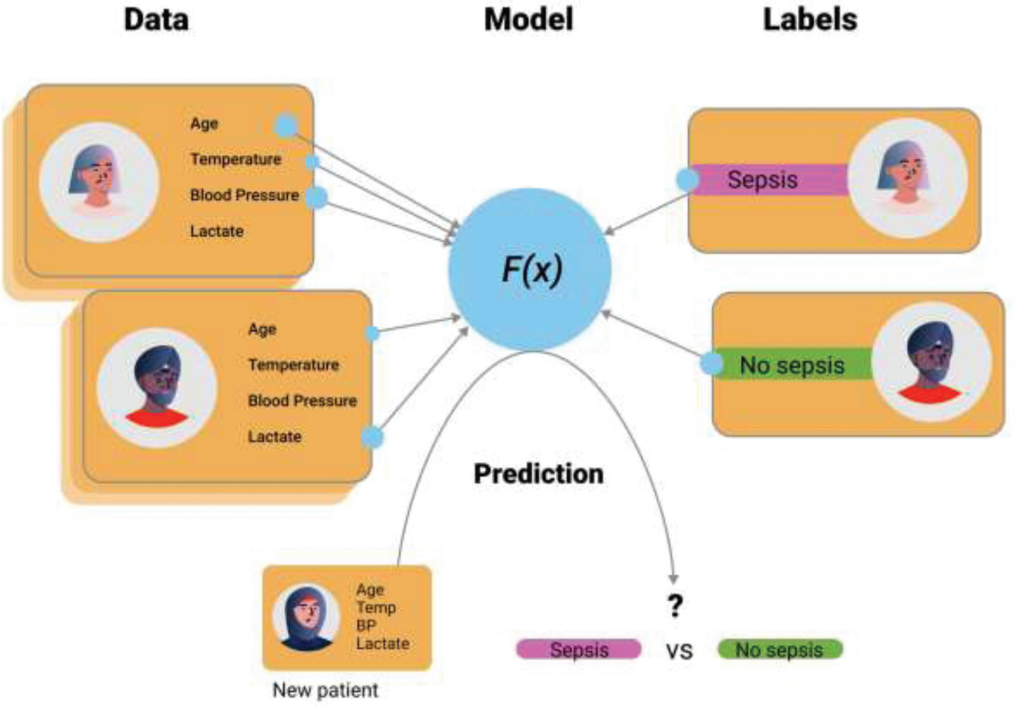

Fig. 1 Data and labels for that data are used to train the ML model. New cases can then be input into the model to make predictions. ML, machine learning.

training? In reality, it is not uncommon to see medical students or residents defining these sources of truth.[41]

It is important to acknowledge that there is often not a perfect definition for many clinical conditions (e.g., cancer).[34] There will always be some uncertainty regarding accurate clinical labels. Still, these decisions are clinical ones and should be scrutinized by experienced clinicians. The key takeaway is that how a "gold standard" label is identified is important when training a supervised ML model, and this information should be transparent so that readers can best determine if the model is reliable and applicable to their clinical needs.[3]

Lastly, it is also important to pay attention to the clinical use case to which the ML model is being applied. ML models are powerful enough to make accurate predictions on any dataset that you have good data and labels for. Unfortunately, such models might end up providing only miniscule benefit if there is little ability to change the outcome even given an accurate prediction (e.g., predicting 1 year mortality in patients with terminal cancer).[31,32]

Crosscheck—C

The next step is to crosscheck the paper for ML modeling best practices. While good external validation of an ML model would trump any forms of internal validation, there are still best practices for internal validation to look for in ML papers that are worth mentioning. One such best practice is to break the original retrospective data into three separate datasets, a training set, a development set, and a test set.[42] The training set is used to train the ML model. The development set is used to optimize it. The test set is used to evaluate the true performance of the model. Think of the test set as the best available alternative to external validation. If the model's performance is significantly worse in the test set compared with the training or validation sets, this suggests that the model will not generalize well and is not worth pursuing further.[42] If the model's performance is high in all datasets, this gives us confidence that the model could generalize well to other settings and is worth pursuing further external validation.

