

1 Diversity and inclusion: A hidden additional benefit of Open Data

2 **Authors**

3 Marie-Laure Charpignon^{1,2}, Leo Anthony Celi^{3,4,5}, Marisa Cobanaj⁶, Rene Eber⁷, Amelia Fiske⁸,
4 Jack Gallifant^{3,9}, Chenyu Li¹⁰, Gurucharan Lingamallu¹¹, Anton Petushkov¹², Robin Pierce¹³

5

6 **Affiliations**

7 1. Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge,
8 MA 02139.

9 2. Broad Institute of MIT and Harvard, Cambridge, MA 02139.

10 3. Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge,
11 MA 02139.

12 4. Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical
13 Center, Boston, MA 02215.

14 5. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115.

15 6. National Center for Radiation Research in Oncology, OncoRay, Helmholtz-Zentrum Dresden-
16 Rossendorf, Dresden, Germany.

17 7. Montpellier Research in Management, Montpellier University, France.

18 8. Institute of History and Ethics in Medicine, Department of Clinical Medicine, TUM School of
19 Medicine and Health, Technical University of Munich, Munich, Germany.

20 9. Department of Critical Care, Guy's and St Thomas' NHS Trust, London, UK.

21 10. University of Pittsburgh School of Medicine Department of Biomedical Informatics, 5607 Baum
22 Blvd, Pittsburgh, PA, US.

23 11. University of Michigan, Ann Arbor, USA.

24 12. University of Washington, Seattle, USA.

25 13. University of Exeter, Exeter, UK.

26

27 **Funding Statement**

28 LAC is funded by the National Institute of Health through R01 EB017205, DS-I Africa U54
29 TW012043-01 and Bridge2AI OT2OD032701, and the National Science Foundation through
30 ITEST #2148451. JG is funded by the National Institute of Health through R01 EB017205, DS-I
31 Africa U54 TW012043-01 and Bridge2AI OT2OD032701.

32

33

34 **Abstract**

35 The recent imperative by the National Institutes of Health to share scientific data publicly
36 underscores a significant shift in academic research. Effective as of January 2023, it emphasizes
37 that transparency in data collection and dedicated efforts towards data sharing are prerequisites
38 for translational research, from the lab to the bedside. Given the role of data access in mitigating
39 potential bias in clinical models, we hypothesize that researchers who leverage open-access
40 datasets rather than privately-owned ones are more diverse. In this brief report, we proposed to
41 test this hypothesis in the transdisciplinary and expanding field of artificial intelligence (AI) for
42 critical care.

43 Specifically, we compared the diversity among authors of publications leveraging open datasets,
44 such as the commonly used MIMIC and eICU databases, with that among authors of publications
45 relying exclusively on private datasets, unavailable to other research investigators (e.g., electronic
46 health records from ICU patients accessible only to Mayo Clinic analysts). To measure the extent
47 of author diversity, we characterized gender balance as well as the presence of researchers from
48 low- and middle-income countries (LMIC) and minority-serving institutions (MSI).

49 Our comparative analysis revealed a greater contribution of authors from LMICs and MSIs among
50 researchers leveraging open critical care datasets than among those relying exclusively on private
51 data resources. The participation of women was similar between the two groups, albeit slightly
52 larger in the former. Notably, although over 70% of all articles included at least one author inferred
53 to be a woman, less than 25% had a woman as a first or last author. Importantly, we found that
54 the proportion of authors from LMICs was substantially higher in the treatment than in the control
55 group (10.1% vs. 6.2%, $p < 0.001$), including as first and last authors. Moreover, we found that the
56 proportion of US-based authors affiliated with a MSI was 1.5 times higher among articles in the
57 treatment than in the control group, suggesting that open data resources attract a larger pool of
58 participants from minority groups (8.6% vs. 5.6%, $p < 0.001$).

59 Thus, our study highlights the valuable contribution of the Open Data strategy to
60 underrepresented groups, while also quantifying persisting gender gaps in academic and clinical
61 research at the intersection of computer science and healthcare. In doing so, we hope our work
62 points to the importance of extending open data practices in deliberate and systematic ways.

63

64 **Author Summary**

65 In light of the significance of data access to the mitigation of bias in clinical models, we
66 hypothesize that researchers who leverage existing open-access datasets rather than privately-
67 owned ones are more diverse. In this brief report, we propose to test this hypothesis in the
68 transdisciplinary and expanding field of artificial intelligence for critical care. Specifically, we
69 compare the diversity among authors of publications leveraging open datasets, such as the
70 commonly used MIMIC and eICU databases, with that among authors of publications relying
71 exclusively on private datasets, unavailable to other research investigators. To measure the
72 extent of author diversity, we characterize gender balance, geographic diversity (i.e., the number
73 of countries with which authors are affiliated and the income categories these countries map to),
74 and the presence of researchers from minority-serving institutions located in the United States.
75 Furthermore, we comment on the challenges of increasing the participation of researchers from
76 underrepresented groups and suggest changes that can be made to the current Open Data
77 strategy to enhance representation in authorship in the next decade. By evaluating the association
78 between data accessibility and author diversity, our study pinpoints actionable steps that the
79 broader field of clinical AI can take to foster inclusion in the scientific community and mitigate blind
80 spots in data preparation and/or model development.

81

82

83 **Introduction**

84 The rapidly expanding field of health data science integrates two established disciplines:
85 computer science and healthcare. It promises to address the growing complexity of healthcare
86 systems arising from (a) the multiplicity of care delivery settings (e.g., hospital, home), (b) the
87 increasing number of data sources – both traditional (e.g., UK BioBank, NIH All Of Us) and non-
88 traditional (e.g., marker trajectories from wearables, social media traces of health-related
89 behavior), and (c) their multi-modality (e.g., structured electronic health records (EHR), medical
90 images, genome sequencing, clinical notes, voice recordings). In parallel with this ambitious
91 endeavor, the emergence of health data science has increased the need for significant changes
92 in the education of health professionals. Recent examples include courses in machine learning in
93 healthcare and opportunities to shadow a team deploying clinical algorithms in hospitals. Such
94 theoretical and practical trainings are foundational as data science plays an increasing role in the
95 provision of healthcare. These combined skills are needed to retrospectively derive novel insights
96 using statistical inference (e.g., estimating treatment effects using observational data), to build
97 interpretable clinical models (e.g., predicting in-hospital mortality in a given time horizon), and to
98 support their prospective implementation (e.g., conducting risk analysis and identifying potential
99 errors that can be systematically addressed). Depending on the professional role, skills in one or
100 more of these areas are becoming increasingly necessary to apprehend real-world data used in
101 clinical models.

