

1 **Title:** Inter-rater reliability of the Infectious Disease Modeling Reproducibility Checklist
2 (IDMRC) as applied to COVID-19 computational modeling research

3
4 Darya Pokutnaya^{1*}, Willem G Van Panhuis², Bruce Childers³, Marquis S Hawkins¹, Alice E
5 Arcury-Quandt¹, Meghan Matlack⁴, Kharlya Carpio¹, Harry Hochheiser⁵

6
7 ¹University of Pittsburgh, Department of Epidemiology; Pittsburgh, Pennsylvania, United States
8 of America

9 ²Office of Data Science and Emerging Technologies, National Institute of Allergy and Infectious
10 Diseases; Rockville, Maryland, United States of America [note that Dr. Van Panhuis completed
11 the research described in this paper during his time at the University of Pittsburgh, before
12 starting his position at NIAID]

13 ³University of Pittsburgh, Department of Computer Science; Pittsburgh, Pennsylvania, United
14 States of America

15 ⁴University of Pittsburgh, Department of Environmental and Occupational Health, Pittsburgh,
16 PA, USA

17 ⁵University of Pittsburgh, Department of Biomedical Informatics, Intelligent Systems Program,
18 and Clinical and Translational Science Institute; Pittsburgh, Pennsylvania, United States of
19 America

20 *Corresponding author.

21 Email: dap184@pitt.edu

22

23 **Abstract**

24

25 **Background:** Infectious disease computational modeling studies have been widely published

26 during the coronavirus disease 2019 (COVID-19) pandemic, yet they have limited

27 reproducibility. Developed through an iterative testing process with multiple reviewers, the

28 Infectious Disease Modeling Reproducibility Checklist (IDMRC) enumerates the minimal

29 elements necessary to support reproducible infectious disease computational modeling

30 publications. The primary objective of this study was to assess the reliability of the IDMRC and

31 to identify which reproducibility elements were unreported in a sample of COVID-19

32 computational modeling publications. **Methods:** Four reviewers used the IDMRC to assess 46

33 preprint and peer reviewed COVID-19 modeling studies published between March 13th, 2020,

34 and July 31st, 2020. The inter-rater reliability was evaluated by mean percent agreement and

35 Fleiss' kappa coefficients (κ). Papers were ranked based on the average number of reported

36 reproducibility elements, and average proportion of papers that reported each checklist item were

37 tabulated. **Results:** Questions related to the computational environment (mean $\kappa = 0.90$, range =

38 $0.90-0.90$), analytical software (mean $\kappa = 0.74$, range = $0.68-0.82$), model description (mean $\kappa =$

39 0.71 , range = $0.58-0.84$), model implementation (mean $\kappa = 0.68$, range = $0.39-0.86$), and

40 experimental protocol (mean $\kappa = 0.63$, range = $0.58-0.69$) had moderate or greater ($\kappa > 0.41$)

41 inter-rater reliability. Questions related to data had the lowest values (mean $\kappa = 0.37$, range =

42 $0.23-0.59$). Reviewers ranked similar papers in the upper and lower quartiles based on the

43 proportion of reproducibility elements each paper reported. While over 70% of the publications

44 provided data used in their models, less than 30% provided the model implementation.

45 **Conclusions:** The IDMRC is the first comprehensive, quality-assessed tool for guiding

46 researchers in reporting reproducible infectious disease computational modeling studies. The

47 inter-rater reliability assessment found that most scores were characterized by moderate or
48 greater agreement. These results suggests that the IDMRC might be used to provide reliable
49 assessments of the potential for reproducibility of published infectious disease modeling
50 publications. Results of this evaluation identified opportunities for improvement to the model
51 implementation and data questions that can further improve the reliability of the checklist.

