Restricting datasets to classifiable samples augments discovery of immune disease markers

Gunther Glehr
UKR https://orcid.org/0000-0002-1495-9162
Paloma Riquelme
University Hospital Regensburg
Katharina Kronenberg
University Hospital Regensburg
Robert Lohmayer
Leibniz Institute for Immunotherapy
Victor Lopez-Madrona
Institute de Neurosciences des systèmes
Michael Kapinsky
Beckman Coulter GmbH
Hans Schlitt
University of Regensburg https://orcid.org/0000-0002-3874-0296
Edward Geissler
University Hospital Regensburg
Rainer Spang
University of Regensburg
Sebastian Haferkamp
University Hospital Regensburg
James Hutchinson (james.hutchinson@klinik.uni-regensburg.de)
UKR https://orcid.org/0000-0002-1199-4210

Article

Keywords:

Posted Date: June 6th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2921819/v1

Additional Declarations: There is **NO** Competing Interest.

- Original Article –

Restricting datasets to classifiable samples augments discovery of immune disease markers

Gunther Glehr,[1] Paloma Riquelme,[1] Katharina Kronenberg,[1] Robert Lohmayer,[2] Víctor J. López-Madrona,[3] Michael Kapinsky,[4] Hans J. Schlitt,[1] Edward K. Geissler,[1] Rainer Spang,[5] Sebastian Haferkamp[6] and James A. Hutchinson[1†]

1. Department of Surgery, University Hospital Regensburg, Regensburg, Germany
2. Algorithmic Bioinformatics Research Group, Leibniz Institute for Immunotherapy, Regensburg, Germany
3. Aix Marseille Univ, INSERM, Institute of Systems Neuroscience, Marseille, France
4. Beckman Coulter Life Sciences GmbH, Krefeld, Germany
5. Department of Statistical Bioinformatics, University of Regensburg, Regensburg, Germany
6. Department of Dermatology, University Hospital Regensburg, Regensburg, Germany

† *Corresponding author*: Prof. Dr. James A. Hutchinson
Department of Surgery
University Hospital Regensburg
Franz-Josef-Strauß-Allee-11
93053 Regensburg, Germany
Tel: +49-941-944-6831
Email: james.hutchinson@ukr.de

Running title: Dataset restriction aids immune disease marker discovery.

Grant support: This work was supported by the Bristol-Myers Squibb (BMS) Immune Oncology Foundation (Award FA19-009) and the Coronavirus Research Fund of the Bavarian State Ministry for Science and Art.

Authorship: GG and JAH designed and performed the project, and wrote the manuscript. KK and PR performed experiments and analyzed data. RL, VJLM and RS gave expert computational and statistical advice. MK gave expert advice about flow cytometry. HJS and EKG gave critical feedback. SH provided clinical samples and information, and gave expert Dermatological Oncology opinion.

Non-standard abbreviations: Area under the ROC curve (AUC); cytomegalovirus (CMV); correct classification rate (CCR); double negatives (DN); double positives (DP); Dirichlet distribution with K categories and its concentration parameters α: $Dir(\alpha_1, \alpha_2, \ldots \alpha_K)$; false positive rate (FPR); glutamic oxaloacetic transaminase (GOT); glutamic pyruvic transaminase (GPT); immune checkpoint inhibitor (ICI); immune-related adverse events (irAE); Kullback-Leibler divergence (KL-divergence); ROC curve with a long tail towards bottom-left (left-skewed ROC curve); samples below restriction (marker^{LOW} samples); samples above restriction (marker^{HIGH} samples); median fluorescence intensity (MFI); Normal distribution with mean μ and variance σ^2 $\left(\mathcal{N}(\mu, \sigma^2)\right)$; negative predictive value (NPV); positive predictive value (PPV); partial area under the curve (pAUC); Receiver operating characteristic (ROC); ROC curve with a long tail towards top-right (right-skewed ROC curve); restricted AUC (rAUC); restricted standardized AUC (rzAUC); rzAUC of marker^{HIGH} samples ($rzAUC_{HIGH}$); rzAUC of marker^{LOW} samples ($rzAUC_{LOW}$); true positive rate (TPR); effector memory T cells (T_{EM}); effector memory T cells re-expressing CD45RA (T_{EMRA}); central memory T cells (T_{CM}); naïve T cells (T_{naive}); true positive proportion asymmetric (TPP-asymmetric); true negative proportion asymmetric (TNP-asymmetric)

ABSTRACT

Immunological diseases are typically heterogeneous in clinical presentation, severity and response to therapy. Markers of immune diseases often reflect this variability, especially compared to their regulated behavior in health. This leads to a common, unarticulated problem that frustrates marker discovery and interpretation: Unequal variance of immune disease marker expression between patient classes necessarily limits a marker's informative range. To solve this problem, we introduce dataset restriction, a procedure that splits datasets into classifiable and unclassifiable samples. Applied to synthetic flow cytometry data, restriction identified markers that were otherwise disregarded. In advanced melanoma, restriction found new markers of immune-related adverse event risk after immunotherapy and enabled multivariate models that accurately predicted immunotherapy-related hepatitis. Hence, dataset restriction augments discovery of immune disease markers, increases predictive certainty for classifiable samples and improves multivariate models incorporating markers with a limited informative range. This principle can be directly extended to any classification task.

INTRODUCTION

The immune system detects pathological challenges with exquisite sensitivity and specificity, which enables it to mount appropriate protective responses[1]. Widely distributed immune cell subsets are responsible for sensing pathogens, tissue injury and cellular stress through diverse receptor systems[2-4]. These disease-related signals are then amplified through humoral and cellular cascades that stimulate the migration, expansion and activation of particular effector cell populations[5]. By capturing information about the precise nature of the immune response, we can draw inferences about the triggering event, allowing us to develop diagnostic or prognostic models to guide personalized treatment decisions[6].

Flow cytometry is a sophisticated, fast and relatively inexpensive method for analyzing the properties of single cells from a uniform cell suspension[7]. In clinical practice, flow cytometry is commonly used to profile leucocyte subset distribution in patient blood samples, especially in the context of hematological malignancies and infectious diseases[8]. Modern cytometers simultaneously collect data about expression of multiple proteins in single cells, while also allowing us to interrogate many millions of cells from a single sample[9]. This enables accurate identification of narrowly defined cell subsets, including rare populations, as well as broadly surveying many leucocyte subsets[10]. This rich information is captured as a data matrix for each sample with an unordered number of rows corresponding to cells and a defined number of cell-associated features as columns[11].

Applications of flow cytometry in clinical diagnostics are growing rapidly[12]. Of special interest, recent reports claim that immunophenotyping of peripheral blood leucocytes can be used to predict immune-related adverse events (irAE) following immune checkpoint inhibitor (ICI)-therapy[13-16]. Combined treatment with anti-PD-1 (Nivolumab) and anti-CTLA-4 antibody

(Ipilimumab) is now first-line therapy for many patients with unresectable metastatic melanoma[17]. Its effectiveness is remarkable in terms of clinical response rates, progression-free survival and overall survival; however, immune-mediated complications, such as colitis or hepatitis, present a significant clinical concern[18]. Life-threatening reactions are uncommon[19], but they often require interruption or discontinuation of immunotherapy, and introduction of glucocorticoids or non-steroidal immunosuppressants[20]. Clinically applicable, robust markers to guide irAE prevention or treatment strategies in patients would be extremely useful[21].

Extracting reliable predictive information from flow cytometry measurements is difficult because disease-related changes are often small compared to typical biological and technical variations[22]. This is especially true when investigating systemic changes in peripheral blood samples for signals that reflect localized disease[23]. Consequently, we often rely upon computational methods to perceive small and multivariate, but consistent changes between patient samples[24]. Most current approaches entail identification of cell populations with clustering methods like FlowSOM[25], extracting sample-wise cell frequencies from each cluster and then comparing between samples to identify significantly differentially represented cell subsets[26]. Alternatively, some methods identify disease related changes at a single-cell level[27].

Compared to the tightly regulated homeostasis of health, immunological diseases are inherently more variable[28]. Generally speaking, it follows that immune disease-related markers are more variably expressed in disease than health[29,30]. As we show, this fundamental biological insight is important because overlapping marker expression with unequal variance between patient classes necessarily implies a range of marker values with no discriminatory potential. This problem is exaggerated when marker distributions with unequal variance substantially overlap between two patient classes, such as health and disease. Critically, we often find that disease-related differences in immunological markers are small in relative and absolute terms[31]. This

inconvenient and unintuitive property, which is typical of markers measured by flow cytometry, masks informative markers in discovery studies and limits their clinical utility[32].

In this report, we examine the problem of finding and interpreting disease markers with a restricted range of informative values from an immunologist's perspective. To do this, we must first disambiguate some key terms with different meanings for immunologists and computer scientists. Properties of single cells measured by flow cytometry, such as cell lineage-associated surface antigen expression, will be called "features." We reserve "marker" to mean a sample-related quantity, such as cell subset frequency, that signifies information relevant to sample classification, hence diagnoses. The distribution of marker values within a set of patient samples is described by its probability density function, or simply "density." Throughout this article, we present plots of densities that compare marker expression in patient subgroups; crucially, these should not be mistaken for histograms showing feature expression within samples.

We provide a computational method to optimally restrict markers to their informative range, which makes them easier to discover and interpret. The power of dataset restriction is demonstrated through its application to flow cytometry markers; in particular, T cell subset frequencies. For each marker, we calculate a restricted standardized AUC (rzAUC) for every marker value by splitting the sample set into $marker^{HIGH}$ and $marker^{LOW}$ parts. We define the optimal restriction according to the maximum absolute rzAUC of either the $marker^{HIGH}$ or $marker^{LOW}$ part. We then assign a permutation p-value to the optimal rzAUC. Finally, we leverage the adapted range of all restricted markers in a multivariate (random forest) model by forcing decision tree cuts within each informative range.

In essence, restriction identifies the informative range of a marker, which allows us to segregate datasets into classifiable and unclassifiable samples. Importantly, using information about the informative range of markers leads to superior multivariate models. We qualify our method

using realistically simulated flow cytometry data, then apply it to real T cell subset analyses to discover new markers of irAE risk in patients receiving immunotherapy for advanced melanoma. Using a restricted dataset, we were able to train and prospectively validate a multivariate model to predict immunotherapy-related hepatitis, which failed when using unrestricted data. Our computational methods can be directly applied to other types of data, not limited to transcriptomic, (epi-)genomic, proteomic, metabolomics or clinical information.

RESULTS

Two-class distributions resulting in skewed ROC curves

We begin by showing how the distribution of a discriminatory marker that differs in its expression between diseased (patients) and unaffected (controls) individuals results in skewed receiver operating characteristic (ROC) curves. ROC curves relate the true positive rate (TPR) and false positive rate (FPR) for a disease marker at every datapoint in a two-class classification problem. Area under the ROC curve (AUC) is often used as a measure of the discriminatory capacity of a disease marker[33]. Throughout this report, we illustrate distributions of marker expression within classes by plotting probability densities. Densitities are normalized to 1 within each class, so the appearance of these plots is independent of class size (see Supplementary Note 1). In the following sections, we consider hypothetical markers whose expression is normally distributed $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2.

Perfectly discriminatory markers result in concave ROC curves with an AUC = 1 (Fig 1a). For imperfect markers, where there is overlap between the distributions of a disease marker expression in patient and control populations, provided that variance is equal in both classes, the ROC curve is symmetric about the anti-diagonal with 1 > AUC > 0.5. In the hypothetical example, marker expression is normally distributed with equal variances in the patient $\mathcal{N}(6, 1)$ and control $\mathcal{N}(5, 1)$ populations, but mean expression is higher in patients (Fig 1b). Entirely uninformative markers result in straight diagonal ROC curves with an AUC = 0.5 (Fig 1c).

Interpreting the area under a ROC curve is more complicated when comparing overlapping marker distributions with unequal variances, which result in ROC curves skewed around the anti-diagonal. Our first hypothetical example of a skewed ROC curve shows that normally

distributed, overlapping marker distributions with a higher mean and variance in the patient population compared to controls leads to a right-skewed ROC curve that crosses the diagonal in a region corresponding to low marker expression values (Fig 1d). Indeed, it is generally true that normally distributed populations with different variances result in non-concave ROC curves that cross the diagonal[34]. To illustrate this point, we simulated 200 samples by drawing random values from Normal distributions to show how varying the mean and variance of marker expression in patient and control distributions affects the shape and AUC of ROC curves (Supplementary Video 1 and 2). In the context of clinical diagnostics, markers of immune diseases usually reflect a change between tightly-regulated homeostasis in health and a disturbed, higher-variability condition in disease. Coupled with the fact that disease-associated changes in cell subset frequencies in blood are typically small, it is perhaps unsurprising disease markers measured by flow cytometry frequently result in skewed ROC curves[13]. In support of this assertion, we present a real-world example of a right-skewed ROC curve with a low AUC (Fig 1e). Specifically, this example shows that erythrocyte counts were elevated in baseline blood samples from patients metastatic melanoma who responded to combined Ipilimumab plus Nivolumab (Ipi-Nivo) therapy compared to non-responders.

Left-skewed ROC curves arise when the negative population has a lower mean, but higher variance than the positive population (Fig 1f). We find a real-world example in the previously unreported association between $CD8^+$ γδ T cells and hepatitis risk after combined Ipi-Nivo therapy (Fig 1g and Supplementary Fig 2). In this case, the higher variance of the control population might be due to technical imprecision in quantifying a rare cell population, since the absolute number of $CD8^+$ γδ T cells in blood was only 25.6 ± 19.3 c/nl.

We next considered the case of a phenotypically heterogeneous positive population, which could reflect multiple aetiologies leading to a common clinical presentation, different stages of a disease that culminate in a common presentation or different treatment responses. In such scenarios, we expect a bimodal distribution of a disease marker in the positive population that leads to a skewed ROC curve (Fig 1h). We previously reported the identification of a subset of patients with advanced melanoma who developed hepatitis after Ipi-Nivo therapy, which was reliably predicted by CMV-associated expansion of $CD4^+$ T_{EM} cells prior to immunotherapy[35]. In our melanoma dataset, we show that using $CD4^+$ T_{EM} frequencies to predict hepatitis after immune checkpoint inhibitor (ICI) therapy leads to a right-skewed ROC curve (Fig 1i). We know from previous work that baseline $CD4^+$ T_{EM} expansion is only a useful marker of hepatitis risk in CMV-infected patients, who constituted just 47.3% of our study cohort; therefore, this is a biologically validated example of alternative immunopathologies contributing to a common pathological presentation that impacts marker performance.

These three hypothetical distributions, and their real-world counterparts, demonstrate an important concept in immune marker discovery – namely, that a disease marker may be highly informative over a restricted range of measured values, but will consistently misclassify samples with marker values outside that range. By extension, using AUC across the entire ROC curve to assess predictive performance leads us to disregard potentially informative markers. Clearly, we need a method of finding such markers and defining their valid ranges.

Dataset restriction is a new method to find disease markers

Disease markers that give rise to skewed ROC curves perform well in a subset of samples, which may belong to either the positive or negative class, but are only informative over a certain

range. This leads us to the idea that particular *samples* may be classifiable or unclassifiable according to any given disease marker. Here, we present and implement a method of marker discovery that relies upon restricting training datasets to classifiable samples[36]. In the given example, we compared the distributions of 2500 positive and 2500 negative simulated samples, in which 20% of positive and 2% of negative samples were drawn from a normal distribution $\mathcal{N}(9,1)$ and all other samples were drawn from $\mathcal{N}(6,1)$ (Fig 2a). This resulted in a right-skewed ROC curve for the complete dataset (Fig 2b). We first generated two ROC curves for every possible "restriction" of the dataset – explicitly, one for samples above the restriction (marker^{HIGH} samples, orange; Fig 2c, d-f) and one for samples beneath (marker^{LOW} samples, blue; Fig 2c, g-i). Marker^{HIGH} samples generally correspond to the bottom-left part of the complete ROC curve (Fig 2d). Considering the densities of only marker^{HIGH} samples (Fig 2e), the restricted ROC curve had a superior "restricted" AUC (rAUC) of 0.692 (Fig 2f). Marker^{LOW} samples generally correspond to the top-right part of the complete ROC curve, but here their densities overlapped substantially; therefore, the restricted ROC curve was close to diagonal (Fig 2g-i). Notably, restricted densities are not the same as those in Fig 2a but are instead re-calculated on either marker^{HIGH} or marker^{LOW} samples.

Standardizing each rAUC according to sample size gave the restricted standardized AUC (rzAUC). The maximum absolute value of rzAUC defined the optimal restriction value (Fig 2c). In our example, rzAUC was maximal at FPR = 0.258, which corresponded to an optimal marker restriction value of 6.8. Consequently, marker^{HIGH} samples should be kept and marker^{LOW} samples should be discarded – that is to say, marker^{HIGH} samples are classifiable, whereas marker^{LOW} samples are unclassifiable. Supplementary Video 3 helps to visualize the $rAUC$ for varying restrictions of the dataset.

