Novel approach to analysis of the immune system using an ungated model of immune surface marker abundance to predict health outcomes

G. Provost
University of Sherbrooke
F. B. Lavoie
University of Sherbrooke
A. Larbi
Singapore Immunology Network
TP. Ng
Yong Loo Lin School of Medicine, National University Health System, National University of Singapore
C. Tan Tze Ying
Singapore Immunology Network
M. Chua
Singapore Immunology Network
T. Fulop
University of Sherbrooke
A.A. Cohen (Alan.Cohen@USherbrooke.ca)
University of Sherbrooke

Research Article

Keywords: Immunology, neural network, complex system

Posted Date: March 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1401703/v1

Abstract

Traditionally, the immune system is understood to be divided into discrete cell types that are identified via surface markers. While some cell type distinctions are no doubt discrete, others may in fact vary on a continuum, and even within discrete types, differences in surface marker abundance could have functional implications. Here we propose a new way of looking at immune data, which is by looking directly at the values of the surface markers without dividing the cells into different subtypes. To assess the merit of this approach, we compared it with manual gating using cytometry data from the Singapore Longitudinal Aging Study (SLAS) database. We used two different neural networks (one for each method) to predict the presence of several health conditions. We found that the model built using raw surface marker abundance outperformed the manual gating one and we were able to identify some markers that contributed more to the predictions. This study is intended as a brief proof-of-concept and was not designed to predict health outcomes in an applied setting; nonetheless, it demonstrates that alternative methods to understand the structure of immune variation hold substantial progress.

Introduction

In recent decades, aspects of immune function have been linked to the outcomes of an increasing number of medical conditions, including cancer [1], diabetes [2], Alzheimer's disease [3], and cardiovascular diseases [4]. This indicates that the immune system is at the forefront of our fight against not only infectious diseases but also a wide range of other conditions: the immune system communicates with, and is an integral part of, global physiological networks that maintain dynamic equilibrium [5][6], and immune perturbations, dysregulations, and adaptations can have wide-ranging effects [7][8][9]. Accordingly, changes in immune state are a crucial feature of the aging process, and likely can contribute to links between aging and age-related disease. In this context, understanding how immune state interacts with other systems is crucial to a broad understanding of how organisms maintain dynamic equilibrium, and of how changes in immune state might contribute to or mitigate aging processes. Precise and valid measures of age-associated changes in immune state are thus a prerequisite both to a sufficient understanding of immunosenescence and to potential clinical applications, such as identifying appropriate immunotherapy regimes for patients [10]. In turn, such precise and valid measures of immune aging require an appropriate way to characterize the immune system more generally.

The current paradigm in immunology is to consider the immune system as composed of a multitude of specialized cells which can be classified into discrete types, each having precise roles in the defense of the organism [11][12][13]. The main way to analyze those cell types and their interactions is by flow cytometry because it allows for the classification by type and quantification of a large number of cells within an organism. The classical way to process the data produced by this technique into comprehensive information is called gating, which is a technique where the cells are placed on two-by-two grids of pairs of surface markers in a sequential manner to classify them [14]. However, this technique is imperfect: it results in the loss of some multivariate relationships among markers, in the loss

of information on levels of cell surface markers, and in the exclusion or misplacement of some cellular populations due to the subjectivity of the technique and the rigid nature of its cut-off [15].

Other methods have been developed to overcome those shortcomings such as viSNE [16] or SPADE [17] which uses semi-supervised clustering methods to categorize cell types using multiple surface markers at the same time. This creates a more flexible way to separate the diverse cell types while conserving the multivariate structure of the data. However, these techniques also have disadvantages such as a lack of reproducibility between runs (SNE) or between algorithms and a lack of an unbiased way to decide if findings made by the algorithms are indeed findings or artefacts [18]. So, there is a need to develop other techniques to analyze immune cytometry data that could overcome these flaws.

In recent studies, it has been shown that immune cell subpopulations are more heterogeneous than was previously believed [19][20][21]. This heterogeneity is even further increased during the aging process, partially driven by naïve T cells, which are functionally different when generated at different stages of life [22][23]. This suggests that immune cell types are less well-defined than was previously believed and has highlighted our lack of understanding of how many different immune cell subtypes might exist and what their precise roles and range of actions are. Despite these findings, there are hardly any studies analyzing the immune system without dividing it into cell types.

Here, we propose a different way of looking at the immune system, which is to directly analyze the values of the surface markers without dividing the cells into different subtypes. This might allow a more global view of the immune system and a less biased way to analyze it since no prior knowledge of the subdivision is assumed. Our goal is to assess if it is possible to obtain relevant biological information using only raw surface marker levels. An ideal method would include both continuous, multivariate information on cell surface markers and the identity of the cells they are on; this is, however, methodologically, and conceptually challenging, so our goal was to assess whether there was potential in going beyond traditional gating methods. The use of immune markers to get relevant biological information is not new [24][25] and we expect that since the markers used to differentiate cell types have specific functions – beyond their ability to classify cells into discrete types – their overall levels could have an impact on health and immune system functioning without having to consider which cell they are on. To test this hypothesis, we used cytometry data from the Singapore Longitudinal Aging Study (SLAS) database and analyzed the distributions of 27 surface markers. We then performed nonlinear regression using two different neural networks to try to predict the presence of several health conditions and found that a model using raw surface marker abundance outperformed a model built using classically gated cell types. We also showed that there was no specific marker that contributed significantly more to the predictions. This study is intended as a brief proof-of-concept and was not designed to predict health outcomes in an applied setting.