102

103 ***Bias in clinical models can emanate from multiple sources***

104 The data underlying clinical models may contain biases that can be unknowingly propagated to
105 downstream inference and prediction tasks [1,2]. Such biases can emanate from multiple sources
106 – ranging from differences in how physicians report information in EHR and clinical notes to
107 artifacts in images to the miscalibration of medical devices – but they can also reflect existing

108 social determinants of health. In other words, biases encountered in clinical data are often
109 intersectional in nature, i.e., both social and technological [3]. For example, skin tone affects the
110 accuracy of pulse oximetry but is rarely considered in trials measuring medical device
111 performance [4,5,6,7]. Interrogating existing datasets and fully understanding the underlying
112 biases requires a multidisciplinary examination involving more than a single data analyst or
113 research group. Instead, the cooperation of clinicians, engineers, data scientists, social scientists,
114 and industry partners is greatly needed. Indeed, studies have shown that research groups with
115 more diverse expertise are more effective in identifying or addressing issues of bias [8].

116

117 ***Study hypothesis and contribution***

118 In this study, we seek to understand the role of open data and diversity in research expertise
119 towards mitigating biases affecting clinical models. We hypothesize that the groups of researchers
120 who leverage existing open-access datasets are more diverse than those using privately-owned
121 datasets. In what follows, we explain our rationale and motivation for analyzing the profile of
122 researchers who use open vs private datasets.

123

124 ***Many existing approaches to mitigate bias occur downstream of model development***

125 The timing of efforts to address bias may be critical. When, in the lifecycle of a clinical model, are
126 interventions to mitigate bias most effective? Healthcare systems can deploy interventions either
127 downstream or upstream of the model development phase to mitigate the repercussions of data
128 biases. Researchers in the field of ethics in artificial intelligence (AI) for health have participated
129 in this effort by exploring several downstream approaches. Notably, biases can be mitigated after
130 model fitting and/or deployment [9] through the use of explainable AI (XAI) tools [10,11] that
131 identify biased features contributing to discriminatory outcomes (e.g., by decomposing individual
132 predicted risk scores).

133 ***The public release of datasets offers an alternative, upstream approach***

134 In contrast to XAI tools and other technology-based solutions, the public release of datasets in
135 science, medicine, and engineering offers an upstream solution. This human-centric approach,
136 which focuses on better understanding biases in health data, maximizes the number of
137 investigators involved and leverages their cognitive and social diversity to examine clinical data.
138 There exist several examples throughout history whereby research necessitated multiple teams
139 to examine the same dataset to ultimately reach an agreement. The case of right heart
140 catheterization was notable: repeated analyses of the original dataset, collected by Connors et
141 al. in 1996 [12], yielded conflicting results. This confusion left clinicians needing clarification about
142 the effect of the procedure for years. Untangling the confounding factors to reach our current
143 understanding took several teams of biostatisticians. In sum, the coordinated efforts of many
144 investigators, in an iterative learning process, are required to achieve consensus in data analysis
145 and interpretation. In recent years [13,14], additional upstream approaches such as the
146 embedded ethics methodology have been proposed to ensure that interdisciplinary ethical inquiry
147 and deliberation are integrated into AI and healthcare technology development processes starting
148 at project ideation [15]. Others have advanced approaches such as algorithmic impact analysis,
149 which seeks to develop robust public interest methodologies to better understand the impact of
150 AI and automated decision-making systems on people's lives and society at large [16].

151

152 ***The current landscape of clinical data science research***

153 In the past five years, several organizations – including the National Institutes of Health (NIH),
154 European Commission, and World Economic Forum [17] – have recommended a shift to Open
155 Data, a movement whose goal is to increase the release of FAIR (Findable, Accessible,
156 Interoperable, and Reusable) [18] datasets in scientific research. In particular, large investments
157 have been made in the biomedical sciences [19,20]. However, the clinical data science landscape

158 remains highly siloed and opaque [21]. Expertise at the intersection of computer science and
159 healthcare is currently concentrated among a few academic and industry research teams
160 responsible for the preprocessing of data and the training of models [22,23]. Only researchers
161 who are fortunate enough to be aware of a dataset's existence, to be granted access to it, and to
162 have sufficient funding to afford the associated licensing fees and computing infrastructure, can
163 effectively leverage it in practice. Therefore, in far too many instances, datasets are often
164 inaccessible to investigators outside the very research team that curated them. For example,
165 researchers who are not clinicians (e.g., computer scientists at MIT working on diabetic
166 retinopathy in Uganda) often have limited access to primary data. Thus, they must rely on second-
167 hand knowledge from clinical investigators at their institution or in their network (e.g., physicians
168 at Harvard Medical School) but whose domain expertise may have been gained from datasets
169 and practices originated in a different context. For example, datasets developed in North America
170 may have significant limitations when used to train models to be implemented in East Africa and
171 vice versa. Such secondary data analysis requires a deep understanding of the data curation
172 process, including biases in data collection and artifacts in clinical measurements, which may vary
173 locally by medical site or by region. Thus, knowledge transfer alone is insufficient to safeguard
174 against the spread of biases. There are numerous reasons why bias occurs in clinical models.
175 For instance, a decision-support model to prioritize screening for diabetic retinopathy that does
176 not appropriately account for differences in the frequency of specialty visits among patients may
177 result in selection bias. This example points to the need for familiarity with relevant socio-
178 demographics and patient care-seeking behavior. Bias also can manifest when analysts inherit
179 datasets without any background about the underlying environment. In such situations, they risk
180 not only using a training dataset that is ill-suited for the target population, but also failing to
181 understand the limitations of their model because local features have not been considered.
182 Ideally, researchers will seek to interface with the team responsible for primary data collection.

183 Without such a dialogue, external teams run the risk of unsafely deploying algorithms that do not
184 generalize well out-of-distribution, for the cohorts of patients they care for.

185

186 ***The promise of new NIH data-sharing policies***

187 The recent imperative by the NIH to share scientific data publicly underscores a significant shift
188 in academic research [24]. Effective as of January 2023, it emphasizes that transparency in data
189 collection and dedicated efforts towards data sharing – with other investigators and the broader
190 public, including citizen scientists – are prerequisites for translational research, from the lab to the
191 bedside. Certain fields of healthcare have paved the way: workshops on data ethics, privacy,
192 consent, and anonymization have been organized in radiology [25]; a common ontology has been
193 developed for data standardization in radiation oncology [26]; multi-center data-sharing platforms
194 have been designed for collaboration in sleep medicine [27]; and distributed learning networks
195 have recently been proposed as a solution to preserve the privacy of patients' EHR [28]. In the
196 long run, requirements such as submitting a Data Management and Sharing Plan along with
197 funding applications [29] will allow a more diverse population of researchers to interrogate both
198 raw, unprocessed, and curated, pre-processed datasets. In light of the significance of data access
199 to the mitigation of bias, we hypothesize that researchers who leverage existing open-access
200 datasets rather than privately-owned ones are more diverse. We reason that the diversity of the
201 backgrounds of open dataset users may in turn result into greater attention to equity in patient
202 data collection and in a more compelling use of research data to address the pressing needs of
203 the global population. This increased attention to issues of equity could also facilitate a more
204 nuanced interpretation of study results and a stronger willingness to translate these results into
205 practice.

