Precision Medicine: Using Artificial Intelligence to Improve Diagnostics and Healthcare

Michelle Whirl-Carrillo
Department of Biomedical Data Science, Stanford School of Medicine
Stanford, CA, United States
Email: mwcarrilllo@stanford.edu

Steven E. Brenner
University of California, 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, OH, United States
Email: dcc64@case.edu

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

David Ouyang
Smidt Heart Institute, Cedars-Sinai Medical Center
Los Angeles, CA, United States
Email: david.ouyang@cshs.org

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

Precision medicine requires a deep understanding of complex biomedical and healthcare data, which is being generated at exponential rates and increasingly made available through public biobanks, electronic medical record systems and biomedical databases and knowledgebases. The complexity

and sheer amount of data prohibit manual manipulation. Instead, the field depends on artificial intelligence approaches to parse, annotate, evaluate and interpret the data to enable applications to patient healthcare At the 2023 Pacific Symposium on Biocomputing (PSB) session entitled "Precision Medicine: Using Artificial Intelligence (AI) to improve diagnostics and healthcare", we spotlight research that develops and applies computational methodologies to solve biomedical problems.

Keywords: Artificial intelligence; Machine learning; Genomics; Multi-omics

1. Introduction

The goal of precision medicine is to tailor medical care to the individual patient, from disease prevention to diagnosis to treatment. It holds the key to improve healthcare for all, diminishing health disparities. The generation of extensive, comprehensive and diverse medical datasets provide the opportunity to develop tools and methods that will advance the medical field through patient-tailored treatment enabling healthcare equity across diverse populations. Below, we summarize research focusing on methodology development and applications to move personalized medicine forward. Based on the accepted submissions for the Precision Medicine: Using Artificial Intelligence (AI) to improve diagnostics and healthcare session at the Pacific Symposium on Biocomputing (PSB) 2023, computational and AI approaches are being used to advance cancer research, aid in pregnancy-related healthcare, reduce bias in biomedical data, enhance medical imaging and improve immunotherapy strategies.

2. AI-driven tools for improving diagnostics and healthcare

As copious amounts of data are generated at rapidly increasing rates, precision medicine research faces the challenge of integrating across the landscape of "multi's", including multi-omics, multi-models, multi-model systems and multi-sample types. (Acosta) The following submissions highlight greatly needed methods for analysis of integrated data across diverse datasets, and one submission addresses population bias in data.

Hashim et al. developed a self-supervised learning approach for cancer type classification based on multi-omics cancer data, particularly for unannotated or unlabeled data. They applied their pre-training paradigm to The Cancer Genome Atlas pan-cancer dataset. Benefits to their approach is that it can handle missing omics data types and is flexible enough to handle different types of datasets for pre-training and downstream training. (Hashim et al.)

Bhattacharyya et al. integrate multi-omics and model systems to study cellular mechanisms of cancer to discover therapeutic associations. Their hierarchical Bayesian evidence synthesis framework, BaySyn, uses Gaussian process models and is suitable for rich datasets. The authors applied their framework to multi-omic cancer cell line and patient datasets for pan-gynecological cancers, implicating multiple functional genes across cancers. (Bhattacharyya et al.)

Trinh et al. address the problem of using multi-omics data from a study to investigate questions beyond the scope of that study. To do this, they develop trans-omic knowledge transfer modeling and apply it to the case of using information from an ulcerative colitis cohort in the Integrative Human Microbiome Project (IHMP) to understand biomarkers for anti-TNF therapy resistance in a different ulcerative colitis cohort. They discuss the advantages and disadvantages of three different approaches to knowledge transfer modeling: using a supervised classifier, relative separation, and signature transfer. Through the application of these methods, they provide insights into implementing trans-omic, cross-cohort biomarker discovery. (Trinh et al.)

An important aspect of precision medicine is efficiently identifying disease and disease risk in a patient, and subsequently predicting treatments and therapies that will be effective for that person. The following submissions focus on improving methods for detecting and predicting disease and treatment efficacy.

Extending conventional causal inference methods, Aoki and colleagues propose a framework to use neural networks to estimate multi-treatment effect size. By training a neural network with inputs that include treatments, covariates, as well as outcomes, the deep learning approach summarizes the impact of each treatment with the expectation that the latent space distills meaningful information regarding true treatment effect. Using three synthetic datasets with known true treatment effect, the authors show their approach best approximates treatment effect compared multiple standard benchmark causal inference methods. (Aoki et al.)

