nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-023-00781-8

Improving generalization of machine learning-identified biomarkers using causal modelling with examples from immune receptor diagnostics

Received: 19 April 2023

Accepted: 4 December 2023

Published online: 24 January 2024

Check for updates

Milena Pavlović[1,2,3] ✉, Ghadi S. Al Hajj[1], Chakravarthi Kanduri[1,3], Johan Pensar[4], Mollie E. Wood[5,6], Ludvig M. Sollid[2,7], Victor Greiff[7] & Geir K. Sandve[1,2,3] ✉

Machine learning is increasingly used to discover diagnostic and prognostic biomarkers from high-dimensional molecular data. However, a variety of factors related to experimental design may affect the ability to learn generalizable and clinically applicable diagnostics. Here we argue that a causal perspective improves the identification of these challenges and formalizes their relation to the robustness and generalization of machine learning-based diagnostics. To make for a concrete discussion, we focus on a specific, recently established high-dimensional biomarker—adaptive immune receptor repertoires (AIRRs). Through simulations, we illustrate how major biological and experimental factors of the AIRR domain may influence the learned biomarkers. In conclusion, we argue that causal modelling improves machine learning-based biomarker robustness by identifying stable relations between variables and guiding the adjustment of the relations and variables that vary between populations.

High-throughput sequencing technologies enable analyses of various patient characteristics, such as genetic variation[1], DNA methylation[2], gene expression[3], gut microbiota[4] and adaptive immune receptor repertoires (AIRRs)[5]. Such molecular and biological markers (biomarkers), defined as objective indications of the medical state that can be accurately and reproducibly measured[6], hold great promise for machine learning (ML) disease diagnostics[2,3,5]. However, several challenges exist to using ML: study participants are selected based on availability ('convenience sampling'), and data collected at multiple locations or distinct time points are combined, which may introduce systematic differences between datasets. A failure to account for such differences (for example, measurement errors or batch effects) can introduce selection and confounding biases that lead to models failing in real-world applications, despite showing promising performance during diagnostic development[7–10]. Finally, sequencing data are typically high-dimensional, making it more challenging to disentangle noise and biases from the true markers of the disease.

ML approaches examine these challenges from a purely statistical perspective by anticipating how the distributions of features or labels will change (a phenomenon called 'dataset shift' or 'distributional shift') and include domain adaptation[11,12] (when some information about the target domain is available) and domain generalization techniques[13–15] (where the target domain is unknown, but (multiple)

[1]Centre for Bioinformatics, Department of Informatics, University of Oslo, Oslo, Norway. [2]K.G. Jebsen Centre for Coeliac Disease Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [3]UiO:RealArt Convergence Environment, University of Oslo, Oslo, Norway. [4]Department of Mathematics, University of Oslo, Oslo, Norway. [5]Department of Pharmacy, University of Oslo, Oslo, Norway. [6]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [7]Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway. ✉e-mail: milenpa@uio.no; geirksa@uio.no

BOX 1

A brief introduction to causal inference

Causal inference aims to estimate causal effects between variables of interest, typically by introducing certain assumptions regarding how the variables are causally related. We say that variable C has a causal effect on variable E if intervening to change C would change (the distribution of) E[16].

Here, we briefly define basic concepts and different types of variable important in the causal inference field. A more detailed account of the field and its connection to ML is available elsewhere[20,29].

A 'structural causal model' consists of a set of variables of interest and a set of functions that describe how the values of the variables are assigned and their dependencies on other variables. Such models describe (often partially) the data-generating process. The causal structure of the model can be represented by a causal graph (Fig. 1b), where the nodes represent variables, and each edge defines the influence of one variable on another. The absence of an edge between two nodes implies no direct causal relation between them.

Assuming a causal graph structure that includes all common causes, typically represented in the form of a directed acyclic graph, the do-calculus framework[16] can be used to identify whether it is possible to estimate the causal effect between two variables C (cause) and E (effect) from the available data. Moreover, when a causal effect is identifiable, do-calculus will also provide an expression that non-parametrically specifies how to estimate the causal effect.

Other variables might influence the effect estimation in different ways[18]: 'confounders' are variables that causally affect both C and E (C←confounder→E), 'colliders' are influenced by both C and E (C→collider←E), 'mediators' are intermediary variables between C and E (C→mediator→E) that describe the mechanisms of how C influences E indirectly. 'Moderators' (effect modifiers) change the relation between C and E depending on the moderators' values[85]. 'Precision covariates'[54] are variables that influence only C or only E and may be used to improve the precision of the estimators in some cases.

Causal graphs can be arbitrarily complex, and variables may not have as clear roles as described above. To enable consistent estimation of the causal effect, the paths in the graph must be analysed to prevent bias. An instrumental technique for preventing bias is the 'backdoor criterion', where the idea is to close all non-causal paths with incoming arrows into both C and E (backdoor paths) while simultaneously keeping all directed (causal) paths from C to E open. A path is open if every collider on the path (or a descendant of the collider) is controlled for and any other variable on the path is not controlled for. Controlling for a confounder (for example, age, Fig. 1b) is a simple example of closing a backdoor path.

source domains are available). Traditionally, these approaches do not consider causal relations between features and labels.

More recently, the causal inference framework[16–18] has also been applied to describe dataset shifts using formal definitions with respect to proposed causal models of the underlying processes[19,20]. Do-calculus[16], the fully non-parametric CI framework described by Pearl (Box 1), can be used to estimate causal effects from non-experimental data whenever the effect is identifiable under a given causal model. The structure of the causal model is typically encoded using a (directed acyclic) causal graph, where the nodes represent the variables of interest, and the directed edges between the nodes represent direct causal relationships.

One way to obtain the causal structure of a biological process is from domain knowledge. However, this is challenging due to the complex and unknown nature of disease mechanisms. Alternatively, obtaining a causal structure can be approached by (1) learning the structure from data using causal structure learning[21,22], or (2) learning only the part of the causal model that results in robust (or invariant) predictions across different environments[23–27]. Causal structure learning typically attempts to infer the complete structure, which is difficult due to the high-dimensional nature of the problem, yet it can be applied to a single set of observational data. Invariant prediction is more focused: it attempts to identify features producing stable predictions for the variable of interest under general interventions (that is, different environments). However, it relies on multiple datasets generated under a sufficiently diverse set of interventions. When building the final ML model, accounting for the inferred causal structure should improve robustness to various dataset shifts[7,19,28].

In addition to potentially improving the performance of learned models, causal inference can help to formally or intuitively analyse diagnostic robustness across application contexts. Causal inference might also be useful when combining multiple datasets sampled under heterogeneous conditions to answer a probabilistic (or causal) query of interest, thus dealing with biases emerging due to environment change, confounding and participant selection[29]. This motivates considering the causal perspective as an essential component for diagnostics[30], medical image analysis[8], decision-making in healthcare[31] and the clinic in general[32].

