AMERICAN ACADEMY
OF OPHTHALMOLOGY®

Editorial

The Growing Need for Ophthalmic Data Standardization

Yusrah Shweikh, MBChB, FRCOphth - *Boston, Massachusetts; Sussex, United Kingdom*
Sayuri Sekimitsu, BS - *Boston, Massachusetts*
Michael V. Boland, MD, PhD - *Boston, Massachusetts*
Nazlee Zebardast, MD, MSc - *Boston, Massachusetts*

It is estimated that data related to health care account for approximately one third of all data in existence globally.[1] Data availability has catalyzed medical research with notable examples in ophthalmology, focused largely on common conditions, including glaucoma, diabetic retinopathy, cataracts, and age-related macular degeneration. Health care data are derived from a variety of sources and may be prospectively collected for specific research studies, collected as part of routine clinical practice, or generated as a by-product of data from industries external but related to health care. Different sources of data certainly have distinct utility and limitations (Table 1). Many factors challenge the use of clinical data for research, including the following: (1) the accuracy and completeness of entered data; (2) adherence to available data standards; (3) accessibility of data in the face of governance and technological barriers; and (4) systematic biases in patient demographics and data entry.

The current lack of centralized data repositories and insufficient adoption of harmonized data structures hinder data sharing and limit the potential for large-scale research in ophthalmology.

Navigating data resources can be time-consuming, and it is difficult for researchers to be certain that the data selected are the best available to test individual hypotheses. The current lack of centralized data repositories and insufficient adoption of harmonized data structures hinder data sharing and limit the potential for large-scale research in ophthalmology. Here, we highlight the need for standards to facilitate high-quality research and collaboration.

One challenge facing ophthalmic research lies in the effective collation of data to permit large-scale analyses. Data standards are source-dependent and vary considerably in eye care. Between institutions, electronic health record (EHR) data can be aggregated using a data model implemented by a shared clinical information system, but sharing data in the absence of a shared system vendor is much more challenging because of differences in the representation of clinical data in each EHR. Even within one institution or shared clinical information system, the same data may be stored in different locations or at different levels of granularity. It is also apparent that the utility of data in eye care is restricted by the accuracy and completeness of data labeling for diagnoses and outcomes. There are inherent difficulties in achieving high-quality labeling, including the availability of suitably skilled personnel, implementing the necessary quality control measures, and meeting the associated costs. As an example, the American Academy of Ophthalmology Intelligent Research in Sight Registry (https://www.aao.org/iris-registry) is the first national comprehensive eye disease clinical registry. While aggregating large numbers of patients and encounters, Intelligent Research in Sight has not overcome the challenges of harmonizing data across practices and suffers from the usual problems of differences in EHR data semantics and inaccurate data entry. Issues with labeling, shareability, and the observed variation in ophthalmology data reporting hinders researchers in the tasks of data acquisition and dataset aggregation.

Internationally agreed upon data standards establishing harmonized data structures and minimum requirements to be satisfied before the publication of datasets would substantially decrease the burden of data acquisition and merging going forward. As an example, in the United States, there is not a single central rule about data sharing, which has led to fragmented decision making by states or institutions. The National Institutes of Health have established a Data Management and Sharing Policy, effective January 25, 2023, to share scientific data (including "omics," imaging, and biological data) supporting a manuscript at the time of publication. However, this policy does not require a specific protocol for data standardization, instead leaving it to the discretion of the individual researcher. In the United Kingdom, while increasing access to health data is listed as part of their National Data Strategy and Data Protection and Digital Information Bill 2022−2023, this bill has only been introduced to Parliament at this time. Steps to harmonize data internationally will also form part of the broad approach needed to aid the clinical translation of machine learning systems currently confined to the research domain, in part due to concerns surrounding transparency, privacy, and reproducibility.[2]

One approach to overcome differences in source data is to remap those data into a standard model. One example has been developed by the Observational Health Data Science and Informatics (www.ohdsi.org), which has put forward the Observational Medical Outcomes Partnership (OMOP) common data model to enable systematic analyses of multiple observational databases.[3] As an example, the National Institutes of Health All of Us Research Program (https://allofus.nih.gov) utilizes OMOP to standardize EHR data

