Precision Medicine: Using Artificial Intelligence to Improve Diagnostics and Healthcare

Roxana Daneshjou
Departments of Dermatology and Biomedical Data Science, Stanford School of Medicine
Stanford, CA, United States
Email: roxanad@stanford.edu

Steven E Brenner
Departments of Bioengineering, Molecular & Cell Biology, Plant & Microbial Biology, University of California at Berkeley
Berkeley, CA, United States
Email: brenner@combio.berkeley.edu

Jonathan H Chen
Department of Medicine and Center for Biomedical Informatics Research, Stanford School of Medicine
Stanford, CA, United States
Email: jonc101@stanford.edu

Dana C Crawford
Department of Population and Quantitative Health Sciences, Case Western Reserve University
Cleveland, Ohio, United States
Email: dcc64@case.edu

Samuel G Finlayson
Harvard School of Medicine and Massachusetts Institute of Technology
Cambridge, MA, United States
Email: samuel_finlayson@hms.harvard.edu

Łukasz Kidziński
Bioclinica and Stanford University
Stanford, CA, United States
Email: lukasz.kidzinski@stanford.edu

Martha L Bulyk
Departments of Medicine and Pathology, Brigham and Women's Hospital and Harvard Medical School
Cambridge, MA United States
Email: mlbulyk@genetics.med.harvard.edu

The continued generation of large amounts of data within healthcare—from imaging to electronic medical health records to genomics and multi-omics —necessitates tools and methods to parse and interpret these data to improve healthcare outcomes. Artificial intelligence, and in particular deep learning, has enabled researchers to gain new insights from large scale and multimodal data. At the 2022

Pacific Symposium on Biocomputing (PSB) session entitled “Precision Medicine: Using Artificial Intelligence to Improve Diagnostics and Healthcare”, we showcase the latest research, influenced and inspired by the idea of using technology to build a more fair, tailored, and cost-effective healthcare system after the COVID-19 pandemic.

Keywords: Artificial intelligence (AI); Machine learning; Genomics; Multi-omics

1. Introduction

The COVID-19 pandemic has highlighted the longstanding inequities and fractures within the healthcare system.(Bambra et al., 2020) Based on insights from the pandemic, the future of healthcare must be more fair, more tailored, and more cost-effective. At the same time, rapid advances in technology across multiple domains has led to the generation of large amounts of data—from medical imaging to sequencing and genomics technologies to electronic health records (EHRs). The ultimate goal of precision medicine is to leverage these large datasets in order to improve healthcare delivery; however, this requires the development of methods and tools for parsing large-scale, multi-modal data.

Here, we review some of the research aimed at addressing these needs with examples from the accepted submissions for the Precision Medicine: Using Artificial Intelligence to Improve Diagnostics and Healthcare session at the Pacific Symposium on Biocomputing (PSB) 2022. Recent trends from this research indicate a growing body of work leveraging artificial intelligence (AI) for analyzing complex medical data, such as interpreting medical imaging or forecasting infectious diseases, and the development of more sophisticated tools for gaining insight into genomics and multi-omics data.

2. AI-driven tools for addressing problems in healthcare

In the last two years, there has been a rapid increase in the number of FDA approved AI tools for medical applications.(FDA, 2021; Wu et al., 2021) While AI is a broad term, we are referring to methods that allow machines to simulate ‘intelligent’ behavior; this term encompasses machine learning and deep learning.(Stone et al., 2016)

Deep learning methods have dominated automated analysis of medical images, and they are routinely used for research on X-ray, CT, and MRI image data.(Esteva et al., 2019) For cardiac imaging, assessing the function of the left ventricle is important for diagnosing patients with cardiovascular disease and assessing their level of cardiac risk.(Duffy et al.) Duffy et al. (2022) developed a method that predicts a 3D depth-map from 2D videos of echocardiograms, which allows the evaluation of left ventricular function and leads to better performance than previous 2D-based methods.(Duffy et al.) Additionally, their method is more interpretable, and allows for human input and fine-tuning.(Duffy et al.)