When analysing sequencing data to discover molecular biomarkers of disease, three possible underlying causal structures may connect a biomarker with a disease. The biomarkers from sequencing data may be causing the disease, may arise as an effect of the disease, or the biomarkers and the disease may both have a common cause. Disregarding the common cause scenario for now, this means that the diagnostic prediction may be in a causal direction (predicting the effect from causes; for example, finding changes in the sequencing data that have played a role in causing a disease) or in an anticausal direction[33] (predicting causes from effects; for example, finding differences in the sequencing data that occurred as a consequence of the disease, Fig. 1c). Depending on this direction, dataset shifts will manifest in different ways. For example, when predicting in the causal direction ($X \to Y$), we might expect the performance to be more stable under changes to $P(X)$ because our target, $P(Y|X)$, is a component in the causal factorization $P(X, Y) = P(X)P(Y|X)$ and thus independent of $P(X)$ due to the principle of independent causal mechanisms[27]. On the other hand, we would expect no such robustness under the anticausal direction ($X \leftarrow Y$) because $P(Y|X)$ does not follow the causal structure.

Because of the opportunities for ML in diagnostics, we build on existing literature on causal inference and ML[20,27] and focus on complex, real-world applications in the field of adaptive immune receptor repertoires (AIRRs), which are increasingly used for diagnostic purposes[5,34]. AIRRs are high-dimensional molecular markers reflecting an individual's past and present immune responses and can be obtained from targeted high-throughput sequencing from a blood sample (Fig. 1a and Box 2). AIRR-based approaches may enable earlier diagnosis and prognosis, complement existing diagnostic tests, and have, in principle, the capacity to diagnose a broad range of diseases with a single test[5]. However, validation studies on external cohorts (for example, expanding on existing approaches[35]) are needed to establish the robustness of existing models.

In the following sections, we define and discuss different challenges in the study design for biomarker discovery, including confounders, batch effects, selection bias, variability of causal models

https://doi.org/10.1038/s42256-023-00781-8

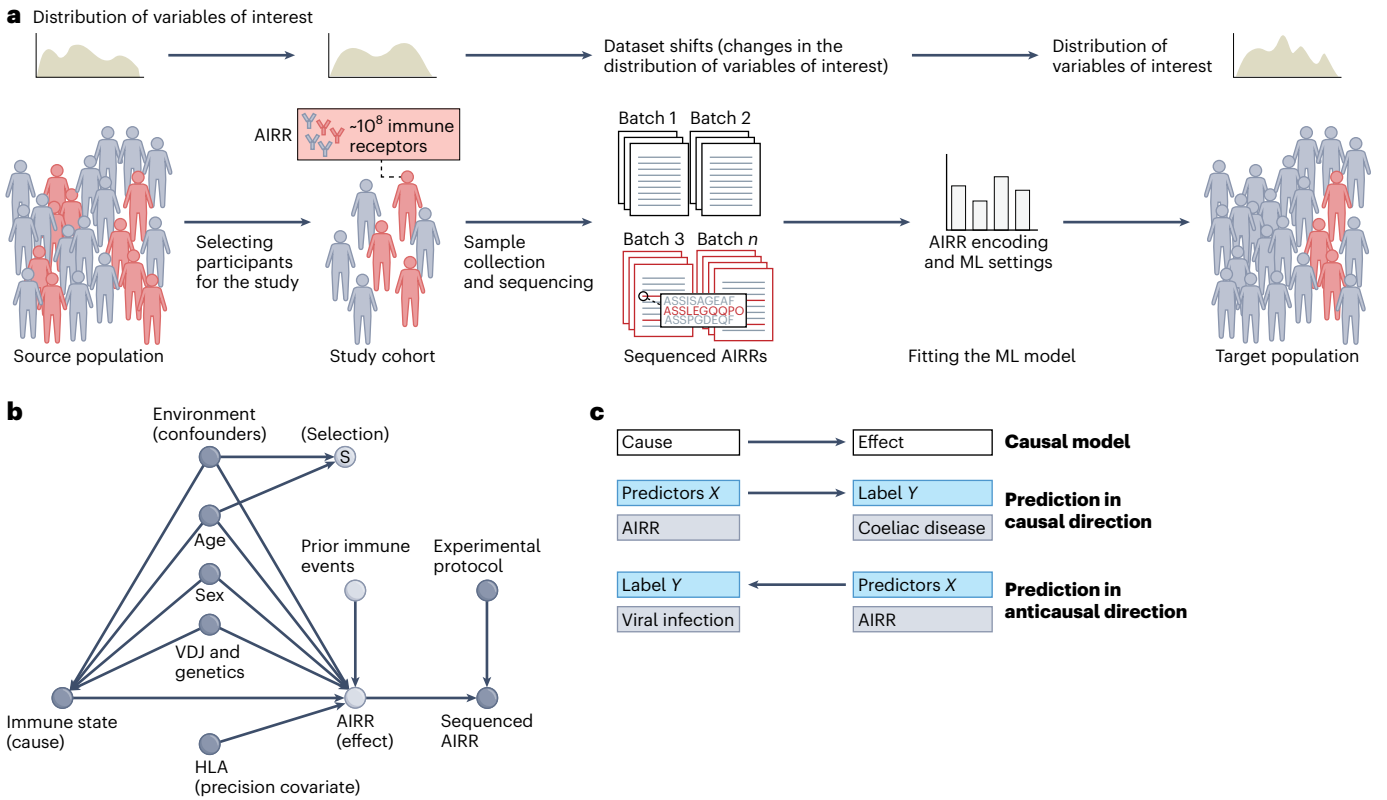


Fig. 1 | Developing an AIRR-based diagnostic. a, Overview of the diagnostic pipeline based on AIRRs, including patient collection from the source population, sampling, sequencing with batch effects, ML method development, and application in the target population. **b**, An example of a causal structure of AIRRs and the immune state, where the nodes represent the different variables involved (repertoire, HLA type, age, immune state) and the arrows represent causal relationships between variables. Filled nodes in the graph (immune state, HLA type, sequenced AIRR) denote observed variables, and open nodes are not observed (prior immune state). The node with S inside is the selection node, with edges showing what variables influenced the selection of participants for the study. **c**, For AIRR-based diagnostics, predictions may be either made in the causal direction (predicting the effect from the cause, for example, in autoimmunity) or in the anticausal direction (predicting the cause from the effect, for example, in infections), making the models predicting in the anticausal direction potentially less stable because they are not modelling the biological mechanism.

across diseases, and the high dimensionality of molecular data. We present a simulation study to illustrate these concerns and conclude with a proposal of reporting standards and study design advice, finally outlining open questions in the field.

Challenges in AIRR diagnostics study design

The main challenge of ML in diagnostic settings is whether the probability distributions learned from the training data will generalize to new application settings.

The set of examples (study participants) available at training time will be called the study sample (or study cohort), sampled from the underlying source (development) population (Fig. 1a). The population where the classifier will be applied is the target (deployment) population (environment or domain). If the source and target populations have the same joint probability distribution (disease prevalence and feature distribution, and the relations between them stay the same) and examples (for example, AIRRs) are independent and identically distributed (i.i.d.), the estimated ML model can be readily applied to the target population, provided that the model is internally valid (Box 3).

