Article published online: 2022-12-04

© 2022 IMIA and Georg Thieme Verlag KG

2021 Bioinformatics and Translational Informatics Best Papers

Mary Lauren Benton[1], Scott Patrick McGrath[2], Section Editors for the IMIA Yearbook Section on Bioinformatics and Translational Informatics

[1] Assistant Professor, Department of Computer Science, Baylor University, Waco, TX, USA

[2] Academic Program Management Officer, CITRIS Health, University of California Berkeley, Missoula, MT, USA

Summary

Objectives: To identify and summarize the top bioinformatics and translational informatics papers published in 2021 for the IMIA Yearbook.

Methods: We performed a broad literature search to retrieve Bioinformatics and Translational Informatics (BTI) papers and coupled this with a series of editorial and peer reviews to identity the top papers in the area.

Results: We identified a final candidate list of 15 BTI papers for peer-review; from these candidates, the top three papers were chosen to highlight in this synopsis. These papers expand the integration of multi-omics data with electronic health records and use advanced machine learning approaches to tailor models to individual patients. In addition, our honorable mention paper foreshadows the growing impact of BTI research on precision medicine through the continued development of large clinical consortia.

Conclusion: In the top BTI papers this year, we observed several important trends, including the use of deep-learning approaches to analyse diverse data types, the development of integrative and web-accessible bioinformatics pipelines, and a continued focus on the power of individual genome sequencing for precision health.

Keywords

Bioinformatics, biomedical informatics, machine learning, precision medicine, genomics

Yearb Med Inform 2022:116-20

http://dx.doi.org/10.1055/s-0042-1742538

1 Introduction

The year 2022 finds both the general public and scientific researchers in a unique state. Having traversed the many ups and downs of the COVID-19 pandemic, we are both more knowledgeable about SARS-CoV-2, and more exhausted by it. While the public may be increasingly disinterested in the virus, scientific publications focused on SARS-CoV-2 are still being released at a significant pace. According to LitCovid there have been 254,669 COVID-19 publications since 2019 [1]. In 2022, there have been 46,203, with an average of 2,100 new papers per week. For the 2021 Bioinformatics and Translational Informatics (BTI) best papers, we did not directly filter out COVID papers, although we did focus our work on identifying important papers that may have been deprived of scientific attention during this time. In this synopsis, we will describe our methodology for selecting the top BTI papers of the year and discuss a few important trends highlighted by the work.

2 Methods

In order to identify high-quality manuscripts to add to our candidate pool, we used the following methodology. We focused our search on the most relevant journals for bioinformatics and translational informatics with electronic publication dates on or after January 1, 2021. The journals surveyed for best papers are as follows: Journal of the American Medical Informatics Association (JAMIA), Journal of Biomedical Informatics (JBI), PLoS Computational Biology, Bioinformatics, BMC Bioinformatics, BMC Systems Biology, Nature, Nature Genetics, Nature Biotechnology, Nature Methods, Science, Science Translational Medicine, Clinical Pharmacology and Therapeutics, New England Journal of Medicine, Journal of the American Medical Association (JAMA), Lancet, PLoS Genetics, and Cell. We then used the following query with the MeSH terms to limit our search to the most relevant papers: (sign OR symptom OR disease OR drug) and (genome OR protein OR small molecule OR RNA OR DNA) AND (computer OR informatics OR statistics). Depending on the journal, each editor was allowed the flexibility to modify or adapt the starting query.

Results from each query were initially scanned by title. Any papers aligned with the domain of bioinformatics or translational informatics were noted, and the abstract was reviewed for possible inclusion. The initial review of papers for inclusion had three primary criteria: novelty, significance, and quality of the paper. Novelty refers to the originality and innovation found in the research question or research methodology (or both), while significance refers to the potential impact of the paper on the field. The quality of the paper was assessed based on the both the technical aspects of the work as well as correctness in the interpretation of the results.

The outcome of these reviews yielded 19 potential papers. The section editors then performed a second review and narrowed this list down to 15 candidate papers for peer-review by IMIA Yearbook editors

and independent experts. Each paper was reviewed by at least three reviewers. Once the volunteer reviewers had submitted their reviews, the editors followed up with their own scores. From the list of 15 candidates, three papers were selected as best papers for the IMIA BTI section. While it was not selected as one of the top three BTI papers, we did want to highlight an important "honorable mention", so a total of four papers are distinguished (Table 1).