Advancements of computational techniques not only allow us to analyze more complex data modalities, but also enable use of sensors that previously had limited role in healthcare research and practice. Vodrahalli et al. (2022) showed that machine learning allows commodity devices like smart-phones to track gaze and near clinical-quality resolution.(Vodrahalli et al., 2022) They developed a set of tasks for assessing visuo-motor diseases and identified a set of tasks most predictive for dysfunction.(Vodrahalli et al., 2022) They showed that the system can discriminate among many common movement-related diseases such as Parkinson's disease, nystagmus, cerebellar ataxia and others., which opens up new opportunities for neuroscience.(Vodrahalli et al., 2022)

Chao et al. (2022) demonstrates the potential for advancing methods to unpack increasingly high-dimensional, heterogeneous biological data.(Chao et al., 2022) They specifically study single-cell sequencing of tissue transcripts, generating data for thousands of cells and tens of thousands of genes.(Chao et al., 2022) This plethora of data opens the possibility for discovery and personalized care, but precision medicine requires respective analytic tools to digest this into interpretable results. In this case, the authors extend beyond traditional machine learning classifiers with signal-extractive neural network architectures with axiomatic feature attribution to produce robust and interpretable means to distinguish signal from noise.(Chao et al., 2022) Specifically they demonstrate in the setting of predicting response of immune checkpoint inhibitors in multiple tissue types and classifying DNA mismatch repair in colorectal cancer.(Chao et al., 2022) Their resulting attribution scores not only validate known biology, but suggest new hypotheses and mechanisms, which may be extended to other hierarchically-structured data.(Chao et al., 2022)

Predicting how a disease will progress in time is one of the key challenges in precision medicine. Cui et al. (2022) analyzed data from over 2000 COVID-19 patients.(Cui et al., 2022) They used Hierarchical Gaussian Processes and Mixture of Experts models to fit trajectories of markers related to disease progression.(Cui et al., 2022) They illustrated predictive performance of temporal models in a case study for albumin, a protein known to be correlated with outcomes in COVID-19 patients.(Cui et al., 2022) Such computational methods open up new opportunities for developing biomarkers, since temporal information is hard to process manually by doctors using EHR data.

To analyze the effect of a treatment on disease progression, Ren et al. (2022) proposed a new regularized matrix factorization method for understanding response to drugs.(Ren et al., 2022) The key problem in this setting is how to combine multiple sources and modalities of information such as chemical structure of drugs, their impact on cellular signaling systems, and cancer cell cellular systems, to predict drug response. Ren et al. (2022) introduced a method that embeds drugs in a latent space and cells in a separate latent space, to find a mapping between the two potentially revealing previously unknown relations.(Ren et al., 2022) They showed the effectiveness of their method in prediction of outcome of chemotherapy in lung cancer patients.(Ren et al., 2022)

Noshad et al (2022) illustrate the potential for AI applications to disrupt medical specialty consultation processes.(Noshad et al., 2022) The authors note the overwhelming decision space and complexity of modern medicine, where patients increasingly depend on consultations with medical specialists.(Noshad et al., 2022) They found patient outcomes are easily compromised as

when they often await months for an in-person consultation visit, and even then, half of those visits do not even have appropriate initial diagnostic tests completed.(Noshad et al., 2022) In this study, they illustrated the capacity for data-driven AI recommender algorithms to predict what a specialist would do for a patient, with greater accuracy and personalization than standard checklists, opening the pathway towards digital consultation systems for artificial intelligence systems to augment human systems.(Noshad et al., 2022)

With the COVID-19 pandemic, infectious disease surveillance has become one of the most important public health tasks. Poonawala-Lohani et al. (2022) developed a novel time series forecasting method, Randomized Ensembles of Auto-regression chains (Reach).(Poonawala-Lohani et al., 2022) They evaluated Reach and previously described methods on influenza-like illness case counts from 2015-2018 in Auckland, New Zealand and found Reach performed better at forecasting influenza-like illnesses.(Poonawala-Lohani et al., 2022) Such tools, if further validated, will be integral to forecasting and monitoring future pandemics.

3. Genomics and multi-omics for improving healthcare

The COVID-19 pandemic highlighted a need for precision medicine, particularly new processes for tailored care and drug discovery. With the continued large scale generation of genomics and multi-omics data, there is an opportunity for the development of tools that can parse this information to discover new insights that improve patient care.

