Perspectives

Clinical Cancer Research

Data Science in Radiology: A Path Forward

Hugo J.W.L. Aerts

Abstract

Artificial intelligence (AI), especially deep learning, has the potential to fundamentally alter clinical radiology. AI algorithms, which excel in quantifying complex patterns in data, have shown remarkable progress in applications ranging from self-driving cars to speech recognition. The AI application within radiology, known as radiomics, can provide detailed quantifications of the radiographic characteristics of underlying tissues. This information can be used throughout the clinical care path to improve diagnosis and treatment planning, as well as assess treatment response. This tremendous potential for clinical translation has led to a vast increase in the number of research studies being conducted in the field, a number that is expected to rise sharply in the future. Many studies have reported robust and meaningful findings; however, a growing number also suffer from flawed experimental or analytic designs. Such errors could not only result in invalid discoveries, but also may lead others to perpetuate similar flaws in their own work. This perspective article aims to increase awareness of the issue, identify potential reasons why this is happening, and provide a path forward. *Clin Cancer Res; 24(3); 532–4. ©2017 AACR.*

Breakthroughs in artificial intelligence (AI) have the potential to fundamentally alter medical image analysis as well as the clinical practice of radiology. These methods excel at identifying complex patterns in images and in using this information to guide clinical decisions (1–3). AI encompasses quantitative image analysis, also known as radiomics (4–9), which involves either the application of predefined engineered algorithms (that often rely on input from expert radiologists) or the use of deep learning technologies that can automatically learn feature representations from example data (4). Consequently, AI is expected to play a key role in automating clinical tasks that presently can only be done by human experts (1, 2). Such applications may aid the radiologist in making reproducible clinical assessments, thereby increasing the speed and accuracy of radiologic interpretation, and help the reader in situations difficult for human observers to interpret, such as in predicting the malignancy of a particular lesion or response to a particular therapy based on the patient's total tumor burden (10–12).

The potential of AI has resulted in an explosion of investigations that utilize various applications of data science to further radiologic research. The magnitude of this transformation is reflected in the large number of research studies published in recent years. Numerous articles have been published describing the automated detection of abnormalities (also known as CADe; refs. 13–15), others the automated quantification of diseases (also known as CADx; refs. 16–19), and still others that link radiologic data with genomic data (also known as imaging–genomics or radiogenomics) to define genotype–phenotype interactions (6, 20–24). With promising results from these early studies and the increasing availability of imaging data (including large retrospective datasets with clinical endpoints), we expect radiomic investigations to continue to grow rapidly in number and complexity in the coming years.

Many examples of robust discoveries have emerged from studies with stringent analytic and experimental designs. However, a number of studies, including many published in high-impact journals, suffer from (avoidable) flaws in experimental design or analytic methodology, which could potentially invalidate the findings. Among the most egregious examples are studies that employ imaging datasets that are too small to power a significant finding, for example, hundreds of parameters are evaluated but in only a couple dozen samples. Others include analyses that lack independent validation and present models that are trained and validated on the same data. Still, others suffer from "information leakage" due to improper separation of training and validation datasets. A common example of this would be when the features are selected from the same data used to evaluate performance. Errors are also being made in statistical analyses, such as improper correction for multiple testing (or a failure to correct at all), which can lead to overoptimistically low *P* values, or the reporting of incorrect statistical outcome measures. Such studies give rise to findings that cannot be replicated, ultimately weakening the perception of data science applications in radiology and threatening the credibility of other investigations in the field.

These problems occur for a plethora of reasons, ranging from inadequate training to inadequate example studies to inadequate reporting and sharing. First, many researchers working in the medical imaging and radiology field have little or no formal background in data analysis and biostatistics. This gap in understanding creates a blind spot in which investigators fail to recognize what is required for good study design or what methods can be used most appropriately to arrive at sound analytic results with a high likelihood of being replicated in an independent study. While humans are capable of understanding

Harvard Medical School, Dana-Farber Cancer Institute, and Brigham and Women's Hospital, Boston, Massachusetts.

Corresponding Author: Hugo J.W.L. Aerts, Harvard–Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02115. Phone: 617-525-7156; Fax: 617-582-6037; E-mail: Hugo_Aerts@dfci.harvard.edu

doi: 10.1158/1078-0432.CCR-17-2804

Translational Relevance

The future of radiology is bright, as novel artificial intelligence (AI) algorithms have the potential to increase the clinical efficacy and quality of radiology, while making it faster and cheaper. Indeed, a large number of research groups are investigating how AI can improve radiologic quantifications, and several compelling discoveries have already been made. Unfortunately, for a variety of reasons, there has also been a large increase in studies with serious experimental design flaws. Prime examples include the use of small datasets that are underpowered or lack independent validation, and flawed statistical analyses such as improper multiple testing correction. These errors can lead to invalid discoveries that are not replicable and only serve to weaken the perception of the field, the credibility of its investigations, and perhaps even slow the clinical introduction of new technologies.

a small number of imaging features in correlation with a limited number of pathologic findings, the thousands of sometimes non-intuitive imaging parameters current technology is capable of extracting from imaging data, compounded by the complex relationships that exist between them, require the use of more sophisticated analytical methods.

