A Systematic Review of Machine Learning-based Prognostic Models for Acute Pancreatitis: Towards Improving Methods and Reporting Quality

Brian Critelli1, Amier Hassan1, Ila Lahooti2, Lydia Noh3, Jun Sung Park2, Kathleen Tong2, Ali Lahooti1, Nate Matzko1, Jan Niklas Adams4, Lukas Liss4, Justin Quion5, David Restrepo5, Melica Nikahd6, Stacey Culp6, Adam Lacy-Hulbert7, Cate Speake8, James Buxbaum9, Jason Bischof9, Cemal Yazici11, Anna Evans Phillips12, Sophie Terp13, Alexandra Weissman14, Darwin Conwell15, Phil Hart2, Mitch Ramsey2, Somashekar Krishna2, Samuel Han2, Erica Park2, Raj Shah2, Venkata Akshintala16, John A Windsor17, Nikhil K Mull18, Georgios I Papachristou2, Leo Anthony Celi5,19, Peter J Lee2*

Short title: A Systematic Review of Machine Learning Prognostic Models for Acute Pancreatitis

Author affiliations:
1. Weill Cornell Medical College, Division of Gastroenterology and Hepatology
2. Ohio State University Wexner Medical Center, Division of Gastroenterology and Hepatology
3. Northeast Ohio Medical School
4. Rheinisch-Westfälische Technische Hochschule Aachen University, Division of Process and Data Science
5. Massachusetts Institute of Technology, Laboratory for Computational Physiology
6. Ohio State University Wexner Medical Center, Division of Bioinformatics
8. Center for Interventional Immunology, Benaroya Research Institute at Virginia Mason, Seattle, Washington.
9. University of Southern California, Division of Gastroenterology
10. Ohio State University Wexner Medical Center, Division of Emergency Medicine
11. University of Illinois at Chicago, Division of Gastroenterology
12. University of Pittsburgh Medical Center, Division of Gastroenterology
13. University of Southern California, Division of Emergency Medicine
14. University of Pittsburgh Medical Center, Division of Emergency Medicine
15. University of Kentucky, Department of Medicine
16. Johns Hopkins Medical Center, Division of Gastroenterology
17. University of Auckland, Surgical and Translational Research Centre
18. University of Pennsylvania, Division of Hospital Medicine and Penn Medicine Center for Evidence-based Practice
19. Beth Israel Medical Center, Division of Critical Care

*: Amier Hassan and Brian Critelli are co-first authors (i.e., these authors contributed equally to this work)

*Corresponding author: Peter J Lee
Email: Peter.Lee@osumc.edu

Conflicts of interest: all authors declare no conflicts of interest
Title: A Systematic Review of Machine Learning-based Prognostic Models for Acute Pancreatitis: Towards Improving Methods and Reporting Quality

Background: An accurate prognostic tool is essential to aid clinical decision making (e.g., patient triage) and to advance personalized medicine. However, such prognostic tool is lacking for acute pancreatitis (AP). Increasingly machine learning (ML) techniques are being used to develop high-performing prognostic models in AP. However, methodologic and reporting quality has received little attention. High-quality reporting and study methodology are critical to model validity, reproducibility, and clinical implementation. In collaboration with content experts in ML methodology, we performed a systematic review critically appraising the quality of methodology and reporting of recently published ML AP prognostic models.

Methods: Using a validated search strategy, we identified ML AP studies from the databases MEDLINE, PubMed, and EMBASE published between January 2021 and December 2023. Eligibility criteria included all retrospective or prospective studies that developed or validated new or existing ML models in patients with AP that predicted an outcome following an episode of AP. Meta-analysis was considered if there was homogeneity in the study design and in the type of outcome predicted. For risk of bias (ROB) assessment, we used the Prediction Model Risk of Bias Assessment Tool (PROBAST). Quality of reporting was assessed using the Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis or Diagnosis – Artificial Intelligence (TRIPOD+AI) statement that defines standards for 27 items that should be reported in publications using ML prognostic models.

