# Machine learning methods for rare diseases

*This manuscript (permalink) was automatically generated from jaybee84/ml-in-rd@216af17 on August 13, 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 is increasingly high-throughput and algorithms or approaches become more transparent and interpretable.

Application of machine learning to any dataset poses challenges, 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.

Rare disease research has not yet significantly benefited from machine learning applications for various reasons, 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. We anticipate higher occurrence of such applications in the near future and aim to highlight the current state-of-art in this perspective. However recent advances in the methodologies to accommodate rarity of samples and increased transparency in model outputs have encouraged application of machine learning in rare disease.

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. Moreover, in the context of rare diseases, special considerations need to be made at each of the above mentioned steps to safeguard against misinterpretation of data. 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.

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

This section will highlight promising approaches for analyzing rare disease data to extract biological insights. We will discuss techniques like transfer learning, representation learning, cascade learning, integrative analysis, and knowledge-graph creation and use that leverage other knowledge and data sources to construct testable hypotheses from rare diseases datasets with limited sample sizes.

## Ensemble Learning

Implementing machine learning on data with low sample size and high label uncertainty can lead to unstable predictions. In such cases where single predictors fail, various machine learning methods together or *ensemble learning* may help increase accuracy of prediction. *Ensemble learning* methods like random forests use bootstrap aggregation (or *bagging*) of independent decision trees that use similar parameters but different paths for the selection of features to form a consensus about the important predictive features[1]. 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 [6]. 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 of uncertainty (or “label-noise”) leading to a *silver standard* dataset[7]. In such cases, ensemble learning with multiple methods leveraging distinct underlying assumptions are used in tandem to capture stable patterns existing in the *silver standard* data and reduce uncertainty. Such *cascade learning* classifiers have been widely used in image recognition where initially a small subset of image features are used to classify images (e.g. features identifying a face like eyes, nose, mouth). The initial classification is then augmented by more complex features and algorithms like AdaBoost ( *boosting* ) that weight the various features implemented to detect the content of the image (e.g. features identifying a human face like relative distance between eyes etc.) [8].

In rare diseases, a variant of *cascade learning* that showed robustness in view of uncertainty in the data was implemented to identify rare disease patients from electronic health records from the general population [11]. This implementation consisted of three steps each employing a different independent learning algorithm: (1) feature extraction to assign text words (from Pubmed literature) to diagnosis using word2vec [12], (2) preliminary prediction using an ensemble of decision trees with penalization for excessive tree-depth, (3) prediction refinement using similarity of data points to resolve sample labels and reiterating step (2). In this implementation the algorithm was able to identify rare disease patients due to the robustness conferred by the independence of the feature extraction step and the prediction refinement step from the preliminary classification of the labeled dataset. The classification step capitalized upon the information learned by the label prediction step preceding it and the prediction refinement step following it, and was able to perform better over other ensemble methods when implemented on silver standard data.

Most cascade classifiers follow *one-classifier-at-a-time* approach where algorithms at each level predict all classes involved. But scenarios where the need for high prediction accuracy for one class outweighs other classes (e.g. malignant tumor-types, or severe psychiatric cases) require further modification of the cascade learning efforts. An example of this was seen implemented for triaging psychiatric patients where the identification of one class of psychiatric patients (“severe”) far outweighed the need for optimized overall classification accuracy[13]. Due to the requirements of the problem, they developed a *one-class-at-a-time* approach for cascade learning, where at each stage a binary classifier is used to predict a specific class against all others. The final model implemented all models together each identifying one class sequentially and the final prediction was the union of the predictions at all the different models. 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 data that is limited in quality or quantity, where single algorithms would otherwise produce unstable predictions. However, the choice of using *bagging*, *boosting*, independent algorithmic steps, or *one-class-at-a-time* approach would strictly depend on the nature of the prediction problem. In most cases involving rare disease data, it seems that *bagging* has had limited success, which has necessitated various modifications of the approaches as discussed above.