The human reference genome does not reflect human diversity, leading to unmapped non-reference reads with short-read sequencing.(Chrisman et al., 2022) Population-scale whole genome sequencing is becoming more common as costs become more reasonable, highlighting the need for the development of methods to more efficiently map as many reads as possible to maximize the data's potential for equitable downstream research or clinical applications.(Chrisman et al., 2022) To address this need, Chrisman et al. (2022) present a method that maps short read sequences (100 base pairs) to the reference genome using family-based data.(Chrisman et al., 2022) Using a multiplex dataset of ~1,000 families consisting of ~4,000 individuals with 30x whole genome sequencing data, Chrisman et al. (2022) leverage the genetic variation data and inheritance patterns to then match the distribution of the unmapped short reads to determine the most likely location of these reads in the human genome.(Chrisman et al., 2022) The localization is performed with a Hidden Markov Model coupled to a Maximum Likelihood estimator, and an assessment of its performance suggests that the algorithm successfully maps most alternative haplotypes to the genome (96% unmapped short reads mapped with 90% accuracy).(Chrisman et al., 2022) Coupling linkage analysis with newer artificial intelligence methods, the algorithm offers a practical near-term solution for orphaned short reads as the field slowly transitions from short to long-read sequencing at scale.(Chrisman et al., 2022)

After nearly 15 years of genome-wide association studies (GWAS), thousands of single nucleotide variants have been identified as associated with hundreds of human phenotypes and traits.(Nam et al., 2022) A major challenge is using this genotype-phenotype knowledge to clinically predict disease susceptibility.(Nam et al., 2022) In recognition that a single variant only contributes a fraction of risk, methods have been developed to summarize the genetic variants and their associated effect sizes for specific human outcomes of interest.(Nam et al., 2022) Known as

polygenic risk scores (PRSs), these genetic risk summaries have been developed for and applied to patient populations to identify those at greatest risk who could possibly benefit from early intervention strategies.(Nam et al., 2022) PRSs have many known limitations.(Nam et al., 2022) A limitation addressed here is that PRSs are developed for a single outcome of interest and do not account for the complexity of known and unknown co-morbidities.(Nam et al., 2022) Nam et al. (2022) draw parallels between GWAS and today's PRSs and offer a common solution: phenome-wide association studies (PheWAS).(Nam et al., 2022) In PheWAS, the entire phenome or set of phenotypes is interrogated for genotype-phenotype associations as opposed to a single outcome or trait.(Nam et al., 2022) Here, Nam et al. (2022) calculate network-based comorbidity risk scores using the UK Biobank PheWAS summary statistics to first construct the disease-variant heterogeneous multi-layered network and then applied graph-based semi-supervised learning to predict possible patient co-morbidity scores that are then combined using logistic regression.(Nam et al., 2022) With access to individual level genotype data from the Penn Medicine Biobank, Nam et al. (2022) provide proof-of-concept data for myocardial infarction that demonstrate more efficient risk stratification compared with now conventional PRSs.(Nam et al., 2022)

Clinical target, whole exome, and whole genome sequencing are becoming standard diagnostic tools for patients presenting with specific phenotypes. Not yet adopted but becoming widely available in research settings is the use of bulk or single cell RNA-sequencing (scRNA-seq) data that provide variation data at the cell population level.(He et al., 2022) In anticipation of their eventual use in clinical settings, He et al. (2022) have developed a method named CloudPred that uses scRNA-seq data to predict patient phenotypes.(He et al., 2022) CloudPred's input is scRNA-seq data from multiple patients and each patient's data is modeled as a mixture of Gaussians.(He et al., 2022) The modeling provides a set of features that can be used to predict patient phenotypes.(He et al., 2022) CloudPred's performance was evaluated using simulations and actual scRNA-seq data generated on patients with (n=120) and without (n=22) lupus.(He et al., 2022) CloudPred performed as well in predicting binary patient status for lupus compared with other tested methods and outperformed the same methods for predicting the quantitative phenotype of monocyte composition.(He et al., 2022) As might be expected, CloudPred and other comparative methods' performances improve with an increase in patient and cell sample size.(He et al., 2022) This biologically motivated mixture model is available https://github.com/bryanhe/CloudPred.(He et al., 2022)