When reading an ML paper that creates a new ML model, it is important to know if the outcomes that are described in the paper are from the training set, development set, or test set. The outcomes with the most value are those from the test set.[42] If there is no clarification on which dataset an ML model's performance came from, or if there was no dataset splitting at all, it raises a red flag about the methodology of the paper. It is important to note that cross validation is another acceptable method of internal validation. In this approach, all available data are divided into a number (n) of partitions. A model is built on ($n-1$) of these partitions, then validated on the remaining partition that was not used to train that model. This is repeated for all possible combinations of ($n-1$) partitions and resultant validation sets. The performance of each model from each partition of data are then averaged together to determine the overall model and model performance.[43] As health care data and predictions often include aspects of time (e.g., changes in practice trends and disease prevalence over time), there is a theoretical concern that cross validation could lead to more error compared with using a test set from the most recent and relevant time period. However, this concern has not been confirmed to be a major issue in practice so far.[44,45]

Another best practice that is worth checking for is temporality. When training an ML model, it is vital to ensure that the data used is consistent with the data that will be available when the ML model is implemented. For example, claims data and ICD codes can offer useful information about patient outcomes, but might not be assigned until long after a patient encounter with the health care system has been completed.[32] Including claims data when training a real-time ML model to predict sepsis onset during hospitalization might seem to improve model accuracy during training, but when tested in the real world would likely underperform. This is because the claims data are providing extra information that will not be present when the model is implemented in the real world. This is referred to as *data leakage*.[46]

Outcomes—O

In clinical trials, physicians are taught to look at primary and secondary outcomes.[47] However, ML papers that create new ML models from retrospective data often use outcome metrics that are different and less familiar to physicians.

There are numerous different outcome metrics that can be seen in ML papers, and a list of common one can be seen in ►Table 1.[48–51] An important distinction to make is whether or not the metric is a classification or regression metric. Classification metrics measure how well a model discriminates between different classes (e.g., sepsis present vs. not present). Common classification metrics include the receiver operating characteristic (ROC) and precision–recall (PR) curves, as well as the area under the curve (AUC) for both respective graphs (►Fig. 2a, b).[49] Regression metrics measure how good a model is at calibration, meaning how close to the true value the model's predictions are (e.g., predicted length of stay).[51] We will mainly focus on the classification metrics mentioned above here, but it is important to note that many other valid metrics can be seen in ML papers other than AUC or ROC.

The AUC is a common metric seen in ML papers. To understand it, one has to understand classification thresholds.[52] Classification models output a probability. For example, the probability that a patient has sepsis could be 70, 63, or 15%. As this is a classification problem in which the model should only output "yes" or "no" to having sepsis, the probability has to be translated to "yes" or "no." This can be done by selecting a threshold. If a threshold of 50% is set for having sepsis, then the first two cases above would be classified as "yes," and the last one "no." If a threshold of 80% is set instead, then all the cases would be classified as "no."

To generate the ROC curve, one calculates and plots the model's performance in terms of the confusion matrix (►Fig. 2c) for every possible threshold between 0 and 100%, the result is the ROC curve (►Fig. 2a). The area under this graph is the AUC (also called AUROC, or C statistic), and is an aggregate measure of overall performance of the model over all possible threshold values. AUC values are between 0 and 1, with higher values being better.[52] A PR curve works