Methods

Dataset

For all the analysis performed in this article, we used the second cohort of the Singapore Longitudinal Ageing Study (SLAS-2), a longitudinal study of aging and health of community-dwelling Singaporeans aged 55 or more at the start of the study, as previously described [26][27][28]. It excludes individuals unable to participate because of severe physical or mental disabilities. It includes 3200 residents of the southwest and central south of Singapore starting in 2010. The study received ethical approval from the National University of Singapore Institutional Review Board and written consent was obtained from all participants (response rate of 78%). The study followed the Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines [29]. Although the dataset presents longitudinal components, the flow cytometry data needed for this study were only available cross-sectionally in all participants.

Health outcome metrics

In our analysis, we looked at the predictive power of our models on 20 health or health-status-related measures: (1) Age; (2) Mortality; (3) Self-assessed health measured on a five-point Likert scale, based on the question “Generally would you say your health is: Excellent, Very good, Good, Fair or Poor”; (4) Frailty evaluated on the 5 criteria from Fried’s phenotypic scale [30]: weakness, slowness, weight loss, low physical activity and exhaustion[31]; (5) Global cognitive function as quantified via the Mini Mental State Evaluation (MMSE)[32]; (6) The number of comorbidities from a list of 23 based on self-report, medication and physical or laboratory tests; (7) High blood pressure; (8) High cholesterol; (9) Diabetes; (10) Stroke; (11) Heart attack; (12) Atrial fibrillation; (13) Eye problem; (14) Asthma; (15) Arthritis; (16) Osteoporosis, (17) Gastrointestinal problems; (18) Thyroid problems; (19) Cancer; (20) Depression. Most of these metrics are dichotomic, but age, self-assessed health, frailty, MMSE, and comorbidities are discrete measures with multiple values. Religion was also included as a negative control.

Cell surface markers

The surface markers used in this article are 6-Sulfo LacNAc (Slan), CD19, Pan-GDT, TCRVg1, TCRVa7.2, CD45RO, CD127, CD56, HLADR, CCR6, CD45, CRTH2, CD34, CD38, CD57, CD25, CD16, CD123, CD27, CD3, CD8, CD14, CXCR3, TCRVg2, IgD, CD4 and CD161. The markers CD19 & Pan-GDT, TCRVg1 & TCRa7-2, CD8 & CD14, and TCRVg2 & IgD were paired together respectively on the same channel, during the panel design Flow Cytometry, as these markers are located on different cell types (mutually exclusive)

Preprocessing and statistical analysis

The Flow Cytometry data were analyzed with primary gating to exclude debris using the FSC-A/SSC-A gate, the FSC-A/FSC-H gate to keep only single cells and excluding cells absorbing the LIVE/DEAD™ Fixable Blue Stain (ThermoFisher Scientific). Finally, cells expressingCD45+ were kept. This gating enabled to work on single living leukocytes for the rest of the analyses.

For the non-gated model, since the number of cells varied between individuals but could often approach half a million, we randomly sampled 5000 cells for each individual to ensure equal representation and

reduce computational time. Before this sampling, the first and last 10% of each individual file were removed to limit inconsistencies during the Flow Cytometry acquisition. Since a few extreme negative outliers were observed for most of the markers, a threshold was set at -50 000 relative fluorescence units for all markers and all cells with markers below that limit were removed. This was done to prevent these outliers from weighing too much on the model, since it is based on distribution. Then, for each individual, the distribution of fluorescence intensity of each marker was divided into 102 different sections. All values below the 2.5th percentile and above the 97.5th percentile were put together into the two lowest and highest sections, respectively, in order to avoid outliers having too strong of an impact on the results. The rest of the distributions were separated into 100 sections of the same width on the absolute scale. The number of cells present in each of these sections was then stored and used as input in the model. The individuals were split into groups of 300 for the calibration and 267 for the validation.

The non-gated model is therefore composed of 23 sets (one for each surface marker) of 102 inputs, each followed by a dense layer of 75 neurons, another dense layer of 50 neurons, another dense layer of 25 neurons, and then a dense layer of 1 neuron for the marker studied. The number of neurons in each layer was selected to be lower than the initial input layer and to form a decreasing gradient so that the later layers represent more generalized patterns. The last 23 layers of 1 neuron are then added and passed to a last dense layer of 1 neuron which gives the final output (fig. 1A). For the first 3 layers of 75, 50, and 25 neurons, the activation function is the exponential linear unit and for the two layers of 1 neuron, the activation function is linear. The non-gated model was run with an epochs of 25000 and a batch size of 100.