Although statistically convenient, the i.i.d. assumption rarely holds in the real world[36]—the probability distribution might change from source to the target population in the marginal or conditional distribution of variables. Marginal distributions may vary due to label shift (for example, change in disease prevalence) or covariate shift (for example, change in age distribution). The conditional distribution of variables may change if it describes an anticausal relation[33] (when predicting the cause from the effect, for example, immune state from AIRR; Fig. 1c) or due to the occurrence of unstable mechanisms[7] (for example, changing the time of sequencing in the course of the disease might result in estimates that only hold for the study cohort). Importantly, these shifts reflect systematic biases that would hold up even if a study cohort was infinitely large, and their extent cannot be quantified based only on information from the source population. The biases may arise from different aspects of the data-generating process, and when related to both AIRR and the immune state, they might introduce spurious correlations.

To illustrate these concerns, we provide an overview of AIRR-based diagnostic development (Fig. 1a). The study cohort is selected, and the targeted cell population (for example, T cells) is DNA-sequenced and analysed using ML. We also introduce an example of an AIRR-based diagnostic for a viral infection (Fig. 1b). In this example, the immune state is defined as the presence of the pathogen that (causally) changes AIRR. In addition to the immune state, previous immune events (for example, infections or vaccinations), age[37], sex[38], genetics (including the V(D)J recombination model[39]), the environment (for example, geographical location) and human leucocyte antigen (HLA)[40,41] also influence the AIRR. Finally, the observed sequencing data reflect only a limited proportion of a patient's full AIRR and introduce sampling variability. The experimental protocol may also introduce systematic biases in terms of which receptors are captured[42–44], which is especially problematic if the experimental protocol varies in a way that correlates with other patient variables. The causal graphs might differ for different types of disease.

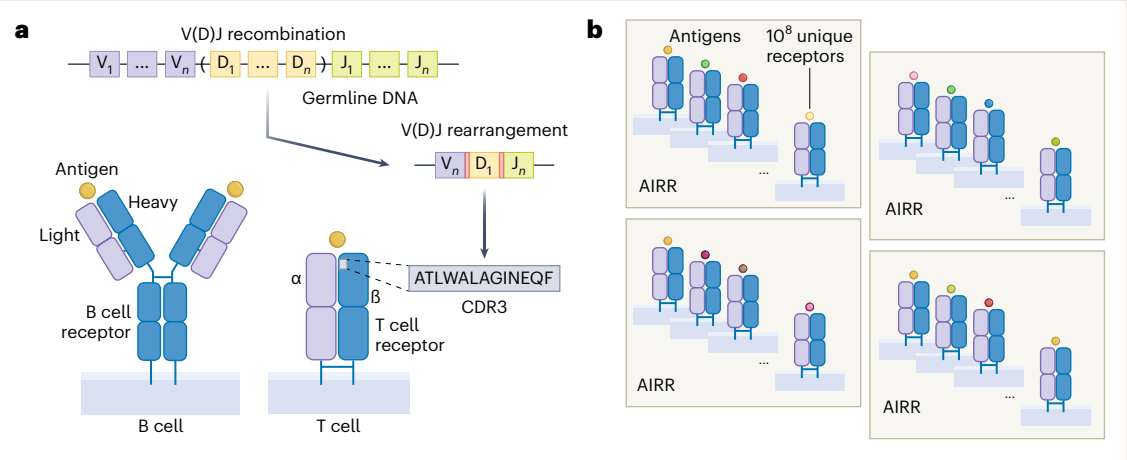
https://doi.org/10.1038/s42256-023-00781-8

BOX 2

Adaptive immune receptor repertoires

Adaptive immune receptors (AIRs) are proteins created by B and T cells that specifically recognize parts of foreign or self-antigens, such as viruses or cancerous cells, and mount an immune response to neutralize them[63,86]. AIRs are highly diverse (approximately 10^{15} different possible receptors[87–89]) and are specific to certain antigens. Their diversity arises from a stochastic process called V(D)J recombination that combines V, D and J gene segments with random insertions and deletions to create receptors able to detect various antigens an individual encounters in their lifetime[39,90,91] (panel **a**). The crucial region of AIRs for recognizing an antigen is the complementarity-determining region 3 (CDR3)[92,93]. On average, it is a 15-amino-acid-long part of the receptor with the highest variability. When examining antigen recognition and binding, it is often the only part of the receptor used in computational AIRR analyses[94].

AIRRs are sets of all AIRs present in an individual (panel **b**). There are approximately an estimated 10^8 unique receptors with frequency distribution specific to an individual[95–97] at any given time, with very few receptors specific to one antigen. Examining the AIRs in AIRRs provides unique insights into disease, rendering AIRRs a major target of current diagnostic biomarker research[5]. This task is challenging due to the low overlap of receptors between AIRRs of different individuals[98] and the unknown specificity of individual AIRs determined by complex sequence patterns. This inspires ML applications for AIRR-based diagnostics[34].



How confounders affect the analysis

Age, sex and genetic background influence the immune repertoire. Repertoire diversity decreases with age[37,45], and sex affects the usage of V genes in T cell receptor (TCR) repertoires[38]. Sex and age also influence the immune state through innate immunity[46,47], representing potential confounders in disease diagnostics (Box 1). Environment, broadly defined as a proxy for socioeconomic background and geographic location, may also be a confounder. Unlike age or sex, it is typically unobserved in AIRR studies (unmeasured or hidden confounder).

For predictive purposes, confounding is not generally problematic and might even improve the performance of the model[48,49]. However, the recovered biomarkers could reflect the confounders as much as the immune state[9]. If the aim is to gain insight into the underlying biological process (or estimate causal effects), confounding should be controlled for. Additionally, if the source and target populations differ, confounder distributions or functional relations may change, potentially changing the perceived (non-causal) relationship between the immune state and AIRRs.

Batch effects and timing of measurements

Batch effects are systematic biases connected to experimental protocols exhibiting different behaviour across conditions, with a certain level of bias always being present in the sequencing (molecular) data[50]. If batch effects are independent of the labels to be predicted, they will not introduce any bias in the learned ML models, leading only to increased variability of the biomarker. Batch effects are more problematic when correlated with the label (for example, immune state) when a predictive model may perform well in the study cohort by learning batch effect associations. Such a model would fail when applied to a population where the batch effect is differently associated with the disease. This is also of interest when multiple datasets need to be combined for a study. Batch effects in AIRRs[42–44,51] might manifest through sequencing errors, differences in gene usage between protocols, the sensitivity of detecting rare receptors, and skewed diversity capture[42].

The timing of measurement, that is, when the sequencing is performed in the course of the disease, also affects diagnostic development. If positive examples (diseased individuals) in the dataset are collected after individuals have received treatment, the collected AIRRs will not be representative of the AIRRs of individuals who will get tested for diagnosis. To mitigate this, the study cohort should be representative of the target population in terms of the timing of measurement or include individuals sequenced across the disease progression spectrum. Alternatively, different disease stages could be modelled separately, making the task a multiclass classification problem.

Selection bias and choosing participants for a study

Selection bias is defined slightly differently in causality and ML. In causality, selection bias is any statistical association resulting from selective inclusion into the study cohort (for example, participants are recruited based on some of the variables of interest in the analysis)[52]. For example, if only individuals with symptoms are tested for a diagnostic of a viral infection, the study cohort is not representative of the source population as it ignores asymptomatic individuals. Selection bias does not depend on the cohort size[53] and can introduce,

increase, decrease or even reverse the sign of existing associations[54]. Given a causal graph, selection bias may be present whenever the data-collection process depends on the cause and effect or the parents in the causal graph[55] (Fig. 2).

