Machine learning methods for rare diseases

This manuscript (permalink) was automatically generated from jaybee84/ml-in-rd@f8c2023 on August 25, 2020.

Authors

- **Jineta Banerjee**
  0000-0002-1775-3645 · jaybee84
  Sage Bionetworks · Funded by Neurofibromatosis Therapeutic Acceleration Program; Children's Tumor Foundation

- **Robert J Allaway**
  0000-0003-3573-3565 · allaway · allawayr
  Sage Bionetworks · Funded by Neurofibromatosis Therapeutic Acceleration Program; Children's Tumor Foundation

- **Jaclyn N Taroni**
  0000-0003-4734-4508 · jaclyn-taroni
  Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

- **Casey Greene**
  0000-0001-8713-9213 · cgreene
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania; Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

- **Justin Guinney**
  0000-0003-1477-1888 · jguinney
  Sage Bionetworks · Funded by Neurofibromatosis Therapeutic Acceleration Program; Children's Tumor Foundation

Synopsis

(Instructions: Describe the background, basic structure of the article, list material to be covered indicating depth of coverage, how they are logically arranged, include recent pubs in the area, 300-500 words)

Substantial technological advances have dramatically changed biomedicine by making deep characterization of patient samples routine. These technologies provide a rich portrait of genes, cellular pathways, and cell types involved in complex phenotypes. Machine learning is often a perfect fit for the types of data now being generated, and Nature Methods routinely has reports of machine learning methods that extract disease-relevant patterns from these high dimensional datasets. Often, these methods require a large number of samples to identify reproducible and biologically meaningful patterns. With rare diseases, biological specimens and consequently data, are limited due to the rarity of the condition. In this perspective, we outline the challenges and emerging solutions for using machine learning in these settings. We aim to spur the development of powerful machine learning techniques for rare diseases. We also note that precision medicine presents a similar challenge, in which a common disease is partitioned into small subsets of patients with shared etiologies and treatment strategies. Advances from rare disease research are likely to be highly informative for other applications as well.

Introduction

Machine learning is gaining momentum in biomedical data analysis as data collection becomes increasingly high-throughput and as novel computational methods for exploring those data are developed. Application of machine learning to any dataset requires careful execution, but the application to biomedical data and subsequent interpretation requires depth of knowledge not only in the biomedical domain but also a clear understanding of the methods and their underlying assumptions. Application of machine learning to any kind of data consists of the following major steps: (1) data evaluation and question formulation, (2) selection of normalization/dimension reduction to mitigate technical differences, (3) selection of appropriate algorithms which select features to answer the formulated question, (4) evaluation of the answers generated by the algorithm. Each of these steps require the practitioner to choose from a variety of methodologies to apply. The selection of the methodologies at each of these steps need to be based upon robust reasoning to ensure stability of the results.

Rare disease research has additional constraints to consider when using machine learning methods, including lack of statistical power in dataset size, heterogeneity in available data, and sensitivity of machine learning methods to misinterpretation in view of small datasets. Moreover, in the context of rare disease, special considerations need to be made to safeguard against misinterpretation of results. Such considerations include incorporation of techniques that build upon prior domain-specific knowledge, methods that are resilient to challenges posed by small datasets, and methods that can mitigate technical disparities in the data. Recent advances in methodologies to accommodate rarity of samples and increased transparency in model outputs have encouraged application of machine learning in rare diseases. In this perspective, we discuss techniques for understanding the nature of rare disease data, methods for addressing some of the limitations of these data, and machine learning methods that can tolerate some of these limitations.

Managing disparities in data generation is required for robust rare disease analyses

Rare disease data from can suffer from artifacts introduced by batch, assay platform, specimen quality or other non-biological phenomena. The consequences of these artifacts are amplified in rare diseases which often have few samples and heterogeneous phenotypes. Furthermore, datasets are often pieced together from multiple small studies where biological characteristics are confounded by technical variables. Collaboration with data generators or domain experts may result in unexpected insight into potential sources of variation. The authors experienced this when studying neurofibromatosis type 1 (NF1). The NF1 datasets were comprised of samples obtained with different surgical techniques, resulting in biological differences that were a consequence of sample collection, rather than … Consequently, careful assessment of and accounting for confounding factors is critical to identifying biologically meaningful features within a dataset.

Assessment of confounding factors and heterogeneity is perhaps most easily performed using unsupervised learning approaches. K-means clustering or hierarchical clustering can be used to characterize the structure present in genomic or imaging data. [1,2]. Similarly, dimensionality reduction methods can be used to visualize heterogeneity and confounders, including multidimensional scaling, principal components analysis, t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP), among others. [3,4,5,6] All of these methods can be used to identify batch effects and other structure in the data, though some (like t-SNE and UMAP) require parameters that can affect the output and it's interpretation [6,7]. Therefore, obtaining a clear interpretation from these methods requires understanding the underlying approach and parameters.

Another important consideration is discussed by Way, et. al. [8]: a single dimensionality reduction method alone may not be sufficient to reveal all of the technical or biological heterogeneity; testing multiple methods may result in a more comprehensive portrait of the data. Dimensionality reduction techniques are not restricted to 'omic' data - they can also be used in rare disease applications to characterize the structure and heterogeneity of imaging data [9], mass cytometry data [10], and others.

Once the nature of the non-biological heterogeneity has been established, different techniques can be used to correct the differences. Common approaches include reprocessing the raw data using a single analysis pipeline if the data are obtained from different sources, application of batch correction methods [11,12], and normalization of raw values[13]. It is also important to be realistic when working with rare disease data. For various reasons including ethics, funding, and limited biospecimens, experimental design and the resulting data will often be less-than-ideal. In these cases, it may be prudent to take a step back, re-evaluate the data, and identify methods that can operate within the constraints of the data.