Table 1 Regression and classification ML outcome metrics

Outcome metric	Description	Pros/Cons
Regression		
Root mean squared error (RMSE) $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$	The square root of the sum of the differences between the prediction values and observed values squared, and divided by the number of samples *n*.	Generally, the standard regression metric used as easy to do certain mathematical operations. Penalizes predictions that are significantly different from observed values more than MAE as values are squared. Values can range from 0 to infinity.
Mean absolute error (MAE) $\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}\lvert y_j - \hat{y}_j\rvert$	The sum of the absolute differences between the predicted values and observed values, divided by the number of samples *n*.	Penalizes predictions that are significantly different from observed values less than RMSE, so less affected by outliers. Values can range from 0 to infinity.
R^2 R^2 $\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$	Is equal to 1 minus the ratio of the sum of differences between the prediction values and observed values squared, and the sum of the differences between the average observed value and observed values squared.	Is a measure to compare how well the model performs relative to predicting the average observed value. An R^2 value of 0 suggests similar performance to predicting the average. An R^2 value below 0 suggests worse performance than predicting the average. An R^2 above 0 suggests better performance.
Classification		
Sensitivity (Recall) $TP/(TP + FN)$	Proportion of patients with the condition who test positive.	Measures test performance without accounting for prevalence of the condition in the population. In conditions with low prevalence, positive predictive value can be low despite high sensitivity.
Positive predictive value (Precision) $TP/(TP + FP)$	Proportion of patients who test positive who are positive for the condition.	In conditions with high prevalence, sensitivity of the test can be low despite high positive predictive values.
F_1 score $F1 = 2 \times Precision \times Recall/(Precision + Recall)$	F1 Score is the harmonic mean between precision and recall, meaning that it is trying to balance the importance of both metrics.	In cases in which precision and recall are both important, F1 score can be used to incorporate the importance of both measures into one. F1 score mainly penalizes lower values for either precision or recall.
Accuracy $(TP + TN)/(N + P)$	Simplest metric to use that looks broadly at correct predictions over all predictions.	Not a good measure when there is a class imbalance. For example, if the prevalence of a condition is only 1%, just guessing the absence of the condition will lead to 99% accuracy.

Abbreviations: FN, number of false negatives; FP, number of false positives; ML, machine learning; *N*, number of actual negatives; *P*, number of actual positives; TN, number of true negatives; TP, number of true positives.

similarly to an ROC curve, except that the y-axis is precision (or positive predictive value), and the x-axis is recall (or sensitivity) (►**Fig. 2b**).

It is natural to want to use higher AUC values as an indicator for better ML models and papers. However, this single number interpretation can be very misleading,[53] and there are some additional considerations that can help evaluate an ML paper more holistically. The first is to determine if the AUC is from a ROC curve, or a PR curve. The second is to have an estimate of the prevalence of the condition being evaluated. ROC curves measure the performance of the model independent of the prevalence of the condition, whereas PR curves are affected by prevalence.[54] This means that as the prevalence of a condition goes under 10%, and especially if it falls under 5%, the AUC from the PR curve will start to become significantly lower compared with the AUC from the ROC curve.[55] As most conditions in medicine have low prevalence (i.e., under 10%),[56–58] it is worth making the argument that the PR curve, and its AUC, might be a more honest and clinically relevant metric to look at than that from the ROC curve.[59,60] When comparing the AUCs from ML papers to each other, it is important to ensure that one is comparing apples to apples, rather than apples to oranges.

Furthermore, when evaluating an ML model, clinicians should also consider what constitutes an acceptable performance level, or clinical utility, for that model.[53,61] In clinical