In ML, selection bias has a less structural definition: it occurs when there is a difference in the marginal distribution of any variable used for prediction (covariate shift[56]) between the study cohort and the source population, or a difference in the marginal distribution of the label (such as the disease status, resulting in label shift[57]). These definitions do not rely on causal graphs. However, considering the underlying causal models can improve ML analysis by specifying how to recover from biases under given assumptions.

Closely related to selection bias is the concept of transportability[29,58] (related to external validity; Box 3). Transportability bias occurs when moving from a source population to a distinct target population, where the target and source populations are at least partially non-overlapping[59]. An example might be an AIRR-based diagnostic built in Norway (source population) that needs to be applied in Serbia (a target population that might arbitrarily differ in the marginal distribution of variables, such as HLA or age), where the causal mechanisms of the disease (the influence of the pathogen on AIRRs given HLA and age) remain stable.

BOX 3

Internal and external validity for ML classification

Internal validity. In causal inference, a conclusion of a scientific study is said to be internally valid if it is true (statistically correct) of the population on which the study was conducted[19]. In ML, we define internal (in-distribution) validity as based on learning the source population distribution instead of noise, and assessed by cross-validation, a leave-out test dataset or bootstrapping[71]. Failure to comply with this requirement leads to a model that is overfitted to the data and is overly optimistic. The correct procedure ensures that the obtained performance estimate reflects how the model is expected to behave when applied to the new data from the same distribution. In the traditional ML literature, the performance on new data independently sampled from the same distribution is called generalization. Previous work provides recommendations for best practices in ML for biology and medicine[71,99].

External validity. External validity, in causal inference, is defined as the ability to generalize results to new environments or populations[58,59]. The ML field typically examines external validity in the context of domain adaptation[12], domain generalization[13,19,100] and out-of-distribution generalization[15]. External validity is often the main aim of scientific analyses, and it is typically achieved by discovering invariant mechanisms across populations.

HLA can take on alternative causal roles

Depending on the disease, biological variables can have different roles in the causal graph, with important implications for the analysis. For example, HLA molecules present peptides derived from pathogens to T cells, and thus shape the T cell repertoire. HLA also influences the TCR repertoire during positive and negative selection during T cell maturation in the thymus[60]. HLA can therefore affect the TCR repertoire composition of both naive and antigen-experienced T cells, making HLA an important variable in diagnostic development.

Depending on the assumptions of the causal model describing the disease, the role of HLA in the analysis will differ. For viral infections, HLA is considered a precision covariate (Fig. 3a)—it will influence the AIRR but not the immune state. Theoretically, adjusting for it will not resolve any bias, but it might improve the precision of the diagnostic. In practice, this might depend on the amount of data available for different HLAs.

HLA might be a confounder in AIRR-mediated autoimmunity, where AIRR causes the immune state (Fig. 3b). Additionally, HLA can act as a moderator (Box 1), for example, in coeliac disease. Coeliac disease is an autoimmune condition occurring due to gluten consumption. For the disease to occur, the subjects have to both carry specific HLA allotypes (HLA-DQ2 or HLA-DQ8) and have gluten-specific TCRs[61]. Therefore, exploiting HLA information might be necessary to develop a diagnostic for this disease.

In some cancers, tumour cells have somatic mutations that affect peptides binding to HLA and help tumour cells evade immune recognition[62]. In this case, HLA acts as a mediator between disease and the AIRR (Fig. 3c) and, as such, does not need to be adjusted for in the analysis when developing a diagnostic.

High dimensionality of AIRR data

Building diagnostics based on AIRRs (or other molecular data) is made more challenging by their high dimensionality. In this Perspective, we have represented AIRRs by a single node (variable) in the causal graph, which then represents millions of individual AIRs.

AIRRs consist of a large number of sequences that are mostly non-overlapping between individuals, with very few of them specific to any one disease[63] (Box 2). Some approaches represent AIRRs by their observed sequences, physicochemical properties or summary statistics. Alternatively, it is possible to learn an AIRR representation. As the sequence specificities are primarily unknown, self-supervised representation learning methods might be the most suitable: pretraining methods[64], fitting generative models that learn the data distribution with latent variables being used in downstream tasks[65], or imposing constraints on the learned representation space via alternative labels or training tasks[66–68]. To ensure the robustness of representations learned in this manner, the data may come from multiple distributions[25,69,70], or algorithms that try to infer latent causal variables may also be used[27]. However, the interpretability of the learned representations and their relations to the causal model remain a challenge[20].

Although causality for ML in the high-dimensional setting of medical imaging[8] might have some parallels with molecular datasets, the imaging causal models differ substantially from molecular biomarkers, posing the question of exactly how the different biases discussed earlier manifest in high-dimensional sequencing data.

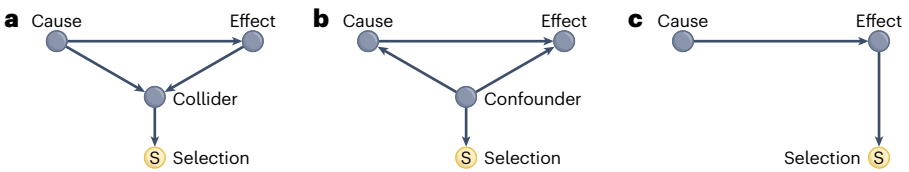


Fig. 2 | Examples of selection bias in causal models. a–**c**, Selection bias may occur by selecting based on a collider (**a**), because a confounder influences selection (**b**) or by selecting based on the effect variable (**c**).

https://doi.org/10.1038/s42256-023-00781-8

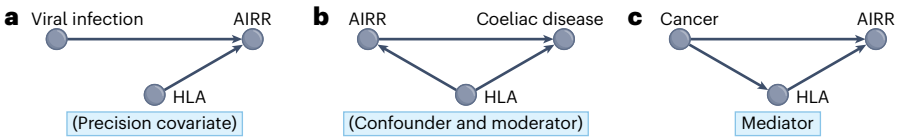


Fig. 3 | The different causal roles HLA can take for different types of immune-related disease. a, In a viral infection, HLA is a precision covariate. **b**, In coeliac disease, HLA is both a confounder and a moderator. **c**, In cancer, HLA can be a mediator.