Techniques and procedures must be implemented to manage model complexity without sacrificing the value of machine learning

Inherent challenges posed by low sample numbers in rare diseases are further aggravated by disease heterogeneity, poorly defined disease phenotypes, and often a lack of control (i.e. normal) data. Machine learning approaches must be carefully designed to address these challenges. We discuss how to implement methodological solutions like bootstrapping sample data, regularization methods for deep learning, and hyper-ensemble techniques to minimize misinterpretation of the data.

Bootstrapping

Bootstrap (resampling) computation is a powerful statistical technique that can be used to estimate population values from datasets of limited sample size by resampling the data to generate an estimated distribution of the population statistic and minimize estimation error [14]. Bootstrap based techniques are used in conjunction with various learning methods to find the most informative models given a specific dataset (e.g. bootstrap aggregating or bagging used in random forests [15,16], bootstrap in neural networks [17], or regression models [18,19]). In addition to model selection, bootstrap can be used to enhance information content of rare disease datasets and generate confidence intervals for the model predictions [20]. In this study, bootstrapping the training sample without replacement simulated generation of separate datasets that helped expose the learning models (in this case random forests) to the incomplete nature of the data. Such bootstrapping of the training data in addition to that included in the model (bagging) helped generate a distribution (and confidence intervals) of the importance scores of the predictive features selected by the model.

Regularization

A common strategy for handling the paucity of data in rare disease is to aggregate data from multiple studies or time points to produce a more comprehensive dataset. Given a dataset with strong preexisting study-specific technical differences between groups of samples, machine learning methods may model dataset-specific features instead of true biology, leading to high prediction accuracy for training data but poor performance in new test data (an "overfit" model) [21]. Minimization of overfitting can be accomplished by cross-validation (to reduce variance in predictions) and regularization (to reduce low bias in models) methodologies. Regularization makes models less reliant on training data by adding a small penalty (determined by cross-validation), and can not only minimize overfitting but can additionally help in predicting outcomes using a limited number of samples.

ML models can be regularized using 3 main methods, each with their particular strengths and weaknesses. Ridge regression aims to minimize the magnitude of the features, but cannot completely remove unimportant features and thus may not be ideal for reducing the feature space. Another method, LASSO or Least Absolute Shrinkage and Selection Operator regression, works well for selecting few important features since it can minimize the magnitude of some features more than the others[22]. Elastic-net regression is a combination of LASSO and ridge regression[23], and helps to select most useful features, especially in presence of large number of correlated features.

While regression based regularization has not been used extensively in rare disease, examples of combinations of above strategies implemented in rare variant discovery and immune cell signature discovery can provide an insight into their possible use in rare disease. In rare variant discovery, adaptive ridge regression was utilized to combine rare variants into a single score to increase the signal of rare variants[24], while LASSO was implemented along with group penalties to identify rare variants/ low frequency predictors [25,26]. Hybrid approaches of LASSO including boosting the signal of rare variants by capturing linear combinations of variants by gene or chromosome location in 5% of subjects [27,28,28,29,30], integration with the probabilistic logistic bayesian approach [31], and combining feature selection methods with a generalized pooling strategy [32] have also been tested in rare variant discovery. Another interesting approach which incorporated prior knowledge into the regularization (called sparse-group LASSO) worked well to select the driver genes in a pathway of interest where only few genes in a pathway were true predictors of a phenotype [33].

In immune cell signature discovery, elastic-net regression (a combination of LASSO and ridge regression) has been used to reduce the feature space and was found to outperform the other regression approaches [23,34,35]. A variation of elastic-net, where a two-step regularized logistic regression was used to pre-select an optimal number of genes before implementing elastic-net regularization for gene selection, identified immune cell signatures in an RNA-seq dataset where the number of cells sampled were far fewer than number of genes profiled [**???** 10.1186/s12859-019-2994-z].
Thus robust regularizations methods like LASSO or elastic-net have been methods of choice where the the profiled feature space have outnumbered the number of samples or patients by a magnitude and should be explored while working with rare disease datasets.

Hyperensemble

Techniques that build on prior knowledge and indirectly related data are necessary for many rare disease applications

Knowledge graphs

An intrinsic constraint in studying rare diseases is the lack of large, normalized datasets, which limits our ability to study key attributes of rare diseases. A potentially powerful strategy for evaluating genotype-phenotype relationships or repurposing drugs when large datasets are scarce is to use knowledge graphs. Knowledge graphs integrate related-but-different data types, creating a rich data source. Examples of public biomedical knowledge graphs and frameworks that could be useful in rare disease include the Monarch Graph Database[36], hetionet[37], PheKnowLator[38], and the Global Network of Biomedical Relationships[39]. These graphs connect information like genetic, functional, chemical, clinical, and ontological data to enable the exploration of relationships of data with disease phenotypes through manual review[40] or computational methods[41,42].

In the academic rare disease space, there a few pioneering examples of ML-based mining of knowledge graphs to repurpose drugs[41] and classify rare diseases[42]. These studies make it clear that there are some challenges in using machine learning using graph databases in rare disease. For example, these papers rely on a gold standard dataset to validate the performance of the models; often, there are not robust gold standard datasets available for individual rare diseases. They also evaluate rare diseases in an unbiased manner, rather than interrogating a specific disease of interest. Consequently, it is not yet clear how effective these approaches, and knowledge graphs in general, are in studying a specific disease of interest; more work needs to be done to identify methods that can provide actionable insights for a specific rare disease application.

