medRxiv preprint doi: https://doi.org/10.1101/2024.07.01.24309803; this version posted July 3, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

GEOMETRY BASED GENE EXPRESSION SIGNATURES DETECT CANCER TREATMENT RESPONDERS IN CLINICAL TRIALS

Wojciech Chachólski
Department of Mathematics
KTH-Royal Institute of Technology
Stockholm, Sweden 10044
wojtek@kth.se

Ryan Ramanujam *
Department of Clinical Neuroscience
Karolinska Institutet
Stockholm, Sweden 17176
ryan.ramanujam@ki.se

July 2, 2024

ABSTRACT

Aim: The overall aim of this project is to determine if gene expression signatures of tumors, constructed from geometrical attributes of data, could be used to predict patient treatment response by detecting subgroups of responders. This is tested in Pfizer clinical trial data and compared with standard clustering methods ($n = 726$).

Results: Geometrical gene expression signature analysis demonstrated high utility to detect subgroups with enhanced treatment response. In the Pfizer trial, gene expression signatures were able to detect three subgroups of responders ($p = 0.012$), containing 52.9% of patients and accounting for nearly all the observed treatment effect. Standard techniques following a similar methodology were able to partition a single subgroup containing 21.3% of patients.

Conclusions: Gene expression based geometrical signatures yielded vastly superior performance over standard clustering techniques, as demonstrated in Pfizer's Phase III clinical trial data. These can be used to determine subgroups of enhanced treatment response in oncology clinical trials, and might lead to personalized treatment recommendations in the future.

Keywords Gene expression · Cancer · TDA

1 Introduction

Precision medicine has the goal of individualizing the treatment for a patient based on that patient's particular characteristics, which may include data on patient demographics and clinical information, as well as experimental data such as biomarkers, genomics, gene expression, proteomics, metabolomics, etc. By using individual response to treatment, a more efficacious treatment approach or compound can be selected based on analysis of this data [1]. In oncology, tumors can be characterized by molecular approaches such as measuring gene expression or mutations. The outcomes for each patient in clinical studies are typically reflected in time to event, with an increased time to progression free survival (PFS), being frequently used to indicate better treatment effect.

Mutations are currently the most frequently used biomarker for dividing patients and tumors into clinical subgroups [2]. This may be a single somatic mutation or a combination of several mutations. This has led to efforts to find biomarkers across tumor types which may indicate treatment response [3]. The ability to detect individuals who may respond to a treatment across various tumor tissues would have strong clinical utility for precision oncology [4]. Nevertheless, identifying such biomarkers to detect a large proportion of patients with efficacious responses remains difficult [1].

Some cancer therapies have been FDA approved based on the presence of certain mutations [5]. However, data from the Dependency Map project (DepMap) [6] presented evidence that gene expression is the datatype most strongly associated with treatment response. Given this, efforts to combine gene expression data into signatures which correspond to

*Corresponding author

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.01.24309803; this version posted July 3, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

tumor biology and/or state are warranted. One method of combining large amounts of data, such as from RNA-seq experiments, is by using techniques being developed in geometrical and topological research.

Topological data analysis (TDA) is a rapidly growing field enabling the ability to structure and organize data based on its geometry. It transforms data into invariants that capture higher order interactions and are amenable for statistical and machine learning analysis. The presence of the higher order information often can yield insights which are not available with standard statistical methods. This field has been applied to biomedical research, with results in breast cancer characterizing tumors with unique characteristics [7]. It also was used to identify Type 2 diabetes (T2D) subgroups based on clinical features which had distinct genetic associations [8].

We have developed several TDA methods based on classical TDA approaches such as extending the mapper algorithm for multi-measurements [9], and creating a new invariant called stable rank [10, 11] which is a key component of the technology which is presented here. Stable rank has further been used in creating supervised learning methods [12], which then have been applied to characterization of sex-based differences in microglia [13]. Extensions to these methods herein create signatures based on gene expression, which contain various unique properties and can be correlated to outcomes of interest such as treatment response.

