

1 Original Article

2

3 **Completeness of reporting of clinical prediction models developed**
4 **using supervised machine learning: A systematic review**

5

6 Constanza L Andaur Navarro^{1,2} *doctoral student* (c.l.andaurnavarro@umcutrecht.nl, 0000-0002-7745-
7 2887), Johanna A A Damen^{1,2} *assistant professor* (j.a.a.damen@umcutrecht.nl, 0000-0001-7401-4593),
8 Toshihiko Takada¹ *assistant professor* (t.takada-3@umcutrecht.nl, 0000-0002-8032-6224), Steven W J
9 Nijman¹ *doctoral student* (S.W.J.Nijman@umcutrecht.nl, 0000-0001-6798-2078), Paula Dhiman^{3,4}
10 *research fellow* (paula.dhiman@ndorms.ox.ac.uk, 0000-0002-0989-0623), Jie Ma³ *medical statistician*
11 (jie.ma@csm.ox.ac.uk, 0000-0002-3900-1903), Gary S Collins^{3,4} *professor* (gary.collins@csm.ox.ac.uk,
12 0000-0002-2772-2316), Ram Bajpai⁵ *research fellow* (r.bajpai@keele.ac.uk, 0000-0002-1227-2703),
13 Richard D Riley⁵ *professor* (r.riley@keele.ac.uk, 0000-0001-8699-0735), Karel GM Moons^{1,2} *professor*
14 (k.g.m.moons@umcutrecht.nl, 0000-0003-2118-004X), Lotty Hooft^{1,2} *professor* (l.hooft@umcutrecht.nl,
15 0000-0002-7950-2980)

16

17

18

19

20

21 ¹Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht,
22 The Netherlands.

23 ²Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.

24 ³Center for Statistics in Medicine, NDORMS, University of Oxford, Oxford, United Kingdom.

25 ⁴NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United
26 Kingdom

27 ⁵Centre for Prognosis Research, School of Medicine, Keele University, Keele, United Kingdom.

28

29

30 **Word Count manuscript : 3624; Word Count abstract : 321**

31 **Keywords:** machine learning, prediction model, diagnosis, prognosis, development, validation,
32 reporting adherence, reporting guideline, TRIPOD.

33 **Correspondance to:** Constanza L Andaur Navarro, c.l.andaurnavarro@umcutrecht.nl

34 Julius Center for Health Sciences and Primary Care, Universiteitsweg 100, P.O. Box 85500, 3508 GA,
35 Utrecht, The Netherlands.

36 **ABSTRACT**

37 **Objective.** While many studies have consistently found incomplete reporting of regression-
38 based prediction model studies, evidence is lacking for machine learning-based prediction
39 model studies. We aim to systematically review the adherence of Machine Learning (ML)-
40 based prediction model studies to the Transparent Reporting of a multivariable prediction
41 model for Individual Prognosis Or Diagnosis (TRIPOD) Statement.

42 **Study design and setting:** We included articles reporting on development or external
43 validation of a multivariable prediction model (either diagnostic or prognostic) developed
44 using supervised ML for individualized predictions across all medical fields (PROSPERO,
45 CRD42019161764). We searched PubMed from 1 January 2018 to 31 December 2019. Data
46 extraction was performed using the 22-item checklist for reporting of prediction model
47 studies (www.TRIPOD-statement.org). We measured the overall adherence per article and per
48 TRIPOD item.

49 **Results:** Our search identified 24 814 articles, of which 152 articles were included: 94 (61.8%)
50 prognostic and 58 (38.2%) diagnostic prediction model studies. Overall, articles adhered to a
51 median of 38.7% (IQR 31.0-46.4) of TRIPOD items. No articles fully adhered to complete
52 reporting of the abstract and very few reported the flow of participants (3.9%, 95% CI 1.8 to
53 8.3), appropriate title (4.6%, 95% CI 2.2 to 9.2), blinding of predictors (4.6%, 95% CI 2.2 to 9.2),
54 model specification (5.2%, 95% CI 2.4 to 10.8), and model's predictive performance (5.9%,
55 95% CI 3.1 to 10.9). There was often complete reporting of source of data (98.0%, 95% CI 94.4
56 to 99.3) and interpretation of the results (94.7%, 95% CI 90.0 to 97.3).

57 **Conclusion.** Similar to prediction model studies developed using conventional regression-
58 based techniques, the completeness of reporting is poor. Essential information to decide to
59 use the model (i.e. model specification and its performance) is rarely reported. However,
60 some items and sub-items of TRIPOD might be less suitable for ML-based prediction model
61 studies and thus, TRIPOD requires extensions. Overall, there is an urgent need to improve the
62 reporting quality and usability of research to avoid research waste.

63 **What is new?**

- 64 • **Key findings:** Similar to prediction model studies developed using regression techniques,
65 machine learning (ML)-based prediction model studies adhered poorly to the TRIPOD
66 statement, the current standard reporting guideline.
- 67 • **What this adds to what is known?** In addition to efforts to improve the completeness of
68 reporting in ML-based prediction model studies, an extension of TRIPOD for these type of
69 studies is needed.
- 70 • **What is the implication, what should change now?** While TRIPOD-AI is under
71 development, we urge authors to follow the recommendations of the TRIPOD statement
72 to improve the completeness of reporting and reduce potential research waste of ML-
73 based prediction model studies.

74 INTRODUCTION

75 Clinical prediction models are used extensively in healthcare to aid patient diagnosis and
76 prognosis of disease and health status. A diagnostic model combines multiple predictors or
77 test results to predict the presence or absence of a certain disorder, whereas a prognostic
78 model estimates the probability of future occurrence of an outcome.¹⁻³ Studies developing,
79 validating, and updating prediction models are abundant in most clinical fields and their
80 number will continue to increase as prediction models developed using artificial intelligence
81 (AI) and machine learning (ML) are receiving substantial interest in the healthcare
82 community.⁴

83 ML, a subset of AI, offers a class of models that can iteratively learn from data, identify
84 complex data patterns, automate model building, and predict outcomes based on what has
85 been learned using computer-based algorithms.^{5,6} ML is often described as more efficient and
86 accurate than conventional regression-based techniques. ML-based prediction models,
87 correctly developed, validated, and implemented, can improve patient benefit, and reduce
88 disease and health system burden. There is increasing concern of the methodological and
89 reporting quality of studies developing prediction models, with research till date focusing on
90 models developed with conventional statistical techniques such as logistic and Cox
91 regression.⁷⁻¹¹ Recent studies have found limited application of ML-based prediction models
92 because of poor study design and reporting.^{12,13}

93 Incomplete (or unclear) reporting makes ML-based prediction models difficult to interpret
94 and impedes validation by independent researchers, thus creating barriers to their use in
95 daily clinical practice. Complete and accurate reporting of ML-based prediction model studies
96 will improve its interpretability, reproducibility, risk of bias assessment, and applicability in
97 daily medical practice and is, therefore, essential for high-quality research.¹⁴ To improve
98 transparency and reporting of prediction model studies, the Transparent Reporting of a
99 multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement, a
100 checklist of 22 items, was designed (www.tripod-statement.org).^{15,16} Specific guidance for ML-
101 based prediction model studies is currently lacking and has initiated the extension of TRIPOD
102 for prediction models developed using ML or AI (TRIPOD-AI).¹⁷

103 We conducted a systematic review to assess the completeness of reporting of ML-based
104 diagnostic and prognostic prediction model studies in recent literature using the TRIPOD

105 Statement.^{15,16} Our results will highlight specific reporting areas that can inform reporting
106 guidelines for ML, such as TRIPOD-AI.¹⁷

107 **METHODS**

108 Our systematic review protocol was registered (PROSPERO, CRD42019161764) and
109 published.¹⁸ We reported this systematic review following the PRISMA statement.¹⁹

110

111 **Data source and search**

112 We searched PubMed on December 19, 2019 to identify primary articles describing
113 prediction models (diagnostic or prognostic) using any supervised ML technique across all
114 clinical domains published between 1 January 2018 and 31 December 2019. The search
115 strategy is provided in the supplemental material.