Beyond the aforementioned studies, there are few examples of studies in the public domain that leverage knowledge graphs to characterize rare disease. Private entities (e.g. healx, Boehringer Ingelheim, DrugBankPlus) are performing an undisclosed amount of work to create proprietary rare disease knowledge graphs for ML-based drug discovery applications.
The existence of private companies pursuing this idea, as well as the availability of several public knowledge graphs with relevance to rare disease, suggests to us that this is a likely fruitful and untapped area of rare disease research in the public sphere. More work needs to be done to assess 1) which graphs and graph features capture the salient information about rare diseases, 2) the utility of ML methods to obtain actionable insights about rare diseases and 3) which problems - like drug discovery, identification of novel rare diseases, or assessment of genotype-phenotype relationships - can be interrogated using ML of knowledge graphs.

Wisdom of the crowd: rare disease applications of ensemble methods

Implementing machine learning on data with low sample size and high label uncertainty can lead to unstable predictions. In such cases various machine learning methods together (also called *ensemble learning*) can help increase accuracy and stability of the predictions. Ensemble learning can use multiple similar approaches stitched together to reach a consensus, or can be a collection of different approaches that perform better compared to any single algorithm. For example, ensemble learning methods like random forests use bootstrap aggregation (or *bagging*) of independent decision trees that use similar parameters but different paths to form a consensus about the important predictive features hidden in the dataset [43]. However, successful application of consensus based ensemble learning requires “gold standard” data where the diagnosis or label of a data point in the training dataset has very little uncertainty (or “label-noise”) associated with it [47]. In most cases of rare disease, due to the inherent nature of being less defined, the symptoms as well as any underlying biology comes with a reasonable amount label-noise leading to a *silver standard* dataset[48]. In such

datasets, the limited success of the *bagging* approach has led to the use of ensemble learning or *cascade learning*, where multiple methods leveraging distinct underlying assumptions are used in tandem and augmented with algorithms like AdaBoost (*boosting*) to capture stable patterns existing in the silver standard data and reduce uncertainty [49]. A variation of cascade learning implemented to identify rare disease patients from electronic health records (EHR) from the general population utilized independent steps for feature extraction (using word2vec [52]), preliminary prediction (ensemble of decision trees with penalization for excessive tree-depth), and prediction refinement (using similarity of data points to resolve sample labels) [53]. This cascade learner benefited from the independence of the feature extraction step and the prediction refinement step from the preliminary classification of the labeled dataset to find stable patterns and perform better than other ensemble methods when implemented on this silver standard dataset.

In datasets with multiple classes, most cascade classifiers follow a *one-classifier-at-a-time* approach where algorithms at each level predict all classes involved. But instances where the need for high prediction accuracy for one class outweighs other classes, further modification of the cascade learning efforts is required. An example of such modification was implemented for triaging psychiatric patients where the identification of one class of psychiatric patients ("severe") far outweighed the need for optimized overall classification accuracy [54]. Due to the requirements of the problem, a *one-class-at-a-time* cascade learning approach was adopted, where at each stage a binary classifier was used to predict a specific class against all others. The final model implemented all models together each identifying one class sequentially and the union of the predictions of all the different models as the final prediction. The cascade classifiers using the one-class-at-a-time approach were found to perform better than multi-class ensemble classifiers in most cases.

Thus ensemble learning can be helpful in producing stable predictions from rare disease data, but the choice of using bagging, boosting, independent algorithmic steps, or one-class-at-a-time approach depends on the nature of the prediction question.

Representation learning

Representation learning, also called feature learning, is the process of learning features from raw data, where a feature is an individual variable. An algorithm or approach will construct features as part of training and, in a supervised application, use those features to predict labels on input data. Using an example from transcriptomics, an unsupervised method such as matrix factorization can be used to extract a low-dimensional representation of the gene-level data, learning features that are a combination of input genes' expression levels [8,55]. Low-dimensional representations trained on a collection of transcriptomic data can then be used as input to supervised machine learning methods [56]. Supervised neural networks used in medical imaging studies [57] (reviewed in [58]), which are trained to predict labels or classes, are also an example of representation learning. Learned features in the medical imaging domain may be a series of edges representing a blood vessel formation that discriminates between disease states. Features learned from transcriptomic data could be coordinated sets of genes involved in a biological process that are descriptive in some way [59].

In the rare disease domain, Dincer et al. leveraged publicly available acute myeloid leukemia (AML) gene expression data to improve the prediction of *in vitro* drug responses [60]. The authors trained a variational autoencoder (VAE) on AML data that had been collected over time without the phenotypic information they were interested in (drug response). (A VAE is an unsupervised neural network that learns a series of representations from data.) The authors used the learned attributes to encode a low-dimensional representation of held-out AML data with phenotype labels of interest, and used this low-dimensional representation as input to a classifier that predicted *in vitro* drug response.

Representation learning tends to be data-intensive; many samples are required. Though there were over 6500 AML samples from many different studies used as part of the training set in Dincer et

al. [60], we expect that in other rare diseases considerably fewer samples will be available or may be from different tissues in systemic diseases. The study by Dincer and colleagues highlights another challenge: samples collected as part of multiple studies may not be associated with the deep phenotypic information that would maximize their scientific value. In the next section, we will introduce methods or approaches that may be more broadly useful in rare diseases; representation learning underlies many of them.

