Article published online: 2022-12-04

Content Summary of Best Papers for the Bioinformatics and Tranlational Informatics Section of the 2022 IMIA Yearbook

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

In this paper, Moses *et al.*, describe how a combination of cortical activity recording and computational modeling can enable communication for people who have lost the ability to speak. The authors collected dozens of hours of cortical activity recordings in a patient with anarthria as the patient was prompted to say 50 different words. They used these recordings to train a deep-learning model to decode the spoken word directly from the cortical activity profiles. In addition, the authors applied a secondary natural language processing model to further refine the prediction output. Ultimately, this approach was able to decode speech in real time with an overall 47.1% accuracy at a speed of approximately 15 words per minute. Although there is considerable work yet to be done in terms of per-word accuracy, the model was able to detect the vast majority of attempts by the patient to say a word (98%) despite the lack of speech output. The model's current performance suggests that future iterations of this approach could be transformational for patients with anarthria.

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

In this paper, the authors use a novel three-stage approach to identify new lipid-associated genes and linked phenotypes in the electronic health record (EHR). This is significant because plasma lipids are a heritable risk factor for heart disease, a highly impactful disease. Improved informatics approaches to better connect data across biological scales and integrate those data with clinical phenotypes, will advance our understanding of the molecular and genetic architecture of lipids and associated diseases. The first stage involved paired genome- and transcriptome-wide association studies (GWAS and TWAS, respectively) to identify genes associated with plasma lipid traits. The second stage used these results as input to a phenome-wide association study (PheWAS) to identify a set of phenotype associations for the genes in two large clinical biobanks. Finally, in the third stage, the authors use Mendelian randomization to identify causal associations between genetic variants associated with both lipids and diseases captured in the EHR. There were 67 new lipid-associated genes identified in this work, in addition to new evidence of pleiotropy between lipids and a number of complex diseases. Future work could focus on lipids as a potentially modifiable risk factor for diseases other than heart disease, and this framework is amenable to similar studies on other molecular traits.

Wei Q, Ramsey SA

Predicting chemotherapy response using a variational autoencoder approach

BMC Bioinformatics 2021 Sep 22;22(1):453

The gene expression profiles of tumors encode useful information about the patient's response to chemotherapy for cancer treatment. However, due to the large feature space, it is difficult to choose the most important features for prediction, or develop a well-performing model using the full feature set. In this paper, Wei and Ramsey develop a variational autoencoder (VAE) to identify latent features from the transcriptome data of five different cancers (colon, pancreatic, bladder, breast, sarcoma). They use the resulting features as input into a second, supervised machine learning model in order to predict the patient's response to chemotherapy directly from the gene expression profiles of the tumor (n = 2,606). They show that the VAE encoding maintains important biological information about the tumors, including information about the cancer type, and also leads to better-performing prediction models than current state-of-the-art feature selection approaches, such as PCA.