Results: The search strategy identified 6480 publications of which 30 met the eligibility criteria. Studies originated from China (22), U.S (4), and other (4). All 30 studies developed a new ML model and none sought to validate an existing ML model, producing a total of 39 new ML models. AP severity (23/39) or mortality (6/39) were the most common outcomes predicted. The mean area-under-the-curve for all models and endpoints was 0.91 (SD 0.08). The ROB was high for at least one domain in all 39 models, particularly for the analysis domain (37/39 models). Steps were not taken to minimize over-optimistic model performance in 27/39 models. Due to heterogeneity in the study design and in how the outcomes were defined and determined, meta-analysis was not performed. Studies reported on only 15/27 items from TRIPOD+AI standards, with only 7/30 justifying sample size and 13/30 assessing data quality. Other reporting deficiencies included omissions regarding human-AI interaction (28/30), handling low-quality or incomplete data in practice (27/30), sharing analytical codes (25/30), study protocols (25/30) and reporting source data (19/30).

Discussion: There are significant deficiencies in the methodology and reporting of recently published ML based prognostic models in AP patients. These undermine the validity, reproducibility and implementation of these prognostic models despite their promise of superior predictive accuracy.

Funding: none

Registration: Research Registry (reviewregistry1727)
INTRODUCTION

Defined as acute inflammation of the pancreas, acute pancreatitis (AP) remains a common and costly cause of gastrointestinal-related hospitalization, with 1 million new cases each year globally and increasing incidence[1, 2]. The etiology of the disease varies across patient demographics, with gallstones and alcohol comprising the majority of adult cases and diverse environmental factors such as hypertriglyceridemia, drugs, infections, or trauma[3]. The severity of AP can be further categorized as mild, moderately severe, or severe, with severe AP being defined by the presence of persistent organ failure [4]. The combination of persistent organ failure and infected pancreatic necrosis defines a ‘critical’ category of AP severity with the highest morbidity and mortality risk[5, 6]. Survivors of AP can suffer from long-term sequelae including diabetes mellitus, recurrent or chronic pancreatitis, and pancreatic exocrine insufficiency[3, 7-10]. Given the significant short- and long-term morbidity and mortality associated with AP, the National Institute of Health has called for an accurate prognostic model in AP for use in research and the clinical setting[11-13]. Benefits of an accurate prognostic model are many, including enablement of cost-efficient clinical trials through cohort enrichment [14, 15], identification of subphenotypes within a cohort that require different treatment strategies [16, 17], and prompt triaging of patients in the emergency room [18].

Current prognostic models for AP were developed using regression-based techniques (e.g., Glasgow Criteria, Bedside Index for Severity in Acute Pancreatitis (BISAP) etc.) which demonstrate suboptimal performance and limited clinical usefulness[19]. For example, in a prospective external evaluation of regression-based models predicting mortality, none of the models tested produced a post-test probability higher than 14% when “positive”[20]. There has been a call for new approaches to improve prediction accuracy [19, 21]. Advances in the subset
of artificial intelligence (AI) known as machine learning (ML) have facilitated the development of non-regression prediction models, which offer advantages over regression-based models by performing better in diseases with non-linear predictor-outcome relationships such as AP[22]. There has been an increasing number of published ML-based prognostic models that appear to outperform regression-based models [23-25]. However, ML experts have cited concerns regarding methodologic quality, model building practices, and lack of transparent reporting [26-28]. We therefore undertook a systematic review and critical appraisal of recent published studies proposing new non-regression ML based prognostic models to detail any methodological shortcomings and/or gaps in reporting. This was a collaborative effort between experienced clinicians and ML experts [19].

**METHODS**

Detailed methodology of this review has been published elsewhere[29] (doi: 10.1186/s41512-024-00169-1). We conducted a systematic review of all studies published between January 2021 and December 2023 in which a non-regression, ML-based prognostic model in AP was developed and/or validated (either internally or externally), with or without model updating. This review included studies of prospective or retrospective design including post-hoc analysis of clinical trials that: a) enrolled only adult patients (i.e., 18 years old or older), b) contained a prognostic model of AP developed with non-regression ML technique(s), c) predicted any outcome(s) of AP, and d) published in English. Studies involving participants with chronic pancreatitis, pancreatic cancer, or post-surgical pancreatitis were excluded, as were studies with animals, regression-based models, or models that predict the development of AP instead of disease outcomes. Studies published in abstract form only and review articles were also excluded.
We searched the databases MEDLINE (OvidSP) and EMBASE (OvidSP) from January 1, 2021 to December 31, 2023 (Date of search for all data sources, January 31st) Our search was limited to the most recent three years for the following reasons 1) Significant advancements in AP management paradigm has led to a significant change in the natural history/prognosis of the disease over the last decade [30-37]. It was important to identify models trained/evaluated on datasets generated from the most recent cohort of AP. 2) New algorithms rapidly emerge, replacing older algorithms and temporal quality degradation is an established phenomenon in AI models[38]. Validated search strategies [39, 40] were used and are listed in Supplementary Tables 1 and 2, respectively. Covidence software (city, country) was used to screened title-abstract and full text in sequential steps. Each stage required concordance between two independent reviewers (LN, IL, KT, JP, AH, BC, NM, or AL). Disagreements were resolved by a third independent reviewer (PJL or LAC). Included studies were then appraised in terms of risk of bias in study design, completeness of reporting, and for summarization of model predictive performances. Necessary data for PROBAST and TRIPOD+AI evaluation were extracted in accordance with the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist[41].