206

207

208

209 **Structure of this paper**

210 In this brief report, we propose to test this hypothesis in the transdisciplinary and expanding field
211 of AI for Critical Care. Specifically, we compare the diversity among authors of publications
212 leveraging open datasets, such as the commonly used MIMIC [30] and eICU databases [31], with
213 that among authors of publications relying exclusively on private datasets, unavailable to other
214 research investigators. To measure the extent of author diversity, we characterize gender
215 balance, geographic diversity (i.e., the number of countries with which authors are affiliated and
216 the income categories these countries map to), and the presence of researchers from minority-
217 serving institutions. Furthermore, we comment on the challenges of increasing the participation
218 of researchers from underrepresented groups and suggest changes that can be made to the
219 current Open Data strategy to enhance representation in authorship in the next decade. By
220 evaluating the association between data accessibility and author diversity, our study pinpoints
221 actionable steps that the broader field of clinical AI can take to foster inclusion in the scientific
222 community and mitigate blind spots in data preparation and/or model development.

223

224 **Methods**

225 Our scripts and datasets are publicly available on GitHub:

226 <https://github.com/anpetushkov/OpenVsPrivateDatasets>

227

228 Data

229 We leveraged PubMed [32] to select research studies at the intersection of AI and Critical Care
230 published between 2010 and 2022. We created two separate queries to derive (1) a list of
231 publications related to AI and (2) a list of publications addressing topics in critical care medicine.
232 We used the same process as Celi et al. [20] for the AI-specific query since the authors' model
233 identified AI-related publications with suitable performance for our task (AUROC=0.96). Query

234 search terms specific to critical care were selected based on Van de Sande et al. [33] and vetted
235 by two physician authors on our team, LAC and JG. Subsequently, we merged the two publication
236 lists, thereby capturing only the studies related to both fields. Further, we split the resulting set of
237 studies into two groups: “treatment” and “control.” The treatment group comprised publications
238 leveraging either of the two major critical care databases currently in open access, i.e., MIMIC
239 and eICU. We used dataset-specific queries from Google Dataset Search [34] to derive a list of
240 works leveraging MIMIC or eICU critical care databases. Conversely, the control group consisted
241 of publications based on privately-owned datasets that are unavailable to researchers other than
242 the primary investigators. To avoid leakage, we confirmed that the two groups were mutually
243 exclusive, i.e., no publication belonged to both. For studies in the control group, we first
244 downloaded their unique PMID identifiers from PubMed, owing to the large sample size. Then,
245 we used the Dimensions AI platform, an interlinked research information system provided by
246 Digital Science, to collect metadata pertaining to each research article [35]. For studies in the
247 treatment group, we performed the query via Dimensions AI directly as the sample size was much
248 smaller. The platform was accessed in June 2023. Finally, we manually filtered the initial set of
249 papers in the treatment group to exclude outliers and include only relevant MIMIC and eICU
250 manuscripts. Three team members (AP, CL, and GL) completed this manual validation task
251 independently before reconvening and reaching a consensus. Details about the creation of the
252 study dataset are available in **Figure 1.a**.

253

254 Labeling of author gender

255 Each author's first name was processed by the Genderize.io Application Programming Interface
256 (API) (**Figure 1.a**). The API is based on a global collection of first names that have been manually
257 annotated and linked to their most likely gender. Building on this international database of first
258 names, the probability that an author is a woman or a man can be derived from the API. The API

259 returns an “unknown” label when the uncertainty is too high. We assigned the most probable
260 gender label associated with each author’s first name (“female,” “male,” or “unknown”).

261

262 Labeling of minority-serving institutions

263 To measure the extent of the representation of minority-serving institutions (MSI) within the control
264 and treatment groups, we developed our own fuzzy-matching pipeline between a pre-specified
265 list of institutions and each author’s affiliation(s). In particular, we built upon the fuzzy-match
266 Python package [36], specifically the Levenshtein Partial Ratio Function with a matching threshold
267 of 97 percent. The list of MSIs used in this study was obtained by combining two data sources:
268 the 2020 list from [37] comprising 774 distinct MSIs and the 2022 list from [38] containing 865
269 distinct MSIs (**Figure 1.a**). A total of 566 institutions were shared by the two sources (exact +
270 fuzzy matching based on the institution’s name), while the remaining MSIs were unique to each
271 list. The integration of these two data sources allowed for a more comprehensive set of MSIs.

272 To confirm the accuracy of the mapping between institutional affiliations and their potential MSI
273 status, we performed manual verification of outputs from the fuzzy-matching process for both the
274 2020 and 2022 MSI datasets. In cases when an author’s institutional affiliation was incorrectly
275 mapped to an MSI, we rectified the mistake manually. Verification was limited to reducing false
276 positives, i.e., we only determined institutional affiliations that were erroneously linked with MSIs.
277 However, our matching process was deemed comprehensive, since we selected a high threshold
278 value of 97 percent to limit the number of false negatives.

279

280 Labeling of institutions based in low- and middle-income countries

281 The 2022 World Bank country classification was used to map countries associated with
282 researcher affiliations to the low- and middle-income category (LMIC) or the high-income category
283 (HIC) (**Figure 1.a**). Countries are ranked according to the gross national income. For authors with

284 multiple affiliations, each was considered separately and mapped to the corresponding income
285 category.

286

287 Diversity metrics

288 A total of three diversity metrics were considered. First, for each paper, we quantified the overall
289 number of authors, the number of probable women among the authors, and whether the first/last
290 author was likely a woman. For both the control and treatment groups, we derived the proportion
291 of probable women by article and the overall percentage of articles featuring an author who was
292 likely a woman as the first and/or last author. Second, for each paper, we measured the number
293 of authors affiliated with an institution based in an LMIC and whether the first/last author was
294 based in an LMIC. For both groups, we derived the proportion of LMIC authors by article and the
295 overall percentage of articles featuring an LMIC author in a leading role. Third, for each article,
296 we quantified the number of authors affiliated with an MSI and the MSI status of the first and/or
297 last author's affiliation. For both groups, we similarly derived the proportion of MSI authors by
298 article and the overall percentage of articles featuring MSI authors. Note that papers whose first
299 or last author had an unknown gender or unidentifiable LMIC or MSI status based on their
300 affiliation were excluded from the corresponding analyses.

