Predictions, Pivots, and a Pandemic: a Review of 2020's Top Translational Bioinformatics Publications

Scott P. McGrath[1], Mary Lauren Benton[2], Maryam Tavakoli[3], Nicholas P. Tatonetti[4]
[1] CITRIS Health, University of California Berkeley, Berkeley, CA, USA
[2] Department of Computer Science, Baylor University, Waco, TX, USA
[3] MTERMS Lab, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[4] Department of Biomedical Informatics, Columbia University, New York, NY, USA

Summary

Objectives: Provide an overview of the emerging themes and notable papers which were published in 2020 in the field of Bioinformatics and Translational Informatics (BTI) for the International Medical Informatics Association Yearbook.

Methods: A team of 16 individuals scanned the literature from the past year. Using a scoring rubric, papers were evaluated on their novelty, importance, and objective quality. 1,224 Medical Subject Headings (MeSH) terms extracted from these papers were used to identify themes and research focuses. The authors then used the scoring results to select notable papers and trends presented in this manuscript.

Results: The search phase identified 263 potential papers and central themes of coronavirus disease 2019 (COVID-19), machine learning, and bioinformatics were examined in greater detail.

Conclusions: When addressing a once in a centruy pandemic, scientists worldwide answered the call, with informaticians playing a critical role. Productivity and innovations reached new heights in both TBI and science, but significant research gaps remain.

Keywords

COVID-19, machine learning, bioinformatics

Yearb Med Inform 2021:219-25
http://dx.doi.org/10.1055/s-0041-1726540

1 Introduction

Each year in the International Medical Informatics Association (IMIA) Yearbook a survey manuscript reviewing notable papers and trends in the field of Bioinformatics and Translational Informatics (BTI).The advancement of knowledge in other areas of BTI continued on, despite the focus being applied to coronavirus disease 2019 (COVID-19) and disruptions to research and work due to precautionary shut-downs. Machine learning and drug repositioning continue to be hot topics, continuing a trend seen in the 2020 Yearbook of Medical Informatics [1]. Significant upheaval occurred over the past year, but there are plenty of published works worthy of praise.

In this year's search, we found exciting pairings of machine learning with systematic immunogenic profiling [2], adapting and integrating multiple data modalities to study disease [3], and examples of drug design and discovery tools in an effort to accelerate treatment options and targets for COVID-19 vaccines [4]. With machine learning, we witnessed an expansion of applying interpretation to a variety of tool sets and the continued concern about data security, privacy, and bias. With bioinformatics, there has been a massive increase in the use of single cell gene expression datasets, in line with the field of molecular and cellular biology as a whole. Drug outcome prediction techniques continue to be refined, and increasing complexity seen in global biobanks are providing richer datasets. However, the need to diversify the populations in these datasets still remains a priority.

For the year 2020, bioinformatics, science, and life in general was disrupted by the COVID-19 global pandemic. The scientific community did not retreat, and in fact, rose to meet the challenge. By collaborating and accelerating the dissemination of scientific knowledge at a pace never seen before, enormous strides were achieved in understanding COVID-19 and how to combat its spread. Informatics methods were often central to the execution, analysis, and presentation of these results. We will take some time to reflect on both the positive and negative outcomes of some of those changes.

2 Methods

We relied on a literature review activity, which serves as the foundation of the translational and bioinformatics year in review presentation at the American Medical Informatics Association (AMIA) Informatics Summit. This has been a recurring annual presentation given over the past decade and is a good barometer for notable papers and trends in the field [5].

In this year's effort, a team of 16 students and young informatics professionals aggregated papers published from December 2019 until January of 2021. The following query was used to search for manuscripts and modified as needed by members of the team:

(sign OR symptom OR disease OR drug) and (genome OR protein OR small molecule OR RNA OR DNA) AND (computer OR informatics OR statistics)

Our initial query identified 263 papers. The group then graded the manuscripts with a rubric that evaluated informatics novelty in their methods and techniques, topic importance, and overall quality. We used this corpus to identify the manuscripts which highlight some of the trends from this year. Trends were identified by using the Medical Subject Headings (MeSH) on Demand website to capture the MeSH terms from the papers. A total of 1,224 MeSH terms were identified from this step. A python script was then used to cluster the terms and identify themes. Table 1 presents the top 10 MeSH terms based on frequency count, and Table 2 shows the top ten themes which emerged from our corpus.

Table 1 MeSH terms frequency.

