Journal of the Association for Research in Otolaryngology (2023) 24:397–399
https://doi.org/10.1007/s10162-023-00908-z

JARO

EDITORIAL

# An Opportunity for Constructing the Future of Data Sharing in Otolaryngology

Mark A. Eckert[1] · Fatima T. Husain[2] · Dona M.P. Jayakody[3,4] · Winfried Schlee[5,6] · Christopher R. Cederroth[7,8]

Published online: 10 August 2023

There is a shifting landscape for how researchers collect, organize, and share data owing to the demand for big data and initiatives to increase innovation, scientific integrity, rigor, and reproducibility. Moreover, there is opportunity to advance understanding in a field by integrating multi-site data, including for complex and rare disorders and/or underfunded areas of study. These changes and opportunities have led to recommendations and guidelines from journals [1] (including *JARO*'s recent updates[1]), as well as funding agencies [2], that researchers share their data. Data sharing policies vary considerably depending on the research field [3], and many are still in the process of establishing guidelines [4]. Here, we summarize the current untidy state of data sharing in the otolaryngology research field and present a call for the community to establish guidelines and mechanisms that will facilitate data sharing and prevent datasets from being siloed into difficult to access and/or difficult to search repositories.

## The Current Data Sharing Landscape

The most common approach for sharing data from published, completed study phase (i.e., longitudinal), and/or fully completed research studies provides researchers with more control over the data and limits data sharing to specific requests through Data Use Agreements (DUA). There are multiple complications with this approach, including the uncertainty about long-term sustained access to data. For example, researchers may have difficulty providing access to the data if they leave the institution where the data has been collected. Hardware failure may also simply jeopardize the long-term accessibility to data. Ransomware attacks can also put the data at stake for the researchers within an institution, as in the December 2019 phishing attack on Maastricht University. The DUA approach can also be time consuming with the need to include ethics, research office, and legal departments across contributing and recipient institutions. Initiatives like the Federal Demonstration Partnership (FDP) can modestly streamline data sharing for FDP members with the use of template DUA, but there is room for improvements to make it easier for international researchers to share and access data.

Sharing data across borders may need special attention by researchers depending on the data protection regulations of the involved countries, particularly for studies involving human participants. For example, within the European Union (EU), data sharing is regulated by the General Data Protection Regulation (GDPR). To the best of our knowledge, the GDPR only regulates the sharing of personal data that relates to an identified or identifiable living individual, which also means that the GDPR does not apply to

[1] See the section on Research Data Policy and Data Availability Statements under https://www.springer.com/journal/10162/submission-guidelines

✉ Mark A. Eckert
eckert@musc.edu

[1] Hearing Research Program, Department of Otolaryngology – Head and Neck Surgery, Medical University of South Carolina, Charleston, SC, USA

[2] Department of Speech and Hearing Science and The Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

[3] Ear Science Institute Australia, Subiaco, WA 6008, Australia

[4] Ear Sciences Centre, Medical School, University of Western Australia, Crawley, Australia

[5] Institute for Information and Process Management, Eastern Switzerland University of Applied Sciences, St. Gallen, Switzerland

[6] Department of Psychiatry and Psychotherapy, University of Regensburg, Regensburg, Germany

[7] Laboratory of Translational Auditory Neuroscience, Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden

[8] Translational Hearing Research, Tübingen Hearing Research Center, Department of Otolaryngology, Head and Neck Surgery, University of Tübingen, Tübingen, Germany

anonymized data where the individual is not identifiable. Nor does it apply to the deceased. In other words, data sharing of anonymized data is not regulated by GDPR. Similarly, research with anonymized data often does not require institutional review board oversight in the USA.

The definition of anonymized data can differ between countries. In the USA, the Department of Health and Human Services provides guidance for expert determination that data is not identifiable or the "Safe Harbor" approach of removing 18 types of identifiers as acceptable data de-identification approaches. The current guideline in the USA is that a de-identified dataset can include a label or code for cases that can be linked back to participant identifiers, provided that the data recipient does not have access to the "key" linking the code to identifiers. In the EU, data is considered de-identified only when new labels are given to data that are unlinked to the original labels and "key." This can be challenging in the context of studies where data collection is still ongoing (e.g., longitudinal). It can also be time consuming to provide unlinked participant codes for complex multi-dimensional datasets, although there is de-identification software available that can limit the effort for data providers to organize complex datasets and will automatically provide new and unlinked case labels.

The risk for re-identification should be considered when sharing data to limit the potential for harm to identified participants. In Australia, this risk is a key factor in determining whether data is de-identified, as defined using a "De-identification Decision-Making Framework," where investigators audit their data situation and responsibilities, identify the risk of disclosure, and consider the potential impacts of a data release. For example, when sharing data belonging to Australian Aboriginal and Torres Strait Islander peoples, researchers must consider the Indigenous self-determination, Indigenous leadership, impact and value, and sustainability and accountability principles of the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) Code of Ethics. Researchers and institutions or other parties must discuss the ownership management and communication of research data and results with Indigenous peoples based on Indigenous data sovereignty and governance principles. Statistical approaches like generalizing variable values to limit unique combinations of values across variables or by creating synthetic datasets [5] are approaches for limiting risk.

Here, we have only considered data generated by researchers in academia or similar institutions (i.e., research foundations). There can be constraints on data generated and curated by commercial entities, including cost to academic user(s) or denial of access by the user's institution when the commercial entity requires a DUA with indemnification language. Institutions are also selling data (e.g., de-identified electronic health records) that provide opportunities for advancing science and health care but also risks to privacy [6], which include the potential for re-identification. Researchers and academic institutions should be clear-eyed when planning to partner with commercial entities and/or access commercial data.