301

302 Statistical analysis

303 For each of the three diversity metrics of interest (i.e., gender representation, geographic diversity,
304 and MSI status), we performed a one-sided proportional Chi-squared test of independence to
305 determine if there was a significant difference between the control and treatment groups. We set
306 the threshold for statistical significance to 0.05, following common practice. A p-value less than
307 0.05 would thus indicate a statistically significant difference between the control and treatment
308 groups (e.g., in terms of the representation of women, LMIC, or MSI authors), in favor of the latter.
309 We conducted three sensitivity analyses to assess the impact of missing data and to test the

310 robustness of our results. In the first counterfactual scenario, we assumed that none of the papers
311 with missing author affiliation had any authors from LMICs. Conversely, in the second
312 counterfactual scenario, we assumed that all of these papers had at least one author affiliated
313 with an institution based in an LMIC. Lastly, in the third counterfactual scenario, we assumed that
314 missing income category labels could be imputed via dataset-specific distributions derived from
315 labeled data, i.e., using either that of the control or treatment group, depending on the group the
316 article belonged to (**Figure 1.b**).

317

318 **Results**

319 Overall, we identified 5,219 Critical Care AI papers, including 2,912 studies in the control group
320 (i.e., 55.8%) and 2,307 studies in the treatment group (i.e., 44.2%). The control and treatment
321 groups comprised 17,999 and 9,959 distinct authors, respectively; among them, 16,743 (93.0%
322 of the control group) and 8,210 (82.4% of the treatment group) had available research affiliation
323 information. A total of 562 authors appeared in both groups, representing 3.4% of the control
324 group and 6.8% of the treatment group. In the treatment group, the three leading venues were all
325 preprint servers, accounting for 29.3% of all articles: arXiv (20.6%), Research Square (6.4%), and
326 medRxiv (2.3%). Following these, the next three most popular venues were journals, accounting
327 for 4.4% of all articles: Frontiers in Medicine (1.8%), Scientific Reports (1.3%), and, notably,
328 Critical Care Medicine (1.3%), the flagship journal in the field. Together, these six venues
329 accounted for 778 (33.7%) of all articles in the treatment group, suggesting that authors
330 leveraging open datasets publish their work in a great diversity of outlets. In contrast, in the control
331 group, the top venue was the IEEE Engineering in Medicine and Biology Society Conference
332 (EMBC, 3.4%), followed by two journals, PLOS One (2.4%) and Scientific Reports (2.3%).
333 Collectively, the top three forums represented 238 (8.2%) of all articles in the control group.
334 Beyond the journal Critical Care Medicine, which accounts for 68 articles (2.3%), other popular
335 venues included the Journal of Clinical Monitoring and Computing (1.5%) and Computers in

336 Biology and Medicine (1.2%). Overall, the six main venues accounted for 385 (13%) of all articles
 337 in the control group, underscoring the heterogeneity of outlets in which authors leveraging private
 338 datasets publish their work as well. The comprehensive breakdown of author characteristics for
 339 each group is detailed in **Table 1**, while the full distribution of conference venues and journals is
 340 available in our GitHub repository. For each group, the distribution of papers among the top 10
 341 venues is available in **Supplementary Figure 1** (treatment) and **Supplementary Figure 2**
 342 (control).

Table 1a. Characteristics of authors in the treatment and control groups.			
Diversity metric and author role		Treatment group MIMIC/eICU (n / non-missing)	Control group (n / non-missing)
Probable woman (inferred gender)	Any author	27.8% (2438 / 8758)	30.8% (4710 / 15285)
LMIC*	Any author	6.7% (547 / 8210)	3.5% (590 / 16743)
MSI*	Any author	8.6% (190 / 2207)	5.6% (277 / 4914)
Table 1b. Characteristics of papers in the treatment and control groups.			
Diversity metric and author role		Treatment group MIMIC/eICU (n / non-missing)	Control group (n / non-missing)
Probable woman (inferred gender)	Any Author	71.8% (1024 / 1426)	73.0% (1330 / 1823)
	First Author	28.1% (401 / 1426)	30.9% (564 / 1823)
	Last Author	23.6% (336 / 1426)	21.7% (395 / 1823)
LMIC*	Any Author	10.1% (150 / 1487)	6.2% (168 / 2694)
	First Author	7.9% (117 / 1487)	4.8% (130 / 2694)
	Last Author	7.8% (116 / 1487)	4.6% (123 / 2694)
MSI*	Any Author	27.6% (126 / 456)	17.7% (175 / 991)
	First Author	7.9% (36 / 456)	4.7% (47 / 991)
	Last Author	6.1% (28 / 456)	5.5% (55 / 991)
LMIC* & woman (inferred gender)	Any Author	6.2% (59 / 945)	3.8% (64 / 1698)
MSI* & woman (inferred gender)	Any Author	20.6% (35 / 170)	34.5% (86 / 249)

*LMIC: Low- and Middle-Income Country, *MSI: Minority Serving Institution.

343 *Note 1:* LMIC-related statistics were determined based on papers with non-missing affiliation and
344 hence non-missing country information. Note that 368 (8%) of all papers had at least one author
345 with missing country information. This missing data pattern affected 185 (11.1%) and 183 (6.4%)
346 papers from the treatment and control groups, respectively.

347 *Note 2:* In the treatment group, 104 papers (7.0%) had only one author. In the control group, 178
348 papers (6.6%) had only one author.

349

350 Gender

351 The proportion of papers with at least one author inferred to be a woman was qualitatively
352 comparable between the two groups, albeit slightly higher in the control group (73.0% vs. 71.8%,
353 $z=-0.726$, $p=0.468$). The representation of women among the first and the last authors was similar
354 between the two groups (28.1% vs. 30.9%, $z=-1.74$, $p=0.0811$; 23.6% vs. 21.7%, $z=1.28$,
355 $p=0.199$, respectively). Importantly, in both groups, the proportion of women serving as a last
356 author (overall average of 22.5%), often reflecting a senior research leadership role, was lower
357 than that of women serving as a first author (overall average of 29.7%), generally awarded to the
358 person leading study design and analysis. This difference was more pronounced (9.2 vs. 4.5
359 percentage points) in the control than in the treatment group.