For the gated model, 67 mutually exclusive different cell types were obtained via a gating strategy shown in Supplemental figure X. The individuals were split into 300 for the calibration and 267 for the validation. The model is composed of 67 inputs, followed by a dense layer of 50 neurons, a dense layer of 30 neurons, a dense layer of 15 neurons, and a dense layer of 1 neuron which gives the final output (fig. 1B). The first three layers of 50, 30, and 15 neurons have an exponential linear unit activation function and the last layer of 1 neuron has a linear activation function. The same reasoning as for the Continuous model was applied to the selection of the number of neurons in each layer for this model. For the gated model, an epochs of 10000 was used since the model converged more easily and a batch size of 100.

For both models, individuals that had missing data in any of the measures were removed to keep the number of people used to calibrate and evaluate each model the same. Models were generated 100 different times using the same settings to create replicates to consider the random variation that can occur during the generation of the model. Both models used the Adam algorithm for there optimisation. All analyses were conducted using R v3.6.3[33], Python 3.7.6 [34] and TensorFlow 1.14.0 [35].

Success of predictions was assessed based on the comparison of the root mean squared error (rmse) score and the mean value. An rmse for a health measure with no predictive capacity would be close to or higher than its mean value. Successful predictions were considered to be health measures for which the rmse was less than one third of the mean value.

Results

Table 1
Sample characteristics.

	n = 567
Age	67.1 ± 7.5
Mean ± SD	55–89
Range (min-max)	
Sex	337 (39)
M (%)	527 (61)
F (%)	
Mortality (%)	33 (5.8)
Self-assessed Health	7 (1.2)
1 (%) – better health	88 (15.5)
2 (%)	334 (58.9)
3 (%)	134 (23.6)
4 (%)	4 (0.7)
5 (%) – worst health	
Frailty	261 (46)
0 (%)	184 (32.5)
1 (%)	81 (14.3)
2 (%)	35 (6.2)
3 (%)	5 (0.9)
4 (%)	1 (0.2)
5 (%)	
MMSE, mean ± SD	27.8 ± 2.8
Comorbidity, mean ± SD	2.4 ± 1.6
High blood pressure (%)	245 (43.2)
High cholesterol (%)	263 (46.4)
Diabetes (%)	77 (13.6)
Stroke (%)	23 (4)
Heart attack (%)	30 (5.3)

	n = 567
Atrial fibrillation (%)	19 (3.4)
Eye problem (%)	175 (30.1)
Asthma (%)	28 (4.9)
Arthritis (%)	80 (14.1)
Osteoporosis (%)	28 (4.9)
Gastrointestinal problem (%)	50 (8.8)
Thyroid problem (%)	28 (4.9)
Cancer (%)	19 (3.4)
Depression (%)	18 (3.2)

Distributions of the surface markers

Figure 2 shows four representative examples of surface marker distributions. CD3 (Fig. 2A) is a clear bimodal distribution, presumably indicating a general presence (right) or absence (left) of the marker; note, however, the substantial quantitative variation in the marker on cells which are positive for it. CD 161 (Fig. 2D) is closer to a normal distribution indicating a more continuous gradient of abundance of the surface marker. This second type of distribution was more common amongst the markers analyzed, but some distributions also fell in between those categories like CD38 (Fig. 2B) and CD45RO (Fig. 2C). All distributions can be found in the supplement (figure X).

Testing the predictive capacity of the raw surface marker abundance

The predictive power of the model using raw surface marker abundance (continuous) and the one using gated cell types (gated) were determined using multiple health measures (Table 2). For most health measures, we were unable to obtain any successful prediction for either the gated or ungated models, which is not unexpected, as it would be surprising if immune markers were able to predict everything we tested. Nevertheless, we were able to successfully predict three health measures: Age, Self-assessed Health, and MMSE. All three scored low rmse for the validation set in comparison with their mean values, which is the value that we would obtain if we had no predictive capacity for this health measure. Religion was added as a negative control to test for overfitting.

The same health measures were significantly predicted in both models, but the rmse scores were higher in all three cases for the gated model, indicating a less precise prediction. It can be seen in the scatter plot of Figs. 3A, B, and C, as the errors of the gated model (in blue) seem to be more extreme. This is especially visible in Fig. 3A as the dots are more clearly visible and in the violin plot of Fig. 3D where we

see that the gated model has both higher median error as well as more extreme errors. There was also a bit more variation in the different iterations of the model for the gated results, especially for age and MMSE with respectively three and two times higher standard deviation values as seen in Table 2. Beyond the significance of the individual models, 15 of 21 models showed lower rmse in the continuous model, even if slightly, a bias which has a p-value of 0.04 based on the binomial distribution.

Table 2

Averages and standard deviations of the rmse on the validation set of 100 separate runs of the non-gated and gated model for the health measure tested and the mean values for these measures. In bold are the health measures for which the models were able to make successful predictions (rmse < mean/3).