**Methodologic Quality Assessment:** The Prediction Model Risk of Bias Assessment Tool (PROBAST) was used to assess both risk of bias in study design of prospective models across four main domains: participants, predictors, outcomes, and analysis[42]. The assessment of Applicability section of PROBAST was planned if meta-data were appropriate and feasible for meta-analysis. To optimize the validity of the PROBAST assessment, all evaluators underwent PROBAST rater training, which entailed weekly meetings with an AP content expert trained by PROBAST developers (PJL) to review all 20 signaling questions. Data scientists (JNA, LL, JQ,
or DR) and ML content experts (LAC) were engaged to accurately complete CHARMS and PROBAST. Each model was assessed via the PROBAST framework by two independent reviewers (LN, IL, KT, JP, AH, BC, NM, AL, JNA, LL, JQ, or DR), and disagreements were resolved by an independent third reviewer (PJL or LAC). The pair of reviewers comprised a clinician and a data scientist. The risk of bias in each domain and overall risk of bias were reported for all studies.

**Reporting Quality Assessment:** To assess the quality of the reporting, we decided to use TRIPD+AI statement, which contains a comprehensive list of items that need to be reported for papers reporting development and/or validation of prognostic AI model[43]. List of sections and items on this list covers every key part of a manuscript including title, abstract, introduction, methods, results, and discussion. Additionally, it contains items related to open science and patient & public involvement. Summary statistics of quality of reporting according to the standards of TRIPD+AI[43] were calculated for each study. This review has been registered at Research Registry (reviewregistry1727).

All data reporting in this systematic review adhered to the guidelines of Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) and the checklist can be found in a separate supplementary file.

**RESULTS**

Metadata used to generate these results can be accessed at DOI:10.6084/m9.figshare.26078743.