116

117 **Study selection**

118 We included articles that described the development or validation of one or more
119 multivariable prediction models using any supervised ML technique aiming for individualized
120 prediction of risk or outcomes. As there is still no consensus on a definition of ML, we defined
121 a 'study using ML' as a study that describes the use of a non-generalized linear models to
122 develop or validate a prediction model (e.g. tree-based models, ensembles, deep learning).
123 Hence, studies that claimed to have used ML, but they reported only regression-based
124 statistical techniques were excluded from this systematic review (e.g. logistic regression, lasso
125 regression, ridge regression and elastic net). Specifically, we focused on supervised ML, a
126 subdomain of ML, that is characterized by using an algorithm that learns to predict from
127 labelled outcome examples. Example are random forest, support vector machine, neural
128 network, naïve bayes, and gradient boosting.

129 Articles reporting on the incremental value or model extension were also included. We
130 included all articles regardless of study design, data source, or patient-related health
131 outcome. Articles that investigated a single predictor, test or biomarker, or its causality with
132 an outcome were excluded. Articles using ML to enhance reading of images or signals, or
133 articles where ML models only used genetic traits or molecular markers as predictors, were
134 also excluded. We also excluded systematic reviews, conference abstracts, tutorials, and
135 articles for which full-text was unavailable via our institution. We restricted the search to
136 human subjects and English-language articles. Further details are stated in our protocol.¹⁸

137

138 Two researchers, from a group of seven (CLAN, TT, SWJN, PD, JM, RB, JAAD), independently
139 screened titles and abstracts to identify potentially eligible studies. Full-text articles were then
140 retrieved, and two independent researchers reviewed them for eligibility using Rayyan.²⁰ One
141 researcher (CLAN) screened all articles and six researchers (TT, SWJN, PD, JM, RB, JAAD)
142 collectively screened the same articles. Disagreements between reviewers were resolved by a
143 third researcher (JAAD).

144 **Data extraction**

145 The data extraction form was based on the TRIPOD adherence assessment form ([www.tripod-](http://www.tripod-statement.org)
146 [statement.org](http://www.tripod-statement.org)).²¹ This form contains several adherence statements (hereafter called sub-
147 items) per TRIPOD item. Some items and sub-items are applicable to all types of studies,
148 while others are only applicable to model development only or external validation only ([Table](#)
149 [1](#)). To judge reporting of the requested information, sub-items were formulated to be
150 answered with 'yes', 'no', 'not applicable'. We amended the published adherence form by
151 omitting the 'referenced' option because we checked the information in the references,
152 supplemental material or appendix. Sub-items related to items 10b and 16 were extracted
153 per model, rather than at study-level, as they refer to model performance.

154

155 We performed a double data extraction for included articles. Two reviewers independently
156 extracted data from each article using the standardized form which was available in REDCap,
157 a data capture tool.²² To accomplish consistent data extraction, the form was piloted by all
158 reviewers on five articles. One researcher (CLAN) extracted data from all articles and six
159 researchers (TT, SWJN, PD, JM, RB, JAAD) collectively extracted data from the same articles.
160 Discrepancies in data extraction were discussed and resolved between each pair of reviewers.

161

162 **Data synthesis and analysis**

163 We categorized prediction model studies as prognosis or diagnosis and classified studies by
164 research aim: development (with or without internal validation), development with external
165 validation (same model), development with external validation (different model), and external
166 validation only. Detailed definition of research aims can be found in the supplemental
167 material. Where articles described the development and/or validation of more than one

168 prediction model, we chose the first ML model reported in the methods section for
169 extraction.

170 We scored each TRIPOD item as 'reported' and 'not reported' based on answers to
171 corresponding sub-items. If the answer to all sub-items of a TRIPOD item is scored 'yes' or
172 'not applicable', the corresponding item was considered 'reported'. Two analyses were
173 conducted: adherence per item and overall adherence per article. We calculated the
174 adherence per TRIPOD item by dividing the number of studies that adhered to a specific item
175 by the number of studies in which the item was applicable. The total number of TRIPOD items
176 varies by the type of prediction model study ([Table 1](#)). We calculated the overall adherence to
177 TRIPOD per article by dividing the sum of reported TRIPOD items by the total number of
178 applicable TRIPOD items for each study. If an item was 'not applicable' for a particular study,
179 it was excluded when calculating the overall adherence, both in the numerator and
180 denominator.²¹ Analyses were performed using R version 3.6.2 (R Core Team, Vienna, Austria).
181 Results were summarized as percentages, medians, ranges, and using visual plots.

182 **RESULTS**

183 We identified 24 814 unique articles, of which we sampled ten random sets of 249 articles
184 each with sampling replacement for screening. We screened the title and abstracts of 2 482
185 articles, screened full-text of 312 articles and included 152 eligible articles ([Figure 1](#)).

186

187 We included 94 (61.8%) prognostic and 58 (38.2%) diagnostic prediction model studies. 132
188 (86.8%) articles described development with internal validation and 19 (12.5%) development
189 with external validation (same model). One (0.6%) article was development with external
190 validation (different model) and was included as a development with internal validation study
191 in the present analysis. Prediction models were developed most often in oncology (21/152
192 [13.8%]). Detailed description of the included studies is provided in supplemental material.

193

194 Across the 152 studies, 1429 models were developed and 219 were validated, with a range of
195 1 to 156 for both types of studies. The most commonly used ML techniques for the first
196 reported model were Classification and Regression Tree (CART [10.1%]), Support Vector
197 Machine (SVM [9.4%]) and Random Forest (RF [9.4%]). Alongside ML techniques, 19.5% of
198 studies reported the development of a model using conventional statistical techniques, such
199 as logistic regression. Five out of 152 studies (3.3%, 95% CI 1.4% to 7.5%) stated following the
200 recommendations of the TRIPOD Statement.

201

202 **Overall adherence per TRIPOD item**

203 Five TRIPOD items reached at least 75% adherence (background, objectives, source of data,
204 limitations, and interpretation), whilst 12 TRIPOD items were below 25% adherence ([Figure 2](#)).
205 Results for the overall adherence per TRIPOD item stratified by study type, diagnosis and
206 prognosis, and publication year are shown in [Table 2](#).

207

208 ***Title and abstract (item 1 and 2)***

209 Seven out of 152 studies (4.6%, 95% CI 2.2 to 9.2) completely adhered to title
210 recommendations. Description of type of prediction model study (sub-item 1.i) was poorly
211 reported (11.2%, CI 7.0 to 17.2), but outcome to be predicted (sub-item 1.iv) was well
212 reported (91.4%, CI 85.9 to 94.9). No study fully reported item 2, abstract (0.0%, CI 0.0%to
213 2.5).

