Data Quality, Data Sharing, and Moving Artificial Intelligence Forward

Joann G. Elmore, MD, MPH,

Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California

Christoph I. Lee, MD, MS, MBA

Department of Radiology, University of Washington School of Medicine, Seattle

Buda et al[1] have curated and annotated a data set of 3-dimensional digital breast tomosynthesis (DBT) examinations obtained from 5060 patients. In using this data set, they developed a deep learning algorithm for breast cancer detection and reached a sensitivity of 65% at 2 false positives per breast on a test set from 418 patients. Compared with the reported performance of several commercial artificial intelligence (AI) products for mammography,[2] the performance of their model is not great. However, tasking AI to detect breast cancer in DBT examinations, in comparison with 2-dimensional digital mammograms, remains notoriously challenging. The large amounts of imaging data produced from DBT contributes to the already complex interpretive task of both radiologists and AI algorithms, yet the additional image-based data may theoretically provide more opportunities to detect meaningful cancers.

While AI holds great promise to improve detection and efficiency, it also requires large amounts of data to be properly trained and tested. Historically, development and evaluation of these algorithms have been hindered by a lack of well-annotated, large-scale, publicly available data sets. Despite organizational proposals and guidance on data sharing, medical data has not been shared to a degree that can "trigger the expected data-driven revolution in precision medicine."[3] Buda et al[1] are bucking the trend and making their annotated image data set publicly available, including their experiments' full code and network architecture with model weights. The authors are to be commended for their scientific spirit and what we see as a sign of forward progress: scientists sharing data and code to advance the field of AI in medicine.

Details, and thus data quality, matter in research. For those not involved in the generation and collection of shared medical data, it may be difficult to understand the choices made in defining cohorts. These choices make documentation a key aspect to quality data sharing. However, documentation of this caliber also requires time and attention to detail. In this instance, the newly public data from Buda et al[1] would be more helpful to future investigators if additional information on the involved cases were made available. As experienced investigators in breast imaging data collection, quality control, and analysis,

Corresponding Author: Joann G. Elmore, MD, MPH, Department of Medicine, David Geffen School of Medicine at UCLA, 1100 Glendon Ave, Ste 900, Los Angeles, CA 90024 (jelmore@mednet.ucla.edu).

we identified important questions concerning their description of the database and the possibility that it is not fully characteristic of a screening DBT cohort. Investigators should be aware of the following limitations before fully embracing this new data set:

1. Without at least 1-year follow-up for presentation of interval cancers, the authors do not adequately describe the longitudinal follow-up of this cohort required to determine whether any imaging examination was conferred a false-negative result. Devoid of adequate cancer follow-up, and relying solely on a radiologist's human interpretation, this can mislead the ground truth used for algorithm training and testing.

2. DBT cases in the study dropped from 16 802 to 5610 cases, a significant number of exclusions that could bias the remaining set of cases. Description of the patient characteristics and distribution of Breast Imaging-Reporting and Data System assessments for these examinations could inform us of whether the cases are representative of a screening population. Moreover, while the authors attempted to exclude diagnostic DBT examinations by excluding those with compression views, it is still likely that this data set includes both screening and diagnostic imaging exams. It would have been more appropriate to include examinations with only a screening clinical indication.

3. The authors did not include any DBT screening examinations for which a diagnostic evaluation was requested due to calcifications. While most malignant calcifications are determined to be ductal carcinoma in situ rather than more aggressive invasive cancers, leaving out cases of suspicious calcifications make this a nonrepresentative data set of a true screening population, in which a significant proportion of callbacks from screening are due to calcifications. This further alters the composition and usability of the data set.

4. Finally, the open access to this small data set brings up the issue of patient privacy concerns and the ethics of sharing patients' medical image data with those who stand to potentially benefit from future commercial development of algorithms using these images. While unlikely that individual women could be identifiable from their DBT examinations, it is unclear whether informed consent should be obtained in this and future studies.[4]

Although the study by Buda et al[1] does not exceed the performance of already available AI algorithms for screening mammography, the positive outcome remains their attempt to openly share data. However, data sets made public must be of better quality and representative of a screening population to be truly useful. Future models will otherwise risk being trained and tested on the wrong ground truth. The quality of data and implications of sharing such information are important questions to consider as we merge shared data science and AI into medical imaging.

Conflict of Interest Disclosures:

REFERENCES

1. Buda M, Saha A, Walsh R, et al. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. JAMA Netw Open. 2021;4(8):e2119100. doi:10.1001/jamanetworkopen.2021.19100 [PubMed: 34398205]

2. Salim M, Wahlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol. 2020;6(10):1581–1588. doi:10.1001/jamaoncol.2020.3321 [PubMed: 32852536]

3. Blasimme A, Fadda M, Schneider M, Vayena E. Data sharing for precision medicine: policy lessons and future directions. Health Aff (Millwood). 2018;37(5):702–709. doi:10.1377/hlthaff.2017.1558 [PubMed: 29733719]

4. Kotsenas AL, Balthazar P, Andrews D, Geis JR, Cook TS. Rethinking patient consent in the era of artificial intelligence and big data. J Am Coll Radiol. 2021;18(1 Pt B):180–184. doi:10.1016/j.jacr.2020.09.022 [PubMed: 33413897]