In other situations, the positive class may have higher or lower marker values, potentially leading to an AUC < 0.5 and accordingly a negative rzAUC. In Supplementary Fig 3, we show that regardless which class is labelled positive or negative, our method identifies the same restriction value. In such cases, $marker^{HIGH}$ and $marker^{LOW}$ rzAUCs are mirrored, meaning the restriction at the optimal absolute rzAUC remains identical. Critically, regardless of marker distribution, because areas under ROC curves are independent of class size, it follows that restriction values are also independent of class size[37].

Restriction identifies classifiable samples in simulated datasets

To test our computational approach, we next applied it to our four preceding examples by simulating 100 samples from each class. In the first example, the negative class $\mathcal{N}(5,1)$ and positive class $\mathcal{N}(6,1)$ gave rise to a symmetrical ROC curve with a maximum rzAUC corresponding to FPR=1; consequently, the optimally informative dataset contained all samples (Fig 3a). In the second example, the negative class $\mathcal{N}(5,1)$ and positive class $\mathcal{N}(6,2)$ produced a right-skewed ROC curve because the variances were unequal (Fig 3b). We see that low marker values led to a consistent misclassification, indicated by the ROC curve crossing the diagonal. The maximum rzAUC of 5.8 for $marker^{HIGH}$ samples indicated that samples with a marker value < 4 must be discarded. In the third example, the negative class $\mathcal{N}(5,2)$ and positive class $\mathcal{N}(6,1)$ produced a left-skewed ROC curve (Fig 3c). Here, high marker values led to consistent misclassification; therefore, the ROC curve deviated below the diagonal. The maximum rzAUC of 5.8 for $marker^{LOW}$ samples indicated that samples with a marker value > 7 must be discarded. In the fourth example, we compared 100 samples from the negative class $\mathcal{N}(5,1)$ and a bimodal positive class consisting of 90 samples from the same distribution $\mathcal{N}(5,1)$, plus 10 samples from a distribution $\mathcal{N}(9,1)$ with a higher mean (Fig 3d). The resulting right-skewed ROC curve

reflected the fact that our simulated marker was only informative for higher sample values. Accordingly, the optimal rzAUC of 2.4 for marker^{HIGH} samples restricted our dataset to samples with a marker value ≥ 6.2. Evidently, our method is able to optimally restrict cleanly simulated patient populations, such that we retain only classifiable samples.

Synthesizing realistic flow cytometry datasets

Realistic synthetic data can be extremely valuable in machine-learning; for instance, for validating new analytical methods, calculating experimental sample sizes or data augmentation. Because no generative model already existed, we developed a new algorithm to create synthetic flow cytometry datasets, which differed from the preceding simulated examples in several key respects – specifically, they comprise multiple covarying markers, incorporate a realistic level of noise, and were adjusted in biologically meaningful ways. Our web-based interactive gating tree allows readers to synthesize their own flow cytometry data (Supplementary Website 1).

In order to validate our restriction method, we needed a way of imitating disease-related differences between groups of samples. In the method described above, any effect that changes the proportion of cells in any gate(s) equates to changing the Dirichlet distribution parameters. In the given example, the originally estimated mean proportions are projected onto the gating tree and corresponding Dirichlet distribution for three example leafs A, B and L. Here, the mean proportion of $CD8^+$ T_{EMRA} cells is 7.17 % (Fig 4a). Now, instead of determining the number of cells in each leaf gate according to the originally estimated distribution, we generate synthetic cells from a modified Dirichlet distribution in which the mean proportion of $CD8^+$ T_{EMRA} cells was arbitrarily changed to 33.23 % (Fig 4b). Using our method, changing the proportion of cells in any gate leads to changes in the proportion of cells in all other gates, which we represent by

the different intensities of red shading in the gating trees and the different Dirichlet distribution for the three example leafs A, B and L. Three examples of gating generated with a mean proportion $CD8^+$ T_{EMRA} cells = 33.23 % are provided (Supplementary Fig 4).

Applying restriction to realistic synthesized flow cytometry datasets

We next applied our restriction method to synthetic flow cytometry datasets that incorporated estimated technical and biological noise typical of real-world measurements. Specifically, we generated synthetic samples that gave rise to marker distributions similar to the preceding simulated examples. Artificial disease-associations were introduced by changing the frequency of $CD4^+$ T_{EM} cells, which had a baseline mean proportion of 7.7 % among healthy donors. We subsequently extracted $CD4^+$ T_{EM} cell frequencies relative to $CD3^+$ T cells from all samples by applying our standard gating strategy, then applied our restriction method. Similar to Figure 3, we simulated marker values from normal distributions. We then generated synthetic flow cytometry datasets by setting the $CD4^+$ T_{EM} cell Dirichlet parameter to each simulated marker value.

In the first example, the negative class $\mathcal{N}(7.7, 1)$ and positive class $\mathcal{N}(10.7, 1)$ gave rise to a symmetrical ROC curve (Fig 5a). As expected, the results were much noisier than those shown in Figure 3; nevertheless, the maximum rzAUC = 7.2 corresponded to FPR = 1, so the optimally informative dataset contained all samples. In the second example, the negative class $\mathcal{N}(7.7, 1)$ and positive class $\mathcal{N}(8.7, 3)$ gave rise to a right-skewed ROC curve (Fig 5b). The maximum rzAUC = 4.3 led us to retain $marker^{HIGH}$ samples with $\geq 5.94\%$ $CD4^+$ T_{EM} cells. In the third example, the negative class $\mathcal{N}(7.7, 3)$ and positive class $\mathcal{N}(8.7, 1)$ gave rise to a left-skewed ROC curve (Fig 5c). The maximum rzAUC = 4.6 led to a restriction of the dataset to

311 marker^{LOW} samples with < 8.37% CD4$^+$ T_{EM}. In the fourth example, we compared the negative
312 class $\mathcal{N}(7.7, 1)$ and a bimodal positive class comprising 80 samples showing no effect
313 $\mathcal{N}(7.7, 1)$ plus 20 samples from a distribution $\mathcal{N}(16.7, 1)$ with a higher mean (Fig 5d). The
314 resulting right-skewed ROC curve with a maximum rzAUC = 3.7 led us to keep marker^{HIGH}
315 samples with $\geq$ 8.24% CD4$^+$ T_{EM}. Hence, our method is able to appropriately restrict
316 realistically synthesized flow cytometry datasets for symmetric or skewed ROC curves, such
317 that we retain only classifiable samples.
318
319 **Restriction method improves findability in realistic synthesized datasets**
320 As explained above, introducing an artificial disease association into realistically synthesized
321 flow cytometry datasets by adjusting the frequency of one cell population (in this case, CD4$^+$
322 T_{EM} cells) leads to changes in all other nodes in our gating tree. We next asked whether our
323 restriction method could also improve the discoverability of these covariant markers in the
324 synthesized datasets presented above. Of note, rzAUC allows us to compare the discriminatory
325 performance of different markers within one dataset; however, rzAUC values are not
326 comparable between datasets, including independent training, validation and test datasets
327 (Supplementary Fig 5). The AUC is equivalent to the Mann-Whitney U-statistic[33] and we can
328 extend this equivalence to the rzAUC; however, this is not helpful in assigning significance
329 values because optimizing for highest rzAUC introduces a bias. Instead, we must calculate
330 permutation p-values[38]. For each of our four realistic synthesized examples, we calculated
331 permutation p-values using the unrestricted sample set and the optimally restricted sample set
332 for every gated cell population. Figure 6 shows these p-values as scatter plots in which the
333 green-shading demarcates unrestricted p-values > 0.05 and optimally restricted p-values < 0.05
334 – that is, markers identified as significant using our restriction method, but missed without it.

In our example of a symmetric ROC curve, we found that $CD4^+$ T_{EM} cells and 3 subordinate populations were significant discriminators in both the unrestricted and restricted datasets (Fig 6a). Two further populations were significant only in the restricted dataset. In the second example, which resulted in a right-skewed ROC curve, we found that two populations (i.e. $CD4^+$ T_{EM} cells and $CD27^+$ $CD28^+$ $CD57^-$ $CD4^+$ T_{EM} cells) had an optimal restriction permutation p-value = 0, whereas the corresponding unrestricted permutation p-value was > 0.05 (Fig 6b). 8 other subsets were significant discriminators after restriction, but were not significant in the unrestricted sample-set. In the third example, which resulted in a left-skewed ROC curve, we found that $CD4^+$ T_{EM} cells had an optimal restriction permutation p-value = 0, but were not significant in the unrestricted dataset (Fig 6c). 11 other subsets were significant discriminators after restriction, but not in the unrestricted sample-set. In the fourth example, $CD4^+$ T_{EM} cells had an optimal restriction permutation p-value = 0.007, but were not significant in the unrestricted dataset (Fig 6d). Hence, dataset restriction enables discovery of disease markers which would otherwise be disregarded in synthesized flow cytometry datasets.

Dataset restriction discovers valid irAE markers

Having qualified our restriction method using synthesized datasets, we next applied it to real clinical data. In previous work, we investigated pre-treatment peripheral blood samples from 110 patients with advanced melanoma who received Ipi-Nivo therapy[13]. Using conventional methods, we found no significant marker after correcting for multiple comparison. Here, we asked whether our restriction method could reveal any novel markers of hepatitis or colitis risk in the same dataset. No markers of colitis survived correction for multiple comparison (Fig 6e and Supplementary Fig 6). However, in predicting hepatitis, our restriction method returned 7 significant markers with an unrestricted permutation p-value > 0.05 (Fig 6f). After correction

for multiple testing, 4 of these 7 hepatitis markers remained significant with an FDR < 0.05. By contrast, no marker identified from the unrestricted dataset returned a significant permutation p-value after correction for multiple testing. Thus, our restriction method returned significant disease-associated markers, which were not found using the unrestricted dataset.

Using our novel restriction method, we identified $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} cell frequency relative to $CD4^+$ in blood as a marker of hepatitis risk after dataset restriction. To illustrate the potential utility of restricted markers, we compared the performance of $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} frequency as a marker of hepatitis risk in our unrestricted and restricted datasets (Fig 7). The discriminatory cut-off for patient classification, defined by the Youden index, was the same for both the restricted and unrestricted datasets, such that samples with more than 9.56% of $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} relative to $CD4^+$ are predicted hepatitis positive. Accordingly, using the unrestricted dataset, $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} (%) correctly predicted incidence of hepatitis in 74 of 110 patients. The unrestricted cell frequency had a sensitivity (TPR) of 45.8% and a specificity (true negative rate, TNR) of 83.9%. The positive predictive value (PPV) was 68.8% and the negative predictive value (NPV) was 66.7%. Our restriction method usually implies that some cases should be considered unclassifiable based upon its marker values. In this example, 58 of 110 patients were deemed unclassifiable. Incidence of hepatitis was correctly predicted in 40 of 52 classifiable samples. The restricted cell frequency had a sensitivity of 91.7% and a specificity of 64.3%. The positive predictive value was 68.8% and the negative predictive value was 90% .

Multivariate analysis of restricted data predicts hepatitis

Although our restriction method leads to discarding many samples as unclassifiable by any particular marker, we found that different markers define noncongruent sets of classifiable samples (Fig 8a). This led us to investigate whether using restricted datasets could improve the predictive performance of multivariate models. First, we built a random forest model[39] using all 84 markers from the unrestricted training dataset of 110 patients (Fig 8b). When this model was applied to a fully independent, prospective validation set of 30 patients, the resulting predictions were inaccurate (CCR=56.7% vs. 54.8% under the no-information model).

By contrast, we observed a significant improvement in predictive performance using the restricted dataset to train our random forest. To leverage information from our restriction method, we assigned a value of -1 to restricted samples across all 84 markers. When this model was applied to the validation set, the resulting predictions were significant (Fisher's Exact p-value = 0.026) and had a correct classification rate of 73.3%. 12 of 16 predictions of hepatitis were correct (PPV=75%) and 10 of 14 negative predictions were correct (NPV=71.4%). Hence, in principle, dataset restriction can improve the training and performance of multivariate predictive models.

DISCUSSION

Immunological diseases are often heterogeneous in clinical presentation and severity, reflecting the variability of their underlying immunopathologies[15]. It follows, we argue, that immune disease-associated markers typically exhibit greater variance among diseased patients than unaffected individuals. This general proposition was broadly corroborated by our real-world examples of patient groups who were prone to immunotherapy-related complications. Unequal variance in marker distribution between patient classes affects our ability to identify markers with discriminatory capacity over a restricted range of marker values. To solve this biological problem, we introduced dataset restriction as a marker discovery tool. In artificial and real-world examples, dataset restriction enabled us to find discriminatory markers that were disregarded by conventional measures of marker performance. Moreover, we showed that dataset restriction improves performance of multivariate predictive models. Our work formalizes a new way of evaluating diagnostic results – namely, that certain markers can only be usefully interpreted over a restricted range of values, and that samples with values outside this range should be considered as unclassifiable.

Flow cytometry is a powerful method for interrogating the phenotype of many single cells within a heterogeneous mixture. This technique allowed us to estimate the relative numbers of accurately defined leucocyte subsets in peripheral blood samples, including T cell subsets, which are direct targets of Ipilimumab (anti-CTLA-4) and Nivolumab (anti-PD-1) therapy[40]. Although flow cytometry generates rich and immunologically interpretable data, it has two key limitations – namely, that blood leucocyte frequencies vary within a narrow dynamic range, and that higher order cell feature combinations may define rare cell subsets[41,42]. Small disease-related changes in markers are problematic because substantially overlapping marker

421 distributions with unequal variance lead to exaggerated skewness of ROC curves. Rare cell
422 subsets are problematic because our estimates of their frequency are less reliable[43]. Crucially,
423 dataset restriction helps to overcome the special difficulties of correctly interpreting flow
424 cytometry data by limiting marker values to a range in which the signal-to-noise ratio is
425 increased relative to the full range. Consequently, we reduce the likelihood of false positive or
426 false negative classification at the cost of discarding some samples as unclassifiable.
427
428 We created an R-package, called *restrictedROC*[36], that calculates restricted standardized AUC
429 scores. The rzAUC is returned together with a value that delimits the marker's informative
430 range. This builds upon earlier ideas about partial AUCs, which were introduced to account for
431 imposed restrictions that capped true and false positive rates[44-46]. Imposed restrictions usually
432 come from domain knowledge; for instance, tests with a high false positive rate are
433 inappropriate for expensive diagnostic screening applications, whereas tests with a high false
434 negative rate are inappropriate when a life-saving treatment is available[47]. McClish introduced
435 a "standardization" for partial AUCs for a given range of false positive rates, such that a
436 randomly selected positive sample has a higher value than a randomly selected negative sample
437 conditional upon the negative sample arising from the false positive range[48]. In our method, we
438 introduced a scaling factor for the two-way partial AUC[47] resulting in the restricted AUC
439 (rAUC). With this scaling factor, the rAUC becomes the probability that a randomly selected
440 positive sample has a higher value than a randomly selected negative sample conditional upon
441 *both samples* arising from a range spanned by a minimum true positive rate and a maximal false
442 positive rate. The restricted standardized AUC (rzAUC) then takes into account both the rAUC
443 and the number of samples in the marker^{HIGH} or marker^{LOW} range leveraging the equivalence
444 between AUC and Mann-Whitney U test[33].

445 We further developed our method to determine the optimal range of marker values that correctly
446 classifies samples. Specifically, we optimize a restriction that either includes samples with
447 higher marker values ($marker^{HIGH}$) or lower marker values ($marker^{LOW}$) and has the highest
448 possible absolute rzAUC. The rzAUC can be directly compared within one dataset, but depends
449 on the total number of samples. By calculating permutation p-values[38] for the rzAUC, we
450 remove this dependence and attribute significance values.
451
452 There are alternative ways of describing the geometric symmetry of ROC curves apart from
453 graphical skewness. Left-skewed ROC curves are also described as True Negative Proportion
454 (TNP)-asymmetric and right-skewed ROC curves as True Positive Proportion (TPP)-
455 asymmetric. These asymmetries can be defined by Kullback-Leibler divergences[49] (KL-
456 divergences). Therefore, KL-divergence could be used to assess whether restriction should be
457 applied to a given marker; however, in the case of symmetric ROC curves, our restriction keeps
458 all samples, so such preselection of markers is unnecessary. Importantly, excluding samples to
459 minimize KL-divergence is not the equivalent of dataset restriction.
460
461 In principle, dataset restriction can be applied to optimize any marker range. However,
462 following from our immunological rationale, restricting the upper or lower range is especially
463 applicable in clinical diagnostics. For completeness of our discussion, we can imagine a marker
464 with both uninformative $marker^{HIGH}$ and $marker^{LOW}$ values (ie. where only mid-range values
465 are informative) that might only be discovered by applying our restriction method twice in
466 succession. In theory, restriction could be iteratively applied to a dataset until no more samples
467 are identified as unclassifiable, but the practical value of multiply restricting datasets is unclear.
468

In order to validate our restriction method, we developed a novel method for synthesizing realistic flow cytometry data with class-related effects. Because no generative method previously existed, our approach represents a significant contribution to cytometry analysis, particularly for benchmarking of diagnostic flow cytometry algorithms, sample size calculations or data augmentation. Our method uses an expert-given hierarchical gating strategy, where the proportions of cells per gate are described with a Dirichlet distribution. Within each terminal (leaf) gate, the cells are described using a normal distribution. Thus, we effectively created a Gaussian mixture distribution with the number of components defined by the number of terminal gates. In cytometry, (Gaussian) mixture models are an established method for unsupervised cell population identification[50,51]. In principle, these earlier approaches could be used to generate new cells from estimated distributions, although their focus was labeling existing cells rather than creating artificial ones. Our use of a hierarchical gating strategy and a Gaussian mixture model allows for the creation of complex data distributions. In future, for certain applications, further adaptions of our approach, such as multivariate skew t-distributions, could be used to improve the accuracy of simulated data[52].