	Continuous		**Gated**		**Mean**
	rmse	**std rmse**	**rmse**	**std rmse**	
Age	**8.705**	**0.336**	**13.732**	**0.961**	**67.105**
Mortality	0.255	0.022	0.264	0.022	0.058
Religion	2.208	0.104	2.335	0.109	2.260
Self-assessed Health	**0.840**	**0.030**	**0.961**	**0.046**	**3.072**
Frailty	1.154	0.048	1.302	0.057	0.830
MMSE	**3.344**	**0.332**	**4.497**	**0.308**	**27.835**
Comorbidity	1.982	0.101	2.291	0.127	2.322
High blood Pressure	0.616	0.029	0.679	0.034	0.432
High cholesterol	0.626	0.024	0.714	0.037	0.464
Diabetes	0.413	0.026	0.481	0.034	0.136
Stroke	0.234	0.020	0.243	0.02	0.040
Heart attack	0.277	0.021	0.305	0.025	0.053
Atrial fibrillation	0.220	0.022	0.234	0.017	0.034
Eye problem	0.594	0.022	0.611	0.032	0.301
Asthma	0,263	0,020	0,298	0,032	0.049
Arthritis	0,428	0,022	0,481	0,023	0.141
Osteoporosis	0,277	0,035	0,301	0,027	0.049
Gastrointestinal problem	0,352	0,026	0,367	0,037	0.088
Thyroid problem	0,275	0,027	0,287	0,027	0.049
Cancer	0,219	0,025	0,216	0,017	0.034
Depression	0,214	0,020	0,241	0,047	0.032

Contributions of the different markers on the gated model's results

We looked at the values added for each marker in the last layer of the non-gated model (informative layer, Fig. 1), because this value represents the contribution of that marker to the prediction. The results (Fig. 4) show us that CD3 contributed a lot to the prediction of age and health, but not that much in MMSE. CD16 contributed very little to all three health measures.

Discussion

Here we have shown that it is possible to extract useful information from the levels of immune cell surface markers without consideration of specific cell types, and that this information could even outperform information extracted from traditional gating techniques. For most of the health measures tested here, neither of our models was able to give a meaningful prediction, a finding that was not unexpected. The fact that both models were able to predict the same health measures indicates that those that were successfully predicted are likely to be linked to the immune system and not just successful by chance.

The distributions of the surface markers observed in this study, consistent with those normally reported in traditional gating studies [36][37][38], indicate that many cannot easily be categorized as present or absent, or minimally that there is substantial variation in the quantities of the markers present even when there may be a distinct class of cells lacking the marker. This, together with our finding that a model built using raw surface marker abundance can outperform one built with traditional gating, points toward the potential of analyses considering cell surface marker abundance as a continuous rather than dichotomous measure. In our analysis, we noted certain markers as more or less important for the prediction of some health measures. CD3 was important for the prediction of age and health. It is a co-receptor that is used to identify T cells, indicating that this cell type might be important to determine these health outcomes. CD16 is present, among many other functionally similar receptors, on natural killer cells, monocytes, and macrophages and is implicated in the activation of those cells during an infection. Since it contributed very little to all three measures, it might mean that this function is not directly linked to these health outcomes. While CD3 is linked to adaptive immunity and CD16 is linked to innate immunity, this study could help discriminate the association of the two arms of the immune system in health outcomes during aging. Results presented in Fig. 4 cannot be replicated for the gated model since it does not take raw surface markers abundance as an entry. This shows that building models using this kind of information can help discover more information about these markers in ways that might be difficult with a more traditional approach.

The approach shown here has several limitations. Most notably, while it appears that continuous information on surface marker abundance is relevant for understanding health, the approach used here does not consider which cells have which joint abundances of markers. Obviously, the relevance of a high abundance of a given marker on a cell may depend on the levels of the other markers on that same cell. It is analytically challenging to generate a portrait of an individual based on a composition of a large number of cells that are not discretely categorized and/or that vary along multiple axes (i.e., markers). Our goal was simply to show that the traditional gating approach implies a loss of relevant information.

Second, despite the high-quality immune data available in SLAS via Flow Cytometry, we did not have the sample sizes needed to properly train models on specific immune pathologies or states (Table 1). The health measures we successfully predicted were all continuous or semi-continuous, suggesting that a lack of power was an important factor in the failure of other predictions. Third, this study was not designed to develop predictive models of health based on Flow Cytometry data, and accordingly, we make no claims about the power or relevance of the predictions made and attempted. Fourth, our analysis only includes people aged 55 and above. This limits the scope of our finding but does not exclude that our results might be generalizable to a broader population. Heterogeneity of the immune subpopulations increases during aging, making continuous analysis of the immune system even more suited for this type of population. Lastly, we note that the 27 surface markers included here are far from an exhaustive catalogue, and much more might be done with a more extensive list.

These are important limitations, and we stress that our key goal here was to briefly demonstrate the potential of new ways to consider the variability of the immune system that might be developed based on the incredible richness of Flow Cytometry data, possibly encouraging more studies to be made with this type of approach. The traditional gating approach clearly results in loss of important biological information. Improvements to gating based on Bayesian clustering [39] or other methods that reduce the dataset to counts of discrete cell types are likely to provide marginal but not massive improvements. Given the limited literature looking at the immune system without dividing it into cell types, this study might be useful to stimulate more research from that angle. We hope our approach will stimulate further thought on how to integrate continuous variation in surface marker abundance into future analyses.