52 **Keywords:** Reproducibility, infectious disease, epidemiology, modeling, COVID-19,
53 coronavirus disease 2019

54 **Background**

55
56 Throughout the coronavirus disease 2019 (COVID-19) pandemic, policy makers relied
57 extensively on epidemiological, biostatistical, and computational infectious disease models to
58 inform decisions regarding public health interventions (1). Although there was increased interest
59 in transparent research prior to the pandemic (2), increasingly complex modeling methods and
60 frequently insufficiently detailed descriptions of those methods have led to increasing
61 reproducibility concerns. We recently proposed the Infectious Disease Modeling Reproducibility
62 Checklist (IDMRC) a comprehensive set of guidelines that researchers can follow to publish
63 reproducible modeling results (3). Our goal in this paper is to assess the reliability of the IDMRC
64 to facilitate the reporting of elements impacting the reproducibility of COVID-19 research.

65 Reproducibility is a cornerstone of the scientific method, enabling the verification of
66 discoveries and protecting against scientific misconduct (4). However, the rapid pace of COVID-
67 19 research has raised concerns about the reproducibility of modeling results. For years
68 governing bodies have published advice to enhance reproducibility of scientific research and
69 proposed lists of elements that should be included in publications to ensure reproducibility have
70 been reported (5–9). Prior to our work, these initiatives have not been synthesized into reliable

71 guidelines for infectious disease computational modeling research. We filled this critical gap in
72 the literature by creating a framework for the implementation of reproducible computational
73 infectious disease models. We formatted the framework into the Infectious Disease Modeling
74 Reproducibility Checklist (IDMRC), a checklist that is applicable to varying types of infectious
75 disease models with ranging complexities (3).

76 Previously developed guidelines, such as the Strengthening the Reporting of
77 Observational Studies in Epidemiology (STROBE) checklist and the EPIFORGE 2020
78 guidelines for epidemic forecasting have been instrumental in enhancing the quality of modeling
79 research (10,11). However, they focus on general recommendations for describing elements in a
80 publication without including specific items related to data, analytical software, operating
81 systems (including both names and version numbers), and other key computational components
82 used to conduct the analyses. The IDMRC overcomes these limitations through the inclusion of
83 specific items relevant to publishing reproducible infectious disease modeling studies. Here, we
84 assess the reliability of the IDMRC with multiple reviewers and a sample of COVID-19
85 computational modeling studies. To our knowledge, this is the first time the reliability of a
86 checklist used to assess the reproducibility of infectious disease modeling studies has been
87 evaluated.

88 The Models of Infectious Disease Agent Study (MIDAS) Coordination Center
89 (midasnetwork.us) is an NIGMS-funded center supporting the infectious disease research
90 community. Four researchers from the MIDAS Coordination Center evaluated the reliability of
91 the checklist by assessing a random selection of preprint and peer-reviewed COVID-19 modeling
92 papers published between March 13th, 2020, and July 31st, 2020. The purpose of this study was
93 to assess the inter-rater reliability of the IDMRC, characterize papers based on the reviewers'

94 qualitative rankings, and determine which reproducibility elements are frequently included or
95 overlooked in COVID-19 computational modeling studies.

96 **Methods**

97
98 The IDMRC was previously developed as a framework for the implementation of
99 reproducible computational infectious disease models (Additional File 1) (3). The IDMRC
100 consists of twenty-two questions grouped into six categories: computational environment,
101 analytical software, model description, model implementation, data, and experimental protocol
102 (Additional File 2). We evaluated the performance of the IDMRC in the COVID-19 modeling
103 literature by measuring the agreement among four reviewers for the overall instrument and for
104 individual questions. Based on the evaluations, we made suggested changes to the IDMRC
105 (Additional File 3).

106 We searched PubMed, medRxiv, arXiv, and bioRxiv using queries for COVID-19
107 modeling papers between March 13th, 2020, and July 31st, 2020 (Additional File 4). As preprint
108 servers were widely used to disseminate COVID-19 models at the beginning of the pandemic
109 (12), we included medRxiv, arXiv, and bioRxiv in our search. We did not restrict to certain types
110 of computational modeling studies in our assessment given that our checklist should be
111 applicable to all computational infectious disease modeling studies ranging from regression
112 models to complex agent-based models. From the search results, we randomly selected 100
113 papers for title and abstract review (Figure 1).

114 Four researchers (DP, AAQ, KC, MM) used the IDMRC to independently review the 46
115 modeling papers to assess which IDMRC elements were included. All four reviewers had
116 experience in reading modeling papers, including training of at least a Master's in Public Health

117 Degree. DP and AAQ were involved with the development of the checklist and had more
118 experience using the IDMRC relative to KC and MM.