https://doi.org/10.1016/j.xops.2022.100262
ISSN 2666-9145/23

Table 1. Types of Clinical Data Used in Ophthalmology Research

Data Type	Utility	Limitations	Examples
Research data			
Population-based studies	Analyzing associations between variables and outcomes, epidemiology, public health	• Limited sample diversity with predominantly White participants • Often only allows for cross-sectional analyses • Access may be restricted for researchers external to host institution(s)	**Publicly available:** UK Biobank, National Health and Nutrition Examination Survey (NHANES), NIH All of Us Research Program **Not publicly available:** Rotterdam Eye Study, Baltimore Eye Study, Los Angeles Latino Eye Study
Aggregated data	Analyses specific to clinical uses or conditions	• Require data entry and maintenance • Nonstandardized data formats among institutions • Eye-specific registries may lack systemic diagnosis and medication information	Intelligent Research in Sight (IRIS), the European Registry of Quality Outcomes for Cataract and Refractive Surgery (EUREQUO), the Japan-Retinal Detachment Registry, the Swedish National Cataract Register
RCTs	Can be used to determine causal associations	• A substantial proportion of major ophthalmology RCTs were conducted 1–2 decades ago • Limited sample sizes • Limited sample ethnic diversity	Early Manifest Glaucoma Trial, Ocular Hypertension Treatment Study, Age-Related Eye Disease Studies, Pediatric Eye Disease Investigator Group
Administrative data			
Institutional data	Longitudinal analyses within an institution	• Limited data quality with missing, duplicate, or misclassified data • Nonstandard nomenclature • Cannot be generalized	Data from all clinical practices
Insurance claims-level data	Longitudinal analyses with large and diverse samples	• Clinical care cannot reliably be inferred from claims data • Data can be missing or misclassified • Recorded diagnosis and treatment codes may be influenced by institutional coding patterns	Optum Medical Claims, National Health Insurance Research Database of Taiwan, The Centers for Medicare and Medicaid Services (CMS) Medicare data

RCT = randomized controlled trial.

from a variety of sources. Of note, there is an Observational Health Data Science and Informatics/OMOP Eye Care & Vision Research working group that aims to improve integration of ophthalmology data with other large-scale datasets mapped to OMOP.

Another example for harmonizing data across systems is the Informatics for Integrating Biology to the Bedside platform (Partners Healthcare Systems, www.i2b2.org), which facilitates cohort discovery using a standard data model and query tools. The remit of Informatics for Integrating Biology to the Bedside has also been extended to facilitate multi-institution collaboration by allowing queries to span multiple institutions. A key challenge with Informatics for Integrating Biology to the Bedside is the configuration and maintenance of the infrastructure required.

Standards developed for use in clinical practice also have relevance to research data acquisition and management. Fast Healthcare Interoperability Resources has been adopted as the preferred means by which clinical information systems in the United States should exchange clinical data.[4] Fast Healthcare Interoperability Resources has been designed for the exchange of data between systems and so has a role in retrieving clinical data for research, but is not yet a solution for the storage of those data.

Finally, the Digital Imaging and Communication in Medicine standard has been developed for biomedical image formatting, management, interpretation, and storage.[5] Digital Imaging and Communication in Medicine has also been extended to accommodate eye care, and standards are available for the most commonly used in-office testing devices. Unfortunately, these standards have not been widely adopted by vendors in eye care, and Digital Imaging and Communication in Medicine compliance is reported to be low for ophthalmic imaging.[6] The National Eye Institute, Office of the National Coordinator for Health Information Technology, and the Food and Drug Administration recently held a joint workshop to promote the adoption of ocular imaging standards across vendors in May 2022.

Taking the aforementioned solutions into account, one comprehensive approach to address the drawbacks of current data practices in eye care involves a combination of Fast Healthcare Interoperability Resources, to enhance clinical-care level interoperability, and the OMOP Common Data Model to harmonize data for research. This hybrid approach has been adopted in other medical fields, with notable success in neurology and cardiology. The International Neuroinformatics Coordinating Facility was established in 2005 to promote neuroscience data sharing and data reuse.[7] The International Neuroinformatics Coordinating Facility has instituted standards including Neuroscience information exchange format, a data model and file format for annotated scientific datasets, and the Computational Neuroscience Ontology, a controlled vocabulary of terms used to describe nervous system models.[7] Similarly, the American College of Cardiology introduced the National Cardiovascular Data Registry in 1997 to record cardiac catheterization and coronary intervention data.[8] The National Cardiovascular Data Registry has a Data Quality Program that mandates filtering for consistency and completeness before data entry.[8] In addition, the Society of Thoracic Surgeons registry has enrolled > 90% of adult cardiac surgical centers in the United States and includes data from 9 countries, with a total of > 5 million patient records.[9] These measures demonstrate how such initiatives can contribute to large-scale data sharing to lead to impactful research within their respective fields.

The implications of poor data standardization, availability, and shareability are potentially far-reaching. Many recent ophthalmology publications use data from the same datasets, which are familiar to researchers due to their prominence in existing literature and visibility compared with alternative data sources.[10] Barriers to data access also contribute to the oversampling of select datasets. This introduces bias because minority groups are underrepresented or unrepresented in these datasets, which are derived from predominantly White populations of European ancestry.[10−12] The participants from which these data are derived may also be more affluent compared with the general population, which can result in possible socioeconomic determinants of health and disease being overlooked.[13] On a global scale, the United States, China, and Western Europe contribute an overwhelmingly disproportionate quantity of accessible eye care clinical data, with some underresourced populations not represented at all. Furthermore, available data frequently relate to diseases of particular relevance in developed countries, often associated with disease screening and multiple sequential patient visits in specialized health care settings where multimodality imaging is routine. This is in contrast to the eye health priorities of developing countries, which include uncorrected refractive error, trachoma, and cataracts.[14]

