Journal of the American Medical Informatics Association, 28(2), 2021, 360–364
doi: 10.1093/jamia/ocaa258
Advance Access Publication Date: 7 October 2020
Brief Communication

Brief Communication

Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset

Carlos Sáez[1], Nekane Romero[1], J. Alberto Conejero[2], and Juan M. García-Gómez[1]

[1]Biomedical Data Science Lab, Instituto Universitario de Tecnologías de la Información y Comunicaciones, Universitat Politècnica de València, Camino de Vera s/n, Valencia 46022, España and [2]Instituto Universitario de Matemática Pura y Aplicada, Universitat Politécnica de València, Valencia, Spain

Corresponding Author: Carlos Sáez, Biomedical Data Science Lab, Instituto Universitario de Tecnologías de la Información y Comunicaciones (ITACA), Building 8G, Access B, Universitat Politècnica de València (UPV), Camino de Vera s/n, Valencia 46022, España (carsaesi@upv.es)

Received 19 June 2020; Revised 7 September 2020; Editorial Decision 27 September 2020; Accepted 28 September 2020

ABSTRACT

Objective: The lack of representative coronavirus disease 2019 (COVID-19) data is a bottleneck for reliable and generalizable machine learning. Data sharing is insufficient without data quality, in which source variability plays an important role. We showcase and discuss potential biases from data source variability for COVID-19 machine learning.
Materials and Methods: We used the publicly available nCov2019 dataset, including patient-level data from several countries. We aimed to the discovery and classification of severity subgroups using symptoms and comorbidities.
Results: Cases from the 2 countries with the highest prevalence were divided into separate subgroups with distinct severity manifestations. This variability can reduce the representativeness of training data with respect the model target populations and increase model complexity at risk of overfitting.
Conclusions: Data source variability is a potential contributor to bias in distributed research networks. We call for systematic assessment and reporting of data source variability and data quality in COVID-19 data sharing, as key information for reliable and generalizable machine learning.

Key words: COVID-19, data quality, machine learning, biases, data sharing, distributed research networks, multi-site data, variability, heterogeneity, dataset shift

INTRODUCTION

The reliability of data for developing robust predictive models and clinical decision support systems to help fight coronavirus disease 2019 (COVID-19) is crucial. It is urgent that we rapidly identify clinical decision that can be informed by learning from large datasets. An editorial in *The BMJ* has recently drawn attention to the potential biases and misuses of COVID-19 predictive models if a certain level of quality is not guaranteed during their development.[1] The editorial is linked to the work by Wynants et al,[2] who found uniformly poor quality in COVID-19 predictive models, discouraging their clinical use. They concluded that there is an urgent need for large patient-level COVID-19 datasets to improve the quality of predictive models.

However, data quality (DQ) has at least as great an impact on predictive models as sample size does. Conventional DQ issues include missing, inconsistent, or replicated data; however, we call at-

© The Author(s) 2020. Published by Oxford University Press on behalf of the American Medical Informatics Association.
All rights reserved. For permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/jamia/article/28/2/360/5919075 by guest on 21 April 2021