214

215 ***Introduction (item 3)***

216 Background and objectives were most often reported TRIPOD items. Background was
217 provided in 123 studies (80.9%, 95% CI 73.9 to 86.4), and the objectives were reported in 124
218 studies (81.6%, CI 74.6 to 86.9).

219

220 ***Methods (item 4-12)***

221 Source of data was the most often reported item in the methods section, and across all
222 TRIPOD items (98.0%, 95% CI 94.4 to 99.3). Study setting was reported in 107 studies (70.4%,
223 CI 62.7 to 77.1), eligibility criteria in 105 (69.1%, CI 61.3 to 75.9), and handling of predictors in
224 105 out of 152 studies (69.1%, CI 61.3 to 75.9). Ten studies assessed risk groups and five
225 reported complete information (50.0%, CI 23.7 to 76.3). Differences between development
226 and validation set were reported in 10 out of 19 applicable studies (52.6%, CI 31.7 to 72.7).
227 For 72 studies, definition of outcome was reported (47.4%, CI 39.6% to 55.3). Key study dates
228 such as start and end date of accrual, and length of follow-up were completely reported in 56
229 studies (36.8%, CI 29.6 to 44.7). Details of treatment were reported in 36 out of applicable 116
230 studies (31.0%, CI 23.3 to 39.9). Blinding of outcome and predictors were reported in 49
231 (32.2%, CI 25.3 to 40.0) and 7 studies (4.6%, CI 2.2 to 9.2), respectively.

232

233 Forty-four studies reported how missing data were handled (28.9%, 95% CI 22.3 to 36.6). The
234 missing data item consists of four sub-items of which three were rarely addressed in included
235 studies. Within 28 studies that reported handling of missing data: three studies reported the
236 software used (10.7%, CI 3.7 to 27.2), four studies reported the variables included in the
237 procedure (14.3%, CI 5.7 to 31.5) and no study reported the number of imputations (0.0%, CI
238 0.0 to 39.0). Predictor definitions were given in 32 out of 152 studies (21.1%, CI 15.3 to 28.2),
239 and justification of study size was reported in 27 studies (17.8%, CI 12.5 to 24.6). Model
240 building procedures, such as predictor selection and internal validation, were reported in 22
241 out of 152 studies (14.5%, CI 9.8 to 20.9). Internal validation, a sub-item of item 10b, was one
242 of the most reported sub-items across studies (91.4%, CI 85.9 to 94.9).

243

244 Reporting of measures used to assess and quantify the predictive performance was complete
245 in 19 studies (12.5%, 95% CI 8.2 to 18.7). Though 106 studies (69.7%, CI 62.0 to 76.5) reported

246 discrimination (sub-item 10d.i), only 19 studies (12.5%, CI 8.2 to 18.7) reported calibration
247 (sub-item 10d.ii). Definitions of discrimination and calibration are stated in supplemental
248 material. Other performance measures (sub-item 10d.iii), for example sensitivity, specificity, or
249 predictive values, were reported in 124 studies (81.6%, CI 74.7 to 86.9).

250

251 **Results (item 13-17)**

252 Study participant characteristics were reported in 38 out of 152 studies (25.0%, 95% CI 18.8 to
253 32.4). Basic demographics, at least age and gender (sub-item 13b.i), were provided in 117
254 studies (77.0%, CI 69.7 to 83.0), while summary information of the predictors (sub-item 13b.ii)
255 was reported in 67 studies (44.1%, CI 36.4 to 52.0). Number of study participants with missing
256 data for predictors (sub-item 13b.iii) was reported in 15 studies (24.2%, CI 15.2 to 36.2).
257 Unadjusted associations were reported in 41 out of the 74 studies that reported regression-
258 based models alongside with ML-models (41.9%, CI 31.3 to 53.3). The number of participants
259 and events were described in 37 studies (24.3%, CI 18.2 to 31.7). In 31 out of 152 studies, an
260 explanation on how to use the developed model to make predictions for new individuals was
261 provided, often in the form of a scoring rule or online calculator (20.4%, CI 14.8 to 27.5). Flow
262 of participants was reported in 6 studies (3.9%, CI 1.8 to 8.3) and model specification was
263 reported in 6 out of 116 applicable studies (5.2%, CI 2.4 to 10.8). Model predictive
264 performance was completely reported in 9 out of 152 studies (5.9%, CI 3.1 to 10.9).

265

266 **Discussion (items 18-20)**

267 Overall interpretation of results was reported in 124/152 studies (81.6%, 95% CI 74.7 to 86.9).
268 Limitations of the study were reported in 144 studies (94.7%, CI 90.0 to 97.3). An
269 interpretation of model performance in the validation set in comparison with the
270 development set was given in 14/19 studies (73.7%, CI 51.2 to 88.2). Potential clinical use and
271 implications for future research was reported in 61 studies (40.1%, CI 32.7 to 48.1).

272

273 **Other information (items 21 and 22)**

274 Availability of supplementary resources was mentioned in 93/152 studies (61.2, 95% CI 53.3
275 to 68.6). Funding information was reported in 42 studies (27.6%, CI 21.1 to 35.2).

276

277 **Overall adherence per article**

278 Overall adherence of studies to items of the TRIPOD Statement ranged between 13.0% and
279 65.0%; median adherence was 38.7% (IQR 31.0 to 46.5). The completeness reporting in
280 prognostic model studies was higher (median adherence=40.0% (IQR 33.3 to 46.8)) than
281 diagnostic model studies (median adherence=35.7% (IQR 30.2 to 45.0)) ([Figure 3](#)). Moreover,
282 median adherence was 40.6% (CI 28.6 to 46.1) in development (with internal validation)
283 studies, compared to 37.9% (CI 31.0 to 46.4) in development with external validation studies.

284 **DISCUSSION**

285 We conducted a systematic review of ML-based diagnostic and prognostic prediction model
286 studies and assessed their adherence to the TRIPOD Statement. We found that ML-based
287 prediction model studies adhere poorly to the TRIPOD Statement reporting items.

288

289 Complete reporting in titles and abstracts is crucial to identify and screen articles. However,
290 titles and abstracts were fully reported in less than 5% of articles. In addition, information
291 about methods was infrequently reported. Complete and accurate reporting of the methods
292 used to develop or validate a prediction model facilitates external validation, as well as
293 replication of study results by independent researchers. For example, to enhance
294 transparency and risk of bias assessment, it is recommended to report the number of
295 participants with missing data and report how missing data were handled in the analysis.
296 Handling of missing data was seldom reported, but this may be partially explained by the fact
297 that some ML techniques can handle missing data by design (e.g. sparsity aware splitting in
298 XGBoost and surrogate splits in decision trees).^{23,24} Also most studies divided a single dataset
299 into three: training, validation and test set; the last is used for internal validation. The split
300 sample approach for internal validation was among the most reported sub-items in our
301 sample, but several methodological studies and guidelines have long discouraged this
302 approach.²⁵ We included diagnostic model studies that used images as one of the predictors,
303 and deep learning. Often, these studies use several numerical variables based on pixels or
304 voxels and build prediction models based on several layers of statistical interaction. Both
305 topics are challenging to report due to number of variables used and poor interpretability of
306 interactions. This may explain why diagnostic ML-based model studies were slightly worse
307 reported compared to prognostic studies.

