A rubric for assessing conformance to the Ten Rules for credible practice of modeling and simulation in healthcare

Alexandra Manchel[1,10], Ahmet Erdemir[2,10], Lealem Mulugeta[3,10], Joy P. Ku[4,10], Bruno V. Rego[5,10], Marc Horner[6,10], William W Lytton[7,8,10], Jerry G. Myers, Jr.[9,10], Rajanikanth Vadigepalli[1,10,*]

[1] Department of Pathology and Genomic Medicine, Thomas Jefferson University, Philadelphia, PA, USA

[2] Department of Biomedical Engineering, Lerner Research Institute, Cleveland Clinic, OH, USA

[3] InSilico Labs LLC, Houston, TX USA

[4] Department of Bioengineering, Stanford University, Stanford, CA, USA

[5] Department of Biological & Agricultural Engineering, Louisiana State University, Baton Rouge, LA, USA

[6] Ansys Inc. Evanston, IL, USA

[7] Downstate Health Sciences University, Brooklyn, NY, USA

[8] Kings County Hospital, Brooklyn, NY, USA

[9] NASA - John H Glenn Research Center, Cleveland, OH, USA

[10] Committee on Credible Practice of Modeling, & Simulation in Healthcare, Interagency Modeling and Analysis Group and Multiscale Modeling Consortium, Bethesda, MD USA

* Corresponding Author

Email: rajanikanth.vadigepalli@jefferson.edu

Abstract

The power of computational modeling and simulation (M&S) is realized when the results are credible, and the workflow generates evidence that supports credibility for the context of use. The Committee on Credible Practice of Modeling & Simulation in Healthcare was established to help address the need for processes and procedures to support the credible use of M&S in healthcare and biomedical research. Our community efforts have led to the Ten Rules (TR) for Credible Practice of M&S in life sciences and healthcare. This framework is an outcome of a multidisciplinary investigation from a wide range of stakeholders beginning in 2012. Here, we present a pragmatic rubric for assessing the conformance of an M&S activity to the TR. This rubric considers the ability of an M&S study to communicate how well the study conforms to the Ten Rules for credible practice and facilitate outreach to a wide range of stakeholders from context-specific M&S practitioners to policymakers. It uses an ordinal scale ranging from Insufficient (zero) to Comprehensive (four) that is applicable to each rule, providing a uniform approach for comparing assessments across different reviewers and different modeling studies. We used the rubric to evaluate the conformance of two computational modeling activities: 1. six viral disease (COVID-19) propagation models, and 2. a model of hepatic glycogenolysis with neural innervation and calcium signaling. These examples were used to evaluate the applicability of the rubric and illustrate rubric usage in real-world M&S scenarios including those that bridge scientific M&S with policymaking. The COVID-19 M&S studies were of particular interest because they needed to be quickly operationalized by government and private decision-makers early in the COVID-19 pandemic and were accessible as open-source tools. Our findings demonstrate that the TR rubric represents a systematic tool for assessing the conformance of an M&S activity to codified good practices and enhances the value of the TR for supporting real-world decision-making.

Introduction

The role of computational modeling and simulation (M&S) in healthcare research and clinical practice is expanding at a rapid pace. M&S approaches have been integral to the progress in biomedical sciences and are starting to enable *in silico* and systems medicine efforts [1,2]. Computational modeling is relatively new in clinical and biomedical settings, necessitating the standardization of M&S efforts. The addition of standardized practices increases the credibility of the practice of M&S in this area as it has done in other disciplines, such as engineering. This also increases the M&S usefulness and widespread adaptation. Multiple standards have been proposed, both in industry and government, for establishing and ensuring credibility of M&S practices in various engineering fields, including medical devices [3–7]. Similarly, multiple standards exist for systems biology applications, which have been reviewed in Tatka et al. 2023 [8], that aim to address conceptual information, nomenclature, data formats, and representations of biochemical systems, and intend to improve the communication and sharing of M&S components.

In order to promote this standardization process in the biomedical community beyond those working in systems biology, the IMAG/MSM Committee on Credible Practice of Modeling and Simulation in Healthcare developed the "Ten Rules for Credible Practice of Modeling and Simulation in Healthcare" [9]. This framework is an outcome of multidisciplinary input from a wide range of stakeholders [10,11]. These rules aim to establish a unified conceptual framework to design, implement, evaluate, and communicate the activities, products, and outcomes of M&S in the biomedical sciences and clinical care domain. In application, the unified framework enables outreach to the entire M&S user community, ranging from model developers to policy makers to clinicians and other non-M&S practitioners.

74   Recently, Tatka et al. [8] reviewed the existing standards for representing and

75   documenting systems biology models. Current standards are limited to agreed-upon modeling

76   formats as a means to share information; however, there is no widely utilized standard for

77   assessing credibility of the practice in this area. As Tatka et al. [8] noted in their review, standards

78   for model annotation must become more widely accepted such that interoperability, reusability,

79   comparability, and comprehension can be improved. Credible practice will also be enhanced

80   when the information needed for simulation and parameter estimation is explicitly defined and

81   stated. Lastly, reproducibility would not be possible without efficient dissemination of all artifacts

82   and proper documentation on an open-source repository platform. When an M&S study conforms

83   to credible modeling practice guidelines at a high level, there is outreach to a wide range of

84   stakeholders, thus with proper dissemination of documentation, one will largely be able to

85   independently reproduce the M&S results. The authors of Tatka et al. [8] note that there is a lack

86   of consensus on quantitative credibility scoring and that a system that addresses this area would

87   provide the community of practice with a metric for comparing the credibility of models and a guide

88   for the development of more credible models.

89

90   The Committee's Ten Rules for Credible Practice of M&S in Healthcare (Table 1)

91   establishes initial standards for systems modeling and beyond [9]. While every effort was made

92   to thoroughly describe and define the rules, it lacks a quantitative, rigorous, and repeatable metric.

93   A consistent application of the rules likely requires a complementary rubric for assessing

94   conformance to the rules and evaluating the credibility of the M&S practice. Such a rubric would

95   be used to assess and communicate various aspects of the Ten Rules (TR), including the validity,

96   level of detail, and overall "correctness" of the M&S practice.

97

98   **Table 1: The Committee's Ten Rules of credible practice of M&S in healthcare** [9]**.**

Rule	Description

1.	Define context clearly	Develop and document the subject, purpose, and intended use(s) of the model or simulation
2.	Use contextually appropriate data	Employ relevant and traceable information in the development or operation of a model or simulation
3.	Evaluate within context	Perform verification, validation, uncertainty quantification, and sensitivity analysis of the model or simulation with respect to the reality of interest and intended use(s) of the model or simulation
4.	List limitations explicitly	Provide restrictions, constraints, or qualifications for or on the use of the model or simulation for consideration by the users or customers of a model or simulation
5.	Use version control	Implement a system to trace the time history of modeling and simulation activities including delineation of each contributors' efforts
6.	Document appropriately	Maintain up-to-date informative records of all modeling and simulation activities, including simulation code, model mark-up, scope and intended use of modeling and simulation activities, as well as users' and developers' guides
7.	Disseminate broadly	Share all components of modeling and simulation activities, including simulation software, models, simulation scenarios and results
8.	Get independent reviews	Have the modeling and simulation activity reviewed by nonpartisan third-party users and developers
9.	Test competing implementations	Use contrasting modeling and simulation implementation strategies to check the conclusions of different strategies against each other
10.	Conform to standards	Adopt and promote generally applicable and discipline specific operating procedures, guidelines, and regulations accepted as best practices

99

100   In principle, a case can be made that each of the Ten Rules in Table 1 needs its own

101   assessment approach. For example, Rule 1 on defining the context of use can be assessed

102   according to factors that quantify the level of detail in the documentation of the M&S subject,

103   scope of the M&S purpose or results and intended use of the M&S results such as to support

104   clinical decision making, inform regulatory evidence, or to inform next research steps. Additionally,

105   Rule 3 on model evaluation requires extensive consideration of how the M&S activity and results

106   are verified and validated, as well as how the assessment is presented to support the intended

107   use. This rule is consistent with U.S. Food and Drug Administration (FDA) guidance and American

108   Society of Mechanical Engineers (ASME) standards for best practices in verification and

109   validation of medical devices [4,12]. However, such a customized, rule-specific assessment

110   approach might become overly complex and unwieldy for consistently evaluating conformance to

111   the Ten Rules.