Restricting markers to an informative range of values is important because it improves classification performance. We emphasize that classification cut-offs and restriction values are different concepts. Classification cutoffs, such as the Youden index[53], divide a sample set into predicted positive and predicted negative classes. By contrast, restriction divides a sample set into classifiable and unclassifiable samples. In the context of individualized patient care, it might seem unproductive to label samples as unclassifiable. On the contrary, we argue that the clinical utility of a predictive marker improves if its certainty is high, even if it is only works in a small subset of patients. Consider a disease-related marker giving a right-skewed ROC curve:

493 Conventional approaches return a reliable positive classification and an unreliable negative
494 classification; in contrast, our restriction method returns a reliable positive classification, a
495 reliable negative classification and a set of unclassifiable samples, which do not necessarily
496 have the most negative values. Of note, the discriminatory cut-off determined by the Youden
497 index is often the same after restriction but may change in some cases. When interpreting a
498 single marker, our restriction method improves either the positive or the negative predictive
499 value, so improves certainty of our predictions.
500
501 Our method may concern some clinicians, who will legitimately ask about unclassifiable
502 patients[54]. Here, we provide an answer by building an informative and prospectively validated
503 random forest model after replacing all restricted values with a constant outside the informative
504 range. Consequently, we force each tree of the random forest to select a cut-off within the
505 informative range or a cut-off between the classifiable and unclassifiable regions. More
506 sophisticated methods may be developed in future, but our experimentally validated random
507 forest is a proof-of-principle that differently restricted markers can be usefully combined in
508 multivariate models.
509
510 To demonstrate the potential clinical utility of dataset restriction, we applied our method to the
511 clinically significant problem of immune-related adverse events following combined
512 immunotherapy. In univariate analyses, dataset restriction identified new markers associated
513 with ICI-related hepatitis, including $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} cells, that were not returned by
514 conventional methods. Of clinical importance, dataset restriction increased NPV without
515 compromising PPV. Combining many restricted markers into a random forest model generated
516 an informative model, whereas training on unrestricted data from the same set of 110 samples

517 returned no valid models. To validate our predictive model, we assessed its performance in an
518 independent, prospectively collected set of 30 samples, where it returned significant
519 predictions, which were superior to the performance of any single marker alone. Beyond the
520 scope of this article, such multivariate models could be extended to include markers from
521 multiple flow cytometry panels or other patient-related information, such as age, sex or clinical
522 chemistry results.
523
524 Clinical manifestations of immune disease are often heterogeneous. This is certainly true of
525 irAE after immunotherapy, which vary greatly in severity, time-of-onset, clinical features and
526 response to treatment[28]. Further, there is increasing evidence that multiple immune aetiologies
527 lead to common clinical presentations, such as colitis[55], myositis[56] or hepatitis[57]. This
528 heterogeneity connotes individual genetic predisposition[58,59], environmental factors[60,61] and
529 past immunological challenges[35]. In particular, we now recognize the contribution of previous
530 viral infections in preconditioning towards adverse reactions. An unanticipated consequence of
531 dataset restriction is that disease markers with a bimodal distribution in the positive class, such
532 as might arise from multiple aetiologies, are findable. Excitingly, combining features from a
533 restricted dataset into multivariate models should, in principle, enable predictions about
534 diseases with multiple aetiotypes – a situation where conventional methods are unsuitable.
535 Extending this idea of dataset restriction as a way of classifying samples with intraclass
536 heterogeneity to unsupervised methods, such as PCA or clustering, could aid discovery of
537 previously unknown patient subsets.
538
539 In summary, clinical markers that can only be interpreted over a restricted range are inherently
540 likely in immune diseases. Where classical methods fail, dataset restriction solves the problem

of discovering and interpreting such markers. Our approach is not limited to prospective data, but can also be used retrospectively to find new markers or improve existing ones. Dataset restriction was developed here to analyze flow cytometry data; however, it is directly applicable to any sample classification problem. In immunological research, this could include transcriptomic, proteomic, (epi-)genomic, metabolomic or imaging data. We hope others will apply our method to existing datasets, perhaps leading to valuable new markers or novel biological insights.

MATERIALS AND METHODS

Collection of clinical information

Data from three sources were used in this study: (1) a training set (n=48) from a cohort of healthy humans used to develop our flow cytometry data simulations; (2) a previously reported training set (n=110) from patients with advanced melanoma used for marker discovery[13]; and (3) a new prospective validation set (n=30) from patients with advanced melanoma. Whole blood was collected from healthy thrombocyte donors with approval from the Ethics Committee of the University of Regensburg (approval 22-2780-01). All donors gave full, written consent to sample and data collection. Clinical samples for the marker training and validation sets were collected within a single-center, non-interventional study[62],which was conducted in accordance with the Declaration of Helsinki and all applicable German and European laws and ethical standards. This observational study was authorized by the Ethics Committee of the University of Regensburg (approval 16-101-0125) and registered with clinicaltrials.gov (NCT04158544). Blood samples were obtained from patients with Stage III/IV melanoma under the care of the Department of Dermatology at University Hospital Regensburg (UKR). Eligible patients were consecutively recruited without stratification or matching. All participants gave full, informed written consent. For the training set, the first reported case was recruited in OCT-2016 and the last reported case was recruited in JUN-2021. For the prospective validation set, the first reported case was recruited in JUN-2021 and the last reported case was recruited in JAN-2023 (Supplementary Table 1). All study participants received standard-of-care treatment according to local guidelines. Specifically, patients with unresectable metastatic disease who received first- or second-line checkpoint inhibitor therapy were initially treated with Nivolumab (αPD-1; 1 mg/kg; Bristol-Myers Squibb) plus Ipilimumab (αCTLA-4; 3 mg/kg; Bristol-Myers

Squibb) for up to four cycles at 3 week intervals. Thereafter, patients received 480 mg Nivolumab monotherapy at 4 week intervals.

Diagnosis and grading of clinical outcomes.

All irAE were evaluated by an expert Dermatological Oncologist. ICI-related hepatitis was diagnosed when: (i) GOT, GPT, γ-GT or total bilirubin substantially deviated from pretreatment values; (ii) this change was not attributable to other causes, such as co-medication or viral disease; and (iii) liver injury was sufficiently severe that ICI therapy was suspended or stopped, or immunosuppression was started. Colitis was diagnosed when increased stool frequency or loose consistency, accompanied by abdominal discomfort led to suspension or cessation of ICI therapy and introduction of immunosuppressive treatment. Clinical responses were assessed using the Response Evaluation Criteria in Solid Tumors (RECIST 1.1)[63]. Patients with progressive disease were categorized as non-responders, whereas those with complete or partial responses, and those with stable disease, were categorized as responders.

Flow cytometry.

Step-by-step protocols for preparing and analyzing clinical samples by flow cytometry can be accessed through Nature Protocol Exchange[64]. Briefly, blood was collected into EDTA-vacutainers by peripheral venepuncture then delivered to the responsible lab at ambient temperature. Samples were stored at 4°C for up to 4 h before processing. Whole blood samples were stained using the DURAClone IM T Cell Subsets Tube (Beckman Coulter, B53328). Data were collected using a Navios™ cytometer running Cytometry List Mode Data Acquisition and Analysis Software version 1.3 (Beckman Coulter). An experienced operator performed blinded

analyses following a conventional workflow that entailed sample-wise recompensation, arcsinh transformation and rescaling before applying a uniform gating strategy (Supplementary Fig 4).

Restriction method

We propose a method for finding markers with high performance in subsets of samples that involves: 1) "restricting" samples into marker^{HIGH} and marker^{LOW} sets for every unique marker value; 2) calculating the corresponding restricted receiver operating characteristic(ROC) curve; 3) calculating the area under the restricted ROC curve; 4) adjusting the restricted AUC (rAUC) for sample size; 5) selecting the optimal restriction level, 6) calculating permutation p-values; and 7) reporting performance and significance. This algorithm is implemented as an R package called *restrictedROC*[36].

To define our nomenclature, we first introduce ROC curve analysis. Let a cut-off $c \in \mathbb{R}$, a continuous marker $Y \in \mathbb{R}$ and a grouping of samples into *diseased* (positive, $D = 1$) and *non-diseased* (negative, $D = 0$). A sample can be classified as diseased if $Y \geq c$ and into non-diseased if $Y < c$. The true positive rate (TPR) and false positive rate (FPR) at cut-off c are defined as $TPR(c) = P[Y \geq c|D = 1] = P[Y_D \geq c]$ and $TPR(c) = P[Y \geq c|D = 0] = P[Y_{\bar{D}} \geq c]$. The ROC curve relates the TPR and FPR for all possible cut-offs c, including $\{\infty, -\infty\}$ (*nb*. compare with Supplementary Fig 7). We can write the value of the ROC curve at any false positive rate $t \in (0,1)$ as $\mathrm{ROC}(t) = TPR\left(FPR^{-1}(t)\right)$. For empirical survival functions S_D and $S_{\bar{D}}$ we write $\mathrm{ROC}(t) = S_D\left(S_{\bar{D}}^{-1}(t)\right)$ by substituting TPR and FPR. The area under the ROC curve (AUC) is then defined as

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(t)\mathrm{d}t \ . \quad (1)$$

Consequently, a perfectly discriminating marker with higher values corresponding to the positive class translates to a perfect ROC curve with AUC = 1. An uninformative marker has an AUC of 0.5, corresponding to $\mathrm{ROC}(t) = t \; \forall t \in (0, 1)$. A perfectly discriminating marker but with higher values corresponding to the negative class has an AUC of 0. From a probabilistic point of view, the AUC equals the probability that the marker value of a random positive sample will be higher than that of a random negative sample: $\mathrm{AUC} = P[Y_D > Y_{\bar{D}}]$ [33,65,66]. The derivation is given in Supplementary Note 2.

Next, we introduce the concept of restricted ROC curves. Our "restriction" is a marker value that splits the samples into marker$^{\mathrm{HIGH}}$ and marker$^{\mathrm{LOW}}$ sets. For both sets, we separately calculate "restricted" ROC curves and their corresponding restricted AUC (rAUC). See supplement for the full derivation. In Supplementary Note 3, we prove that calculating rAUC is identical to scaling a partial AUC (pAUC). Therefore, before we describe our computational method, we consider the (two-way) pAUC[47,67]. The partial AUC (pAUC) is defined as the AUC up to a certain false positive rate. Its probabilistic correspondence has been shown[45,66]:

$$\mathrm{pAUC}(t_0) = \int_0^{t_0} \mathrm{ROC}(t)\mathrm{d}t = P\left[Y_D > Y_{\bar{D}} \mid Y_{\bar{D}} > {S_{\bar{D}}}^{-1}(t_0)\right] \cdot t_0 \tag{2}$$

The pAUC was recently extended to two-way partial AUCs[47]. Here, the area is calculated between $S_D(t) \geq 1 - \alpha$ and $S_{\bar{D}}(t) \leq \beta$. This area, shown in Supplementary Fig 8 as shaded area **A**, can be written as

$$\mathrm{AUC}_{\alpha}^{\beta} = \int_{S_{\bar{D}}\left(S_D^{-1}(1-\alpha)\right)}^{\beta} \mathrm{ROC}(t)\,\mathrm{d}t \; - \; (1-\alpha)\left(\beta - S_{\bar{D}}\left({S_D}^{-1}(1-\alpha)\right)\right) \tag{3}$$

$$= P\left[Y_D > Y_{\bar{D}}, Y_D \leq \; S_D^{-1}(1-\alpha), Y_{\bar{D}} \geq \; S_{\bar{D}}^{-1}(\beta)\right] \tag{4}$$

Our restriction method uses two special cases of AUC_{α}^{β}, shown in Supplementary Fig 9:

1) The left part of the area under the curve up to a false positive rate β, which is identical to the pAUC described earlier

$$\mathrm{AUC}_{high}(\beta) = \mathrm{AUC}_{\alpha\equiv 1}^{\beta} = \tag{5}$$

$$= \int_0^{\beta} \mathrm{ROC}(t)\,\mathrm{d}t \tag{6}$$

$$= P\left[Y_D > Y_{\bar{D}}, Y_{\bar{D}} > {S_{\bar{D}}}^{-1}(\beta)\right] \tag{7}$$

2) The right part of the area under the curve with at least a true positive rate of $1-\alpha$

$$\mathrm{AUC}_{low}(\beta) = \mathrm{AUC}_{\alpha}^{\beta\equiv 1} = \tag{8}$$

$$= \int_{S_{\bar{D}}\left(S_D^{-1}(1-\alpha)\right)}^{1} \mathrm{ROC}(t)\,\mathrm{d}t\,(1-\alpha)\left(1 - S_{\bar{D}}\left({S_D}^{-1}(1-\alpha)\right)\right) \tag{9}$$

$$= P\left[Y_D > Y_{\bar{D}}, Y_D \leq {S_D}^{-1}(1-\alpha)\right] \tag{10}$$

Partial AUCs consider only a specific part of the original ROC curve, therefore the interpretation of perfect ($\mathrm{AUC} = 1$) or non-informative ($\mathrm{AUC} = 0.5$) becomes invalid. For pAUC, the following standardization was proposed to restore this interpretation[48]

$$standardized\ \mathrm{pAUC} = \frac{1}{2}\left(1 + \frac{\mathrm{pAUC} - \min}{\max - \min}\right) \tag{11}$$

where, min is the pAUC given an non-informative marker $\left(\min = \frac{\beta^2}{2}\right)$, and max is the pAUC given a perfect marker ($\max = \beta$) up to an false positive rate of β.

In contrast, our restriction method applies the following two scaling factors to any two-way partial $\mathrm{AUC}_{\alpha}^{\beta}$

$$\mathrm{rAUC}_{\alpha}^{\beta} \coloneqq \mathrm{AUC}_{\alpha}^{\beta} \cdot \frac{1}{\beta - S_{\bar{D}}\left(S_D^{-1}(1-\alpha)\right)} \cdot \frac{1}{S_D\left(S_{\bar{D}}^{-1}(\beta)\right) - (1-\alpha)} \tag{12}$$

Effectively, these two scaling factors rescale the area spanned through α and β to 1. Importantly, this is equivalent to calculating rAUC considering only samples with $S_{\bar{D}}^{-1}(\beta) < t < S_D^{-1}(1-\alpha)$. This has a probabilistic interpretation of

$$\mathrm{rAUC}_{\alpha}^{\beta} = P\left[Y_D > Y_{\bar{D}} \mid S_{\bar{D}}^{-1}(\beta) \leq Y \leq S_D^{-1}(1-\alpha)\right] \tag{13}$$

Here, the $\mathrm{rAUC}_{\alpha}^{\beta}$ is defined in terms of maximum false positive rate $1-\alpha$ and minimum true positive rate β. Alternatively, we introduce a "restriction" $r \in \mathbb{R}$ which splits the data into marker$^{\mathrm{HIGH}}$ and marker$^{\mathrm{LOW}}$ sets where $\alpha \coloneqq 1 - S_D(r)$ and $\beta \coloneqq S_{\bar{D}}(r)$. With this, our two special cases become

$$\mathrm{rAUC}_{high}(r) = \mathrm{rAUC}_{\alpha\equiv 1}^{\beta = S_{\bar{D}}(r)} = \mathrm{AUC}_{high}\left(S_{\bar{D}}(r)\right) \cdot \frac{1}{S_{\bar{D}}(r)} \cdot \frac{1}{S_D(r)} \tag{14}$$

$$\mathrm{rAUC}_{low}(r) = \mathrm{rAUC}_{\alpha = 1 - S_D(r)}^{\beta\equiv 1} = \mathrm{AUC}_{low}(1 - S_D(r)) \cdot \frac{1}{1 - S_{\bar{D}}(r)} \cdot \frac{1}{1 - S_D(r)} \tag{15}$$

This is equivalent to keeping marker$^{\mathrm{HIGH}}$ samples with values $> r$ $\left(\mathrm{rAUC}_{high}\right)$ or to keeping marker$^{\mathrm{LOW}}$ samples with values $\leq r$ (rAUC_{low}), then calculating AUC on the restricted dataset. Supplementary Video 3 uses a hypothetical dataset to visualize the rAUC and show the visual equivalence of our scaling factor and when restricting the dataset.