308

309 Overall, most articles adhered to less than half of the applicable items considered essential
310 for complete reporting. Authors may have avoided reporting specific details about methods
311 and results because their objective may be to explore the data and modeling technique
312 accuracy, rather than build models for individualized predictions in “real world” clinical
313 settings. However, high-quality reporting is also essential for reproducibility and replication.
314 Also, most developed models were unavailable for replication, assessment, or clinical
315 application. Only five studies reported using the TRIPOD Statement for reporting their

316 research. Although TRIPOD was published and disseminated in 2015, it is infrequently used
317 for reporting of ML-based prediction model studies.

318

319 Previous systematic reviews have shown poor reporting of regression-based prediction
320 model studies.^{7,8,10} One study assessed the completeness of reporting of articles published in
321 high impact journals during 2014 within 37 different clinical fields. In 146 studies, over half of
322 TRIPOD items were not fully reported, obtaining an overall adherence of 44% (IQR 35.0 to
323 52.0). Comparable to our study, the review found poor reporting of the title, abstract, model
324 building, model specification and model performance.⁷ A recent study assessed the
325 completeness of reporting of deep learning-based diagnostic model studies. Although they
326 developed their own data extraction for reporting quality, authors found poor reporting of
327 demographics, distribution of disease severity, patient flow, and distribution of alternative
328 diagnosis.²⁶ These items were also inappropriately reported in our study with a median
329 adherence between 0.0% and 47.3%. Another systematic review that assessed studies
330 comparing the performance of diagnostic deep learning algorithms for medical imaging
331 versus expert clinicians reported the overall adherence to TRIPOD was poor with a median of
332 62.0% (45.0 to 69.0).²⁷ In line with our results, a study about the performance of ML models
333 showed that 68.0% of included articles had unclear reporting.¹²

334

335 To our knowledge, this is the first systematic review evaluating the completeness of reporting
336 of supervised ML-based prediction model studies in a broad sample of articles. We ran a
337 validated search strategy and performed paired screening. We also used a contemporary
338 sample of studies in our review (2018-2019). Though some eligible articles may have been
339 missed, it is unlikely they would change the conclusions of this review.

340

341 We used a systematic scoring-system enhancing the objectivity and consistency for the
342 evaluation of adherence to a reporting guideline.²¹ We used the formal TRIPOD adherence
343 form and checklist for data extraction and assessment; however, these were developed for
344 studies developing prediction models with regression techniques. Although we applied the
345 option 'not applicable' for items that were unrelated to ML and items were excluded when
346 calculating overall adherence, our results should be interpreted within this context.

347 While some items and sub-items may be less relevant for prediction models developed with
348 ML techniques, other items are more relevant for transparent reporting in these studies. For
349 example, source of data (4a), study size (8), missing data (9), transformation of predictors
350 (10a.i), internal validation (10b.iv), and availability of the model (15b) acquire new relevance
351 within the context of ML-based prediction model studies. As ML techniques are prone to
352 overfitting, we recommend to extend item 10b of the TRIPOD adherence form to include a
353 new sub-item specifically related to penalization or shrinkage techniques. New reporting
354 items such as the hardware (i.e. technical aspects) that was used to develop or validate an
355 algorithm in images studies are needed, as well as data clustering. New practices such as
356 explaining models through feature importance plot or tuning of hyper-parameters could be
357 also added to the extension of TRIPOD for ML-based prediction models. Items such as
358 testing of interaction terms (Item 10b-iv), unadjusted associations (14b), and regression
359 coefficients (15a) require updating. Despite these recommendations, most TRIPOD items and
360 sub-items are still applicable for both, regression and ML techniques and should be used to
361 improve reporting quality.

362

363 We identified nearly 25 000 articles with prediction and ML-related terms within 2 years,
364 similar to previous systematic reviews about deep learning models.^{28,29} The literature has
365 become saturated with ML-based studies; thus, their identification, reporting and assessment
366 becomes even more relevant. If studies are presented without essential details to make
367 predictions in new patients, subsequent researchers will develop a new model, rather than
368 validating or updating an existing model. Reporting guidelines aim to increase the
369 transparent evaluation, replication and translation of prediction models into clinical
370 practice.³⁰ Some reporting guidelines for ML clinical prediction models have been developed.
371 ^{31,32} However, these guidelines are limited and do not follow the EQUATOR recommendations
372 for developing consensus-based reporting guidelines.³³ The improvement in reporting after
373 the introduction of a guideline has shown to be slow.³⁰ Improving the completeness of
374 reporting of ML-based studies might be even more challenging given the number of
375 techniques and associated details that need to be reported. There are also practical issues,
376 like terminology used, word limits, or journal requirements, that are acting as barriers to
377 complete reporting. To overcome these barriers, the use of online repositories for data, script,
378 and complete pipeline could help researchers share their models with enough details to

379 make predictions in new patients and to allow external validation of the model. Our results
380 will provide input and support for the development of TRIPOD-AI, an initiative launched in
381 2019.¹⁷ We call for a collaborative effort between algorithm developers, researchers, and
382 journal editors to improve the adoption of good scientific practices related to reporting
383 quality.

384 **CONCLUSION**

385 ML-based prediction model studies currently do not adhere well to the TRIPOD reporting
386 guideline. More than half of the TRIPOD items considered essential for transparent reporting
387 were inadequately reported, especially regarding details of title, abstract, blinding, model
388 building procedures, model specifications and model performance. Whilst ML brings new
389 challenges to the development of tailored reporting guidelines, our study serves as a baseline
390 measure to define future updates or extensions of TRIPOD tailored to ML modelling
391 strategies.

392 **Contributors**

393 The study concept and design were conceived by CLAN, JAAD, PD, LH, RDR, GSC, and KGMM.
394 CLAN, JAAD, TT, SN, PD, JM and RB conducted article screening and data extraction. CLAN
395 performed data analysis and JAAD verified the underlying data. CLAN wrote the first draft of
396 this manuscript, which was critically revised for important intellectual content by all authors
397 who have provided the final approval of this version. CLAN, the corresponding author, is the
398 guarantor of the review. The corresponding author attests that all listed authors meet
399 authorship criteria and that no others meeting the criteria have been omitted.

400 **Disclosures**

401 GSC, RDR and KGMM are members of the TRIPOD Group. All authors have nothing to
402 disclose.

403 **Data sharing**

404 The study protocol is available at doi: [10.1136/bmjopen-2020-038832](https://doi.org/10.1136/bmjopen-2020-038832). The search strategy is
405 available in appendix; detailed extracted data are available upon reasonable request.

406 **Acknowledgements**

407 The authors would like to thank and acknowledge the support of René Spijker, information
408 specialist.

409 **Funding support**

410 This study did not receive any specific grant from funding agencies in the public, commercial,
411 or not-for-profit sectors. GSC is supported by the National Institute for Health Research
412 (NIHR) Oxford Biomedical Research Centre (BRC) and by Cancer Research UK program grant
413 (C49297/A27294). PD is supported by the NIHR Oxford BRC. The views expressed are those of
414 the authors and not necessarily those of the NHS, NIHR, or Department of Health.