Declarations

Acknowledgments

We thank the following voluntary welfare organizations for their support in kind: Geylang East Home for the Aged, Presbyterian Community Services, St Luke's Eldercare Services, Thye Hua Kwan Moral Society (Moral Neighbourhood Links), Yuhua Neighbourhood Link, Henderson Senior Citizens' Home, NTUC Eldercare Co-op Ltd, Thong Kheng Seniors Activity Centre (Queenstown Centre) and Redhill Moral Seniors Activity Centre.

Ethics approval

Ethics approval was obtained from National University of Singapore IRB(Ref:04–140). All participants gave written informed consent to participate in the study. Ethics approval for secondary analysis of the data was obtained from the University of Sherbrooke ethic board (Ref number: 2019-2657).

Consent for publication

Not applicable

Authors' contributions

GP analysed the data, generated the results, and wrote the manuscript. GP and FBL created the model to make the predictions. AL, TPN, TTY and MC contributed to the generation of the SLAS dataset. GP, TF and AAC designed the study. TF and AAC were major contributor in writing the manuscript. All authors read and approved the final manuscript.

Funding Sources

This work was supported by research grants from the Agency for Science Technology and Research (A*STAR) Biomedical Research Council (grant number BMRC/08/1/21/19/567) and the National Medical Research Council (grant numbers NMRC/1108/2007, NMRC/CIRG/1409/2014) and the Canadian Institutes of Health Research (CIHR) (No. 106634) and No. PJT-162366) to AL and TF, National Science and Engineering Research Council Grant # RGPIN-2018-06096, the Société des médecins de l'Université de Sherbrooke and the Research Center on Aging of the CIUSSS-CHUS, Sherbrooke to TF, and the FRQS Audace grant to AAC, TF, and AL (grant number 2020/AUDC/270433). AAC is a Senior Research Fellow of the *Fonds de recherche du Québec—Santé* (FRQS), and a member of the FRQS-supported *Centre de recherche sur le vieillissement* and *Centre de recherche du CHUS*.

Conflict of Interest

AAC is founder and CEO at Oken Health.

References

1. Melief, C. J. M., Toes, R. M., Medema, J. P., Van Der Burg, S. H., Ossendorp, F., & Offringa, R: Strategies for immunotherapy of cancer. Advances in Immunology (2000) 235–282.
2. BS Nikolajczyk, M Jagannathan-Bogdan, H Shin1 and R Gyurko: State of the union between metabolism and the immune system in type 2 diabetes. Genes and Immunity (2011) 12, 239–250.
3. Guerriero, F., Sgarlata, C., Francis, M., Maurizi, N., Faragli, A., Perna, S., Rondanelli M., Rollone M., Ricevuti, G. Neuroinflammation, immune system and Alzheimer disease: searching for the missing link. Aging Clinical and Experimental Research, (2016) 29(5), 821–831.
4. François M. Abboud, Sailesh C. Harwani, Mark W. Chapleau: Autonomic Neural Regulation of the Immune System Implications for Hypertension and Cardiovascular Disease. Hypertension (2012) Apr;59(4):755 – 62.
5. Fulop T, Le Page A, Fortin C, Witkowski JM, Dupuis G, Larbi A. Cellular signaling in the aging immune system. Curr Opin Immunol. (2014) Aug;29:105–11.
6. Cohen, A. A., Martin, L. B., Wingfield, J. C., McWilliams, S. R., & Dunne, J. A. (2012). Physiological regulatory networks: ecological roles and evolutionary constraints. Trends in Ecology & Evolution, 27(8), 428–435.
7. Fülöp T, Dupuis G, Witkowski JM, Larbi A. The Role of Immunosenescence in the Development of Age-Related Diseases. Rev Invest Clin. (2016) Mar-Apr;68(2):84–91.