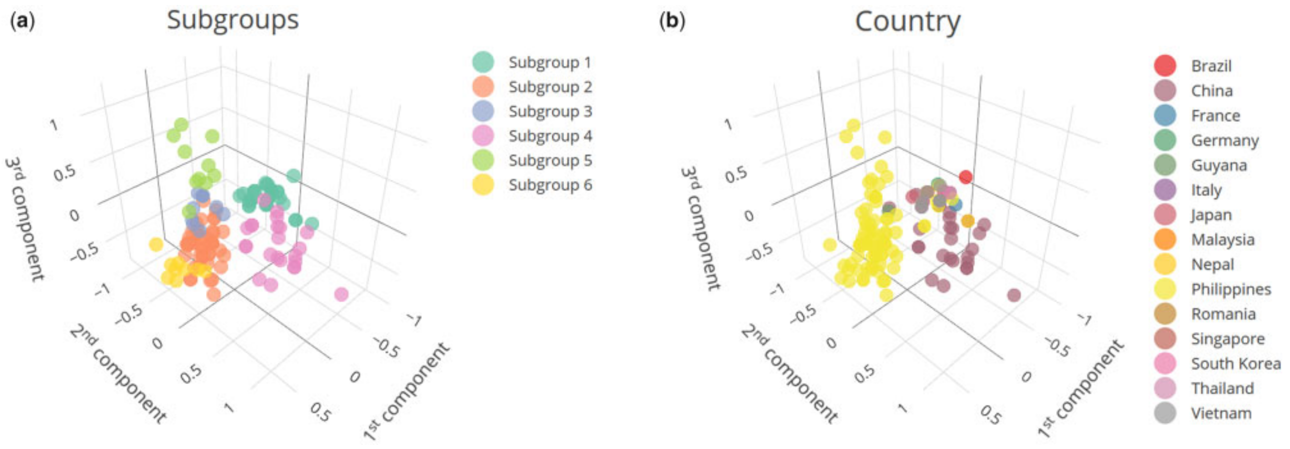

Figure 1. COVID-19 (coronavirus disease 2019) patient subgroups in the nCov2019 dataset, in which the 2 most prevalent countries, China and the Philippines, divided into separate subgroups with distinct severity manifestations. (A) Scatterplot of subgroups embedded by multiple correspondence analysis on 3 dimensions from symptoms and comorbidities. (B) The same scatterplot but labeled by the country of the case. Subgroups 2, 3, 5, and 6 belong to data from the Philippines. Subgroups 4 and 1 mostly represent data from China. Subgroup 1 comprised young patients with mild disease (acute nasopharyngitis) and no comorbidities. Subgroup 2 comprised elderly patients with severe pulmonary disease (pneumonia, acute respiratory distress syndrome) and comorbidities (hypertension, diabetes mellitus, chronic kidney disease). Subgroup 3 comprised middle-aged patients with severe pulmonary disease (pneumonia, acute respiratory distress syndrome) and no comorbidities—similar to subgroup 2, with no remarkable comorbidities. Subgroup 4 comprised elderly patients with mild disease (acute nasopharyngitis) and no comorbidities. The negative outcome within this subgroup might be explained by either poor in-hospital evolution or unreported comorbidities, in which the lack of complete patient data might lie at the root of potential bias. Subgroup 5 comprised elderly patients with severe systemic disease (septic shock, acute kidney injury) and comorbidities. Subgroup 6 comprised elderly patients with severe pulmonary disease (pneumonia) and heart failure due to acute coronary syndrome (most likely diagnosed on admission). For further details, see http://covid19sdetool.upv.es/?tab=ncov2019.

tention to an issue that is particularly harmful to machine learning (ML) and predictive modeling in multisite distributed research networks (DRNs): the variability of data among data sources.

We showcase the potential limitations that multisource variability may have for COVID-19 ML research on large international DRNs. We present our findings in the nCov2019 dataset, recently published in *Scientific Data*.[3]

MATERIALS AND METHODS

The nCov2019 dataset comprises a collection of individual-level COVID-19 epidemiological data publicly available in several countries. In addition to geographic data, it includes, where available, symptoms, outcomes, comorbidities, and key dates. By May 11, 2020, the dataset reported more than 500 000 cases. We included those cases in which at least 1 symptom and an outcome were available.

Our initial aim to help in the fight against COVID-19 was to develop a model to classify patients into severity subgroups given their symptoms, comorbidities, age, and sex. We included those cases in which at least 1 symptom and an outcome were available, reducing the sample to 1568 cases. We next homogenized and filtered the outcome values into "recovered" and "deceased," leading to 214 cases—many outcomes were Twitter posts. Finally, we removed duplicates and homogenized values in comorbidities and symptoms, mapping the latter to International Classification of Diseases–Tenth Revision, leading to a final sample of 170 cases.

To find factors that could help split the population into subgroups relevant to clinical prognosis, we applied a multiple correspondence analysis 3-dimensional embedding[4,5] of symptoms and comorbidities and a hierarchical clustering. The proper number of clusters for both age-independent and age group analyses were selected by supervised inspection of group consistency. The code is available for replication (https://github.com/carsaesi/covid19sdtool).

RESULTS AND DISCUSSION

Findings

Figure 1 describes the results for the age-independent analysis. The resulting subgroups appeared, in general, to be clinically meaningful. What surprised us, however, was that we found remarkable variability between the 2 most prevalent data country sources: China and the Philippines. Their separability was so severe that China and the Philippines were split into distinct subgroups, and consistently for different age groups.

Specifically, subgroups 2, 3, 5, and 6 belong to data from the Philippines, generally elderly patients with a severe disease presentation and comorbidities—except subgroup 3, middle-aged patients. Subgroup 4 mostly represent data from China, with elderly patients with a mild disease presentation. Subgroup 1 combines data from China with cases from other countries, with a mild disease presentation and no comorbidities. Remarkably, only subgroup 1 showed a recovery rate significantly distinct from zero (68.52%; $\alpha = 0.05$), probably related to its younger sample. In the other subgroups, we found no significant difference in survival days after admission, although the survival outcome was mostly missing in the Philippines groups. Full results can be explored online in our COVID-19 Subgroup Discovery and Exploration Tool (http://covid19sdetool.upv.es/?tab=ncov2019).

