The AI Generalization Gap: One Size Does Not Fit All

Merel Huisman, MD, PhD • Gerjon Hannink, PhD

Merel Huisman, MD, PhD, is a radiologist at Radboud University Medical Center in Nijmegen, the Netherlands. Her clinical subspecialty interest is cardiothoracic and musculoskeletal radiology. She is a clinical epidemiologist by training, a European Society of Medical Imaging Informatics board member, and involved in several national and international initiatives concerning artificial intelligence in health care.

Gerjon Hannink, PhD, is a clinical epidemiologist and biostatistician at Radboud University Medical Center in Nijmegen, the Netherlands. His research interests focus on clinical prediction models, artificial intelligence, and medical decision-making.

To ensure safe, effective, equitable, and trustworthy deployment, diagnostic artificial intelligence (AI) tools must perform robustly across the target population at their initial use and over time (1). Robust model performance is a complex and multifaceted challenge that requires careful consideration of various highly context-dependent factors. The AI research and development community is beginning to recognize and address this challenge more comprehensively. Increasingly, external testing studies are being conducted that aim to bridge the existing generalization gap, and ultimately, the implementation gap (2). The goal is to establish standardized methodologies and guidelines that can help overcome the challenges associated with the limited applicability of AI tools and ensure their broader effectiveness and reliability.

One important step toward robust deployment is subgroup analysis. This analysis can reveal hidden stratification, providing insight into the inherent weaknesses of the algorithm that may adversely affect subgroups and make a certain tool unfit for a certain clinical context. In this issue of *Radiology: Artificial Intelligence*, Ahluwalia et al (3) show that the four tested chest radiograph classifiers are unfit for their specific setting to accurately identify subtle findings and merely serve as a "gross abnormality detector." The group performed a pragmatic study illustrating the subgroup problem in, to our knowledge, the largest comparative external testing study to date. One classifier was a proprietary third-party classifier, while the rest were open source and trained on well-known, large, publicly available datasets (ie, CheXpert, MIMIC-CXR, and Chest X-ray-14). The classifiers were tested on a large dataset of chest radiographs consecutively collected at their regional teaching hospital from patients with temporally, geographically, and demographically different makeup (n = 197 540 adults with a posteroanterior chest radiograph; age range, 18–105 years; 50% males; 67% with ancestry from "greater Europe"). The primary outcome was binary (normal or abnormal) with algorithm composition and operating thresholds "as-is"; for example, the overall prevalence of abnormalities was 49% of cases versus 91% in CheXpert.

Subgroup analysis was conducted by patient setting (eg, emergency), age categories, sex, and name-based ancestry. A pragmatic, natural language processing–based, semisupervised method to establish ground truth was chosen and locally validated. Overall sensitivity ranged from 50% to 72%; the third-party classifier had the highest overall performance. We could not find the intended performance of the studied classifiers in the article nor in the literature, but we assume that the overall performance as reported by Ahluwalia et al (3) was considerably less than anticipated.

Subgroup analysis revealed weak performance of the classifiers for solitary findings, younger patients (<40 years), and emergency settings, with absolute sensitivity drops of 27%, 33%, and 12%, respectively, that could not be attributed to calibration differences. Figure 2 of the article nicely depicts performance shifts per subgroup; such information is rarely reported, especially regarding ethnicity (4), but is highly insightful. Significant sensitivity drops were reported in females (-3%) and those with names of African/Indian ancestry (-6%). These drops likely represent movements along the receiver operating characteristic curve, which could lead to underdiagnosis in these groups if models are deployed when uncalibrated.

The results clearly disqualify the tested tools for screening tasks such as triage in all patients, but especially in the scenario of younger patients presenting to the emergency department with one abnormality (eg, pneumothorax). This represents a critical insight, as filtering out normal studies is currently considered a popular scenario for deployment of chest radiograph detection tools. A possible part of the explanation is that the current dataset contains relatively healthier patients than the open-source databases the tools were trained on. Studies like this are needed to define the suitable context for decision-support tools, especially when stand-alone deployment is intended.

From Radboudumc, Oudwijk 49, Nijmegen, Utrecht 6500, the Netherlands. Received July 6, 2023; revision requested July 10; revision received July 12; accepted August 10.
Address correspondence to M.H. (email: *merel.huisman@radboudumc.nl*).

Authors declared no funding for this work.

Conflicts of interest are listed at the end of this article.

See also article by Ahluwalia et al in this issue.

Ethically sound deployment of AI tools has been a topic of much debate even outside the imaging space. A famous example is the Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS, risk assessment tool used by the U.S. criminal justice system to predict the likelihood of reoffense based on several personal factors. The tool turned out to be biased against Black defendants, especially females (5), and is no longer used in several locations. The same phenomenon has been shown in radiology, and the topic has gained increasing attention, as the relatively homogeneous background of developers and authors may (subconsciously) influence algorithm generalizability (6). Recent efforts in breast radiology have established a benchmark dataset that is racially diverse to obtain unbiased performance (7). One hopes that other subspecialty areas will follow.

The pursuit of generalizability can lead to applications that sacrifice strong performance at individual sites or specific patient populations in favor of applications with mediocre or poor performance across multiple sites or subgroups of patients. Clinicians and researchers face a trade-off between improving system performance locally or for specific subgroups and having systems that can generalize across different contexts.