More extreme restrictions result in fewer samples, so our estimates of $\mathrm{rAUC}(r)$ become increasingly unreliable; therefore, we adjust $\mathrm{rAUC}(r)$ for sample size after restriction. Here, we leverage the equality of the AUC to the Mann-Whitney U test[33] in order to calculate the restricted standardized AUC (rzAUC_X) for X either marker$^{\mathrm{HIGH}}$ and marker$^{\mathrm{LOW}}$ sets by calculating the test statistic

$$\mathrm{rzAUC}_X(r) = \frac{\mathrm{rAUC}_X(r) - 0.5}{\sqrt{var_{H_0}\left(\mathrm{rAUC}_X(r)\right)}} \tag{16}$$

where $var_{H_0}\left(\mathrm{rAUC}_X(r)\right)$ is the variance under the null hypothesis H_0 that positive and negative samples are independent and identically distributed. This demands no assumption of normality. Then this var_{H_0} is[68,69]

$$var_{H_0}\left(\mathrm{rAUC}_X(r)\right) = \frac{m+n+1}{12mn} \tag{17}$$

where m is the number of positive samples and n is the number of negative samples with marker values higher (rzAUC_{high}) or lower or equal (rzAUC_{low}) than the restriction r. With this adjustment, a higher number of samples reduces variance, hence rzAUC_x becomes more reliable. For a visual example, see Supplementary Fig 10 where the rAUC and rzAUC are shown for all possible restrictions in terms of the false positive rate. The rzAUC_X can be negative if the corresponding rAUC_X is below 0.5, decreases with fewer samples and increases in absolute value the further rAUC_X is from 0.5.

After calculating the $rzAUC$, we next identify the optimal restriction, which is defined as the highest absolute value of rzAUC_{high} or rzAUC_{low}. Including more samples would result in a smaller rAUC_X and therefore smaller rzAUC_X. Excluding more samples would result in an equal or higher rzAUC_X but also a higher variance and therefore also a smaller rzAUC_X. With this restriction we include some and potentially, but not necessarily, exclude other samples in the calculation of the rAUC. We describe the excluded samples as "unclassifiable" and remove them from further calculation of usual performance measures like accuracy, specificity or sensitivity.

Finally, we calculate permutation p-values for the unrestricted AUC and rzAUC. After obtaining the *unrestricted* AUC for an unrestricted dataset or the rzAUC_X for an optimized subset of samples, we need to assign a p-value using permutation tests. This a non-parametric way to determine statistical significance based upon a null hypothesis that class labels assigned to samples are interchangeable[38]. Following this approach, we first calculate unrestricted AUC, rzAUC_{high} and rzAUC_{low} using the correct labels. Then we permutate the labels 10,000 times before recalculating unrestricted AUC, rzAUC_{high} and rzAUC_{low}. The unrestricted permutation

p-value is then the number of times the permutated unrestricted AUC is above the original unrestricted AUC. Likwise, the restricted permutation p-value is then the number of times either rzAUC_{high} or rzAUC_{low} are absolutely higher than the optimal rzAUC_X.

Multivariate restriction analysis

Our restriction method identifies only a part of the samples as classifiable and cannot make predictions for the unclassifiables. This potentially excludes many samples, so constrains predictive power. To circumvent this problem, we replace the marker values of unclassifiable samples with a clearly distinct value (-1) and then apply a random forest. With this substitution, we can predict all given samples, regardless if they are unclassifiable by some markers. In our melanoma dataset, per sample we first downsampled 10,000 $CD3^+$ T cells. We then restricted our set of features to 84 gates where at least 10% of 110 training samples contained more than 10 counts. Then we calculated relative proportion to either $CD4^+$ $CD8^-$ or $CD4^-$ $CD8^+$ T cells. We also used $CD4^+CD8^+$ (double positive), $CD4^-$ $CD8^-$ (double negative), $CD4^+$ $CD8^-$ and $CD4^-$ $CD8^+$ T cell counts, which are relative to the fixed parent gate of 10,000 $CD3^+$ T cells.

For our unrestricted, classical multivariate approach, we used the proportions and counts of all 110 previously published training samples. We then trained a random forest[39] model using the H2O R library[70] with 1000 trees and the number of bins of 100, a random manual seed for reproducibility of the results the remaining default parameters. Explicitly, a maximum depth of 20, a minimum number of samples in a node of 1, logloss stopping metric, the number of randomly sampled candidate markers as floor of the square root of 84 (9), a sample rate of 0.632, minimum split improvement of 10^{-5} and an automatic histogram type. Finally, we applied the random forest on a prospective cohort of n=30 patients.

For our restricted multivariate approach, we performed a marker-wise restriction to samples, then replaced all unclassifiable marker values with -1. (We chose this value because all classifiable values are strictly positive as they represent either proportions of $CD4^+$ or $CD8^+$ T cells, or absolute T cell counts.) This substitution forces each tree in the random forest to select discriminatory cutoffs within the range of informative marker values. We then trained a random forest model with the same settings as for the unrestricted multivariate approach. We finally applied the restriction values obtained from the training set to the prospective validation set, replaced the unclassifiable marker values with -1 and applied the random forest on the prospective cohort.

Synthesizing realistic flow cytometry data

Our method to synthesize realistic flow cytometry data is accessible as python[71] package *NBNode* via github https://github.com/ggrlab/NBNode[72]. The process of hierarchically gating cells and simulating data with any given effect in any cell population involves five steps. In the following, **bold** letters or arrows above the letter ($\vec{a}$) denote vectors, regular letters single values.

In the first step, we applied a uniform manual gating to 48 human peripheral blood samples stained with the DURAClone IM T Cell Subsets Tube (Beckman Coulter GmbH). Data were preprocessed by manually recompensating the samples, removing the TIME feature, and asinh transforming all cell features x

$$asinh_{cofactor}(x) = asinh\left(\frac{x}{cofactor}\right) \quad (18)$$

with the following cofactors: FS INT: 1, FS TOF: 1, SS INT: 1, CD45RA FITC: 1000, CCR7 PE: 2000, CD28 ECD: 2000, PD1 PC5.5: 800, CD27 PC7: 3000, CD4 APC: 4000, CD8 AF700: 10000, CD3 AA750: 500, CD57 PB: 2000, CD45 KrO: 20. Because the channel-wise median fluorescence intensity (MFI) varied between samples, this alone was not sufficient to apply the exact same gating to all samples. Therefore, we performed a sample-wise rescaling (Supplementary Fig 11 and Supplementary Video 4). For every cell feature x, we identified the positive and negative population of all cells and found the corresponding MFI_x^+ and MFI_x^-. Using these, the rescaling min-max standardizes all cells per sample,

$$rescale(x) \coloneqq \frac{x - MFI_x^-}{MFI_x^+ - MFI_x^-} \tag{19}$$

leading to a rescaled $MFI^+_{rescale(x)}$ of 1 and a rescaled$MFI^-_{rescale(x)}$ of 0.

We then applied a standard gating strategy, which is shown schematically (Supplementary Fig 12a) and explicitly for a real-world sample (Supplementary Fig 4). This hierarchical gating of biaxial scatter plots is effectively a decision tree with 98 "leaf" gates (Supplementary Fig 12a). Each leaf gate corresponds to a terminal gating node and all supraordinate nodes are "intermediate" gates. Importantly, every cell must fall into one, and only one, of the subordinate 98 leaf gates.

In the second step, we model the proportion of cells in each leaf gate after uniformly gating all cells from all samples. Specifically, we describe the proportion of cells in each gate according to a Dirichlet distribution $Dir(\boldsymbol{\alpha})$ (Supplementary Fig 12b,c). The Dirichlet distribution is a suitable choice after its mass is only on non-negative compositions that sum up to one. Following Minka *et al.*[73], let $\boldsymbol{p} \in (0,1)^K$ be one random vector of proportions such that $\sum_k^K p_k = 1$ for $k \in \{1, \dots, K\}$ for K cell populations. In our case, all cells of a sample fall into one and only one of the 98 terminal gates. Therefore, the sum of the cell percentages in each

terminal gate adds up to 100%. The probability density under the Dirichlet model with a parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^K_{>0}$ is defined as

$$\mathrm{p}(\boldsymbol{p}) \sim Dir(\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_k^{\alpha_k - 1} \tag{20}$$

$$\text{with } p_k > 0 \text{ and } \sum_k^K p_k = 1$$

More intuitively, the $\boldsymbol{\alpha}$ parameters can be split into mean proportions per cell population and a precision:

$$\boldsymbol{m} = E[\boldsymbol{p}] = \frac{\boldsymbol{\alpha}}{\sum_k^K \alpha_k} \quad \text{(mean vector)} \tag{21}$$

$$\mathrm{s} = \sum_k^K \alpha_k \quad \text{(precision)} \tag{22}$$

Hence, a useful explanation of the parameters is that the higher the precision, the more localized the probability becomes around the means. $\alpha_x > \alpha_y$ indicates that, on average, the proportion of cell population x is higher than the proportion of cell population y. If $0 < \alpha_k < 1$, the distribution is effectively pushed away from the corresponding cell population. See Supplementary Fig 13 and Supplementary Table 2 for examples of the Dirichlet distribution with K=3 and different parametrizations of $\boldsymbol{\alpha}$. Plots were built using the R-package[74]. We set a maximum likelihood estimation of the distribution parameters $\boldsymbol{\alpha}$ [75] based on $N_{samples} = 48$ measured cell population proportions $\boldsymbol{p}^{(i)}$ for $i \in \{1, \ldots, N_{samples}\}$. In some cell populations and samples there were no cells so the proportion became zero. Because the estimation cannot handle proportions equal to zero, we added a pseudo-proportion to all proportions and normalized to 1 before applying maximum likelihood estimation. With this, the zero-adjusted proportion $\boldsymbol{p}_k^{(i)''}$ of sample i and cell population k becomes

$$\boldsymbol{p}_k^{(i)''} = \frac{\boldsymbol{p}_k^{(i)'}}{\sum_k^K \boldsymbol{p}_k^{(i)'}} \tag{23}$$

$$\text{with } \boldsymbol{p}_k^{(i)'} = \boldsymbol{p}_k^{(i)} + 0.001 \cdot \min(all\ proportions) \tag{24}$$

We end up with a Dirichlet distribution with estimates for the parameter $\hat{\boldsymbol{\alpha}}$

$$Dir(\hat{\alpha}_1, \hat{\alpha}_2, \dots \hat{\alpha}_K) \tag{25}$$

In the third step, we build a gating hierarchy using the estimated $\boldsymbol{\alpha}$ *parameters* corresponding to the leaf nodes. We used the estimated Dirichlet parameters and manual gating structure to create a probabilistic representation of the gating hierarchy. In this structure, all cells fall into one and only one gate. To calculate intermediate nodes, we sum the estimated α parameters according to the manual gating tree, starting from the bottom and working to the top. Given a Dirichlet distributed variable with K cell populations

$$\mathrm{p}(\boldsymbol{p}) = (p_1, \dots, p_K) \sim Dir(\alpha_1, \alpha_2, \dots, \alpha_K) \tag{26}$$

the sum of any two cell populations is again Dirichlet distributed

$$(p_1, \dots, p_i + p_j, \dots, p_K) \sim Dir(\alpha_1, \dots, \alpha_i + \alpha_j, \dots, \alpha_K) \tag{27}$$

Therefore, every intermediate or leaf node is described by a Dirichlet distribution. Intuitively, all cells of any gate must fall in one of the subsequent gates and can, therefore, reflect a Dirichlet distribution. To visualize proportions corresponding to these parameters, the decision tree was shaded in red, such that deeper red indicates a higher proportion of cells in that gate (Supplementary Fig 12a).

In the fourth step, we fit a cell feature distribution using cells from all samples per leaf gate (Supplementary Fig 12d,e). The Dirichlet distributions only describe the number of cells in every gate – that is, a vector of K cell population proportions $\boldsymbol{p} \in (0,1)^K$. However, a flow cytometry measurement results in a $\mathbb{R}^{n \times m}$ matrix with n cells and m cell features where every

cell comes from a specific cell population. Each such cell population is defined by the m continuous cell feature values. Accordingly, we model the cells for each leaf node l by a multivariate normal distribution $\mathcal{N}(\vec{\mu}_l, \Sigma_l)$ with mean $\vec{\mu}_l \in \mathbb{R}^m$ and covariance matrix $\Sigma_l \in \mathbb{R}^{m \times m}$. In the illustrated example, we show the parameters of one gate's normal distribution with the centers of the ellipsoids $\vec{\mu}_l$ and the shaded areas $\vec{\mu}_l \pm \sigma$ (Supplementary Fig 12e). We estimated the normal distributions using all cells from n=48 samples. For populations with < 2 cells, a covariance matrix was not calculable, so such populations were removed.

In the fifth step, we use the estimated cell population and cell feature distributions to generate realistic flow cytometry datasets. We use the estimated parameters of the Dirichlet distribution and the normal distributions of each leaf node to generate cells. As shown in Supplementary Fig 14, this simulation involves: (a) drawing a vector $\boldsymbol{p} \in \mathbb{R}^K$ from the estimated Dirichlet distribution $Dir(\hat{\alpha}_1, \hat{\alpha}_2, \ldots \hat{\alpha}_K)$, which represents the proportion of cells in each leaf node; (b) calculating the number of synthetic cells per leaf node using the expected number of cells for the sample (e.g. 10,000 cells); and (c) Finally, drawing the required number of synthetic cells from the normal distribution of each corresponding leaf node for each sample. By repeating this process for each sample, we generate a synthetic dataset that reflects the underlying population of cells. We visualize our complete decision tree as an interactive online tool (https://vissim.gunthergl.com/) (Supplementary Website 1).

Realistic imitation of disease associated effects

We can now introduce any given effect in any given cell population and obtain cells from a realistic synthetic sample. For that, we change the underlying Dirichlet distribution and then sample from the existing normal distributions as before. To change the proportion of cell

population x, we have to change its parameter α_x. However, simply changing α_x, e.g. by a factor $f \in \mathbb{R}_{>0}$ ($\alpha'_x := f \cdot \alpha_x$) also changes the precision and therefore the effective change of the population proportion is different than multiplying with f

$$E[p'_x] = \frac{\alpha'_x}{s'} = \frac{f \cdot \alpha_x}{s - \alpha_x + f \cdot \alpha_x} = \frac{f \cdot \alpha_x}{s + (1-f) \cdot \alpha_x} \neq f \frac{\alpha_x}{s} \tag{28}$$

Therefore, we calculate the new α_x by the share of the expected target proportion in the total old precision and the remaining precision is shared across all other nodes

$$\alpha'_x := target_{\%} \cdot s \tag{29}$$

$$\alpha'_{not\ x} := (1 - target_{\%}) \cdot s = \sum_{k \in K \setminus \{x\}} \alpha_k \tag{30}$$

where $not\ x$ corresponds to all nodes which are not the changed node x **nor** subordinate nodes. After a single synthetic cell comes from a specific leaf node distribution, we still have to express the changed intermediate node x by its leaf nodes. After parameter α_k of any node is the sum of all leaf node parameters α_l below node k, we calculate the new leaf node parameter α'_l as the old α_l multiplied with the ratio of the new and old changed node above

$$\alpha'_l = \alpha_l \frac{\alpha'_x}{\alpha_x} \ \ or \ \ \alpha'_l = \alpha_l \frac{\alpha'_{not\ x}}{\alpha_{not\ x}} \tag{31}$$

This finally leads us to a change in the expected proportion of the target population x.

ACKNOWLEDGMENTS

The authors thank The Bristol Myers Squibb Foundation for Immuno-Oncology (Award FA-19-009) and the Bavarian State Ministry for Science and Art's Coronavirus Research Programme for funding this work. We thank Beckman Coulter Life Sciences GmbH for its continuing support of our research. We are grateful to Ian V. Hutchinson (Providence St John's Cancer Institute, CA, USA) for proof-reading our manuscript. This work was only possible with Erika Ostermeier's outstanding technical support.

DATA AVAILABILITY STATEMENT

The authors declare that all data supporting the findings of this study are available within the paper, its supplementary information files and downloadable files deposited at figshare (private reviewer link: https://figshare.com/s/27719825e2afdb5f5a14).