8. Fulop, T., Larbi, A., Dupuis, G., Le Page, A., Frost, E. H., Cohen, A. A., … Franceschi, C. (2018). Immunosenescence and Inflamm-Aging As Two Sides of the Same Coin: Friends or Foes? Frontiers in Immunology, 8.
9. Fülöp T, Larbi A, Witkowski JM. Human Inflammaging. Gerontology. 2019;65(5):495–504.
10. Curiel TJ. Tregs and rethinking cancer immunotherapy. J Clin Invest. 2007;117(5):1167–1174.
11. Nicholson LB. The immune system. Essays Biochem. (2016) 60(3):275–301.
12. Alam R. A brief review of the immune system. Prim Care. (1998) Dec;25(4):727 – 38.
13. Jerne NK. The immune system. Sci Am. 1973 Jul;229(1):52–60
14. McCoy, Jr, J. P. (Ed.). Immunophenotyping. Methods in Molecular Biology. (2019).
15. Gonder S, Fernandez Botana I, Wierz M, Pagano G, Gargiulo E, Cosma A, Moussay E, Paggetti J, Largeot A. Method for the Analysis of the Tumor Microenvironment by Mass Cytometry: Application to Chronic Lymphocytic Leukemia. Front Immunol. (2020) 20;11:578176.
16. Amir el-AD, Davis KL, Tadmor MD, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat Biotechnol. (2013); 31(6):545–552.
17. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Jr, Bruggner RV, Linderman MD, Sachs K, Nolan GP, Plevritis SK. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol. (2011);29:886–891.
18. Pedersen, C.B. and Olsen, L.R. Algorithmic Clustering Of Single-Cell Cytometry Data—How Unsupervised Are These Analyses Really? Cytometry (2020), 97: 219–221.
19. Norah L. Smith, Ravi K. Patel, Arnold Reynaldi, Jennifer K. Grenier, Jocelyn Wang, Neva B. Watson, Kito Nzingha, Kristel J. Yee Mon, Seth A. Peng, Andrew Grimson, Miles P. Davenport, Brian D. Rudd. Developmental Origin Governs CD8 + T Cell Fate Decisions during Infection, Cell (2018), Volume 174, Issue 1, Pages 117–130.e14.
20. Reynaldi A, Smith NL, Schlub TE, Venturi V, Rudd BD, Davenport MP. Modeling the dynamics of neonatal CD8 + T-cell responses. Immunol Cell Biol. (2016);94(9):838–848.
21. Gerlach, C., Moseman, E. A., Loughhead, S. M., Alvarez, D., Zwijnenburg, A. J., Waanders, L., Garg, R., de la Torre, J. C., & von Andrian, U. H. (2016). The Chemokine Receptor CX3CR1 Defines Three Antigen-Experienced CD8 T Cell Subsets with Distinct Roles in Immune Surveillance and Homeostasis. Immunity, 45(6), 1270–1284.
22. Zhang H, Weyand CM, Goronzy JJ. Hallmarks of the aging T-cell system. FEBS J. 2021 Dec;288(24):7123–7142.
23. Zhang H, Weyand CM, Goronzy JJ, Gustafson CE. Understanding T cell aging to improve anti-viral immunity. Curr Opin Virol. 2021 Dec;51:127–133
24. Catacchio, I., Scattone, A., Silvestris, N., & Mangia, A. Immune Prophets of Lung Cancer: The Prognostic and Predictive Landscape of Cellular and Molecular Immune Markers. Translational Oncology (2018), 11(3), 825–835.

25. Seiler C., Kronstad L.M., Simpson L.J., Le Gars M., Vendrame E., Blish C.A., Holmes S. Uncertainty Quantification in Multivariate Mixed Models for Mass Cytometry Data. arXiv (2019), 1903.07976.
26. Ng TP, Feng L, Nyunt MS, Larbi A, Yap KB, Frailty in older persons: multisystem risk factors and the Frailty Risk Index (FRI). J Am Med Dir Assoc. 2014 Sep; 15(9):635–42.
27. Feng L, Zin Nyunt MS, Gao Q, Feng L, Yap KB, Ng TP, Cognitive Frailty and Adverse Health Outcomes: Findings From the Singapore Longitudinal Ageing Studies (SLAS). J Am Med Dir Assoc. 2017 Mar 1; 18(3):252–258.
28. Kai Wei, Ma-Shwe-Zin Nyunt, Qi Gao, Shiou-Liang Wee, Keng-Bee Yap and Tze-Pin Ng, Association of Frailty and Malnutrition With Long-term Functional and Mortality Outcomes Among Community-Dwelling Older Adults, Results From the Singapore Longitudinal Aging Study 1, JAMA Netw Open. 2018 Jul; 1(3): e180650.
29. Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. Epidemiology (2007), 18(6):805–35.
30. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, Seeman T, Tracy R, Kop WJ, Burke G, McBurnie MA; Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci. (2001) Mar;56(3):M146-56.
31. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. Frailty in older adults: evidence for a phenotype. The journals of gerontology Series A, Biological sciences and medical sciences.2001;56:M146-156.
32. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975 Nov;12(3):189–98.
33. R Core Team. (2017) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/
34. Van Rossum G, Drake F L (2009) Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.
35. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2016).Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.
36. Aurélie Poli 1, Tatiana Michel, Maud Thérésine, Emmanuel Andrès, François Hentges, Jacques Zimmer, CD56bright natural killer (NK) cells: an important NK cell subset, Immunology (2009) Apr;126(4):458 – 65.
37. Sara M Centuori, Cecil J Gomes, Samuel S Kim, Charles W Putnam, Brandon T Larsen, Linda L Garland, David W Mount, Jesse D Martinez, Double-negative (CD27 - IgD -) B cells are expanded in NSCLC and inversely correlate with affinity-matured B cell populations, J Transl Med (2018) Feb 15;16(1):30.
38. Irina Yu Nikitina, Alexander V Panteleev, George A Kosmiadi, Yana V Serdyuk, Tatiana A Nenasheva, Alexander A Nikolaev, Lubov A Gorelova, Tatiana V Radaeva, Yana Yu Kiseleva, Vladimir K Bozhenko, Irina V Lyadova, Th1, Th17, and Th1Th17 Lymphocytes during Tuberculosis: Th1 Lymphocytes

Predominate and Appear as Low-Differentiated CXCR3 + CCR6 + Cells in the Blood and Highly Differentiated CXCR3 +/- CCR6 - Cells in the Lungs, J Immunol (2018) Mar 15;200(6):2090–2103.

