

1 A Systematic Review of Machine Learning-based Prognostic Models for Acute 2 Pancreatitis: Towards Improving Methods and Reporting Quality

3
4 **Brian Critelli[¶], Amier Hassan[¶]**, Ila Lahooti², Lydia Noh³, Jun Sung Park², Kathleen Tong², Ali
5 Lahooti¹, Nate Matzko¹, Jan Niklas Adams⁴, Lukas Liss⁴, Justin Quion⁵, David Restrepo⁵, Melica
6 Nikahd⁶, Stacey Culp⁶, Adam Lacy-Hulbert⁷, Cate Speake⁸, James Buxbaum⁹, Jason Bischof¹⁰, Cemal
7 Yazici¹¹, Anna Evans Phillips¹², Sophie Terp¹³, Alexandra Weissman¹⁴, Darwin Conwell¹⁵, Phil Hart²,
8 Mitch Ramsey², Somashekar Krishna², Samuel Han², Erica Park², Raj Shah², Venkata Akshintala¹⁶, John
9 A Windsor¹⁷, Nikhil K Mull¹⁸, Georgios I Papachristou², Leo Anthony Celi^{5, 19}, Peter J Lee^{2*}

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

11 Author affiliations:

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

33
34 [¶]: Amier Hassan and Brian Critelli are co-first authors (i.e., these authors contributed equally to this work)

35 *Corresponding author: Peter J Lee

36 Email: Peter.Lee@osumc.edu

37

38 Conflicts of interest: all authors declare no conflicts of interest

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

41
42 **Background:** An accurate prognostic tool is essential to aid clinical decision making (e.g., patient triage)
43 and to advance personalized medicine. However, such prognostic tool is lacking for acute pancreatitis (AP).
44 Increasingly machine learning (ML) techniques are being used to develop high-performing prognostic
45 models in AP. However, **methodologic and reporting quality has received little attention**. High-quality
46 reporting and study methodology are critical to model validity, reproducibility, and clinical implementation.
47 In collaboration with content experts in ML methodology, we performed a systematic review critically
48 appraising the quality of methodology and reporting of recently published ML AP prognostic models.

49
50 **Methods:** Using a validated search strategy, we identified ML AP studies from the databases MEDLINE,
51 PubMed, and EMBASE published between January 2021 and December 2023. Eligibility criteria included
52 all retrospective or prospective studies that developed or validated new or existing ML models in patients
53 with AP that predicted an outcome following an episode of AP. Meta-analysis was considered if there was
54 homogeneity in the study design and in the type of outcome predicted. For risk of bias (ROB) assessment,
55 we used the Prediction Model Risk of Bias Assessment Tool (PROBAST). Quality of reporting was
56 assessed using the Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis or
57 Diagnosis – Artificial Intelligence (TRIPOD+AI) statement that defines standards for 27 items that should
58 be reported in publications using ML prognostic models.

59
60 **Results:** The search strategy identified 6480 publications of which 30 met the eligibility criteria. Studies
61 originated from China (22), U.S (4), and other (4). All 30 studies developed a new ML model and none
62 sought to validate an existing ML model, producing a total of 39 new ML models. AP severity (23/39) or
63 mortality (6/39) were the most common outcomes predicted. The mean area-under-the-curve for all models
64 and endpoints was 0.91 (SD 0.08). The ROB was high for at least one domain in all 39 models, particularly
65 for the analysis domain (37/39 models). Steps were not taken to minimize over-optimistic model
66 performance in 27/39 models. Due to heterogeneity in the study design and in how the outcomes were
67 defined and determined, meta-analysis was not performed.
68 Studies reported on only 15/27 items from TRIPOD+AI standards, with only 7/30 justifying sample size
69 and 13/30 assessing data quality. Other reporting deficiencies included omissions regarding human-AI
70 interaction (28/30), handling low-quality or incomplete data in practice (27/30), sharing analytical codes
71 (25/30), study protocols (25/30) and reporting source data (19/30),.

72
73 **Discussion:** There are significant deficiencies in the methodology and reporting of recently published ML
74 based prognostic models in AP patients. These undermine the validity, reproducibility and implementation
75 of these prognostic models despite their promise of superior predictive accuracy.

76
77 **Funding:** none

78 **Registration:** Research Registry ([reviewregistry1727](https://www.crd.york.ac.uk/PROSPERO/record/2024062600001727))

79

80 **INTRODUCTION**

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

97 Current prognostic models for AP were developed using regression-based techniques (e.g.,
98 Glasgow Criteria, Bedside Index for Severity in Acute Pancreatitis (BISAP) etc.) which
99 demonstrate suboptimal performance and limited clinical usefulness[19]. For example, in a
100 prospective external evaluation of regression-based models predicting mortality, none of the
101 models tested produced a post-test probability higher than 14% when “positive”[20]. There has
102 been a call for new approaches to improve prediction accuracy [19, 21]. Advances in the subset

103 of artificial intelligence (AI) known as machine learning (ML) have facilitated the development of
104 non-regression prediction models, which offer advantages over regression-based models by
105 performing better in diseases with non-linear predictor-outcome relationships such as AP[22].
106 There has been an increasing number of published ML-based prognostic models that appear to
107 outperform regression-based models [23-25]. However, ML experts have cited concerns regarding
108 methodologic quality, model building practices, and lack of transparent reporting [26-28]. We
109 therefore undertook a systematic review and critical appraisal of recent published studies
110 proposing new non-regression ML based prognostic models to detail any methodological
111 shortcomings and/or gaps in reporting. This was a collaborative effort between experienced
112 clinicians and ML experts [19].

113

114 **METHODS**

115 Detailed methodology of this review has been published elsewhere[29] (doi: 10.1186/s41512-024-
116 00169-1). We conducted a systematic review of all studies published between January 2021 and
117 December 2023 in which a non-regression, ML-based prognostic model in AP was developed
118 and/or validated (either internally or externally), with or without model updating. This review
119 included studies of prospective or retrospective design including post-hoc analysis of clinical trials
120 that: a) enrolled only adult patients (i.e., 18 years old or older), b) contained a prognostic model of
121 AP developed with non-regression ML technique(s), c) predicted any outcome(s) of AP, and d)
122 published in English. Studies involving participants with chronic pancreatitis, pancreatic cancer,
123 or post-surgical pancreatitis were excluded, as were studies with animals, regression-based
124 models, or models that predict the development of AP instead of disease outcomes. Studies
125 published in abstract form only and review articles were also excluded.