Term	Paper count	% of Papers
Machine Learning	49	20
Pharmaceutical Preparations	44	17.96
Algorithms	42	17.14
Genomics	41	16.73
Neoplasms	37	15.1
Phenotype	33	13.47
Transcriptome	31	12.65
SARS-CoV-2	26	10.61
COVID-19	26	10.61
Electronic Health Records	23	9.39
Animals	21	8.57

Table 2 Paper themes.

Term	Paper count	% of Papers
Investigative Techniques	208	84.9
Environment and Public Health	163	66.53
Information Science	155	63.27
Health Care Quality, Access, and Evaluation	143	58.37
Genetic Phenomena	133	54.29
Natural Science Disciplines	112	45.71
Mathematical Concepts	109	44.49
Amino Acids, Peptides, and Proteins	76	31.02
Neoplasms	74	30.2
Health Care Facilities, Manpower, and Services	73	29.8
Health Services Administration	69	28.16

3 Results

3.1 SARS CoV-2

The scope of this paper is to perform a survey of the literature from the past year in the areas of bioinformatics and translational informatics. However, we believe that before starting any recent survey of scientific literature, one must address the largest sudden health crisis in modern history.

3.1.1 A Pandemic Arrives

The World Health Organization (WHO) formally declared coronavirus disease 2019 (COVID-19) a Public Health Emergency of International Concern (PHEIC) on January 30th 2020 [6]. PHEICs are the WHO's highest level of alarm and set the stage for the year to come. Since 2009, there have been nine events assessed for potential PHEIC declarations with six formal declarations: the 2009 H1N1 pandemic, the 2014 polio decleration, the 2014 Ebola outbreak, the 2018 Kivu Ebola outbreak, and the ongoing COVID-19 pandemic [7]. COVID-19 is not the longest PHEIC (the 2014 polio PHEIC still remains in effect in 2021), but it does stand apart in its global impact. In March of 2021, global cases of COVID-19 had exceeded 126 million and caused 2.77 million deaths worldwide. The largest impacts have been seen in the United States and Brazil, with deaths in excess of 559,000 and 340,000 respectively as of April of 2021 [8]. Comparatively, the swine flu (H1N1) was estimated to cause 284,000 deaths worldwide (from a range of 150,000 to 575,000 deaths) [9]. Global cost estimates of the COVID-19 pandemic have been set at $28 trillion by the International Monetary Fund [10], and the impact to the United States alone is estimated at $16 trillion [11]. This, unsurprisingly, has caused the COVID-19 pandemic to be labeled the worst global crisis since the Great Depression [12].

The ways COVID-19 has impacted daily life, science included, have been profound. Changes observed in the publication of scientific manuscripts were of particular relevance to our topic here. Scientific globalism suddenly found a largely unfettered path, a heightened focus on a singular topic, and a rich variety of research targets, all with a growing sense of urgency [13].

Scientists worldwide engaged in a collective action that became the largest research pivot in modern science. The pace of research across many fronts was astounding, with massive intellectual horsepower harnessed in this effort. Within one month of the first COVID-19 outbreak in Wuhan, China, in December of 2019, there were multiple full viral genomes sequenced [14, 15]. Vaccine development typically faces a 10-15 year research and testing window [16]. In 1967, the mumps vaccine was developed just in just four years, a record that would stand for over 50 years [17]. Less than a year into the COVID-19 pandemic, 19 vaccine candidates yielded two different and highly effective vaccines [18]. By March of 2021,

there were 76 SARS-COV-2 vaccines in clinical trials and six vaccines approved for emergency use [19]. Scientific publications on the pandemic also reached an unprecedented level. New curated literature sites emerged, like LitCovid, which includes over 116,000 COVID-19 articles as of early April 2021 [20].

3.1.2 Scientific Publishing's Transmutation

The scientific publishing industry also had to adapt in extraordinary ways. With the world's research focus targeting a single topic, there was sudden deluge of paper submissions. For context, since its discovery in 1976, there have been ~9,700 Ebola-related papers published [21]. According to LitCovid over the past year (March 16th 2020 – March 14th 2021), there has been an average of 2,075 COVID papers published per week, with 4,322 appearing in the week of August 24th alone. The only significant dip occurred the week of Christmas (December 21st – December 20th), where only 1,057 new papers came out.