112

113   To address this issue, the Committee formulated a rubric based on the ability of the M&S

114   to facilitate outreach to a wide range of stakeholders from context-specific M&S practitioners to

115   policymakers. Since there exist various levels of M&S expertise in the healthcare domain, the

116   need for direct and clear communication of M&S results is essential. The development of the Ten

117   Rules rubric facilitates such communication and understanding of computational modeling

118   implementation and simulation results between stakeholders. This includes, but is not limited to,

119   communication between the model developers, M&S practitioners, model end-users, as well as

120   clinicians, policy makers, and other decision makers who depend on the knowledge generated by

121   the M&S. Therefore, the purpose of the rubric is to evaluate and assess an M&S study's

122   conformance to credible practices outlined in the Ten Rules as it pertains to outreach capability.

123

124   Most recently, the global response to the COVID-19 pandemic highlights the need for a

125   systematic assessment of credible practice of M&S across this entire spectrum of stakeholders

126   [13–17]. The role of M&S in providing quantitative insight for COVID-19 spread in the general

127   population was called into question due to a failure to predict early (circa 2020) outbreak dynamics

128   [13]. Nonetheless, the model predictions strongly influenced decision makers due to the ability of

129   M&S practitioners to quickly generate results with a perceived to be high degree of precision

130   superior to available observational statistical analyses. In retrospect, although results were

131   computationally precise, they exhibited lower accuracy than initially anticipated. Model

132   transparency, which includes explicit documentation of model choices, assumptions, the steps in

133   the modeling process, and the expectations for the outputs, provides a reasonable defense

134   against the propagation of misinformation and misunderstanding, such as what occurred during

135   the pandemic [15-16]. In several instances during the pandemic, a model developed for

136   population level COVID-19 spread in a large geographical region was applied to a less

137   appropriately applicable region without significant tuning and modification to account for

138   population-specific demographic, clinical, epidemiological, and other influencing factors [18].

139   Such activities during the early part of the COVID-19 pandemic illustrate how the lack of

140   transparency and independent evaluation reduces the utility of models to inform critical decisions

141   [17]. Informing the user of the M&S results of model context and its intended use, such as through

142   the assessment of its conformance to the "Ten Rules in Healthcare" will greatly minimize the

143   negative impacts on model utility at all levels of application [9]. Such an assessment would not

144   just improve transparency but would enable communication of credible practice of M&S in a

145   comprehensive manner (Table 1; [9]).

146

147   The remainder of the manuscript is organized as follows: First, we describe our process

148   for developing and utilizing the rubric. We then present the rubric and an explanation of its

149   components. Next, we illustrate the application of the rubric in multiple use cases to evaluate its

150   utility in assessing the conformance to the Ten Rules in a consistent manner across multiple

151   reviewers and M&S studies. Finally, we discuss best practices for applying the rubric and possible

152   future extensions.

153

154   **Materials and Methods**

155   **Development of the rubric framework**

156

157   Our proposed rubric can be used to assess and communicate the extent of conformance

158   to the Ten Rules for Credible Practice of M&S based on the capability of outreach to the

159   biomedical and healthcare community (Table 2). The concept of using a rubric for communicating

160   the credibility state of an M&S evolved from challenges in communicating the ten simple

161   assessments at different decision-making levels and in different contextual applications. Thus,

162   the intent of the rubric development, is to bring a concise communication tool to the M&S

163   healthcare community.

164

165   To develop the rubric, the Committee considered requirements of outreach to a wide range

166   of stakeholders (Fig 1A), each of whom has their own distinct use cases and priorities in

167   evaluating an M&S model. For instance, M&S practitioners may want to conduct granular analysis

168   of their own M&S practices, while clinicians are primarily concerned with whether they can trust

169   M&S to inform a clinical practice decision.

170

171   **Table 2. rubric for assessing conformance to the Ten Rules.**

Outreach Capability	Outreach to application-domain experts who may not be M&S practitioners	Outreach to M&S practitioners who may not be application-domain experts	Outreach to application-domain specific M&S practitioners	Outreach to application-domain specific M&S practitioners	None or very limited
Conformance Level	Comprehensive	Extensive	Adequate	Partial	Insufficient
	4	3	2	1	0
Description Level	Can be understood by <u>non-M&S practitioners</u> familiar with the application	Can be understood by M&S practitioners <u>not</u> familiar	Can be understood by M&S practitioners familiar with	<u>Unclear</u> to the M&S practitioners familiar with the application	Missing or grossly incomplete information to properly evaluate the

	domain and the intended context of use	with the application domain and the intended context of use	application domain and the intended context of use	domain and the intended context of use	conformance with the rule

172

173
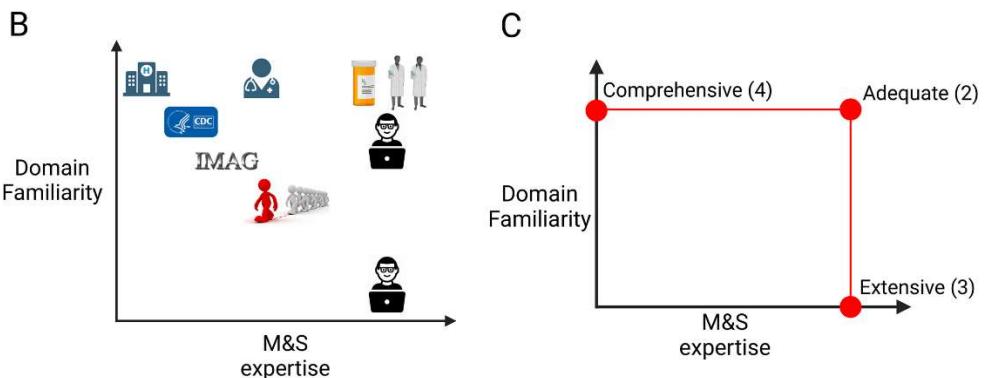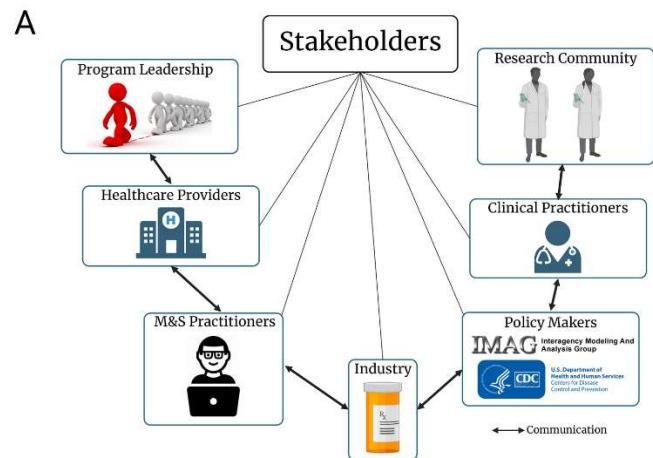
Fig 1. Assessing TR rubric conformance to the Ten Rules based on the extent of outreach to the stakeholder's domain expertise. (A) The range of stakeholders that may utilize the rubric. (B) Example distribution of the stakeholders in the stakeholder assessment chart. (C) The conformance levels to each of the Ten Rules based on the extent of outreach to stakeholders along the M&S expertise and domain familiarity axes.