39. Minoura K, Abe K, Maeda Y, Nishikawa H, Shimamura T. Model-based cell clustering and population tracking for time-series flow cytometry data. BMC Bioinformatics. 2019 Dec 27;20(Suppl 23):633.

Figures

Figure 1

Representation of the models. A: The continuous model. **B:** The gated model.

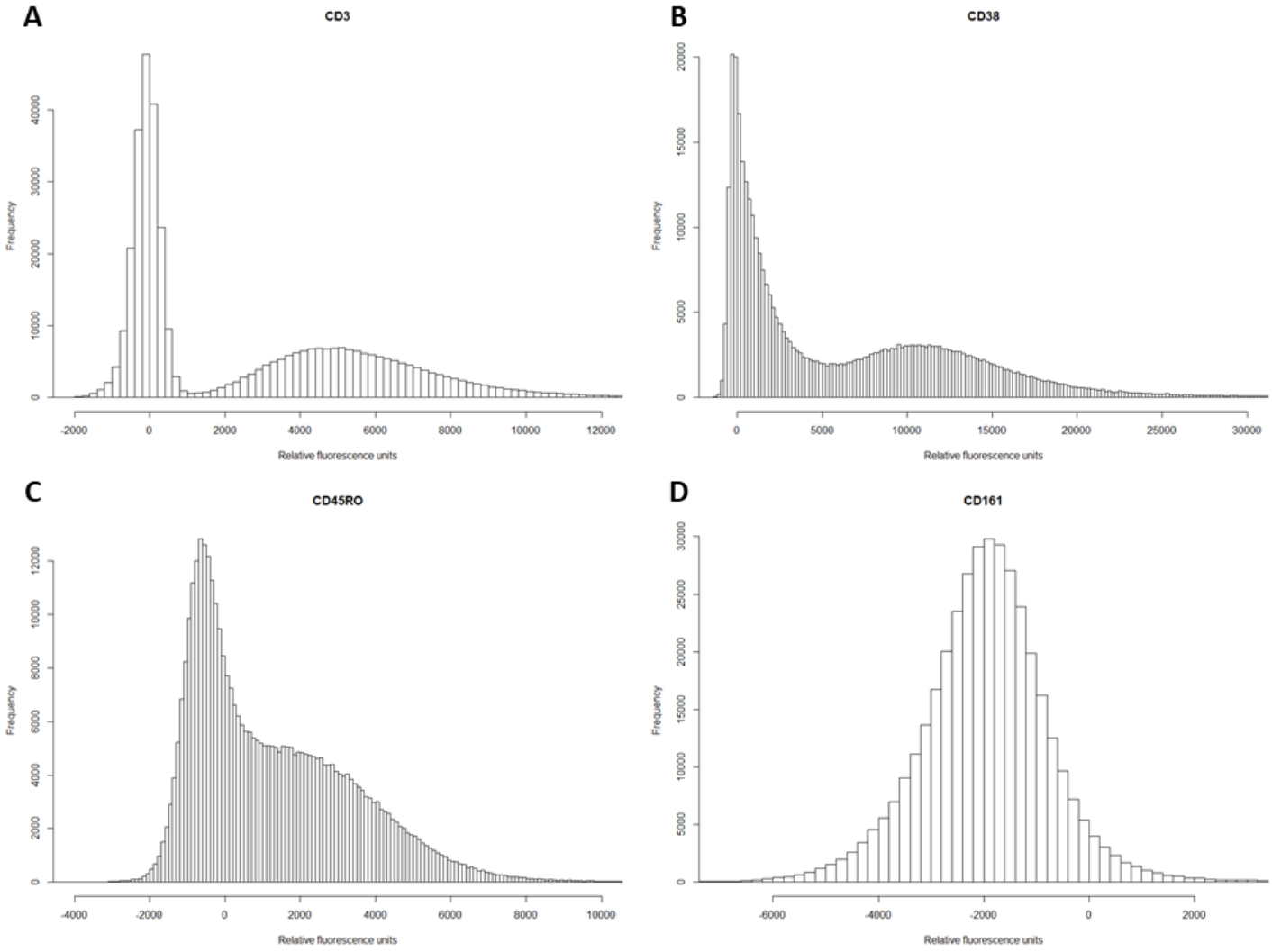

Figure 2

Example distributions of four of the surface markers tested in this article. A: Distribution of CD3, used to identify T cells. **B**: Distribution of CD38, used to identify B cell subsets. **C**: Distribution of CD45RO, used to identify memory T cells. **D**: Distribution of CD161, which can help define various T cell subsets.

