Lack of Data Sharing Despite Data Availability Statements in Studies Using Machine Learning Models for Prediction of Gestational Diabetes Mellitus

Mark Germaine,[1,2,3] Graham Healy,[1] and Brendan Egan[2,4]

Diabetes Care 2024;47:e78–e79 | https://doi.org/10.2337/dc24-1483

Recent advancements in artificial intelligence and machine learning (ML) research allow for the mining of electronic health records (EHRs) for predicting health outcomes. One application is that ML models can be developed to predict likelihood of gestational diabetes mellitus (GDM) by using data taken from EHRs obtained early in pregnancy. We have completed preliminary work developing such models using EHR data collected in the first trimester (1). An important feature of ML modeling is the use of an independent data set for external validation to ensure the model's generalizability across different data sets. However, obtaining such a data set has proven challenging and illustrates broader issues regarding data sharing and the implementation of open science principles.

In an attempt to acquire data for external validation, we contacted authors from 22 published articles describing studies that aimed to predict GDM using EHRs (Table 1), and we sought access to a sample subset of their data sets for the purpose of external validation of our model. These studies were identified from a systematic literature search, which we performed up to March 2024, for ML models developed to predict GDM using data from EHRs.

We contacted the respective authors on three separate occasions between 18 April and 17 June 2024. All listed e-mail addresses were contacted simultaneously. The first e-mail detailed the purpose of our request, the importance of external validation for improving model reliability, and assurances regarding data confidentiality. Follow-up emails served as concise reminders of the significance of their contribution to advancing research in ML models for GDM prediction.

Of the 22 articles, 14 had data availability statements indicating that data were available upon request, 3 stated data were not available, and 5 did not have any data availability statement. Despite our efforts, the response rate was unequivocally low. Only one author group (corresponding to two articles) responded positively, expressing a willingness to validate our model independently using their data set, but were unable to share their data directly. Out of the remaining 20 articles, one e-mail address was no longer valid, and another e-mail address elicited an automatic reply but without further response with follow-up emails. Authors from the remaining 18 articles did not provide any response (Table 1).

While it is recognized that the availability of research data declines rapidly with article age (2), the median publication date of the identified articles was 2021. Only 4 out of 22 articles were published prior to 2020. We expected that this recent publication timeframe would result in greater likelihood of data availability.

Difficulties in acquiring data despite the presence of data availability statements is not uncommon. Nonresponses and refusals when attempting to conduct an individual patient data meta-analysis have been previously reported (3). Analysis of data sharing practices in *The BMJ* found that despite a strong data sharing policy, actual sharing rates were low (4). Only 4.5% of the articles shared their data sets, although a higher rate of 24% was observed for articles describing clinical trials. Ambiguous policy wording and a lack of incentives for researchers were identified as being among several barriers to data sharing (4).

Our experience, along with findings above and from others (5), suggests that these data availability statements often do not translate into actual data sharing and highlights two major issues. First, it underscores poor practices around data availability statements. Despite these statements, the sharing of data described in published articles remains inconsistent and unreliable. Second, the lack of data sharing poses a substantial barrier to the external validation of predictive models in ML. Without access to external data sets, it is challenging to ensure the generalizability and robustness of ML models, which will ultimately affect the utility of ML for advancing digital health and artificial intelligence–driven health care.

[1]*School of Computing, Dublin City University, Dublin, Ireland*
[2]*School of Health and Human Performance, Dublin City University, Dublin, Ireland*
[3]*SFI Centre for Research Training in Machine Learning, Dublin City University, Dublin, Ireland*
[4]*Florida Institute for Human and Machine Cognition, Pensacola, FL*

Corresponding author: Brendan Egan, brendan.egan@dcu.ie

Received 18 July 2024 and accepted 25 July 2024

Table 1—Details of studies included in data access inquiries

Article no.	Year	Sample size range	Country	Data availability statement	Author response	Outcome of request
1	2010	<1,000	The Netherlands	N	N	E-mail address no longer valid; NDS
2	2013	1,000–5,000	Vietnam	N	N	NDS
3	2017	1,000–5,000	China	N	N	NDS
4	2017	<1,000	Australia	N	N	NDS
5	2019	>500,000	U.S.	Y	N	Contains "Accessible Data" link, but no data available in repository; NDS
6	2020	>500,000	Israel	Y	N	NDS
7	2020	1,000–5,000	China	Y	N	NDS
8	2020	5,001–50,000	China	Y	N	NDS
9	2021	5,001–50,000	China	Y	N	NDS
10	2021	5,001–50,000	China	Y	N	NDS
11	2021	5,001–50,000	China	Y	N	Automatic reply acknowledging emails, but no further response; NDS
12	2021	1,000–5,000	China	Y	N	NDS
13	2022	1,000–5,000	China	Y	N	NDS
14	2023	50,001–500,000	Japan	Y	N	NDS
15	2023	5,001–50,000	South Korea	DNA	N	NDS
16	2023	<1,000	China	DNA	N	NDS
17	2023	5,001–50,000	Australia	Y	Y	Cannot share data but willing to validate model in own data set; NDS
18	2023	5,001–50,000	Australia	DNA	Y	Cannot share data but willing to validate model in own data set; NDS
19	2023	1,000–5,000	Chile	Y	N	NDS
20	2023	<1,000	China	Y	N	NDS
21	2024	5,001–50,000	China	Y	N	NDS
22	2024	5,001–50,000	China	N	N	NDS

Shown are details of 22 studies, identified by a systematic literature search, which aimed to develop ML models to predict GDM from data in EHRs and to whose authors we sent data access inquiries. DNA, data not available; N, no; NDS, no data shared; Y, yes.

These challenges highlight the need for clearer policies and better incentives to promote data sharing and support the open science movement. This gap between the ideal of open science and the reality of data accessibility emphasizes the need for more robust mechanisms to ensure data availability and to support the reproducibility of scientific findings. To advance the field, it is important to establish more dependable mechanisms for data sharing. This includes reinforcing the commitment of authors and journals to uphold data availability statements in practice as well as developing clearer policies and incentives to promote data sharing. The move toward open science has encouraged the inclusion of data availability statements to promote transparency and reproducibility in research, but the cultural shift toward open data is still evolving, and there remains significant room for improvement.

Funding. This work has emanated from research supported in part by a grant from Science Foundation Ireland under grant number 18/CRT/6183.
Duality of Interest. No potential conflicts of interest relevant to this article were reported.
Author Contributions. M.G. contacted the authors, compiled the data, and wrote the manuscript. G.H. and B.E. contributed to the discussion and reviewed and edited the manuscript. B.E. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Handling Editors. The journal editors responsible for overseeing the review of the manuscript were Steven E. Kahn and Matthew J. Crowley.

References

1. Germaine MA, O'Higgins AC, Healy G, Egan B. 1968-LB: Early prediction of gestational diabetes mellitus using electronic health records and machine learning. Diabetes 2024;73(Supplement_1):1968-LB
2. Vines TH, Albert AYK, Andrew RL, et al. The availability of research data declines rapidly with article age. Curr Biol 2014;24:94–97
3. Jaspers GJ, Degraeuwe PLJ. A failed attempt to conduct an individual patient data meta-analysis. Syst Rev 2014;3:97
4. Rowhani-Farid A, Barnett AG. Has open data arrived at the British Medical Journal (BMJ)? An observational study. BMJ Open 2016;6:e011784
5. Obels P, Lakens D, Coles NA, Gottfried J, Green SA. Analysis of open data and computational reproducibility in registered reports in psychology. Adv Methods Pract Psychol Sci 2020;3:229–237