415 **Ethical approval**

416 Not required.

REFERENCES

1. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? *BMJ*. 2009;338(7706):1317-1320. doi:10.1136/bmj.b375
2. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med*. 2013;10(2). doi:10.1371/journal.pmed.1001381
3. Riley, Richard D; van der Windt, Danielle; Croft, Peter; Moons KGM. *Prognosis Research in Health Care: Concepts, Methods, and Impact*. Oxford University Press; 2019. doi:10.1093/med/9780198796619.001.0001
4. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*. 2016;353. doi:10.1136/bmj.i2416
5. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol*. 2019;188(12):2222-2239. doi:10.1093/aje/kwz189
6. Mitchell T. *Machine Learning*. McGraw Hill; 1997.
7. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: Towards a targeted implementation strategy of the TRIPOD statement. *BMC Med*. 2018;16(1):1-12. doi:10.1186/s12916-018-1099-2
8. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: A systematic review. *PLoS Med*. 2012;9(5). doi:10.1371/journal.pmed.1001221
9. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. *BMC Med*. 2011;9. doi:10.1186/1741-7015-9-103
10. Collins GS, De Groot JA, Dutton S, et al. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14(1):40. doi:10.1186/1471-2288-14-40
11. Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. *BMJ Open*. 2020;10(9):e041537. doi:10.1136/bmjopen-2020-041537
12. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
13. Gravesteyn BY, Nieboer D, Ercole A, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol*. 2020;122:95-107. doi:10.1016/j.jclinepi.2020.03.005
14. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383(9913):267-276. doi:10.1016/S0140-6736(13)62228-X
15. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698
16. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015;162(1):55. doi:10.7326/M14-0697
17. Collins GS, Moons KG. Reporting of artificial intelligence prediction models. Published online 2019. doi:10.1016/S0140-6736(19)30235-1
18. Andaur Navarro CL, Damen JAAG, Takada T, et al. Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. *BMJ Open*. 2020;10(11):1-6. doi:10.1136/bmjopen-2020-038832
19. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*. 2009;6(7). doi:10.1371/journal.pmed.1000097
20. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210. doi:10.1186/s13643-016-0384-4
21. Heus P, Damen JAAG, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: The example of TRIPOD for assessing completeness of reporting of prediction

- model studies. *BMJ Open*. 2019;9(4). doi:10.1136/bmjopen-2018-025611
22. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. doi:10.1016/j.jbi.2019.103208
 23. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol 13-17-August-2016. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785
 24. Therneau TM, Atkinson EJ. *An Introduction to Recursive Partitioning Using the RPART Routines*; 1997.
 25. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res*. 2017;26(2):796-808. doi:10.1177/0962280214558972
 26. Yusuf M, Atal I, Li J, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open*. 2020;10(3):e034568. doi:10.1136/bmjopen-2019-034568
 27. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ*. 2020;368. doi:10.1136/bmj.m689
 28. Faes L, Liu X, Wagner SK, et al. A clinician's guide to artificial intelligence: How to critically appraise machine learning studies. *Transl Vis Sci Technol*. 2020;9(2):7-7. doi:10.1167/tvst.9.2.7
 29. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Heal*. 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2
 30. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. *BMC Med*. 2010;8(1):24. doi:10.1186/1741-7015-8-24
 31. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J Med Internet Res*. 2016;18(12). doi:10.2196/jmir.5870
 32. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320-1324. doi:10.1038/s41591-020-1041-y
 33. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med*. 2010;7(2). doi:10.1371/journal.pmed.1000217

Table 1. TRIPOD adherence reporting items

Reporting Items	Study design	If applicable to studies	Reporting items for TRIPOD adherence		
			Development only	Development and validation	
1. Title	D,V		✓	✓	
2. Abstract	D,V		✓	✓	
Introduction					
3. Background and objectives	a. Context and rationale	D,V	✓	✓	
	b. Objectives	D,V	✓	✓	
Methods					
4. Source of data	a. Source of data	D,V	✓	✓	
	b. Key dates	D,V	✓	✓	
5. Participants	a. Study setting	D,V	✓	✓	
	b. Eligibility criteria	D,V	✓	✓	
	c. Details of treatment	D,V	✓	✓	
6. Outcome	a. Outcome definition	D,V	✓	✓	
	b. Blinding of outcome assessment	D,V	✓	✓	
7. Predictors	a. Predictors definition	D,V	✓	✓	
	b. Blinding of predictor assessment	D,V	✓	✓	
8. Sample size	Arrival at study size	D,V	✓	✓	
9. Missing Data	Handling of missing data	D,V	✓	✓	
10. Statistical analysis	a. Handling of predictors in the analysis	D	✓	✓	
	b. Specification of the model, all model building procedures, and internal validation methods	D	✓	✓	
	c. For validation, description of how predictions were calculated	V	✓	n.a.	
	d. Specification of all measures used to assess model performance	D,V	✓	✓	
	e. Description of model updating	V	✓	✓	n.a.
11. Risk groups	Details of how risk groups were created	D,V	✓	✓	
12. Development vs. validation	For validation, description of differences between development and validation data	V	✓	✓	
Results					
13. Participants	a. Flow of participants through the study	D,V	✓	✓	
	b. Description of characteristics of participants	D,V	✓	✓	
	c. For validation, comparison with development data	V	✓	✓	
14. Model development	a. Number of participants and outcome in each analysis	D	✓	✓	
	b. Unadjusted association between each candidate predictor and outcome	D	✓	✓	
15. Model specification	a. Presentation of full prediction model	D	✓	✓	
	b. Explanation of how to use the prediction model	D	✓	✓	
16. Model performance	Report of model performance measures	D,V	✓	✓	
17. Model updating	Results from any model updating	V	✓	✓	n.a.
Discussion					
18. Limitations	Limitations	D,V	✓	✓	

19. Interpretation	a. For validation, interpretation of performance measure results	V			✓
	b. Overall interpretation of results	D,V		✓	✓
20. Implications	Potential clinical use of the model and implications for future research	D,V		✓	✓
Other information				✓	✓
21. Supplementary information	Availability of supplementary resources	D,V		✓	✓
22. Funding	Source of funding and role of funders	D,V		✓	✓
Total number of applicable items for TRIPOD adherence score				31	37

(n.a) No included studies reported external validation only or model updating (Item 10c, 10e, and 17)