## Data Repositories

Whereas journals have strived for data deposition in repositories, this mainly pertains to genomics, transcriptomics, proteomics, and crystallography. *JARO*, for instance, recommends data to be included as supplemental material. However, for a research field as broad as otolaryngology, which encompasses a large number of sub-research areas where data can reach terabytes (e.g., audiology, neuroimaging, genetics, metabolomics, modeling, and recently cytometry and immunology), it can be useful to consider existing repositories. Publishers like Springer and Elsevier have research data policies and examples of repositories to be used. Springer also launched a journal called *Scientific Data* that allows authors to describe their data sets in a format called Data Descriptors. JARO too may become the future platform for the publication of Data Descriptors of the otolaryngology community. We also note that code sharing through platforms like GitHub is important for facilitating the reproducibility of data analyses.

General purpose data repositories, like the Open Science Framework, provide another reasonable data sharing option. However, datasets from different studies within a research field can often be scattered across various data sharing resources, thus making it difficult to find and aggregate data across studies for novel investigations on existing data. Moreover, such unstructured data sharing approaches can be limiting if researchers do not provide sufficient information for understanding the data, including information about best-use practices for analysis and interpretation of the data. Open access data repositories should require clearly defined data, or meta-data, and guidance about how to best use the data. For example, the Australian Data Archive is organized within a Dataverse platform for the storage of data from 1000s of studies where meta-data about sampling, data collection approaches, demographics, and participant response rates are available in an open access format prior to a formal request for data access. Similarly, the Zenodo platform allows for sharing of curated data like a European Union–supported multi-site project on tinnitus (UNITI) with data descriptors for the full data set that will optimize usability.

Perhaps the most useful multi-site data sharing approach to date has been to generate data repositories dedicated to model organisms (e.g., Mouse Genome Organisms) and domain-specific data (e.g., gEAR) [7]. Here, datasets can be curated with common variable types, as well as integrated and analyzed using methods to deal with missing data and different measurements for the same construct. Moreover, a community of collaboration can develop within these resources. These types of resources can also provide secondary data to researchers as a form of incentive to share data and to facilitate replication. For example, the Dyslexia Data Consortium and Hearing Health Institute's data repositories are emerging resources where researchers can share their neuroimaging data, which is automatically processed to provide visualization and secondary data generation (e.g., regional brain volume predictors of dyslexia). That is, contributors can leverage data processing functions that would otherwise require computational resources and personnel training, in addition to providing access to these important datasets in one location. Development of these types of resources is labor intensive and requires significant resources to ensure long-term viability, which has been addressed with data access fees for some data repositories (e.g., UK Biobank). Ideally, there would be no cost to limited data access under the findable, accessible, interoperable, and reusable (FAIR), as well as collective benefit, authority, responsibility, and ethics (CARE) principles. There are many potential solutions for limiting costs (e.g., sliding fee structures for academic to industry access) that will depend on the organizations housing the data.

## The Future Data Sharing Landscape

Institutions place high value on data generated within its infrastructure and thus there can be conflict between open data access and institutional priorities. At least for institutions competing for funding from the National Institutes of Health, it seems likely that new data sharing requirements will guide the development of institution-specific data sharing infrastructures, thereby allowing researchers to be compliant with data sharing requirements, while maintaining intellectual property for the institution. This scenario may require a federated database system (FDS) that stores metadata describing the data available across institutional repositories. Here, researchers would search the FDS to identify data of interest and to determine where that data is housed. There may be an opportunity for professional associations like ARO, of course with commitment from association members and an investment in the necessary infrastructure, to establish an FDS. The benefit to members and the broader community is the relatively rapid access to multi-site data for hypothesis driven and discovery research, education, and establishing collaboration. This type of resource would allow for historical record of research in the field and could be used to track or identify new research directions. That is, an FDS effectively becomes a domain-specific search engine for research materials that allows for detailed understanding of a field at different scales of resolution that can advance member research, allow for data sharing policy compliance, provide a mechanism to index data used for *JARO* publications, and perhaps raise the profile of the association and value to members.

## Conclusions

Now is the time for establishing guidelines and procedures to share data in a way that is equitable and moves a field forward, while considering stakeholders, including study participants and researchers contributing the data and in service of the health of a scientific field. There is opportunity for the ARO community to determine how their data will be shared and how they wish to access data, perhaps through an ARO-supported mechanism. We understand that while there may not be an ideal common approach for all research areas within otolaryngology, there is an exciting opportunity for creating a resource(s) to promote the core science, diversity, integrity, collaboration, and education values of ARO.

## References

1. Announcement: Where are the data? (2016) Nature 5377619: 138–138
2. Kozlov M (2022) NIH issues a seismic mandate: share data publicly. Nature 6027898:558–559
3. Tedersoo L et al (2021) Data sharing practices and data availability upon request differ across scientific disciplines. Sci Data 81:192
4. Data sharing is the future (2023) Nat Methods 204:471
5. Vaden KI Jr et al (2020) Fully synthetic neuroimaging data for replication and exploration. Neuroimage 223:117284
6. Mandl KD, Perakslis ED (2021) HIPAA and the leak of "deidentified" EHR data. NEJM 38423:2171–2173
7. Orvis J et al (2021) gEAR: Gene Expression Analysis Resource portal for community-driven, multi-omic data exploration. Nat Methods 188:843–844

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.