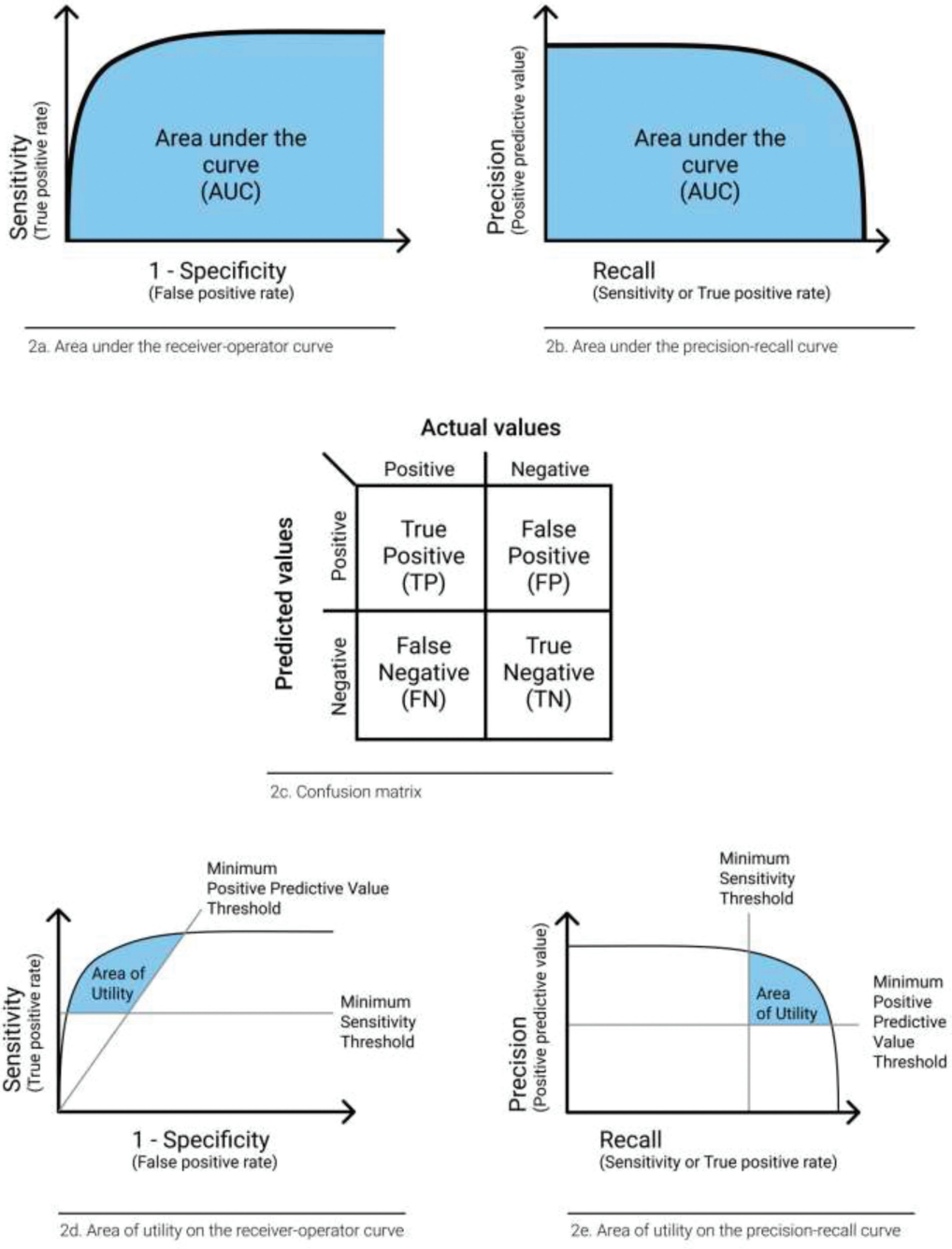

Fig. 2 (**a**) Area under the curve for the ROC curve. (**b**) Area under the curve for the PR curve. (**c**) Example of the confusion matrix. (**d**) Area of utility for ROC curve. (**e**) Area of utility for PR curve. PR, precision–recall; ROC, receiver operating characteristic.

medicine, it is frowned upon to order tests that have poor utility, and so the same expectations should exist for ML models. When an ML model is being considered for implementation, there should be expectations on what minimal sensitivity and positive predictive value (PPV) thresholds are clinically acceptable. These thresholds are easy to draw on the PR curve (►**Fig. 2d**) and help visually identify if the model meets minimal requirements (area of utility). Assuming a disease prevalence is known or able to be estimated, the area of utility can also be drawn out on the ROC curve based on minimal thresholds for sensitivity and PPV (►**Fig. 2e**).[62] More details on this can be found in ►**Appendix A**.

If the ML model is not able to meet the minimal sensitivity and PPV thresholds required to be clinically useful, then it may not be worth implementing. Sensitivity and PPV are emphasized above as they are familiar metrics for clinicians. However, recent work has introduced more advanced ideas for measuring the clinical utility of an ML model by including both the effectiveness of the model at making predictions as well as the effectiveness of the downstream intervention(s). This metric is called number needed to benefit and is the product of the number needed to screen (NNS) and the number needed to treat (NNT). The NNS is equal to the reciprocal of the PPV and is purely a

Table 2 Summary of important questions to consider when reading clinical ML papers, organized by ML-PICO