119 We performed an inter-rater reliability analysis to assess the concordance of the ratings of
120 the ordinal categorical items. Reliability was assessed based on the mean percent agreement with
121 Wald 95% confidence intervals and Fleiss' kappa (κ) estimates. Fleiss' kappa is the observed
122 agreement corrected for the agreement expected by chance and is appropriate when there are
123 more than two raters assessing ordinal or nominal data (13). We used the Power4Cats function in
124 the kappaSize package in R version 4.0.2, RStudio Version 1.3.107 to determine that 46
125 publications could reliably produce a lower limit for a kappa estimate of 0.293 (1). Fleiss' kappa
126 was computed with linear weights using the wlin.conc function in the R raters package 2.0.1
127 (14,15). Monte Carlo simulations were used to calculate percentile bootstrap confidence
128 intervals. Results were interpreted using previously published guidelines: $\kappa < 0.01$ indicates no
129 agreement; $\kappa = 0.01-0.20$, slight; $\kappa = 0.21-0.40$, fair; $\kappa = 0.41-0.60$, moderate; $\kappa = 0.61-0.80$
130 substantial; and $\kappa = 0.81-1$ almost perfect agreement (15). A kappa score below 0.41 falls into
131 the category of "slight" agreement and was deemed by the authors as indicative of questions that
132 needed to be reviewed and revised.

133 For each reviewer, we qualitatively ranked the papers by the number of reported elements
134 in each publication. Publications with the most elements included (as rated by the reviewers)
135 were ranked the highest. We also averaged the reviewers' rankings to report the average
136 qualitative rankings of the 46 publications. To assess if the potential impact of the peer-review on
137 the number of reproducibility elements, DP independently reviewed the five highest-rated and 5
138 lowest-rated publications to determine 1) if any of the publications that were published in
139 preprint servers at the time of the review had since been published in peer-reviewed journals, and

140 2) if the papers that had been published in peer-reviewed journals had reported more
141 reproducibility elements in those publications. Finally, we tabulated the proportion of papers that
142 reported each checklist element as well as the proportion of checklist elements reported in all
143 publications (both averaged across all reviewers).

144 **Results**

145 Four MIDAS researchers used the IDMRC to review 46 COVID-19 computational
146 modeling papers published between March 13th, 2020, and July 31st, 2020. After title and
147 abstract review, 48 papers were excluded based on the following exclusion criteria:
148 observational, genomic, immunological, and molecular studies, commentaries, reviews,
149 retractions, letters to editor, response papers, papers not related to COVID-19, and descriptions
150 of software applications. Of the remaining 48 papers, two publications reviewing previously
151 developed COVID-19 models were excluded after full text review (Figure 1). The final 46 paper
152 sample consisted of 39 (85%) publications published in preprint servers (n = 34 from medRxiv; n
153 = 5 from arXiv) and 7 (15%) from peer-reviewed journals (Additional File 5).

154 **[Figure 1. Publications included in the inter-rater reliability analysis of the Infectious Disease**
155 **Modeling Reproducibility Checklist. Abbreviations: COVID-19, coronavirus disease 2019]**

156

157

158 *Inter-rater reliability of the IDMRC*

159

160 The inter-rater reliability evaluation indicated that the IDMRC was a reliable tool with
161 most questions characterized by moderate or better ($\kappa > 0.41$) agreement between the four
162 reviewers. Overall, the mean percent agreement ranged from 54% (data question 5.3) to 94%
163 (computational environment 1.1, 1.2; model implementation 4.6). Fleiss' kappa estimates ranged
164 from 0.23 (95%CI 0.10, 0.40) for IDMRC data question 5.5 to 0.90 (95%CI 0.79, 0.98) for both

165 computational environment questions. Several Fleiss' kappa estimates in the model

166 implementation and data categories fell below moderate agreement (Table 1).