In the context of buying AI solutions, there is no one-size-fits-all solution, and the adage "buyer beware" is applicable. Therefore, it is important for the end user to be aware of possible performance shifts in certain subgroups that were underrepresented in the development data and to realize that performance monitoring over time is a necessity to avoid degrading performance. The end user, the radiologist or clinician, needs to decide what serves patients best, and it is the vendor's responsibility to be transparent about model development and validation. To that end, electronic health record vendors should enable easily accessible demographics and general disease characteristics of the target population to enable decision-making regarding the translatability of a certain AI tool to the intended setting. For vendors to be able to reach their commercial goals, it is important that they understand the granularity of the health care context to better steer business efforts and avoid technical push (ie, products being developed based on technical opportunities rather than a clinical need).

To alleviate some of the issues encountered with underrepresented patient groups in the data, domain adaptation and/or transfer learning may be used. In medical imaging, domain adaptation refers to the process of adjusting machine learning models trained on medical image data from one domain (eg, specific hospital or imaging center) to perform well on a different but related domain. Due to variations in imaging protocols, equipment, patient populations, or demographics, there can be differences in the data distributions between these domains, known as domain shift (or dataset shift).

The objective of domain adaptation techniques is to address these domain differences and enhance model generalizability across diverse imaging settings, aiming to leverage the knowledge gained from the source domain to improve model performance in the target domain. Recent advances in domain adaptation methods have been explored extensively in the field of medical imaging analysis. These advancements encompass a range of approaches, including feature alignment, image transformation, unsupervised domain adaptation, and the use of generative adversarial networks (8).

Transfer learning is a broader concept that involves domain adaptation and refers to the process of transferring knowledge or learned representations from one task or domain to another (9). While domain adaptation specifically deals with adapting models to different domains, transfer learning can involve transferring knowledge between tasks within the same domain or across different domains.

Another known tool in AI that can help assure broad generalizability of AI models is stress testing (10). Stress tests help determine if a model will perform consistently when applied to different datasets. It aims to ensure that the AI application can handle unexpected scenarios, large volumes of data, or adverse conditions without breaking down or producing unreliable results. These stress tests involve introducing artificial shifts or examining the model's predictions in specific subsets of the testing data. However, it is not feasible to create stress tests for every possible shift, especially in radiology, where various factors like image acquisition parameters and patient characteristics can result in a wide range of potential changes. Ideally, stress tests should be carefully designed to replicate specific shifts that can cause the model to fail. In addition to considering how well the model performs its intended task, researchers must also anticipate the conditions in which the model will be used to ensure stable performance. By subjecting the system to challenging conditions, stress testing helps uncover potential issues with, for example, shifted data, subgroups, or noise, artifacts, and different acquisition parameters, and allows developers to address them proactively, leading to more reliable and trustworthy AI applications.

Limitations of the study by Ahluwalia et al (3) are well addressed and include semisupervised natural language processing–based labeling techniques and using the report instead of the image as a basis for the reference standard. Also, only a selected number (n = 12) of abnormalities was investigated. Notwithstanding, we believe that the most important take-away of this study is its demonstration of the overarching principle of inconsistent subgroup performance in a large-scale dataset.

In summary, the issues related to generalizability of AI applications calls for a collaborative effort between researchers, clinicians, and vendors to ensure that these technologies are thoroughly validated and carefully implemented to maximize their benefits and minimize potential risks. As the field of AI in radiology continues to grow and evolve, it is essential that the scientific community develops a rigorous, standardized, and evidence-based approach to ensure ethically sound and tailored deployment.

Disclosures of conflicts of interest: M.H. Speakers honoraria from DeepC, Bayer, and MedicalPhit; leadership or fiduciary role with EuSoMII, ECR scientific subcommittee, FMS (Dutch), and *Radiology: Artificial Intelligence* trainee editorial board advisory panel (all unpaid). **G.H.** No relevant relationships.

References

1. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018;378(11):981–983.

2. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: A systematic review. Radiol Artif Intell 2022;4(3):e210064.
3. Ahluwalia M, Abdalla M, Sanayei J, et al. The subgroup imperative: Chest radiograph classifier generalization gaps in patient, setting, and pathology subgroups. Radiol Artif Intell 2023; 5(5):e220270.
4. Driessen R, Bhatia N, Gichoya JW, Safdar NM, Balthazar P. Sociodemographic variables reporting in human radiology artificial intelligence research. J Am Coll Radiol 2023;20(6):554–560.
5. Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. Sci Adv 2018;4(1):eaao5580.
6. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health 2022;4(6):e406–e414.
7. Jeong JJ, Vey BL, Bhimireddy A, et al. The Emory Breast Imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. Radiol Artif Intell 2023;5(1):e220047.
8. Guan H, Liu M. Domain adaptation for medical image analysis: A survey. IEEE Trans Biomed Eng 2022;69(3):1173–1185.
9. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. BMC Med Imaging 2022;22(1):69.
10. Eche T, Schwartz LH, Mokrane FZ, Dercle L. Toward generalizability in the deployment of artificial intelligence in radiology: Role of computation stress testing to overcome underspecification. Radiol Artif Intell 2021;3(6):e210097.