	Description	Important questions to ask when reading clinical ML papers
M L	Machine learning: Type of ML paper	• Is the ML paper focused on creating a new ML model (diagnostic test), or does it follow an outcomes-based research format? ◦ If following outcomes-based research format, the traditional PICO framework can be used to appraise the paper. ◦ If creating a new ML model, proceed with using the ML-PICO framework.
P	Population: Characteristics of the population of data used to create the ML model	• Do the characteristics of the original model's data match that of the setting where the model is anticipated to be implemented into (i.e., similar Table 1)? ◦ Is there risk for database shift as a result? • What is the quality of the data used to create the model? ◦ How accurate is it, and how do you know? ◦ Is there missing data? If so, is that acceptable or would it introduce significant bias? ◦ Can the clinical question being addressed be answered with the type(s) of data sources being used?
I	Identification: How gold standard labels are identified from the data?	• Is the use case that this ML model will be applied to clinically relevant? ◦ Can outcomes be improved by the presence of this new prediction? ◦ Can this ML model be implemented seamlessly into workflow? • How is the clinical concept defined in the data (i.e., how are gold standard labels identified from the data)? ◦ If via ICD-10 codes, do they appropriately capture the target condition? ◦ If via "clinical criteria," is it consistent with established clinical definitions? ◦ Who is responsible for giving the final label and is his or her expertise appropriate?
C	Crosscheck: Checking for ML best practices	• If the paper is creating a new ML model, did the authors describe how they divided the data into training, development, and test sets? Or did they use cross validation? ◦ What dataset did the outcomes described in the paper come from? • Is the data used to train the ML model reflective of the actual data that the model will have access to when implemented in the real world? • Are there signs of database leakage?
O	Outcome: What are the outcomes that are emphasized in the paper?	• What ML outcomes are being emphasized in the paper? ◦ If AUC is the main outcome, is it from the ROC curve or the PR curve? ◦ Is the medical condition being targeted common or rare (i.e., under 10% prevalence). ◦ Based on the prevalence, is the PR curve or ROC curve and its respective AUC more appropriate? • For the use case that this ML model is being applied to, what are the minimally accepted sensitivity and PPV that would make this model clinically useful? ◦ Does the model meet these thresholds (i.e., intersect the area of utility on ROC or PR curve)? • What is the number needed to benefit from the ML model and its workflow implementation? ◦ Is this worth pursuing?

Abbreviations: ICD, International Classification of Diseases, Tenth Revision; ML, machine learning; PICO, Population, Identification, Crosscheck, Outcomes; PR, precision–recall; ROC, receiver operating characteristic.

result of the performance of the model. The NNT, on the other hand, is a result of the effectiveness of expected intervention which results from a positive alert.[48] This allows clinicians to weigh the benefit of an ML model and its implementation more rigorously, and whether or not it can truly address their needs.

Conclusion

As the importance of ML to clinical medicine increases in the future, the number of papers written in this area will continue to grow. It would be increasingly important for clinicians to be familiar with such publications and how to properly assess and utilize them in their practice. In this paper we have described a practical framework on how to read ML papers that create a new ML model (or diagnostic test): ML-PICO. We hope that this can be used by clinicians to better evaluate the quality and utility of ML papers. A summary of our key points organized by ML-PICO can be found in ►Table 2.

Limitations

The main limitation to this paper is that the ML-PICO framework still needs formal validation.

Clinical Relevance Statement

The ML-PICO (Machine Learning, Population, Identification, Crosscheck, Outcomes) framework can be adapted to help clinicians systematically read ML papers, so that they can better evaluate the quality and utility of ML papers for their clinical practice and use case.

Multiple Choice Questions

1. Machine learnings papers that focus on the creation of a new ML model from retrospective data provide a level of evidence similar to which of the following?
   a. Randomized control trials
   b. Prospective cohort study
   c. Retrospective diagnostic test creation
   d. Meta-Review

   Correct Answer: The correct answer is option c. Most machine learning papers are only creating new diagnostic test(s) (e.g., Wells Criteria) based on retrospective data, and are not prospectively validated.