167 **[Table 1. Infectious Disease Modeling Reproducibility Checklist elements reported in COVID-**
168 **19 modeling papers.]**

169

170 *Characterization of papers based on reviewer qualitative rankings*

171 Reviewers identified similar publications as reporting the most reproducibility elements

172 (i.e., reviewers reported “yes” for more questions) or the least number of elements (i.e.,

173 reviewers reported “yes” less often). KC and MM, the two reviewers with the least experience

174 using the IDMRC, agreed upon eight publications in the top 25% (n = 13) and the eight

175 publications in the bottom 25% (n = 13) (Additional File 6). DP and AAQ, the two reviewers

176 with more experience using the checklist, agreed on nine publications in the top 25% and ten

177 publications in the bottom 25% (Additional File 6). All four reviewers agreed on six publications

178 in the top 25% and seven publications were reported in the bottom 25% (Additional File 7). The

179 publications with the most reproducibility elements based on average scores (publications 9, 15,

180 16, 19, and 27) and the least reported reproducibility elements (2, 12, 20, 23, and 37) were all

181 originally published as preprints. Four publications (2, 19, 20, and 27) have since been published

182 in peer-reviewed journals. An independent review of these four papers by DP determined that the

183 peer-reviewed versions did not have a significantly increased number of reported reproducibility

184 elements.

185 **[Figure 2. Average quantitative paper ranking (n = 46) among four reviewers. Green bars**

186 **correspond to the average number of reported elements in each publication (“yes” responses);**

187 **yellow indicates partially reported elements; red indicates not reported elements, gray indicates**

188 **not applicable responses.]**

189

190 *Average proportion of papers that reported each checklist item*

191 Rates of inclusion of the 22 checklist elements varied from 2% of papers reporting the
192 operating system version (question 1.2) to 92% providing the model description in the journal or
193 publication as opposed to referencing a previously developed model (question 3.2, Figure 3A).
194 Fifty percent (n = 23) of publications provided less than 40% of all checklist categories (Figure
195 3B). Over 94% of studies did not provide either the name or the version of the operating system
196 used in their analysis (questions 1.2, 1.3, respectively) (Figure 3A). The analytical software name
197 (e.g., R, STATA, SAS) was provided in 62% of publications (question 2.1), but only 41% of the
198 software tools were openly accessible without a licensing fee (question 2.2). Most studies
199 provided the input data (70%; question 5.4); however, only 25% provided the model
200 implementation, or code, used to generate the data (question 4.1). Averaged across the raters,
201 over 50% of the publications provided all five data elements (questions 5.1–5.5), but less than
202 50% of the publications provided all six model implementation elements (questions 4.1–4.6).
203 Thirty-nine percent of publications provided the parameters used in their models (questions 6.1)
204 while 19% provided a clear explanation for how categories 1–5 were used together to create the
205 model results (question 6.2).

206 **[Figure 3.** Proportion of coronavirus disease 2019 (COVID-19) modeling publications (n = 46)
207 that reported each infectious disease modeling reproducibility checklist (IDMRC) component
208 elements. A, average percentage of papers that reported each checklist element; B, average
209 proportion of checklist elements that were reported in all publications. Dashed line in B indicates
210 the mean.]

211

212 **Discussion**

213

214 Improved reproducibility of infectious disease computational models will help

215 researchers efficiently build upon previous studies and accelerate the pace of scientific

216 advancements. We previously developed the Infectious Disease Modeling Reproducibility
217 Checklist (IDMRC) to enumerate the elements necessary to support reproducible infectious
218 disease computational modeling studies (3). Our evaluation indicated that the IDMRC is a
219 reliable tool with the majority of the inter-rater reliability estimates reporting moderate or greater
220 agreement between the four reviewers. Participating reviewers placed similar publications in the
221 top and bottom reproducibility score quantiles based on the number of elements missing in each
222 publication. The two experienced and the two novice checklist users had more similar rankings,
223 suggesting that formally training researchers to use the IDMRC prior to evaluating a study may
224 produce more consistent results. Furthermore, revisions to the checklist questions, primarily in
225 the model implementation and data sections, may increase reliability of future evaluations.