Our search strategy identified 6480 studies published between January 2021 and December 2023, of which 30 met eligibility criteria (S1 Figure). Studies originated from China (22), the United States (4), Hungary (2), Turkey (1), and New Zealand (1) (Table 1).
<table>
<thead>
<tr>
<th>Author</th>
<th>Publication year</th>
<th>Study site</th>
<th>Type of study</th>
<th>Number of centers</th>
<th>Number of participants</th>
<th>Racial category</th>
<th>Machine learning algorithms</th>
<th>AUC</th>
<th>Type of predictors included*</th>
<th>Outcome predicted†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen[44]</td>
<td>2023</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>978</td>
<td>NR</td>
<td>Neural Network (incl. deep learning)</td>
<td>0.82, 0.92</td>
<td>2, 3, 4</td>
<td>1, 2</td>
</tr>
<tr>
<td>Ding[45]</td>
<td>2021</td>
<td>United States of America</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>337</td>
<td>Reported</td>
<td>Neural Network (incl. deep learning)</td>
<td>0.77</td>
<td>1, 2, 3, 4</td>
<td>3</td>
</tr>
<tr>
<td>Hameed[46]</td>
<td>2022</td>
<td>United States of America</td>
<td>Administrative database</td>
<td>2</td>
<td>6326</td>
<td>NR</td>
<td>Tree-based models</td>
<td>0.94</td>
<td>1, 4</td>
<td>3</td>
</tr>
<tr>
<td>Hong[47]</td>
<td>2022</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>648</td>
<td>NR</td>
<td>Tree-based models</td>
<td>0.96</td>
<td>1, 3, 4</td>
<td>1</td>
</tr>
<tr>
<td>Ince[48]</td>
<td>2022</td>
<td>Turkey</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>1334</td>
<td>NR</td>
<td>Other (Gradient Boost)</td>
<td>0.91-0.98</td>
<td>1, 3, 4</td>
<td>1, 3, 4</td>
</tr>
<tr>
<td>Jin[49]</td>
<td>2021</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>369</td>
<td>NR</td>
<td>Neural Network (incl. deep learning)</td>
<td>0.98</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Kimita[50]</td>
<td>2022</td>
<td>New Zealand</td>
<td>Prospective cohort</td>
<td>1</td>
<td>160</td>
<td>Reported</td>
<td>Tree-based models</td>
<td>0.67</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Kiss[51]</td>
<td>2022</td>
<td>Hungary</td>
<td>Prospective cohort</td>
<td>30</td>
<td>2387</td>
<td>NR</td>
<td>Tree-based models</td>
<td>0.76</td>
<td>1, 4</td>
<td>7</td>
</tr>
<tr>
<td>Ku[52]</td>
<td>2022</td>
<td>Hungary</td>
<td>Prospective cohort</td>
<td>28</td>
<td>1184</td>
<td>NR</td>
<td>K-nearest neighbor</td>
<td>0.81</td>
<td>1, 2, 4</td>
<td>1</td>
</tr>
<tr>
<td>Langmead[24]</td>
<td>2021</td>
<td>United States of America</td>
<td>Secondary analysis of prospective cohort study designed for another reason</td>
<td>1</td>
<td>133</td>
<td>Reported</td>
<td>Tree-based models</td>
<td>0.91</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Li[53]</td>
<td>2022</td>
<td>China</td>
<td>Prospective cohort</td>
<td>7</td>
<td>915</td>
<td>NR</td>
<td>Tree-based models Support vector machine Random Forest LightGBM Ensemble</td>
<td>0.79-0.90</td>
<td>1, 2, 4</td>
<td>2, 3, 4, 8, 9</td>
</tr>
<tr>
<td>Liang[54]</td>
<td>2023</td>
<td>China</td>
<td>Administrative database</td>
<td>1</td>
<td>1798</td>
<td>NR</td>
<td>Neural Network (incl. deep learning)</td>
<td>0.98</td>
<td>3</td>
<td>2, 5</td>
</tr>
<tr>
<td>Luo, Z[55]</td>
<td>2023</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>2</td>
<td>673</td>
<td>NR</td>
<td>Naive Bayes</td>
<td>0.96</td>
<td>1, 3, 4</td>
<td>1</td>
</tr>
<tr>
<td>Luo, J[56]</td>
<td>2023</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>13645</td>
<td>NR</td>
<td>Neural Network (incl. deep learning)</td>
<td>0.91</td>
<td>2, 4</td>
<td>10</td>
</tr>
<tr>
<td>Ren[57]</td>
<td>2023</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>531</td>
<td>NR</td>
<td>Tree-based models</td>
<td>0.81</td>
<td>1, 3, 4</td>
<td>11</td>
</tr>
<tr>
<td>Author</td>
<td>Year</td>
<td>Country</td>
<td>Study Type</td>
<td>N</td>
<td>NR</td>
<td>Type of predictors included</td>
<td>AUC</td>
<td>Models</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>---------</td>
<td>------------</td>
<td>------</td>
<td>----</td>
<td>----------------------------</td>
<td>-----</td>
<td>--------</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>Shi[58]</td>
<td>2022</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>3</td>
<td>NR</td>
<td>Tree-based models</td>
<td>0.90, 0.98</td>
<td>1, 4</td>
<td>3, 5</td>
<td></td>
</tr>
<tr>
<td>Thapa[59]</td>
<td>2022</td>
<td>United States of America</td>
<td>Administrative database</td>
<td>700</td>
<td>371885</td>
<td>Reported Tree-based models</td>
<td>0.92</td>
<td>1, 2, 4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Xu[60]</td>
<td>2021</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>3</td>
<td>447</td>
<td>Other (Adaptive Boost)</td>
<td>0.83</td>
<td>4, 5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Yan[61]</td>
<td>2022</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>151</td>
<td>Tree-based models</td>
<td>NR</td>
<td>2, 4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Yang, D[62]</td>
<td>2022</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>996</td>
<td>Tree-based models Neural Network (incl. deep learning) XGBoost</td>
<td>0.73-0.91</td>
<td>1, 2, 4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Yang, Y[63]</td>
<td>2022</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>2</td>
<td>424</td>
<td>Tree-based models</td>
<td>0.91</td>
<td>1, 3, 4, 5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Yang, D[64]</td>
<td>2023</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>292</td>
<td>Tree-based models*</td>
<td>0.995</td>
<td>4, 5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Yin[65]</td>
<td>2022</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>3</td>
<td>1012</td>
<td>Tree-based models Gradient Boosting Machines Neural Networks XGBoost</td>
<td>0.87-0.95</td>
<td>1, 3, 4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Yuan[66]</td>
<td>2022</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>2</td>
<td>5280</td>
<td>Tree-based models</td>
<td>0.87</td>
<td>1, 2, 3, 4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Zhang, W[67]</td>
<td>2023</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>440</td>
<td>Tree-based models</td>
<td>0.93</td>
<td>1, 3, 4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Zhang, J[68]</td>
<td>2023</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>4</td>
<td>820</td>
<td>CatBoost Random Forest Neural Network</td>
<td>0.52-0.75</td>
<td>1, 4</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Zhang, M[69]</td>
<td>2023</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>460</td>
<td>Bayesian Support Vector Machine Ensembles of Decision Tree</td>
<td>0.81-0.89</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Zhao[70]</td>
<td>2023</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>215</td>
<td>Tree-based models</td>
<td>0.89</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Zhou[71]</td>
<td>2022</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>1</td>
<td>441</td>
<td>XGBoost</td>
<td>0.91</td>
<td>1, 3, 4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Zhu[72]</td>
<td>2021</td>
<td>China</td>
<td>Retrospective cohort</td>
<td>6</td>
<td>711</td>
<td>Tree-based models Neural network</td>
<td>0.99</td>
<td>1, 3, 4</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