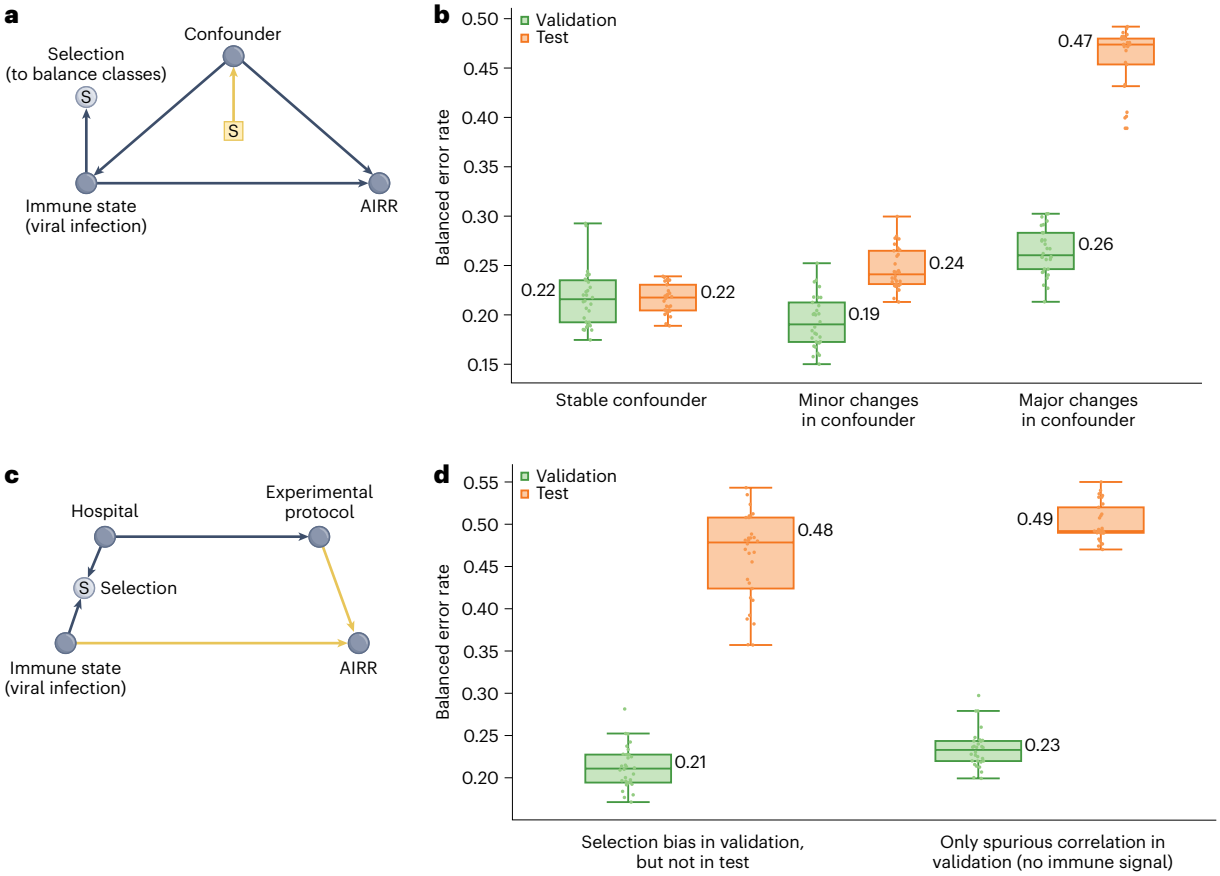


Fig. 4 | Experiments showing immune-state prediction performance under different causal models. a–**d**, Results are shown for 500 AIRRs for training and validation and 500 for testing, with 500 TCRβs per repertoire across 30 repetitions. The median values of the balanced error rate are shown in the plots. The immune signal indicative of the immune state consisted of 3-mers implanted approximately in the middle of the receptor sequence. For scenarios **b**,**d**, 3-mer frequencies and logistic regression were used. **a**, Causal model for experiment 1. The yellow arrow denotes that the confounder distribution changes between the training (validation) and test population. The immune state is balanced in both populations. **b**, The balanced error rate when the training (validation) and test sets originate from the same distribution (the confounder distribution is stable), when the distribution changes slightly ($P(\text{confounder})_{\text{validation}} = 0.6$, $P(\text{confounder})_{\text{test}} = 0.8$), and when the confounder distribution substantially changes ($P(\text{confounder})_{\text{validation}} = 0.8$, $P(\text{confounder})_{\text{test}} = 0.01$). **c**, The causal model for experiment 2. The edges in yellow denote the relations modified in the experiment. **d**, The balanced error rate when the selection bias is present in the training population but not in the test population.

So far, we have used AIRRs to refer to both single- and paired-chain TCR and B cell receptor repertoires, although often only single-chain receptors are sequenced. The causal model could be extended to include paired-chain TCR and B cell receptor repertoire data as separate variables depending on the research question. Each of these receptor populations might be further split for causal modelling to allow for complex interactions between different cell subtypes or localizations.

Simulation study for AIRR diagnostics study design

To illustrate the influence of different variables in the causal model on the performance of ML algorithms for diagnostics, we performed three simulation experiments where we systematically varied the causal parameters. In the first experiment, we trained a model to predict the immune state based on AIRRs without taking confounders (for example, age or sex) into account. We showed that, with a changing confounder distribution and keeping the classes balanced, the performance (measured by balanced error rate) might drop substantially (Fig. 4a,b). In the second experiment, we showed how selection bias may lead to poor performance on an independent target population due to spurious correlations (Fig. 4c,d). In the third experiment, we contrasted the handling of batch effects for the AIR setting against a different molecular biomarker where batch effect handling was more established. We showed that batch effects might lead to higher error rates and result in classifiers learning spurious signals, especially in AIRR settings (Fig. 5). A description of the experiments is provided in the Supplementary Information.