Publishers adopted several different techniques to help streamline the publication pipeline. The journal eLife announced it would cut back on requests for additional experiments during revisions, suspend revision deadlines, and require all submissions to post preprints to bioRxiv or medRxiv [22]. The Royal Society Open Publishing recruited a group of 700 reviewers who committed to reviewing fast-tracked COVID-19 papers in 24 to 48 hours [23]. Efforts to expedite the publication process were found to be very effective across the board. Typically, a biomedical manuscript takes a median of 100 days from submission to acceptance [24]. Studies found that the time between submission and publication for COVID-19 papers decreased by 49% on average [23]. Palayew et al. found there was a 6-day median time for submission to publication in the early stages of the pandemic [24]. This highlights the demand for the most recent data on COVID-19 and the lengths publishers went to ensure data reached scientists and medical professionals quickly.

Demand for the newest information on SARS-COV-2 was not contained to scientific circles. The general public was also ravenous for any new material they could find. The social web aggregate site Reddit.com had two dedicated communities, known as subreddits, materialize during the pandemic: /r/Coronavirus[1] and /r/COVID-19[2]. The /r/Coronavirus subreddit has over 2.36 million members and is dedicated to general information and news about the pandemic. The sister subreddit, /r/COVID-19, was focused on the emerging science on the virus and had over 317k members. The science-focused /r/COVID-19 subreddit had additional rules for sharing material and was more heavily moderated. The massive interest in pre-print servers would often be reflected in these communities, as members would share and discuss the latest pre-print manuscripts in parallel with the latest published papers. The enthusiasm for the science is a bright spot to appear from this pandemic, with younger generations expressing more interest in STEM careers [25]. However, this enthusiasm may be somewhat tempered by concerns over the rapid pace of pre-print and publication and the potential for some corners to be cut.

3.1.3 Pitfalls and Pratfalls

For all the advancement and acceleration of the science focused on COVID-19, there were significant errors caused by removing some of the traditional guardrails in scientific publication. The website Retraction Watch, which monitors retracted manuscripts, has been tracking COVID-19 papers and noted 75 fully retracted papers, 11 retracted to journal error, four retracted and reinstated, and five flagged with expressions of concern [26]. Pre-print servers like medRxiv[3] and bioRxiv[4] were platforms to help accelerate publications and witnessed exponential growth during this pandemic [27]. However, concerns about medical preprints were validated as some papers went viral before there was adequate review [28]. There was a pre-print paper about seroprevalence in Santa Clara County that got national media attention when it first appeared on April 17th, 2020 [29]. However, just a few days later, people were expressing serious concerns about potential flaws in the study [30], but only after it had captured the attention of the general public [31]. Traditional peer review should have addressed these concerns prior to publication, but the new and faster process may have led to more errors by reviewers and editors. Rushed and flawed papers were not the only concerning outcome from this pandemic. There are signs that the gender gap in science may be further exacerbated, as female scientists, particularly those with young dependents, reported significant declines in the time they could devote to their research over the past year, which could impact their careers for years to come [32]. A period of reflection will be needed to further identify what elements helped advance science during this pandemic, and what issues require repair or removal to prevent additional harm in the future. This sets the stage for the environment we encountered when beginning our survey of bioinformatics and translational informatics papers. COVID-19 caused tectonic shifts in how science and the world adjusted during a modern pandemic. Scientific information saw the arrival of new pathways for dissemination. While the impact COVID-19 has been profound, we do not want it to steal the spotlight from other notable papers and trends from the past year. After reviewing the MeSH term frequency results in Table 1, we decided to organize the manuscripts we wanted to highlight into two categories: machine learning and bioinformatics.

3.2 Machine Learning

We reviewed novel machine learning methods proposed by the top-scored manuscripts with Information System (L01) and Mathematical Concepts (G-17) MeSH headers and identified a few significant perspectives to further discuss in this section.

3.2.1 Representation

Designing a meaningful and suitable representation for the data is one of the most crucial steps in a machine learning pipeline. It takes a lot of time, hypothesis analysis, and domain expertise to engineer meaningful and useful features. Recent deep learning

[1] https://reddit.com/r/Coronavirus
[2] https://reddit.com/r/COVID-19
[3] https://medrxiv.org
[4] https://biorxiv.org

models have offered automatic feature extraction potentials with relatively high performance. Nevertheless, it is extremely crucial to interpret and validate the extracted features properly.

On this year's top scored manuscripts, using embedding and distributed representation remains a popular alternative or addition to classic feature engineering in predictive tasks. The representations are mainly extracted by deep learning [33-38] or latent probabilistic [35] methods. These distributed representations, i.e., embedding, are used to encode various modalities of data, including gene expressions [39, 40], events [36], images [33], and other relational graph data [37, 41]. The embedding methods are data-driven representations that can capture semantic and contextual information and incorporate them into a numerical representation. However, the high dependency of data-driven methods on data quality and the detachment of domain knowledge and validation methods from the feature extraction process suggests a broad range of potential improvements for the research in this area.