In "Nonlinear post-selection inference for genome-wide association studies (GWAS)", Slim et al presents a new method for GWAS, kernelPSI, to improve statistical performance and interpretability of GWAS.(Slim et al., 2022) Single nucleotide polymorphism (SNP) based analyses of GWAS can lack statistical power due to the low effect sizes of each individual SNP.(Daneshjou et al., 2013; Slim et al., 2022) Gene-based methods that aggregate SNPs across a gene have been proposed as a way to mitigate this issue.(Slim et al., 2022) Slim et al proposes a methodology for SNP selection using an adaptation of kerenelPSI, which uses kernel-based post-selection inference; their method models epistatic interactions between SNPs and uses post selection inference to identify regions that can help predict a phenotype.(Slim et al., 2022)

Tissue-specific gene expression data coupled with genome-wide single nucleotide variation data such as GTEx have become *the* reference dataset to infer gene expression in genome-wide datasets

that are sans expression data.(Mahoney et al., 2022) Most inference approaches such as PrediXcan do not offer sex-specific models despite evidence that gene expression may differ by sex.(Mahoney et al., 2022) Mahoney et al. (2022) explore the possible value added when sex-specific models are considered using PrediXcan.(Mahoney et al., 2022) Although sex-specific gene expression effects were observed in GTEx, these effects were relatively small and could not be replicated using a gene expression dataset from the 1000 Genomes Project.(Mahoney et al., 2022) Mahoney et al. (2022) suggest that larger gene expression datasets linked to genome-wide data are needed to better characterize and leverage sex-specific effects that underlie tissue-specific gene expression patterns that may ultimately impact human outcomes relevant to precision medicine.(Mahoney et al., 2022)

Kaczmarek et al. (2022) tackled a different aspect of data integration, presenting a novel cancer classifier that uses a multi-omic graph transformer to directly leverage known miRNA-mRNA interactions.(Kaczmarek et al., 2022) The authors first constructed a knowledge graph encoding biological communication and cellular interactions between miRNAs and mRNAs, which served as the backbone for a graph neural network.(Kaczmarek et al., 2022) Their machine learning algorithm then took in patient cancer samples as input, and applied message-passing and attention mechanisms over the above knowledge graph to make new predictions.(Kaczmarek et al., 2022) In addition to achieving impressive accuracy, the method's use of the knowledge graph also allowed for explainability studies that identify important targeting pathways and molecular biomarkers.(Kaczmarek et al., 2022) While applied to mRNA-miRNA datasets, the method can be easily extended to other multi-omic data problems.(Kaczmarek et al., 2022)

Li et al's (2022) work notes the disconnect in precision medicine between pharmacogenetics and disease genetics.(Li et al., 2022) To address this, they analyzed pharmacogenetic and clinical genetics databases (e.g., PharmGKB, Clinvar) to identify overlaps between genes and variants responsible for disease and those responsible for affecting drug response.(Li et al., 2022) They found 26 genes associated with disease that also have a strong pharmacogenetic association, demonstrating the importance of joint analysis between disease and pharmacogenetic domains.(Li et al., 2022)

4. Conclusion

In submissions to the Precision Medicine: Using Artificial Intelligence to Improve Diagnostics and Healthcare session in PSB 2022, we observed the adaptation and development of new computational methods towards a healthcare system that is more streamlined, tailored, and fair. This body of research covers a wide range of topics from novel methods for forecasting influenza-like illnesses to new tools for predicting disease outcomes to methods for discovering new insights from multi-omics data.

References

Bambra, C., Riordan, R., Ford, J., & Matthews, F. (2020). The COVID-19 pandemic and health inequalities. *J Epidemiol Community Health*, *74*(11), 964-968. https://doi.org/10.1136/jech-2020-214401

Chao, S., Brenner, M. P., & Hacohen, N. (2022). Identifying Cell Type-Specific Chemokine Correlates with Hierarchical Signal Extraction from Single-Cell Transcriptomes In. Pacific Symposium on Biocomputing 2022.

Chrisman, B. S., Paskov, K. M., He, C., Jung, J.-Y., Stockham, N., Yigitcan, W., & Wall, D. P. (2022). Towards a More Diverse Human Reference Genome: A Method for Localizing Non-Reference Sequences to the Human Genome In. Pacific Symposium on Biocomputing 2022.