2. When a machine learning algorithm is created for a clinical condition with a natural low prevalence rate (e.g., <10%), which of the following are important outcome metrics to pay attention to?
   a. Sensitivity
   b. Positive predictive value
   c. Area under the precision-recall curve (AUPR)
   d. All of the above

   Correct Answer: The correct answer is option d. There is not one metric that is always more important than the other ones. This depends on the use case as well as the goals of algorithm. However, when the prevalence of the condition, that the machine learning algorithm is designed to address, is naturally low, metrics like positive predictive value and AUPR become more relevant as they highlight the benefit versus cost of high false alarm rates.

Protection of Human and Animal Subjects

Human and/or animal subjects were not included in this project.

Conflict of Interest

None declared.

Acknowledgments

We would like to thank and acknowledge Dr. Michael Wang and Dr. Dana Ludwig for their contribution, advice, and support throughout this process.

References

1 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(01):44–56
2 Peterson ED. Machine learning, predictive analytics, and clinical practice: can the past inform the present? JAMA 2019;322(23): 2283–2284
3 Sevakula RK, Au-Yeung WM, Singh JP, Heist EK, Isselbacher EM, Armoundas AA. State-of-the-art machine learning techniques aiming to improve patient outcomes pertaining to the cardiovascular system. J Am Heart Assoc 2020;9(04):e013924
4 Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88
5 Forney MC, McBride AF. Artificial intelligence in radiology residency training. Semin Musculoskelet Radiol 2020;24(01):74–80
6 Weisberg EM, Fishman EK. Developing a curriculum in artificial intelligence for emergency radiology. Emerg Radiol 2020;27(04): 359–360
7 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380(14):1347–1358
8 Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019;322(18):1806–1816
9 Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. JAMA 2018;320(01):27–28
10 Aslam S, Emmanuel P. Formulating a researchable question: a critical step for facilitating good clinical research. Indian J Sex Transm Dis AIDS 2010;31(01):47–50
11 Leeflang MMG, Allerberger F. How to: evaluate a diagnostic test. Clin Microbiol Infect 2019;25(01):54–59
12 Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ 1986;134(06):587–594
13 Thiese MS. Observational and interventional study design types; an overview. Biochem Med (Zagreb) 2014;24(02):199–210
14 Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689
15 Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9(05):1–12
16 Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353: i3140
17 van der Ploeg T, Nieboer D, Steyerberg EW. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. J Clin Epidemiol 2016;78:83–89
18 Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc 2019;26(12):1651–1654
19 Antonelli M, Johnston EW, Dikaios N, et al. Machine learning classifiers can predict Gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists. Eur Radiol 2019; 29(09):4754–4764
20 Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. NPJ Digit Med 2020;3:107
21 Bowman L, Mafham M, Wallendszus K, et al; ASCEND Study Collaborative Group. Effects of aspirin for primary prevention in persons with diabetes mellitus. N Engl J Med 2018;379(16):1529–1539
22 Calvert JS, Price DA, Chettipally UK, et al. A computational approach to early sepsis detection. Comput Biol Med 2016;74:69–73
23 Desautels T, Calvert J, Hoffman J, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform 2016;4(03):e28
24 Shashikumar SP, Stanley MD, Sadiq I, et al. Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. J Electrocardiol 2017;50(06):739–743
25 MIT Critical Data. Secondary Analysis of Electronic Health Records Cham: Springer International Publishing; 2016
26 Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics 2020;21(02):345–352
27 He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25(01):30–36
28 Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020;2020:191–200
29 Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol 2015;15:30