Transfer, multitask, and few-shot learning

We focus on a series of approaches that are centered on the following concept: to realize the potential of machine learning for biological discovery in rare diseases, we often cannot study an individual rare disease alone as samples are limited. Instead, we can build on prior knowledge and large volumes of data that do not directly assay our disease of interest, but are similar enough to be valuable for discovery. We can leverage shared features, whether they are normal developmental processes that are aberrant in disease or an imaging anomaly present in rare and common diseases, for advancing our understanding. Methods that leverage shared features include transfer learning, multitask learning, and few-shot learning approaches.

Transfer learning

Transfer learning is an approach where a model trained for one task or domain (source domain) is applied to another, typically related task or domain (target domain). Transfer learning can be supervised (one or both of the source and target domains have labels), or unsupervised (both domains are unlabeled). Though there are multiple types of transfer learning, we will focus on feature-representation-transfer [61] here. Feature-representation-transfer approaches learn representations from the source domain and apply them to a target domain [61]. This concept is embodied in Dincer et al., where features are learned from unlabeled AML data and then used to encode a low-dimensional representation of AML data with *in vitro* drug response labels [60]. The authors then used this low-dimensional representation as input to predict drug response labels–a supervised example.

In an unsupervised case, Taroni et al. trained a Pathway-Level Information ExtractoR (PLIER) [62] on a large generic collection of human transcriptomic data (recount2 [63]) and used the latent variables learned by the model to describe transcriptomic data from the unseen rare diseases antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis (AAV) and medulloblastoma in an approach termed MultiPLIER [64]. (Here “unseen” refers to the fact that these diseases were not in the training set.) PLIER is a matrix factorization approach that takes prior knowledge in the form of gene sets or pathways and gene expression data as input; some latent variables learned by the model will align with input gene sets [62]. We demonstrated that training on larger collections of randomly selected samples produced models that captured a larger proportion of input gene sets and better distinguished closely related signals, which suggests that larger training sets produced models that are more suitable for biological discovery [64].

Though models trained on generic compendia had appealing properties, that alone does not guarantee suitability for describing rare diseases. We must examine the relevance of learned features to the disease under study. In Taroni et al., we found that the expression of latent variables that could be matched between the MultiPLIER model and a dataset-specific model were well-correlated, particularly when latent variables were associated with input gene sets [64]. Despite the absence of AAV from the training set, MultiPLIER was able to learn a latent variable where the genes with the highest contributions encode antigens that the antineutrophil cytoplasmic antibodies (ANCA) form against in AAV and with higher expression in more severe disease [65]. The utility of this approach stems from the fact that biological processes are often *shared* between conditions–the same ANCA antigen genes are components of neutrophilic granule development that is likely captured or assayed

in the collection of transcriptomic data used for training. MultiPLIER has additional attributes that make it practical for studying rare diseases: latent variables that are not associated with input gene sets may capture technical noise separately from biological signal and we can use one model to describe multiple datasets instead of reconciling output from multiple models (see *05.heterogeneity.md*).

Taken together, DeepProfile [60] and MultiPLIER [64] suggest transfer learning can be beneficial for studying rare diseases. In the natural images field, researchers have demonstrated that the transferability of features depends on relatedness of tasks [66]. The limits of transfer learning for and the concept of relatedness in high-dimensional biomedical data assaying rare diseases are open research questions. In the authors' opinion, selecting an appropriate model for a given task and evaluations that are well-aligned with a research goal are crucial for applying these approaches in rare diseases.

Multitask and few-shot learning

Where transfer learning can be supervised or unsupervised, the related approaches multitask and few-shot learning are forms of supervised learning that generally rely on deep neural networks. Multitask learning is an approach where classifiers are learned for *related tasks* at the same time using a shared representation [67], where task refers to an individual prediction being made. Few-shot learning is the generalization of a model trained on related tasks to a new task with limited labeled data (e.g., the detection of a patient with a rare disease from a low number of examples of that rare disease).

Multitask neural networks that predict multiple tasks simultaneously are generally thought to improve performance over models that make predictions for a single task by learning a shared representation and effectively being exposed to more training data than the single task case [**???**,67]. Kearnes, Goldman, and Pande set out to examine the effects of dataset size and task relatedness on multitask learning performance improvements ("multitask effect") in drug discovery–an area that also suffers from insufficient data [**???**]. The authors found that the multitask performance gains were highly dataset-specific: smaller datasets tended to benefit most from multitask learning and the addition of more training data did not guarantee improved performance for multitask models. In predicting phenotypes from EHR data, Ding et al. demonstrated that multitask neural networks outperformed single-task networks for predicting complex rare phenotypes but not common phenotypes [**???**]. Liu et al. developed a method to train long short-term memory networks, a type of recurrent neural network, to predict mortality in rare diseases using EHR data as input [68]. Their method, Ada-SiT (Adaptation to Similar Tasks), was specifically designed for many tasks with insufficient data and allowed for task similarity to be measured during training.

In contrast, one-shot or few-shot learning relies on using prior knowledge to generalize to new prediction tasks where there are a low number of examples [69], where a distance metric is learned from input data and used to compare new examples for prediction [70]. Altae-Tran et al. developed a method for predicting small molecule activity that learned a meaningful distance metric over the properties of various compounds [70]. However, the authors' results suggested underperformance of one-shot learning methods relative to baseline random forest models when structural similarity could not be exploited and did not show support for generalization of models trained on very different contexts from the target task. Quellec et al. presented a few-shot learning approach for detecting rare pathologies in fundus photographs [71]. The authors trained a convolutional neural network (CNN) to predict common pathologies, which tended to cluster similar conditions in feature space. The learned feature space was then used to train a probabilistic model for each rare pathology. This approach outperformed multitask learning, which suggests few-shot learning provides an advantage in contexts where predicting common conditions simultaneously results in a loss of performance [71].