*Type of predictors included: 1 = Clinical history (incl. demographics, social, medical history), 2 = Physical Exam Findings, 3 = Radiologic features, 4 = Laboratory values, 5 = Cytokines/new biomarker
†Outcome(s) predicted: 1 = severe pancreatitis, 2 = mild acute pancreatitis, 3 = mortality (all-cause, acute pancreatitis specific, does not specify), 4 = intensive care unit admission, 5 = moderately severe and severe pancreatitis, 6 = other, 7 = pancreatic necrosis, 8 = length of stay, 9 = pancreatic necrosis – infected, 10 = multisystem organ dysfunction/failure, 11 = recurrent pancreatitis, 12 = new onset diabetes, 13 = Intra-abdominal infection;
All 30 studies reported the development of a new ML-based prognostic model, but only one study included external validation step of the newly developed model. Nearly three-fourths (22/30) of included studies were retrospective cohort, while only five studies were prospective, of which one was a secondary analysis. Five studies developed more than one model, resulting in a total of 39 models developed in 30 studies. The most common machine learning algorithms were tree-based models (20/39) and neural networks (7/39). AP severity (21/39) or mortality (6/39) were the most common outcomes predicted. The most common methods of internal validation were cross-validation (23/39) and bootstrapping (17/39). For 31/39 models, shrinkage methods were not used to evaluate for or adjust for optimism (shrinkage methods: techniques used to account for magnitude of noise in the dataset contributing to overinflation of predictive performance). A summary of pertinent descriptive statistics collected as per the CHARMS checklist is provided in Table 1. Overall, for the 39 models the mean area-under-the-curve (AUC) was 0.91 (SD 0.08). Six studies developed more than one ML-model using the same dataset, presenting the parameters of the “best performing” model (Table 1). Every model had at least one domain in which the risk of bias was classified as high (Fig 1), meaning that all 39 models were assessed to be at high risk of bias by PROBAST standards (see S3 Table for individual model’s ROB rating). The median number of TRIPOD+AI items that were reported on in the 30 studies was 15/27 (range 6-20). No study reported on all the items. A comprehensive breakdown of the number of TRIPOD+AI items reported on in each study is given in Supplementary Table 4 and on the heatmap for visual presentation of the data (Fig 2).