In some drug-related studies, graph convolution network variations (GCN) [42] are used to incorporate domain knowledge of topological chemical structures into the representation learning process. Use of GCN in DeepCDR [43] and use of directed-message passing deep neural network model [44] for antibiotic drug discovery [37] are among these practices. In multimodal studies [41, 45, 46] the information fusion is designed in a graph-based form according to a domain-driven information flow. Wang et al. proposed a bipartite GCN for drug re-purposing prediction, which accounts for the central role of proteins in drug-disease association [41]. These methods are examples of a more general direction in incorporating the domain knowledge to refining the data-driven approaches.

3.2.2 Interpretation

It is notable that in many studies with deep learning, interpretation approaches were applied either by using toolsets such as SHapley Additive exPlanations (SHAP) [47] or by applying a parallel traditional machine learning method. Zhang et al. used a surrogate support vector machine (SVM) for convolution neural network predictions as an interpretation method in a pyrazinamide resistance prediction study to identify important genetic factors for Mycobacterium Tuberculosis [48]. Smedley et al. trained a transformer model and used gene masking and saliency to interpret and understand the mapping between gene and MRI image traits of cancer tumors [49].

In a pioneering article by Ashdown et al., informatics and molecular biology were integrated to produce a system for predicting and evaluating antimalarial drug-action [33]. While the goal of the study itself is laudable, the execution is what makes it so notable. In this study, the authors use laboratory experiments to generate fluorescence imaging data of normal plasmodium falciparum cell growth. They first demonstrated the use of deep neural networks (DNN) to process this data into an interpretable quantitative feature that couples tightly with the cell cycle. Using this new analytical representation, they then show how disruptions to the cell cycle (by chemical agents, for example) can be easily identified in their new feature. The authors round out the study by using their DNN representation to accurately reveal the mechanisms of action of the chemical agents. This well-written and performed study serves as an exemplar of impactful and understandable neural network-based research.

3.2.3 Data Security, Privacy, and Bias Concern

The growing demand for data-centered analyses raises two important concerns. On the one hand, the prediction bias is caused by the models trained on datasets that are not representative of all race and population characteristics. This issue naturally calls for a more systematic data collection and data sharing practice. On the other hand, it remains a significant concern for the institutions to preserve individual and population-level information privacy and prevent unintended information leakage during this data era. Gao et al. suggested transfer learning as an alternative method for mixture and stratification-based models for partial bias recovery [34]. The authors elegantly demonstrate the utility of transfer learning to address underrepresentation in existing data and how to identify its source. Other studies provide solutions for a better data sharing practice and moving toward federated machine learning [50] methods to preserve security [4] and privacy [51, 52] while seeking data-centered research.

3.3 Bioinformatics

One of the main themes from our highly-ranked bioinformatics papers was the use of informatics to decipher data from more advanced experimental techniques. In order to better capture relevant variability in traits, single-cell gene expression datasets are becoming increasingly common. Single-cell RNA-sequencing is better able to account for dynamics across cell states, even when using simple linear models. For example, Li et al. predicted breast cancer prognosis by modeling gene expression from single-cell RNA-seq during an important cellular transition [53]. Similarly, other studies leveraged single-cell techniques to study populations of cells across time and space, from mapping pathway activation in response to stimuli [54] and contrasting expression profiles across developmental stages [55] to profiling chromatin accessibility across brain regions [56]. Ultimately, this shift away from bulk sequencing assays allows for a more nuanced view of multi-omics data, greatly improving our ability to measure the dynamic processes influencing disease progression and outcomes.

Informatics is also commonly applied to develop clinically-relevant prediction models using genomics data. Given the diverse range of -omics datasets available, studies from this year considered novel ways to integrate data from multiple experimental sources in order to build more accurate models and highlight mechanisms underlying disease. One striking example is the multi-omics approach designed by Su et al. to tease apart the immunological differences between mild, moderate, and severe COVID-19 [54]. The authors linked gene expression to changes in immune signaling and clinical measures that differentiate between patients with mild versus moderate disease. The biomarkers discovered through this analysis provide a starting point for developing prognostic metrics and targeted treatments for COVID-19.