360

361 Low- and middle-income countries (LMIC)

362 Overall, out of the 4,181 articles with non-missing affiliations included in our study, 318 (i.e., 7.6%)
363 had at least one LMIC author. The proportion of authors from LMICs was substantially higher in
364 the treatment than in the control group (10.1% vs. 6.2%, $z=4.5$, $p<0.001$). Moreover, we found
365 that the diversity of first and last authors in terms of country of affiliation was greater in the
366 treatment group, i.e., among studies leveraging MIMIC and eICU open critical care datasets.
367 Indeed, 7.9% (vs 4.8%, $z=4.30$, $p<0.001$) of papers in the treatment group had their first author
368 affiliated with an LMIC country. Furthermore, 7.8% (vs 4.5% $z=3.14$, $p<0.001$) had their last author

369 affiliated with an LMIC country. The first sensitivity analysis, imputing missing data using the
370 distribution based on labeled samples, also led to the conclusion of a greater representation of
371 LMIC authors among researchers leveraging open datasets (overall: 10.1 % vs 8.3%, $z=4.5$,
372 $p<0.001$). The second sensitivity analysis, which made the optimistic assumption that all papers
373 with missing affiliation information had authors from LMICs, confirmed the robustness of our
374 results (overall: 42.0% vs 16.7%, $z=23.5$, $p<0.001$). The third sensitivity analysis, which made the
375 pessimistic assumption that none of the papers with missing country information had authors from
376 LMICs, yielded results qualitatively similar to the main analysis, albeit not statistically significant
377 (overall: 6.5% vs 7.3%, $z=1.1$, $p=0.14$; first: 4.6% vs. 4.9%, $z=1.3$, $p=0.10$; last: 4.6% vs 5.2%,
378 $z=0.77$, $p=0.22$).

379

380 Minority-serving institutions (MSI)

381 Our analysis of authorship among MSIs was restricted to the United States (US). The control
382 group comprised 4,914 distinct authors with institutional affiliations within the US for a total of
383 16,743 distinct authors with an affiliation worldwide (i.e., 29.3%). Among them, 277 different
384 authors were affiliated with MSIs, accounting for approximately 5.6% of the total. In contrast, the
385 treatment group comprised 2,207 authors with institutional affiliations within the US, for a total of
386 8,210 distinct authors with non-missing affiliations worldwide (i.e., 26.9%). Among them, 190
387 different authors were affiliated with MSIs, accounting for approximately 8.6% of the total. Thus,
388 the proportion of US-based authors affiliated with an MSI was 1.5 times higher among articles in
389 the treatment than in the control group; this difference was statistically significant, suggesting that
390 open data resources attract a larger pool of participants from minority groups ($z=4.69$, $p<0.001$).

391 In addition to overall statistics, we characterized the involvement of MSI researchers at the team
392 level, i.e., per paper. The control (resp. treatment) group consisted of 991 (resp. 456) distinct
393 papers with at least one author having an institutional affiliation in the US, out of 2,877 (resp.

394 1,672) papers worldwide (i.e., 34.4% and 27.3%, respectively). Among these papers, 175 (resp.
395 126) had at least one author affiliated with an MSI, representing approximately 17.7% (resp.
396 27.6%) of the total US research output in critical care AI involving the use of private (resp. open)
397 databases. Thus, the proportion of papers featuring US-based authors affiliated with an MSI was
398 1.6 times higher in the treatment than in the control group; this difference was statistically
399 significant, suggesting that MSI researchers effectively benefit from open data resources, which
400 further translates into publications and preprints ($z=4.34$, $p<0.001$).

401 Out of the 991 distinct papers in the control group, the percentage of papers with an MSI-affiliated
402 first author reached only 4.7% (47 / 991). The proportion of MSI researchers serving as first
403 authors was significantly greater ($z=2.40$, $p=0.008$) in the treatment group, reaching 7.9% (36 /
404 456). This result suggests that barriers remain for MSI-affiliated authors to lead research studies
405 when the underlying datasets are inaccessible to the broader public. In contrast, opening critical
406 care datasets can bolster the participation of MSI researchers as first authors. Of note, the
407 percentage of papers with an MSI-affiliated last author was larger ($z=0.45$, $p=0.327$) in the
408 treatment group (28 / 456, i.e., 6.1%) than in the control group (55 / 991, i.e., 5.5%), but this
409 difference was statistically insignificant owing to a reduced sample size in the analysis focused
410 on MSI representation.

411 Intersectionality

412 ***Gender and LMIC***

413 Of the 2,643 papers with complete author data regarding both gender and LMIC status, a clear
414 difference emerged between the control and treatment groups, with respect to the intersectional
415 representation of researchers. In the treatment group comprising 945 papers, 59 (i.e., 6.2%)
416 featured at least one woman and at least one LMIC-based researcher among the authors. In
417 contrast, in the control group comprising 1,698 papers, only 64 (i.e., 3.8%) included both a woman
418 and an LMIC-based researcher. This difference was statistically significant ($z=2.78$, $p=0.003$),

419 underscoring greater intersectional diversity among authors who leveraged the publicly available
420 MIMIC and eICU databases than among authors in the control group, who relied exclusively on
421 private datasets.

422 ***Gender and MSI***

423 Among the 277 authors in the control group affiliated with an MSI, 249 authors had non-missing
424 gender information and 86 (i.e., 35%) were women. In contrast, among the 190 authors in the
425 treatment group, 170 authors had non-missing gender, and 35 were women (i.e., 21%). The
426 difference was statistically insignificant ($z=-3.09$, $p=0.999$); hence there was no evidence of
427 greater intersectional diversity, by gender and MSI, in the treatment group. Nonetheless, the
428 sample sizes resulting from multiple stratifications were quite small, therefore limiting statistical
429 power. Thus, future efforts should focus on monitoring trends over time to gather more evidence
430 about differences in the representation of authors, with a focus on the intersectionality of their
431 identities.

432

433 **Discussion**

434 Our comparative analysis revealed a greater contribution of authors from LMICs and MSIs among
435 researchers leveraging open critical care datasets than among those relying exclusively on private
436 data resources. The participation of women was similar between the two groups, albeit slightly
437 larger in the treatment group. Notably, although over 70% of all articles included at least one
438 author likely to be a woman, they served as a first or last author in less than 25% of those articles.
439 Thus, our study highlights the value of the Open Data strategy for underrepresented groups, while
440 also quantifying persisting gender gaps in academic and clinical research at the intersection of
441 computer science and healthcare. In doing so, we hope our work points to the importance of
442 extending open data practices in deliberate and systematic ways.

443

444 While incorporating AI into healthcare is a technically challenging endeavor, its success depends
445 not only on the performance of clinical models but also on the humans interfacing with them.
446 Clinical models are a reflection of the patient data they are trained upon. The people collecting,
447 processing, and analyzing the data all play a role in rendering the final representation of patients
448 underlying inference and prediction tasks. Therefore, cognitive diversity among researchers
449 responsible for study design and data examination will facilitate a more thoughtful investigation of
450 potential pitfalls encoded within clinical data.

451 Critical care research is still highly imbalanced. For example, while the incidence and mortality of
452 sepsis is the highest in sub-Saharan Africa and other low- and middle-income countries, over
453 75% of clinical studies underlying the 2021 sepsis guidelines were conducted in high-income
454 countries [39]. As the complexity of critical care data has increased, so has the complexity of the
455 biases introduced: because of pronounced imbalances in the patient populations featured in
456 research datasets [40], their identification can be difficult. Our research shows that open access
457 to critical care data can change the status quo. With the public release of datasets such as MIMIC
458 and eICU, we found that participation from authors based in LMICs or affiliated with MSIs can be
459 greatly improved.