Fig 1. Summary of Risk of Bias in 4 Domains Assessed by PROBAST

Fig 2. Heatmap depicting common areas of deficiencies in reporting standards as assessed by TRIPOD+AI *Publication has same first author and year as another paper listed; PMID of each * in ascending order: Yang et al, 2022: 35430680, 35607360. Luo et al, 2023: 36653317, 36773821. Zhang et al, 2023: 36902504, 36964219, 37196588.
Risk of Bias in Four Domains of Methodology as Assessed by PROBAST

PROBAST ratings of the 39 models based on individual studies are summarized in Supplementary Table 3. Assessment of Applicability was not applicable to the objectives of this review. As the primary objective was to assess the methodologic quality and because of marked heterogeneity of the cohorts and the different definitions and determination of outcomes, a synthesis of the meta-data was not undertaken.

Participants Domain: In this domain there was a high risk of bias with 35/39 models. The data source was not appropriate with 31/39 models. The inclusions and exclusions of participants was not appropriate in 26/39 models.

Predictors Domain: In this domain there was a high risk of bias with 18/39 models. The predictors were not defined and measured in a similar way for all participants in 12/39 models. Assessor blinding to the outcome data was not done with 30/39 models. In 8/39 studies predictors were included when the result would not be available at the time of applying the prognostic model.

Outcomes Domain: In this domain there was a high risk of bias with 24/39 models. While outcomes were defined in a standard way in 33/39 models, they were not determined appropriately in 20/39 models. The way that outcomes were determined was not reported for 1/39 models. Outcomes were not defined and determined in a similar way in 13/39 models. Blinding was not performed in 24/39 models. Outcomes were included as predictors in 17/39 models.

Analysis Domain: In this domain there was a high risk of bias with 37/39 models (Fig 1). The common deficiencies in this domain were no accounting for overfitting and optimism (i.e. no shrinkage methods employed) in 31/39 models, none or inappropriate reporting of data complexity in 38/39 models (Fig 2), insufficient sample size in 28/39 models, and selection of predictors relied solely on univariate analysis in 26/39 models.
Quality of Reporting as Assessed by TRIPOD+AI

Title, Abstract, Introduction Section: All 30 studies reported to the standards of TRIPOD+AI except in one important sub-item. No study reported the health inequalities that may exist in outcomes between sociodemographic groups (Fig 2 and S4 Table).

Methods Section: Twenty-eight studies described the sources of data, study dates, setting and eligibility were described in 28/30 studies but only 5/30 studies reported details of any treatment received where treatment might have influences the outcome of interest. Other frequent omissions included no description of model fairness and their rationale (28/30), no sample size justification (23/30), no blinding of assessors (20/30), no reporting differences between training and evaluation data (16/30), no outcome measurement (15/30), no description of data preparation and pre-processing (13/30), no reporting of elements pertinent to outcome definition (13/30), and no assessment of study quality (13/30).

Open Science and Patient/Public Involvement Section: There was no reporting on whether a protocol was prepared, available or accessed in 25/30 studies. There was no report as to the availability of study data (9/30) or analytical code (28/30). There was comment on whether patients and public were involved in 26/30 studies.

Results Section: There was insufficient detail of the prediction model to allow external validation in 25/30 studies. Reporting details of the prediction model performance in key subgroups (e.g. sociodemographic) was not available in 15/30 studies.

Discussion Section: Items pertaining to the usability of the model in the context of current care were usually not discussed. Only 3/30 studies described how poor quality or missing data should be handled with clinical implementation of the model. Only 1/30 study specified whether users
will be required to interact with handling of the input data or use of the model and what level of
expertise is required to use the model.

**Rationale against performing subgroup analyses:** Even though several of the included studies
developed models predicting similar outcomes, decision was made not to perform subgroup
analyses stratified by similar endpoints. All but one model was judged to have high risk of bias in
at least two out of the four PROBAST domains and none of the models were at low risk of bias in
the statistical analyses domain. With such limitations in the methodology across the board,
subgroup analyses were felt not to lead to meaningful discoveries or different conclusions.

**DISCUSSION**

In this systematic review, we assessed the quality of the methodology and reporting of studies
that develop and/or validated non-regression ML-based models in AP literature. While the
performance of the published models was high (mean AUC 0.91), we identified several key
limitations in the recently published models. Unfortunately, these shortcomings are like those
identified in other fields such as oncology[28] and anesthesiology[73]. First, the concern relates
to the high risk of bias most notably in the statistical analysis section, which can undermine the
validity of the models. Second, due to the lack of external validation studies, generalizability of
the ML models may be limited. Third relates to open science practice, where in over 90% of the
studies, the code was not shared and no information was provided on how the model was built.
Additionally, there was a lack of reporting on how the ML model can be implemented in clinical
practice. Lastly, none of the studies described potential health inequities among different
sociodemographic groups, which risks widening disparities in healthcare, if implemented in real
clinical practice.
The quality of the statistical analyses is one of the most important facets of model development. The PROBAST ROB tool dedicates 9 signaling questions to this domain[42]. Two particularly deficient areas were sample size justification and guarding against overfitting. A robust sample size (especially for a ML model) and guarding against overfitting are critically important. When these steps are omitted, a model may perform well in the development dataset but the predictive performance may not be reproducible[74]. We found that most published studies developed a model with a sample size of less than 1,000 participants and median events per variable was 9.5. Even for regression-based models, the minimum recommended events per variable is 20[42].