179

180   The Committee developed the rubric framework through an iterative approach. The initial

181   framework had reviewers assess models qualitatively, ranking the conformance of a model to

182   each of the Ten Rules as insufficient, partial, adequate, extensive, or comprehensive (Table 2).

183   The qualitative assessment made it challenging to compare reviewer assessments and derive an

184   overall rating for the model when there existed variability between the individual reviewers'

185   assessments, as in the case of the first COVID-19 model (UPenn's COVID-19 model) to which

186   the Committee applied the rubric. Therefore, a second development of the Ten Rules rubric was

187   implemented. In this development, a scoring system was included such that for each rule, the

188   level of conformance is given a numerical score.

189

190   Application of the rubric for different use cases

191

192   We applied the Ten Rules and the rubric to evaluate the M&S practices of several COVID-

193   19 modeling studies with versions released early in the pandemic and available at the time of this

194   study: MIT model [19], IHME model [20,21], CU model [22], NE model [23], ICL model [24], UPenn

195   model [25]. Model details can be found in Table 3. Independent reviews by persons with significant

196   experience in M&S credibility assessment and with some familiarity of the application of M&S in

197   supporting government and medical industry decision making are used to assess each model.

198

199   **Table 3. COVID-19 models tested for their conformance to the Ten Rules.**

Location of Model Development	Model Description	Website	Dates Accessed	References

Massachusetts Institute of Technology (MIT)	A novel epidemiological model for predicting detected cases and deaths in the pre-vaccination era of the COVID-19 pandemic	https://www.covidanalytics.io/	June 2020 September 2020 February 2021	[19]
Institute for Health Metrics and Evaluation (IHME)	A model for predicting possible trajectories of COVID-19 infections and the effects of non-pharmaceutical interventions in the United States	https://www.healthdata.org/covid	June 2020 September 2020 February 2021	[20,21]
Columbia University (CU)	A model to infer critical epidemiological characteristics associated with COVID-19	https://columbia.maps.arcgis.com/apps/webappviewer/index.html?id=ade6ba85450c4325a12a5b9c09ba796c	June 2020 September 2020 February 2021	[22]
Northeastern University (NE)	A model used to study spatiotemporal COVID-19 spread	https://covid19.gleamproject.org/#model	June 2020 September 2020 February 2021	[23]
Imperial College London (ICL)	A model used to study the effect of non-pharmaceutical interventions in controlling the COVID-19 epidemic	https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/disease-areas/covid-19/covid-19-planning-tools/	June 2020 September 2020 February 2021	[24]

University of Pennsylvania (UPenn)	COVID-19 Hospital Impact Model for Epidemics (CHIME), which was designed to assist hospitals and public health officials with understanding hospital capacity needs during the pandemic	https://penn-chime.phl.io/	April 2020 June 2020 September 2020 February 2021	[25]

200

201   The selection process for the COVID-19 models we evaluated did not take into account

202   potential conformance to the Ten Rules, but rather addresses the availability of the model and

203   related information present at the time of the study. As none of the models explicitly followed our

204   recommended credibility practice, we did not seek to compare the COVID-19 models to determine

205   which one is the most conformant to the Ten Rules but rather to express our assessment based

206   on the information provided. Finally, this is not an endorsement or criticism of the M&S practices

207   utilized for the models tested for conformance to the Ten Rules. Instead, we seek to exemplify

208   how the rubric is to be employed when testing for M&S conformance and outreach. Additionally,

209   we note that no attempts were made to reproduce any of the results reported by these models,

210   nor was any attempt made to assess the scientific validity of the models, assumptions, or

211   limitations. Instead, we assessed the outreach capability of the models and were interested in

212   their representation and dissemination.

213

214   Two independent reviewers evaluated the conformance of the UPenn COVID-19 model

215   (details can be found in Table 3) to the Ten Rules in April 2020. They used the initial, qualitative

216   rubric. A separate independent reviewer (Reviewer 3) assessed the conformance of the remaining

217   COVID-19 models to the Ten Rules using the numerical scoring version of the rubric. Reviewer

218 3's original assessment was performed on June 24, 2020 and repeated twice: once on September

219 7, 2020 and again on February 5, 2021.

220

221   Finally, the rubric was applied in a clinically relevant M&S study to evaluate the utility of

222 our approach in assessing the extent of credibility of M&S practices in these contexts of use. In

223 the study, the authors developed a multi-scale, multi-organ model of hepatic metabolism. The

224 authors performed a self-assessment of their model's conformance to the Ten Rules prior to their

225 initial manuscript submission on this model and then reassessed their model during the

226 manuscript revision process.

227

228 RESULTS

229 A generalized rubric based on outreach capability

230

231   The Committee recognized that the rubric needed to account for the different stakeholders

232 who may be interested in utilizing a model. Assessment of a model's ability to communicate how

233 and if it satisfied each of the Ten Rules would differ greatly depending on the stakeholder.

234 Therefore, in the proposed rubric, the key stakeholder traits are distributed along two axes: their

235 level of M&S expertise and their familiarity with the biological domain (Fig 1B). The stakeholder

236 communities can have different mixtures of M&S expertise and domain familiarity. Individuals with

237 expertise in M&S and the biological domain relevant to the context of use are positioned towards

238 the upper right, while individuals with very little M&S expertise and domain knowledge are

239 positioned towards the lower left. The rubric assesses the conformance to each of the Ten Rules

240 based on the extent of outreach to each group (Fig 1C). For a given rule, if the M&S practice was

241 conducted at a level that is primarily accessible to only those with M&S expertise and domain

242 familiarity, we deem this practice to be conformant to the rule at the Adequate level. If the M&S

243   practice of a given rule is more broadly understood by individuals with M&S expertise without

244   familiarity of the specific biological domain, we deem this practice to be conformant to the rule at

245   the Extensive level. If the M&S practice of a given rule is understood by those familiar with the

246   biological domain but do not have M&S expertise, we deem this practice to be conformant to the

247   rule at the Comprehensive level. Lastly, the M&S practice that is unclear to the M&S practitioners

248   with familiarity of the biological domain is considered as a Partial level of conformance, with

249   missing information assessed as an Insufficient level.

250

251   In this rubric, the model with the highest conformance level (Comprehensive) provides

252   outreach to domain experts who may not be M&S practitioners while the lowest conformance level

253   (Insufficient) does not provide sufficient outreach to any community level. Taken together, this

254   rubric provides a generalized and graded approach to assess the conformance to the Ten Rules

255   (Fig 1C). Table 2 shows a concise representation of the proposed conformance rubric to the Ten

256   Rules of credible practice of M&S in healthcare. The extended rubric can be found in S1 File.

257

258   The rubric does not assess the "correctness" (i.e. the validity or accuracy) of the

259   computational models, but rather analyzes M&S credible practice conformance based largely on

260   two dimensions: M&S experience and scientific domain expertise (Fig 1). We note that the rubric

261   for conformance to the Ten Simples Rules is not an M&S practice accreditation process, but rather

262   a communication tool for analyzing the robustness of the M&S practice employed for a

263   computational model within specifically stated context of use.

264

265   **Assessment criteria for each rule**

266

267   Within each of the Ten Rules (i.e., guiding principles of M&S practice), we specified the

268   detailed criteria to assess the level of conformance and outreach capability to all stakeholders

269   across different application contexts (S1 Table). For instance, Rule #1 is to define the context of

270   use clearly by developing and documenting the application, purpose, and intended uses of the

271   model and simulation (Table 1). In our proposed rubric, an M&S practice conforms to this rule at

272   the highest level (Comprehensive) if:

273   1. a summary of the context definition can be understood by non-M&S practitioners,

274   2. detailed explanation is understandable by experts from the application domain that may

275   not be M&S practitioners, and

276   3. many relevant details are included in the documentation that enable adequate

277   understanding by both application domain-specific and non-domain M&S experts.