To investigate the reason for this variability, we checked the "source" variable in the nCov2019 dataset. The source of the Philippines data was a "COVID-19 tracker" from the Department of Health of the Republic of the Philippines.[6] The sources of China data were diverse, but most of them came from patient reports dated January 2020 from the National Health Commission of the People's Republic of China[7] and from a Weixin post by DXY.cn.[8] However, this information proved insufficient to identify a cause for the distinct patterns.

On the one hand, if we were to deliver a severity predictive model based on the current training sample, the population of the

Downloaded from https://academic.oup.com/jamia/article/28/2/360/5919075 by guest on 21 April 2021

Philippines would be assigned only to the high-severity subgroups 2, 3, 5, and 6. On the other hand, Chinese patients would by classified by default in the milder severity subgroups 1 and 4. To evaluate how these sharp differences between data sources could bias a predictive model, the initial question is whether the training data represent the target populations in which the model is to be used.

ML in the presence of data source variability

Multisource variability can potentially limit the optimization of ML model parameters and their generalization. Multiple modes can coexist in the parameters likelihood associated to each data source (eg, the optimum weights of a neural network or the coefficients of a logistic regression can be different for each data source). In addition, the resulting models may poorly generalize to new data due to variability between training data and new data during model use, in form of dataset shifts.[9,10]

In the presence of multisource variability, we could argue for building local predictive models for each data source or, otherwise, building a global model including all the data sources. A local model might better fit its target population according to the training set, but it might have a knowledge gap in the sense of general modeling of the problem. A global model could fill that gap with a wider variable casuistic, ie, covering more combinations of variable values and, particularly, in their outcome-conditional probabilities. However, this approach might have 2 main drawbacks. First, it might lead to a more complex model, eg, requiring more parameters or including a mixed effect for the source to capture all the outcome-conditional probabilities of each source, with the risk of overfitting and loss of generalization. Second, it might lead to a more generalizable model, eg, with a smaller number of parameters, but with the risk of underfitting and losing country-specific performance.

A subgroup analysis like the one we showed for the nCov2019 dataset could be the first step toward making a decision on building a local or a global model. In our case, the difference was evident visually. Additionally, we could measure the dependency between the cluster groups and the source variable. Besides, source variability metrics such as the global probabilistic deviation and source probabilistic outlyingness[11] could help by quantifying the separability between the statistical distributions of the data sources. Therefore, while a reduced separability between sources suggests a global model, a large separability could motivate both approaches. What can ultimately define the best strategy are the results of a model evaluation through an independent test set and cost function that reliably represent how it will perform in its end use. However, this might not be straightforward: the immediate question is if we should use a global test set or local test sets.

In the nCov2019 dataset, we may decide building local models for the Philippines and China, and test them with their local data. However, if a prospective patient in the Philippines presents those patterns of a Chinese one in the current data, or the other way around, the separate local predictive models would be out-of-sample forecasting, providing unreliable answers.

In contrast, we may decide building a global model including all the countries, increasing the sample size and hoping a better fit to a wider prospective variable representation. However, it is difficult to know at the training stage whether learning this extra casuistry will have a benefit or otherwise hinder the model generalization to specific sources. Testing a global model with hold out country specific tests from current data might provide a poorer performance in comparison to a local model. In addition, this might result on different best models for each source.

Against this uncertainty, a timely and proper design of the model evaluation is critical, as well as is planning a continuous monitoring of model performance and updates once in routine use. Last, but not least, in ML modeling sometimes we should consider if it is fairer to stop until more data are available for reliable testing and retraining.

The problem of data variability and heterogeneity in distributed research networks

Multisource variability is closely related to the concepts of statistical frailty or heterogeneity, with recognized potential biases in statistical modeling and interpretation. As noted by Aalen et al,[12] “Heterogeneity often manifests itself as clustering of cases in families more than would be expected by chance.” Cannot be closer to what we described, which is in fact confirmed by the relationship between countries and subgroups.