While events per variable is not a singular reflection of sufficient sample size, it is generally accepted that ML models require much larger sample size (than regression-based models) due to the risk of model instability[75]. Potentially limited generalizability of the published models need to be highlighted. Only one study conducted external validation but with limitations, all but 5 studies were single-center design. While AP is a common gastrointestinal disease, with an annual worldwide 1 million new cases a year[76], international or large multi-center consortiums with efforts to build a generalizable model have been lacking. Lack of such collaboration results in siloed attempts at building models that may not be clinically utilized due to poor reproducibility and generalizability. As with the case with the regression-based models[21], we are seeing a similar trend in ML-based models in AP.

Ultimately, prognostic models are built to aid clinical decision making or enhance cohort enrichment in a research study. Therefore, steps need to be taken to thoughtfully consider real-life issues we will face when trying to deploy these models (e.g., ways to deal with missing values in real clinical practice when patients won't have the data elements necessary for the ML
model). We also found key missing items relevant to open science, that limit external validation studies by other investigators and clinical implementation by the hospitals. For example, only 5 studies shared the code to permit third-party evaluation and implementation, only 3 studies gave guidance on how to handle missing data, and one study detailed the specifics of what constitutes human-AI interaction. As important, aspects of model building relevant healthcare equity (e.g., comparison of performance estimates among different sociodemographic subgroups) were not evaluated. Such deficiency leads to a potential to produce a model that widens the socioeconomic disparities[77].

Our study has several strengths. For transparency and rigor of our methodology we have published our methods and adhered strictly to the standards of TRIPOD-SR/MA. Our work was conducted in collaboration between data scientists, ML methodologist, and content experts in AP, which we believe enhances the reliability of our findings. There are multiple aspects to PROBAST and TRIPOD+AI assessment that require both AP content and ML methodology expertise. Third, rigorous internal training for PROBAST assessment preceded the project, enhancing the validity of our ROS assessment.

Several limitations deserve mention. Our search strategy extended only the last 3 years so it is possible that our findings may not be fully representative of all the ML models published for AP thus far. Second, while PROBAST was developed by expert methodologists, it is possible that models deemed high ROB by PROBAST may still be valid, reproducible, and generalizable in AP. However, there is emerging data from other diseases that suggest models deemed high ROB by PROBAST perform poorly external validation studies[78, 79]

In conclusion, the potential benefit of ML-based prognostic models is evident with an overall high AUC (mean 0.91±0.8SD). However, this study indicates that there should be great caution
in implementing the reported models because of the very major concerns with the quality of the methodology and reporting. These raise questions about the validity, reproducibility, and generalizability of the prognostic models. It is recommended that AP-specific, standardized methodology that covers all 4 PROBAST domains and all items within TRIPOD+AI be used in developing and validating ML-based prognostic models. Only then should implementation be considered. Our study findings provide valuable baseline assessment of the quality of methods and reporting of ML-based models in AP. It is also timely given the recent publication of TRIPOD+AI[43], which we hope will encourage future investigators to utilize.

ACKNOWLEDGEMENTS: none


21


SUPPORTING INFORMATION CAPTIONS

S1 Figure: PRISMA Flow diagram
S1 Table: Search Strategy in Medline
S2 Table: Search Strategy in EMBASE
S3 Table: PROBAST rating of models from individual studies
S4 Table: Responses on TRIPOD+AI and overall fidelity to transparent reporting standards for machine learning studies (N=30)
PROBAST Assessment of Published AI Prognostic Models in Acute Pancreatitis

- Risk of Bias: High risk of bias
- Analysis: High risk of bias
- Outcomes: High risk of bias
- Predictors: High risk of bias
- Participants: High risk of bias