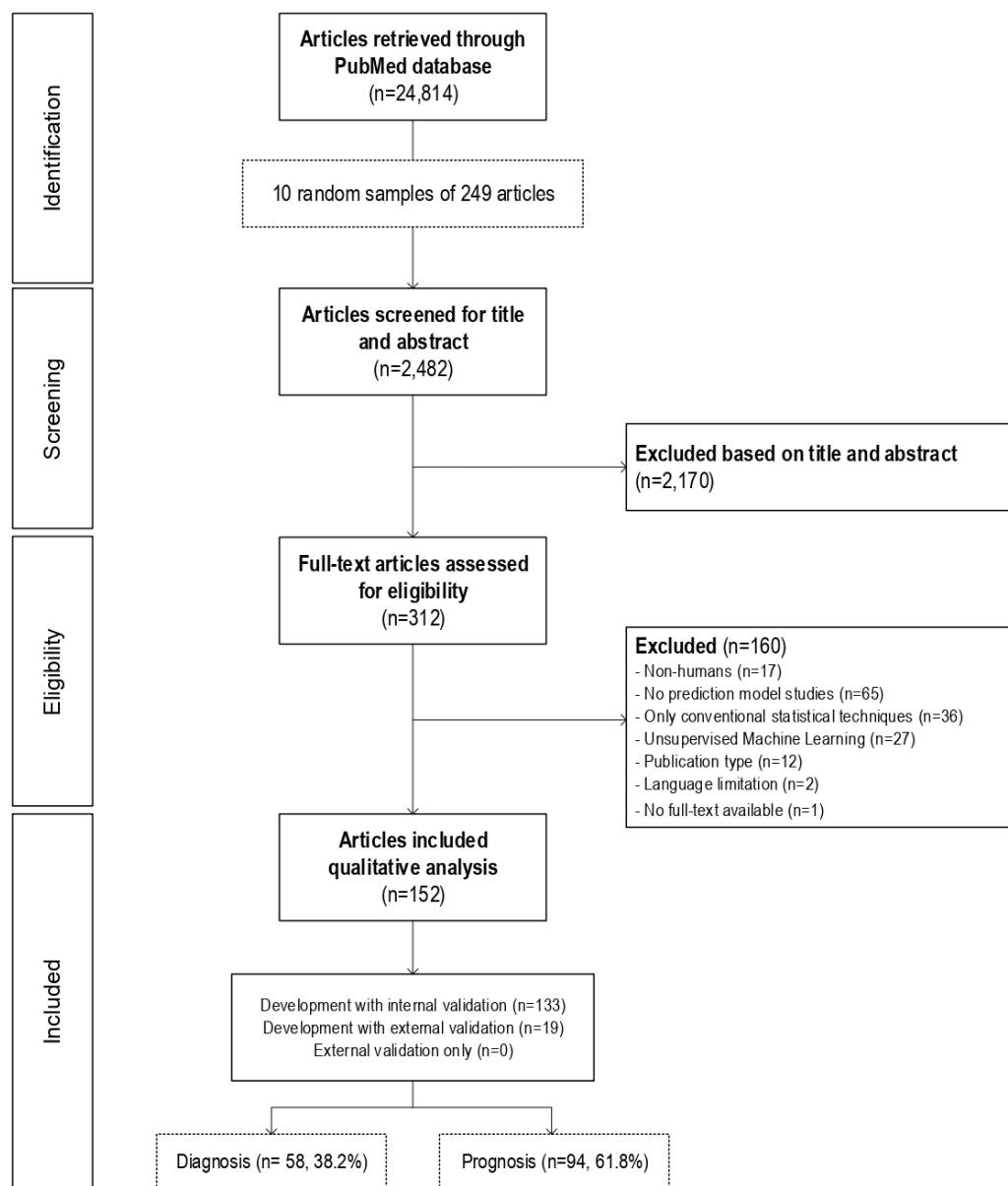


Figure 1. Flowchart of included studies

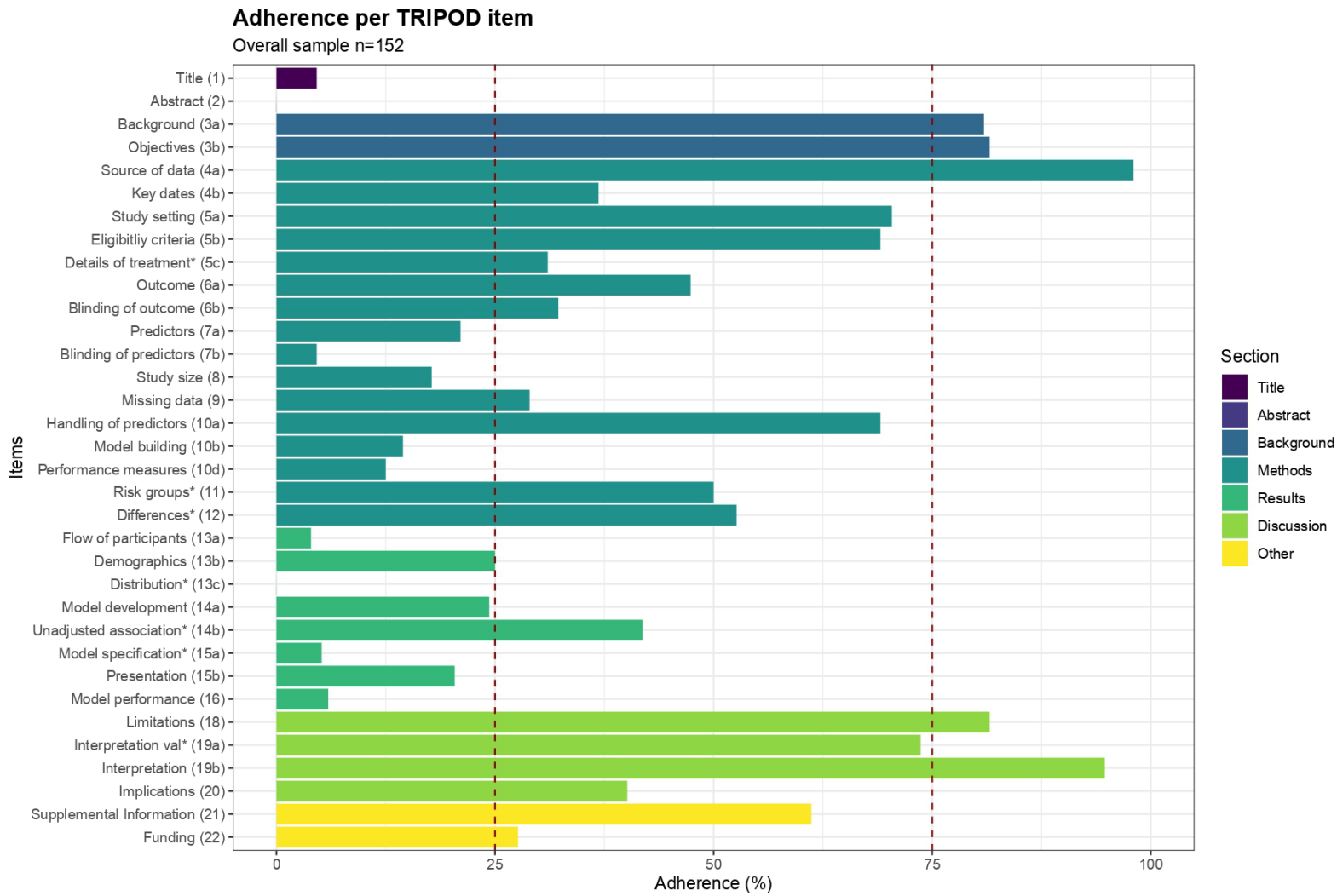
Table 2. Adherence to TRIPOD items

TRIPOD Items	Adherence to TRIPOD items (%)						
	Overall	Development only	Development with external validation	Prognosis	Diagnosis	2018	2019
	n=152 [95% CI]	n=133 [95% CI]	n=19 [95% CI]	n=94 [95% CI]	n=58 [95% CI]	n=65 [95% CI]	n=87 [95% CI]
Title (1)	4.6 [2.2 to 9.2]	3.8 [1.6 to 8.5]	10.5 [2.9 to 31.4]	7.4 [3.7 to 14.6]	0.0 [0.0 to 6.2]	3.1 [0.8 to 10.5]	5.7 [2.5 to 12.8]
Abstract (2)	0.0 [0.0 to 2.5]	0.00 [0.0 to 2.8]	0.00 [0 to 16.8]	0.0 [0.0 to 3.9]	0.0 [0.0 to 6.2]	0.0 [0.0 to 5.6]	0.0 [0.0 to 4.2]
Background (3a)	80.9 [73.9 to 86.4]	79.7 [72.1 to 85.7]	89.5 [68.6 to 97.1]	83.0 [74.1 to 89.2]	77.6 [65.3 to 86.4]	84.6 [73.9 to 91.4]	78.2 [68.4 to 85.5]
Objectives (3b)	81.6 [74.7 to 86.9]	78.9 [71.3 to 85.0]	100 [83.2 to 100]	83.0 [74.1 to 89.2]	79.3 [67.2 to 87.7]	84.6 [73.9 to 91.4]	79.3 [69.6 to 86.5]
Source of data (4a)	98.0 [94.4 to 99.3]	98.5 [94.7 to 99.6]	94.7 [75.4 to 99.7]	98.9 [94.2 to 99.9]	96.6 [88.3 to 99.0]	98.5 [91.8 to 99.9]	97.7 [92.0 to 99.4]
Key dates (4b)	36.8 [29.6 to 44.7]	38.3 [30.5 to 46.8]	26.3 [11.8 to 48.8]	33.0 [24.3 to 43.0]	43.1 [31.2 to 55.9]	40.0 [29.0 to 52.1]	34.5 [25.3 to 44.9]
Study settings (5a)	70.4 [62.7 to 77.1]	72.2 [64.0 to 79.1]	57.9 [36.3 to 76.9]	73.4 [63.7 to 81.3]	65.5 [52.7 to 76.4]	75.4 [63.7 to 84.2]	66.7 [56.2 to 75.7]
Eligibility criteria (5b)	69.1 [61.3 to 75.9]	71.4 [63.2 to 78.4]	52.6 [31.7 to 72.7]	72.3 [62.6 to 80.4]	63.8 [50.9 to 74.9]	69.2 [57.2 to 79.1]	69.0 [58.6 to 77.7]
Details of treatment* (5c)	31.0 [23.3 to 39.9]	28.7 [20.8 to 38.2]	46.7 [24.8 to 69.9]	30. [22.0 to 41.0]	32.1 [17.9 to 50.7]	25.5 [15.5 to 38.9]	35.4 [24.9 to 47.5]
Outcome (6a)	47.4 [39.6 to 55.3]	47.4 [39.1 to 55.8]	47.4 [27.3 to 68.3]	52.1 [42.1 to 61.9]	39.7 [28.1 to 52.5]	47.7 [36.0 to 59.6]	47.1 [37.0 to 57.5]
Blinding of outcome (6b)	32.2 [25.3 to 40.0]	33.1 [25.7 to 41.5]	26.3 [11.8 to 48.8]	40.4 [31.1 to 50.5]	19.0 [10.9 to 30.9]	35.4 [24.9 to 47.5]	29.9 [21.3 to 40.2]
Predictors (7a)	21.1 [15.3 to 28.2]	22.6 [16.3 to 30.4]	10.5 [2.9 to 31.4]	18.1 [11.6 to 27.1]	25.9 [16.3 to 38.4]	18.5 [10.9 to 29.6]	23.0 [15.4 to 32.9]
Blinding of predictors (7b)	4.6 [2.2 to 9.2]	5.3 [2.6 to 10.5]	0.00 [0.0 to 16.8]	2.1 [0.6 to 7.4]	8.6 [3.7 to 18.6]	0.0 [0.0 to 5.6]	8.0 [4.0 to 15.7]
Study size (8)	17.8 [12.5 to 24.6]	18.0 [12.4 to 22.4]	15.8 [5.5 to 37.6]	20.2 [13.3 to 29.4]	13.7 [7.2 to 24.9]	16.9 [9.7 to 27.8]	18.4 [11.6 to 27.8]