In addition, a multisource variability problem, as described in our case, is a potential source of socioeconomic or race/ethnicity disparities in predictive modeling. Gianfrancesco et al[13] notably remarked that “algorithms may not offer benefit to people whose data are missing from the data set. [. . .] When an algorithm cannot observe and identify certain individuals, a machine learning model cannot assign an outcome to them. [. . .] As a result, if models trained at one institution are applied to data at another institution, inaccurate analyses and outputs may result.” Similarly, Galvin et al[14] questioned about possible source variability biases in the public health decisions taken by the United States and the United Kingdom to let COVID-19 run its course without intervention, as well as in the suggestions to flatten the curve of cases to spread the hospital resources demand, solely based on cases reported from China and Italy. Recently, the COVID-19 4CE (Consortium for Clinical Characterization of COVID-19 by EHR) DRN found significant variation between countries and between their hospitals regarding laboratory results.[15]

Wynants et al[2] and Sperrin et al[1] argued that their reported findings on poor quality predictive models could easily generalize to other domains than COVID-19. Unfortunately, our experience shows that the potential biases of multisource variability for ML can also be generalized, especially in large cross-border DRNs. In a previous work, we found that predictive models for brain tumor diagnosis and grading trained on multisite European data showed an average 10% increase in error rates when validated on new European datasets, even using the same acquisition protocols, possibly due to uncontrolled variability factors.[16] These factors can generally be attributed to different data formats or coding, dissimilar populations, differences in clinical practice, or observer variability.[11,13,17,18] It follows that even when standardized data formats or protocols are used, variability may persist in the statistical distributions of data.

We recall that, in addition to among sources, data variability can also exist over time, especially when sharing data over long periods. Temporal variability might occur due to, eg, changes in clinical practice or coding.[17,19,20] Temporal dataset shifts set out similar biases than source-related shifts for ML and model generalization: shall we train a model with data from all the available timespan or select recent data? Could a model in routine clinical use become obsolete? Given the rapidly evolving knowledge and practice changes in COVID-19, using change detection methods or tools like the

Downloaded from https://academic.oup.com/jamia/article/28/2/360/5919075 by guest on 21 April 2021

EHRtemporalVariability could be of great benefit to assess temporal variability in COVID-19 DRNs.[20,21]

Preventing biases through DQ and variability assessment and reporting

A routine assessment of the variability among data sources in ML and statistical methodologies could potentially reduce biases or extra costs of unexpectedly discovering heterogeneous subgroups among sources. Because other DQ dimensions can be equally important sources of bias, a complete DQ assessment including source, and also temporal variability, could be an efficient manner of preventing those biases.

Wynants et al[2] and Sperrin et al[1] proposed extending predictive modeling methodological and reporting guidelines, such as TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis),[22] by "recommending when and how it is appropriate to use historical data from similar populations." We also recently claimed for the reporting of DQ, as well as its actual and potential impacts, as a routine practice in reporting the results of data science.[23] DQ reports should be part of ML methodological and reporting guidelines, including TRIPOD or the CRISP-DM (cross-industry standard process for data mining).[24] As described in this work, source variability should be part of these. In addition, a disclaimer about the capability of the model to generalize to newly observed populations would be advised, which could even assist in the development of randomized controlled trials for clinical decision support systems.[25]

Finally, as an alternative approach to curating data for artificial intelligence (AI), we propose establishing new generation AI with built in consciousness about DQ and variability. Both in training and against the observation of new cases, this AI should behave robustly to real world DQ issues (eg, against missing or inconsistent patient information). It should generalize against data variability (eg, to transfer models between locations or reduce their obsolescence through continual learning). An AI in which automatic explainability regarding these DQ and variability might play an important role, and all this could be operationalized on the trending MLOps methodology.[26]

Limitations

The used sample size of 170 cases of the nCov2019 dataset can be a limitation in this study. However, at this size, the difference between the 2 most prevalent countries was evident. Through this specific case study we intend to timely warn and prevent the potential complications of data-source variability for ML, an ongoing problem that can potentially occur in the large multisource COVID-19 datasets being currently collected worldwide.

CONCLUSION

The emergence of COVID-19 international data sharing initiatives and DRNs have a huge potential benefit for COVID-19 research.[15,27,28] The ideal goal is to provide high-quality, homogeneous data, but this is not always straightforward. Using standard formats is the first step toward consistent COVID-19 data representation in multisite settings.[14] However, variability in the statistical distributions among the different sources and over time can potentially persist. As such, COVID-19 DRNs should consider routine variability assessment and reporting, especially if the data are to be used in predictive modeling, given the potential risks of bias.

In conclusion, there is an urgent need to share full population data for COVID-19 research, but just as important is to report on its DQ and, particularly in multisite settings, to report the variability among sources. This will be the only way that the ML community can help this crisis with reliable predicting modeling.