Creating and testing these methods is difficult due to a lack of real-world data from drug trials. However, a trial was conducted by Pfizer to test Avelumab plus Axitnib in patients with advanced renal cell carcinoma as part of the JAVELIN Renal 101 program. The Phase III trial reached clinical endpoints and the drug combination was approved. The data from this trial was released publicly, and includes information on patient clinical features and outcomes, as well as high throughput data such as gene expression, mutation, HLA, etc. The release of this data provides an ideal test case for using geometrical methods to predict treatment response in an oncology clincial trial.

The objectives of this study are to create gene expression signatures and apply related methodology to the Pfizer clinical trial to determine if subgroups of responders can be identified which are statistically significant, and to compare subgrouping results from gene expression signatures using TDA methods with more conventional approaches.

2 Materials and methods

2.1 Pfizer JAVELIN Renal 101 trial

The Pfizer JAVELIN program consisted of long-term monitoring of patients with advanced renal cell carcinoma [14]. A Phase III trial in avelumab plus axitinib was conducted with a comparator arm of sunitinib as the standard of care. The total number of patients enrolled was 886, with 442 assigned to the treatment arm and 444 assigned to the sunitinib comparator arm.

Data for this trial was provided for multiple data types, including 726 for which gene expression data was available along with clinical parameters and outcomes. Gene expression was given in the form of $\log_2$ TPM, ie. Transcripts Per Million, with minimum values set to 0.01. Data were not further processed in any way before being used in topological data analysis.

Additional data such as gene mutations, HLA, pathway, etc. were not considered in this analysis.

Clinical information included baseline parameters such as age, PD-L1 positive status (PD-L1+), sex, time to PFS, and binary indication of censor at PFS (that is, inverse of event at PFS). The Median time to PFS was reported as 13.3 months in the treatment arm and 8.0 in the comparator arm, resulting in a stratified HR of 0.69 (p-value < 0.0001). However, it must be noted that in the data provided by Pfizer, using data from the 726 individuals with both clinical and gene expression data, median time to PFS is 4.07 months between arms as opposed to the 5.3 months reported. We therefore used the data in its current form, while considering the more modest difference in median time to PFS present for the present study.

2.2 TDA approach

In this article we build from the local approach, as described in [15]. This process assigns to every element in a distance space, a non increasing function called a stable rank, according to some predefined parameters. A key feature of this signature of a point in a distance space is that it captures geometrical properties of the embedding of this point into the entire distance space. In particular, points whose embeddings have similar geometrical properties have similar stable ranks, and the embeddings of points whose stable ranks are far apart have different geometrical properties.

The first step of our analysis is to choose a collection of sets of parameters (called geometrical parameters) for extracting stable ranks. Extracting stable ranks, with respect to each set of geometrical parameters, is the second step of our analysis. In the third step, for every set of geometrical parameters, we consider the distribution of integrals of the

medRxiv preprint doi: https://doi.org/10.1101/2024.07.01.24309803; this version posted July 3, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

obtained stable ranks. We think about these integrals as a filter function and apply local clustering according to this filter and the l_2 distance in a process similar to the mapper procedure [7]. In this way, for every set of geometrical parameters, we construct various subgroups of the obtained stable ranks, and hence subgroups of the points in the distance space from which we started.

Thus the points in the distance space are subgrouped according to the geometrical properties of their embeddings as captured by the obtained stable ranks: the embedings of the points in different subgroups have different geometrical character and the embeddings of the points in the same subgroup have similar geometrical character.

We applied this three step process to the gene expression data with the distance given by the Euclidean metric and 22 standard sets of geometrical parameters in homological degees 0 and 1. The outcome is a collection of subgroups of patients based on geometrical similarities of the embeddings of the associated gene expression in the Euclidean distance space given by the entire gene expression data. These subgroups are organized into 44 groups, corresponding to our choice of sets of geometrical parameters.

2.3 Statistical analysis

The significance threshold alpha was set to 0.05 for all testing.