CODE AVAILABILITY STATEMENT

The authors declare that all computer code supporting the findings of this study are available as supplementary information files and downloadable files deposited at figshare (private reviewer link: https://figshare.com/s/27719825e2afdb5f5a14). The python package NBNode is accessible at https://github.com/ggrlab/NBNode. The R package restrictedROC is accessible at https://github.com/ggrlab/restrictedROC.

DISCLOSURES

M.K. is a Beckman Coulter Life Sciences associate. S.H. has received consulting fees and speaker's honoraria from BMS and Merck Sharp & Dohme (MSD). J.A.H. has received consulting fees and speaker's honoraria from Mallinckrodt Pharmaceuticals. Other authors have no conflict-of-interest to declare.

REFERENCES

1 Medzhitov, R. The spectrum of inflammatory responses. *Science* **374**, 1070-1075 (2021).

2 Bartok, E. & Hartmann, G. Immune Sensing Mechanisms that Discriminate Self from Altered Self and Foreign Nucleic Acids. *Immunity* **53**, 54-77 (2020).

3 Du, H. *et al.* Tuning immunity through tissue mechanotransduction. *Nat Rev Immunol* **23**, 174-188 (2023).

4 Rumpret, M., von Richthofen, H. J., Peperzak, V. & Meyaard, L. Inhibitory pattern recognition receptors. *J Exp Med* **219** (2022).

5 Deets, K. A. & Vance, R. E. Inflammasomes and adaptive immune responses. *Nat Immunol* **22**, 412-422 (2021).

6 Willis, J. C. & Lord, G. M. Immune biomarkers: the promises and pitfalls of personalized medicine. *Nat Rev Immunol* **15**, 323-329 (2015).

7 Scheffold, A. & Kern, F. Recent developments in flow cytometry. *J Clin Immunol* **20**, 400-407 (2000).

8 Cossarizza, A. *et al.* Guidelines for the use of flow cytometry and cell sorting in immunological studies (third edition). *Eur J Immunol* **51**, 2708-3145 (2021).

9 Liechti, T. *et al.* An updated guide for the perplexed: cytometry in the high-dimensional era. *Nat Immunol* **22**, 1190-1197 (2021).

10 Maecker, H. T., McCoy, J. P. & Nussenblatt, R. Standardizing immunophenotyping for the Human Immunology Project. *Nat Rev Immunol* **12**, 191-200 (2012).

11 Spidlen, J. *et al.* Data File Standard for Flow Cytometry, Version FCS 3.2. *Cytometry A* **99**, 100-102 (2021).

12 Liechti, T. *et al.* Immune phenotypes that are associated with subsequent COVID-19 severity inferred from post-recovery samples. *Nat Commun* **13**, 7255 (2022).

13 Glehr, G. *et al.* External validation of biomarkers for immune-related adverse events after immune checkpoint inhibition. *Front Immunol* **13**, 1011040 (2022).

14 Das, R. *et al.* Early B cell changes predict autoimmunity following combination immune checkpoint blockade. *J Clin Invest* **128**, 715-720 (2018).

15 Bukhari, S. *et al.* Single-cell RNA sequencing reveals distinct T cell populations in immune-related adverse events of checkpoint inhibitors. *Cell Rep Med* **4**, 100868 (2023).

16 Lozano, A. X. *et al.* T cell characteristics associated with toxicity to immune checkpoint blockade in patients with melanoma. *Nat Med* **28**, 353-362 (2022).

17 Livingstone, E. *et al.* Adjuvant nivolumab plus ipilimumab or nivolumab alone versus placebo in patients with resected stage IV melanoma with no evidence of disease (IMMUNED): final results of a randomised, double-blind, phase 2 trial. *Lancet* **400**, 1117-1129 (2022).

18 Schneider, B. J. *et al.* Management of Immune-Related Adverse Events in Patients Treated With Immune Checkpoint Inhibitor Therapy: ASCO Guideline Update. *J Clin Oncol* **39**, 4073-4126 (2021).

19 Wang, D. Y. *et al.* Fatal Toxic Effects Associated With Immune Checkpoint Inhibitors: A Systematic Review and Meta-analysis. *JAMA Oncol* **4**, 1721-1728 (2018).

20 Conroy, M. & Naidoo, J. Immune-related adverse events and the balancing act of immunotherapy. *Nat Commun* **13**, 392 (2022).

21 Esfahani, K. *et al.* Moving towards personalized treatments of immune-related adverse events. *Nat Rev Clin Oncol* **17**, 504-515 (2020).

22 Ganesan, S. & Mehnert, J. Biomarkers for Response to Immune Checkpoint Blockade. *Annual Review of Cancer Biology* **4**, 331-351 (2020).

23 Maecker, H. T. *et al.* New tools for classification and monitoring of autoimmune diseases. *Nat Rev Rheumatol* **8**, 317-328 (2012).

24 Fox, A. *et al.* Cyto-Feature Engineering: A Pipeline for Flow Cytometry Analysis to Uncover Immune Populations and Associations with Disease. *Sci Rep* **10**, 7651 (2020).

25 Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636-645 (2015).

26 Brummelman, J. *et al.* Development, application and computational analysis of high-dimensional fluorescent antibody panels for single-cell flow cytometry. *Nat Protoc* **14**, 1946-1969 (2019).

27 Hu, Z., Tang, A., Singh, J., Bhattacharya, S. & Butte, A. J. A robust and interpretable end-to-end deep learning model for cytometry data. *Proc Natl Acad Sci U S A* **117**, 21373-21380 (2020).

28 Martins, F. *et al.* Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance. *Nat Rev Clin Oncol* **16**, 563-580 (2019).

29 McKean, W. B., Moser, J. C., Rimm, D. & Hu-Lieskovan, S. Biomarkers in Precision Cancer Immunotherapy: Promise and Challenges. *Am Soc Clin Oncol Educ Book* **40**, e275-e291 (2020).

30 Lakshmikanth, T. *et al.* Human Immune System Variation during 1 Year. *Cell Rep* **32**, 107923 (2020).

31 Harrington, C. *et al.* Noninvasive biomarkers for the diagnosis and management of autoimmune hepatitis. *Hepatology* **76**, 1862-1879 (2022).

32 Schilling, H. L. *et al.* Development of a Flow Cytometry Assay to Predict Immune Checkpoint Blockade-Related Complications. *Front Immunol* **12**, 765644 (2021).

33 Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29-36 (1982).

34 Gneiting, T. & Vogel, P. Receiver operating characteristic (ROC) curves: equivalences, beta model, and minimum distance estimation. *Machine Learning* **111**, 2147-2159 (2022).

35 Hutchinson, J. A. *et al.* Virus-specific memory T cell responses unmasked by immune checkpoint blockade cause hepatitis. *Nat Commun* **12**, 1439 (2021).

36 Glehr, G. Restriction method for marker discovery in R. R package version 2.3.4. https://github.com/ggrlab/restrictedROC (2023).

37 Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861-874 (2006).

38 Good, P. *Permutation Tests*. 2 edn, (Springer New York, 2000).

39 Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).

40 Burke, K. P., Grebinoski, S., Sharpe, A. H. & Vignali, D. A. A. Understanding adverse events of immunotherapy: A mechanistic perspective. *J Exp Med* **218** (2021).

41 Kverneland, A. H. *et al.* Age and gender leucocytes variances and references values generated using the standardized ONE-Study protocol. *Cytometry A* **89**, 543-564 (2016).

42 Arvaniti, E. & Claassen, M. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat Commun* **8**, 14825 (2017).

43 Roederer, M. How many events is enough? Are you positive? *Cytometry A* **73**, 384-385 (2008).

44 Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

45 Dodd, L. E. & Pepe, M. S. Partial AUC estimation and regression. *Biometrics* **59**, 614-623 (2003).

46 Parodi, S. *et al.* Restricted ROC curves are useful tools to evaluate the performance of tumour markers. *Stat Methods Med Res* **25**, 294-314 (2016).

47 Yang, H., Lu, K., Lyu, X. & Hu, F. Two-way partial AUC and its properties. *Stat Methods Med Res* **28**, 184-195 (2019).

48 McClish, D. K. Analyzing a portion of the ROC curve. *Med Decis Making* **9**, 190-195 (1989).

49 Bhattacharya, B. & Hughes, G. Symmetry of receiver operating characteristic curves and Kullback–Leibler divergences between the signal and noise populations. *Journal of Mathematical Psychology* **55**, 365-367 (2011).

50 Cron, A. *et al.* Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. *PLoS Comput Biol* **9**, e1003130 (2013).

51 Johnsson, K., Wallin, J. & Fontes, M. BayesFlow: latent modeling of flow cytometry cell populations. *BMC Bioinformatics* **17**, 25 (2016).

52 Boris, P. H., Chariff, A., Raphael, G., François, C. & Rodolphe, T. Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data. *The Annals of Applied Statistics* **13**, 638-660 (2019).

53 Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32-35 (1950).

54 Cassel, C. K. & Guest, J. A. Choosing wisely: helping physicians and patients make smart decisions about their care. *JAMA* **307**, 1801-1802 (2012).

55 Perez-Ruiz, E. *et al.* Prophylactic TNF blockade uncouples efficacy and toxicity in dual CTLA-4 and PD-1 immunotherapy. *Nature* **569**, 428-432 (2019).

56 Pinal-Fernandez, I. *et al.* Transcriptomic profiling reveals distinct subsets of immune checkpoint inhibitor induced myositis. *Ann Rheum Dis* (2023).

57 De Martin, E. *et al.* Characterization of liver injury induced by cancer immunotherapy using immune checkpoint inhibitors. *J Hepatol* **68**, 1181-1190 (2018).

58 Groha, S. *et al.* Germline variants associated with toxicity to immune checkpoint blockade. *Nat Med* **28**, 2584-2591 (2022).

59 Khan, Z. *et al.* Genetic variation associated with thyroid autoimmunity shapes the systemic immune response to PD-1 checkpoint blockade. *Nat Commun* **12**, 3355 (2021).

60 McQuade, J. L. *et al.* Association of Body Mass Index With the Safety Profile of Nivolumab With or Without Ipilimumab. *JAMA Oncol* **9**, 102-111 (2023).

61 McCulloch, J. A. *et al.* Intestinal microbiota signatures of clinical response and immune-related adverse events in melanoma patients treated with anti-PD-1. *Nat Med* **28**, 545-556 (2022).

62 McShane, L. M. *et al.* Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst* **97**, 1180-1184 (2005).

63 Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* **45**, 228-247 (2009).

64 Kronenberg, K., Riquelme, P. & Hutchinson, J. A. Standard protocols for immune profiling of peripheral blood leucocyte subsets by flow cytometry using DuraClone IM reagents. *Protocol Exchange* (2019).

65 Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**, 387-415 (1975).

66 Sullivan Pepe, M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. (Oxford University Press, 2004).

67 Yang, Z. *et al.* Optimizing Two-way Partial AUC with an End-to-end Framework. *IEEE Trans Pattern Anal Mach Intell* **PP** (2022).

68 Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18**, 50-60 (1947).

69 Rosner, B. & Glynn, R. J. Power and sample size estimation for the Wilcoxon rank sum test with application to comparisons of C statistics from alternative prediction models. *Biometrics* **65**, 188-197 (2009).

70 LeDell, E. & Poirier, S. Scalable Automatic Machine Learning. *7th ICML Workshop on Automated Machine Learning (AutoML)* (2020).

71 Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, 2009).

72 Glehr, G. Non-binary node and flow cytometry simulation python library. Python package version 1.0.0. https://github.com/ggrlab/NBNode. (2023).

73 Minka, T. Estimating a Dirichlet distribution. (2003).

74 Kahle, D. dirichlet: A light weight package for the (d/r)dirichlet functions for the Dirichlet distribution in R. R package version 0.0.999. https://github.com/dkahle/dirichlet. (2023).

75 Suh, E. J. Dirichlet MLE python library. Python package version v0.9.0. https://github.com/ericsuh/dirichlet. (2023).

FIGURE LEGENDS

FIGURE 1. *Two-class distributions resulting in asymmetric ROC curves*. We present hypothetical and real-world examples of marker distributions in two classes, which are intended to represent sets of patients with different clinical outcomes. The distribution of values from positive (i.e. diseased) class are coloured green and values from negative (i.e. control) class are coloured red; the overlapping density areas are coloured in purple. For each example, we present the corresponding ROC curve. (**a**) A hypothetical example of perfectly discriminatory marker with negative $\mathcal{N}(5,1)$ and positive $\mathcal{N}(15,1)$ populations that give rise to a symmetrical ROC curve. The area under the ROC curve (AUC) is 1.0. (**b**) A hypothetical example of substantially overlapping marker distributions in the negative $\mathcal{N}(5,1)$ and positive $\mathcal{N}(6,1)$ populations that gives rise to a symmetrical ROC curve with AUC=0.76. (**c**) A hypothetical example of a non-informative marker distribution with negative $\mathcal{N}(5,1)$ and positive $\mathcal{N}(5,1)$ populations that gives rise to a diagonal ROC curve with AUC = 0.5. (**d**) A hypothetical example of substantially overlapping marker distributions in the negative $\mathcal{N}(5,1)$ and positive $\mathcal{N}(6,2)$ populations with unequal variance that gives rise to right-skewed ROC curve with AUC = 0.67. (**e**) A real-world example of substantially overlapping distributions of absolute erythrocyte counts with unequal variance in patients with metastatic melanoma who responded (n = 61) or did not respond (n = 44) to combined Ipilimumab and Nivolumab (Ipi-Nivo) therapy. We observe a right-skewed ROC curve with AUC = 0.62. **(f)** A hypothetical example of substantially overlapping marker distributions in the negative $\mathcal{N}(5,2)$ and positive $\mathcal{N}(6,1)$ populations with unequal variance that gives rise to a left-skewed ROC curve with AUC = 0.67. (**g**) A real-world example of substantially overlapping distributions of $CD8^+$ γδ T cells with unequal variance in patients with metastatic melanoma who did (n = 22) or did not (n = 42) develop treatment-related hepatitis after Ipi-Nivo therapy. We observe a left-skewed

ROC curve with AUC = 0.69. **(h)** A hypothetical example of substantially overlapping marker distributions in the normally distributed negative $\mathcal{N}(5,1)$ and bimodally distributed positive populations. In this example, the positive population comprises 10 % cases with elevated marker expression $\mathcal{N}(9,1)$ and 90 % cases with unaltered marker expression $\mathcal{N}(5,1)$. Phenotypic heterogeneity in the diseased cases gives rise to a right-skewed ROC curve with AUC = 0.55. (**i**) A real-world example of a phenotypically heterogeneous set of patients with metastatic melanoma who did (n = 48) or did not (n=62) develop treatment-related hepatitis after Ipi-Nivo therapy. As previously described, a subset of these patients exhibited a baseline expansion of $CD4^+$ T_{EM} cells, which was likely driven by subclinical cytomegalovirus (CMV) reactivation. Consequently, $CD4^+$ T_{EM} cell frequency before therapy is a weakly discriminatory marker of hepatitis-risk that gives rise to a right-skewed ROC curve with AUC = 0.64.

FIGURE 2. *Method to optimally restrict datasets to classifiable samples*. We present a simulated example of marker distributions in two classes, which are intended to represent sets of patients with different clinical outcomes. **(a)** The distribution of values from positive (i.e. diseased, n=2500) class are coloured green and values from negative (i.e. control, n=2500) class are coloured red; the overlapping density areas are coloured in purple. In this example, 20 % of positive samples and 2 % of negative samples were drawn from a population with elevated marker expression $\mathcal{N}(9,1)$. All other samples were drawn from a population with unaltered marker expression $\mathcal{N}(6,1)$. The optimal restriction of this dataset lies at a marker value of 6.8, which is marked with a red line. Restriction of the dataset defines two subsets of samples – explicitly, $\text{marker}^{\text{HIGH}}$ (orange) and $\text{marker}^{\text{LOW}}$ (blue) samples. **(b)** A complete ROC curve marked at the optimal restriction point (red lines) that corresponds to FPR = 0.258. Restricting the parts of the ROC curve corresponding to $\text{marker}^{\text{HIGH}}$ or $\text{marker}^{\text{LOW}}$ samples gives us restricted ROC curves for which restricted AUCs (rAUCs) can be calculated. This is equivalent to rescaling the respective part of the ROC curve. **(c)** Adjusting the rAUC for the number of samples delimited by the restriction gives the restricted standardized AUC (rzAUC). Hence, we can plot rzAUC for $\text{marker}^{\text{HIGH}}$ and $\text{marker}^{\text{LOW}}$ samples at all possible restriction values. The optimal restriction value is defined as the maximum absolute rzAUC for either the $\text{marker}^{\text{HIGH}}$ or $\text{marker}^{\text{LOW}}$ samples. **(d)** A complete ROC curve to illustrate the delimitation of $\text{marker}^{\text{HIGH}}$ values (orange rectangle) according to the optimal restriction. **(e)** Densities of the negative and positive classes after restriction to $\text{marker}^{\text{HIGH}}$ values. **(f)** ROC curve constructed from $\text{marker}^{\text{HIGH}}$ samples. Intuitively, we see that recalculating the ROC curve using only $\text{marker}^{\text{HIGH}}$ samples is equivalent to rescaling the partial ROC curve. **(g)** A complete ROC curve to illustrate the delimitation of $\text{marker}^{\text{LOW}}$ values (blue rectangle) according to the optimal restriction. **(h)**

Densities of the negative and positive classes after restriction to $marker^{LOW}$ values. **(i)** ROC curve constructed from $marker^{LOW}$ samples.