https://doi.org/10.1038/s42256-023-00781-8

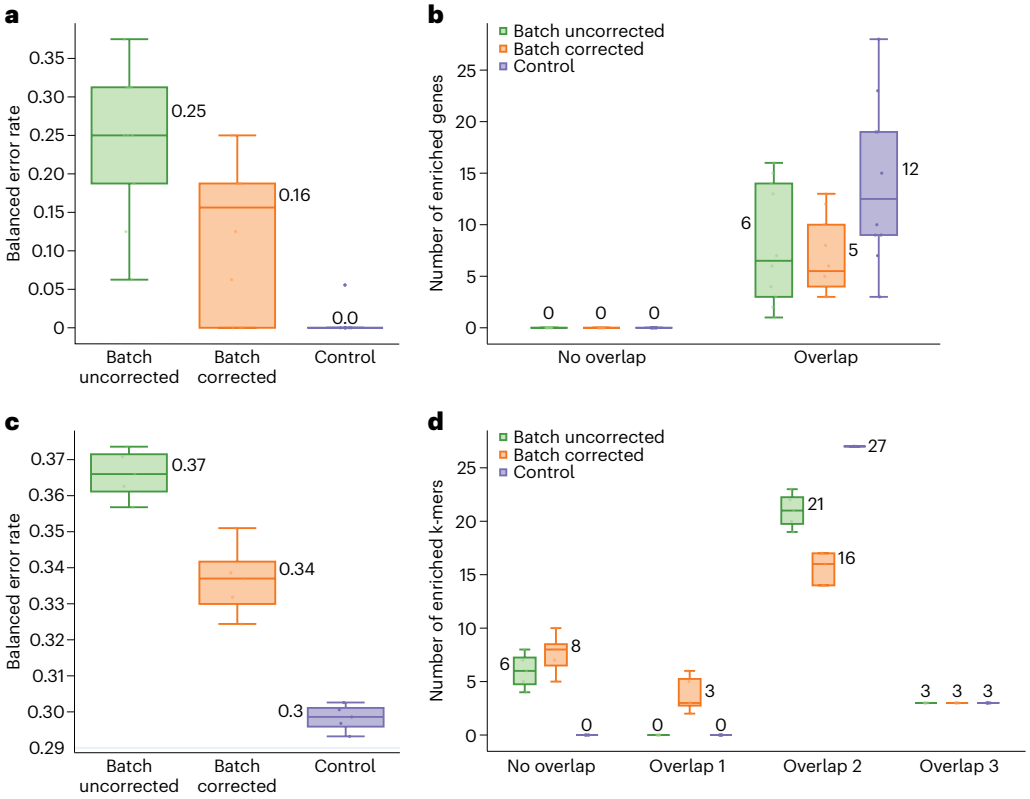


Fig. 5 | Batch effects may lead to higher error rates in transcriptomic and AIR settings and might result in classifiers learning spurious signals, especially in the AIR setting. a–**d**, Results are shown for three scenarios: batch effects present but uncorrected, corrected batch effects using linear regression on *k*-mer frequencies, and no batch effects in the data (control). Median values are shown throughout. **a**, Transcriptomic classification performance, where uncorrected batch effects lead to the highest balanced error rate. **b**, The number of enriched genes detected by the L1-regularized logistic regression model and their overlap with the simulated biological signal. **c**, AIR classification performance using *k*-mer frequencies for data representation and logistic regression, where uncorrected batch effects lead to the highest error rate, although the difference is small. **d**, The recovered 3-mers from the logistic regression model are grouped by how much they overlap with the 3-mers of the immune signal across the three scenarios. The recovered 3-mers from the model are obtained as the features corresponding to the 30 largest coefficients in the L1-regularized logistic regression model, in absolute values.

Conclusion

Advice on AIRR study design and computational processing

We propose the following guidelines to learn AIRR-based biomarkers that generalize well to clinical settings. (1) Ensure that batch effects, although nearly always present, only influence the observed AIRRs and are not correlated with the immune state. Avoid systematically different protocols and recruitment periods across different labels. (2) Internal validity, occurring when the targeted probability distribution is learned instead of noise (Box 3), has to be achieved through appropriate assessment strategies and sufficiently large study cohorts[71]. Total cohort sizes are not the main focus of this Perspective because small cohorts would only increase the variability of estimates but not introduce bias per se. However, an insufficient number of participants makes it hard to achieve internal validity and, for domain adaptation or transfer learning, to determine whether there are true systematic differences between settings. Additionally, recruiting a sufficient number of participants for each confounder value group is necessary. (3) Avoid selection biases that may introduce spurious associations when recruiting study participants. One exception to avoiding selection biases is when they are deliberately introduced (and compensated for) to enrich signals for ML, for example, by balancing the classes when training a prediction model. Furthermore, in the case where the target population is known to differ from the source (training) population in a variable that has a major influence on the AIRRs or immune state, we advise considering techniques such as pretraining that might help with dataset shifts[72], using data from multiple environments, if available, to obtain more robust representations[25,69,70], and exploring importance weighting[73]. We illustrate how failing to follow these recommendations might influence the prediction task in worked examples (Fig. 4).

Proposed reporting standards for AIRR diagnostic study design

We propose the following reporting standards to increase the trustworthiness of AIRR diagnostic studies and ensure their applicability in future use cases, such as meta-analyses where multiple studies are examined together to answer the research question better. (1) Report the sets of AIRR samples that have been processed together in batches. (2) For each AIRR, provide information on recruitment source, experimental protocol and institution. (3) If external validity is anticipated, define the target (deployment) setting where the diagnostic could be applied. (4) Report metadata, including sex, age, HLA and similar properties outlined by the MiAIRR standard[74,75]. Results per strata should be provided for any variable considered to have a major impact on AIRR and immune state (consult the state of the art in the AIRR field and disease field at the time of publication). Include information on genetic ancestry and aim to cover diverse patient cohorts[76,77]. Additionally, reviewing study protocols in advance, for example, through Registered Reports[78], may alleviate some of the concerns described previously.

Suggested research directions for the AIRR field

One major open question is how the HLA influences AIRRs[5,34,41,79]. Strong correlations between HLA and the CDR3 regions of TCRs have recently been observed, indicating that HLA risk allotypes might increase the frequency of autoreactive TCRs already during T cell development[41].

From a diagnostic perspective, the HLA influence can be seen as two sub-questions: (1) the degree to which HLA leaves a detectable mark in the overall AIRR that can be leveraged to capture the disease-predictive information of HLA by AIRR sequencing alone (leveraging the indirect path AIRR ← HLA → disease), and (2) the degree to which HLA moderates the direct AIRR–disease relation so that ML models need to learn distinct predictive patterns for individuals with different HLAs.

So far, we have considered diagnostics development as a binary classification problem. However, it could be extended to consider multiple classes illustrative of disease stages for a single disease or multiple diseases[80]. One way to handle multiple disease stages might be to estimate a ML model in a one-versus-all fashion (with possibly shared representation of sequencing data), thus allowing distinct features to be learned as relevant for each disease stage. Multiple diseases could also interact, making the analysis more challenging[81]. Interactions could lead to structural causal models with cycles[82]. Finally, in dynamic treatment regime settings[83], biomarkers could support adaptive treatment decisions through multiple stages of disease progression for individual patients.

Although we argue that causality is important for ML robustness and diagnostic development study design, causality is also an aim in itself in terms of describing biological mechanisms[84]. Establishing causal AIRR models would enable improvement of AIRR-based diagnostics and allow for causal interpretations and estimations of the effects of interventions. For example, a sufficiently detailed AIRR model and the methodology to successfully handle high-dimensional data may allow computational screening of new candidate therapies and vaccination procedures.

Data availability

All data and results for the analysis presented in the manuscript are openly available on Zenodo at https://zenodo.org/record/7756163 (experiment 1), https://zenodo.org/record/7752837 (experiment 2), https://zenodo.org/record/7752115 (experiment 3, AIRR setting) and https://zenodo.org/record/7727894 (experiment 3, transcriptomic setting).

Code availability

The source code for the experiments is openly available on GitHub at https://github.com/uio-bmi/causalairr.