Furthermore, many of the journal editors and reviewers assigned to evaluate and critique these scientific studies are experts in radiology but have no expertise in data analysis methods. Consequently, potential mistakes in the analytic pipeline may go unnoticed during the editorial and review process, resulting in publications with findings that may not be fully supported or reproduced by the data. This scenario sets up a vicious cycle in which other readers also fail to recognize experimental or analytic flaws, mistake the reported results for truth, and repeat the methodologic errors in their own work. Indeed, a number of proof-of-concept radiomic studies with methodologic shortcomings have been published, and one can now see these same errors repeated by others.

As is true for other scientific fields, current mechanisms for correcting published studies are inadequate. Although study flaws in published works may be quickly recognized by investigators with quantitative training or experience, those studies are rarely publicly challenged. There are few incentives to call into question the results of published articles, as doing so can cause great animosity. Nevertheless, there are some positive examples where a careful review of published studies can help us understand how to do better science. Chalkidou and colleagues (25) performed an independent reanalysis of several published imaging studies and found that the majority had significant issues with the experimental design, potentially leading to incorrect results. It is important for the community to take notice of and support reanalyses such as these and to recognize their value in advancing our science.

A straightforward approach to advance the quality of analytics in quantitative imaging is to create a culture in which there is a willingness to share primary data and software code. Much can be done to stimulate this culture by making data sharing a requirement for publication. Indeed, the technical aspects of sharing radiologic data are often feasible, and initiatives exist that can support investigators with this process, such as the Cancer Imaging Archive (26). Proper sharing assures that the results of any single study can be recapitulated, as other investigators can test the reproducibility of the main findings. It also facilitates rapid development of new, more robust methods as more data become available for integrated analyses.

As is true in other fields, such as genomics, editors and reviewers should demand the publication of all data (including medical images), code, and results to ensure full reproducibility. This level of disclosure is consistent with the guidelines of many scientific journals for other types of study and reflects the NIH's requirements for data sharing and reproducible research. Integrating these "best practices" into quantitative imaging will help assure the quality and reliability of our studies and will increase the strength and influence of our work, as others use and cite our data and software.

As the saying goes, "the devil is in the details." This is especially true for the field of data science where confounding errors are easy to generate, but hard to identify, and require expertise and experience to identify and resolve. The most important steps investigators must take are to acknowledge their limitations, know when to ask for external expert advice, and recognize that a poorly designed and analyzed study is of little value in the long run. Better awareness and education in data science and statistical methodologies will increase the overall quality of discoveries within radiology.

It is also important to establish analysis guidelines to avoid pitfalls and provide recommendations for analysis strategies for medical imaging data. Guidelines related to data acquisition, data normalization, development of robust features and models, and rigorous statistical analyses will also increase the quality of these studies and allow for better evaluation of imaging data with other data sources such as clinical and genomic data.

Although the points raised here may seem overly critical of radiology research, this is not the first field to face such challenges. The most direct example in my view is the field of genomics, where early studies were underpowered, poorly analyzed, and nonreplicable. As faith in genomic assays began to wane, the community came together and recognized the need for better standards for experimental design, reproducible research, data and code sharing, and good analytic practices (27–35). The widespread institution of these practices by academic leaders and scholarly journals has led to genomic assays that are far more reproducible between laboratories, a necessity for potential clinical application.

Fortunately, there is growing appreciation of these issues and of the importance of better training in quantitative methodologies. Data science is becoming an important subject at leading radiology and image analysis conferences, and educational seminars are stimulating learning and the acquisition of new skills. It is likely the knowledge base will continue to increase for both investigators and editors, improving the overall quality of new research studies.

If we do this right, and keep on emphasizing the importance of data science training, I believe our field has a bright future. We will never rid ourselves of all our mistakes; however, we can, if we avoid the major pitfalls, improve our credibility. This will not only lead to good science, but also could ultimately reshape clinical radiology practice. Most important, it will lead to improved treatment and better outcomes for the patients we serve.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We acknowledge financial support from the NIH (U24CA194354 and U01CA190234) and would like to thank Dr. Lawrence H. Schwartz and Dr. John Quackenbush for their insightful and helpful comments.