3 Current Trends

Below, we provide a brief discussion of the three important themes we discovered while evaluating the top BTI papers. These themes highlight some important areas of research in this field where we expect to see continued growth in the coming years.

3.1 Data Integration Unlocks New Research Directions

In the past year, researchers have identified new research questions that were previously inaccessible by integrating novel data types into existing models and better leveraging available data. Many of the notable papers here leveraged written and spoken data in order to improve disease diagnosis or identify adverse drug events. For example, in one of the top three papers, Moses *et al.*, used recordings of cortical activity to train a machine learning model to decode words and phrases directly from the brain of a patient with anarthria [2]. This model was able to detect word attempts 98% of the time, with 47.1% accuracy. By incorporating brain activity data with state-of-the-art computational approaches from machine learning and natural language processing (NLP), this model has the potential to transform communication for people who are unable to speak.

Alkenani *et al.*, also leveraged techniques from NLP to train a machine learning model to recognize Alzheimer's disease by analyzing an individual's words (both written and spoken) [3]. The model used an ensemble learning approach and techniques repurposed from the field of NLP to identify the signs of cognitive decline based on linguistic features. The strong performance of the model indicates that it could be a versatile—and fully automated—tool to screen for Alzheimer's disease in the future. In their work, Letinier *et al.*, developed a set of machine learning models to scan patient reports for evidence of adverse drug reactions (ADR) [4]. This approach could improve automated surveillance of ADR from routinely collected data. Both of these methods are innovative because they use written data collected directly (and often routinely) from patients to improve diagnosis and symptom evaluation. Similarly, Lu *et al.*, developed a deep-learning model using routinely collected histology images of tumors to predict the type (primary or metastatic) and site of origin of tumor in cancers of unknown primary [5]. This model could provide physicians with important information about the tumor's origins, thus better directing treatment. The model further represents an advance because it leverages data that is frequently collected, even in clinical settings with fewer resources, leading to a result that allows for both improved patient care and increased equity.

3.2 Advanced Genome-Wide Analysis Pipelines Improve Links between Data and Diagnosis

Genome-wide analyses such as genome-wide and phenome-wide association studies (GWAS and PheWAS, respectively) have been at the forefront of bioinformatics and translational informatics work for many years. However, as demonstrated by the top paper from Veturi *et al.*, developing more advanced bioinformatics pipelines to analyze existing data can provide novel insight and identify signals that previous approaches miss [6]. For example, Veturi *et al.*, developed a three-stage pipeline to combine genome-, transcriptome-, and phenome-wide association studies to study plasma lipids [6]. They were able to identify more than 60 novel signals, and connect lipid-associated variation to a range of other complex diseases through Mendelian randomization. Other notable papers this year also focus on the need to improve our analytic pipelines, from leveraging integrative meta-analyses to identify pleiotropy in disease [7] and rank risk genes [8], to developing temporal analysis techniques that use a variety of biological and clinical data types [9].

Table 1 Top BTI papers chosen for the IMIA Yearbook. Papers are listed in alphabetical order by the last name of the first author. The * denotes the one honorable mention.