References

1. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
2. Locke, W. J. et al. DNA methylation cancer biomarkers: translation to the clinic. *Front. Genet.* **10**, 1150 (2019).
3. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271 (2016).
4. Huang, K., Wu, L. & Yang, Y. Gut microbiota: an emerging biological diagnostic and treatment approach for gastrointestinal diseases. *JGH Open* **5**, 973–975 (2021).
5. Arnaout, R. A. et al. The future of blood testing is the immunome. *Front. Immunol* **12**, 626793 (2021).
6. Strimbu, K. & Tavel, J. A. What are biomarkers? *Curr. Opin. HIV AIDS* **5**, 463–466 (2010).
7. Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality and shift-stable models in health AI. *Biostatistics* **21**, 345–352 (2020).
8. Castro, D. C., Walker, I. & Glocker, B. Causality matters in medical imaging. *Nat. Commun.* **11**, 3673 (2020).
9. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.* **23**, 169–181 (2021).
10. Dockès, J., Varoquaux, G. & Poline, J.-B. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience.* **10**, giab055 (2021).
11. Daumé, H. & Marcu, D. Domain adaptation for statistical classifiers. *J. Artif. Intell. Res.* **26**, 101–126 (2006).
12. Kouw, W. M. & Loog, M. A review of domain adaptation without target labels. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 766–785 (2021).
13. Wang, J. et al. Generalizing to unseen domains: a survey on domain generalization. *IEEE Trans. Knowl. Data Eng.* **35**, 8052–8072 (2023).
14. Gulrajani, I. & Lopez-Paz, D. In search of lost domain generalization. Preprint at https://arxiv.org/abs/2007.01434 (2020).
15. Liu, J. et al. Towards out-of-distribution generalization: a survey. Preprint at https://doi.org/10.48550/arXiv.2108.13624 (2023).
16. Pearl, J. *Causality* (Cambridge Univ. Press, 2009); https://doi.org/10.1017/CBO9780511803161
17. Peters, J., Janzing, D. & Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms* (MIT Press, 2017).
18. Hernán, M. & Robins, J. *Causal Inference: What If* (Chapman & Hall/CRC, 2020).
19. Rothenhäusler, D. & Bühlmann, P. Distributionally robust and generalizable inference. *Statist. Sci.* **38**, 527–542 (2023).
20. Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J. & Silva, R. Causal machine learning: a survey and open problems. Preprint at https://doi.org/10.48550/arXiv.2206.15475 (2022).
21. Heinze-Deml, C., Maathuis, M. H. & Meinshausen, N. Causal structure learning. *Annu. Rev. Stat. Appl.* **5**, 371–391 (2018).
22. Squires, C. & Uhler, C. Causal structure learning: a combinatorial perspective. *Found. Comput. Math.* https://doi.org/10.1007/s10208-022-09581-9 (2022).
23. Peters, J., Bühlmann, P. & Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. B Stat. Methodol.* **78**, 947–1012 (2016).
24. Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. Invariant risk minimization. Preprint at https://doi.org/10.48550/arXiv.1907.02893 (2020).
25. Jiang, Y. & Veitch, V. Invariant and transportable representations for anti-causal domain shifts. *Adv. Neural Inf. Process Syst.* **35**, 20782–20794 (2022).
26. Magliacane, S. et al. Domain adaptation by using causal inference to predict invariant conditional distributions. *Adv. Neural Inf. Process Syst.* **31**, 10846–10856 (2018).
27. Schölkopf, B. et al. Toward causal representation learning. *Proc. IEEE* **109**, 612–634 (2021).
28. Cui, P. & Athey, S. Stable learning establishes some common ground between causal inference and machine learning. *Nat. Mach. Intell.* **4**, 110–115 (2022).
29. Bareinboim, E. & Pearl, J. Causal inference and the data-fusion problem. *Proc. Natl Acad. Sci. USA* **113**, 7345–7352 (2016).
30. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun.* **11**, 3923 (2020).
31. Prosperi, M. et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat. Mach. Intell.* **2**, 369–375 (2020).
32. Raita, Y., Camargo, C. A., Liang, L. & Hasegawa, K. Big data, data science and causal inference: a primer for clinicians. *Front. Med.* **8**, 678047 (2021).
33. Schölkopf, B. et al. On causal and anticausal learning. In *Proc. 29th International Conference on Machine Learning* 459–466 (Omnipress, 2012).
34. Greiff, V., Yaari, G. & Cowell, L. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Curr. Opin. Syst. Biol.* https://doi.org/10.1016/j.coisb.2020.10.010 (2020).

35. Emerson, R. O. et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017).
36. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
37. Britanova, O. V. et al. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* **192**, 2689–2698 (2014).
38. Schneider-Hohendorf, T. et al. Sex bias in MHC I-associated shaping of the adaptive immune system. *Proc. Natl Acad. Sci. USA* **115**, 2168–2173 (2018).
39. Slabodkin, A. et al. Individualized VDJ recombination predisposes the available Ig sequence space. *Genome Res.* **31**, 2209–2224 (2021).
40. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).
41. Ishigaki, K. et al. HLA autoimmune risk alleles restrict the hypervariable region of T cell receptors. *Nat. Genet.* **54**, 393–402 (2022).
42. Barennes, P. et al. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nat. Biotechnol.* **39**, 236–245 (2021).
43. Trück, J. et al. Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling. *eLife* **10**, e66274 (2021).
44. Smirnova, A. O. et al. The use of non-functional clonotypes as a natural calibrator for quantitative bias correction in adaptive immune receptor repertoire profiling. *eLife* **12**, e69157 (2023).
45. Krishna, C., Chowell, D., Gönen, M., Elhanati, Y. & Chan, T. A. Genetic and environmental determinants of human TCR repertoire diversity. *Immun. Ageing* **17**, 26 (2020).
46. Klein, S. L. & Flanagan, K. L. Sex differences in immune responses. *Nat. Rev. Immunol.* **16**, 626–638 (2016).
47. Castelo-Branco, C. & Soveral, I. The immune system and aging: a review. *Gynecol. Endocrinol.* **30**, 16–22 (2014).
48. Hernán, M. A., Hsu, J. & Healy, B. A second chance to get causal inference right: a classification of data science tasks. *Chance* **32**, 42–49 (2019).
49. Blaas, A., Miller, A., Zappella, L., Jacobsen, J.-H. & Heinze-Deml, C. Considerations for distribution shift robustness in health. In *Proc. Machine Learning for Healthcare Workshop* (ICLR, 2023).
50. Leek, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
51. Bonaguro, L. et al. A guide to systems-level immunomics. *Nat. Immunol.* **23**, 1412–1423 (2022).
52. Bareinboim, E. & Pearl, J. Controlling selection bias in causal inference. In *Proc. 15th International Conference on Artificial Intelligence and Statistics* Vol. 22 (eds Lawrence, N. et al.), 100–108 (PMLR, 2012).
53. Correa, J., Tian, J. & Bareinboim, E. Generalized adjustment under confounding and selection biases. In *Proc. 32nd AAAI Conference on Artificial Intelligence* Vol. 32, 6335–6342 (AAAI, 2018).
54. Laubach, Z. M., Murray, E. J., Hoke, K. L., Safran, R. J. & Perng, W. A biologist's guide to model selection and causal inference. *Proc. R. Soc. B Biol. Sci.* **288**, 20202815 (2021).
55. Hernán, M. A., Hernández-Díaz, S. & Robins, J. M. A structural approach to selection bias. *Epidemiology* **15**, 615–625 (2004).
56. Zhang, K., Schölkopf, B., Muandet, K. & Wang, Z. Domain adaptation under target and conditional shift. In *Proc. International Conference on Machine Learning* 28 (eds Dasgupta, S. et al.) 819–827 (PMLR, 2013).
57. Garg, S., Wu, Y., Balakrishnan, S. & Lipton, Z. C. A unified view of label shift estimation. *Adv. Neural Inf. Proc. Syst.* **33**, 3290–3300 (2020).
58. Pearl, J. & Bareinboim, E. External validity: from *Do*-calculus to transportability across populations. *Stat. Sci.* **29**, 579–595 (2014).
59. Degtiar, I. & Rose, S. A review of generalizability and transportability. *Annu. Rev. Stat. Appl.* **10**, 501–524 (2023).
60. Sharon, E. et al. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat. Genet.* **48**, 995–1002 (2016).
61. Jabri, B. & Sollid, L. M. T cells in Celiac disease. *J. Immunol.* **198**, 3005–3014 (2017).
62. Schaafsma, E., Fugle, C. M., Wang, X. & Cheng, C. Pan-cancer association of HLA gene expression with cancer prognosis and immunotherapy efficacy. *Br. J. Cancer* **125**, 422–432 (2021).
63. Rappazzo, C. G. et al. Defining and studying B cell receptor and TCR interactions. *J. Immunol.* **211**, 311–322 (2023).
64. Hendrycks, D., Lee, K. & Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *Proc. 36th International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 2712–2721 (PMLR, 2019).
65. Pradier, M. F. et al. AIRIVA: a deep generative model of adaptive immune repertoires. Preprint at https://doi.org/10.48550/arXiv.2304.13737 (2023).
66. Gao, Y. et al. Pan-Peptide meta learning for T-cell receptor–antigen binding recognition. *Nat. Mach. Intell.* **5**, 236–249 (2023).
67. Ostrovsky-Berman, M., Frankel, B., Polak, P. & Yaari, G. Immune2vec: embedding B/T cell receptor sequences in RN using natural language processing. *Front. Immunol.* **12**, 680687 (2021).
68. Fang, Y., Liu, X. & Liu, H. Attention-aware contrastive learning for predicting T cell receptor–antigen binding specificity. *Brief. Bioinform.* **23**, bbac378 (2022).
69. Gupta, G., Kapila, R., Gupta, K. & Raskar, R. Domain generalization in robust invariant representation. Preprint at https://doi.org/10.48550/arXiv.2304.03431 (2023).
70. Zhang, J. & Bottou, L. Learning useful representations for shifting tasks and distributions. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A et al.), 40830–40850 (PMLR, 2023).
71. Walsh, I. et al. DOME: recommendations for supervised machine learning validation in biology. *Nat. Methods* **18**, 1122–1127 (2021).
72. Wiles, O. et al. A fine-grained analysis on distribution shift. Preprint at https://arxiv.org/abs/2110.11328 (2021).
73. Byrd, J. & Lipton, Z. What is the effect of importance weighting in deep learning? In *Proc. 36th International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 872–881 (PMLR, 2019).
74. Rubelt, F. et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat. Immunol.* **18**, 1274–1278 (2017).
75. Vander Heiden, J. A. et al. AIRR community standardized representations for annotated immune repertoires. *Front. Immunol.* **9**, 2206 (2018).
76. Peng, K. et al. Diversity in immunogenomics: the value and the challenge. *Nat. Methods* **18**, 588–591 (2021).
77. Huang, Y.-N. et al. Ancestral diversity is limited in published T cell receptor sequencing studies. *Immunity* **54**, 2177–2179 (2021).
78. *Registered Reports* (Center for Open Science); https://www.cos.io/initiatives/registered-reports
79. DeWitt, W. S. III et al. Human T cell receptor occurrence patterns encode immune history, genetic background and receptor specificity. *eLife* **7**, e38358 (2018).
80. Zaslavsky, M. E. et al. Disease diagnostics using machine learning of immune receptors. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.26.489314 (2023).
81. Langenberg, C., Hingorani, A. D. & Whitty, C. J. M. Biological and functional multimorbidity—from mechanisms to management. *Nat. Med.* **29**, 1649–1657 (2023).