For each subgroup created via our TDA pipeline described in Section 2.2, the Kaplan-Meier (KM) estimators was created and a log-rank test was employed. To avoid spurious associations arising from multiple testing error, Bonferroni correction was employed by altering alpha based on the number of total subgroups for a given choice of a set of geometrcial parameters. Subgroups that passed the revised threshold were then passed to step two.

Strictly speaking, Bonferroni correction may be too conservative when partitioning a dataset into hypothetical subgroups to test. This is due to the reduced number of samples available and lowered p-value; in some instances when too many subgroups are created, the variation may be insufficient to attain statistical significance due to lower sample size. Nevertheless, this provides an even basis for comparison of results within a robust and well characterized statistical framework.

In the second step, we determined the p-value of the entire subgrouping process. Since the entirety of the data has a statistically significant p-value, we considered the null hypothesis to be that there are no subgroups with a statistically significant p-value in the KMs. We then created 1000 permutations by shuffling the samples to create the approximate test distribution statistic based on the number of significant subgroups in the KMs created. We then assigned a p-value to results by determining the number of results more extreme with respect to significant subgroups.

In the last step, subgroups were merged using the following procedure. First, all subgroups with KM derived p-values less than the threshold calculated using Bonferroni correction based on the number of blocks in a given geometry were included, and non-significant subgroups were discarded. Next, subgroups were merge if and only if the Gower distance between them indicated less than 50% overlap of the smaller subgroup. The final subgroups were then reported, with all samples not in one such subgroup lumped together and considered as "non-subgroup".

This entire procedure was repeated using a standard gene-expression clustering procedure. This consisted of first using t-SNE to reduce the data to two dimensions, using "TSNE" from the python package "sklearn". Then, the DBSCAN algorithm from "sklearn" was used to cluster according to similarity. This was conducted with several parameter choices, with an ad-hoc choice of the optimum number to promote maximum clusters. A similar approach of Bonferroni correction using log-rank test of KM was used, although determining statistical significance was not conducted since only a single partitioning of the data was created which is not amenable to permutation testing. Since t-SNE is not deterministic, a representative embedding was selected. We queried and tested other parameter choices to ensure that superior results to the selection were not likely.

3 Results

The unsupervised geometrical transformation into gene-expression signatures resulted in six statistically significant subgroups. Permutation testing determined that this partitioning had an empirical p-value of 0.012. The subgroups were then merged according to the $a\ priori$ methodology, resulting in three subgroups ($n = 258$, $n = 107$, $n = 91$, union=384). The remaining patients ($n = 352$) were then grouped into a "non-subgroup" for comparison purposes. The standard method (t-SNE with DBSCAN) resulted in one significant subgroup being detected ($n = 155$, 21.3% of patients). The remaining patients were likewise grouped together ($n = 571$).

medRxiv preprint doi: https://doi.org/10.1101/2024.07.01.24309803; this version posted July 3, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

3.1 Kaplan-Meier plots

Figure 1 shows the Kaplan-Meier plots for the subgroups generated by geometrical gene expression signatures, for each of the three subgroups detected (n=384) as well as the remaining non-subgroup patients (n=342).

In Figure 2, the Kaplan-Meier estimators for the conventional subgrouping method is presented for the identified subgroup of patient responders (n=155) and all other patients (n=571).

The baseline characteristics of subgroups created by both geometric and conventional methods are shown in the table in Figure 3. The number of patients, mean age, proportion female, PD-L1+ proportion, proportion in the treatment arm and proportion that reached PFS are given for each of the subgroups and non-subgroups.

		Geometrical results				***t*-SNE & DBSCAN**	
	All	**Subgr 1**	**Subgr 2**	**Subgr 3**	**Non-subgr**	**Subgr 1**	**Non-subgr**
Number	726	258	107	91	342	155	571
Age (mean)	60.69	59.60	59.70	60.10	61.30	59.42	61.03
Sex (proportion female)	0.25	0.28	0.30	0.19	0.23	0.24	0.25
PD-L1+ (proportion)	0.71	0.66	0.75	0.77	0.72	0.71	0.71
Treatment arm (proportion)	0.49	0.52	0.59	0.47	0.46	0.45	0.50
Reach PFS (proportion)	0.51	0.45	0.52	0.55	0.53	0.46	0.52

Figure 3: Baseline clinical parameters per subgroup and non-subgroup, for both the subgroups/non-subgroups created with geometrical methods, as well as for those created with t-SNE and DBSCAN

4 Discussion

We have created a method of creating geometry based gene expression signatures for cancer tumors. This allows for comparison of signatures within a clinical trial to identify subgroups which are correlated with improved treatment. This use case shows strong utility with data from a real-world cancer trial and has applications in precision oncology, ie. getting the correct treatment to the correct patient.