Diversifying reference data and improving our ability to predict diseases in a greater number of ethnic groups will help to protect against the potential for research to exacerbate existing health disparities. Investing in standardizing and rationalizing data structures for shareability and the public availability of datasets with the capability to support research for data-poor regions will be positive steps to address inequities. The standardized curation of metadata could enhance data discoverability and facilitate dataset comparisons and merging. There is also potential for an online, searchable dataset repository to not only increase the visibility of lesser-used data but also to mitigate against aborted research efforts where appropriate data may be available but not easily found. A standardized data model may also facilitate the ability to conduct federated learning approaches, which allow for interfacility analyses when direct data sharing is not possible. Enhancing the sharing and diversity of data informing research conclusions will reduce bias resulting from the repeated use of the same potentially skewed datasets and is likely to improve the generalizability of research findings.

In summary, current data practices in eye care are discordant and therefore limit the potential for representative research discoveries. We call for the implementation of eye care data standards to support the efficient use of data to inform decision making, large-scale data sharing, and greater research collaboration.

Footnotes and Disclosures

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s): M.V.B.: Consultant − Carl Zeiss Meditec, Topcon, Allergan, Janssen; Lecture fees − Carl Zeiss Meditec.

N.Z.: Financial Support − NIH K23 career development award (K23EY032634), Research to Prevent Blindness career development award.

Correspondence:
Nazlee Zebardast, MD, MSc, Massachusetts Eye and Ear, 243 Charles St, Boston, MA 02114. E-mail: nazlee_zebardast@meei.harvard.edu.

References

1. Coughlin S, Roberts D, O'Neill K, Brooks P. Looking to tomorrow's healthcare today: a participatory health perspective. *Intern Med J*. 2018;48(1):92−96. https://doi.org/10.1111/imj.13661.
2. Ng WY, Zhang S, Wang Z, et al. Updates in deep learning research in ophthalmology. *Clin Sci (Lond)*. 2021;135(20): 2357−2376. https://doi.org/10.1042/CS20210207.
3. Kent S, Burn E, Dawoud D, et al. Common problems, common data model solutions: evidence generation for health technology assessment. *Pharmacoeconomics*. 2021;39(3): 275−285. https://doi.org/10.1007/s40273-020-00981-9.
4. Braunstein ML. Healthcare in the age of interoperability: the promise of fast healthcare interoperability resources. *IEEE*

Pulse. 2018;9(6):24−27. https://doi.org/10.1109/MPUL.2018.2869317.

5. Bidgood WD, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. *J Am Med Inform Assoc*. 1997;4(3):199−212. https://doi.org/10.1136/jamia.1997.0040199.
6. Lee AY, Campbell JP, Hwang TS, et al. Recommendations for standardization of images in ophthalmology. *Ophthalmology*. 2021;128(7):969−970. https://doi.org/10.1016/j.ophtha.2021.03.003.
7. Abrams MB, Bjaalie JG, Das S, et al. A standards organization for open and FAIR neuroscience: the international neuroinformatics coordinating facility. *Neuroinformatics*. 2022;20(1):25−36. https://doi.org/10.1007/s12021-020-09509-0.
8. Messenger JC, Ho KKL, Young CH, et al. The National Cardiovascular Data Registry (NCDR) data quality brief: the NCDR Data Quality Program in 2012. *J Am Coll Cardiol*. 2012;60(16):1484−1488. https://doi.org/10.1016/j.jacc.2012.07.020.
9. Fernandez FG, Shahian DM, Kormos R, et al. The Society of Thoracic Surgeons national database 2019 annual report. *Ann Thorac Surg*. 2019;108(6):1625−1632. https://doi.org/10.1016/j.athoracsur.2019.09.034.
10. Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health*. 2021;3(1):e51−e66. https://doi.org/10.1016/S2589-7500(20)30240-5.
11. Ziemssen F, Feltgen N, Holz FG, et al. Demographics of patients receiving intravitreal anti-VEGF treatment in real-world practice: healthcare research data versus randomized controlled trials. *BMC Ophthalmol*. 2017;17(1):7. https://doi.org/10.1186/s12886-017-0401-y.
12. Allison K, Patel DG, Greene L. Racial and ethnic disparities in primary open-angle glaucoma clinical trials: a systematic review and meta-analysis. *JAMA Netw Open*. 2021;4(5):e218348. https://doi.org/10.1001/jamanetworkopen.2021.8348.
13. Shweikh Y, Ko F, Chan MP, et al. Measures of socioeconomic status and self-reported glaucoma in the U.K. Biobank cohort. *Eye (Lond)*. 2015;29(10):1360−1367. https://doi.org/10.1038/eye.2015.157.
14. Burton MJ, Ramke J, Marques AP, et al. The Lancet global health commission on global eye health: vision beyond 2020. *Lancet Glob Health*. 2021;9(4):e489−e551. https://doi.org/10.1016/S2214-109X(20)30488-5.