

Biomedical Data Science Regulation through Systems Immunology in Immune Disease

Sadia Tasneem

Jun 1, 2025

Abstract

The article describes biomedical data science as a way to convert high-dimensional systems-immunology measurements into interpretable models of immune disease. Single-cell transcriptomics, cytometry, receptor sequencing, and proteomics measure different layers of immune state, but technical noise and biological heterogeneity can create misleading patterns. The article emphasizes prospective sampling before outcomes, careful preprocessing, stable assay design, independent external validation, and functional testing of computationally identified mechanisms. Models should perform across institutions, demographic groups, instruments, and workflows rather than only within the discovery environment. Federated analysis is discussed when privacy limits direct data pooling. Uncertainty, provenance, interpretability, and subgroup robustness are essential because predictions may influence immunosuppression or invasive testing. The central goal is not merely to classify immune disease, but to identify a reproducible biological pattern that supports a specific clinical decision, explains its evidentiary limits, and remains reliable when deployed beyond the original dataset.



Biomedical data science transforms systems immunology into models of coordinated immune behavior from millions of measurements of cells. One cytokine or cell count can not be used to understand the immune system; responses occur at multiple time points, with multiple cell types, signaling pathways, and tissues all interacting. Today, technologies like high dimensional technologies measure that complexity directly. The analytical challenge is to separate signal from noise and to relate computational patterns with mechanisms that are relevant to immune disease.

Single cell technologies come from the raw material. Immune populations can be described at an unprecedented depth using mass cytometry, flow cytometry, single cell RNA sequencing, receptor sequencing and proteomic assays. A different layer is represented on each platform. Transcriptomics is the measure of gene expression, cytometry of proteins and signaling states, receptor sequencing of clonal lineages and soluble assays of extracellular environment.

Systems immunology does not interpret each layer alone, but rather integrates them. All these changes can occur simultaneously in a disease. Multivariate models may be able to find combinations that are more discriminating in distinguishing disease stages or predicting response, than any single feature. The combination is useful only if the model is interpretable and reproducible.

The advantage of the longitudinal design is its ability to capture changes in the immune system in the natural state. The coordinated immune features change predictably

throughout pregnancy, and pregnancy studies have shown that complex immune variation can have a biological clock. The same applies to infection, vaccination, autoimmune flare and treatment response. Repeated sampling yields a separation of stable personal characteristics and state dependent changes.

While machine learning can identify structure in these measurements, the accuracy isn't enough. Technical artifacts, batch effects, age-related effects, or medication effects that coincidentally relate to disease may be used by a classifier. However, several steps must be taken to ensure proper validation, such as using a held out data set, a separate cohort, careful data preprocessing, and biological checking of influential features. A model that works well in one set of data doesn't necessarily work well in another set of data.

That can be mitigated by a prospective immune phenotyping. Cleaner prediction testing occurs when samples are taken prior to clinical outcomes on a predetermined protocol. Baseline immune states can be used to compare to later remission, relapse, infection severity or vaccine response. A signature that pre-predicts outcome before treatment, has greater translational value than a signature found after outcome is apparent.

Single cell multiomics is a tool for linking prediction to mechanism. While a model can be used to recognize a pattern of interferon, multiomic data can be used to determine the cell types that contain the signal and whether or not the chromatin state is conducive to sustained activation. A change in a T cell can be determined by receptor sequencing to know whether it is a clonal expansion. Functional stimulation can test and determine if signaling capacity is changed, from descriptive association to causal biology.

In addition, antigen specific assays provide a further dimension as immune competence is not completely reflected in resting state. Cells may look normal but when challenged will show abnormalities. Signaling reserve, cytokine production, proliferation, and memory, are revealed by stimulation with defined antigens or cytokines. Sometimes dynamic responses can be better at discriminating disease states than baseline abundance.

Scale is a critical factor when dealing with data integration. There may be thousands of cells in a data set that belong to one person, and only a dozen or so people in the data set. When you treat each cell as an independent player, certainty will be overrated. A hierarchical structure is required in statistical models to account for cells being nested in samples and samples in individuals. Biological replication is at the participant level and not just at the level of the number of cells measured.