## Knowledge graphs for rare disease

An intrinsic constraint in the study of rare disease is the availability of large, normalized datasets. This limits our ability to study genotype-phenotype relationships or other key attributes of rare diseases. A potentially powerful strategy for evaluating genotype-phenotype relationships or repurposing drugs when large datasets are scarce or nonexistent is to develop and use knowledge graphs. Knowledge graphs integrate related-but-different data types, creating a rich and complex data source. Examples of well-known public biomedical knowledge graphs and graph frameworks that could be useful in the rare disease context include the Monarch Graph Database[14], hetionet[15], PheKnowLator[16], and the Global Network of Biomedical Relationships[17]. These graphs connect information like genetic, functional, chemical, clinical, and ontological data to enable the end user to explore many types of data and their relationships with disease phenotypes whether through manual review[18] or computational methods[19,20]. In the academic rare disease space, there a few pioneering examples of machine learning-based mining of knowledge graphs to repurpose drugs[19] and classify rare diseases[20]. 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. These methods also evaluate a broad swath of rare diseases in a relatively unguided manner, rather than interrogating a pre-defined 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 very few examples of studies in the public domain that leverage knowledge graphs to characterize rare disease. However, private entities (e.g. healx, Boehringer Ingelheim, DrugBankPlus) have established partnerships and are performing an undisclosed amount of work to create and explore proprietary rare disease knowledge graphs for machine learning-based drug discovery applications. The formation 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 but generally untapped area of rare disease research in the public sphere. More work needs to be done to assess 1) which graph networks and network features best capture the salient information about rare diseases, 2) the utility of an array of knowledge graph analysis methodologies (such as graph neural nets, reinforcement learning approaches, and adversarial learning approaches) [doi:10.1109/TKDE.2020.2981333] 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 meaningfully interrogated using machine learning of knowledge graphs.

## Representation learning

Representation learning, also called feature learning, is the process of learning features from raw data, where a feature is an individual variable or property. An algorithm or approach will construct features as part of training and, in the case of supervised feature learning, 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 [21,22]. Low-dimensional representations trained on a collection of transcriptomic data can then be used as input to supervised machine learning methods [23]. Supervised neural networks used in medical imaging studies [24] (reviewed in [25] and [26]), which are trained to predict labels or classes, are also an example of representation learning.

Whether or not a learned representation or set of features is *useful* depends on the task at hand. In a supervised setting, it may be sufficient for a feature to distinguish between classes, but if we hope to use a feature for biological discovery, we may prioritize intepretability. For example, learned features in the medical imaging domain may be a series of edges that constitute a blood vessel formation that discriminates between disease states or the learned features may not align with characteristics recognized as important by domain experts at all. From transcriptomics data, learned features could be coordinated sets of genes involved in a biological process that are descriptive in some way [27], but do not necessarily distinguish cases from controls in our study.

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 [28]. The authors trained a variational autoencoder (VAE) on AML data that had been collected over time without the phenotypic information they were interested in–in this case, drug response. A VAE is an unsupervised neural network that learns a series of representations or encodings from data, where each learned attribute will have a probability distribution associated with it rather than a single value. 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 publicly available AML samples from many different studies used as part of the training set in Dincer et al. [28], we expect that in other rare diseases considerably fewer samples will be available for training or, in the case of systemic diseases, studies may be from different tissues. The study by Dincer and colleagues highlights another challenge: even if enough samples have been collected over time to support representation learning, these samples 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 biological patterns that are a normal part of development aberrant in a particular disease context or an imaging anomaly present in both 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 principally on feature-representation-transfer [29] here. Feature-representation-transfer approaches learn representations from the source domain and apply them to a target domain [29]. 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 [28]. 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) [30] on a large generic collection of human transcriptomic data (recount2 [31]) 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 [32]. (Here “unseen” refers to the fact that no AAV or medulloblastoma data were included 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; PLIER utilizes regularization such that some latent variables learned by the model will align with input gene sets and, for those latent variables that are aligned with gene sets, latent variables will be associated with a limited number of gene sets [30]. We demonstrated that training on larger collections of randomly selected samples produced models that captured a larger proportion of input gene sets and were more suitable for disentangling closely related signals (e.g., type I and type II interferon signaling), suggesting that larger training sets produced PLIER models that are more valuable for biological discovery [32].