278

279   The next levels of conformance are based on whether the context definition was

280   communicated at a level that is a) understandable only by M&S experts, even if they are from

281   outside of the application domain (Extensive); b) restricted to M&S experts with experience in the

282   specific application domain (Adequate); or c) achieved only partially (Partial).

283

284   As another example, Rule #2 is to use contextually appropriate data by employing relevant

285   and traceable information in the development or operation of a model or simulation (Table 1). In

286   our proposed rubric, M&S practice conforms to this rule at the highest level (Comprehensive) if:

287   1. the data used in the M&S development is contextually appropriate,

288   2. all the data used in M&S development and/or operation is traceable to its original source,

289   and

290   3. application-domain experts that are not M&S practitioners can understand which and how

291   the data was used.

292

293   As a general guideline, the level of conformance of an M&S practice to each rule should

294   be assessed systematically by answering the following three questions:

295   1. Does the M&S practice employ the rule as defined by Erdemir et al. (2020)?

296   2. What is its outreach capability, or which types of stakeholders can the M&S practice

297   effectively support?

298   3. How easily can different stakeholders understand the extent to which the rule was applied

299   to the M&S practice?

300

301   It is also important to note that if the answer to the first question is "no" or "uncertain," the

302   conformance level of the M&S practice for that rule must be scored as 0 (Insufficient). For

303   example, in the case of Rule #2, if the data used in the M&S practice is entirely inappropriate, the

304   answer to the first question would be "no." As a result, the conformance score for Rule #2 would

305   be 0, regardless of how traceable or well-documented the data is in that M&S practice.

306

307   The detailed criteria corresponding to all of the Ten Rules can be found in S1 Table.

308

Numerical scoring used in the rubric

310

311   A numerical scoring system was included in the rubric to quantify the assessments with

312   each level of conformance. A conformance level of Insufficient is given a score of 0, while a

313   conformance level of Comprehensive is given a score of 4.

314

315   After the reviewer has completed their assessment of the model's conformance to the

316   rules, a total numeric score can be calculated, thereby allowing for a higher-level understanding

317   of the model's conformance and providing a means of easily comparing assessments between

reviewers. Assessment of a model which reaches an overall conformance level of Comprehensive will have a total score in the range [35, 40], while a model with Insufficient conformance will have a total score in the range [0, 5)) (Table 5). The total score is most informative at the high and low extremities, as the M&S study's overall performance can be easily evaluated

Table 5. Numerical scoring system for assessing conformance to the Ten Rules.

Conformance Level	Score for each Rule	Score Range for Averaging across Rule	Score Range for Summing all Ten Rules
Comprehensive	4	[3.5, 4]	[35, 40]
Extensive	3	[2.5, 3.5)	[25, 35)
Adequate	2	[1.5, 2.5)	[15, 25)
Partial	1	[0.5, 1.5)	[5, 15)
Insufficient	0	[0, 0.5)	[0, 5)

If a subset of the TR is not included in a model assessment, the overall scores associated with a given conformance level will need to be adjusted. For example, if two Rules are omitted in the assessment, a Comprehensive conforming model will then have a total score in the range [28, 32], rather than [35, 40]. Similarly, Extensive will have a score in the range [20, 28), Adequate will have a score in the range [12, 20), Partial will have a score in the range [4, 12), and Insufficient will have a score in the range [0, 4).

Numerical scoring also enables the calculation of statistics, such as averages and standard deviations, across multiple assessments for a single rule. When averaging scores across reviewers for a single rule, the score may not be a whole number, which we have accounted for in Table 5, which details the range of conformance scores for each rule.

Recommended process for implementing rubric

339   The recommended process for implementing the rubric throughout the M&S life cycle

340   begins with clearly identifying the M&S intended context of use, including M&S domain of use,

341   use capacity, and strength of influence (Fig 2) [9]. Next, the conformance threshold must be

342   established according to the rubric and TR. It is expected that throughout the M&S lifecycle there

343   is to be further development and refinement of the model, thereby necessitating evaluation of the

344   updated M&S per the Ten Rules and rubric thresholds. Following this assessment, there should

345   be clear documentation and then implementation of the M&S. Additionally, when implementing

346   the M&S activities, further reporting and documentation may be needed.
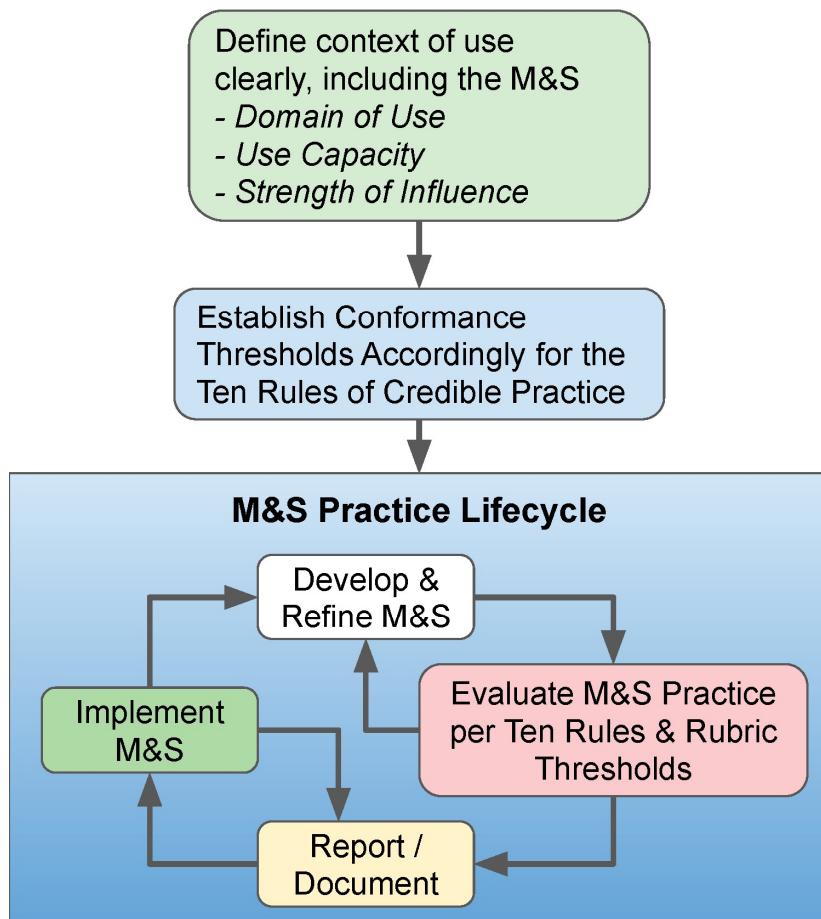
347

348

349   **Fig 2. Recommended process for implementing the TR rubric throughout the M&S**

350   **lifecycle.**

351

Illustrative application of the rubric to assess COVID-19 M&S practice

We applied the Ten Rules and the rubric to evaluate the M&S practices of several COVID-19 modeling studies, as described in the Methods. Table 4 illustrates a summary of our two independent reviewer processes to evaluate the conformance to the Ten Rules of the UPenn COVID-19 model in April 2020. The complete assessment and conformance testing made by Reviewer 1 can be found in S2 File, and by Reviewer 2 in S3 File. Briefly, Reviewer 1 and 2 disagreed on the conformance level of the UPenn COVID-19 model for five of the ten rules (Rule #1, #2, #4, #8, and #10). Despite this variability, the overall conformance of the model as tested using the numeric scoring system resulted in Reviewer 1's overall score of 21 and Reviewer 2's overall score of 20. Both reviewers agreed that the overall conformance of the model was Adequate in that the model can be understood by those with expertise in M&S and the biological domain. Thus, the scoring system facilitates comparisons between reviewers, assessments for each individual rule, and also a model's overall conformance. In the present rubric, the total assessment acts like an average, rather than a sum, of the individual rule assessments and is presented as representing an overall conformance using the same category scale as used for each individual rule.

