Digestive Diseases and Sciences (2024) 69:18–21
https://doi.org/10.1007/s10620-023-08136-4

FELLOWS AND YOUNG GIS

Gastroenterology Fellowship and Postdoctoral Training in Omics and Statistics—Part I: Why Is It Needed?

Madeline Alizadeh[1] · Natalia Sampaio Moura[2] · Alyssa Schledwitz[2] · Seema A. Patil[2] · Jacques Ravel[1] · Jean-Pierre Raufman[2,3,4,5]

Received: 13 August 2023 / Accepted: 27 September 2023 / Published online: 2 November 2023

Abstract

A multitude of federally and industry-funded efforts are underway to generate and collect human, animal, microbial, and other sources of data on an unprecedented scale; the results are commonly referred to as "big data." Often vaguely defined, big data refers to large and complex datasets consisting of myriad datatypes that can be integrated to address complex questions. Big data offers a wealth of information that can be accessed only by those who pose the right questions and have sufficient technical knowhow and analytical skills. The intersection comprised of the gut-brain axis, the intestinal microbiome and multi-ome, and several other interconnected organ systems poses particular challenges and opportunities for those engaged in gastrointestinal and liver research. Unfortunately, there is currently a shortage of clinicians, scientists, and physician-scientists with the training needed to use and analyze big data at the scale necessary for widespread implementation of precision medicine. Here, we review the importance of training in the use of big data, the perils of insufficient training, and potential solutions that exist or can be developed to address the dearth of individuals in GI and hepatology research with the necessary level of big data expertise.

Keywords Medical informatics · Medical education · Big data · Gastroenterology · Hepatology

Introduction

Although large datasets are encountered with increasing frequency in both the sciences and medicine, the term 'big data' is commonly undefined and open to some degree of interpretation. In the present communication we use this term to refer to large, complex datasets consisting of several different datatypes that can often be integrated to address complex questions and must almost always be sorted and analyzed using sophisticated computer algorithms and programs. Accessing, analyzing, and interpreting big data becomes an increasingly daunting challenge given that asking and answering the right questions requires technical knowledge to parse these datasets and extract the required information using validated methods. Moreover, the interconnection between the gut-brain axis, the gut microbiome and multi-ome, and connected digestive organ systems (i.e., hepatobiliary tract and pancreas) poses additional challenges to those involved in gastrointestinal (GI) and liver research [1]. Indeed, training in the methods needed to integrate and correctly analyze and interpret the information in multiple large, non-linear datasets consisting of multi-omics and other datatypes remains far behind data acquisition [1]. Even the use of artificial intelligence (AI) to address these needs requires the development, oversight, and validation of these novel tools by appropriately trained and experienced people. These factors contribute to the

✉ Jean-Pierre Raufman
jraufman@som.umaryland.edu

[1] The Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 20201, USA

[2] Division of Gastroenterology and Hepatology, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[3] VA Maryland Healthcare System, Baltimore, MD 21201, USA

[4] Marlene and Stewart Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[5] Department of Biochemistry and Molecular Biology, University of Maryland School of Medicine, Baltimore, MD 21201, USA

perceived failure to capitalize on unrivalled opportunities to apply new knowledge to inform the mechanisms underlying health and disease and, thereby, develop novel diagnostics and therapeutics [2].

The use of big data is critical to advancing precision medicine; unfortunately, current approaches to provide adequate training in big data acquisition and analysis in GI and hepatology research are insufficient, resulting in datasets that are frequently inaccurately and inadequately analyzed and commonly underutilized. These deficiencies can be addressed by improving and expanding training opportunities in this field. Here, we review the role of big data in gastroenterology and hepatology research, current inadequacies in training regarding its use, and the need for additional training opportunities across the spectrum of research careers. In a subsequent communication, we review a variety of currently available avenues to obtaining such training.

Role of Big Data in GI and Liver Research

While substantial variation in data characteristics and structures exist, 'big data' represents sizeable datasets defined by their shared traits of large volume and high complexity [3]. The term applies to multiple different data types, including genetics and genomics, metabolomics and proteomics, transcriptomics, microbiome sequencing, imaging, and clinical data. While differences in analytic methods exist between these datasets, and can overwhelm those unfamiliar with the field, similar statistical principles underlie those methods (Fig. 1). These datasets are key to understanding health and disease at an unprecedented level of granularity, especially at the degree required to develop and employ precision medicine. As a result, big data has become increasingly common in every facet of biomedical research—GI and hepatology research being a prime example.