Missing data are another practical problem. It is rare that all assays are done for all

samples, in multiomic studies. There are also tissues which are not easily retrievable more than once and which may fail to be assayable more often than others. Computational imputation can be useful, but imputed variables should not silently be treated as measured biology. Results are sensitive to the assumptions about missingness and this can be displayed by performing sensitivity analyses.

Calibration is the cornerstone for precision medicine. A high risk prediction model should be associated with a known likelihood of the clinical outcome, rather than just a ranking score. The decision threshold should be set taking into account the risk of false positive and false negative. In cases of immune disease, over-suppression is a risk and delayed treatment can allow irreversible damage. Therefore, the clinical utility of the model should also be assessed in addition to the discrimination.

The link between data science and medicine is external validation. If a signature has been found at one institution and is to be tested, it must be done utilizing independent samples, preferably different technical and demographic conditions. When performance drops, the failure can expose what features are cohort specific. Strong signals in system immunology benefit from being platform, geographic, and workflow independent.

Test for feature stability should be conducted over technical replicates. The single cell clustering may change if the algorithm changes, the random seeds change, or the annotation rules change. This should not be based on one arbitrary cluster boundary, which would be meaningless in a clinically useful signature. By mapping predictive features to reproducible cell states or measured proteins, the result can be more robust.

Immunology is a difficult area in which to establish reference ranges due to the broad healthy ranges. Individual baseline can sometimes be more informative than population baselines. Multiple measurements prior to the onset of disease activity can create an individual's immune range, and then when the immune response shifts, it can be detected, but with a margin of error smaller than the margin of error in the population.

Causal graphs can be used to structure the hypotheses prior to modelling. Medication, disease activity, immune cells, cytokines and outcomes may form directed associations that can also help identify which factors may be confounders and which may be mediators. The graph does not prove causality, but it helps ensure that decisions for adjusting are not made that will inadvertently disrupt the pathway that is being studied.

One of the scientific validities are reproducible pipelines. Before external testing, Code, Preprocessing Rules, Annotation Dictionaries, and Model Versions should be fixed and

documented. High dimensional results can be influenced by small analytical choices. Transparency of pipelines allows identification if failure to replicate is biological or computational.

When privacy and governance constraints mean that immune datasets cannot be combined, federated analysis can be helpful. Models can be trained institution-wide without the need for data to be shared, except for differences in assay platforms which must be harmonized. External testing across sites is useful as it provides insight into possible dependency of one laboratory environment.

The uncertainty must also be conveyed when interpreting. Probabilistic models can give plausible bounds on cell state estimates or risk predictions. A prediction that is only borderline should lead to a different response than a high confidence prediction. This is particularly important if the downstream action is immunosuppression or invasive testing.

Each feature should have its data provenance. Researchers should know the instrument, reagent lot, processing pipeline and sample condition that produced a measurement. With Provenance, technical drift can be traced in case of model failure. It also helps to ensure reproducibility if the assay is attempted by another lab. Metadata quality is crucial in high dimensional immunology, as is the molecular depth.

Interpretability should be taken to the level of the patient. The immune features that underpinned a prediction, the direction of change of those features, and the uncertainty of each immune feature can all be included with the prediction. These explanations can uncover improbable predictions, inform review, and indicate if the model is reacting to the biology of disease or to an unusual artifact of the model itself.

This evolution from isolated biomarkers to interpretable immune systems is reflected in Nima Aghaeepour's research interests in biomedical data science, systems immunology, and precision medicine. The most effective strategy is to include prospective sampling, single cell measurements, functional assays along with computer models which can be validated on an independent discovery dataset. Clinically meaningful biomedical data science is one that does more than classify disease. It should reveal a consistent immune mechanism, measure the uncertainty, and provide the support for a decision that can be tested in a prospective manner.

That discipline enhances scientific reproducibility and, ultimately, clinical confidence.