460 Open data offers a resource for data scientists and healthcare specialists to develop skills that
461 are essential to patient care in the digital health era. However, the transparent release of datasets
462 on freely-accessible cloud platforms is not sufficient in itself to generate meaningful knowledge in
463 the biomedical sciences. Beyond data sharing and collaboration across institutions,
464 improvements in education and research should be sought at multiple stages, starting with
465 outreach programs aimed at diversifying teams of clinicians and engineers as well as continued
466 education to raise awareness about both persisting and evolving health disparities. For instance,
467 the INFORMED fellowship in oncology data science offered by the National Cancer Institute (NCI)
468 constitutes an excellent model to be replicated elsewhere [41,42]. To sustain LMIC participation

469 in critical care research, reducing the barrier to learning, engaging with, and publishing in, the
470 digital health field is vital. Models are often built in one site but deployed in others. Therefore, it is
471 crucial to enable teams serving at medical centers with fewer resources to examine distributional
472 shifts between their local data and the data originally used for training and validation and to
473 evaluate model performance locally [43]. Furthermore, temporal evaluations of subpopulation
474 shifts (i.e., related to variation in patient sociodemographics and/or clinical profiles) and calibration
475 drifts (possibly related to the former or to changes in clinical practice, outcome detection tools
476 etc.) must be continuously performed. These checks can help detect the emergence of new
477 disparities and measure the effectiveness of interventions aiming to correct for those that were
478 previously identified [44]. With the increasing digitization of medical records, the hope is that the
479 community of health informatics researchers will broaden and help break institutional silos at each
480 site. While continuing to advocate for the collection of comprehensive clinical datasets
481 appropriately reflecting the target populations, investing in implementation science and striving to
482 integrate clinical models into healthcare systems should also be prioritized.

483 Although the differences observed between works leveraging open vs. private datasets are
484 striking, we acknowledge three key limitations that affect the precision of the prevalences reported
485 in our study. First, the algorithm that classified an author's likely gender was trained using binary
486 gender labels on a researcher's first name, which provides an imperfect proxy for a complex
487 attribute such as gender identity. While common in bibliometric analyses, this method overlooks
488 nuances in how individuals self-identify and excludes those who do not fit into binary gender
489 categories. Going forward, integrating survey data from individual researchers could enable a
490 more accurate categorization, especially as gender model performance varies across languages
491 and cultures. Second, representation was assessed only in terms of the three following
492 dimensions: gender, country income level, and minority-serving status of the author's institutions.
493 Future work should move beyond such unidimensional definitions of diversity to capture the
494 intersectional relationships that shape experiences in academia and clinical research. Third, the

495 definition of minority-serving institutions was based on a US-centric designation. Future research
496 should seek to assess geographic diversity more comprehensively, extending the analysis to
497 countries outside the US; a globally inclusive framework is needed to understand how researchers
498 from under-resourced institutions worldwide engage with open data within and across nations.

499 **Conclusion**

500 Without a concerted commitment to diversifying authorship, clinical AI research risks being
501 confined to a limited group of institutions and individuals. Such homogeneity may introduce and
502 perpetuate biases within AI systems, potentially exacerbating health disparities and reinforcing
503 existing inequities in healthcare delivery. In response, we must actively promote gender
504 representation and include voices from institutions that serve underrepresented populations,
505 thereby incorporating essential perspectives that address the multifaceted dimensions of
506 inequality.

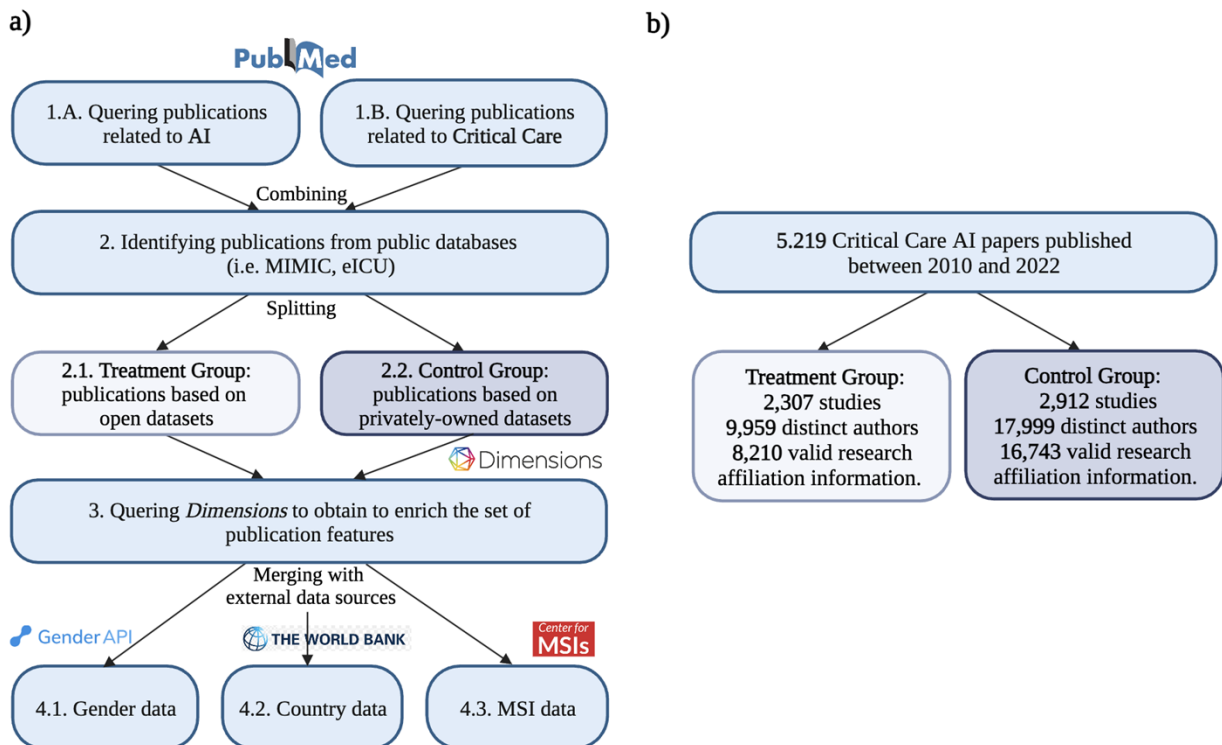
507 The rise of open data platforms adhering to FAIR (Findable, Accessible, Interoperable, and
508 Reusable) principles brings new opportunities for investigators worldwide to participate in
509 biomedical research and knowledge creation. Future policy interventions, including by institutions
510 and editorial boards, should consider the complex associations among access to open data, bias
511 in the development and use of clinical models, and diversity in research groups. Going forward, it
512 will be key to monitor progress frequently, particularly with respect to the intersectional
513 representation of authors – not only by gender, geography, and MSI status, but across the multiple
514 dimensions that constitute a scientist’s identity. Despite recent advances, sustaining an open and
515 inclusive clinical AI ecosystem will require the retention of diverse talent, across geographies and
516 career stages, in part through the provision of dedicated training. Policy and technology
517 innovations must continue lowering barriers that prevent the broader, collaborative engagement
518 of researchers with clinical data resources.