126 We searched the databases MEDLINE (OvidSP) and EMBASE (OvidSP) from January 1, 2021 to
127 December 31, 2023 (Date of search for all data sources, January 31st) Our search was limited to
128 the most recent three years for the following reasons 1) Significant advancements in AP
129 management paradigm has led to a significant change in the natural history/prognosis of the
130 disease over the last decade [30-37]. It was important to identify models trained/evaluated on
131 datasets generated from the most recent cohort of AP. 2) New algorithms rapidly emerge, replacing
132 older algorithms and temporal quality degradation is an established phenomenon in AI models[38].
133 Validated search strategies [39, 40] were used and are listed in Supplementary Tables 1 and 2,
134 respectively. Covidence software (city, country) was used to screened title-abstract and full text in
135 sequential steps. Each stage required concordance between two independent reviewers (LN, IL,
136 KT, JP, AH, BC, NM, or AL). Disagreements were resolved by a third independent reviewer (PJJ
137 or LAC). Included studies were then appraised in terms of risk of bias in study design,
138 completeness of reporting, and for summarization of model predictive performances. Necessary
139 data for PROBAST and TRIPOD+AI evaluation were extracted in accordance with the Critical
140 Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies
141 (CHARMS) checklist[41].

142 **Methodologic Quality Assessment:** The Prediction Model Risk of Bias Assessment Tool
143 (PROBAST) was used to assess both risk of bias in study design of prospective models across four
144 main domains: participants, predictors, outcomes, and analysis[42]. The assessment of
145 Applicability section of PROBAST was planned if meta-data were appropriate and feasible for
146 meta-analysis. To optimize the validity of the PROBAST assessment, all evaluators underwent
147 PROBAST rater training, which entailed weekly meetings with an AP content expert trained by
148 PROBAST developers (PJJ) to review all 20 signaling questions. Data scientists (JNA, LL, JQ,

149 or DR) and ML content experts (LAC) were engaged to accurately complete CHARMS and
150 PROBAST. Each model was assessed via the PROBAST framework by two independent
151 reviewers (LN, IL, KT, JP, AH, BC, NM, AL, JNA, LL, JQ, or DR), and disagreements were
152 resolved by an independent third reviewer (PJL or LAC). The pair of reviewers comprised a
153 clinician and a data scientist. The risk of bias in each domain and overall risk of bias were reported
154 for all studies.

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

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

166 **RESULTS**

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

168 Our search strategy identified 6480 studies published between January 2021 and December 2023,
169 of which 30 met eligibility criteria (S1 Figure). Studies originated from China (22), the United
170 States (4), Hungary (2), Turkey (1), and New Zealand (1) (Table 1).

171

Table 1: Basic characteristics of included studies

<u>Author</u>	<u>Publication year</u>	<u>Study site</u>	<u>Type of study</u>	<u>Number of centers</u>	<u>Number of participants</u>	<u>Racial category</u>	<u>Machine learning algorithms</u>	<u>AUC</u>	<u>Type of predictors included*</u>	<u>Outcome predicted†</u>
Chen[44]	2023	China	Retrospective cohort	1	978	NR	Neural Network (incl. deep learning)	0.82, 0.92	2, 3, 4	1, 2
Ding[45]	2021	United States of America	Retrospective cohort	1	337	Reported	Neural Network (incl. deep learning)	0.77	1, 2, 3, 4	3
Hameed[46]	2022	United States of America	Administrative database	2	6326	NR	Tree-based models	0.94	1, 4	3
Hong[47]	2022	China	Retrospective cohort	1	648	NR	Tree-based models	0.96	1, 3, 4	1
Ince[48]	2022	Turkey	Retrospective cohort	1	1334	NR	Other (Gradient Boost)	0.91-0.98	1, 3, 4	1,3,4
Jin[49]	2021	China	Retrospective cohort	1	369	NR	Neural Network (incl. deep learning)	0.98	4	5
Kimita[50]	2022	New Zealand	Prospective cohort	1	160	Reported	Tree-based models	0.67	5	6
Kiss[51]	2022	Hungary	Prospective cohort	30	2387	NR	Tree-based models	0.76	1, 4	7
Kui[52]	2022	Hungary	Prospective cohort	28	1184	NR	K-nearest neighbor	0.81	1, 2, 4	1
Langmead[24]	2021	United States of America	Secondary analysis of prospective cohort study designed for another reason	1	133	Reported	Tree-based models	0.91	5	1
Li[53]	2022	China	Prospective cohort	7	915	NR	Tree-based models Support vector machine Random Forest LightGBM Ensemble	0.79-0.90	1, 2, 4	2, 3, 4, 8, 9
Liang[54]	2023	China	Administrative database	1	1798	NR	Neural Network (incl. deep learning)	0.98	3	2, 5
Luo, Z[55]	2023	China	Retrospective cohort	2	673	NR	Naive Bayes	0.96	1, 3, 4	1
Luo, J[56]	2023	China	Retrospective cohort	1	13645	NR	Neural Network (incl. deep learning)	0.91	2, 4	10
Ren[57]	2023	China	Retrospective cohort	1	531	NR	Tree-based models	0.81	1, 3, 4	11

Shi[58]	2022	China	Retrospective cohort	3	2846	NR	Tree-based models	0.90, 0.98	1, 4	3, 5
Thapa[59]	2022	United States of America	Administrative database	700	371885	Reported	Tree-based models	0.92	1, 2, 4	1
Xu[60]	2021	China	Retrospective cohort	3	447	NR	Other (Adaptive Boost)	0.83	4, 5	10
Yan[61]	2022	China	Retrospective cohort	1	151	NR	Tree-based models	NR	2, 4	3
Yang, D[62]	2022	China	Retrospective cohort	1	996	NR	Tree-based models Neural Network (incl. deep learning) XGBoost	0.73-0.91	1, 2, 4	5
Yang, Y[63]	2022	China	Retrospective cohort	2	424	NR	Tree-based models	0.91	1, 3, 4, 5	5
Yang, D[64]	2023	China	Retrospective cohort	1	292	NR	Tree-based models*	0.995	4, 5	5
Yin[65]	2022	China	Retrospective cohort	3	1012	NR	Tree-based models Gradient Boosting Machines Neural Networks XGBoost	0.87-0.95	1, 3, 4	1
Yuan[66]	2022	China	Retrospective cohort	2	5280	NR	Tree-based models	0.87	1, 2, 3, 4	4
Zhang, W[67]	2023	China	Retrospective cohort	1	440	NR	Tree-based models	0.93	1, 3, 4	5
Zhang, J[68]	2023	China	Retrospective cohort	4	820	NR	CatBoost Random Forest Neural Network	0.52-0.75	1, 4	12
Zhang, M[69]	2023	China	Retrospective cohort	1	460	NR	Bayesian Support Vector Machine Ensembles of Decision Tree	0.81-0.89	4	5
Zhao[70]	2023	China	Retrospective cohort	1	215	NR	Tree-based models	0.89	3	2
Zhou[71]	2022	China	Retrospective cohort	1	441	NR	XGBoost	0.91	1, 3, 4	1
Zhu[72]	2021	China	Retrospective cohort	6	711	NR	Tree-based models Neural network	0.99	1, 3, 4	13