30 Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020;27 (03):491–497
31 Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Invest 2020;130(02):565–574
32 Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018;25(10):1419–1428
33 Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178(11):1544–1547
34 Adamson AS, Welch HG. Machine learning and the cancer-diagnosis problem—no gold standard. N Engl J Med 2019;381(24): 2285–2287
35 Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317–1318
36 Mao Q, Jay M, Hoffman JL, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open 2018;8(01): e017833
37 Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. Acad Emerg Med 2016;23(03):269–278
38 Rudrapatna VA, Glicksberg BS, Avila P, Harding-Theobald E, Wang C, Butte AJ. Accuracy of medical billing data against the electronic health record in the measurement of colorectal cancer screening rates. BMJ Open Qual 2020;9(01):e000856
39 Iwashyna TJ, Odden A, Rohde J, et al. Identifying patients with severe sepsis using administrative claims: patient-level validation of the angus implementation of the international consensus conference definition of severe sepsis. Med Care 2014;52(06): e39–e43
40 Rhee C, Dantes R, Epstein L, et al; CDC Prevention Epicenter Program. Incidence and trends of sepsis in US Hospitals using clinical vs claims data, 2009-2014. JAMA 2017;318(13):1241–1249
41 Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the short-duration 12-lead ECG using a deep neural network: the CODE Study. ArXiv190401949 Cs Eess Stat. Published online April 1, 2019. Accessed May 11, 2019 at: http://arxiv.org/abs/1904.01949
42 Ng A. Machine Learning Yearning. deeplearning.ai; 2017. Accessed May 11, 2019 at: https://www.deeplearning.ai/machine-learning-yearning/
43 Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper presented at: Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 2. Montreal, Canada: Morgan Kaufmann Publishers Inc.;1995:1137–1143
44 Roberts DR, Bahn V, Ciuti S, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 2017;40(08):913–929
45 Rabinowicz A, Rosset S. Cross-validation for correlated data. J Am Stat Assoc 2020;97:883–897
46 Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data 2012;6(04):1–21
47 Vetter TR, Mascha EJ. Defining the primary outcomes and justifying secondary outcomes of a study: usually, the fewer, the better. Anesth Analg 2017;125(02):678–681
48 Liu VX, Bates DW, Wiens J, Shah NH. The number needed to benefit: estimating the value of predictive analytics in healthcare. J Am Med Inform Assoc 2019;26(12):1655–1659
49 Rácz A, Bajusz D, Héberger K. Multi-level comparison of machine learning classifiers and their performance metrics. Molecules 2019;24(15):E2811
50 Botchkarev A. A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdiscip J Inf Knowl Manag 2019;14:45–76
51 Dangeti P. Statistics for Machine Learning. Birmingham, United Kingdom: Packt Publishing Ltd; 2017
52 Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA 2017;318(14):1377–1384
53 Shah NH, Milstein A, Bagley PhD SC. Making machine learning models clinically useful. JAMA 2019. Doi: 10.1001/jama.2019.10306
54 Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. Crit Care 2015;19(01):285
55 Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol 2015;68(08):855–859
56 Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin 2015;65(02): 87–108
57 Ziaeian B, Fonarow GC. Epidemiology and aetiology of heart failure. Nat Rev Cardiol 2016;13(06):368–378
58 Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care 2004;27(05):1047–1053
59 Rough K, Dai AM, Zhang K, et al. Predicting inpatient medication orders from electronic health record data. Clin Pharmacol Ther 2020;108(01):145–154
60 Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Paper presented at: Proceedings of the 23rd International Conference on Machine Learning—ICML '06. Pittsburgh, Pennsylvania: ACM Press;2006:233–240
61 Matheny M, Thadaney Israni S. Artificial Intelligence Special Publication. NAM Special Publication; 2019. Accessed February 14, 2020 at: https://nam.edu/artificial-intelligence-special-publication/
62 Thomas G, Kenny LC, Baker PN, Tuytten R. A novel method for interrogating receiver operating characteristic curves for assessing prognostic tests. Diagn Progn Res 2017;1(01):17

Appendix A

Finding the positive predictive value (PPV) threshold to the area of utility on the receiver operating characteristic (ROC) curve

Assuming it is possible to provide a reasonable estimate of the prevalence of clinical condition that the machine learning (ML) model is built to address, the minimum PPV threshold border can be drawn on the ROC curve by the equation:

"Equation 1"

"Equation 1"

$$Sensitivity = \frac{PPV(min\_goal)}{1 - PPV(\min\_goal)} * \frac{1 - Prevalence}{Prevalence} * (1 - Specificity)$$

The area of utility (►Fig. 2D) can then be defined as the area bordered by the above equation, the minimum sensitivity threshold, and the ROC curve itself.

An html program that can help visualize the area of utility on both the ROC and PR curves based on minimum sensitivity, PPV thresholds, and estimated prevalence was also developed, and available for sharing by request. Please email xinran.liu@ucsf.edu.