Table 4. Independent reviewer assessment of the UPenn COVID-19 model's conformance to the Ten Rules.

Rule	Conformance Level		Same Conformance across Reviewers?
	Reviewer 1	Reviewer 2	
1. Define context clearly	Comprehensive	Adequate	No
2. Use contextually appropriate data	Adequate	Partial	No
3. Evaluate within context	Insufficient	Insufficient	Yes
4. List limitations explicitly	Adequate	Partial	No

5.	Use version control	Extensive	Extensive	Yes
6.	Document appropriately	Extensive	Extensive	Yes
7.	Disseminate broadly	Comprehensive	Comprehensive	Yes
8.	Get independent reviews	Insufficient	Partial	No
9.	Test competing implementations	Partial	Partial	Yes
10.	Conform to standards	Adequate	Comprehensive	No

373

374   A separate independent reviewer (Reviewer 3) assessed the conformance of the

375   remaining COVID-19 models to the Ten Rules. For each of these model assessments, Rule #3

376   (Evaluate within context) and Rule #9 (Testing competing implementations) were not examined

377   as these rules posed the greatest uncertainty and required in-depth knowledge of the model

378   structure and development. Here, for simplicity, we discuss the results from the most recent

379   assessment on February 5, 2021; however, extensive commentary from each assessment is

380   documented in S4 File. The most recent assessment of the MIT COVID-19 model's conformance

381   to the Ten Rules is exemplified in Table 6, and the reviewer's reasoning for each conformance

382   score is highlighted in the Comments column of the table. Despite the reviewer-assessed overall

383   model conformance level being Extensive, the total score was 18-19, which would point towards

384   an overall conformance level of Adequate. The scoring system introduced in the Methods section

385   provides reviewers with the ability to properly evaluate their complete assessment of model

386   conformance to the Ten Rules in a more quantitative manner. Reviewer 3 followed the same

387   protocol for assessing conformance of the remaining COVID-19 models to the Ten Rules. The

388   models' conformances to the Ten Rules can be found in Table 7 and the detailed assessments

389   can be found in S4 File. The overall conformance across the COVID-19 models assessed by

390   Reviewer 3 were between Adequate and Extensive. All of the COVID-19 models have shown

391   improvement according to the Ten Rules rubric following reassessment.

392

Table 6. Reviewer 3's assessment of the MIT COVID-19 model's conformance to the Ten Rules.

Rule	Conformance Level	Conformance Level Score	Comments
1. Define context clearly	Extensive	3	- A critical tool for COVID-19 planning that charts the progression of the pandemic across the United States and the world.
2. Use contextually appropriate data	Extensive	3	- Country-level projections are modeled based on historical data to increase the accuracy of future predictions. - Data is aggregated over 100 published clinical studies and preprints released between December 2019 and March 2020. - Declaration of collaborators and partner institutions that provide data and insights to model development team.
4. List limitations explicitly	Adequate	2	- Differences between Johns Hopkins University map and MIT cases reported - Total counts only account for countries in which they have sufficient data and where the pandemic is currently active. - Limitations explicitly stated in the paper
5. Use version control	Extensive	3	- Extensive documentation on differences between versions and updates. - Model codes are available on Github: https://github.com/COVIDAnalytics/DELPHI
6. Document appropriately	Extensive	3	- Model documentation contains the system of equations and rates. - Code is well documented and there are detailed instructions on how to reproduce the results.
7. Disseminate broadly	Extensive	3	- Model results regularly published with interactive graphics.

			- Results, data, models and simulations are openly available to the public and scientific community.
8. Get independent reviews	Insufficient/Partial	0-1	- Paper with scientific findings published in [19]
10. Conform to standards	Partial	1	- Codes are written in Python and Mathematica and data is provided in comma-separated variables (CSV) format.
Overall Conformance	**Extensive**	**18-19**	- The epidemiologically based model is disseminated broadly and provides links to model descriptions and data sources from their project website. - The code utilized in the research is accessible to the public via a GitHub repository and the model limitations are adequately described.

395

396 **Table 7. Reviewer 3 assessment of COVID-19 model conformances to the Ten Rules.**

Rule	Conformance Level (Score)				
	IHME Model	**CU Model**	**NU Model**	**ICL Model**	**Average Conformance across models**
1. Define context clearly	Extensive (3)	Extensive (3)	Extensive (3)	Adequate/ Extensive (2-3)	Extensive (2.75)
2. Use contextually appropriate data	Extensive (3)	Extensive (3)	Adequate/ Extensive (2-3)	Adequate/ Extensive (2-3)	Extensive (2.5)
4. List limitations explicitly	Partial (1)	Adequate (2)	Adequate (2)	Adequate (2)	Adequate (1.75)
5. Use version control	Adequate/ Extensive (2-3)	Adequate (2)	Insufficient (0)	Extensive (3)	Adequate (1.75)
6. Document appropriately	Adequate (2)	Adequate (2)	Extensive (3)	Extensive (3)	Extensive (2.5)
7. Disseminate broadly	Adequate/ Extensive (2-3)	Adequate/ Extensive (2-3)	Adequate/ Extensive (2-3)	Adequate/ Extensive (2-3)	Adequate (2)

8.	Get independent reviews	Adequate/ Extensive (2-3)	Extensive (3)	Adequate (2)	Extensive (3)	Extensive (2.5)
10.	Conform to standards	Partial (1)	Partial (1)	Insufficient (0)	Partial (1)	Partial (0.75)
Overall Conformance		**Adequate (16)**	**Adequate (18)**	**Adequate (14)**	**Adequate (18)**	**Adequate (16.5)**

397

398   Following Reviewer 3's assessment of the five COVID-19 model conformances to the Ten

399 Rules, we analyzed the results to identify the rules to which the models were least conformant. If

400 a model was between two conformance levels for a given rule, the lower conformance level was

401 used in the comparative analysis. The average numerical conformance score across all models

402 for each rule was assessed. Those with an average score lower than 1.5 meant that the

403 conformance level was at most Partial. The only rule that met this requirement was Rule #10:

404 Conform to standards. In all COVID-19 models assessed by Reviewer 3, the conformance levels

405 of the models to Rule #10 was either Partial or Insufficient, meaning the M&S practice of

406 conforming to standards was incompletely stated (Partial conformance score) or insufficient

407 evidence was presented to support conformance to standards (Insufficient conformance score)

408 as assessed by M&S practitioners familiar with the application domain and the intended context

409 of use. In order for the COVID-19 models to increase their conformance levels to Rule #10 of the

410 Ten Rules, the models should have been implemented using community standards and formats.

411 The associated documentation should lay out the details on the standards including version

412 numbers and any exceptions or deviations that influence the use of the model. For instance, the

413 IHME COVID-19 model is written in the widely used Python programming language; however, the

414 model has not yet been configured for use outside of the internal IHME infrastructure. The current

415 Partial conformance to Rule 10 can be increased to Adequate and possibly Extensive if IHME

416 provides sufficient evidence for following appropriate standards such as Python style guides and

417 statistical modeling standards (e.g., The American Statistical Association's Ethical Guidelines for

418 Statistical Practice).

419

Illustrative application of the rubric to M&S of liver metabolism

421

422   The rubric was applied to a clinically relevant M&S study by Verma et al. [26] in which the

423   authors developed a multi-scale, multi-organ model of hepatic metabolism. The authors

424   performed a self-assessment of their model's conformance to the Ten Rules prior to their initial

425   manuscript submission and then re-assessed their model during the manuscript revision process.