*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;

172 All 30 studies reported the development of a new ML-based prognostic model, but only one study
173 included external validation step of the newly developed model. Nearly three-fourths (22/30) of
174 included studies were retrospective cohort, while only five studies were prospective, of which one
175 was a secondary analysis. Five studies developed more than one model, resulting in a total of 39
176 models developed in 30 studies. The most common machine learning algorithms were tree-based
177 models (20/39) and neural networks (7/39). AP severity (21/39) or mortality (6/39) were the most
178 common outcomes predicted. The most common methods of internal validation were cross-
179 validation (23/39) and bootstrapping (17/39). For 31/39 models, shrinkage methods were not used
180 to evaluate for or adjust for optimism (shrinkage methods: techniques used to account for
181 magnitude of noise in the dataset contributing to overinflation of predictive performance). A
182 summary of pertinent descriptive statistics collected as per the CHARMS checklist is provided in
183 Table 1. Overall, for the 39 models the mean area-under-the-curve (AUC) was 0.91 (SD 0.08).
184 Six studies developed more than one ML-model using the same dataset, presenting the parameters
185 of the “best performing” model (Table 1). Every model had at least one domain in which the risk
186 of bias was classified as high (Fig 1), meaning that all 39 models were assessed to be at high risk
187 of bias by PROBAST standards (see S3 Table for individual model’s ROB rating). The median
188 number of TRIPOD+AI items that were reported on in the 30 studies was 15/27 (range 6-20). No
189 study reported on all the items. A comprehensive breakdown of the number of TRIPOD+AI items
190 reported on in each study is given in Supplementary Table 4 and on the heatmap for visual
191 presentation of the data (Fig 2).

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

193 **Fig 2. Heatmap depicting common areas of deficiencies in reporting standards as assessed**

194 **by TRIPOD+AI** *Publication has same first author and year as another paper listed; PMID of each * in
195 ascending order: Yang et al, 2022: 35430680, 35607360. Luo et al, 2023: 36653317, 36773821. Zhang et al, 2023:
196 36902504, 36964219, 37196588.

197

198 ***Risk of Bias in Four Domains of Methodology as Assessed by PROBAST***

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

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

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

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

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

221 ***Quality of Reporting as Assessed by TRIPOD+AI***

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

225 *Methods Section:* Twenty-eight studies described The sources of data, study dates, setting and
226 eligibility were described in 28/30 studies but only 5/30 studies reported details of any treatment
227 received where treatment might have influences the outcome of interest. Other frequent omissions
228 included no description of model fairness and their rationale (28/30), no sample size justification
229 (23/30), no blinding of assessors (20/30), no reporting differences between training and evaluation
230 data (16/30), no outcome measurement (15/30), no description of data preparation and pre-
231 processing(13/30), no reporting of elements pertinent to outcome definition (13/30), and no
232 assessment of study quality (13/30)

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

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

240 *Discussion Section:* Items pertaining to the usability of the model in the context of current care
241 were usually not discussed. Only 3/30 studies described how poor quality or missing data should
242 be handled with clinical implementation of the model. Only 1/30 study specified whether users

243 will be required to interact with handling of the input data or use of the model and what level of
244 expertise is required to use the model.

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

251 **DISCUSSION**

252 In this systematic review, we assessed the quality of the methodology and reporting of studies
253 that develop and/or validated non-regression ML-based models in AP literature. While the
254 performance of the published models was high (mean AUC 0.91), we identified several key
255 limitations in the recently published models. Unfortunately, these shortcomings are like those
256 identified in other fields such as oncology[28] and anesthesiology[73]. First, the concern relates
257 to the high risk of bias most notably in the statistical analysis section, which can undermine the
258 validity of the models. Second, due to the lack of external validation studies, generalizability of
259 the ML models may be limited. Third relates to open science practice, where in over 90% of the
260 studies, the code was not shared and no information was provided on how the model was built.
261 Additionally, there was a lack of reporting on how the ML model can be implemented in clinical
262 practice. Lastly, none of the studies described potential health inequities among different
263 sociodemographic groups, which risks widening disparities in healthcare, if implemented in real
264 clinical practice.

265 The quality of the statistical analyses is one of the most important facets of model development.
266 The PROBAST ROB tool dedicates 9 signaling questions to this domain[42]. Two particularly
267 deficient areas were sample size justification and guarding against overfitting. A robust sample
268 size (especially for a ML model) and guarding against overfitting are critically important. When
269 these steps are omitted, a model may perform well in the development dataset but the predictive
270 performance may not be reproducible[74]. We found that most published studies developed a
271 model with a sample size of less than 1,000 participants and median events per variable was 9.5.
272 Even for regression-based models, the minimum recommended events per variable is 20[42].
273 While events per variable is not a singular reflection of sufficient sample size, it is generally
274 accepted that ML models require much larger sample size (than regression-based models) due to
275 the risk of model instability[75].

276 Potentially limited generalizability of the published models need to be highlighted. Only one
277 study conducted external validation but with limitations, all but 5 studies were single-center
278 design. While AP is a common gastrointestinal disease, with an annual worldwide 1 million new
279 cases a year[76], international or large multi-center consortiums with efforts to build a
280 generalizable model have been lacking. Lack of such collaboration results in siloed attempts at
281 building models that may not be clinically utilized due to poor reproducibility and
282 generalizability. As with the case with the regression-based models[21], we are seeing a similar
283 trend in ML-based models in AP.