226 Our experience with the application of the checklist suggests that the question ordering
227 may have impacted reliability. The question regarding whether the model implementation
228 computer language was documented (question 4.4) may have received a lower score due to its
229 positioning after the analytical software name question (question 2.1). In most instances if a
230 publication reported the analytical software name (e.g., R, STATA), the model implementation
231 computer language would be evident (e.g., R uses R coding language, STATA uses STATA
232 coding language). However, occasionally the two may differ, such as when researchers use their
233 own developed software or utilize packages to develop scripts in languages that are not original
234 to the analytical software (e.g., writing Python scripts in R with the use of the *reticulate*
235 package). Additionally, if a reviewer had already selected “no” or “not applicable” for prior a
236 model implementation question (questions 4.1–4.3), the reviewer may have automatically
237 selected the same response for question 4.4 without independent thought to the question. Moving
238 the question from the model implementation section to the analytical software section (after

239 question 2.1) could improve the checklist reliability. We propose a revised version of the
240 checklist which includes question 4.4 directly after question 2.1 (Additional File 3).

241 Reliability assessments can also help highlight ambiguities in definitions commonly used
242 in infectious disease computational modeling literature. Lower κ estimates in the data section
243 may have been due to uncertainty regarding the definition of input data (question 5.1). For
244 example, some of the publications described susceptible-infected-recovered (SIR)
245 compartmental models which can be parametrized using input data, simulated data, or by
246 referencing previously reported parameters. In these situations, reviewers may have not
247 considered parameters as input data. To improve the reliability of the checklist, we defined input
248 data as “any data, including parameters, used to generate a model or initial conditions” in the
249 updated version of the checklist (Additional File 3). Furthermore, we originally included a “not
250 applicable” answer choice in question 5.1; however, after the reliability assessment, we deemed
251 that this answer choice was not warranted because a response of “yes” or “no” should capture all
252 possible answer choices. Thus, we removed the “not applicable” answer choice from question
253 5.1 in the newest version of the checklist (Additional File 3). Given that conditional nature of the
254 checklist questions (i.e., subsequent checklist questions are affected by prior responses),
255 including a definition of the input data as well as correcting the answer choices in question 5.1
256 could improve the reliability of the following data questions.

257 The computational environment, which comprises the operating system name and
258 version, was least reported. Failure to reproduce modeling studies, even if the data and code have
259 been made available, can be due to incompatibilities or specific requirements in the
260 computational environment (16). For example, SAS software is not compatible with the macOS
261 operating system unless it is run in a virtual machine. Software developed by the authors of a

262 given paper or analyses that require high-performance computing may also require specific types
263 of operating systems to be functional. We recommend including a short statement specifying the
264 name and version of the operating system in future infectious disease computational modeling
265 studies .

266 Two of the top qualitatively ranked publications (19 and 27) as well as two of the lower
267 ranked publications (2 and 20) were initially published in medRxiv, during the time of review,
268 but have since been published in peer reviewed journals. An independent review by DP indicated
269 that the peer reviewed versions of these paper did not include significant improvement in the
270 number of reported reproducibility elements. This suggests that the peer review process does not
271 necessarily improve the reproducibility of papers in our sample. Despite an increase in the
272 adoption of data and code sharing policies by journals, stricter application of the IDMRC or
273 similar guidelines may be needed to further improve reproducibility during the peer review
274 process (17). Some suggestions include the complementary submission of checklists, such as the
275 IDMRC, or dynamic computational notebooks (17,18).

276 Over 70% of the publications provided the data used for their analysis. Our estimate was
277 similar to the 60% (n = 29) of CDC-compiled COVID-19 modeling studies analyzed by Jalali et
278 al. and much higher than the 24.8% (n = 332) reviewed by Ioannidis et al. that reported to share
279 their data. However, Ioannidis et al., used a text mining algorithm which may not have picked up
280 publications that shared their data (19,20). Many journals now require researchers to provide a
281 data availability statement when submitting a publication but allow researchers to
282 circumnavigate the provision by stating “the datasets and code are available from the
283 corresponding author on reasonable request.” Some publishers require authors to make their
284 publication data publicly available (21). As of January 25, 2023, National Institute of Health-

285 supported research requires researchers to include a plan for data sharing within their funding
286 applications. While our review included publications published on preprint servers, which have
287 less strict reporting guidelines, we reason that preprint COVID-19 computational models should
288 have been just as transparent with their data as peer-reviewed publications given their
289 widespread use by policymakers and news outlets during the start of the pandemic (12).