In the Pfizer Phase III clinical trial, the methodology displays superior subgrouping of patient responders than conventional clustering. Using a completely generic and strict a priori analysis structure, the methodology detected three subgroups of enhanced response of 384 patients in the treatment arm, which contained nearly all the treatment effect. Permutation testing confirmed this result was unlikely to be a chance finding ($p = 0.012$). The alternative methodology detected a single subgroup of 155 patients with improved response. However, it should be noted that a p-value using permutation testing was not practically feasible. If one assumes that these results are robust, this method of detecting subgroups still identified only 40.3% of the number of patient responders as the geometrical method.

The subgroups did not vary substantially in terms of mean age, proportion of females, PD-L1+ status, or proportions of treated or those that achieved PFS. This was not surprising, given that the subgroups were created purely from geometrical information in an unsupervised manner, such that clinical and baseline characteristics were not considered. Hence the subgroups appear to be distinct molecular signatures of response which correlate to treatment outcome and not to baseline characteristics. This reinforces the notion that subgrouping based on some clinical parameter, such as PD-L1 positivity, may not be optimal for detecting subgroups of enhanced response. Using the techniques described to detect and identify subgroups based on geometrical gene expression signatures could provide important information for Phase III clinical trial endpoints.

The method described herein first transforms the data and then analyses the obtained signatures. These signatures integrate certain geometrical aspects of how individual points embed into the space of gene expression data. This has key advantages over standard machine learning (ML) and (artificial intelligence) AI implementations applied directly to the data. For example our method is able to find results with only a few hundred data points when the number of features is on the order of tens of thousands. Typically, ML and AI techniques requires a large number of datapoints to train models, and the presence of huge numbers of variables creates a noisy point cloud where extraneous features may reduce model performance. This is in contrast to our TDA method which has denoising properties, enabling to identify signals embedded due to otherwise non-observable geometrical patterns.

The methods also have been extended to inherently anonymize data, meaning that the geometrically transformed gene expression signatures cannot be reconverted to raw data by any means. While largely not a concern with tumor gene expression, these can therefore be used without revealing patient data if it is of a sensitive nature, for example in genomics data which may theoretically be used to re-identify individuals. This is also a useful additional component if

medRxiv preprint doi: https://doi.org/10.1101/2024.07.01.24309803; this version posted July 3, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

data must be transferred across various country of origin lines, where legal regulations may forbid the use of identifiable data, or raw data even when re-identification is not a concern.

Current methods in precision medicine for cancer typically prioritize mutations and addressable oncogenes, which derives from basic research and a need to identify the biological mechanisms underlying model of action of a compound. However, as the dependency map project illustrated, gene expression shows the strongest correlation with treatment response in cell lines [6]. Therefore, we take a more holistic view of determining treatment response from the aggregation of all gene expression data without consideration of underlying biological mechanisms. Using the measurements at all genes leverages the power of geometry where complicated relationships between variables may exist and is difficult to model, and where the use of some subset of variables is likely to reduce signal substantially. Because geometries are built in a completely unsupervised fashion and without any input from treatment response, these geometries are robust and downstream correlation with treatment is therefore unbiased. Nevertheless, permutation testing is utilized to determine which associations are statistically significant and thus less likely to be spurious. With this and any method, the gold standard is an independent replication cohort.

This study is limited by the accuracy of the data provided by Pfizer in the publication. Any errors in the collection, processing, measurement and analysis of the patient and tumor derived data may reduce the accuracy of our results. Characterization of clinical variables, such as the primary outcome PFS or times to censor may likewise alter results. Given that this is a published clinical trial, it is assumed that these errors will be minimized to those that are present in any strictly controlled study.