284 Ultimately, prognostic models are built to aid clinical decision making or enhance cohort
285 enrichment in a research study. Therefore, steps need to be taken to thoughtfully consider real-
286 life issues we will face when trying to deploy these models (e.g., ways to deal with missing
287 values in real clinical practice when patients won't have the data elements necessary for the ML

288 model). We also found key missing items relevant to open science, that limit external validation
289 studies by other investigators and clinical implementation by the hospitals. For example, only 5
290 studies shared the code to permit third-party evaluation and implementation, only 3 studies gave
291 guidance on how to handle missing data, and one study detailed the specifics of what constitutes
292 human-AI interaction. As important, aspects of model building relevant healthcare equity (e.g.,
293 comparison of performance estimates among different sociodemographic subgroups) were not
294 evaluated. Such deficiency leads to a potential to produce a model that widens the
295 socioeconomic disparities[77].

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

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

309 In conclusion, the potential benefit of ML-based prognostic models is evident with an overall
310 high AUC (mean $0.91 \pm 0.8SD$). However, this study indicates that there should be great caution

311 in implementing the reported models because of the very major concerns with the quality of the
312 methodology and reporting. These raise questions about the validity, reproducibility, and
313 generalizability of the prognostic models. It is recommended that AP-specific, standardized
314 methodology that covers all 4 PROBAST domains and all items within TRIPOD+AI be used in
315 developing and validating ML-based prognostic models. Only then should implementation be
316 considered. Our study findings provide valuable baseline assessment of the quality of methods
317 and reporting of ML-based models in AP. It is also timely given the recent publication of
318 TRIPOD+AI[43], which we hope will encourage future investigators to utilize.

319

320 **ACKNOWLEDGEMENTS:** none

321

322

323

324

325

326

327

328

329

330

331

332

333

334 **REFERENCES**

- 335 1. Xiao AY, Tan ML, Wu LM, Asrani VM, Windsor JA, Yadav D, Petrov MS. Global incidence
336 and mortality of pancreatic diseases: a systematic review, meta-analysis, and meta-regression of
337 population-based cohort studies. *Lancet Gastroenterol Hepatol.* 2016;1(1):45-55. Epub
338 20160628. doi: 10.1016/s2468-1253(16)30004-8. PubMed PMID: 28404111.
- 339 2. Iannuzzi JP, King JA, Leong JH, Quan J, Windsor JW, Tanyingoh D, et al. Global
340 Incidence of Acute Pancreatitis Is Increasing Over Time: A Systematic Review and Meta-Analysis.
341 *Gastroenterology.* 2022;162(1):122-34. Epub 20210925. doi: 10.1053/j.gastro.2021.09.043.
342 PubMed PMID: 34571026.
- 343 3. Lee PJ, Papachristou GI. New insights into acute pancreatitis. *Nature reviews*
344 *Gastroenterology & hepatology.* 2019. doi: 10.1038/s41575-019-0158-2.
- 345 4. Banks PA, Bollen TL, Dervenis C, Gooszen HG, Johnson CD, Sarr MG, et al.
346 Classification of acute pancreatitis--2012: revision of the Atlanta classification and definitions by
347 international consensus. *Gut.* 2013;62(1):102-11. doi: 10.1136/gutjnl-2012-302779 [doi].
- 348 5. Dellinger EP, Forsmark CE, Layer P, Levy P, Maravi-Poma E, Petrov MS, et al.
349 Determinant-based classification of acute pancreatitis severity: an international multidisciplinary
350 consultation. *Annals of surgery.* 2012;256(6):875-80. doi: 10.1097/SLA.0b013e318256f778.
- 351 6. Wu D, Lu B, Xue H-D, Yang H, Qian J-M, Lee P, Windsor JA. Validation of Modified
352 Determinant-Based Classification of severity for acute pancreatitis in a tertiary teaching hospital.
353 *Pancreatology : official journal of the International Association of Pancreatology (IAP)* [et al].
354 2019;19(2):217-23. doi: 10.1016/j.pan.2019.01.003.
- 355 7. Petrov MS, Yadav D. Global epidemiology and holistic prevention of pancreatitis. *Nat Rev*
356 *Gastroenterol Hepatol.* 2019;16(3):175-84. doi: 10.1038/s41575-018-0087-5. PubMed PMID:
357 30482911; PubMed Central PMCID: PMC6597260.
- 358 8. Das SL, Singh PP, Phillips AR, Murphy R, Windsor JA, Petrov MS. Newly diagnosed
359 diabetes mellitus after acute pancreatitis: a systematic review and meta-analysis. *Gut.*
360 2014;63(5):818-31. doi: 10.1136/gutjnl-2013-305062 [doi].
- 361 9. Huang W, de la Iglesia-García D, Baston-Rey I, Calviño-Suarez C, Lariño-Noia J, Iglesias-
362 Garcia J, et al. Exocrine Pancreatic Insufficiency Following Acute Pancreatitis: Systematic Review
363 and Meta-Analysis. *Digestive diseases and sciences.* 2019;64(7):1985-2005. doi:
364 10.1007/s10620-019-05568-9.
- 365 10. Zhi M, Zhu X, Lugea A, Waldron RT, Pandol SJ, Li L. Incidence of New Onset Diabetes
366 Mellitus Secondary to Acute Pancreatitis: A Systematic Review and Meta-Analysis. *Front Physiol.*
367 2019;10:637. Epub 20190531. doi: 10.3389/fphys.2019.00637. PubMed PMID: 31231233;
368 PubMed Central PMCID: PMC6558372.
- 369 11. Abu-El-Hajja M, Gukovskaya AS, Andersen DK, Gardner TB, Hegyi P, Pandol SJ, et al.
370 Accelerating the Drug Delivery Pipeline for Acute and Chronic Pancreatitis: Summary of the
371 Working Group on Drug Development and Trials in Acute Pancreatitis at the National Institute of
372 Diabetes and Digestive and Kidney Diseases Workshop. *Pancreas*2018. p. 1185-92.
- 373 12. Uc A, Andersen DK, Borowitz D, Glesby MJ, Mayerle J, Sutton R, Pandol SJ. Accelerating
374 the Drug Delivery Pipeline for Acute and Chronic Pancreatitis-Knowledge Gaps and Research
375 Opportunities: Overview Summary of a National Institute of Diabetes and Digestive and Kidney
376 Diseases Workshop. *Pancreas*2018. p. 1180-4.
- 377 13. Serrano J, Laughlin MR, Bellin MD, Yadav D, Chinchilli VM, Andersen DK. Type 1
378 Diabetes in Acute Pancreatitis Consortium: From Concept to Reality. *Pancreas.* 2022;51(6):563-
379 7. doi: 10.1097/mpa.0000000000002073. PubMed PMID: 36206459; PubMed Central PMCID:
380 PMC69555854.
- 381 14. van Brunschot S, van Grinsven J, Voermans RP, Bakker OJ, Besselink MG, Boermeester
382 MA, et al. Transluminal endoscopic step-up approach versus minimally invasive surgical step-up
383 approach in patients with infected necrotising pancreatitis (TENSION trial): design and rationale