3.3.1 Drug Development and Clinical Outcomes

Drug development is another major application area for such technology. Predicting drug response for individual patients remains challenging, especially for notoriously heterogeneous diseases such as cancer. Liu et al. developed a deep learning framework to predict drug response by modeling the molecular structures of the drugs themselves [43]. These networks of structural properties were further integrated with networks derived from genomic, transcriptomic, and epigenomic data. The features informed a final model that was able to accurately predict drug response across multiple cancer cell lines, either as the IC50 sensitivity value or classification as sensitive/resistant. When coupled with heterogeneous networks to assist with biological interpretation, predictive multi-omics models (such as the one presented in [43]) are interpretable and can perform well. Combining novel features with existing -omics networks will refine future models as the networks continue to evolve.

Genomics potentially impacts other clinically relevant health outcomes. Christian et al. found that patients prescribed medications that were incongruent with their genetics were more likely to have low adherence to those medications [44]. This study provides an interesting perspective on the impact of genomic information on other aspects of disease treatment, and suggests that including genomic information in routine clinical care can positively impact health behaviors.

It remains important to disentangle the effects of genetic variation on disease, especially variation in non-protein-coding genomic regions thought to regulate the expression of genes. Mediated expression score regression is a new approach that aims to quantify the contribution of variants to disease by calculating the proportion of disease heritability mediated by gene expression [57]. Although the absolute value is low, the authors found that a significant proportion of disease heritability from GWAS is mediated by gene expression in cis. Similarly, PhenomeXcan linked functional genomics and transcriptomics with trait-associated variation to connect genetically regulated gene expression with phenotype [58]. A deeper understanding of the relationship between genetic variation, gene expression, and phenotype will not only enable further improvements to variant effect prediction algorithms but will also generate useful hypotheses for future analysis.

2020 also saw the rise of whole-omics approaches to understanding SARS-CoV-2 infection. Ramlall et al. discovered a critical role for the complement system in COVID-19 through a hybrid analysis combining clinical data from EHRs with genomic data from the UK Biobank [59]. Given the urgency of the COVID-19 pandemic, researchers turned en masse to informatics and data-driven approaches to find possible therapeutics. Studies that integrated chemical informatics based lead prioritization were quite notable. Panda et al. conducted exhaustive molecular dynamics simulations to several compounds with activity against SARS-CoV-2's viral receptor binding domain [4]. The authors used available data in ChEMBL (a database of compound-target activities) to identify 38 drug-like compounds with activity against coronavirus targets. They then followed up with molecular dynamics models to identify the specific binding pockets and possible mechanisms of action. This type of rapid therapeutic hypothesis generation is made possible by the tireless work of informaticians over the past 20 years to structure, organize, and release data and analytical methods.

3.3.2 Biobanks

With the continued growth of EHR-linked biobanks, increasing numbers of individuals are available with matched genomic and clinical data. Algorithms applied to these datasets can define populations based on similar attributes and highlight shared disease biology. For example, Cortes et al. clustered patients in the UK Biobank based on disease associations derived from TreeWAS [60]. Similar to the multi-omics approaches described earlier, the authors leveraged gene ontology hierarchies to implicate specific underlying biological processes in the disease clusters. Genetic risk scores applied to individual clusters revealed separation based on comorbidities and biological processes, both of which provide insight into disease sub-phenotypes and potential avenues of treatment. This article highlights the continued movement towards incorporating genomic data to improve our clinical understanding of disease.

Although many EHR-linked biobanks exist, individual-level data is not widely shared between sites due to patient privacy concerns. However, data sharing between biobanks would increase power for informatics studies and enable larger research efforts. Statistical methods may be able to overcome the challenges involved with data sharing. For example, Sum-Share is a method developed to detect pleiotropic genetic variants without requiring access to individual-level data [61]. Instead, the approach uses only summary statistics from multiple EHR-linked biobanks to detect pleiotropic effects. The authors demonstrate that this method detects pleotropic variants with the same accuracy as a full analysis of individual-level data and increased power compared to PheWAS approaches. This work demonstrates the potential for novel informatics approaches to expand the universe of accessible data and improve power for association studies without compromising patient privacy.

3.3.3 Genomic Diversity

One theme was notable for its absence from most of the top-scored articles discussed here. It is well documented that historical biases in data collection and analysis have led to the overrepresentation of populations of European descent in genomic studies [62-64]. Health disparities can result from the lack of diversity in existing genomic datasets, especially when computing polygenic risk scores for future clinical use [65, 66]. The authors of a polygenic risk score for glaucoma mentioned the need to develop and validate such scores in additional populations to ensure generalizability [67]. However, despite the use of genetic risk scores and other forms of predictive modeling based on genomic data in other articles, discussion of diversity and health disparities is not at the forefront. In order to make equitable advances in healthcare moving forward, we must consider potential historical biases

in the underlying datasets and prioritize the inclusion of underrepresented populations in modeling and validation efforts. This is especially true in times of global crisis as we have witnessed this past year. In the meantime, machine learning techniques, such as transfer learning, may help to mitigate some of these disparities while we continue to push for increased diversity in our datasets [34].