Machine learning algorithms optimize certain accuracy metrics by finding best low-dimensional representation of the data. While this approach leads to high predictive power, it can lead to biased conclusions when, for example, training data does not represent the target population. This is problem is of particular importance in biomedical data when patient's health is at stake. De Paolis Kaluza et al. propose a method to identify and quantify bias in a setting where labeled data is known to be drawn from a biased population and unlabeled data is drawn from target population. Under a mild assumption that data comes from a mixture of Gaussian distribution, they developed a multi-sample expectation-maximization algorithm to identify and quantify the bias. (De Paolis Kaluza et al., 2023)

In genetic testing for disease diagnosis or risk, genes with functional significance for the given phenotype are tested to identify what variants a patient possesses. Variant classification following guidelines from the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) (Richards) is used to determine if variants found are potentially pathogenic, benign or a “variant of unknown significance” (VUS). VUS are inconclusive for diagnoses but are commonly assigned due to limited clinical evidence regarding many variants. High throughput assays can be leveraged to molecularly characterize variants lacking sufficient clinical evidence to improve variant classification. The work of Chen and Jain et al. aims to use clinical objectives and *in silico* variant pathogenicity prediction to prioritize genes for high throughput assays. The authors found they could improve on current knowledge-driven and data-

driven strategies for variant classification by using a combined score from three metrics quantifying the importance of genes in satisfying specific clinical objectives. (Chen and Jain et al.)

As researchers find ways to vectorize different data modalities we observe more and more creative applications of machine learning for detecting health conditions. In particular, Aryal et al. developed a set of algorithms for quantifying acoustic-linguistic signals and used them to predict status of Alzheimer's disease. Given the dataset of over 1000 patients, they found that in their setting human-engineered linguistic features were more predictive of the disease than acoustic and learned features. (Aryal et al.)

Zhang et al. focused on improving immunotherapy strategies by developing a pipeline for predicting binding affinity for T cell receptor (TCR) and epitope sequences. Computational binding prediction could help streamline the T cell design process. The authors created PiTE -- Pipeline leveraging Transformer-like Encoders -- that uses large numbers of TCR amino acid sequences to pre-train the model and an advanced sequence encoder. (Zhang et al.)

In the past several years, a spotlight has been shown on racial and ethnic disparities in pregnancy-related conditions. (Carty et al.) In fact, pregnancy-related complications and deaths in general in the U.S. continue to rise. (Heavey) The following two submissions discuss computational applications to gestational diabetes and pre-eclampsia.

Mathur et al. demonstrate reasonable and useful applications of Bayesian network modeling approaches that can incorporate both data-driven learning and domain knowledge in the form of network constraints (independence and monotonicity). The methods are well-summarized and demonstrated in a concrete application for gestational diabetes that illustrate the value of multiple different learning and knowledge modeling techniques beyond purely data-driven models. (Mathur et al.)

For many diseases, transcriptional profiling has been used to identify differentially expressed genes (DEGs). The ignorome is the set of genes that have been experimentally identified as associated with disease but for which no established mechanistic relationship exists. In “Knowledge-Driven Mechanistic Enrichment of the Preecomplasia Ignorome”, Callahan et al. use a biomedical knowledge graph to gain insights into the molecular mechanisms behind pre-eclampsia and to connect experimental findings with previously described disease mechanisms in the literature. Their model provides an approach that could be generalizable to other complex disease processes. (Callahan et al.)

Additional submissions in this session focus on improving data representation of genomic variation and deep learning segmentation modeling in medical imaging.

In practice, the use of genomics terms and vocabulary can be community or context dependent. Annotation and representation of genetic variants and their states (e.g., genotypes, alleles,

haplotypes) vary widely across domains including somatic cancer, Mendelian disease and pharmacogenomics. There are multiple formats for genetic data exchange, some predominantly used in each domain, but each has its limitations especially for application to a different domain. (Pawliczek et al., Holmes et al., den Dunnen et al., Gaedigk et al.) To promote standardized and interoperable representation of genetic variants for precision medicine, the Global Alliance for Genomics and Health (GA4GH) Variation Representation Specification (VRS) developed a Genotype model designed to unambiguously represent the allelic composition of a genetic locus. Here, the Goar et al. describe their Genotype model along with their Haplotype model in the context of several relevant precision medicine settings, including pharmacogenomics. (Goar et al.)

Despite extensive progress in segmentation models in medical imaging, deep learning segmentation models are prone to catastrophic mis-annotation in out-of-domain or foreign examples. Given known clinical priors (such as there is only one prostate or most biological structures are convex), Wooten et al. propose a set of shape features that can identify poor quality segmentation in medical imaging. Features related to area, perimeter, volume, compactness, and convexity are shown to be able to distinguish between acceptable and unacceptable segmentation of the kidney. Using a set of acceptable and unacceptable segmentations of the kidney on CT imaging from radiotherapy treatment plans, the authors show simple heuristics and clustering algorithms can partition between acceptable and unacceptable segmentations, which can be used to quality check deep learning models. (Wooten et al.)

3. Conclusion

Paralleling continued progress in general artificial intelligence, we find there is steady and rapid progress in the application of machine learning to healthcare. Precision medicine is a combination of precision therapeutics - targeting the right treatments to address specific mechanisms of action or response - as well as precision diagnostics - identifying the right patients to the right therapeutics. In this year's session of Precision Medicine: Using Artificial Intelligence to improve diagnostics and healthcare at PSB 2023, we find wide ranging innovations in many modalities and medical datasets. From imaging to genetic data to modeling of clinical treatments, the application of algorithms in the space of healthcare allows a deeper understanding of complex questions.