Though models trained on general collections of transcriptomic data had more appealing properties, that alone does not guarantee suitability for describing rare diseases. To assess suitability, we must find ways to examine the relevance of learned features (latent variables) to the rare disease gene expression data. In Taroni et al., we found that the expression of latent variables that could be matched between the MultiPLIER model and a model trained on a rare disease dataset alone were well-correlated, particularly in the case of latent variables that were biologically relevant (i.e., significantly associated with input gene sets) [32]. 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 a group of samples from patients with AAV that were reported to have more severe disease [33]. 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 normal 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: not all latent variables learned by a PLIER model are associated with input gene sets, and therefore may capture technical noise separately from biological signal, and we can use a single model to describe datasets from multiple tissues or cohorts that are not obviously directly comparable instead of attempting to reconcile results from multiple models (see *05.heterogeneity.md*).

Taken together, the DeepProfile [28] and MultiPLIER [32] results 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 [34]. 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 (e.g., using PLIER for biological discovery) and evaluation strategies that are well-aligned with a research goal are crucial for successful application of these approaches in rare diseases.

## 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 or resampling computation is a powerful statistical technique that can be used for estimating population values from datasets of limited sample size [35]. The technique utilizes random sampling of data points from a dataset of limited sample size with replacement to approximate a larger population and estimate various population statistics (e.g. mean). Subsequent iterations of resampling generates a distribution of the statistical value (mean) which minimizes the error of the estimate. 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 [2,36], bootstrap in neural networks [37], or regression models [38,39]).

While most datasets in practice are of finite sample size and can benefit from bootstrapping, rare disease datasets with limited number of samples necessitate the use of bootstrap to form an informative dataset in addition to model selection [40]. In this study, bootstrapping the training sample without replacement simulated formation of different incomplete datasets that helped expose the learning models (in this case random forests) to the incompleteness of the data. Such additional bootstrapping of the training data helped create confidence intervals for the predictions and the important predictors originating from unstable ensemble models run on the incomplete training data

## Regularization

Machine learning algorithms are optimized to find patterns among data points and prioritizes the strongest patterns that exist in a dataset. Given a limited dataset with strong pre-existing technical differences between groups of samples, this optimization may lead to the model learning technical differences thus lowering its predictive accuracy [41]. For example, in a set of 1000 samples where 700 samples are from one healthcare site and 300 from another, it is likely that there will remain site-specific differences between them even after normalization of the samples. If the site-specific differences are more pronounced than the underlying patterns differentiating the samples, any machine learning model trained with these data will preferentially learn the site-specific differences to classify the samples, and rank them higher than the underlying patterns leading to a model showing high prediction accuracy of training data (termed low bias in model). When new test data points are introduced to the model, possibly coming from a third site, the model is unable to locate the earlier differences in the new data points and fails to classify them accurately causing a significant drop in accuracy of the model (termed high variance in model prediction). Such a model is termed “overfit” to its training data. Overfitting can lead to misinterpretation of the site-specific differences as true patterns in the limited data points and thus needs to be minimized. Minimization of overfitting can be accomplished by cross-validation and regularization methodologies.

While cross-validation aims to reduce the variance in prediction, regularization adds a small amount of bias to the initial model to minimize its dependence and sensitivity to training data. Regularization makes models less reliant on training data by adding a penalty (determined by cross-validation), and then minimizes the error between the model’s prediction and ground truth of the test data. Regularization can not only minimize overfitting but can additionally help in predicting outcomes using a limited number of samples.