4 Conclusion

Informatics, science, and life at large have been forever shifted by the global coronavirus pandemic, SARS-CoV-2. For science generally, we have witnessed unprecedented productivity, made possible by the groundwork laid by a generation of informaticians. In this review, we highlight some of the year's most influential and inspiring informatics work. These works address the most important challenges of our time: the pandemic, underrepresentation bias, high-through mult-omics integration – among others. Even so, significant research gaps remain. Biases in biomedical data limit our understanding of disease and contribute to higher morbidity and mortality for minority populations. Global warming and climate change will have severe impacts on the incidence of disease and the equitable distribution of healthcare. If these past 14 months have demonstrated anything, however, it is that the bioinformatics community is ready and willing to face these challenges head on.

Acknowledgements

The authors would like to thank the 2021 AMIA Year in Review research team for their work developing the source material for this paper and Melanie McGrath (PhD, LAT, ATC) for aid in the preperation of this manuscript.

References

1. Smaïl-Tabbone M, Rance B. Contributions from the 2019 Literature on Bioinformatics and Translational Informatics, Yearb Med Inform 2020(29):188–92.
2. Shrock E, Fujimura E, Kula T, Timms RT, Lee IH, Leng Y, et al. Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. Science 2020;370(6520):eabd4250.
3. Su Y, Chen D, Yuan D, Lausted C, Choi J, Dai CL, et al. Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. Cell 2020;183(6):1479-95.e20.
4. Panda PK, Arul MN, Patel P, Verma SK, Luo W, Rubahn HG, et al. Structure-based drug designing and immunoinformatics approach for SARS-CoV-2. Sci Adv 2020;6(8):eabb8097.
5. Romano JD, Bernauer M, McGrath SP, Nagar SD, Freimuth DD. A Decade of Translational Bioinformatics: A Retrospective Analysis of "Year-in-Review" Presentations. AMIA Jt Summits Transl Sci Proc [Internet] 2019 May 6 [cited 2021 March 24];2019;335–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31258986
6. World Health Organization (WHO). Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV), (2020). Available from: https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov)
7. Mullen L, Potter C, Gostin LO, Cicero A, Nuzzo JB. An analysis of International Health Regulations Emergency Committees and Public Health Emergency of International Concern Designations. BMJ Glob Health 2020;5(6):e002502.
8. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20(5):533–4.
9. Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng PY, et al. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: A modelling study. Lancet Infect Dis 2012;12 (9):687–95.
10. International Monetary Fund. World Economic Outlook Update: June 2020 - A Crisis Like No Other, An Uncertain Recovery. World Econ Outlook Reports 2020;20. Available from: https://www.imf.org/en/Publications/WEO/Issues/2020/06/24/WEOUpdateJune2020
11. Cutler DM, Summers LH. The COVID-19 Pandemic and the $16 Trillion Virus, JAMA 2020;324(15):1495–6.
12. Zumbrun J. Coronavirus Slump Is Worst Since Great Depression. Will It Be as Painful? Wall Str J 2020. Available from: https://www.wsj.com/articles/coronavirus-slump-is-worst-since-great-depression-will-it-be-as-painful-11589115601