426   Below is a summary of the author's self-assessment included with the manuscript as a way to

427   illustrate the application of the rubric.

428

429   **Rule #1 (Define context clearly):** The authors provided a detailed description of the

430   model's context written using terminology familiar to non-M&S practitioners who are

431   knowledgeable about the application domain, so the authors described the model's conformance

432   to Rule #1 (Define context clearly) as Comprehensive (score = 4). Briefly, the authors described

433   that the primary goal of the model was to evaluate the role of neural signals in controlling the

434   metabolic functionality of the liver, particularly in regulating the glycogenolysis to maintain

435   appropriate responses to hormonal signals to meet the systemic glucose demands. The biological

436   domain, structures, spatial scales, and time scales are explicitly stated. Additionally, the authors

437   included an explanation of other uses for the model, which include exploration of the effect of

438   dietary intake and insulin resistance in promoting a hepatic steatosis-like phenotype in the context

439   of innervation, calcium signaling and central nervous system (CNS) activation.

440

441   **Rule #2 (Use contextually appropriate data):** The authors believed their model

442   conformed to an Extensive (score = 3) level for Rule #2 since the in vitro, ex vivo, in vivo preclinical

443   and human subject data used for model building and validation was confirmed to meet the detailed

444   data requirements for consistency and explicit description of data heritage.

445

446   **Rule #3 (Evaluate within context):** The authors' self-assessed conformance level was

447   Extensive (score = 3) since verification and validation of the model output was explicitly described

448   and performed by both the developer and a third-party lab member not involved in the study.

449   Additionally, the authors state that during the revision process, there was extensive validation

450   performed as the model was recalibrated based on experimental hepatic calcium dynamics and

451   catecholamine secretion in humans during periods of increased exercise.

452

453   **Rule #4 (List limitations explicitly**): The model's conformance was considered to be

454   Comprehensive (score = 4) as all limitations were explicitly provided. In addition, the limitations

455   were detailed in a manner that is understandable by experts in the liver physiology and pathology

456   domain, even if they are not M&S experts. An example limitation was that the model was

457   parameterized the same for simulating human-like and rodent-like extents of innervation and only

458   differed by the extent of innervation, which does not use M&S terminology but states the issue in

459   biomedical terms. Note that in the study Verma et al. [26] did not explicitly state the quantitative

460   levels of M&S prediction error arising from the explicitly stated limitations. Under the rubric, those

461   details are not required. There just needs to be sufficient information for an individual to

462   understand under which conditions a model should not be used.

463

464   **Rule #5 (Use version control):** The model's conformance was considered to be

465   Extensive (score = 3) as the evolution of the model and the various versions are explicitly

466   documented on GitHub. GitHub is a platform familiar to M&S practitioners but not necessarily to

467   experts in the liver physiology and pathology domain. Hence, the conformance level was not

468   considered Comprehensive (score = 4).

469

Rule #6 (Document appropriately): The model's conformance level is Extensive (score = 3) as comments were provided in the model code, the scope and intended use were described in the main text, and a user guide for M&S practitioners was described in the main text and supplemental files. During the revision process, the model alternative was explained in the main text and an additional supplemental figure was included to detail the results of the model alternative. The user guide was utilized by the independent reviewer (see Rule 8 below) with M&S expertise but little domain familiarity, demonstrating the Extensive level of conformance to Rule 6.

478

Rule #7 (Disseminate broadly): The conformance level was considered as Extensive (score = 3) as the simulations, results and implications were described in the main text and the model files are present in the supplementary material and on GitHub. The software used for this M&S study (Matlab, XPP and a Matlab/XPP interface) are all publicly available either freely or for a fee. The links to these resources and code files were included in the manuscript, enabling potentially Extensive dissemination.

485

Rule #8 (Get independent reviews): The self-assessed model conformance was Extensive (score = 3), as a member of the research group not involved in the study or field performed a review. We note that in order to minimize the bias in the assessment, an internal review, even by a member of the group not involved in the study, is more appropriately scored as a 2 (Adequate). An outside review (outside the primary research groups that conducted the study or even outside the study authors' institutions) could be considered as a 3 (Extensive), and a multi-person independent cross-institutional review, particularly by non-M&S practitioners, could be scored at 4 (Comprehensive).

494

Rule #9 (Test competing implementations): The conformance level only reached a conformance level of Adequate (score = 2) as competing implementations were tested and compared by the first three authors of the paper during the initial manuscript preparation. Furthermore, the competing implementations could only be understood by M&S practitioners familiar with the application domain and the intended context of use, thus justifying the Adequate conformance level. During the manuscript revision stage, the model was further revised, leading to its refinement and improvement. The main text was also updated to include the changes made to the model during revision.

Rule #10 (Conform to standards): The model's conformance was considered Adequate (score = 2) as the model was implemented and simulated in a widely used platform for multiscale modeling (Matlab, in this case), along with another freely available and popular software, XPP. Following best coding practices described in Wilson et al. [27] and Hunter-Zinck et al. [28], the model code is commented at critical locations to aid the reader as well. Although the model was documented and disseminated using publicly available online platforms such as GitHub and open access manuscript supplementary material in conformance with rule #7, there was limited evidence of following the operating procedures, guidelines and standards as described in the credible practice of M&S in healthcare: ten rules from a multidisciplinary perspective [9].

The complete self-assessment for this model is included as a supplement to this manuscript (S5 File).

The computational modeling and simulation study of hepatic metabolism has an overall conformance level of Extensive (total numeric score = 30). Therefore, the overall practice of M&S for this biological scenario can be understood by M&S practitioners not familiar with the application domain and intended context of use. For this example, M&S practice to reach a

521 Comprehensive level of conformance to the TR, the authors would need to incorporate additional

522 features into the study. For example, a detailed step-by-step user's and developer's guide such

523 that a non-M&S practitioner can replicate the M&S results would improve the score corresponding

524 to Rule #6. Additionally, the authors could follow a stricter set of operating procedures and

525 guidelines such that the M&S study appropriately conforms to modeling standards in

526 representation, software code and documentation (Rule #10). Lastly, the authors could more

527 comprehensively test and formally document competing implementations of their model for

528 improving the score on Rule #9.

529

530 # Discussion

531

532 We have described a rubric that specifies detailed criteria for assessing the level of

533 conformance to the Ten Rules for Credible Practice of M&S in Healthcare. The rubric is based on

534 the outreach capability of an M&S practice across a wide range of stakeholder communities

535 including program leadership, healthcare providers, policy makers and clinical practitioners. The

536 rubric establishes a generalized and graded approach to assess levels of conformance from

537 lowest (Insufficient) to highest (Comprehensive). We have illustrated the application of this rubric

538 in two contexts of use including COVID-19 studies and a liver metabolism model. In the context

539 of assessing COVID-19 studies, we evaluated the consistency of applying the rubric across

540 multiple reviewers. We proposed a scoring scheme that provides a consistent process for M&S

541 assessments and identification of critical credibility conformance gaps across a range of

542 reviewers' familiarity levels. The Ten Rules augmented with the rubric aims to provide a

543 generalized approach for the development and evaluation of the credible practice of M&S in

544 translational and fundamental research endeavors aimed at *in silico* support of systems medicine

545 efforts.