Missing data (9)	28.9 [22.3 to 36.6]	30.8 [23.6 to 39.1]	15.8 [5.5 to 37.6]	29.8 [21.5 to 39.7]	27.6 [17.8 to 40.2]	30.8 [20.9 to 42.8]	27.6 [19.3 to 37.8]
Handling of predictors (10a)	69.1 [61.3 to 75.9]	72.2 [64.0 to 79.1]	47.4 [27.3 to 68.3]	66.0 [55.9 to 74.7]	74.1 [61.6 to 83.7]	69.2 [57.2 to 79.1]	69.0 [58.6 to 77.7]
Model building (10b)	14.5 [9.8 to 20.9]	12.8 [8.1 to 19.5]	26.3 [11.8 to 48.8]	13.8 [8.3 to 22.2]	15.5 [8.4 to 26.9]	18.5 [10.9 to 29.6]	11.5 [6.4 to 19.9]
Predictor's calculation (10c)	-	-	-	-	-	-	-
Performance measures (10d)	12.5 [8.2 to 18.7]	12.0 [7.5 to 18.6]	15.8 [5.5 to 37.6]	19.1 [12.5 to 28.3]	1.7 [0.1 to 9.1]	10.8 [5.3 to 20.6]	13.8 [8.1 to 22.6]
Model updating (10e)	-	-	-	-	-	-	-
Risk groups* (11)	50.0 [23.7 to 76.3]	50.0 [23.7 to 76.3]	0.0	28.6 [8.2 to 64.1]	100.0 [43.9 to 100]	50.0 [15.0 to 85.0]	50.0 [18.8 to 81.2]
Development vs. validation (12)	52.6 [31.7 to 72.7]	NA	52.6 [31.7 to 72.7]	33.3 [13.8 to 60.9]	85.7 [48.7 to 99.3]	44.4 [18.9 to 73.3]	60.0 [31.3 to 83.2]
Flow of participants (13a)	3.9 [1.8 to 8.3]	4.5 [2.1 to 9.5]	0.0 [0.0 to 16.8]	2.1 [0.6 to 7.4]	6.9 [2.7 to 16.4]	3.1 [0.8 to 10.5]	4.6 [1.8 to 11.2]
Demographics (13b)	25.0 [18.8 to 32.4]	26.3 [19.6 to 34.4]	15.8 [5.5 to 37.6]	22.3 [15.1 to 31.8]	29.3 [19.2 to 42.0]	29.2 [19.6 to 41.2]	21.8 [14.5 to 31.6]
Distribution (13c)	0.0 [0.0 to 16.8]	NA	0.0 [0.0 to 16.8]	0.0 [0.0 to 24.2]	0.0 [0.0 to 35.4]	0.0 [0.0 to 29.9]	0.0 [0 to 27.8]
Model development (14a)	24.3 [18.2 to 31.7]	24.8 [18.2 to 32.8]	21.1 [8.5 to 43.3]	19.1 [12.5 to 28.3]	32.8 [22.1 to 45.6]	26.2 [17.0 to 38.0]	23.0 [15.4 to 32.9]
Unadjusted association* (14b)	41.9 [31.3 to 53.3]	41.2 [30.3 to 53.0]	50.0 [18.8 to 81.2]	50.0 [35.8 to 64.2]	30.0 [16.7 to 47.9]	37.5 [22.9 to 54.7]	45.2 [31.2 to 60.1]
Model specification (15a)	5.2 [2.4 to 10.8]	4.0 [1.6 to 9.8]	12.5 [3.5 to 36.0]	5.6 [3.5 to 12.4]	4.5 [1.3 to 15.1]	4.0 [1.1 to 13.5]	6.1 [2.4 to 14.6]
Presentation (15b)	20.4 [14.8 to 27.5]	20.3 [14.3 to 27.9]	21.1 [8.5 to 43.3]	21.3 [14.2 to 30.6]	19.0 [10.9 to 30.9]	23.1 [14.5 to 34.6]	18.4 [11.6 to 27.8]
Model performance (16)	5.9 [3.1 to 10.9]	5.3 [2.6 to 10.5]	10.5 [2.9 to 31.4]	9.6 [5.1 to 17.2]	0.0 [0.0 to 6.2]	7.7 [3.3 to 16.8]	4.6 [1.8 to 11.2]
Updating results (17)	-	-	-	-	-	-	-
Limitations (18)	81.6 [74.7 to 86.9]	80.5 [72.9 to 86.3]	89.5 [68.6 to 97.1]	83.0 [74.1 to 89.2]	79.3 [67.2 to 87.7]	86.2 [75.7 to 92.5]	78.2 [68.4 to 85.5]