- 384 of a randomised controlled multicenter trial [ISRCTN09186711. BMC gastroenterology.
385 2013;13:161-. doi: 10.1186/1471-230X-13-161 [doi].
- 386 15. van Santvoort HC, Besselink MG, Bakker OJ, Hofker HS, Boermeester MA, Dejong CH,
387 et al. A step-up approach or open necrosectomy for necrotizing pancreatitis. The New England
388 journal of medicine. 2010;362(16):1491-502. doi: 10.1056/NEJMoa0908821 [doi].
- 389 16. Giamarellos-Bourboulis EJ, Aschenbrenner AC, Bauer M, Bock C, Calandra T, Gat-Viks
390 I, et al. The pathophysiology of sepsis and precision-medicine-based immunotherapy. Nature
391 Immunology. 2024;25(1):19-28. doi: 10.1038/s41590-023-01660-5.
- 392 17. Rosenson RS, Gaudet D, Ballantyne CM, Baum SJ, Bergeron J, Kershaw EE, et al.
393 Evinacumab in severe hypertriglyceridemia with or without lipoprotein lipase pathway mutations:
394 a phase 2 randomized trial. Nat Med. 2023;29(3):729-37. Epub 20230306. doi: 10.1038/s41591-
395 023-02222-w. PubMed PMID: 36879129; PubMed Central PMCID: PMCPCMC10033404.
- 396 18. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Jr., Hasegawa K. Emergency
397 department triage prediction of clinical outcomes using machine learning models. Critical care
398 (London, England). 2019;23(1):64-. doi: 10.1186/s13054-019-2351-7.
- 399 19. Capurso G, Ponz de Leon Pisani R, Lauri G, Archibugi L, Hegyi P, Papachristou GI, et al.
400 Clinical usefulness of scoring systems to predict severe acute pancreatitis: A systematic review
401 and meta-analysis with pre and post-test probability assessment. United European Gastroenterol
402 J. 2023;11(9):825-36. Epub 20230927. doi: 10.1002/ueg2.12464. PubMed PMID: 37755341;
403 PubMed Central PMCID: PMCPCMC10637128.
- 404 20. Papachristou GI, Muddana V, Yadav D, O'Connell M, Sanders MK, Slivka A, Whitcomb
405 DC. Comparison of BISAP, Ranson's, APACHE-II, and CTSI scores in predicting organ failure,
406 complications, and mortality in acute pancreatitis. The American journal of gastroenterology.
407 2010;105(2):435-41; quiz 42. doi: 10.1038/ajg.2009.622.
- 408 21. Mounzer R, Langmead CJ, Wu BU, Evans AC, Bishehsari F, Muddana V, et al.
409 Comparison of existing clinical scoring systems to predict persistent organ failure in patients with
410 acute pancreatitis. Gastroenterology. 2012;142(7):1476-. doi: 10.1053/j.gastro.2012.03.005.
- 411 22. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical
412 introduction. BMC Med Res Methodol. 2019;19(1):64. Epub 20190319. doi: 10.1186/s12874-019-
413 0681-4. PubMed PMID: 30890124; PubMed Central PMCID: PMCPCMC6425557.
- 414 23. Zhou Y, Ge YT, Shi XL, Wu KY, Chen WW, Ding YB, et al. Machine learning predictive
415 models for acute pancreatitis: A systematic review. Int J Med Inform. 2022;157:104641. Epub
416 20211110. doi: 10.1016/j.ijmedinf.2021.104641. PubMed PMID: 34785488.
- 417 24. Langmead C, Lee PJ, Paragomi P, Greer P, Stello K, Hart PA, et al. A Novel 5-Cytokine
418 Panel Outperforms Conventional Predictive Markers of Persistent Organ Failure in Acute
419 Pancreatitis. Clinical and translational gastroenterology. 2021;12(5):e00351-e. doi:
420 10.14309/ctg.0000000000000351.
- 421 25. Fei Y, Gao K, Li W-Q. Artificial neural network algorithm model as powerful tool to predict
422 acute lung injury following to severe acute pancreatitis. Pancreatology : official journal of the
423 International Association of Pancreatology (IAP) [et al]. 2018;18(8):892-9. doi:
424 10.1016/j.pan.2018.09.007.
- 425 26. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al.
426 Systematic review finds "spin" practices and poor reporting standards in studies on machine
427 learning-based prediction models. J Clin Epidemiol. 2023;158:99-110. Epub 20230405. doi:
428 10.1016/j.jclinepi.2023.03.024. PubMed PMID: 37024020.
- 429 27. Andaur Navarro CL, Damen JAA, van Smeden M, Takada T, Nijman SWJ, Dhiman P, et
430 al. Systematic review identifies the design and methodological conduct of studies on machine
431 learning-based prediction models. J Clin Epidemiol. 2023;154:8-22. Epub 20221125. doi:
432 10.1016/j.jclinepi.2022.11.015. PubMed PMID: 36436815.
- 433 28. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al.
434 Overinterpretation of findings in machine learning prediction model studies in oncology: a