Received September 25, 2017; revised October 20, 2017; accepted October 31, 2017; published OnlineFirst November 2, 2017.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
2. Rusk N. Deep learning. Nat Methods 2015;13:35–35.
3. Lewis-Kraus G. The Great AI Awakening. New York Times. 2016 Dec 14.
4. Aerts HJWL. The potential of radiomic-based phenotyping in precision medicine: a review. JAMA Oncol 2016;2:1636–42.
5. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278:563–77.
6. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006.
7. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. Phys Med Biol 2016;61:R150–66.
8. Shaikh FA, Kolowitz BJ, Awan O, Aerts HJ, von Reden A, Halabi S, et al. Technical challenges in the clinical application of radiomics. JCO Clinical Cancer Informatics 2017;1–8. Available from: http://ascopubs.org/doi/abs/10.1200/CCI.17.00004
9. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441–6.
10. Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. JAMA 2016;316:2353–4.
11. Beam AL, Kohane IS. Translating artificial intelligence into clinical care. JAMA 2016;316:2368–9.
12. Liu Y, Balagurunathan Y, Atwater T, Antic S, Li Q, Walker RC, et al. Radiological image traits predictive of cancer status in pulmonary nodules. Clin Cancer Res 2017;23:1442–9.
13. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35:1285–98.
14. Summers RM. Deep learning and computer-aided diagnosis for medical image processing: a personal perspective. In: Lu L, Zheng Y, Carneiro G, Yang L, eds. Deep learning and convolutional neural networks for medical image computing. Advances in computer vision and pattern recognition. Cham, Switzerland: Springer; 2017. p. 3–10.
15. Firmino M, Morais AH, Mendoça RM, Dantas MR, Hekis HR, Valentim R. Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects. Biomed Eng Online 2014;13:41.
16. Amir GJ, Lehmann HP. After detection: the improved accuracy of lung cancer assessment using radiologic computer-aided diagnosis. Acad Radiol 2016;23:186–91.
17. Gupta S, Chyn PF, Markey MK. Breast cancer CADx based on BI-RADS™ descriptors from two mammographic views. Med Phys 2006;33:1810–7.
18. Gruszauskas NP, Drukker K, Giger ML. Robustness studies of ultrasound CADx in breast cancer diagnosis. In: Suzuki K, ed. Machine learning in computer-aided diagnosis: medical imaging intelligence and analysis. Hershey, PA: IGI Global; 2012. p. 1–22.
19. Lo S-CB, Freedman MT, Kinnard L, Makariou E. Mammographic CADx system using an image library with an intelligent agent: A pattern matching approach. In: Proceedings of Medical Imaging 2006: Image Processing; 2006 Feb 11–16; San Diego, CA. Bellingham, WA: SPIE; 2006.
20. O'Connor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol 2017;14:169–86.
21. Rios Velazquez E, Parmar C, Liu Y, Coroller TP, Cruz G, Stringfield O, et al. Somatic mutations drive distinct imaging phenotypes in lung cancer. Cancer Res 2017;77:3922–30.
22. Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, et al. Defining the biological basis of radiomic phenotypes in lung cancer. Elife 2017;6:pii:e23421.
23. Gutman DA, Dunn WD Jr, Grossmann P, Cooper LAD, Holder CA, Ligon KL, et al. Somatic mutations associated with MRI-derived volumetric features in glioblastoma. Neuroradiology 2015;57:1227–37.
24. Aerts HJWL, Grossmann P, Tan Y, Oxnard GG, Rizvi N, Schwartz LH, et al. Defining a radiomic response phenotype: a pilot study using targeted therapy in NSCLC. Sci Rep 2016;6:33860.
25. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. PLoS One 2015;10:e0124165.
26. Prior F, Smith K, Sharma A, Kirby J, Tarbox L, Clark K, et al. The public cancer radiology imaging collections of The Cancer Imaging Archive. Sci Data 2017;4:170124.
27. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365–71.
28. Berrar DP, Dubitzky W, Granzow M. A practical approach to microarray data analysis. Springer Science & Business Media; 2007.
29. Stekel D. Data standards, storage and sharing. In: Microarray Bioinformatics. Cambridge, United Kingdom: Cambridge University Press; 2003. p. 231–52.
30. Quackenbush J. Microarray data normalization and transformation. Nat Genet 2002;32:496–501.
31. Quackenbush J. Computational genetics: computational analysis of microarray data. Nat Rev Genet 2001;2:418–27.
32. Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, et al. A concise guide to cDNA microarray analysis. Biotechniques 2000;29:548–50.
33. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7:55–65.
34. Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci U S A 2000;97:9834–9.
35. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010;11:733–9.