FIGURE 3. *Optimal restriction of two-class distributions results in asymmetric ROC curves.* We present four simulated examples of marker distributions in two classes, which are intended to represent sets of patients with different clinical outcomes. The distribution of values from the positive (i.e. diseased) class are coloured green and values from the negative (i.e. control) class are coloured red; the overlapping density areas are coloured in purple. For each example, we present the following: a plot of positive and negative class densities; the complete ROC curve; a plot of marker values against FPR; a plot of rzAUC calculated for $\text{marker}^{\text{HIGH}}$ (orange) and $\text{marker}^{\text{LOW}}$ (blue) samples at all FPR values. In each plot, red lines indicate the optimal restriction as a marker value or FPR value. **(a)** A simulated example of a symmetric ROC curve from 100 negative $\mathcal{N}(5,1)$ and 100 positive $\mathcal{N}(6,1)$ samples. **(b)** A simulated example of a right-skewed ROC curve from 100 negative $\mathcal{N}(5,1)$ and 100 positive $\mathcal{N}(6,2)$ samples. **(c)** A simulated example of a left-skewed ROC curve from 100 negative $\mathcal{N}(5,2)$ and 100 positive $\mathcal{N}(6,1)$ samples. **(d)** Results for a right-skewed ROC curve from 100 negative $\mathcal{N}(5,1)$ samples and 100 positive samples from a bimodally distributed positive population. In this example, the positive population comprises 10 % cases with elevated marker expression $\mathcal{N}(9,1)$ and 90 % cases with unaltered marker expression $\mathcal{N}(5,1)$.

Figure 4. *Realistically synthesizing flow cytometry data from two class distributions*. Various applications of our synthetic flow cytometry data depend upon generating samples with differences in cell subset distributions. Here, we provide an example of increasing the proportion of $CD8^+$ T_{EMRA} cells in synthetic samples from a baseline value of 7.17 % in the negative class to an altered value of 33.23 % in the positive class. The intensity of red shading in the gating trees illustrates this change in $CD8^+$ T_{EMRA} cells and contingent changes in other populations. **(a)** Gating tree with 7.17% $CD8^+$ T_{EMRA} cells, its corresponding Dirichlet distribution parameter ($\alpha = 12.34$) and the density for three example gates: A, B and L. The cell count table for three samples drawn from this distribution is shown. **(b)** Gating tree with 33.23% $CD8^+$ T_{EMRA} cells, its corresponding Dirichlet distribution parameter ($\alpha = 57.18$) and the density for three example gates: A, B and L. The cell count table for three samples drawn from this distribution is shown. Of special note, percentages of cells in all other gates also changed according to the Dirichlet distribution, leading to changes in simulated cell counts across all leaf gates.

FIGURE 5. *Restriction of synthetheized flow cytometry datasets*. We present examples of marker distributions in two classes, which are intended to represent sets of patients with different clinical outcomes. The distribution of values from positive (i.e. diseased) class are coloured green and values from negative (i.e. control) class are coloured red; overlapping density areas are coloured in purple. For each example, we present the following: a plot of positive and negative class densities; the complete ROC curve; a plot of marker values against FPR; a plot of rzAUC calculated for $\text{marker}^{\text{HIGH}}$ (orange) and $\text{marker}^{\text{LOW}}$ (blue) samples at all FPR values. In each plot, red lines indicate the optimal restriction as a marker value or FPR value. **(a)** A synthetic example of a symmetrical ROC curve from 100 negative $\mathcal{N}(7.7, 1)$ and 100 positive $\mathcal{N}(10.7, 1)$ samples. **(b)** A synthetic example of a right-skewed ROC curve from from 100 negative $\mathcal{N}(7.7, 1)$ and 100 positive $\mathcal{N}(8.7, 3)$ samples. **(c)** A synthetic example of a left-skewed ROC curve from 100 negative $\mathcal{N}(7.7, 3)$ and 100 positive $\mathcal{N}(8.7, 1)$ samples. **(d)** Results for a synthetic right-skewed ROC curve from 100 negative $\mathcal{N}(7.7, 1)$ samples and 100 positive samples from a bimodally distributed positive population. In this example, the positive population comprises 20 % cases with elevated marker expression $\mathcal{N}(16.7, 1)$ and 80 % cases with unaltered marker expression $\mathcal{N}(7.7, 1)$.

FIGURE 6. *Restriction aids marker discovery in synthetic and real-world flow cytometry datasets*. Using synthetic and real-world datasets, we demonstrate that dataset restriction augments discovery of novel immune markers with discriminatory capacity over a limited range of values. In each plot, the x-axis shows permutation p-values for the AUC of complete ROC curves for every gated cell population; the y-axis shows permutation p-values for the AUC of optimally restricted ROC curves. Points within the green-shaded rectangles represent cell subsets for which p-values derived from unrestricted data are not significant ($p \geq 0.05$), but p-values derived from optimally restricted data are significant ($p < 0.05$). Named subsets falling in the green-shaded area are disease-related markers discovered exclusively through dataset restriction. **(a)** Permutation p-values from synthetic samples in which a disease-related effect was introduced into $CD4^+$ T_{EM} resulting in a symmetric ROC curve. 100 samples in the negative $\mathcal{N}(7.7, 1)$ class and 100 samples in the positive $\mathcal{N}(10.7, 1)$ class were generated. **(b)** Permutation p-values from synthetic samples in which a disease-related effect was introduced into $CD4^+$ T_{EM} resulting in a right-skewed ROC curve. 100 samples in the negative $\mathcal{N}(7.7, 1)$ class and 100 samples in the positive $\mathcal{N}(8.7, 3)$ class were generated. **(c)** Permutation p-values from synthetic samples in which a disease-related effect was introduced into $CD4^+$ T_{EM} resulting in a left-skewed ROC curve. 100 samples in the negative $\mathcal{N}(7.7, 3)$ class and 100 samples in the positive $\mathcal{N}(8.7, 1)$ class were generated. **(d)** Permutation p-values from synthetic samples in which a disease-related effect was introduced into $CD4^+$ T_{EM} resulting in a right-skewed ROC curve. 100 samples in the negative $\mathcal{N}(7.7, 1)$ class and 100 samples from a bimodally distributed positive class were generated. In this example, the positive population comprises 20 % cases with elevated marker expression $\mathcal{N}(16.7, 1)$ and 80% cases with unaltered marker expression $\mathcal{N}(7.7, 1)$. **(e)** Permutation p-values from a training set of real-world clinical flow cytometry samples (n=110). 84 markers

were selected where ≥ 10 % of samples had more than 10 counts. Dataset restriction reveals 4 previously undescribed markers of treatment-related colitis risk in metastatic melanoma patients receiving Ipi-Nivo therapy. **(f)** Permutation p-values from a training set of real-world clinical flow cytometry samples (n=110). 84 markers were selected where ≥ 10 % of samples had more than 10 counts. Dataset restriction reveals 7 previously undescribed markers of treatment-related hepatitis risk in metastatic melanoma patients receiving Ipi-Nivo therapy.

FIGURE 7. *Clinical interpretation of restricted markers in predicting disease.* Our method of dataset restriction leads to counterintuitive clinical interpretations of marker values. This is illustrated by our discovery of $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} cells as a univariate marker of hepatitis risk after immunotherapy. Here, we illustrate the conventional evaluation of marker performance across all samples with evaluation of marker performance in a restricted dataset. **(a)** Densities of $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} cells in all samples from patients with metastatic melanoma who developed hepatitis (n = 48) or did not (n = 62) after starting Ipi-Nivo therapy. **(b)** Following the classical approach of determing a classification cut-off for $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} frequency relative to $CD4^+$ T cells using the Youden Index, we predict hepatitis if > 9.62% and then assess the correct classification rate (CCR), negative predictive value (NPV), positive predictive value (PPV), sensitivity (or true positive rate, TPR) and specificity (or true negative rate, TNR) for all samples. **(c)** Our restriction method is predicated on there being a range of values over which a marker provides no discriminatory information. Optimally restricting $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} cell values leads us to discard 58 of 110 samples as "unclassifiable." For the remaining 42 samples where $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} frequency relative to $CD4^+$ T cells > 7.62% , we determine a classification cut-off using the Youden Index, again predicting hepatitis if > 9.62%. Accordingly, we obtain a confusion table with CCR = 76.9%, specificity = 64.3% sensitivity = 91.7%, PPV = 68.8% and NPV = 90% across the classifiable samples.

Figure 8. *Significant markers according to classical and our analysis predicting hepatitis*. **(a)** Heatmap showing significant markers of hepatitis risk after Ipi-Nivo therapy based upon permutation p-values for unrestricted and restricted AUC. Only permutation p-values for markers discovered using our restriction method remained significant after correction for multiple testing, indicated by marker names in red text. Each further column reflects one sample, each row a feature. The samples are grouped into patients who did (green) or did not (red) develop treatment-related hepatitis, shown in the very first row. The main matrix consists of three values: Those excluded according by restriction (white); those included and predicted positive (dark green); and those included and predicted negative (dark red). Columns were clustered, rows in increasing order according to the number of excluded samples. **(b)** Random forest predictions and performances on the prospective validation cohort (n=30) trained on unrestricted marker values from the 110 training samples. **(c)** Random forest model predictions and performances on the prospective validation cohort (n=30) trained on restricted marker values from the 110 training samples. In this case, restricted marker values were replaced with -1 before training the random forest.

Figures

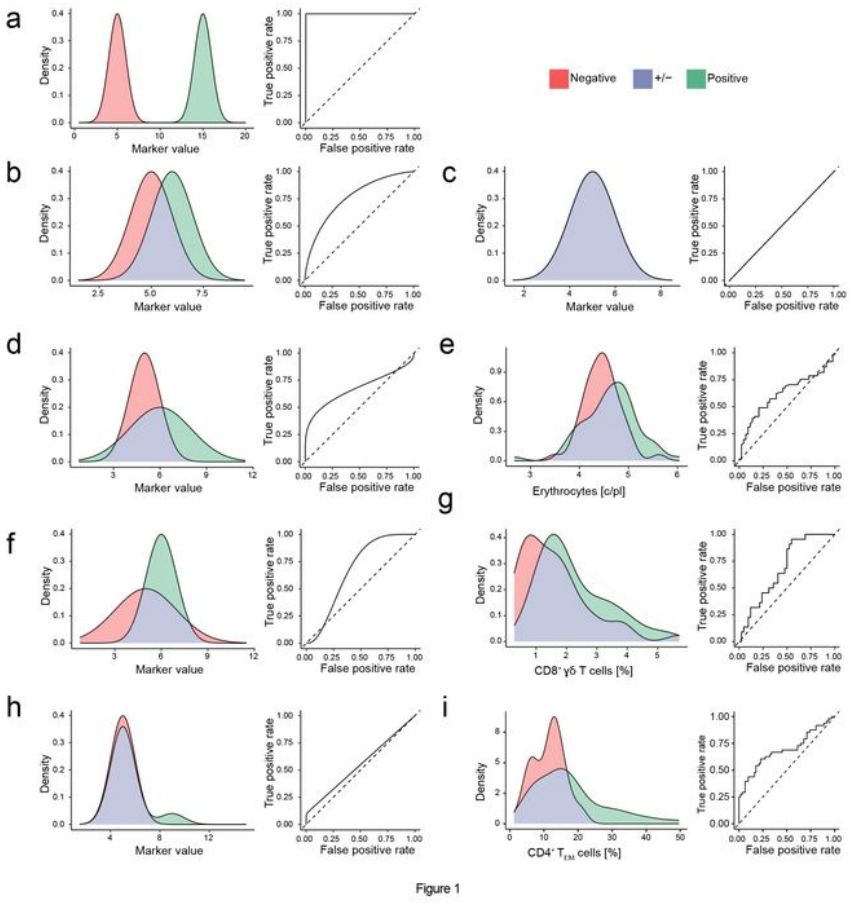

Figure 1

FIGURE 1. *Two-class distributions resulting in asymmetric ROC curves.* We present hypothetical and real-world examples of marker distributions in two classes, which are intended to represent sets of patients with different clinical outcomes. The distribution of values from positive (i.e. diseased) class are coloured green and values from negative (i.e. control) class are coloured *red*; the overlapping density areas are coloured in purple. For each example, we present the corresponding ROC curve. (**a**) A hypothetical example of perfectly discriminatory marker with negative $N(5, 1)$ and positive $N(15, 1)$ populations that give rise to a symmetrical ROC curve. The area under the ROC curve (AUC) is 1.0. (**b**) A hypothetical example of substantially overlapping marker distributions in the negative $N(5, 1)$ and positive $N(6, 1)$ populations that gives rise to a symmetrical ROC curve with AUC=0.76. (**c**) A hypothetical example of a non-informative marker distribution with negative $N(5, 1)$ and positive $N(5, 1)$ populations that gives rise to a diagonal ROC curve with AUC = 0.5. (**d**) A hypothetical example of substantially overlapping marker distributions in the negative $N(5, 1)$ and positive $N(6, 2)$ populations with unequal variance that gives rise to right-skewed ROC curve with AUC = 0.67. (**e**) A real-world example of substantially overlapping distributions of absolute erythrocyte counts with unequal variance in patients with metastatic melanoma who responded (n = 61) or did not respond (n = 44) to combined Ipilimumab and Nivolumab (Ipi-Nivo) therapy. We observe a right-skewed ROC curve with AUC = 0.62. (**f**) A hypothetical example of substantially overlapping marker distributions in the negative $N(5, 2)$ and positive $N(6, 1)$ populations with unequal variance that gives rise to a left-skewed ROC curve with AUC = 0.67. (**g**) A real-world example of substantially overlapping distributions of $CD8^+$ γδ T cells with unequal variance in patients with metastatic melanoma who did (n = 22) or did not (n = 42) develop treatment-related hepatitis after Ipi-Nivo therapy. We observe a left-skewed ROC curve with AUC = 0.69. (**h**) A hypothetical example of substantially overlapping marker distributions in the normally distributed negative $N(5, 1)$ and bimodally distributed positive populations. In this example, the positive population comprises 10 % cases with elevated marker expression $N(9, 1)$ and 90 % cases with unaltered marker expression $N(5, 1)$. Phenotypic heterogeneity in the diseased cases gives rise to a right-skewed ROC curve with AUC = 0.55. (**i**) A real-world example of a phenotypically heterogeneous set of patients with metastatic melanoma who did (n = 48) or did not (n=62) develop treatment-related hepatitis after Ipi-Nivo therapy. As previously described, a subset of these patients exhibited a baseline expansion of $CD4^+$ T_{EM} cells, which was likely driven by subclinical cytomegalovirus (CMV) reactivation. Consequently, $CD4^+$ T_{EM} cell frequency before therapy is a weakly discriminatory marker of hepatitis-risk that gives rise to a right-skewed ROC curve with AUC = 0.64.

Figure 1

See image above for figure legend.