- 435 systematic review. *J Clin Epidemiol.* 2023;157:120-33. Epub 20230317. doi:
436 10.1016/j.jclinepi.2023.03.012. PubMed PMID: 36935090.
- 437 29. Hassan A, Critelli B, Lahooti I, Lahooti A, Matzko N, Adams JN, et al. Critical appraisal of
438 machine learning prognostic models for acute pancreatitis: protocol for a systematic review. *Diagn*
439 *Progn Res.* 2024;8(1):6. Epub 20240402. doi: 10.1186/s41512-024-00169-1. PubMed PMID:
440 38561864; PubMed Central PMCID: PMCPCMC10986113.
- 441 30. van Dijk SM, Hallensleben ND, van Santvoort HC, Fockens P, van Goor H, Bruno MJ,
442 Besselink MG. Acute pancreatitis: recent advances through randomised trials. *Gut.*
443 2017;66(11):2024-32. doi: 10.1136/gutjnl-2016-313595.
- 444 31. de-Madaria E, Buxbaum JL, Maisonneuve P, García García de Paredes A, Zapater P,
445 Guilabert L, et al. Aggressive or Moderate Fluid Resuscitation in Acute Pancreatitis. *The New*
446 *England journal of medicine.* 2022;387(11):989-1000. doi: 10.1056/NEJMoa2202884.
- 447 32. Wolbrink DRJ, van de Poll MCG, Termorshuizen F, de Keizer NF, van der Horst ICC,
448 Schnabel R, et al. Trends in Early and Late Mortality in Patients With Severe Acute Pancreatitis
449 Admitted to ICUs: A Nationwide Cohort Study. *Crit Care Med.* 2022;50(10):1513-21. Epub
450 20220725. doi: 10.1097/ccm.0000000000005629. PubMed PMID: 35876365.
- 451 33. Sorrento C, Shah I, Yakah W, Ahmed A, Tintara S, Kandasamy C, et al. Inpatient Alcohol
452 Cessation Counseling Is Associated With a Lower 30-Day Hospital Readmission in Acute
453 Alcoholic Pancreatitis. *J Clin Gastroenterol.* 2022;56(9):e313-e7. Epub 20220110. doi:
454 10.1097/mcg.0000000000001666. PubMed PMID: 34999646.
- 455 34. Onnekink AM, Boxhoorn L, Timmerhuis HC, Bac ST, Besselink MG, Boermeester MA, et
456 al. Endoscopic Versus Surgical Step-Up Approach for Infected Necrotizing Pancreatitis
457 (EXTENSION): Long-term Follow-up of a Randomized Trial. *Gastroenterology.* 2022;163(3):712-
458 22.e14. Epub 20220514. doi: 10.1053/j.gastro.2022.05.015. PubMed PMID: 35580661.
- 459 35. Hallensleben ND, Timmerhuis HC, Hollemans RA, Pocornie S, van Grinsven J, van
460 Brunschot S, et al. Optimal timing of cholecystectomy after necrotising biliary pancreatitis. *Gut.*
461 2022;71(5):974-82. Epub 20210716. doi: 10.1136/gutjnl-2021-324239. PubMed PMID:
462 34272261.
- 463 36. Sissingh NJ, Groen JV, Koole D, Klok FA, Boekestijn B, Bollen TL, et al. Therapeutic
464 anticoagulation for splanchnic vein thrombosis in acute pancreatitis: A systematic review and
465 meta-analysis. *Pancreatology.* 2022;22(2):235-43. Epub 20211222. doi:
466 10.1016/j.pan.2021.12.008. PubMed PMID: 35012902.
- 467 37. Schepers NJ, Bakker OJ, Besselink MG, Ahmed Ali U, Bollen TL, Gooszen HG, et al.
468 Impact of characteristics of organ failure and infected necrosis on mortality in necrotising
469 pancreatitis. *Gut.* 2018.
- 470 38. Vela D, Sharp A, Zhang R, Nguyen T, Hoang A, Pinykh OS. Temporal quality degradation
471 in AI models. *Scientific Reports.* 2022;12(1):11654. doi: 10.1038/s41598-022-15245-z.
- 472 39. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters
473 for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews.
474 *PLoS One.* 2012;7(2):e32844. Epub 20120229. doi: 10.1371/journal.pone.0032844. PubMed
475 PMID: 22393453; PubMed Central PMCID: PMCPCMC3290602.
- 476 40. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform*
477 *Assoc.* 2001;8(4):391-7. doi: 10.1136/jamia.2001.0080391. PubMed PMID: 11418546; PubMed
478 Central PMCID: PMCPCMC130084.
- 479 41. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al.
480 Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies:
481 The CHARMS Checklist. *PLOS Medicine.* 2014;11(10):e1001744. doi:
482 10.1371/journal.pmed.1001744.
- 483 42. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST:
484 A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of internal*
485 *medicine.* 2019;170(1):51-8. doi: 10.7326/M18-1376.