290 Although providing data access is becoming a common practice in infectious disease
291 computational studies, progress in sharing model implementations is lagging. In our sample,
292 most papers provided the model description; however, the code used to implement the model and
293 create the results was reported in less than 25% of the studies. In the previous review of COVID-
294 19 computational modeling studies, researchers found that a similar 21.5% of publications
295 reported the code ($n = 288$) (20). With increasingly complex computational methodologies in
296 infectious disease modeling literature, withholding the exact data manipulation and analysis steps
297 can impede the consistent regeneration of modeling results. Researchers should aim to provide
298 open-source access to appropriately versioned model implementations accompanied by
299 comprehensible annotations in online repositories.

300 Sharing a reproducible model consists of more than just sharing the data or code. Each
301 component in the checklist works together to produce the final modeling result. With each
302 additional missing component, the time and effort that it takes for future reproduction attempts
303 increases (22). Amid a pandemic, timely, reproducible research is critical in informing policies
304 and life-saving interventions.

305 The present study has several limitations. First, we sampled COVID-19 computational
306 modeling studies published early in the pandemic when authors may have reported fewer
307 reproducibility elements compared to publications published in later periods. In future work,

308 assessing the reproducibility of publications reported during various stages of the pandemic may
309 lead to insights regarding timing of publications and reproducibility of modeling literature.
310 Second, given that two of the reviewers had limited experience using the IDMRC prior to the
311 assessment, we may have underestimated the true reliability of the IDMRC. Furthermore,
312 differences in reviewer experience may have led to an under- or over-estimation of the average
313 number of reported reproducibility elements in our sample of publications. Finally, while the
314 reliability assessment of the IDMRC goes a step beyond most checklists, we did not assess the
315 reliability of the proposed changes to the IDMRC.

316 **Conclusions**

317 Our review focused on evaluating the performance of the IDMRC in COVID-19
318 computational modeling publications. Additional rounds of review with more reviewers and
319 modeling studies outside of COVID-19 might generalize reliability. Furthermore, lower inter-
320 rater reliability scores on some of the elements may have impacted the reported frequencies of
321 missing reproducibility elements. To address these issues, we proposed a revised version of the
322 IDMRC. Tools such as the IDMRC can encourage the documentation and sharing of all elements
323 necessary to reproduce a computational modeling study, thus supporting reproducible
324 computational infectious disease studies and accelerating scientific discoveries by allowing
325 others to validate results as well as by providing resources that might be reused in future studies.

326 **Declarations**

327 **Availability of data and materials**

328 All data generated or analyzed during this study are included in this published article and its
329 supplementary information files.

330 **Competing interests:** Authors declare that they have no competing interests. Dr. Van Panhuis
331 conducted the research while at the University of Pittsburgh, prior to starting his position at
332 NIAID.

333 **Funding:** This work was supported by the National Institute of General Medical Sciences
334 (NIGMS) grant U24GM132013.

335 **Authors' contributions**

336
337 Conceptualization: DP, BC, WVP; Methodology: DP, BC, WVP; Software: DP; Validation: DP;
338 Formal analysis: DP; Investigation: DP, BC, WVP, AAQ, MM, KC; Data curation: DP; Writing
339 – original draft: DP, BC, WVP; Writing – review & editing: DP, BC, WVP, HH, MH, AAQ,
340 MM, KC, MR; Visualization: DP, BC, WVP, HH; Supervision: BC, WVP, HH; Project
341 administration: WVP, HH; Funding acquisition: WVP

342 **Acknowledgements**

343 We thank current and past members of the Public Health Dynamics Lab, including and
344 Dr. Mark S. Roberts, Jessica Kerr, Lucie Contamin, Anne Cross, John Levander, Jeffrey Stazer,
345 Inngide Osirus, and Lizz Piccoli for critical discussions and feedback. We would also like to
346 thank Dr. Anne Newman for her feedback during the early development stages of this
347 manuscript.