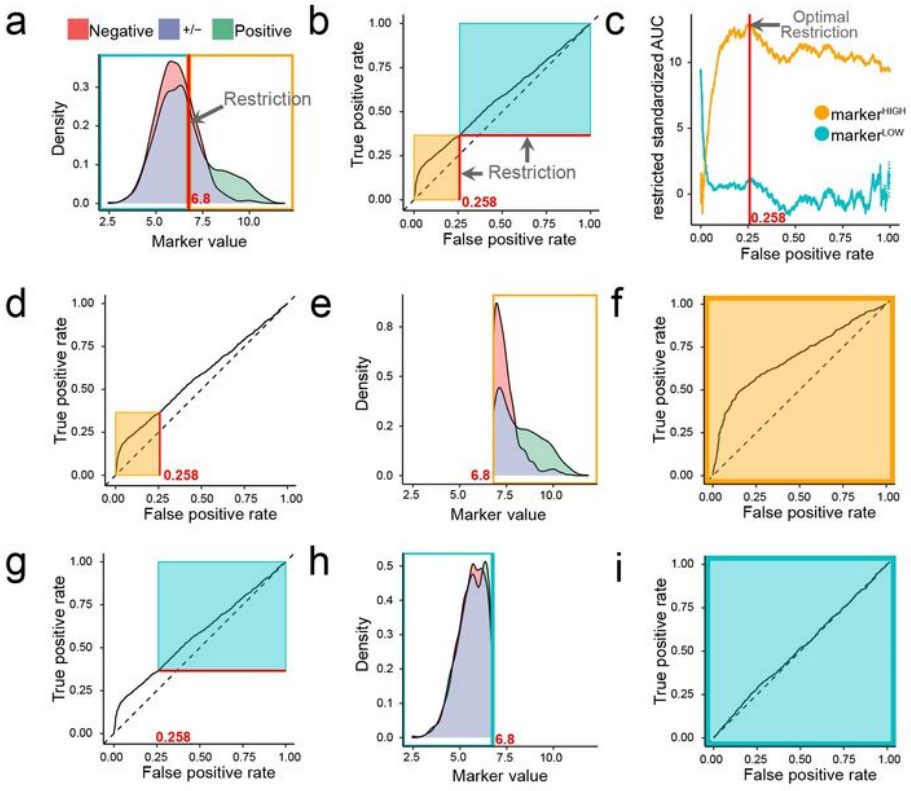

Figure 2

FIGURE 2. *Method to optimally restrict datasets to classifiable samples.* We present a simulated example of marker distributions in two classes, which are intended to represent sets of patients with different clinical outcomes. **(a)** The distribution of values from positive (i.e. diseased, n=2500) class are coloured green and values from negative (i.e. control, n=2500) class are coloured red; the overlapping density areas are coloured in purple. In this example, 20 % of positive samples and 2 % of negative samples were drawn from a population with elevated marker expression $\mathcal{N}(9,1)$. All other samples were drawn from a population with unaltered marker expression $\mathcal{N}(6,1)$. The optimal restriction of this dataset lies at a marker value of 6.8, which is marked with a red line. Restriction of the dataset defines two subsets of samples – explicitly, $\text{marker}^{\text{HIGH}}$ (orange) and $\text{marker}^{\text{LOW}}$ (blue) samples. **(b)** A complete ROC curve marked at the optimal restriction point (red lines) that corresponds to FPR = 0.258. Restricting the parts of the ROC curve corresponding to $\text{marker}^{\text{HIGH}}$ or $\text{marker}^{\text{LOW}}$ samples gives us restricted ROC curves for which restricted AUCs (rAUCs) can be calculated. This is equivalent to rescaling the respective part of the ROC curve. **(c)** Adjusting the rAUC for the number of samples delimited by the restriction gives the restricted standardized AUC (rzAUC). Hence, we can plot rzAUC for $\text{marker}^{\text{HIGH}}$ and $\text{marker}^{\text{LOW}}$ samples at all possible restriction values. The optimal restriction value is defined as the maximum absolute rzAUC for either the $\text{marker}^{\text{HIGH}}$ or $\text{marker}^{\text{LOW}}$ samples. **(d)** A complete ROC curve to illustrate the delimitation of $\text{marker}^{\text{HIGH}}$ values (orange rectangle) according to the optimal restriction. **(e)** Densities of the negative and positive classes after restriction to $\text{marker}^{\text{HIGH}}$ values. **(f)** ROC curve constructed from $\text{marker}^{\text{HIGH}}$ samples. Intuitively, we see that recalculating the ROC curve using only $\text{marker}^{\text{HIGH}}$ samples is equivalent to rescaling the partial ROC curve. **(g)** A complete ROC curve to illustrate the delimitation of $\text{marker}^{\text{LOW}}$ values (blue rectangle) according to the optimal restriction. **(h)** Densities of the negative and positive classes after restriction to $\text{marker}^{\text{LOW}}$ values. **(i)** ROC curve constructed from $\text{marker}^{\text{LOW}}$ samples.

Figure 2

See image above for figure legend.

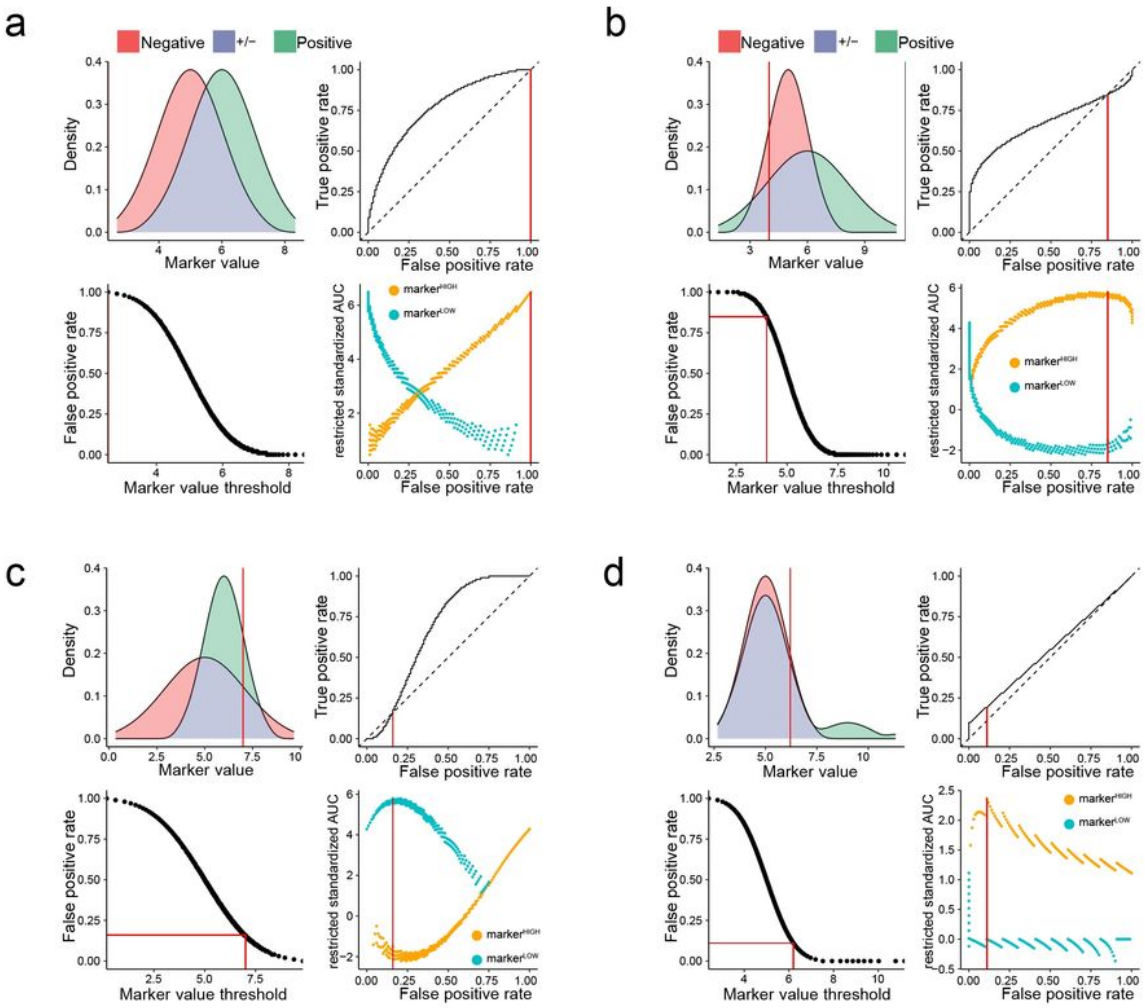

Figure 3

FIGURE 3. *Optimal restriction of two-class distributions results in asymmetric ROC curves.* We present four simulated examples of marker distributions in two classes, which are intended to represent sets of patients with different clinical outcomes. The distribution of values from the positive (i.e. diseased) class are coloured green and values from the negative (i.e. control) class are coloured red; the overlapping density areas are coloured in purple. For each example, we present the following: a plot of positive and negative class densities; the complete ROC curve; a plot of marker values against FPR; a plot of rzAUC calculated for $\text{marker}^{\text{HIGH}}$ (orange) and $\text{marker}^{\text{LOW}}$ (blue) samples at all FPR values. In each plot, red lines indicate the optimal restriction as a marker value or FPR value. **(a)** A simulated example of a symmetric ROC curve from 100 negative $\mathcal{N}(5, 1)$ and 100 positive $\mathcal{N}(6, 1)$ samples. **(b)** A simulated example of a right-skewed ROC curve from 100 negative $\mathcal{N}(5, 1)$ and 100 positive $\mathcal{N}(6, 2)$ samples. **(c)** A simulated example of a left-skewed ROC curve from 100 negative $\mathcal{N}(5, 2)$ and 100 positive $\mathcal{N}(6, 1)$ samples. **(d)** Results for a right-skewed ROC curve from 100 negative $\mathcal{N}(5, 1)$ samples and 100 positive samples from a bimodally distributed positive population. In this example, the positive population comprises 10 % cases with elevated marker expression $\mathcal{N}(9, 1)$ and 90 % cases with unaltered marker expression $\mathcal{N}(5, 1)$.

Figure 3

See image above for figure legend.

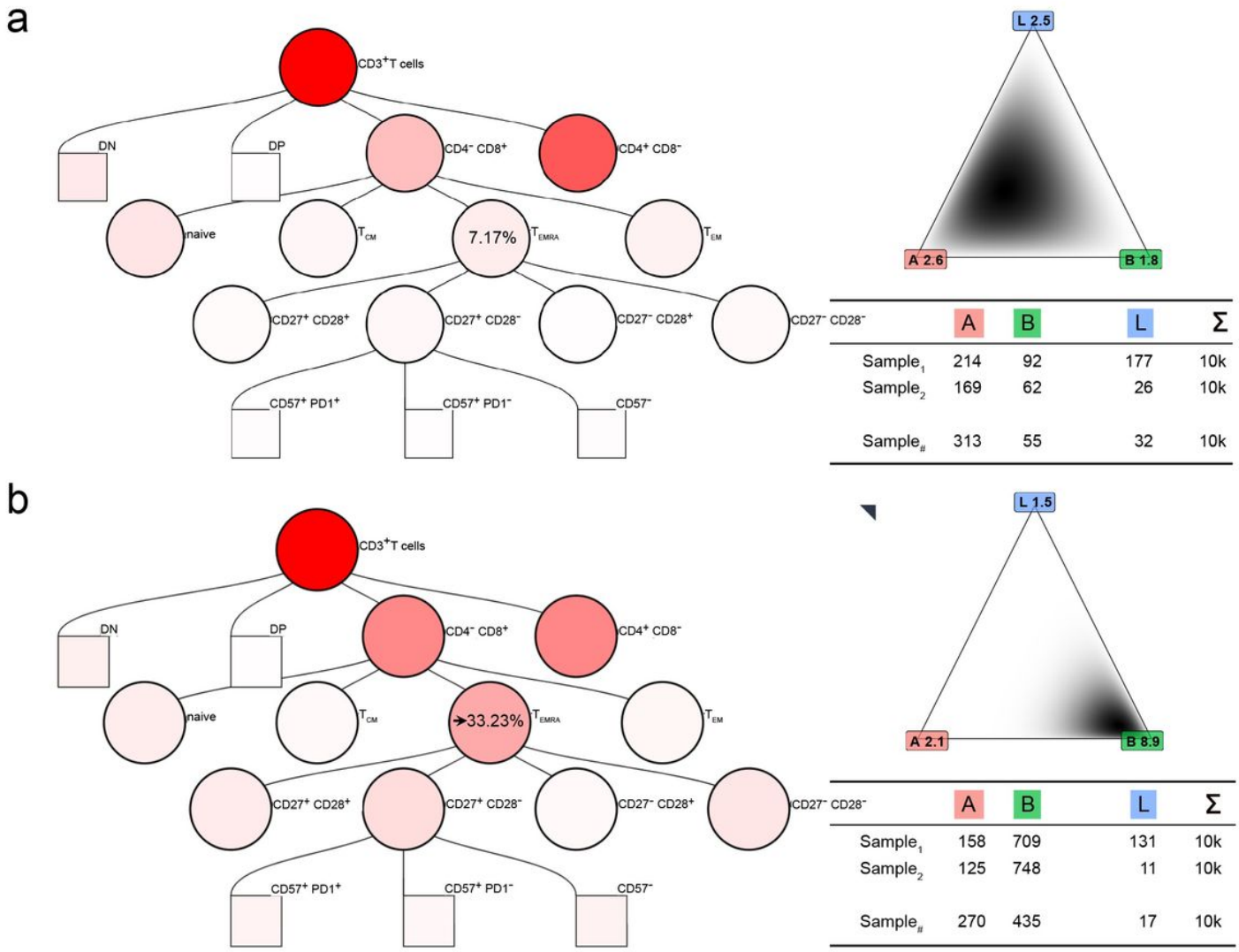

Figure 4

Figure 4

Realistically synthesizing flow cytometry data from two class distributions. Various applications of our synthetic flow cytometry data depend upon generating samples with differences in cell subset distributions. Here, we provide an example of increasing the proportion of $CD8^+$ T_{EMRA} cells in synthetic samples from a baseline value of 7.17 % in the negative class to an altered value of 33.23 % in the positive class. The intensity of red shading in the gating trees illustrates this change in $CD8^+$ T_{EMRA} cells and contingent changes in other populations. **(a)** Gating tree with 7.17% $CD8^+$ T_{EMRA} cells, its corresponding Dirichlet distribution parameter (α = 12.34) and the density for three example gates: A, B and L. The cell count table for three samples drawn from this distribution is shown. **(b)** Gating tree with 33.23% $CD8^+$ T_{EMRA} cells, its corresponding Dirichlet distribution parameter (α = 57.18) and the density for three example gates: A, B and L. The cell count table for three samples drawn from this distribution is

shown. Of special note, percentages of cells in all other gates also changed according to the Dirichlet distribution, leading to changes in simulated cell counts across all leaf gates.

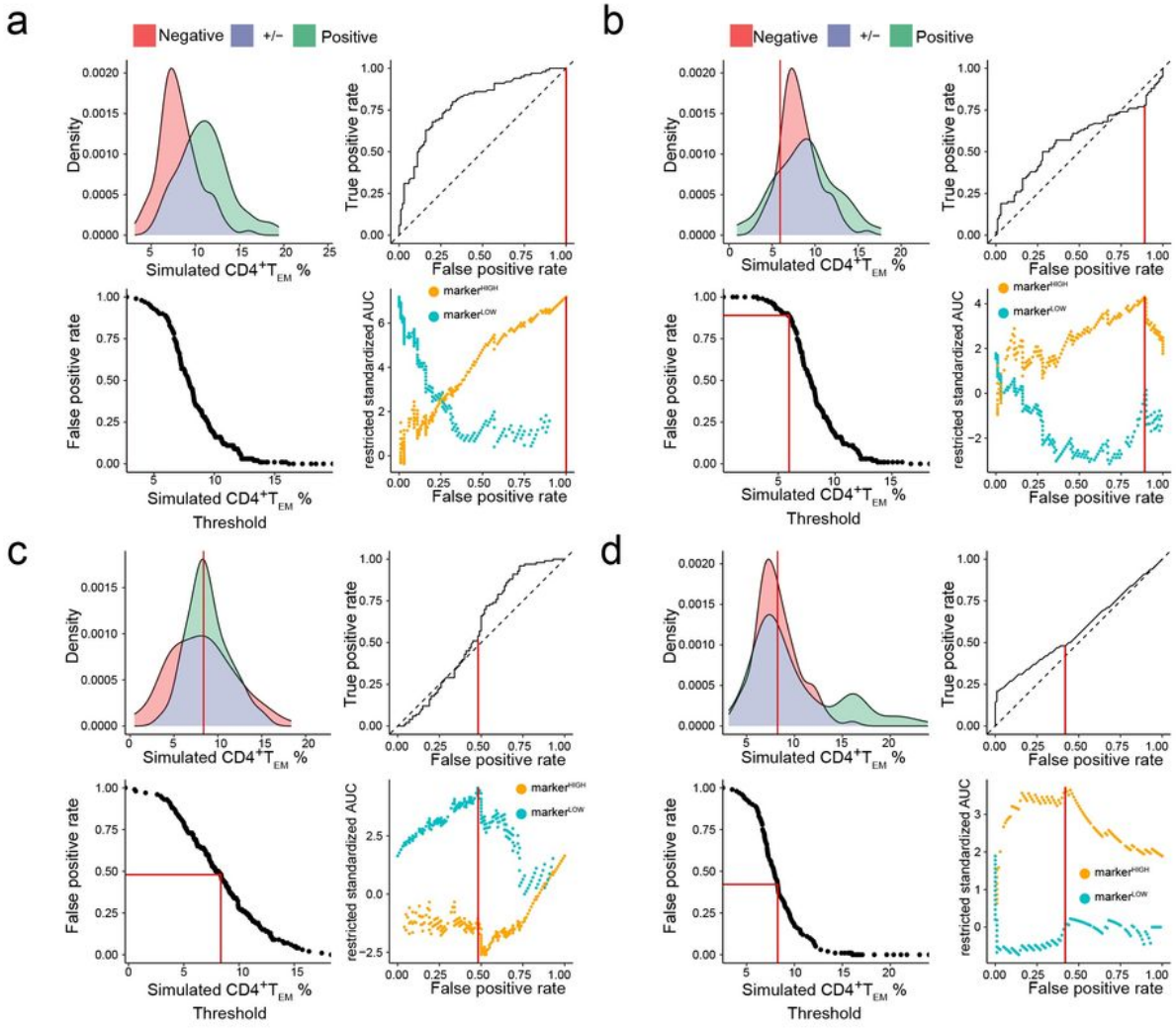

Figure 5

FIGURE 5. *Restriction of synthetheized flow cytometry datasets.* We present examples of marker distributions in two classes, which are intended to represent sets of patients with different clinical outcomes. The distribution of values from positive (i.e. diseased) class are coloured green and values from negative (i.e. control) class are coloured red; overlapping density areas are coloured in purple. For each example, we present the following: a plot of positive and negative class densities; the complete ROC curve; a plot of marker values against FPR; a plot of rzAUC calculated for $\text{marker}^{\text{HIGH}}$ (orange) and $\text{marker}^{\text{LOW}}$ (blue) samples at all FPR values. In each plot, red lines indicate the optimal restriction as a marker value or FPR value. **(a)** A synthetic example of a symmetrical ROC curve from 100 negative $\mathcal{N}(7.7, 1)$ and 100 positive $\mathcal{N}(10.7, 1)$ samples. **(b)** A synthetic example of a right-skewed ROC curve from from 100 negative $\mathcal{N}(7.7, 1)$ and 100 positive $\mathcal{N}(8.7, 3)$ samples. **(c)** A synthetic example of a left-skewed ROC curve from 100 negative $\mathcal{N}(7.7, 3)$ and 100 positive $\mathcal{N}(8.7, 1)$ samples. **(d)** Results for a synthetic right-skewed ROC curve from 100 negative $\mathcal{N}(7.7, 1)$ samples and 100 positive samples from a bimodally distributed positive population. In this example, the positive population comprises 20 % cases with elevated marker expression $\mathcal{N}(16.7, 1)$ and 80 % cases with unaltered marker expression $\mathcal{N}(7.7, 1)$.