Regularization can be of three main types, each with their particular strengths and weaknesses. (1) Ridge regression aims to minimize the magnitude of the features, but in models that try to select the most important features for accurate prediction of sample labels, ridge regression shrinks all features equally, but cannot completely remove unimportant features. Thus in presence of many correlated parameters (e.g. gene expression networks), ridge regression may not be ideal in reducing the feature space. (2) LASSO or least absolute shrinkage and selection operator regression on the other hand works well for selecting few important features since its effect can minimize the magnitude of some features more than the others. Thus it helps in selecting most important features while the magnitude of irrelevant features are shrunk to 0 and eventually removed. This selection attribute of LASSO (in a sample set of size “n”, LASSO can select “n” features for the model) may be an advantage in reducing model complexity, but a disadvantage in cases where identification of all possible collinear features is important (e.g. all biomarkers correlating to a particular disease phenotype) [42]. (3) Elastic-Net regression is a combination of LASSO and ridge regression[43]. Both of the methodologies when applied together helps to select most useful features, specially where there are a lot of correlated features. In this setup, LASSO leads to selection of one of the correlated features and reduces the others to 0 (grouping of features), and the magnitude of the selected features are then minimized through ridge regression.

Any supervised learning implementation in rare disease would require robustness towards feature selection from a small number of samples, i.e. the features selected by a model as important should be stable in view of new data points added to a dataset, even though their relative importance may change due to additional evidence. This robustness is mostly acquired through the combination of various regression strategies. Since machine learning applications in rare disease are infrequent, combination strategies used for rare variant discovery and immune cell signature discovery can serve as good case studies to examine. Many deleterious genomic variants can be extremely rare due to the constant selection pressure working against them. Since the frequency of a rare variant is so low (less than 1%) applying routine statistical procedures that were extensively developed for common variant association, to analyze a low minor allele frequency (MAF) seem inappropriate [44]. For its feature selection attribute, LASSO has been widely applied in microarray and GWAS data for common variants. But since LASSO by itself is too stringent for rare variants, it has been employed along with group penalties to help identify rare variants/ low frequency predictors [45]. Variations of LASSO have also been implemented to aggregate or group the occurrence of rare variants together by gene or chromosome location [46,47,48]. In this strategy, a 0–1 dummy variable was created for each SNP based on the presence or absence of the rare variant. Then linear combinations of the selected dummy variables were considered by using the LASSO procedure. Even though most of the dummy variables were 0, their linear combination was more likely to be nonzero thus leading to increased signal to noise ratio for the rare variants. Only those linear combinations that were non-zero in at least 5% of the subjects were then included to ensure that the new markers were not rare [47,49]. While ridge regression is not usually utilized for feature selection, adaptive ridge regression has been utilized to help combine rare variants into a single score analogous to feature engineering for increasing the signal of rare variants[50]. Another variation of LASSO included its integration with the probabilistic logistic bayesian approach to identify a protective rare variant in lung cancer[51]. Xu et al. on the other hand combined the feature selection methods with a generalized pooling strategy, and evaluated the performance of these hybrid approaches for detection of rare genetic variants[52]. Another interesting approach is the sparse-group LASSO approach which incorporates prior knowledge into the regularization[53]. This approach works well for a scenario where only few genes in a pathway are true predictors of a phenotype, where it helps select the driving genes in a pathway of interest.

Alternatively, Elastic-net regression (a combination of LASSO and ridge regression) has also been used to reduce the feature space in various types of cancer datasets [54,55]. In cases where the number of features were far greater than the number of samples, elastic-net has usually been found to outperform the other regression approaches [43]. A variation of the elastic-net regression was used for identifying 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]. This two-step regularized logistic regression technique included a pre-filtering phase to select the optimal number of genes and then implemented elastic-net regularization for gene selection. The second step generated gene signatures for individual cell types using selected genes from first step and then implemented a binary regularized logistic regression for each cell type against all other samples

Still to add: techniques in deep learning e.g. Deep and shallow architecture: https://ieeexplore.ieee.org/document/7863293