Multitask and few-shot learning are comprised of a variety of approaches and architectures that are beyond this scope of this work (see [**???**,72] and [69] for an overview). As with transfer learning, the utility of these approaches to rare disease research is an open question and is likely to be highly dependent on dataset availability and research goals.

Conclusions

> We will conclude by discussing the potential of the above-mentioned approaches in rare diseases and other biomedical areas where data is scarce.

Outlook

References

1. **Clustering cancer gene expression data: a comparative study**
   Marcilio CP de Souto, Ivan G Costa, Daniel SA de Araujo, Teresa B Ludermir, Alexander Schliep
   BMC Bioinformatics (2008-11-27) https://doi.org/dqqbn6
   DOI: 10.1186/1471-2105-9-497 · PMID: 19038021 · PMCID: PMC2632677

2. **Removing Batch Effects From Histopathological Images for Enhanced Cancer Diagnosis**
   Sonal Kothari, John H. Phan, Todd H. Stokes, Adeboye O. Osunkoya, Andrew N. Young, May D. Wang
   IEEE Journal of Biomedical and Health Informatics (2014-05) https://doi.org/gdm9jd
   DOI: 10.1109/jbhi.2013.2276766 · PMID: 24808220 · PMCID: PMC5003052

3. **Multidimensional Scaling**
   Michael A. A. Cox, Trevor F. Cox
   Springer Berlin Heidelberg (2008) https://doi.org/dg9m4f
   DOI: 10.1007/978-3-540-33037-0_14

4. **Principal component analysis: a review and recent developments**
   Ian T. Jolliffe, Jorge Cadima
   Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (2016-04-13) https://doi.org/gcsfk7
   DOI: 10.1098/rsta.2015.0202 · PMID: 26953178 · PMCID: PMC4792409

5. (2020-06-01) https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

6. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**
   Leland McInnes, John Healy, James Melville
   arXiv (2018-12-07) https://arxiv.org/abs/1802.03426

7. **How to Use t-SNE Effectively**
   Martin Wattenberg, Fernanda Viégas, Ian Johnson
   Distill (2016-10-13) https://doi.org/gffk7g
   DOI: 10.23915/distill.00002

8. **Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations**
   Gregory P. Way, Michael Zietz, Vincent Rubinetti, Daniel S. Himmelstein, Casey S. Greene
   Genome Biology (2020-05-11) https://doi.org/gg2mjh
   DOI: 10.1186/s13059-020-02021-3 · PMID: 32393369 · PMCID: PMC7212571

9. **Automatic detection of rare pathologies in fundus photographs using few-shot learning**
   Gwenolé Quellec, Mathieu Lamard, Pierre-Henri Conze, Pascale Massin, Béatrice Cochener
   Medical Image Analysis (2020-04) https://doi.org/ggsrc7
   DOI: 10.1016/j.media.2020.101660 · PMID: 32028213

10. **Sensitive detection of rare disease-associated cell subsets via representation learning**
    Eirini Arvaniti, Manfred Claassen
    Nature Communications (2017-04-06) https://doi.org/gf9t7w
    DOI: 10.1038/ncomms14825 · PMID: 28382969 · PMCID: PMC5384229

11. **Adjusting batch effects in microarray expression data using empirical Bayes methods**
W. Evan Johnson, Cheng Li, Ariel Rabinovic
Biostatistics (2007-01) https://doi.org/dsf386
DOI: 10.1093/biostatistics/kxj037 · PMID: 16632515

12. **svaseq: removing batch effects and other unwanted noise from sequencing data**
Jeffrey T. Leek
Nucleic Acids Research (2014-12-01) https://doi.org/f8k8kf
DOI: 10.1093/nar/gku864 · PMID: 25294822 · PMCID: PMC4245966

13. **A scaling normalization method for differential expression analysis of RNA-seq data**
Mark D Robinson, Alicia Oshlack
Genome Biology (2010) https://doi.org/cq6f8b
DOI: 10.1186/gb-2010-11-3-r25 · PMID: 20196867 · PMCID: PMC2864565

14. **Improvements on Cross-Validation: The 632+ Bootstrap Method**
Bradley Efron, Robert Tibshirani
Journal of the American Statistical Association (1997-06) https://doi.org/gfts5c
DOI: 10.1080/01621459.1997.10474007

15.**:{unav)**
Leo Breiman
Machine Learning (2001) https://doi.org/d8zjwq
DOI: 10.1023/a:1010933404324

16. **Bootstrap Methods for Developing Predictive Models**
Peter C Austin, Jack V Tu
The American Statistician (2004-05) https://doi.org/bzjjxt
DOI: 10.1198/0003130043277

17. **Bootstrap for neural model selection**
Riadh Kallel, Marie Cottrell, Vincent Vigneron
Neurocomputing (2002-10) https://doi.org/c8xpqz
DOI: 10.1016/s0925-2312(01)00650-6

18. **Fast bootstrap methodology for regression model selection**
A. Lendasse, G. Simon, V. Wertz, M. Verleysen
Neurocomputing (2005-03) https://doi.org/dx5c3p
DOI: 10.1016/j.neucom.2004.11.017

19. **A bootstrap resampling procedure for model building: Application to the cox regression model**
Willi Sauerbrei, Martin Schumacher
Statistics in Medicine (1992) https://doi.org/cnpg3d
DOI: 10.1002/sim.4780111607 · PMID: 1293671