Interpretation validation (19a)	73.7 [51.2 to 88.2]	NA	73.7 [51.2 to 88.2]	75.0 [46.8 to 91.1]	71.4 [35.9 to 91.8]	77.8 [45.3 to 93.7]	70.0 [39.7 to 89.2]
Interpretation (19b)	94.7 [90.0 to 97.3]	94.0 [88.6 to 96.9]	100 [83.2 to 100]	95.7 [89.6 to 98.3]	93.1 [83.6 to 97.3]	93.8 [85.2 to 97.6]	95.4 [88.8 to 98.2]
Implications (20)	40.1 [32.7 to 48.1]	39.1 [31.2 to 47.6]	47.4 [27.3 to 68.3]	41.5 [32.1 to 51.6]	37.9 [26.6 to 50.8]	43.1 [31.8 to 55.2]	37.9 [28.5 to 48.4]
Supplemental Information (21)	61.2 [53.3 to 68.6]	58.6 [50.1 to 66.7]	78.9 [56.7 to 91.5]	63.8 [53.8 to 72.8]	56.9 [44.1 to 68.8]	61.5 [49.4 to 72.4]	60.9 [50.4 to 70.2]
Funding (22)	27.6 [21.1 to 35.2]	26.3 [19.6 to 34.4]	36.8 [19.1 to 59.0]	28.7 [20.6 to 38.6]	25.9 [16.3 to 38.4]	24.6 [15.8 to 36.3]	29.9 [21.3 to 40.2]

Item **10c**, **10e** and **17**, could not be assessed as they are only applicable to studies reporting on *external validation only* and studies including *model update* which were unavailable in our sample; **(*)** If applicable to studies; **(NA)** Item not applicable to study type; **red cells** are items with reporting quality *below 25%*; **green cells** are items with reporting quality *above 75%*. CI: Confidence Interval. Results refer to first model reported.



(*) If applicable to studies. Items 10c, 10e, and 17 are not applicable. Results section considered first model reported

Figure 2. Overall adherence per TRIPOD item

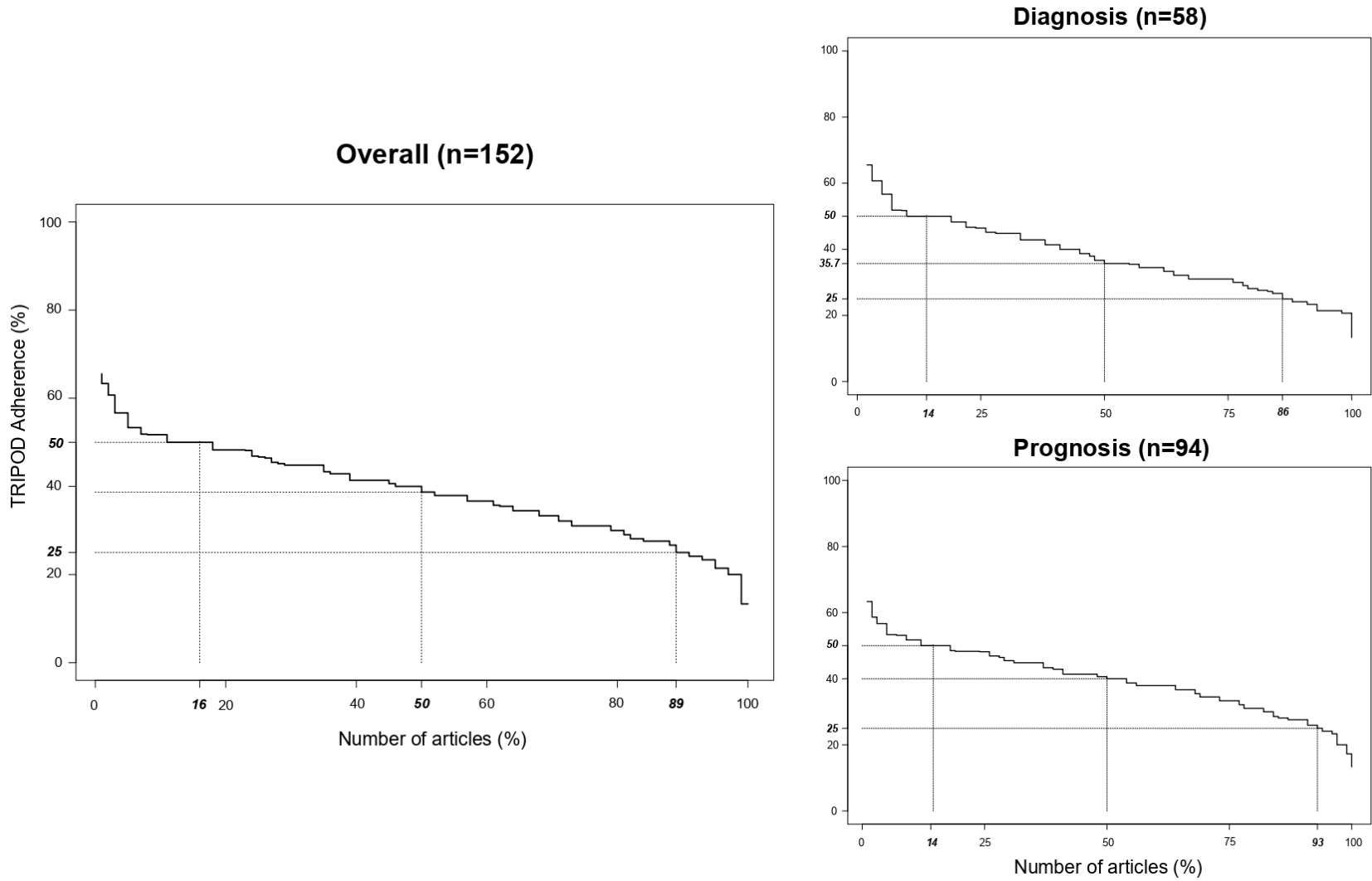


Figure 3. Overall adherence per article