FUNDING

This work was supported by Universitat Politècnica de València contract no. UPV-SUB.2-1302 and FONDO SUPERA COVID-19 by CRUE-Santander Bank grant "Severity Subgroup Discovery and Classification on COVID-19 Real World Data through Machine Learning and Data Quality assessment (SUBCOVERWD-19)."

AUTHOR CONTRIBUTIONS

CS and JMG-G conceived and designed the work. CS and NR performed data collection, processing and analysis. CS, NR, JAC, and JMG-G reviewed and interpreted the results. CS drafted the article. CS, NR, JAC, and JMG-G provided critical revision of the article and approved the final version to be published.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no competing interests.

REFERENCES

1. Sperrin M, Grant SW, Peek N. Prediction models for diagnosis and prognosis in COVID-19. *BMJ* 2020; 369: m1464.
2. Wynants L, Van Calster B, Bonten MM, *et al.* Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. *BMJ* 2020; 369: m1328.
3. Xu B, Gutierrez B, Mekaru S, *et al.* Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci Data* 2020; 7 (1): 106.
4. Hervé A, Williams LJ. Principal component analysis. *WIREs Comput Stat* 2010; 2: 433–59.
5. Husson F, Le S, Pagès J. *Exploratory Multivariate Analysis by Example Using R*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC; 2017.
6. COVID-19 tracker Philippines. 2020. Accessed May 25, 2020.
7. COVID-19 report of the National Health Commission of PRC. 2020. Accessed May 25, 2020.
8. COVID-19 report on. Weixin by DXY.cn. 2020. Accessed May 25, 2020.
9. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. *Pattern Recognit* 2012; 45 (1): 521–30.
10. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019; 380 (14): 1347–58.
11. Sáez C, Robles M, García-Gómez JM. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Stat Methods Med Res* 2017; 26 (1): 312–36.
12. Aalen OO, Valberg M, Grotmol T, Tretli S. Understanding variation in disease risk: the elusive concept of frailty. *Int J Epidemiol* 2015; 44 (4): 1408–21.
13. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018; 178 (11): 1544–7.
14. Galvin CJ, Fernandez-Luque L, Li YC. Accelerating the global response against the exponentially growing COVID-19 outbreak through decent data sharing. *Diagn Microbiol Infect Dis* 2020 May 7 [E-pub ahead of print].
15. Brat GA, Weber GM, Gehlenborg N, *et al.* International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit Med* 2020; 3 (1): 109.

Downloaded from https://academic.oup.com/jamia/article/28/2/360/5919075 by guest on 21 April 2021

16. García-Gómez JM, Luts J, Julià-Sapé M, *et al.* Multiproject-multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. *MAGMA* 2009; 22 (1): 5–18.
17. Sáez C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez JM. Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories. *J Am Med Inform Assoc* 2016; 23 (6): 1085–95.
18. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017; 318 (6): 517–8.
19. Rockenschaub P, Nguyen V, Aldridge RW, Acosta D, García-Gómez JM, Sáez C. Data-driven discovery of changes in clinical code usage over time: a case-study on changes in cardiovascular disease recording in two English electronic health records databases (2001–2015). *BMJ Open* 2020; 10 (2): e034396.
20. Sáez C, Gutiérrez-Sacristán A, Kohane I, García-Gómez JM, Avillach P. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. *GigaScience* 2020; 9 (8): giaa079
21. Sáez C, Rodrigues PP, Gama J, Robles M, García-Gómez JM. Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. *Data Min Knowl Discov* 2015; 29 (4): 950–75.
22. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg* 2015; 102 (3): 148–58.
23. Sáez C, Liaw ST, Kimura E, Coorevits P, Garcia-Gomez JM. Guest editorial: Special issue in biomedical data quality assessment methods. *Comput Methods Programs Biomed* 2019; 181: 104954.
24. Wirth R, Hipp J. CRISP-DM: Towards a standard process model for data mining. In: proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining; 2000: 29–39.
25. Angus DC. Randomized clinical trials of artificial intelligence. *JAMA* 2020; 323 (11): 1043–5.
26. Stenac C, Dreyfus-Schmidt L, Lefevre K, Omont N, Treveil M. *Introducing MLOps*. Newton, MA: O'Reilly Media; 2021.
27. Moorthy V, Henao Restrepo AM, Preziosi MP, Swaminathan S. Data sharing for novel coronavirus (COVID-19). *Bull World Health Organ* 2020; 98 (3): 150.
28. Haendel M, Chute C, Gersing K. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2020 Aug 17 [E-pub ahead of print].

Downloaded from https://academic.oup.com/jamia/article/28/2/360/5919075 by guest on 21 April 2021