Statistics in Biosciences
https://doi.org/10.1007/s12561-024-09422-z

COMMENTARY

Building an Enhanced Publication Ecosystem for Statistical Innovation

Hongkai Ji[1]

Received: 15 July 2023 / Revised: 13 February 2024 / Accepted: 16 February 2024

Keywords Software · Data · Resource · Benchmark · Crowdsource · Artificial intelligence

Statistical research functions much like a supply chain, beginning with the identification of demands, such as open problems in science and technology, and the gathering or generation of data (e.g., through surveys, clinical trials, or scientific experiments). These demands and data fuel the creation of innovative analytical solutions, encompassing fresh theories and methodologies for study design and data analysis, and, increasingly, the development of software tools to disseminate these solutions to end-users. To ensure proper and effective utilization of new methods and tools, it is crucial to systematically evaluate them, enabling users to understand their properties, capabilities, and limitations. Ultimately, their value must be tested and demonstrated by solving real-world problems. Thus, demands, data, theory, methods, software, as well as their evaluation, validation, and application, all serve as vital components within this supply chain. It is important to note that this supply chain cannot function well without the presence of well-trained and creative statisticians.

Publication plays a significant role in disseminating new research ideas and results. However, within the current publication ecosystem in the discipline of statistics, not all components of the supply chain receive equal support. Most statistical journals primarily focus on publishing theories, methods, and their innovative applications to new problems. While some statistical journals have recently begun featuring software tools, manuscripts that describe new demands, data, and resources for method evaluation often receive much less attention.

When it comes to data, most statistical journals currently do not consider the publication of a new dataset in itself, as editors and reviewers tend to perceive limited novelty in terms of theory and methods in such a manuscript. However, data is the

✉ Hongkai Ji
hji@jhu.edu

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205, USA

Published online: 16 March 2024

new oil for the twenty-first century. For example, the ImageNet dataset with tens of millions of human-annotated images has been an important driving force for the recent advances in computer vision and deep learning [1, 2]. The massive amounts of text data on the Internet enable training large language models (LLMs) such as ChatGPT [3–5], revolutionizing artificial intelligence (AI). Besides large datasets, high-quality small datasets also have tremendous value. For instance, finding optimal data analysis solutions for emerging new technologies (e.g., single-cell or spatial omics technologies) requires benchmark datasets with ground truth information, but many such benchmark datasets are small due to the difficulty of conducting experiments to collect ground truth information. Compiling these foundational datasets is non-trivial, requiring vast amounts of time and effort. Timely dissemination of these data, along with instructions on how to use them, will benefit the whole research community and accelerate innovation. While journals for publishing data have emerged in several other disciplines, it remains an uncommon practice in statistics. This has at least two ramifications for our discipline: First, it slows down the dissemination of data for innovation. While there are public repositories for certain data types (e.g., the Gene Expression Omnibus [6, 7] and Sequence Read Archive [8] database for microarray and high-throughput sequencing data), unified data repositories are not yet available for many others. Without a pathway to publish a valuable dataset (e.g., a benchmark dataset that can be used to compare different statistical methods) without new theory and methods, one has limited options to publicize the data and allow the community to benefit from it. Second, it weakens our workforce. It disincentivizes statisticians from leading the efforts to build critical data resources, which can take tremendous amounts of time but too frequently are not recognized enough by, for example, hiring or promotion committees who evaluate candidates based on their publication list. In the long run, this will result in a loss of related talent, creating a bottleneck in the supply chain of our discipline. Without timely access to high-quality data, our innovation will lag behind.

One likely solution is to enhance our publication ecosystem to support the publication of valuable datasets along with their usage instructions. For instance, the Journal of Statistics and Data Science Education considers “Datasets and Stories” articles, which describe “the pedagogical uses of multivariate dataset(s)” [9]. Statistics in Biosciences now welcomes submissions of articles that present broadly useful new datasets and resources for statistical research and education (e.g., [10]). These resource articles are published as part of the journal’s “Case Studies and Practice Articles.” Embracing this new publication type will offer new opportunities for innovation. For example, as more datasets with well-annotated ground truth are published, they may be used to build a crowd-sourced benchmark data compendium for comparing different statistical methods developed for a common task. This would allow for a more robust and less biased assessment of different methods. Improved benchmarks, in turn, can more effectively guide method developers in identifying effective models, techniques, and optimal directions to improve their methods. The publication of data will also allow us to better recognize the vital contribution of those who worked hard on generating, collecting, cleaning, compiling, and annotating the data, thus helping to cultivate the expertise indispensable for the prosperity of the discipline.

Data is just one example of how we can strengthen our supply chain. Similarly, improved support for the other traditionally under-supported supply chain components is also crucial. Timely introduction of new problems and challenges emerging from science and technology to the statistical community, for instance, could help statisticians participate in and contribute to important scientific innovations in the first place. This could be facilitated by publishing commentaries or perspectives by those who work at the interface between statistics and domain sciences. Besides benchmark datasets, there are many other types of resources critical for statistical research, such as computational pipelines for method evaluation and educational materials for emerging topics. They can be naturally published as resource articles. By encouraging these new types of publications, we can create a more balanced publication ecosystem to support statistical innovation. There are various avenues to facilitate this process. One option is for the statistical community to consider establishing new specialized journals tailored for these emerging content types. Alternatively, recognizing that the creation of new journals may involve substantial time and resources, existing journals could incorporate new article types, allowing the community to capitalize on their established platforms for a swift response to emerging needs and opportunities. Regardless of the chosen approach, the success of these new publication types in statistical journals hinges on editors, reviewers, and authors embracing a more inclusive mindset. It is crucial to acknowledge the vital role and significant contributions of each component in our supply chain. Ultimately, the productivity of a supply chain is determined by its bottlenecks. To enhance the success of our discipline, we must ensure that there are no bottlenecks in our supply chain.

Acknowledgements The author would like to thank the handling editor and reviewer for their invaluable feedback. The language in the manuscript was polished by ChatGPT.

Funding This work was supported by the National Institutes of Health Grant R01HG010889 to HJ.

Declarations

Conflict of interest HJ is currently Editor-in-Chief of Statistics in Biosciences.

References

1. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF (2009) ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248–255, doi: https://doi.org/10.1109/CVPR.2009.5206848
2. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Li FF (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115:211–252
3. OpenAI (2021) ChatGPT. https://chat.openai.com/. Accessed 13 Feb 2024.
4. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, Du Y, Yang C, Chen Y, Chen Z, Jiang J, Ren R, Li Y, Tang X, Liu Z, Liu P, Nie JY, Wen JR (2023) A survey of large language models. arXiv:2303.18223v10 [cs.CL]. https://doi.org/10.48550/arXiv.2303.18223
5. Fan L, Li L, Ma Z, Lee S, Yu H, Hemphill L (2023) A Bibliometric review of large language models research from 2017 to 2023. arXiv:2304.02020v1[cs.DL]. https://doi.org/10.48550/arXiv.2304.02020
6. Edgar R, Domrachev M, Lash AE (2002) Gene Expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30(1):207–210
7. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 41:D991–D995
8. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C (2022) The sequence read archive: a decade more of explosive growth. Nucleic Acids Res 50(D1):D387–D390
9. Journal of Statistics and Data Science Education. https://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=ujse21
10. Sugolov A, Emmenegger E, Paterson AD, Sun L (2023) Statistical learning of large-scale genetic data: how to run a genome-wide association study of gene-expression data using the 1000 genomes project data. Stat Biosci. https://doi.org/10.1007/s12561-023-09375-9