519

521 **Figures**

522 **Figure 1.** Flow diagram illustrating **a)** the methodology for the analysis of authorship diversity in
523 scientific publications at the intersection of AI and Critical care and **b)** the number of
524 publications considered.

525



526

527

528

529

530

531

532

533

534

535

536 **References**

- 537
- 538 [1] L. H. Nazer *et al.*, “Bias in artificial intelligence algorithms and recommendations for
539 mitigation,” *PLOS Digital Health*, vol. 2, no. 6, Jun. 2023, doi: 10.1371/journal.pdig.0000278.
- 540 [2] M.-L. Charpignon *et al.*, “Critical Bias in Critical Care Devices,” *Critical Care Clinics*, vol.
541 39, no. 4, pp. 795–813, Oct. 2023, doi: 10.1016/j.ccc.2023.02.005.
- 542 [3] J. Gallifant, L. A. Celi, and R. L. Pierce, “Digital determinants of health: opportunities and
543 risks amidst health inequities,” *Nature Reviews Nephrology*, vol. 19, no. 12, pp. 749–750, Aug.
544 2023, doi: 10.1038/s41581-023-00763-4.
- 545 [4] E. R. Gottlieb, J. Ziegler, K. Morley, B. Rush, and L. A. Celi, “Assessment of racial and
546 ethnic differences in oxygen supplementation among patients in the intensive care unit,” *JAMA*
547 *internal medicine*, vol. 182, no. 8, pp. 849–858, Aug. 2022, doi:
548 10.1001/jamainternmed.2022.2587.
- 549 [5] A.-K. I. Wong *et al.*, “Analysis of Discrepancies Between Pulse Oximetry and Arterial
550 Oxygen Saturation Measurements by Race and Ethnicity and Association With Organ
551 Dysfunction and Mortality,” *JAMA network open*, vol. 4, no. 11, p. e2131674, Nov. 2021, doi:
552 10.1001/jamanetworkopen.2021.31674.
- 553 [6] T. Choy, E. Baker, and K. Stavropoulos, “Systemic Racism in EEG Research:
554 Considerations and Potential Solutions,” *Affective science*, vol. 3, no. 1, pp. 14–20, May 2021,
555 doi: 10.1007/s42761-021-00050-0.
- 556 [7] D. Judelson, “Examining the Gender Bias in Evaluating Coronary Disease in Women,”
557 *Medscape Women’s Health*, vol. 2, no. 2, p. 5, Feb. 1997.
- 558 [8] T. H. Swartz, A. G. Palermo, S. K. Masur, J. A. Aberg. The science and value of
559 diversity: closing the gaps in our understanding of inclusion and diversity. *The Journal of*
560 *infectious diseases*. 2019 Aug 20;220(Supplement_2):S33-41.
- 561 [9] L. A. Celi, L. Citi, M. Ghassemi, and T. J. Pollard, “The PLOS ONE collection on
562 machine learning in health and biomedicine: Towards open code and open data,” *PLOS ONE*,
563 vol. 14, no. 1, Jan. 2019, doi: 10.1371/journal.pone.0210232.
- 564 [10] C. Agarwal *et al.*, “OpenXAI: Towards a Transparent Evaluation of Model Explanations,”
565 *arXiv.org*, Jun. 22, 2022. <https://arxiv.org/abs/2206.11104>
- 566 [11] U. Bhalla, S. Srinivas, and H. Lakkaraju, “Verifiable Feature Attributions: A Bridge
567 between Post Hoc Explainability and Inherent Interpretability,” *arXiv.org*, Jul. 27, 2023.
568 <https://arxiv.org/abs/2307.15007>
- 569 [12] M. A. F. C. Jr, “The Effectiveness of Right Heart Catheterization in the Initial Care of

- 570 Critically Ill Patients,” *JAMA*, vol. 276, no. 11, pp. 889–897, Sep. 1996, doi:
571 10.1001/jama.1996.03540110043030.
- 572 [13] L. Groves, “Algorithmic impact assessment: a case study in healthcare,” *Ada Lovelace*
573 *Institute*, Feb. 22, 2022. [https://www.adalovelaceinstitute.org/report/algorithmic-impact-](https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/)
574 [assessment-case-study-healthcare/](https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/)
- 575 [14] J. Metcalf, E. Moss, E. A. Watkins, R. Singh, and M. C. Elish, “Algorithmic Impact
576 Assessments and Accountability,” in *Proceedings of the 2021 ACM Conference on Fairness,*
577 *Accountability, and Transparency*, Mar. 2021. Accessed: Feb. 04, 2024. [Online]. Available:
578 <http://dx.doi.org/10.1145/3442188.3445935>
- 579 [15] S. McLennan *et al.*, “An embedded ethics approach for AI development,” *Nature*
580 *Machine Intelligence*, vol. 2, no. 9, pp. 488–490, Jul. 2020, doi: 10.1038/s42256-020-0214-1.
- 581 [16] “Algorithmic impact assessment: a case study in healthcare,” *Ada Lovelace Institute*.
582 [https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-](https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/)
583 [healthcare/](https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/) (accessed Feb. 29, 2024).
- 584 [17] W. E. Forum, “Strategic Intelligence,” *Strategic Intelligence*.
585 [https://intelligence.weforum.org/topics/3f1279d958164a90958565e5f456b57a/key-](https://intelligence.weforum.org/topics/3f1279d958164a90958565e5f456b57a/key-issues/1097becddef548838db7f57bb57a48ce)
586 [issues/1097becddef548838db7f57bb57a48ce](https://intelligence.weforum.org/topics/3f1279d958164a90958565e5f456b57a/key-issues/1097becddef548838db7f57bb57a48ce) (accessed Nov. 04, 2023).
- 587 [18] M. D. Wilkinson *et al.*, “The FAIR Guiding Principles for scientific data management and
588 stewardship,” *Scientific Data*, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.1038/sdata.2016.18.
- 589 [19] T. C. Knepper and H. L. McLeod, “When will clinical trials finally reflect diversity?,”
590 *Nature*, vol. 557, no. 7704, pp. 157–159, May 2018, doi: 10.1038/d41586-018-05049-5.
- 591 [20] L. A. Celi *et al.*, “Sources of bias in artificial intelligence that perpetuate healthcare
592 disparities—A global review,” *PLOS Digital Health*, vol. 1, no. 3, Mar. 2022, doi:
593 10.1371/journal.pdig.0000022.
- 594 [21] J. Zhang *et al.*, “Mapping and evaluating national data flows: transparency, privacy, and
595 guiding infrastructural transformation,” *The Lancet Digital Health*, vol. 5, no. 10, pp. e737–e748,
596 Oct. 2023, doi: 10.1016/S2589-7500(23)00157-7.
- 597 [22] J. Matos *et al.*, “The Medical Knowledge Oligarchies,” Cold Spring Harbor Laboratory,
598 Jun. 2023. Accessed: Dec. 26, 2023. [Online]. Available:
599 <http://dx.doi.org/10.1101/2023.06.02.23290881>
- 600 [23] J. Zhang *et al.*, “An interactive dashboard to track themes, development maturity, and
601 global equity in clinical artificial intelligence research,” *The Lancet Digital Health*, vol. 4, no. 4,
602 pp. e212–e213, Apr. 2022, doi: 10.1016/S2589-7500(22)00032-2.
- 603 [24] M. Kozlov, “NIH issues a seismic mandate: share data publicly,” *Nature*, Feb. 16, 2022.