Section
Bioinformatics and Translational Informatics
▪ Moses DA, Metzger SL, Liu JR, Anumanchipalli GK, Makin JG, Sun PF, Chartier J, Dougherty ME, Liu PM, Abrams GM, Tu-Chan A, Ganguly K, Chang EF. Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. N Engl J Med 2021 Jul 15;385(3):217-27.
▪ *100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, Ellingford JM, Arno G, Tucci A, Vandrovcova J, Chan G, Williams HJ, Ratnaike T, Wei W, Stirrups K, Ibanez K, Moutsianas L, Wielscher M, Need A, Barnes MR, Vestito L, Buchanan J, Wordsworth S, Ashford S, Rehmström K, Li E, Fuller G, Twiss P, Spasic-Boskovic O, Halsall S, Floto RA, Poole K, Wagner A, Mehta SG, Gurnell M, Burrows N, James R, Penkett C, Dewhurst E, Gräf S, Mapeta R, Kasanicki M, Haworth A, Savage H, Babcock M, Reese MG, Bale M, … , Caulfield M. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med 2021 Nov 11;385(20):1868-80.
▪ Veturi Y, Lucas A, Bradford Y, Hui D, Dudek S, Theusch E, Verma A, Miller JE, Kullo I, Hakonarson H, Sleiman P, Schaid D, Stein CM, Edwards DRV, Feng Q, Wei WQ, Medina MW, Krauss RM, Hoffmann TJ, Risch N, Voight BF, Rader DJ, Ritchie MD. A unified framework identifies new links between plasma lipids and diseases from electronic medical records across large-scale cohorts. Nat Genet 2021 Jul;53(7):972-81.
▪ Wei Q, Ramsey SA. Predicting chemotherapy response using a variational autoencoder approach. BMC Bioinformatics 2021 Sep 22;22(1):453.

Perhaps most importantly, however, is the increasing number of studies that also explicitly publish a web portal or other platform specifically designed to enable widespread use of these integrative pipelines. For example, EpiGraphDB, is a graph-based database containing a wide range of biological relationships, including literature-defined mechanisms, genetic associations, and causal relationships from Mendelian randomization [10]. The database is easily queried either programmatically or through the web portal, allowing users to easily mine the wealth of available data and discover new relationships between entities of interest. Similarly, Mountjoy *et al.*, developed OpenTargets, a pipeline and accompanying web portal that connects trait-associated variants to their most likely gene targets, ensuring the accessibility of pipelines to prioritize loci implicated by genome-wide association studies [11]. The exploration of causality via Mendelian randomization in many of these papers is another important feature; genome-wide studies must continue to validate and establish causality for the associations discovered in today's large clinical cohorts.

3.3 Sophisticated Machine Learning Approaches Derive Meaning from Large Biological Datasets

Much like genome-wide association studies, the use of machine learning has become increasingly ingrained in the bioinformatics literature. As the breadth and depth of the available datasets continue to grow, however, researchers are able to apply increasingly sophisticated computational approaches to model their data. In some cases, as in the top paper by Wei and Ramsey [12], deep-learning models can act as an important unsupervised feature selection approach for downstream prediction algorithms. Using transcriptomic data from five different cancer types, Wei and Ramsey developed a variational autoencoder (VAE) to generate a set of latent features that could be used as input to a second machine learning model to predict chemotherapy response in patients undergoing cancer treatment. The VAE was able to capture known biological information in the samples, and ultimately outperformed other state-of-the-art dimensionality reduction approaches.

The Wei and Ramsey paper represents just one of many employing deep-learning to improve classification, regression, and prediction models using biological and clinical data [2, 4, 5]. While we expect to see the application of increasingly complicated algorithms to biological data in the future, we hope that the increasing amount of biological and clinical data will also provide researchers with the opportunity to validate the performance of existing approaches, develop explainable machine learning models, and spur the development of novel feature selection algorithms.

3.4 New Steps Towards Precision Health in Rare Disease and Increased Health Equity

The final articles we want to highlight focus on the ascertainment of cohorts to help identify and diagnose rare disease, and the development of guidelines to improve health equity when reporting polygenic risk scores. The 100,000 Genomes Project is an effort based in the United Kingdom that, in part, aims to use genome sequencing to aid in the diagnosis of patients with rare disease [13]. Rare diseases collectively affect a large number of individuals worldwide, and majority of these have a genetic component. Rare disease imposes a significant health burden on both the patients and their families, so timely diagnosis can have a significant clinical impact. In the pilot study published in 2021, the 100,000 Genomes Project reports on the sequencing and analysis of 4,660 participants across more than 2,000 families. Importantly, they found that obtaining genome sequencing improved the diagnostic yield (25% overall) for patients with a wide range of rare diseases, where a quarter of these diagnoses had some clinically actionable outcome. These results complement a similar project in the United States, the Undiagnosed Diseases Network [14], and support the use of genome sequencing and analysis to improve the rate of diagnosis for individuals with rare disease.