Several techniques, methods, and tools have been developed rapidly in recent years specifically for the generation

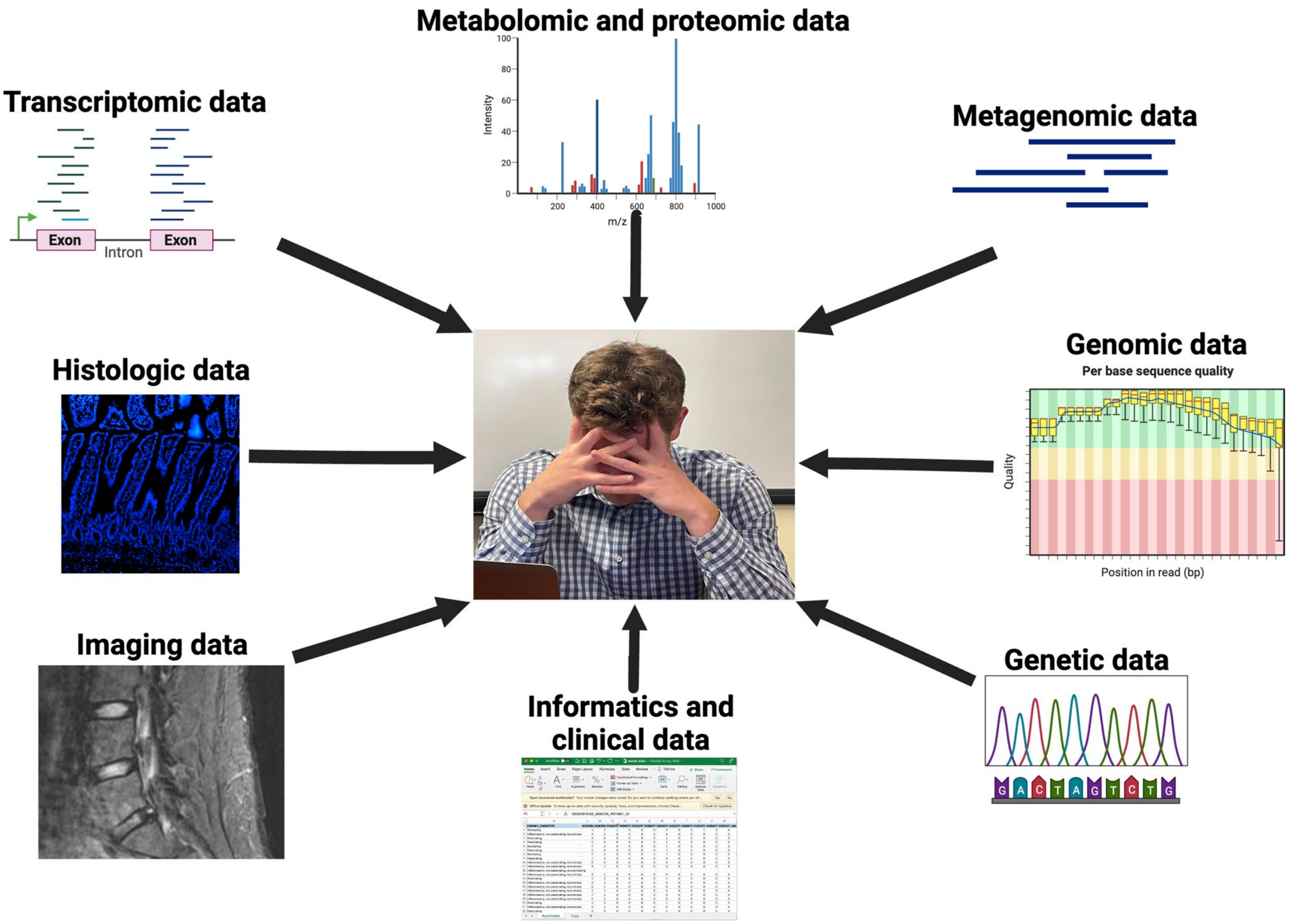

Fig. 1 Many investigators lack the knowhow and resources to adequately assess, interrogate, analyze, and integrate the vast amounts of big data that are currently being generated. Without proper training, the tremendous amount of data bombarding investigators can be overwhelming and uninterpretable. Created using Biorender.com

of these data [3] and many are specific to data collected from the GI tract, e.g., considering stool as a "tissue" type and using intestinal microbiome-specific analytical tools [1]. This is in large part due to the importance of the GI and hepatobiliary tracts as sources of big data, particularly in the context of their substantial interconnection with other organ systems and the gut microbiome. As a source for the production and systemic absorption of metabolites, nutrients, and other products that are then delivered throughout the body, the GI and hepatobiliary tracts are particularly important to incorporate into precision medicine research. Unfortunately, the development of training opportunities has not kept pace with the generation of big data, resulting in a discrepancy between the enormous amount of available data and the relative paucity of researchers equipped to analyze it.