- 604 <https://www.nature.com/articles/d41586-022-00402-1>
- 605 [25] J. C. Battle *et al.*, “Data Sharing of Imaging in an Evolving Health Care World: Report of
606 the ACR Data Sharing Workgroup, Part 1: Data Ethics of Privacy, Consent, and
607 Anonymization,” *Journal of the American College of Radiology*, vol. 18, no. 12, pp. 1646–1654,
608 Dec. 2021, doi: 10.1016/j.jacr.2021.07.014.
- 609 [26] A. Traverso, J. van Soest, L. Wee, and A. Dekker, “The radiation oncology ontology
610 (ROO): Publishing linked data in radiation oncology using semantic web and ontology
611 techniques,” *Medical Physics*, vol. 45, no. 10, Aug. 2018, doi: 10.1002/mp.12879.
- 612 [27] M. Beier *et al.*, “Multicenter data sharing for collaboration in sleep medicine,” *Future
613 Generation Computer Systems*, vol. 67, pp. 466–480, Feb. 2017, doi:
614 10.1016/j.future.2016.03.025.
- 615 [28] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. Ch. Paschalidis, and W. Shi, “Federated
616 learning of predictive models from federated Electronic Health Records,” *International Journal of
617 Medical Informatics*, vol. 112, pp. 59–67, Apr. 2018, doi: 10.1016/j.ijmedinf.2018.01.007.
- 618 [29] S. Gonzales, M. B. Carson, and K. Holmes, “Ten simple rules for maximizing the
619 recommendations of the NIH data management and sharing plan,” *PLOS Computational
620 Biology*, vol. 18, no. 8, p. e1010397, Aug. 2022, doi: 10.1371/journal.pcbi.1010397.
- 621 [30] A. Johnson, T. Pollard, and R. Mark, “MIMIC-III Clinical Database,” Sep. 04, 2016.
622 <https://physionet.org/content/mimiciii/1.4/> (accessed Nov. 04, 2023).
- 623 [31] “eICU.” <https://eicu-crd.mit.edu/about/eicu/> (accessed Nov. 04, 2023).
- 624 [32] “PubMed,” *PubMed*. <https://pubmed.ncbi.nlm.nih.gov> (accessed Dec. 26, 2023).
- 625 [33] D. van de Sande, M. E. van Genderen, J. Huisken, D. Gommers, and J. van Bommel,
626 “Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the
627 intensive care unit,” *Intensive Care Medicine*, vol. 47, no. 7, pp. 750–760, Jun. 2021, doi:
628 10.1007/s00134-021-06446-7.
- 629 [34] “Dataset Search,” *Google Dataset Search*. <https://datasetsearch.research.google.com/>
630 (accessed Nov. 04, 2023).
- 631 [35] “Dimensions AI,” *Dimensions*. <https://www.dimensions.ai> (accessed Nov. 04, 2023).
- 632 [36] “fuzzy-match,” *PyPI*. <https://pypi.org/project/fuzzy-match/> (accessed Nov. 04, 2023).
- 633 [37] “List of minority serving institutions 2020.”
634 https://github.com/anpetushkov/OpenVsPrivateDatasets/blob/main/Data/GuruData/2020_Minorit
635 [y_Serving_Institutions-1.csv](https://github.com/anpetushkov/OpenVsPrivateDatasets/blob/main/Data/GuruData/2020_Minorit_y_Serving_Institutions-1.csv) (accessed Jul. 04, 2023).
- 636 [38] “List of minority serving institutions 2022.”
637 <https://github.com/anpetushkov/OpenVsPrivateDatasets/blob/main/Data/GuruData/2022%20CM>

- 638 SI%20Eligibility%20Matrix%20.csv (accessed Jul. 04, 2023).
- 639 [39] L. Nazer *et al.*, “Patient diversity and author representation in clinical studies supporting
640 the Surviving Sepsis Campaign guidelines for management of sepsis and septic shock 2021: a
641 systematic review of citations,” *BMC Infectious Diseases*, vol. 23, no. 1, pp. 1–10, Nov. 2023,
642 doi: 10.1186/s12879-023-08745-4.
- 643 [40] J. Zhang *et al.*, “Quantifying digital health inequality across a national healthcare
644 system,” *BMJ Health & Care Informatics*, vol. 30, no. 1, Nov. 2023, doi: 10.1136/bmjhci-2023-
645 100809.
- 646 [41] S. Khozin, G. Kim, and R. Pazdur, “From big data to smart data: FDA’s INFORMED
647 initiative,” *Nature Reviews Drug Discovery*, vol. 16, no. 5, pp. 306–306, Feb. 2017, doi:
648 10.1038/nrd.2017.26.
- 649 [42] “Growing the Field—NCI Fellowship Opportunities in Data Science,” *CBIIT*.
650 [https://datascience.cancer.gov/news-events/blog/growing-field-nci-fellowship-opportunities-data-](https://datascience.cancer.gov/news-events/blog/growing-field-nci-fellowship-opportunities-data-science)
651 [science](https://datascience.cancer.gov/news-events/blog/growing-field-nci-fellowship-opportunities-data-science)
- 652 [43] A. Youssef, M. Pencina, A. Thakur, T. Zhu, D. Clifton, and N. H. Shah, “External
653 validation of AI models in health should be replaced with recurring local validation,” *Nature*
654 *Medicine*, vol. 29, no. 11, pp. 2686–2687, Oct. 2023, doi: 10.1038/s41591-023-02540-z.
- 655 [44] J. Gallifant *et al.*, “Disparity dashboards: an evaluation of the literature and framework for
656 health equity improvement,” *The Lancet Digital Health*, vol. 5, no. 11, pp. e831–e839, Nov.
657 2023, doi: 10.1016/S2589-7500(23)00150-4.