As a step towards precision medicine, polygenic risk scores (PRS) are popular way to sum the effects of trait-associated genetic variants from GWAS into a single score that quantifies the genetic predisposition of a trait across individuals. However, PRS are known to generalize poorly across populations, which can worsen health disparities [15], and there is little standardization in their construction and application. Wand *et al.*, present a perspective article outlining the Polygenic Risk Score Reporting Standards to help mitigate these issues [16]. Such guidelines are essential in order to enable the introduction of PRS into the clinic in the most interpretable and equitable way.

Conclusions

The translational and bioinformatics papers highlighted here represent important advances in the field that improve our understanding of the biological basis of disease, and will ultimately enable more precise and equitable healthcare. We anticipate there will be continued progress in the future, specifically along the themes of innovative data integration, increasing sophistication of computational modeling strategies, and increasing accessibility by focusing on open-source analysis pipelines and generating new standards to improve health equity.

Acknowledgements

We would like to thank the other IMIA Yearbook editors and staff for their support and the reviewers for their participation in the selection of best papers for the BTI section.

References

1. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. Nature 2020 Mar;579(7798):193.
2. Moses DA, Metzger SL, Liu JR, Anumanchipalli GK, Makin JG, Sun PF, et al. Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. N Engl J Med 2021 Jul 15;385(3):217-27.
3. Alkenani AH, Li Y, Xu Y, Zhang Q. Predicting Alzheimer's Disease from Spoken and Written Language Using Fusion-Based Stacked Generalization. J Biomed Inform 2021 Jun;118:103803.
4. Létinier L, Jouganous J, Benkebil M, Bel-Létoile A, Goehrs C, Singier A, et al. Artificial Intelligence for Unstructured Healthcare Data:

Application to Coding of Patient Reporting of Adverse Drug Reactions. Clin Pharmacol Ther 2021 Aug;110(2):392-400.
5. Lu MY, Chen TY, Williamson DFK, Zhao M, Shady M, Lipkova J, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature 2021 Jun;594(7861):106-10.
6. Veturi Y, Lucas A, Bradford Y, Hui D, Dudek S, Theusch E, et al. A unified framework identifies new links between plasma lipids and diseases from electronic medical records across large-scale cohorts. Nat Genet 2021 Jul;53(7):972-81.
7. Eijsbouts C, Zheng T, Kennedy NA, Bonfiglio F, Anderson CA, Moutsianas L, et al. Genome-wide analysis of 53,400 people with irritable bowel syndrome highlights shared genetic pathways with mood and anxiety disorders. Nat Genet 2021 Nov;53(11):1543-52.
8. Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. Nat Genet 2021 Mar;53(3):392-402.
9. Giannoula A, Centeno E, Mayer MA, Sanz F, Furlong LI. A system-level analysis of patient disease trajectories based on clinical, phenotypic and molecular similarities. Bioinformatics 2021 Jun 16;37(10):1435-43.
10. Liu Y, Elsworth B, Erola P, Haberland V, Hemani G, Lyon M, et al. EpiGraphDB: a database and data mining platform for health data science. Bioinformatics 2021 Jun 9;37(9):1304-11.
11. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat Genet 2021 Nov;53(11):1527-33.
12. Wei Q, Ramsey SA. Predicting chemotherapy response using a variational autoencoder approach. BMC Bioinformatics 2021 Sep 22;22(1):453.
13. The 100,000 Genomes Project Pilot Investigators. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. N Engl J Med 2021 Nov 11;385(20):1868–80.
14. Ramoni RB, Mulvihill JJ, Adams DR, Allard P, Ashley EA, Bernstein JA, et al. The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. Am J Hum Genet 2017 Feb 2;100(2):185-92.
15. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 2019 Apr;51(4):584-91.
16. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, et al. Improving reporting standards for polygenic scores in risk prediction studies. Nature 2021 Mar;591(7849):211-9.

Correspondence to:
Mary Lauren Benton
One Bear Place #97141
Waco, TX, 76798, USA
E-mail: marylauren_benton@baylor.edu

Scott McGrath
5689 Cattle Drive
Missoula, MT, 59808, USA
E-mail:smcgrath@berkeley.edu