546

547   Assessing the outreach capability of an M&S study is useful for those within and outside

548 of a specific scientific discipline. It enables clear communication and application across various

549 stakeholder groups. For example, through the use of the Ten Rules and TR rubric, those working

550 in an industrial setting can easily understand and implement the M&S practices undertaken by

551 the academic research community. Additionally, these parties can communicate to policy makers

552 and higher-level stakeholders that can take action and employ a new development of the Ten

553 Rules and TR rubric to suit their needs. The continuous evolution of the rubric as seen with the

554 implementation of a numerical scoring system for conformance illustrates a framework that is

555 driven by refinement and improvement by the healthcare community.

556

557   The TR rubric was introduced to expand the reviewer's utilization of the Ten Rules.

558 Specifically, the introduction of the rubric concept is aimed at expanding the focus of the reviewer

559 from solely evaluating a model based on its validity and accuracy, to including the assessments

560 of how supporting information regarding the M&S credibility engages the community beyond those

561 who are familiar with M&S and the context of use. It is important to note, however, that not every

562 M&S needs to meet a score of Comprehensive to be acceptable. For example, for a

563 Comprehensive conformance level, the outreach is to non-M&S practitioners familiar with the

564 application, while a conformance level of Extensive can be understood by M&S practitioners not

565 familiar with the domain and context of use. Therefore, depending on the use of the model, an

566 Extensive conformance level may be more appropriate than a Comprehensive conformance level.

567

568   Assessment of the five discussed COVID-19 model conformances to the Ten Rules shows

569 the value of utilizing such a rubric that prioritizes outreach capability. Specifically, it shows the

570 Ten Rules and TR rubric can establish a cumulative assessment of the TR that has improved

571 consistency in evaluation at each competency level, which was a critical need for decision making

572   support as illustrated by the application to COVID-19 models. Multiple reviewers assessed the

573   conformance of the UPenn COVID-19 model to the Ten Rules. There were notable differences in

574   how the reviewers viewed the supporting credibility evidence, which illustrated that the reviewer's

575   experience level, and their understanding of the context of use relative to the models' intended

576   use, can influence the evaluation. This influence appeared to be nearly orthogonal to the

577   underlying credibility factor domains, leading to the investigation into a more granular and defined

578   TR rubric.

579

580   Following the updated rubric application, the consistency of findings between reviewers of

581   similar backgrounds was improved, especially if we consider the consolidated or summed

582   conformance scoring where both reviewers' scores correspond to an overall model conformance

583   level of Adequate.

584

585   The assumption that each rule's contribution is equally weighted with respect to the global

586   conformance introduces a limitation in the assessment scheme. For instance, it is possible to

587   accumulate an overall score in the Adequate or Extensive range and still have conformance to

588   one or more individual TR be characterized as Insufficient. This suggests a comprehensive

589   reporting that is more representative of the individual scores may be necessary to communicate

590   the complete M&S credibility outreach picture. An option is to use tailored decision ranking tools

591   such as pairwise comparison and analytical hierarchical processes (AHP) to capture specific

592   community best practice principles by effectively weighting the individual credibility rules.

593   Although the pairwise and AHP approaches may provide domain specific consistency, it is a

594   recommended best practice to provide the set of conformance scores for individual rules as well

595   as the global conformance score when delivering these assessments to decision makers in order

596   to ensure appropriate communications levels. In this case, the rubric assessed 10 rules that can

597   be grouped into representations tailored for the technical or decision-making community. A

598   proposed method is illustrated in Table 8 representing a grouping of the Ten Rules to derive

599   categorical scores for Development, Application and Supporting Evidence aspects for use in

600   regulatory applications.

601

602   **Table 8. Example of categorizing the TR and rubric assessments to support regulatory**
603   **applications according to Reviewer 1's assessment of the UPenn COVID-19 model's**
604   **conformance to the Ten Rules.**

Development		Application		Supporting Evidence		Overall
Rule 1	4	Rule 2	2	Rule 6	3	
Rule 3	0	Rule 4	2	Rule 7	4	
Rule 5	3			Rule 8	0	Summary Score 2.1
Rule 10	2			Rule 9	1	
Development Score 2.25		Application Score 2		Supporting Evidence Score 2		

605

606   The proposed numerical score has its limitations, i.e., a moderate score does not directly

607   inform the intended audience of the individual conformance levels for each of the TR. However,

608   at the high and low extremities, one can readily determine exceptional vs. poor M&S conformance

609   to credible modeling practices using the proposed summative scoring method. A more granular

610   numerical scoring approach can help develop a deeper understanding of the M&S study's

611   conformance level when the summative total numerical score across the TR is in the moderate

612   range. Our alternative approach for a categorized rubric with combined numerical scores for M&S

613   Development (including Rules 1, 3, 5, and 10), M&S Application (Rules 2 and 4), and M&S

614   Supporting Evidence (Rules 6, 7, 8, and 9) strikes a balance. Another alternative is to develop

615   visual representations, for example using radar/spider plots, that can illustrate the multiple levels

616   of conformance across the TR without combining into a single numerical score.

617

618   There is additional need and opportunity for streamlining the assessment of M&S activities

619   using the Ten Rules and TR rubric in addition to other associated frameworks. For instance, it

620   may take a significant amount of time to perform the assessment manually. Therefore, automating

621   components of the assessment may provide a capability of assessing the M&S results and

622   associated literature in an unbiased manner. This would be a boon to many communities of

623   practice, especially the healthcare community. A more systematic approach could be taken such

624   that the wider scientific community and stakeholders of the Ten Rules and TR rubric can be

625   included. The Interagency Modeling and Analysis Group (IMAG) and the Multiscale Modeling

626   (MSM) consortium are examples of two groups with significant roles in formulating and developing

627   the Ten Rules and TR rubric. As they both serve a joint purpose of serving the scientific

628   community and providing a collaborative platform to outline good practice of simulation-based

629   medicine, it may be possible to look to their leadership and guidance in systematizing and

630   automating unbiased assessment processes [10].

631

632   While a community effort is valuable to progress and implement the ideologies of the Ten

633   Rules and TR rubric, a specific set of guidelines must be established to ensure proper

634   employment. An excellent example of successful first steps in this direction lies with The

635   Physiome scientific journal. The Physiome is an open access journal that, for a small fee, confirms

636   the reproducibility and reusability of the models according to the Ten Rules. By adopting the Ten

637   Rules and TR rubric for M&S credibility, journal curators established that published models

638   generally only conform to an Adequate level of outreach. Implementing an additional guideline in

639   which the model must meet an overall conformance level of Extensive for publication into the

640   journal may promote the benefits of M&S outreach capability to the scientific community.

641

642   Future implementations of the Ten Rules and TR rubric could consider how the credible

643   practice for assessment of an M&S practice may be transferable from one context to another. For

644   example, another context of use was noted for the model of liver metabolism. However, the

645   reviewers did not assess the model in this alternate context. It is an open question as to how the

646   assessments of the Ten Rules can be applied to the alternate contexts of use and under what

647   conditions this can occur. It may be the case where the previous assessments of only some of

648   the Rules can be transferred while others may be "non-transferable". Additionally, M&S practices

649   may be altered at different user levels. For instance, it may not be appropriate to use a model

650   built on data from a local hospital system and apply the model at the national scale. Furthermore,

651   the transferability issue has implications for assessing the conformance of ensemble models, or

652   a single model that contains multiple diverse models, to the Ten Rules. One potential solution

653   may be to provide reasoning for including each of the models into the greater ensemble model in

654   the same way that a single equation within an ordinary differential equation (ODE)-based model

655   would be explained.