Figure 5

See image above for figure legend.

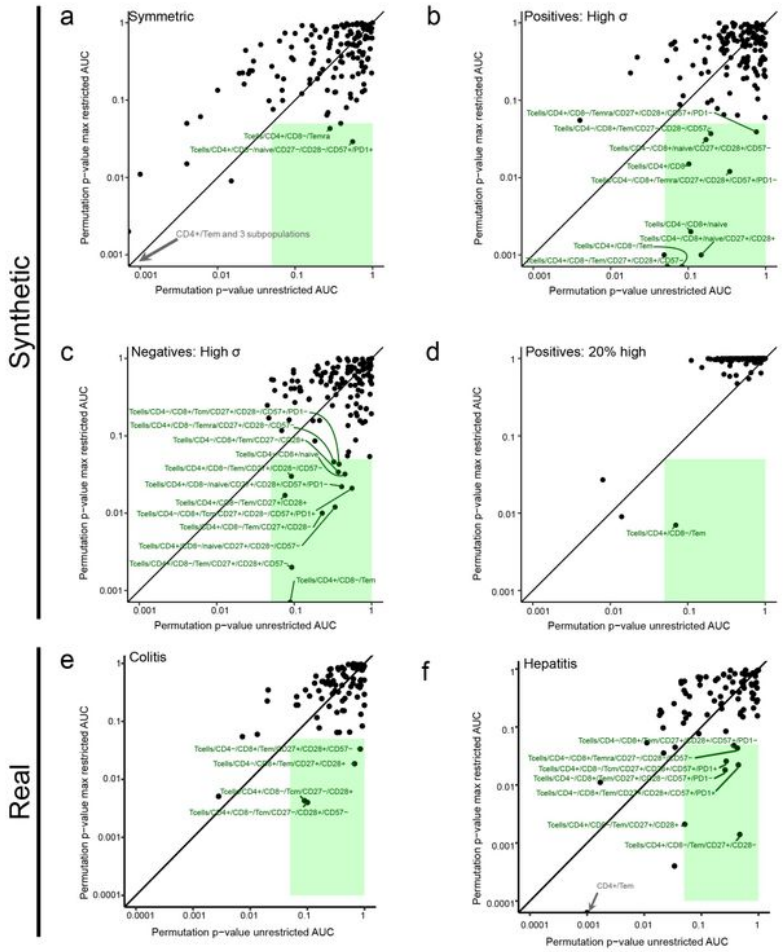

Figure 6

FIGURE 6 *Restriction aids marker discovery in synthetic and real-world flow cytometry datasets.* Using synthetic and real-world datasets, we demonstrate that dataset restriction augments discovery of novel immune markers with discriminatory capacity over a limited range of values. In each plot, the x-axis shows permutation p-values for the *AUC* of complete ROC curves for every gated cell population; the y-axis shows permutation p-values for the *AUC* of optimally restricted ROC curves. Points within the green-shaded rectangles represent cell subsets for which p-values derived from unrestricted data are not significant ($p \geq 0.05$), but p-values derived from optimally restricted data are significant ($p < 0.05$). Named subsets falling in the green-shaded area are disease-related markers discovered exclusively through dataset restriction. **(a)** Permutation p-values from synthetic samples in which a disease-related effect was introduced into $CD4^+$ T_{EM} resulting in a symmetric ROC curve. 100 samples in the negative $N(7.7, 1)$ class and 100 samples in the positive $N(10.7, 1)$ class were generated. **(b)** Permutation p-values from synthetic samples in which a disease-related effect was introduced into $CD4^+$ T_{EM} resulting in a right-skewed ROC curve. 100 samples in the negative $N(7.7, 1)$ class and 100 samples in the positive $N(8.7, 3)$ class were generated. **(c)** Permutation p-values from synthetic samples in which a disease-related effect was introduced into $CD4^+$ T_{EM} resulting in a left-skewed ROC curve. 100 samples in the negative $N(7.7, 3)$ class and 100 samples in the positive $N(8.7, 1)$ class were generated. **(d)** Permutation p-values from synthetic samples in which a disease-related effect was introduced into $CD4^+$ T_{EM} resulting in a right-skewed ROC curve. 100 samples in the negative $N(7.7, 1)$ class and 100 samples from a bimodally distributed positive class were generated. In this example, the positive population comprises 20 % cases with elevated marker expression $N(16.7, 1)$ and 80% cases with unaltered marker expression $N(7.7, 1)$. **(e)** Permutation p-values from a training set of real-world clinical flow cytometry samples (n=110). 84 markers were selected where ≥ 10 % of samples had more than 10 counts. Dataset restriction reveals 4 previously undescribed markers of treatment-related colitis risk in metastatic melanoma patients receiving Ipi-Nivo therapy. **(f)** Permutation p-values from a training set of real-world clinical flow cytometry samples (n=110). 84 markers were selected where ≥ 10 % of samples had more than 10 counts. Dataset restriction reveals 7 previously undescribed markers of treatment-related hepatitis risk in metastatic melanoma patients receiving Ipi-Nivo therapy.

Figure 6

See image above for figure legend.

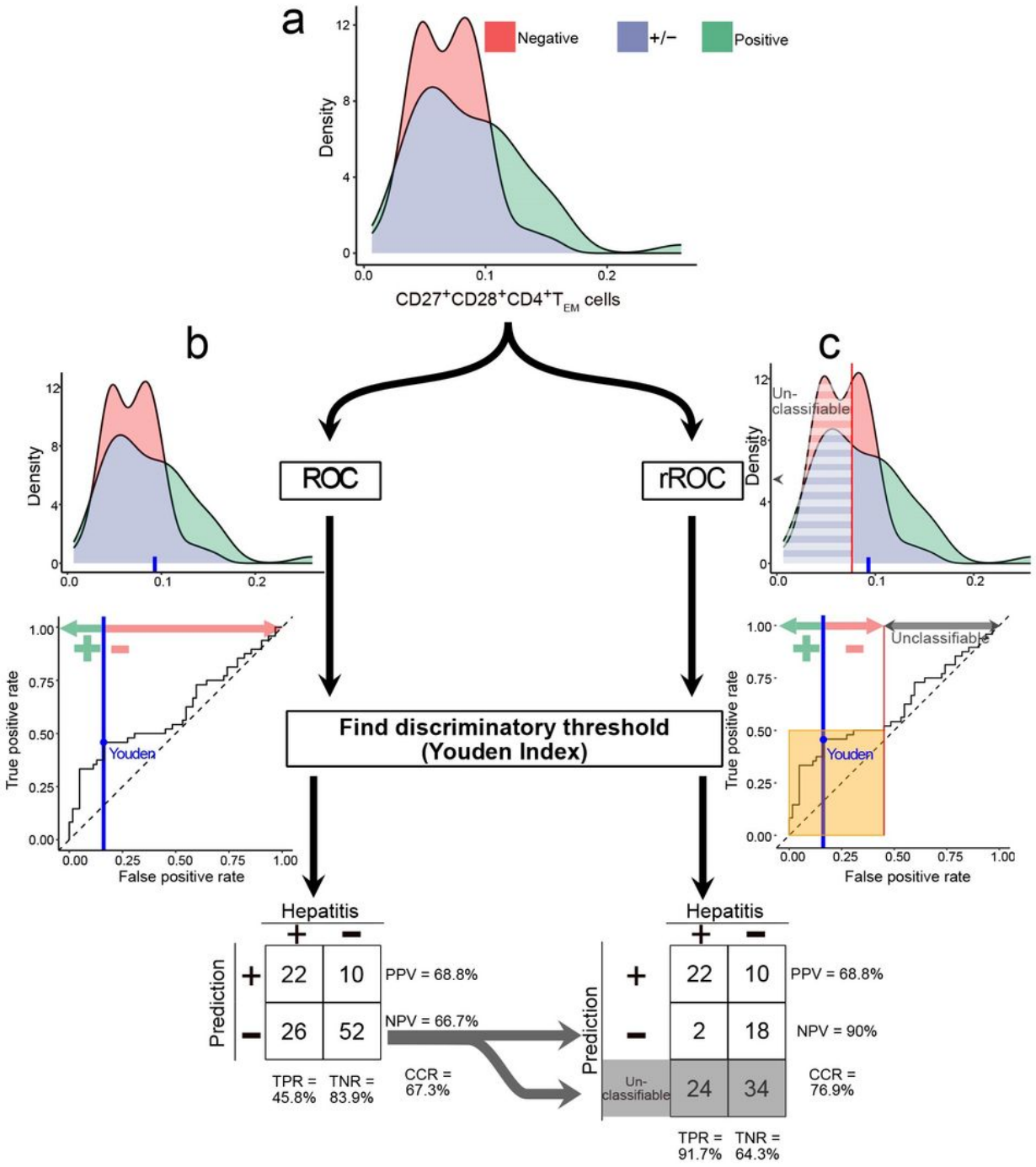

Figure 7

Figure 7

Clinical interpretation of restricted markers in predicting disease. Our method of dataset restriction leads to counterintuitive clinical interpretations of marker values. This is illustrated by our discovery of $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} cells as a univariate marker of hepatitis risk after immunotherapy. Here, we illustrate the conventional evaluation of marker performance across all samples with evaluation of marker

performance in a restricted dataset. **(a)** Densities of $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} cells in all samples from patients with metastatic melanoma who developed hepatitis (n = 48) or did not (n = 62) after starting Ipi-Nivo therapy. **(b)** Following the classical approach of determing a classification cut-off for $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} frequency relative to $CD4^+$ T cells using the Youden Index, we predict hepatitis if > 9.62% and then assess the correct classification rate (CCR), negative predictive value (NPV), positive predictive value (PPV), sensitivity (or true positive rate, TPR) and specificity (or true negative rate, TNR) for all samples. **(c)** Our restriction method is predicated on there being a range of values over which a marker provides no discriminatory information. Optimally restricting $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} cell values leads us to discard 58 of 110 samples as "unclassifiable." For the remaining 42 samples where $CD27^+$ $CD28^+$ $CD4^+$ T_{EM} frequency relative to $CD4^+$ T cells > 7.62% , we determine a classification cut-off using the Youden Index, again predicting hepatitis if > 9.62%. Accordingly, we obtain a confusion table with CCR = 76.9%, specificity = 64.3% sensitivity = 91.7%, PPV = 68.8% and NPV = 90% across the classifiable samples.

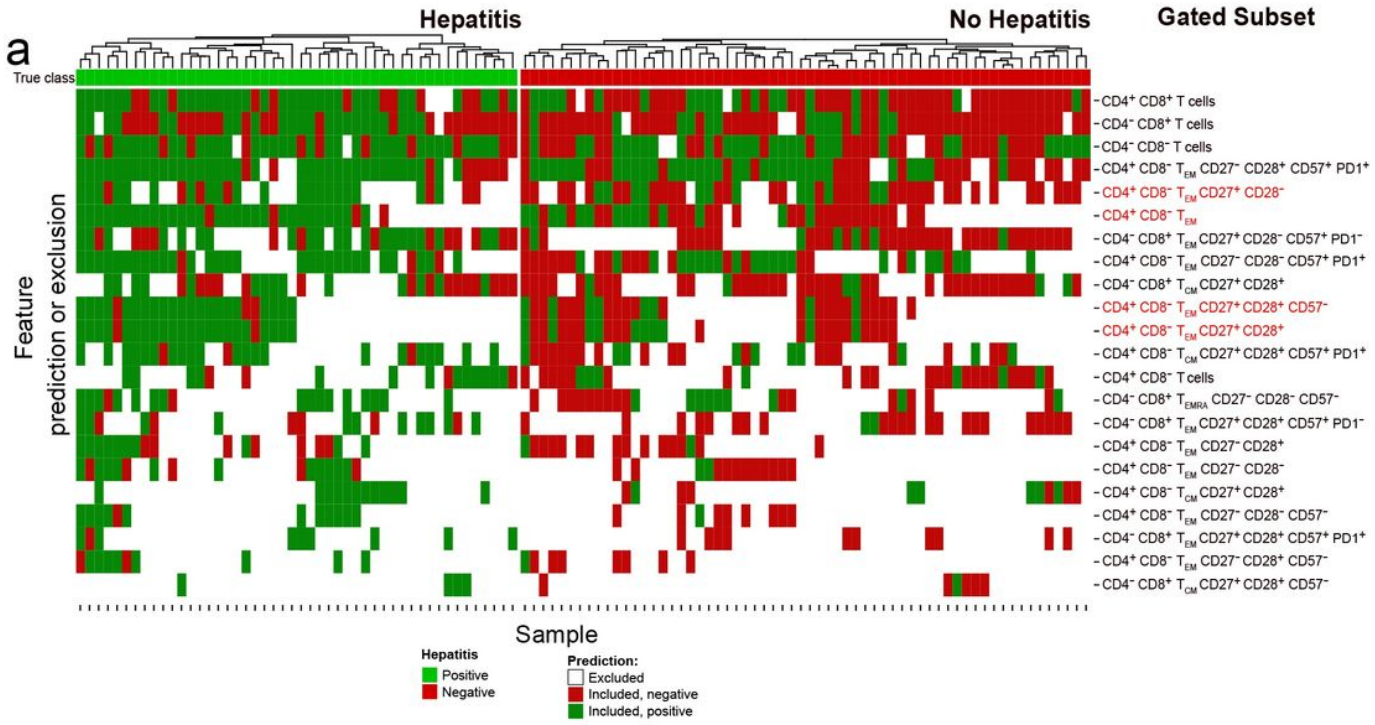

Figure 8

Figure 8

Significant markers according to classical and our analysis predicting hepatitis. **(a)** Heatmap showing significant markers of hepatitis risk after Ipi-Nivo therapy based upon permutation p-values for unrestricted and restricted AUC. Only permutation p-values for markers discovered using our restriction method remained significant after correction for multiple testing, indicated by marker names in red text. Each further column reflects one sample, each row a feature. The samples are grouped into patients who did (green) or did not (red) develop treatment-related hepatitis, shown in the very first row. The main matrix consists of three values: Those excluded according by restriction (white); those included and predicted positive (dark green); and those included and predicted negative (dark red). Columns were clustered, rows in increasing order according to the number of excluded samples. **(b)** Random forest predictions and performances on the prospective validation cohort (n=30) trained on unrestricted marker values from the 110 training samples. **(c)** Random forest model predictions and performances on the prospective validation cohort (n=30) trained on restricted marker values from the 110 training samples. In this case, restricted marker values were replaced with -1 before training the random forest.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Glehr2023ReportingSummaryFINAL.pdf
- Glehr2023SupplementaryMaterialFINAL.pdf
- Glehr2023SoftwarePolicyFINAL.pdf
- Supplementaryvideo1.gif
- Supplementaryvideo2.gif
- Supplementaryvideo3.gif
- Supplementaryvideo4.gif