Cui, S., Yoo, E. C., Li, D., Laudanski, K., & Engelhardt, B. E. (2022). Hierarchical Gaussian Processes and Mixtures of Experts to Model COVID-19 Patient Trajectories. In. Pacific Symposium on Biocomputing 2022.

Daneshjou, R., Tatonetti, N. P., Karczewski, K. J., Sagreiya, H., Bourgeois, S., Drozda, K., . . . Altman, R. B. (2013). Pathway analysis of genome-wide data improves warfarin dose prediction. *BMC Genomics*, *14 Suppl 3*, S11. https://doi.org/10.1186/1471-2164-14-S3-S11

Duffy, G., Jain, I., He, B., & Ouyang, D. INTERPRETABLE DEEP LEARNING PREDICTION OF 3D ASSESSMENT OF CARDIAC FUNCTION In. Pacific Symposium on Biocomputing 2022.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., . . . Dean, J. (2019). A guide to deep learning in healthcare. *Nat Med*, *25*(1), 24-29. https://doi.org/10.1038/s41591-018-0316-z

FDA. (2021). *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. Retrieved Oct 3 2021 from https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices?utm_source=STAT+Newsletters&utm_campaign=36f8a74035-health_tech_COPY_01&utm_medium=email&utm_term=0_8cab1d7961-36f8a74035-153888666

He, B., Thomson, M., Subramaniam, M., Perez, R., Yi, C. J., & Zou, J. (2022). CloudPred: Predicting Patient Phenotypes From Single-cell RNA-seq. In. P acific Symposium on Biocomputing 2022.

Kaczmarek, E., Jamzad, A., Imtiaz, T., Nanayakkara, J., Renwick, N., & Mousavi, P. (2022). Multi-Omic Graph Transformers for Cancer Classication and Interpretation. In. Pacific Symposium on Biocomputing 2022.

Li, B., Whirl-Carrillo, M., Wright, M. W., Babb, L., Rehm, H. L., & Klein, T. E. (2022). An Investigation of the Knowledge Overlap between Pharmacogenomics and Disease Genetics. In. Pacific Symposium on Biocomputing 2022.

Mahoney, E., Janve, V., Hohman, T. J., & Dumitreescu, L. (2022). Evaluation of Sex-Aware PrediXcan Models for Predicting Gene Expression. In. Pacific Symposium on Biocomputing 2022.

Nam, Y., Jung, S.-H., Verma, A., Sriram, V., Wong, H.-H., Yun, J.-S., & Kim, D. (2022). netCRS: Network-based comorbidity risk score for prediction of myocardial infarction using biobank-scaled PheWAS data In. P acific Symposium on Biocomputing 2022.

Noshad, M., Jankovic, I., & Chen, J. H. (2022). Clinical Recommender Algorithms to Simulate Digital Specialty Consultations. In. Pacific Symposium on Biocomputing 2022.

Poonawala-Lohani, N., Riddle, P., Adnan, M., & Wicker, J. (2022). A Novel Approach for Time Series Forecasting of Influenza-like Illness Using a Regression Chain Method In. Pacific Symposium on Biocomputing 2022.

Ren, S., Tao, Y., Yu, K., Xue, Y., Schwartz, R., & Lu, X. (2022). Prediction of Cell-Drug Sensitivities Using Deep Learning-based Graph Regularized Matrix Factorization In. Pacific Symposium on Biocomputing 2022.

Slim, L., Chatelain, C., & Azencott, C.-A. (2022). Nonlinear post-selection inference for genome-wide association studies. In. Pacific Symposium on Biocomputing 2022.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., . . . Teller, A. (2016). Artificial Intelligence and Life in 2030." One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. In: Stanford University.

Vodrahalli, K., Filipkowski, M., Chen, T., Zou, J., & Liao, Y. J. (2022). Predicting Visuo-Motor Diseases From Eye Tracking Data. In. Pacific Symposium on Biocomputing 2022.

Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D. E., & Zou, J. (2021). How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med*, *27*(4), 582-584. https://doi.org/10.1038/s41591-021-01312-x