## Techniques to manage disparities in data generation are required to power robust analyses in rare diseases

As with common diseases, genomic and transcriptomic data from rare diseases can suffer from artifacts introduced by batch, processing methodology, sequencing platform, specimen/data quality or other non-biological phenomena. The consequences of these non-biological artifacts are amplified in rare diseases which often have few samples and heterogeneous phenotypes. Furthermore, because datasets are many times pieced together from multiple small studies, in which disease phenotypes or other important biological characteristics are often confounded by the previously mentioned “batch” factors. A key consideration here is, if possible, active dialogue with the data generators or experts in the field who may have unexpected insights into potential sources of variation. One example of the value of this, experienced by the authors, occurred when studying tumors associated with the disease neurofibromatosis type 1. These datasets were, unbeknownst to the computational biologists, generated from samples obtained with vastly different surgical techniques (laser ablation and excision vs standard excision), resulting in substantial biological differences that are a consequence of process, not reality. One might expect, in this example, that this technical decision would result in profound changes in the underlying biology, such as the activation of heat shock protein related pathways, unfolded protein responses, and so on. Consequently, careful assessment of confounding factors and implementation of normalization methods is important to identifying biologically meaningful features within a dataset. Assessment of confounding factors and heterogeneity in rare disease datasets is perhaps most easily performed using unsupervised learning approaches such as clustering and dimensionality reduction. Clustering methods like k-means clustering or hierarchical clustering can be used to characterize the structure present in many different types of data such as genomic or imaging data. [56,57]. Similarly, a variety of dimensionality reduction methods are can be used to visualize sample heterogeneity and potential confounding variables, including multidimensional scaling (MDS), principal components analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP), among many others. [58,59,60,61] All of these methods can be effectively used to identify batch effects and other structure in the data, though some, like t-SNE and UMAP, have parameters, such as perplexity (number of nearest neighbors), that can substantially affect the output, and thus the interpretation, of the analysis [61,62]. Therefore, successful application of these methods requires a sufficient understanding of the underlying method and parameter sweeping to get a clear picture of the structure of the underlying data. Another important consideration is that, as discussed by Way, et. al. [22], 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 [63], mass cytometry data [64], and others. Once the nature of the non-biological heterogeneity has been established, different techniques can be used to correct the differences. Common approaches to ameliorate non-biological effects include the assessment of data quality using robust metrics, reprocessing the raw data using a single analysis pipeline if the data are obtained from different sources, application of batch correction methods [65,66], normalization of raw values (e.g. z-scores, trimmed mean of M-values [67]). It can also be helpful to be fatalistic, in some sense, when working with rare disease data. For various reasons including ethical considerations, limited funding, and limited biospecimen availability, experimental design and the resulting data will be less-than-ideal - for example - when batch variables and biological variables are confounded. In these cases, it may be prudent to take a step back, re-evaluate the data, and identify methods that can operate within these constraints.

## Conclusions

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

# References

1. **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

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

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

4. **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

5. **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

6. **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

7. **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

8. **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

9. **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

10. **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

11. **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

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

18. **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

19. **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

20. **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

21. **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

22. **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

23. **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

24. **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

25. **An overview of deep learning in medical imaging focusing on MRI**
Alexander Selvikvåg Lundervold, Arvid Lundervold
*Zeitschrift für Medizinische Physik* (2019-05) https://doi.org/ggp8vt
DOI: 10.1016/j.zemedi.2018.11.002 · PMID: 30553609

26. **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

27. **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

28. **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

29. **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

30. **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

31. **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

32. **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

33. **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

34. **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

35. **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

36. **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

37. **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

38. **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

39. **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

40. **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

41. **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

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

43. **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

44. **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

45. **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

46. **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

47. **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

48. **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

49. **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

50. **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

51. **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

52. **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

53. **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

54. **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

55. **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

56. **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

57. **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

58. **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

59. **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

60. (2020-06-01) https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

61. **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

62. **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

63. **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

64. **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

65. **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

66. **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

67. **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