- 486 43. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI
487 statement: updated guidance for reporting clinical prediction models that use regression or
488 machine learning methods. *BMJ*. 2024;385:e078378. doi: 10.1136/bmj-2023-078378.
- 489 44. Chen Z, Wang Y, Zhang H, Yin H, Hu C, Huang Z, et al. Deep Learning Models for Severity
490 Prediction of Acute Pancreatitis in the Early Phase From Abdominal Nonenhanced Computed
491 Tomography Images. *Pancreas*. 2023;52(1):e45-e53. doi: 10.1097/mpa.0000000000002216.
492 PubMed PMID: 37378899.
- 493 45. Ding N, Guo C, Li C, Zhou Y, Chai X. An Artificial Neural Networks Model for Early
494 Predicting In-Hospital Mortality in Acute Pancreatitis in MIMIC-III. *Biomed Res Int*.
495 2021;2021:6638919. Epub 20210128. doi: 10.1155/2021/6638919. PubMed PMID: 33575333;
496 PubMed Central PMCID: PMCPMC7864739.
- 497 46. Hameed MAB, Alamgir Z. Improving mortality prediction in Acute Pancreatitis by machine
498 learning and data augmentation. *Comput Biol Med*. 2022;150:106077. Epub 20220911. doi:
499 10.1016/j.compbimed.2022.106077. PubMed PMID: 36137318.
- 500 47. Hong W, Lu Y, Zhou X, Jin S, Pan J, Lin Q, et al. Usefulness of Random Forest Algorithm
501 in Predicting Severe Acute Pancreatitis. *Front Cell Infect Microbiol*. 2022;12:893294. Epub
502 20220610. doi: 10.3389/fcimb.2022.893294. PubMed PMID: 35755843; PubMed Central PMCID:
503 PMCPMC9226542.
- 504 48. İnce AT, Silahtaroglu G, Seven G, Koçhan K, Yıldız K, Şentürk H. Early prediction of the
505 severe course, survival, and ICU requirements in acute pancreatitis by artificial intelligence.
506 *Pancreatol*. 2023;23(2):176-86. Epub 20221230. doi: 10.1016/j.pan.2022.12.005. PubMed
507 PMID: 36610872.
- 508 49. Jin X, Ding Z, Li T, Xiong J, Tian G, Liu J. Comparison of MPL-ANN and PLS-DA models
509 for predicting the severity of patients with acute pancreatitis: An exploratory study. *Am J Emerg*
510 *Med*. 2021;44:85-91. Epub 20210122. doi: 10.1016/j.ajem.2021.01.044. PubMed PMID:
511 33582613.
- 512 50. Kimita W, Bharmal SH, Ko J, Petrov MS. Identifying endotypes of individuals after an
513 attack of pancreatitis based on unsupervised machine learning of multiplex cytokine profiles.
514 *Transl Res*. 2023;251:54-62. Epub 20220718. doi: 10.1016/j.trsl.2022.07.001. PubMed PMID:
515 35863673.
- 516 51. Kiss S, Pintér J, Molontay R, Nagy M, Farkas N, Sipos Z, et al. Early prediction of acute
517 necrotizing pancreatitis by artificial intelligence: a prospective cohort-analysis of 2387 cases. *Sci*
518 *Rep*. 2022;12(1):7827. Epub 20220512. doi: 10.1038/s41598-022-11517-w. PubMed PMID:
519 35552440; PubMed Central PMCID: PMCPMC9098474.
- 520 52. Kui B, Pintér J, Molontay R, Nagy M, Farkas N, Gede N, et al. EASY-APP: An artificial
521 intelligence model and application for early and easy prediction of severity in acute pancreatitis.
522 *Clin Transl Med*. 2022;12(6):e842. doi: 10.1002/ctm2.842. PubMed PMID: 35653504; PubMed
523 Central PMCID: PMCPMC9162438.
- 524 53. Li JN, Mu D, Zheng SC, Tian W, Wu ZY, Meng J, et al. Machine learning improves
525 prediction of severity and outcomes of acute pancreatitis: a prospective multi-center cohort study.
526 *Sci China Life Sci*. 2023;66(8):1934-7. Epub 20230516. doi: 10.1007/s11427-022-2333-8.
527 PubMed PMID: 37209250.
- 528 54. Liang H, Wang M, Wen Y, Du F, Jiang L, Geng X, et al. Predicting acute pancreatitis
529 severity with enhanced computed tomography scans using convolutional neural networks. *Sci*
530 *Rep*. 2023;13(1):17514. Epub 20231016. doi: 10.1038/s41598-023-44828-7. PubMed PMID:
531 37845380; PubMed Central PMCID: PMCPMC10579320.
- 532 55. Luo Z, Shi J, Fang Y, Pei S, Lu Y, Zhang R, et al. Development and evaluation of machine
533 learning models and nomogram for the prediction of severe acute pancreatitis. *J Gastroenterol*
534 *Hepatol*. 2023;38(3):468-75. Epub 20230127. doi: 10.1111/jgh.16125. PubMed PMID: 36653317.

- 535 56. Luo J, Lan L, Huang S, Zeng X, Xiang Q, Li M, et al. Real-time prediction of organ failures
536 in patients with acute pancreatitis using longitudinal irregular data. *J Biomed Inform.*
537 2023;139:104310. Epub 20230210. doi: 10.1016/j.jbi.2023.104310. PubMed PMID: 36773821.
- 538 57. Ren W, Zou K, Chen Y, Huang S, Luo B, Jiang J, et al. Application of a Machine Learning
539 Predictive Model for Recurrent Acute Pancreatitis. *J Clin Gastroenterol.* 2023. Epub 20231103.
540 doi: 10.1097/mcg.0000000000001936. PubMed PMID: 37983784.
- 541 58. Shi N, Lan L, Luo J, Zhu P, Ward TRW, Szatmary P, et al. Predicting the Need for
542 Therapeutic Intervention and Mortality in Acute Pancreatitis: A Two-Center International Study
543 Using Machine Learning. *J Pers Med.* 2022;12(4). Epub 20220411. doi: 10.3390/jpm12040616.
544 PubMed PMID: 35455733; PubMed Central PMCID: PMCPCMC9031087.
- 545 59. Thapa R, Iqbal Z, Garikipati A, Siefkas A, Hoffman J, Mao Q, Das R. Early prediction of
546 severe acute pancreatitis using machine learning. *Pancreatol.* 2022;22(1):43-50. Epub
547 20211016. doi: 10.1016/j.pan.2021.10.003. PubMed PMID: 34690046.
- 548 60. Xu F, Chen X, Li C, Liu J, Qiu Q, He M, et al. Prediction of Multiple Organ Failure
549 Complicated by Moderately Severe or Severe Acute Pancreatitis Based on Machine Learning: A
550 Multicenter Cohort Study. *Mediators Inflamm.* 2021;2021:5525118. Epub 20210503. doi:
551 10.1155/2021/5525118. PubMed PMID: 34054342; PubMed Central PMCID: PMCPCMC8112913.
- 552 61. Yan J, Yilin H, Di W, Jie W, Hanyue W, Ya L, Jie P. A nomogram for predicting the risk of
553 mortality in patients with acute pancreatitis and Gram-negative bacilli infection. *Front Cell Infect*
554 *Microbiol.* 2022;12:1032375. Epub 20221110. doi: 10.3389/fcimb.2022.1032375. PubMed PMID:
555 36439207; PubMed Central PMCID: PMCPCMC9685314.
- 556 62. Yang D, Zhao L, Kang J, Wen C, Li Y, Ren Y, et al. Development and validation of a
557 predictive model for acute kidney injury in patients with moderately severe and severe acute
558 pancreatitis. *Clin Exp Nephrol.* 2022;26(8):770-87. Epub 20220416. doi: 10.1007/s10157-022-
559 02219-8. PubMed PMID: 35430680.
- 560 63. Yang Y, Xiao W, Liu X, Zhang Y, Jin X, Li X. Machine Learning-Assisted Ensemble
561 Analysis for the Prediction of Acute Pancreatitis with Acute Kidney Injury. *Int J Gen Med.*
562 2022;15:5061-72. Epub 20220517. doi: 10.2147/ijgm.S361330. PubMed PMID: 35607360;
563 PubMed Central PMCID: PMCPCMC9123915.
- 564 64. Yang D, Kang J, Li Y, Wen C, Yang S, Ren Y, et al. Development of a predictive nomogram
565 for acute respiratory distress syndrome in patients with acute pancreatitis complicated with acute
566 kidney injury. *Ren Fail.* 2023;45(2):2251591. Epub 20230919. doi:
567 10.1080/0886022x.2023.2251591. PubMed PMID: 37724533; PubMed Central PMCID:
568 PMCPCMC10512859.
- 569 65. Yin M, Zhang R, Zhou Z, Liu L, Gao J, Xu W, et al. Automated Machine Learning for the
570 Early Prediction of the Severity of Acute Pancreatitis in Hospitals. *Front Cell Infect Microbiol.*
571 2022;12:886935. Epub 20220610. doi: 10.3389/fcimb.2022.886935. PubMed PMID: 35755847;
572 PubMed Central PMCID: PMCPCMC9226483.
- 573 66. Yuan L, Ji M, Wang S, Wen X, Huang P, Shen L, Xu J. Machine learning model identifies
574 aggressive acute pancreatitis within 48 h of admission: a large retrospective study. *BMC Med*
575 *Inform Decis Mak.* 2022;22(1):312. Epub 20221129. doi: 10.1186/s12911-022-02066-3. PubMed
576 PMID: 36447180; PubMed Central PMCID: PMCPCMC9707001.
- 577 67. Zhang W, Chang Y, Ding Y, Zhu Y, Zhao Y, Shi R. To Establish an Early Prediction Model
578 for Acute Respiratory Distress Syndrome in Severe Acute Pancreatitis Using Machine Learning
579 Algorithm. *J Clin Med.* 2023;12(5). Epub 20230221. doi: 10.3390/jcm12051718. PubMed PMID:
580 36902504; PubMed Central PMCID: PMCPCMC10002486.
- 581 68. Zhang J, Lv Y, Hou J, Zhang C, Yua X, Wang Y, et al. Machine learning for post-acute
582 pancreatitis diabetes mellitus prediction and personalized treatment recommendations. *Sci Rep.*
583 2023;13(1):4857. Epub 20230324. doi: 10.1038/s41598-023-31947-4. PubMed PMID: 36964219;
584 PubMed Central PMCID: PMCPCMC10038980.