82. Bongers, S., Forré, P., Peters, J. & Mooij, J. M. Foundations of structural causal models with cycles and latent variables. *Ann. Stat.* **49**, 2885–2915 (2021).
83. Chakraborty, B. & Murphy, S. A. Dynamic treatment regimes. *Annu. Rev. Stat. Appl.* **1**, 447–464 (2014).
84. Bizzarri, M. et al. A call for a better understanding of causation in cell biology. *Nat. Rev. Mol. Cell Biol.* **20**, 261–262 (2019).
85. Baron, R. M. & Kenny, D. A. The moderator–mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations. *J. Pers. Soc. Psychol.* **51**, 1173–1182 (1986).
86. Greiff, V., Miho, E., Menzel, U. & Reddy, S. T. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.* **36**, 738–749 (2015).
87. Nikolich-Žugich, J., Slifka, M. K. & Messaoudi, I. The many important facets of T-cell repertoire diversity. *Nat. Rev. Immunol.* **4**, 123–132 (2004).
88. Zarnitsyna, V., Evavold, B., Schoettle, L., Blattman, J. & Antia, R. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Front. Immunol.* **4**, 485 (2013).
89. Murugan, A., Mora, T., Walczak, A. M. & Callan, C. G. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl Acad. Sci. USA* **109**, 16161–16166 (2012).
90. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302**, 575–581 (1983).
91. Weinstein, J. A., Jiang, N., White, R. A., Fisher, D. S. & Quake, S. R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807–810 (2009).
92. Xu, J. L. & Davis, M. M. Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity* **13**, 37–45 (2000).
93. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
94. Brown, A. J. et al. Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. *Mol. Syst. Des. Eng.* **4**, 701–736 (2019).
95. Qi, Q. et al. Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl Acad. Sci. USA* **111**, 13139–13144 (2014).
96. Elhanati, Y. et al. Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**, 20140243 (2015).
97. Greiff, V. et al. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* **7**, 49 (2015).
98. Elhanati, Y., Sethna, Z., Callan, C. G. Jr, Mora, T. & Walczak, A. M. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol. Rev.* **284**, 167–179 (2018).
99. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *Npj Digit. Med.* **5**, 48 (2022).
100. Ben-David, S. et al. A theory of learning from different domains. *Mach. Learn.* **79**, 151–175 (2010).

Acknowledgements

We acknowledge generous support by The Leona M. and Harry B. Helmsley Charitable Trust (grant no. 2019PG-T1D011 to V.G.), the UiO World-Leading Research Community (to V.G. and L.M.S.), the UiO:LifeScience Convergence Environment Immunolingo (to V.G. and G.K.S.), the UiO:LifeScience Convergence Environment RealArt (to G.K.S. and C.K.), EU Horizon 2020 iReceptorplus (grant no. 825821 to V.G. and L.M.S.), a Research Council of Norway FRIPRO project (grant no. 300740 to V.G.), a Norwegian Cancer Society Grant (215817 to V.G.), a Research Council of Norway IKTPLUSS project (grant no. 311341 to V.G. and G.K.S.) and Stiftelsen Kristian Gerhard Jebsen (K.G. Jebsen Coeliac Disease Research Centre, to L.M.S. and G.K.S.).

Author contributions

M.P., V.G. and G.K.S. conceived the study. M.P., G.S.A.H. and C.K. performed the experiments. J.P., M.E.W., L.M.S., C.K. and G.S.A.H. provided critical feedback. M.P., V.G. and G.K.S. drafted the manuscript. G.K.S. supervised the project. All authors read and approved the final manuscript and are personally accountable for its content.

Competing interests

V.G. declares advisory board positions in aiNET GmbH, Enpicom BV, Absci, Omniscope and Diagonal Therapeutics. V.G. is a consultant for Adaptyv Biosystems, Specifica Inc., Roche/Genentech, immunai and LabGenius. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00781-8.

Correspondence should be addressed to Milena Pavlović or Geir K. Sandve.

Peer review information *Nature Machine Intelligence* thanks Anna Susmelj and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.