13. Lee JJ, Haupt JP. Scientific globalism during a global crisis: research collaboration and open access publications on COVID-19. High Educ (Dordr) 2020 24;1-18.
14. Institut Pasteur. Whole genome of novel coronavirus, 2019-nCoV, sequenced. ScienceDaily 2020. Available from: https://www.sciencedaily.com/releases/2020/01/200131114748.htm
15. Zhang Y-Z. Novel 2019 coronavirus genome - SARS-CoV-2 coronavirus - Virological (n.d.) [cited 2021 March 15]. Available from: https://virological.org/t/novel-2019-coronavirus-genome/319
16. Vaccine Development. Testing, and Regulation | History of Vaccines, (n.d.) .) [cited 2021 March 14]. Available from: https://www.historyofvaccines.org/content/articles/vaccine-development-testing-and-regulation
17. CDC. Pinkbook | Mumps | Epidemiology of Vaccine Preventable Diseases | CDC, (n.d.) [cited 2021 March 14]. Available from: https://www.cdc.gov/vaccines/pubs/pinkbook/mumps.html#vaccines
18. Kaur SP, Gupta V. COVID-19 Vaccine: A comprehensive status report. Virus Res 2020;288:198114.
19. Zimmer C, Corum J, Wee S-L. Covid-19 Vaccine Tracker Updates. New York Times [Internet] 2021[cited 2021 March 14]. Available from: https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html
20. Chen Q, Allot A, Lu Z. LitCovid: An open database of COVID-19 literature. Nucleic Acids Res 2021;49(D1) D1534–D1540.
21. Yong E. How Science Beat the Virus. Atl [Internet] 2020. Available from: https://www.theatlantic.com/magazine/archive/2021/01/science-covid-19-manhattan-project/617262/
22. Eisen MB, Akhmanova A, Behrens TE, Weigel D. Publishing in the time of COVID-19. Elife 2020;9:e57162.
23. Horbach SPJM. Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. Quant Sci Stud 2020;1:1056–67.
24. Palayew A, Norgaard O, Safreed-Harmon K, Andersen TH, Rasmussen LN, Lazarus JV. Pandemic publishing poses a new COVID-19 challenge. Nat Hum Behav 2020;4(7):666–9.
25. EngineeringUK. Young people and Covid-19 : How the pandemic has affected careers experiences and aspirations. [Internet] 2020 [cited 2021 March 23]. Available from: https://www.voced.edu.au/content/ngv%3A87723
26. Retracted coronavirus (COVID-19) papers – Retraction Watch [Internet](n.d.) [cited 2021 March 23]. Available from: https://retractionwatch.com/retracted-coronavirus-covid-19-papers/
27. Vlasschaert C, Topf JM, Hiremath S. Proliferation of Papers and Preprints During the Coronavirus Disease 2019 Pandemic: Progress or Problems With Peer Review? Adv Chronic Kidney Dis 2020;27(5):418–26.
28. King A. Fast news or fake news?: The advantages and the pitfalls of rapid publication through preprint servers during a pandemic. EMBO Rep 2020;21(6):e50817.
29. Bendavid E, Mulaney B, Sood N, Shah S, Ling E, Bromley-Dulfano R, et al. COVID-19 antibody seroprevalence in Santa Clara County, California. In J Epidemiol 2021;50(2):410-9.
30. Gelman A. Concerns with that Stanford study of coronavirus prevalence [Internet] 2020. [cited 2021 March 23]. Available from: https://statmodeling.stat.columbia.edu/2020/04/19/fatal-flaws-in-stanford-study-of-coronavirus-prevalence/
31. Vogel G. Antibody surveys suggesting vast undercount of coronavirus infections may be unreliable. Science [Internet] 2020. doi:10.1126/science.abc3831.
32. Myers KR, Tham WY, Yin Y, Cohodes N, Thursby JG, Thursby MC, et al. Unequal effects of the COVID-19 pandemic on scientists. Nat Hum Behav 2020;4(9):880–3.
33. Ashdown GW, Dimon M, Fan M, Terán FSR,