References

Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J. (2022) Multimodal biomedical AI. *Nat Med.*, 28(9):1773-1784. DOI: 10.1038/s41591-022-01981-2

Aoki, R., Chen, Y., Ester, M. (2023) Multi-treatment Effect Estimation from Biomedical Data. *Pacific Symposium on Biocomputing 2023*.

Aryal, S.K., Prioleau, H., Burge, L. (2023) Acoustic-Linguistic Features for Modeling Neurological Task Score in Alzheimer's. *Pacific Symposium on Biocomputing 2023*.

Bhattacharyya, R., Henderson, N., Baladandayuthapani V. (2023) BaySyn: Bayesian Evidence Synthesis for Multi-system Multiomic Integration. *Pacific Symposium on Biocomputing 2023*.

Callahan, T.J., Stefanski, A.L., Kim J.-D., Baumgartner Jr., W.A., Wyrwa, J.M., Hunter, L.E. (2023) Knowledge-Driven Mechanistic Enrichment of the Preeclampsia Ignorome. *Pacific Symposium on Biocomputing 2023*.

Carty, D.C., Mpofu, J.J., Kress, A.C., Robinson, D., Miller, S.A. (2022) Addressing Racial Disparities in Pregnancy-Related Deaths: An Analysis of Maternal Mortality-Related Federal Legislation, 2017-2021. *J Womens Health*, 31(9):122-1231. DOI: 10.1089/jwh.2022.0336

Chen, Y., Jain, S., Zeiberg, D., Iakoucheva, L., Mooney, S.D., Radivojac, P., Pejaver ,V. (2023) Multi-objective prioritization of genes for high-throughput functional assays towards improved clinical variant classification. *Pacific Symposium on Biocomputing 2023*.

De Paolis Kaluza, M.C., Jain, S., Radivojac. P. (2023) An Approach to Identifying and Quantifying Bias in Biomedical Data. *Pacific Symposium on Biocomputing 2023*.

den Dunnen, J.T., Dalgleish, R., Maglott, D.R., et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat.*, 37(6):564-569. DOI: 10.1002/humu.22981

Gaedigk A., Sangkuhl, K., Whirl-Carrillo, M., et al. The Evolution of PharmVar. *Clin Pharmacol Ther.*, 105(1):29-32. DOI: 10.1002/cpt.1275

Goar, W., Babb, L., Chamala, S., Cline, M., Freimuth, R.R., Hart, R.K., Kuzma, K., Lee, J., Nelson, T., Prlic, A., Riehle, K., Smith, A., Stahl, K., Yates, A.D, Rehm, H., Wagner, A.H. (2023) Development and application of a computable genotype model in the GA4GH Variation Representation Specification. *Pacific Symposium on Biocomputing 2023*.

Hashim, S., Nandakumar, K., Yaqub, M. (2023) Self-omics: A Self-supervised Learning Framework for Multi-omics Cancer Data. *Pacific Symposium on Biocomputing 2023*.

Heavey, E. (2022) Rising US pregnancy-related deaths. *Nursing*, 52(8):36-39. DOI: 10.1097/01.NURSE.0000839800.71201.d8

Holmes, J.B., Moyer, E., Phan, L., et al. SPDI: data model for variants and applications at NCBI. Bioinformatics, 36(6):1902-1907. DOI: 10.1093/bioinformatics/btz856

Mathur, S., Karanam, A., Radivojac, P., Haas, D.M., Kersting, K., Natarajan, S. (2023) Exploiting Domain Knowledge as Causal Independencies in Modeling Gestational Diabetes. *Pacific Symposium on Biocomputing 2023*.

Pawliczek, P., Patel, R.Y., Ashmore, L.R. et al. (2018) ClinGen Allele Registry links information about genetic variants. *Hum Mutat*., 39(11):1690-1701. DOI: 10.1002/humu.23637

Richards, S., Aziz, N., Bale, S., et al. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.*, 17(5):405-424. DOI: 10.1038/gim.2015.30

Trinh, A., Ran, R., Brubaker, D.K. (2023) Trans-Omic Knowledge Transfer Modeling Infers Gut Microbiome Biomarkers of Anti-TNF Resistance in Ulcerative Colitis. *Pacific Symposium on Biocomputing 2023*.

Wooten, Z., Yu, C., Court, L., Peterson, C. (2023) Predictive modeling using shape statistics for interpretable and robust quality assurance of automated contours in radiation treatment planning. *Pacific Symposium on Biocomputing 2023*.

Zhang, P., Bang, S., Lee, H. (2023) PiTE: TCR-epitope binding affinity prediction pipeline using Transformer-based Sequence Encoder. *Pacific Symposium on Biocomputing 2023*.