- 585 69. Zhang M, Pang M. Early prediction of acute respiratory distress syndrome complicated by
586 acute pancreatitis based on four machine learning models. *Clinics (Sao Paulo)*. 2023;78:100215.
587 Epub 20230503. doi: 10.1016/j.clinsp.2023.100215. PubMed PMID: 37196588; PubMed Central
588 PMCID: PMC10199163.
- 589 70. Zhao Y, Wei J, Xiao B, Wang L, Jiang X, Zhu Y, He W. Early prediction of acute
590 pancreatitis severity based on changes in pancreatic and peripancreatic computed tomography
591 radiomics nomogram. *Quant Imaging Med Surg*. 2023;13(3):1927-36. Epub 20230201. doi:
592 10.21037/qims-22-821. PubMed PMID: 36915340; PubMed Central PMCID:
593 PMC10006146.
- 594 71. Zhou Y, Han F, Shi XL, Zhang JX, Li GY, Yuan CC, et al. Prediction of the severity of
595 acute pancreatitis using machine learning models. *Postgrad Med*. 2022;134(7):703-10. Epub
596 20220712. doi: 10.1080/00325481.2022.2099193. PubMed PMID: 35801388.
- 597 72. Zhu C, Zhang S, Zhong H, Gu Z, Kang Y, Pan C, et al. Intra-abdominal infection in acute
598 pancreatitis in eastern China: microbiological features and a prediction model. *Ann Transl Med*.
599 2021;9(6):477. doi: 10.21037/atm-21-399. PubMed PMID: 33850874; PubMed Central PMCID:
600 PMC10006146.
- 601 73. Arina P, Kaczorek MR, Hofmaenner DA, Pisciotta W, Refinetti P, Singer M, et al.
602 Prediction of Complications and Prognostication in Perioperative Medicine: A Systematic Review
603 and PROBAST Assessment of Machine Learning Tools. *Anesthesiology*. 2024;140(1):85-101.
604 doi: 10.1097/aln.0000000000004764. PubMed PMID: 37944114.
- 605 74. Kakarmath S, Esteva A, Arnaout R, Harvey H, Kumar S, Muse E, et al. Best practices for
606 authors of healthcare-related artificial intelligence manuscripts. *NPJ Digit Med*. 2020;3:134. Epub
607 20201016. doi: 10.1038/s41746-020-00336-w. PubMed PMID: 33083569; PubMed Central
608 PMCID: PMC10006146.
- 609 75. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data
610 hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research*
611 *Methodology*. 2014;14(1):137. doi: 10.1186/1471-2288-14-137.
- 612 76. Li CL, Jiang M, Pan CQ, Li J, Xu LG. The global, regional, and national burden of acute
613 pancreatitis in 204 countries and territories, 1990-2019. *BMC Gastroenterol*. 2021;21(1):332.
614 Epub 20210825. doi: 10.1186/s12876-021-01906-2. PubMed PMID: 34433418; PubMed Central
615 PMCID: PMC10006146.
- 616 77. Celi LA, Cellini J, Charpignon ML, Dee EC, Dernoncourt F, Eber R, et al. Sources of bias
617 in artificial intelligence that perpetuate healthcare disparities-A global review. *PLOS Digit Health*.
618 2022;1(3):e0000022. Epub 20220331. doi: 10.1371/journal.pdig.0000022. PubMed PMID:
619 36812532; PubMed Central PMCID: PMC10006146.
- 620 78. Venema E, Wessler BS, Paulus JK, Salah R, Raman G, Leung LY, et al. Large-scale
621 validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form:
622 high risk of bias models show poorer discrimination. *J Clin Epidemiol*. 2021;138:32-9. Epub
623 20210624. doi: 10.1016/j.jclinepi.2021.06.017. PubMed PMID: 34175377.
- 624 79. Helmrich I, Mikolić A, Kent DM, Lingsma HF, Wynants L, Steyerberg EW, van Klaveren
625 D. Does poor methodological quality of prediction modeling studies translate to poor model
626 performance? An illustration in traumatic brain injury. *Diagn Progn Res*. 2022;6(1):8. Epub
627 20220505. doi: 10.1186/s41512-022-00122-0. PubMed PMID: 35509061; PubMed Central
628 PMCID: PMC10006146.
- 629
630
631
632
633

634 **SUPPORTING INFORMATION CAPTIONS**

635 **S1 Figure: PRISMA Flow diagram**

636 **S1 Table: Search Strategy in Medline**

637 **S2 Table: Search Strategy in EMBASE**

638 **S3 Table: PROBAST rating of models from individual studies**

639 **S4 Table: Responses on TRIPOD+AI and overall fidelity to transparent reporting**
640 **standards for machine learning studies (N=30)**

641

PROBAST Assessment of Published AI Prognostic Models in Acute Pancreatitis