656

657   In an additional future implementation, we propose that the current rules and rubric can

658   be adjusted to more explicitly account for patient-specific/digital twin models as they begin to be

659   utilized in the clinical setting. An updated and extended set of rules and practices may be

660   developed for assessing and ensuring the credibility of these models. The need for an updated

661   list of rules is essential in the personalized modeling realm as the current methods lack

662   consistency and credibility, especially within the clinic. Additionally, the Ten Rules may not be

663   adequate in assessing the complexity and detail required for digital twin modeling. The updated

664   and extended rules for digital twin modeling can then be used as a guide during the developmental

665   stages of model development to avoid the problems seen with current digital twin efforts as

666   previously discussed. Future modeling efforts that are guided by future implementations of the

667   rules may also establish more trust and interaction between the modeler and clinician, thereby

668   bridging the gap that currently exists in translating computational models from research into the

669   healthcare field.

670

671   The TR rubric is used to assess a model's conformance to the Ten Rules for credible

672   practice in M&S in healthcare. It is highly recommended that the M&S activity in the healthcare

673   domain reaches either a conformance level of Comprehensive or Extensive. Both conformance

674   levels have their own intended outreach capability as Comprehensive models can be understood

675   by **non**-M&S practitioners familiar with the application domain and the intended context of use

676   while Extensive models can be understood by M&S practitioners **not** familiar with the application

677   domain and the intended context of use. Thus, defining which group must use the M&S results to

678   support their decision is of utmost importance. The outreach goal for a given model is to be as

679   clear and comprehensible to as broad an audience as possible such that the model can be widely

680   adopted.

681

682   In conclusion, we formulated a rubric that promotes consistent and continuous evolution

683   and testing of M&S practices such that one can reach the appropriate outreach level. In addition

684   to the evolution of individual models, the TR rubric may evolve to meet the needs of its users as

685   one continues to test its conformance to the Ten Rules. The development of the TR rubric has

686   initiated a large community effort to assess the outreach, reproducibility, replicability, and

687   credibility of M&S studies in the scientific healthcare domain.

688

689 # Acknowledgments

References

1. VPH Institute | Virtual Physiological Human - International non-profit organisation [Internet]. [cited 2023 Feb 14]. Available from: https://www.vph-institute.org/

2. European Virtual Human Twin [Internet]. [cited 2023 Apr 21]. Available from: https://www.edith-csa.eu/

3. National Aeronautics and Space Administration. Standard for Models and Simulations. NASA. 2024;Report No.: NASA-STD-7009.

4. V&V40 A. Assessing credibility of computational modeling through verification and validation: application to medical devices. The American Society of Mechanical Engineers (ASME). 2018;

5. National Research Council of National Academies. Assessing the Reliability of Complex Models. National Research Council. 2012;

6. Pathmanathan P, Aycock K, Badal A, Bighamian R, Bodner J, Craven BA, et al. Credibility assessment of in silico clinical trials for medical devices. PLoS Comput Biol. 2024 Aug 8;20(8):e1012289.

7. FDA. Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions. 2023;

8. Tatka LT, Smith LP, Hellerstein JL, Sauro HM. Adapting modeling and simulation credibility standards to computational systems biology. J Transl Med. 2023 Jul 26;21(1):501.

9. Erdemir A, Mulugeta L, Ku JP, Drach A, Horner M, Morrison TM, et al. Credible practice of modeling and simulation in healthcare: ten rules from a multidisciplinary perspective. J Transl Med. 2020 Sep 29;18(1):369.

10. Mulugeta L, Erdemir A. Committee on credible practice of modeling and simulation in healthcare. ASME 2013 Conference on Frontiers in Medical Devices: Applications of Computer Modeling and Simulation. ASME; 2013. p. V001T10A015.

11. Erdemir A, Guess TM, Halloran JP, Modenese L, Reinbolt JA, Thelen DG, et al. Commentary on the integration of model sharing and reproducibility analysis to scholarly publishing workflow in computational biomechanics. IEEE Trans Biomed Eng. 2016 Oct;63(10):2080–5.

12. Strategic Plan for Regulatory Science | FDA [Internet]. [cited 2023 Oct 16]. Available from: https://cacmap.fda.gov/science-research/advancing-regulatory-science/strategic-plan-regulatory-science

13. Kuhl E. Data-driven modeling of COVID-19-Lessons learned. Extreme Mechanics Letters. 2020 Oct;40:100921.

14. Panovska-Griffiths J. Can mathematical modelling solve the current Covid-19 crisis? BMC Public Health. 2020 Apr 24;20(1):551.

732 15. Barton CM, Alberti M, Ames D, Atkinson J-A, Bales J, Burke E, et al. Call for
733   transparency of COVID-19 models. Science. 2020 May 1;368(6490):482–3.

734 16. Bunker D. Who do you trust? The digital destruction of shared situational awareness and
735   the COVID-19 infodemic. Int J Inf Manage. 2020 Dec;55:102201.

736 17. Reps JM, Kim C, Williams RD, Markus AF, Yang C, Duarte-Salles T, et al.
737   Implementation of the COVID-19 Vulnerability Index Across an International Network of
738   Health Care Data Sets: Collaborative External Validation Study. JMIR Med Inform. 2021
739   Apr 5;9(4):e21547.

740 18. Newton AJH, Chartash D, Kleinstein SH, McDougal RA. A pipeline for the retrieval and
741   extraction of domain-specific information with application to COVID-19 immune
742   signatures. BMC Bioinformatics. 2023 Jul 20;24(1):292.

743 19. Li ML, Bouardi HT, Lami OS, Trikalinos TA, Trichakis N, Bertsimas D. Forecasting
744   COVID-19 and Analyzing the Effect of Government Interventions. Oper Res. 2022 Jun
745   10;

746 20. IHME COVID-19 health service utilization forecasting team, Murray CJ. Forecasting the
747   impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for
748   the USA and European Economic Area countries. medRxiv. 2020 Apr 26;

749 21. IHME COVID-19 Forecasting Team. Modeling COVID-19 scenarios for the United
750   States. Nat Med. 2021 Jan;27(1):94–105.

751 22. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented
752   infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science.
753   2020 May 1;368(6490):489–93.

754 23. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of
755   travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak.
756   Science. 2020 Apr 24;368(6489):395–400.

757 24. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating
758   the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020
759   Aug;584(7820):257–61.

760 25. Weissman GE, Crane-Droesch A, Chivers C, Luong T, Hanish A, Levy MZ, et al. Locally
761   Informed Simulation to Predict Hospital Capacity Needs During the COVID-19 Pandemic.
762   Ann Intern Med. 2020 Jul 7;173(1):21–8.

763 26. Verma A, Manchel A, Narayanan R, Hoek JB, Ogunnaike BA, Vadigepalli R. A spatial
764   model of hepatic calcium signaling and glucose metabolism under autonomic control
765   reveals functional consequences of varying liver innervation patterns across species.
766   Front Physiol. 2021 Nov 26;12:748962.

767 27. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, et al. Best practices
768   for scientific computing. PLoS Biol. 2014 Jan 7;12(1):e1001745.

769 28. Hunter-Zinck H, de Siqueira AF, Vásquez VN, Barnes R, Martinez CC. Ten simple rules
770   on writing clean and reliable open-source scientific software. PLoS Comput Biol. 2021
771   Nov 11;17(11):e1009481.

Supporting Information

773   **S1 File.** The extended Ten Simple Rules rubric with detailed criteria.

774   **S2 File.** Reviewer 1's complete assessment and conformance testing to the Ten Simple Rules

775   using the rubric applied to the COVID-19 modeling studies.

776   **S3 File.** Reviewer 2's complete assessment and conformance testing to the Ten Simple Rules

777   using the rubric applied to the COVID-19 modeling studies.

778   **S4 File.** Reviewer 3's complete assessment and conformance testing to the Ten Simple Rules

779   using the rubric applied to the COVID-19 modeling studies.

780   **S5 File.** The complete self-assessment and conformance testing to the Ten Simple Rules using

781   the rubric applied to the Verma et al. (26) multi-scale, multi-organ model of hepatic metabolism.