Inadequacies in Big Data Training and Approaches to Increasing the Numbers of Biostatistics- and Programming-Trained Investigators in GI and Liver Research

Compressed 'bootcamps' or one-semester courses may offer an introduction but are insufficient by themselves to provide the required training for big data analytics. The appropriate use of big data to improve patient outcomes requires research team members appropriately trained for the task. While medical schools may require mathematics and biostatistics prerequisites, and basic statistical knowledge is tested by board examinations, most students are trained to understand results, but not to analyze the underlying data. Further, this training is often inadequate. A 2007 study revealed that only 41.4% of residents across 11 residency programs were able to correctly interpret simple statistical results tested on board examinations [4].

Master's level programs in clinical research may offer instruction in the fundamentals of clinical research, that may include coursework in biostatistics and bioinformatics. However, most do not progress past teaching a subset of epidemiological techniques, and commonly fail to address omics data and the analysis and integration of different types of high-throughput omics data. Lastly, they rarely test graduates to ascertain their acquisition and retention of such knowledge. Given the time constraints of such training programs, a condensed version of relevant statistics is typically provided, commonly excluding vital theoretical background information and tools necessary to address complex problems that cannot be solved with simple solutions. Generally speaking, incomplete training in big data analyses has resulted in numerous instances of inappropriate large scale clinical analyses impacting public health, including inaccurate 2013 Google Flu Trend predictions [5].

Proposed solutions to these concerns include incorporating trained data scientists into research teams; some suggest the creation of a new role, referred to as the health information specialist, a team member possessing a comprehensive data analytics background who can be more readily trained in the specifics of clinical data science [6]. Such specialists would possess a strong theoretical and computational background, the ability to make correct analytical choices, e.g., modeling continuous versus discrete variables. This is crucial for generating reproducible, trustworthy results, particularly in the context of applying more advanced machine learning and AI methods, where the quality of initial input may exponentially impact the quality of downstream results. Additionally, the leaders of such research teams must themselves possess sufficient knowledge to have confidence in the scientific rigor and accuracy underlying the acquisition and analysis of these complex big datasets and take ownership and responsibility for the integrity of the results and conclusions.

Advanced clinical training reframes the lens with which clinical data are viewed. Moreover, the current academic infrastructure in the United States makes it difficult to support investigators not directly involved in clinical care. Appropriately training clinicians in the big data analysis methods may provide an alternative approach. As statistical techniques continue to advance and become more complex, so do the skillsets needed to apply them. Hence, ideal candidates might include those with engineering or mathematics backgrounds who desire to transition to health science careers. Participants in medical scientist, i.e., MD/PhD, training programs, have more time during their research years to invest in formal (coursework, symposia, and workshops) and informal (mentorship or project-related) statistical training. Another approach may be to expand post-doctoral opportunities, perhaps using T32 training grants. For instance, at the University of Pittsburgh, the GI division manages a T32 training grant in big data and precision medicine, that provides intense classroom and practical training; developing more programs of this nature would expand access and post-graduate training paths to those interested in big data research. Although there is currently a paucity of such opportunities, to enhance the training experience, asynchronous online courses geared at residents or fellows would be advantageous. Regardless of the precise path followed, training must acknowledge that statistical analysis is also an art; no methodology or approach is etched in stone—the same problem can be addressed by multiple methods, none of which are necessarily the best approach. This conundrum highlights the importance of cultivating a strong background in biostatistics and bioinformatics before developing and using statistical tools.

Conclusion and Future Directions

Although big data science represents the future of GI and hepatology research, there is an unmet need for investigators trained in the use and analysis of existing and accumulating large datasets, with the potential to profoundly impact healthcare. Using big data to personalize medicine has great potential to improve quality of life and clinical outcomes while substantially lowering healthcare expenditures; the potential benefits should neither be underestimated nor underutilized. However, addressing this need requires maximizing the use of current training vehicles, additional training opportunities, and commensurate investment by federal agencies, academic institutions, foundations, and even the pharmaceutical industry to support the cost of extended training. Our subsequent communication addresses existing and promising training opportunities.

Funding M.A. and N.S.M. were supported by an award from the National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases (T32 DK067872-19; J-P Raufman, PI).

Declarations

Conflict of interest The authors have no conflicts to disclose.

References

1. Alizadeh M, Sampaio Moura N, Schledwitz A, Patil SA, Ravel J, Raufman JP. Big Data in Gastroenterology Research. *Int J Mol Sci*. 2023;24.
2. Wooden B, Goossens N, Hoshida Y, Friedman SL. Using Big Data to Discover Diagnostics and Therapeutics for Gastrointestinal and Liver Diseases. *Gastroenterology*. 2017;152:53-67.e53.
3. Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights*. 2016;8:BII.S31559.
4. Windish DM, Huot SJ, Green ML. Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature. *JAMA*. 2007;298:1010–1022.
5. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. *Science*. 2014;343:1203–1205.
6. Fiske A, Buyx A, Prainsack B. Health Information Counselors: A New Profession for the Age of Big Data. *Acad Med*. 2019;94:37–41.