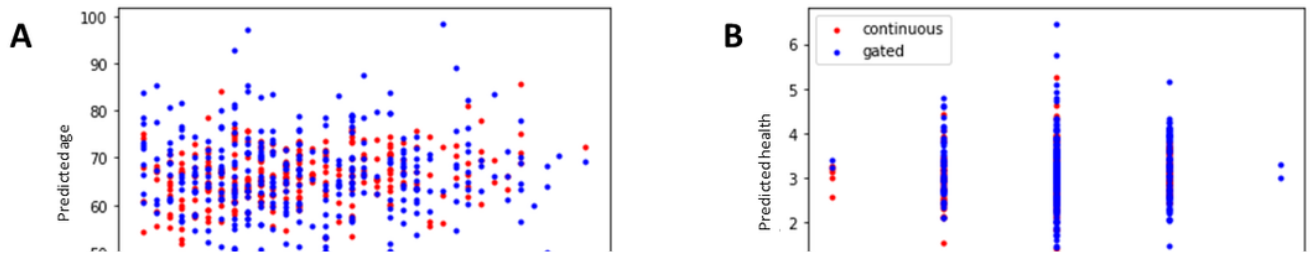

Figure 3

Comparison of the errors of the predictions between the continuous and the gated models for Age, Self-assessed health, and MMSE. **A, B, and C:** Scatter plots of the difference between the observed value and the predicted value for Age, Self-assessed health, and MMSE respectively. **D:** Violin plot of the difference between the observed value and the predicted value, with the middle bar representing the median.

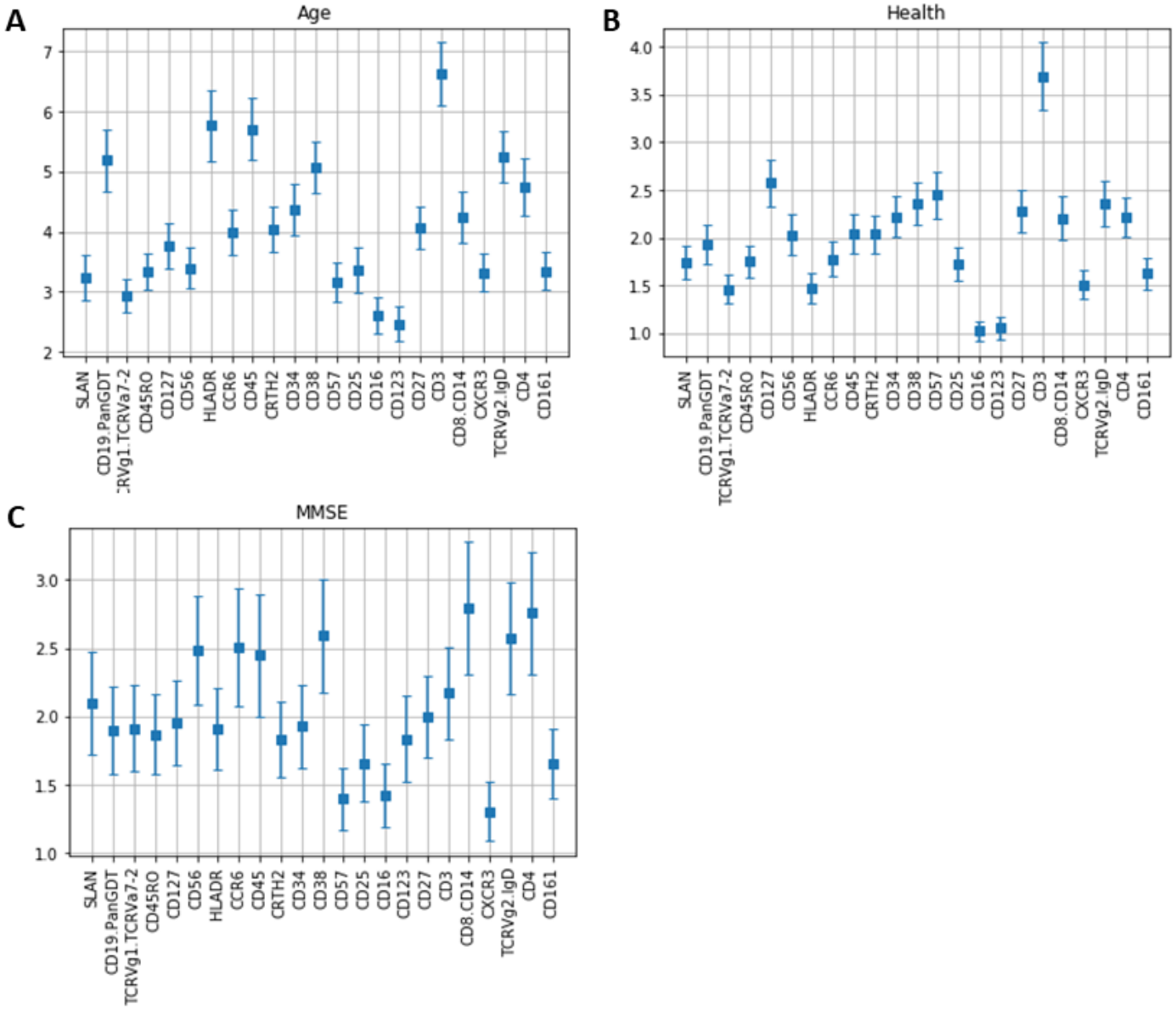

Figure 4

Values from the last layer of the non-gated model (informative layer) for successfully predicted outcomes, representing the contribution of that specific marker to the overall prediction. The middle square represents the mean value obtained for the 100 separate runs of the model and the bars are the standard error.

Supplementary Files

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

- Additionalfile1.docx

- Additionalfile2.pdf