20. **Integrative Analysis Identifies Candidate Tumor Microenvironment and Intracellular Signaling Pathways that Define Tumor Heterogeneity in NF1**
Jineta Banerjee, Robert J Allaway, Jaclyn N Taroni, Aaron Baker, Xiaochun Zhang, Chang In Moon, Christine A Pratilas, Jaishri O Blakeley, Justin Guinney, Angela Hirbe, … Sara JC Gosline
Genes (2020-02-21) https://doi.org/gg4rbj
DOI: 10.3390/genes11020226 · PMID: 32098059 · PMCID: PMC7073563

21. **Definitions, methods, and applications in interpretable machine learning**
W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, Bin Yu
Proceedings of the National Academy of Sciences (2019-10-29) https://doi.org/ggbhmq
DOI: 10.1073/pnas.1900654116 · PMID: 31619572 · PMCID: PMC6825274

22. **Regularization**
Jake Lever, Martin Krzywinski, Naomi Altman
Nature Methods (2016-09-29) https://doi.org/gf3zrr
DOI: 10.1038/nmeth.4014

23. **Regularization and variable selection via the elastic net**
Hui Zou, Trevor Hastie
Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2005-04)
https://doi.org/b8cwwr
DOI: 10.1111/j.1467-9868.2005.00503.x

24. **Adaptive Ridge Regression for Rare Variant Detection**
Haimao Zhan, Shizhong Xu
PLoS ONE (2012-08-28) https://doi.org/f36tm5
DOI: 10.1371/journal.pone.0044173 · PMID: 22952918 · PMCID: PMC3429469

25. **Statistical analysis strategies for association studies involving rare variants**
Vikas Bansal, Ondrej Libiger, Ali Torkamani, Nicholas J. Schork
Nature Reviews Genetics (2010-10-13) https://doi.org/dn4jtz
DOI: 10.1038/nrg2867 · PMID: 20940738 · PMCID: PMC3743540

26. **Association screening of common and rare genetic variants by penalized regression**
H. Zhou, M. E. Sehl, J. S. Sinsheimer, K. Lange
Bioinformatics (2010-08-06) https://doi.org/c7ndkx
DOI: 10.1093/bioinformatics/btq448 · PMID: 20693321 · PMCID: PMC3025646

27. **Identification of Grouped Rare and Common Variants via Penalized Logistic Regression**
Kristin L. Ayers, Heather J. Cordell
Genetic Epidemiology (2013-09) https://doi.org/f5cw72
DOI: 10.1002/gepi.21746 · PMID: 23836590 · PMCID: PMC3842118

28. **A LASSO-based approach to analyzing rare variants in genetic association studies**
Jennifer S Brennan, Yunxiao He, Rose Calixte, Epiphanie Nyirabahizi, Yuan Jiang, Heping Zhang
BMC Proceedings (2011-11-29) https://doi.org/bjcndj
DOI: 10.1186/1753-6561-5-s9-s100 · PMID: 22373373 · PMCID: PMC3287823

29. **Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data**
Bingshan Li, Suzanne M. Leal
The American Journal of Human Genetics (2008-09) https://doi.org/d4jpcb
DOI: 10.1016/j.ajhg.2008.06.024 · PMID: 18691683 · PMCID: PMC2842185

30. **Comparison of statistical approaches to rare variant analysis for quantitative traits**
Han Chen, Audrey E Hendricks, Yansong Cheng, Adrienne L Cupples, Josée Dupuis, Ching-Ti Liu
BMC Proceedings (2011-11-29) https://doi.org/b9mf4x
DOI: 10.1186/1753-6561-5-s9-s113 · PMID: 22373209 · PMCID: PMC3287837

31. **An Improved Version of Logistic Bayesian LASSO for Detecting Rare Haplotype-Environment Interactions with Application to Lung Cancer**
Yuan Zhang, Swati Biswas
Cancer Informatics (2015-02-09) https://doi.org/ggxxfp
DOI: 10.4137/cin.s17290 · PMID: 25733797 · PMCID: PMC4332044

32. **Multiple Regression Methods Show Great Potential for Rare Variant Association Tests**
ChangJiang Xu, Martin Ladouceur, Zari Dastani, J. Brent Richards, Antonio Ciampi, Celia M. T. Greenwood
PLoS ONE (2012-08-08) https://doi.org/f35726
DOI: 10.1371/journal.pone.0041694 · PMID: 22916111 · PMCID: PMC3420665

33. **A Sparse-Group Lasso**
Noah Simon, Jerome Friedman, Trevor Hastie, Robert Tibshirani
Journal of Computational and Graphical Statistics (2013-04) https://doi.org/gcvjw8
DOI: 10.1080/10618600.2012.681250

34. **Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification**
Zakariya Yahya Algamal, Muhammad Hisyam Lee
Computers in Biology and Medicine (2015-12) https://doi.org/f73xvj
DOI: 10.1016/j.compbiomed.2015.10.008 · PMID: 26520484

35. **Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification**
Yong Liang, Cheng Liu, Xin-Ze Luan, Kwong-Sak Leung, Tak-Ming Chan, Zong-Ben Xu, Hai Zhang
BMC Bioinformatics (2013-06-19) https://doi.org/gb8v2x
DOI: 10.1186/1471-2105-14-198 · PMID: 23777239 · PMCID: PMC3718705

36. **The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species**
Christopher J. Mungall, Julie A. McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, Tom Conlin, Nathan Dunn, Mark Engelstad, … Melissa A. Haendel
Nucleic Acids Research (2017-01-04) https://doi.org/f9v7bz
DOI: 10.1093/nar/gkw1128 · PMID: 27899636 · PMCID: PMC5210586

37. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini
eLife (2017-09-22) https://doi.org/cdfk
DOI: 10.7554/elife.26726 · PMID: 28936969 · PMCID: PMC5640425

38. **A Framework for Automated Construction of Heterogeneous Large-Scale Biomedical Knowledge Graphs**
Tiffany J. Callahan, Ignacio J. Tripodi, Lawrence E. Hunter, William A. Baumgartner
bioRxiv (2020-05-02) https://doi.org/gg338z
DOI: 10.1101/2020.04.30.071407

39. **A global network of biomedical relationships derived from text**
Bethany Percha, Russ B Altman
Bioinformatics (2018-08-01) https://doi.org/gc3ndk
DOI: 10.1093/bioinformatics/bty114 · PMID: 29490008 · PMCID: PMC6061699

40. **Structured reviews for data and knowledge-driven research**
Núria Queralt-Rosinach, Gregory S Stupp, Tong Shu Li, Michael Mayers, Maureen E Hoatlin, Matthew Might, Benjamin M Good, Andrew I Su
Database (2020) https://doi.org/ggsdkj
DOI: 10.1093/database/baaa015 · PMID: 32283553 · PMCID: PMC7153956

41. **A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases**
Daniel N. Sosa, Alexander Derry, Margaret Guo, Eric Wei, Connor Brinton, Russ B. Altman
bioRxiv (2019-08-08) https://doi.org/gg5j64
DOI: 10.1101/727925

42. **Improving rare disease classification using imperfect knowledge graph**
Xuedong Li, Yue Wang, Dongwu Wang, Walter Yuan, Dezhong Peng, Qiaozhu Mei
BMC Medical Informatics and Decision Making (2019-12-05) https://doi.org/gg5j65
DOI: 10.1186/s12911-019-0938-1 · PMID: 31801534 · PMCID: PMC6894101

43. **Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data**
Felix Köpcke, Dorota Lubgan, Rainer Fietkau, Axel Scholler, Carla Nau, Michael Stürzl, Roland Croner, Hans-Ulrich Prokosch, Dennis Toddenroth
BMC Medical Informatics and Decision Making (2013-12-09) https://doi.org/f5jqvh
DOI: 10.1186/1472-6947-13-134 · PMID: 24321610 · PMCID: PMC4029400

44. **Analyzing bagging**
Peter Bühlmann, Bin Yu
The Annals of Statistics (2002-08) https://doi.org/btmtjp
DOI: 10.1214/aos/1031689014

45. **Utilising artificial intelligence to determine patients at risk of a rare disease: idiopathic pulmonary arterial hypertension**
David G. Kiely, Orla Doyle, Edmund Drage, Harvey Jenner, Valentina Salvatelli, Flora A. Daniels, John Rigg, Claude Schmitt, Yevgeniy Samyshkin, Allan Lawrie, Rito Bergemann
Pulmonary Circulation (2019-11-20) https://doi.org/gg4jc7
DOI: 10.1177/2045894019890549 · PMID: 31798836 · PMCID: PMC6868581

46. **Double-bagging: combining classifiers by bootstrap aggregation**
Torsten Hothorn, Berthold Lausen
Pattern Recognition (2003-06) https://doi.org/btzfh6
DOI: 10.1016/s0031-3203(02)00169-3

47. **Learning statistical models of phenotypes using noisy labeled training data**
Vibhu Agarwal, Tanya Podchiyska, Juan M Banda, Veena Goel, Tiffany I Leung, Evan P Minty, Timothy E Sweeney, Elsie Gyang, Nigam H Shah
Journal of the American Medical Informatics Association (2016-11) https://doi.org/f9bxf9
DOI: 10.1093/jamia/ocw028 · PMID: 27174893 · PMCID: PMC5070523

48. **Classification in the Presence of Label Noise: A Survey**
Benoit Frenay, Michel Verleysen
IEEE Transactions on Neural Networks and Learning Systems (2014-05) https://doi.org/f5zdgq
DOI: 10.1109/tnnls.2013.2292894 · PMID: 24808033

49. **Component-based face detection**
B. Heiselet, T. Serre, M. Pontil, T. Poggio
Institute of Electrical and Electronics Engineers (IEEE) (2005-08-25) https://doi.org/c89p2b
DOI: 10.1109/cvpr.2001.990537

50. **The Architecture of the Face and Eyes Detection System Based on Cascade Classifiers**
Andrzej Kasinski, Adam Schmidt
Advances in Soft Computing (2007) https://doi.org/cbzq9n
DOI: 10.1007/978-3-540-75175-5_16

51. **Real time facial expression recognition with AdaBoost**
Yubo Wang, Haizhou Ai, Bo Wu, Chang Huang
Institute of Electrical and Electronics Engineers (IEEE) (2004) https://doi.org/crv3sq
DOI: 10.1109/icpr.2004.1334680

52. **Efficient Estimation of Word Representations in Vector Space**
Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
arXiv (2013-01-16) https://arxiv.org/abs/1301.3781v3

53. **Learning to Identify Rare Disease Patients from Electronic Health Records.**
Rich Colbaugh, Kristin Glass, Christopher Rudolf, Mike Tremblay Volv Global Lausanne Switzerland
AMIA ... Annual Symposium proceedings. AMIA Symposium (2018-12-05)
https://www.ncbi.nlm.nih.gov/pubmed/30815073
PMID: 30815073 · PMCID: PMC6371307