348 **References**

- 349 1. Moore S, Hill EM, Tildesley MJ, Dyson L, Keeling MJ. Vaccination and non-
350 pharmaceutical interventions for COVID-19: a mathematical modelling study. *Lancet*
351 *Infect Dis.* 2021 Jun;21(6):793–802.
- 352 2. Michael Barton C, Alberti M, Ames D, Atkinson JA, Bales J, Burke E, et al. Call for
353 transparency of COVID-19 models. *Science (1979).* 2020;368(6490):482–3.

- 354 3. Pokutnaya D, Childers B, Arcury-Quandt AE, Hochheiser H, van Panhuis WG. An
355 implementation framework to improve the transparency and reproducibility of
356 computational models of infectious diseases. *PLoS Computational Biology* (forthcoming).
357 2023.
- 358 4. Plesser HE. Reproducibility vs. Replicability: A Brief History of a Confused
359 Terminology. *Front Neuroinform.* 2017;11:76.
- 360 5. Commission E. Goals of research and innovation policy. 2015.
- 361 6. Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data
362 stewardship: Fair enough? *European Journal of Human Genetics.* 2018;26(7):931–6.
- 363 7. U.S Government Accountability Office. Opportunities to Improve Coordination and
364 Ensure Reproducibility. 2020.
- 365 8. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. Vol. 163,
366 *American Journal of Epidemiology.* 2006. p. 783–9.
- 367 9. Peng RD. Reproducible research in computational science. Vol. 334, *Science.* American
368 Association for the Advancement of Science; 2011. p. 1226–7.
- 369 10. Cuschieri S. The STROBE guidelines. *Saudi J Anaesth.* 2019 Apr;13(Suppl 1):S31–4.
- 370 11. Pollett S, Johansson MA, Reich NG, Brett-Major D, del Valle SY, Venkatramanan S, et
371 al. Recommended reporting items for epidemic forecasting and prediction research: The
372 EPIFORGE 2020 guidelines. *PLoS Med.* 2021 Oct;18(10):e1003793.
- 373 12. Fraser N, Brierley L, Dey G, Polka JK, Pálffy M, Nanni F, et al. The evolving role of
374 preprints in the dissemination of COVID-19 research and their impact on the science
375 communication landscape. *PLoS Biol.* 2021 Apr;19(4):e3000959.

- 376 13. Nelson KP, Edwards D. Measures of agreement between many raters for ordinal
377 classifications. *Stat Med*. 2015 Oct 15;34(23):3116–32.
- 378 14. R Core Team. R: A language and environment for statistical computing. Vienna, Austria;
379 2020.
- 380 15. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal
381 data - which coefficients and confidence intervals are appropriate? *BMC Med Res*
382 *Methodol* [Internet]. 2016 Aug 5;16:93. Available from:
383 <https://pubmed.ncbi.nlm.nih.gov/27495131>
- 384 16. Gronenschild EHBM, Habets P, Jacobs HIL, Mengelers R, Rozendaal N, van Os J, et al.
385 The effects of FreeSurfer version, workstation type, and Macintosh operating system
386 version on anatomical volume and cortical thickness measurements. *PLoS One*. 2012 Jun
387 1;7(6).
- 388 17. Schnell S. “Reproducible” Research in Mathematical Sciences Requires Changes in our
389 Peer Review Culture and Modernization of our Current Publication Approach. *Bull Math*
390 *Biol*. 2018 Dec 1;80(12):3095–105.
- 391 18. Kenall A, Edmunds S, Goodman L, Bal L, Flintoft L, Shanahan DR, et al. Better reporting
392 for better research: A checklist for reproducibility. Vol. 4, *GigaScience*. BioMed Central
393 Ltd.; 2015.
- 394 19. Jalali MS, DiGennaro C, Sridhar D. Transparency assessment of COVID-19 models. Vol.
395 8, *The Lancet Global Health*. Elsevier Ltd; 2020. p. e1459–60.
- 396 20. Zavalis EA, Ioannidis JPA. A meta-epidemiological assessment of transparency indicators
397 of infectious disease models. *PLoS One*. 2022 Oct 1;17(10 October).