Witmer K, Gaboriau DCA, et al. A machine learning approach to define antimalarial drug action from heterogeneous cell-based screens. Sci Adv 2020;6(39):eaba9338.
34. Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. Nat Commun 2020;11(1):5131.
35. Ha J, Park C, Park C, Park S. IMIPMF: Inferring miRNA-disease interactions using probabilistic matrix factorization. J Biomed Inform 2020;102:103358.
36. Raket LL, Jaskolowski J, Kinon BJ, Brasen JC, Jönsson L, Wehnert A, et al. Dynamic ElecTronic hEalth reCord deTection (DETECT) of individuals at risk of a first episode of psychosis: a case-control development and validation study, Lancet Digit Health 2020;2(5):e229–e239.
37. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell 2020;180(4)688–702.
38. Vinks AA, Punt NC, Menke F, Kirkendall E, Butler D, Duggan TJ, et al. Electronic Health Record–Embedded Decision Support Platform for Morphine Precision Dosing in Neonates. Clin Pharmacol Ther 2020;107(1):186-94.
39. Kuang S, Wei Y, Wang L. Expression-based prediction of human essential genes and candidate lncRNAs in cancer cells. Bioinformatics 2021;37(3):396-403.
40. Tripodi IJ, Callahan TJ, Westfall JT, Meitzer NS, Dowell RD, Hunter LE. Applying knowledge-driven mechanistic inference to toxicogenomics. Toxicol In Vitro 2020;66:104877.
41. Wang Z, Zhou M, Arnold C. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing. Bioinformatics 2020;36(Suppl_1):i525–i533.
42. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. ArXiv Prepr 2016. ArXiv1609.02907.
43. Liu Q, Hu Z, Jiang R, Zhou M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics 2020;36(Suppl_2):i911–i918.
44. Christian C, Borden BA, Danahey K, Yeo KTJ, van Wijk XMR, Ratain MJ, et al. Pharmacogenomic-Based Decision Support to Predict Adherence to Medications. Clin Pharmacol Ther 2020;108(2):368–76.
45. Fatima N, Rueda L. iSOM-GSN: an integrative approach for transforming multi-omic data into gene similarity networks via self-organizing maps. Bioinformatics 2020;36(15):4248–54.
46. Hasan MM, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B. HLPpred-Fuse: Improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics 2020;36(11):3350–6.
47. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst 2017-December. [Internet] 2017 [cited 2021 April 10]:4766–75. Available from: http://arxiv.org/abs/1705.07874
48. Zhang A, Teng L, Alterovitz G. An explainable machine learning platform for pyrazinamide resistance prediction and genetic feature identification of Mycobacterium tuberculosis, J Am Med Inform Assoc 2021;28(3):533–40.
49. Smedley NF, El-Saden S, Hsu W. Discovering and interpreting transcriptomic drivers of imaging traits using neural networks. Bioinformatics 2020;36(11):3537–48.
50. Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RTH, Jochems A, et al. Systematic review of privacy-preserving distributed machine learning from federated databases in health care. JCO Clin Cancer Inform 2020;4:184–200.
51. Li R, Duan R, Zhang X, Lumley T, Pendergrass S, Bauer C, et al. Lossless integration of multiple electronic health records for identifying pleiotropy using summary statistics. Nat Commun 2021;12(1):168.
52. Raisaro JL, Marino F, Troncoso-Pastoriza J, Beau-Lejdstrom R, Bellazzi R, Murphy R, et al. SCOR: A secure international informatics infrastructure to investigate COVID-19. J Am Med Inform Assoc 2020;27(11):1721–6.
53. Li X, Liu L, Goodall GJ, Schreiber A, Xu T, Li J, et al. A novel single-cell based method for breast cancer prognosis. PLOS Comput Biol 2020;16(8):e1008133.
54. Cao J, Zhou W, Steemers F, Trapnell C, Shendure J. Sci-fate characterizes the dynamics of gene expression in single cells. Nat. Biotechnol 2020;38(8):980–8.
55. Wang S, Zheng Y, Li J, Yu Y, Zhang W, Song M, et al. Single-Cell Transcriptomic Atlas of Primate Ovarian Aging. Cell 2020;180(3):585-600.e19.
56. Corces MR, Shcherbina A, Kundu S, Gloudemans MJ, Frésard L, Granja JM, et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. Nat Genet 2020;52(11):1158–68.
57. Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat Genet 2020;52(6):626–33.
58. Pividori M, Rajagopal PS, Barbeira A, Liang YMelia O, Bastarache L, et al; GTEx. Consortium. PhenomeXcan: Mapping the genome to the phenome through the transcriptome. Sci Adv 2020;6(37):eaba2083.
59. Ramlall V, Thangaraj PM, Meydan C, Foox J, Butler D, Kim J, et al. Immune complement and coagulation dysfunction in adverse outcomes of SARS-CoV-2 infection. Nat Med 2020;26(10):1609–15.
60. Cortes A, Albers PK, Dendrou CA, Fugger L, McVean G. Identifying cross-disease components of genetic risk across hospital data in the UK Biobank. Nat Genet 2020;52(1):126–34.
61. Li R, Duan R, Zhang X, Lumley T, Pendergrass S, Bauer C, et al. Lossless integration of multiple electronic health records for identifying pleiotropy using summary statistics Nat Commun 2021;12(1):168.
62. Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, et al. Prioritizing diversity in human genomics research. Nat Rev Genet 2018;19(3):175–85.
63. Kelly DE, Hansen MEB, Tishkoff SA. Global variation in gene expression and the value of diverse sampling. Curr Opin Syst Biol 2017;1:102–8. doi.
64. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature 2016;538(7624):161–4.
65. 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;51(4):584–91.
66. Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. Variable prediction accuracy of polygenic scores within an ancestry group. Elife 2020;9:e48376.
67. Craig JE, Han X, Qassim A, Hassall M, Cooke Bailey JN, Kinzy TG, et al. Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. Nat Genet 2020;52(2):160–6.

Correspondence to:
Scott McGrath, PhD
E-mail: smcgrath@berkeley.edu

Nicholas Tatonetti, PhD
E-mail: nick.tatonetti@columbia.edu