54. **Machine learning for psychiatric patient triaging: an investigation of cascading classifiers.**
Vivek Kumar Singh, Utkarsh Shrivastava, Lina Bouayad, Balaji Padmanabhan, Anna Ialynytchev, Susan K Schultz
Journal of the American Medical Informatics Association : JAMIA (2018-11-01)
https://www.ncbi.nlm.nih.gov/pubmed/30380082
DOI: 10.1093/jamia/ocy109 · PMID: 30380082 · PMCID: PMC6213089

55. **CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data**
Elana J. Fertig, Jie Ding, Alexander V. Favorov, Giovanni Parmigiani, Michael F. Ochs
Bioinformatics (2010-11-01) https://doi.org/cwqsv4
DOI: 10.1093/bioinformatics/btq503 · PMID: 20810601 · PMCID: PMC3025742

56. **Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data**
Aaron M. Smith, Jonathan R. Walsh, John Long, Craig B. Davis, Peter Henstock, Martin R. Hodge, Mateusz Maciejewski, Xinmeng Jasmine Mu, Stephen Ra, Shanrong Zhao, ... Charles K. Fisher
BMC Bioinformatics (2020-03-20) https://doi.org/ggpc9d
DOI: 10.1186/s12859-020-3427-8 · PMID: 32197580 · PMCID: PMC7085143

57. **Convolutional Neural Networks for Diabetic Retinopathy**
Harry Pratt, Frans Coenen, Deborah M. Broadbent, Simon P. Harding, Yalin Zheng
Procedia Computer Science (2016) https://doi.org/gcgk75
DOI: 10.1016/j.procs.2016.07.014

58. **Opportunities and obstacles for deep learning in biology and medicine**
Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, ... Casey

S. Greene
Journal of The Royal Society Interface (2018-04-04) https://doi.org/gddkhn
DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

59. **Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder**
Sanjiv K. Dwivedi, Andreas Tjärnberg, Jesper Tegnér, Mika Gustafsson
Nature Communications (2020-02-12) https://doi.org/gg7krm
DOI: 10.1038/s41467-020-14666-6 · PMID: 32051402 · PMCID: PMC7016183

60. **DeepProfile: Deep learning of cancer molecular profiles for precision medicine**
Ayse Berceste Dincer, Safiye Celik, Naozumi Hiranuma, Su-In Lee
bioRxiv (2018-05-26) https://doi.org/gdj2j4
DOI: 10.1101/278739

61. **A Survey on Transfer Learning**
Sinno Jialin Pan, Qiang Yang
IEEE Transactions on Knowledge and Data Engineering (2010-10) https://doi.org/bc4vws
DOI: 10.1109/tkde.2009.191

62. **Pathway-level information extractor (PLIER) for gene expression data**
Weiguang Mao, Elena Zaslavsky, Boris M. Hartmann, Stuart C. Sealfon, Maria Chikina
Nature Methods (2019-06-27) https://doi.org/gf75g6
DOI: 10.1038/s41592-019-0456-1 · PMID: 31249421 · PMCID: PMC7262669

63. **Reproducible RNA-seq analysis using recount2**
Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, Jeffrey T Leek
Nature Biotechnology (2017-04-01) https://doi.org/gf75hp
DOI: 10.1038/nbt.3838 · PMID: 28398307 · PMCID: PMC6742427

64. **MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease**
Jaclyn N. Taroni, Peter C. Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A. Merkel, Casey S. Greene
Cell Systems (2019-05) https://doi.org/gf75g5
DOI: 10.1016/j.cels.2019.04.003 · PMID: 31121115 · PMCID: PMC6538307

65. **Transcription of proteinase 3 and related myelopoiesis genes in peripheral blood mononuclear cells of patients with active Wegener's granulomatosis**
Chris Cheadle, Alan E. Berger, Felipe Andrade, Regina James, Kristen Johnson, Tonya Watkins, Jin Kyun Park, Yu-Chi Chen, Eva Ehrlich, Marissa Mullins, … Stuart M. Levine
Arthritis & Rheumatism (2010-02-12) https://doi.org/chfbtv
DOI: 10.1002/art.27398 · PMID: 20155833 · PMCID: PMC2887718

66. **How transferable are features in deep neural networks?**
Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson
arXiv (2014-12-09) https://arxiv.org/abs/1411.1792

67.**:{unav)**
Rich Caruana
Machine Learning (1997) https://doi.org/d3gsgj
DOI: 10.1023/a:1007379606734

68. **Multi-task Learning via Adaptation to Similar Tasks for Mortality Prediction of Diverse Rare Diseases**
Luchen Liu, Zequn Liu, Haoxian Wu, Zichang Wang, Jianhao Shen, Yiping Song, Ming Zhang
arXiv (2020-04-11) https://arxiv.org/abs/2004.05318v2

69. **Generalizing from a Few Examples: A Survey on Few-Shot Learning**
Yaqing Wang, Quanming Yao, James Kwok, Lionel M. Ni
arXiv (2019-04-10) https://arxiv.org/abs/1904.05046v3

70. **Low Data Drug Discovery with One-Shot Learning**
Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, Vijay Pande
ACS Central Science (2017-04-03) https://doi.org/f95dnd
DOI: 10.1021/acscentsci.6b00367 · PMID: 28470045 · PMCID: PMC5408335

71. **Automatic detection of rare pathologies in fundus photographs using few-shot learning**
Gwenolé Quellec, Mathieu Lamard, Pierre-Henri Conze, Pascale Massin, Béatrice Cochener
arXiv (2019-07-22) https://arxiv.org/abs/1907.09449v3
DOI: 10.1016/j.media.2020.101660

72. **A Survey on Multi-Task Learning**
Yu Zhang, Qiang Yang
arXiv (2017-07-25) https://arxiv.org/abs/1707.08114v2