- 398 21. PLOS ONE. Data Availability [Internet]. 2018. Available from:
 399 <https://journals.plos.org/plosone/s/data-availability>
 400 22. Garijo D, Kinnings S, Xie L, Xie L, Zhang Y, Bourne PE, et al. Quantifying
 401 reproducibility in computational biology: The case of the tuberculosis drugome. PLoS
 402 One. 2013;8(11).
 403

404 **Table 1.** Infectious Disease Modeling Reproducibility Checklist elements reported in COVID-19
 405 modeling papers.

Question	Mean Percent Agreement (95% CI)	Fleiss Kappa (95% CI)
Computational Environment		
1.1) Is the operating system documented?	0.94 (0.87, 1.00)	0.90 (0.81, 0.97)
1.2) Is the operating system version documented?	0.94 (0.87, 1.00)	0.90 (0.79, 0.98)
Analytical Software		
2.1) Is the name of the analytical software documented (e.g., the programming language name)?	0.88 (0.79, 0.97)	0.82 (0.71, 0.92)
2.2) Is the analytical software accessible for free?	0.82 (0.71, 0.93)	0.68 (0.54, 0.81)
2.3) Is the version of the analytical software documented?	0.79 (0.67, 0.90)	0.75 (0.64, 0.85)
2.4) Do the authors include a specific identifier (DOI, URL, citation) that points to the analytical software that was used?	0.76 (0.64, 0.89)	0.72 (0.61, 0.81)
2.5) Is the analytical software installation guide accessible online?	0.89 (0.80, 0.98)	0.75 (0.60, 0.89)
Model Description		
3.1) Is the complete, structured model description provided in the publication, supplement, or referenced publication?	0.53 (0.38, 0.67)	0.58 (0.51, 0.66)
3.2) Is the model specified in the publication or supplement (contrary to being referenced in other papers)?	0.91 (0.83, 0.99)	0.84 (0.70, 0.94)
Model Implementation (“Code”)		
4.1) Is the model implementation (e.g., code, workflow) openly accessible online?	0.86 (0.75, 0.96)	0.86 (0.77, 0.94)
4.2) Does the model implementation (e.g., code, workflow) have a version or modification date?	0.84 (0.73, 0.94)	0.69 (0.55, 0.84)
4.3) Does the model implementation (e.g., code, workflow) have an identifier?	0.79 (0.67, 0.90)	0.74 (0.62, 0.85)
4.4) Is the computer language of the model implementation (e.g., code, workflow) documented?	0.62 (0.48, 0.76)	0.39 (0.22, 0.54)
4.5) Are all model implementation (e.g., code, workflow) dependencies clearly specified in either the publication or supplemental files?	0.69 (0.56, 0.83)	0.63 (0.50, 0.75)
4.6) Are the model implementations (e.g., code, workflow) annotated with comments?	0.94 (0.87, 1.00)	0.80 (0.66, 0.92)
Data		
5.1) Does the model in the publication use input data?	0.75 (0.62, 0.87)	0.59 (0.47, 0.70)
5.2) Has the source and content of the input data been described in the publication or supplement?	0.59 (0.45, 0.74)	0.36 (0.20, 0.52)
5.3) Does the paper cite a specific, unique, and persistent identifier to refer to	0.54 (0.39, 0.68)	0.36 (0.20, 0.52)

each input dataset?

5.4 Is the input data openly accessible?	0.66 (0.52, 0.80)	0.34 (0.16, 0.52)
5.5 Is the data in a format that can be easily re-formatted (or “parsable”) to meet the input specifications of the model implementation?	0.55 (0.41, 0.69)	0.23 (0.10, 0.40)

Experimental Protocol

6.1 Are all the mentioned parameter values for the model implementation (e.g., code, workflow) documented in a single location (e.g., table or list in the publication or supplement)?	0.65 (0.51, 0.79)	0.69 (0.60, 0.77)
6.2 Is there an explanation of how the described/mentioned categories (computational environment, analytical software, model implementation, and data) were used together to create the results (e.g., figures and/ or tables)?	0.50 (0.36, 0.64)	0.58 (0.52, 0.64)

406 *Abbreviations:* CI, confidence intervals