These methods can be applied in Phase II trials to determine inclusion/exclusion criteria for Phase III studies, as well as to create additional primary or secondary endpoints for analysis. This might increase the chances of passing Phase III substantially. In failed Phase III trials, these techniques may be able to detect similar subgroups of responders not visible with the analytical methods used at the time of the trial. This represents an enormous opportunity, since thousands of oncology drugs have failed Phase III trials, the vast majority due to lack of efficacy. Identifying the most efficacious subgroups in a post-mortem fashion may allow for application of a fast-tracked trial to confirm the finding. This could potentially bring a large number of previous failed cancer drugs to market as well as enhancing the treatment effect in subgroups of response, thereby benefiting patients suffering from an incurable disease.

An additional and related use case could involve combining and jointly analyzing data from various clinical trials, in order to find drug repurposing targets or expanded markets for existing therapies. Conducting such an endeavor across multiple cancer types and tissues of origin might provide valuable information to categorize oncology response using gene expression signature instead of testing based on a cancer type/tissue. This could expand the use and potential of many oncology drugs across a variety of cancer phenotypes.

Repeating this procedure on additional Phase III data is desirable, but difficult due to a lack of publicly available datasets. Initiation of collaborations with pharmaceutical consortia may be effective at creating a larger scale testing framework. Such efforts may be utilized not only in single compounds, but also across multiple oncology clinical trials to identify expanded use cases and potential drug repurposing across oncology phenotypes.

5 Conclusion

The application of our gene expression signatures and geometrical methods on Pfizer avelumab plus axitinib data identified three subgroups including slightly more than half of patients, which accounts for nearly all the observed treatment effect.

While this clinical trial met primary endpoints, in Phase II and Phase III studies the ability to detect subgroups of response may improve chances of a compound passing clinical trials dramatically.

6 Data and software availability

Pfizer's avlemuab plus axitinib clinical trial data are freely available at: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8493486/` under "Supplemental Materials". The analyses conducted in this paper are part of a proprietary codebase which is closed source and therefore not publicly available. A web application which will allow for some of the analyses contained within is under preparation.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.01.24309803; this version posted July 3, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

7 Acknowledgements

WC was partially supported by VR, the Wallenberg AI, Autonomous System and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation, and MultipleMS funded by the European Union under the Horizon 2020 program, grant agreement 733,161. Both WC and RR own equity in DatAnon Corporation, which owns the codebase on which this platform has been created as well as related intellectual property. WC and RR have applied for a patent on these methods for anonymizing data and creating gene expression signatures.

8 Author contribution

WC and RR designed the study, conducted experiments, analyzed results, and wrote the paper.

References

[1] Christophe Le Tourneau, Camille Perret, Allan Hackshaw, Jean-Yves Blay, Christoph Nabholz, Jan Geissler, Thy Do, Martina von Meyenn, and Rodrigo Dienstmann. An Approach to Solving the Complex Clinicogenomic Data Landscape in Precision Oncology: Learnings From the Design of WAYFIND-R, a Global Precision Oncology Registry. *JCO Precision Oncology*, (6):e2200019, December 2022.

[2] Giulia C. Napoli, William D. Figg, and Cindy H. Chau. Functional Drug Screening in the Era of Precision Medicine. *Frontiers in Medicine*, 9, July 2022.

[3] Vivek Subbiah, Mohamed A. Gouda, Bettina Ryll, Howard A. Burris III, and Razelle Kurzrock. The evolving landscape of tissue-agnostic therapies in precision oncology. *CA: A Cancer Journal for Clinicians*, n/a(n/a).

[4] Bcps Bcop Dpla Fccp Donald C. Moore, PharmD and Andrew S. Guinigundo, Msn, Rn, Cnp, Anp-Bc. Biomarker-Driven Oncology Clinical Trials: Novel Designs in the Era of Precision Medicine. *Journal of the Advanced Practitioner in Oncology*, 14(3):9–13, April 2023.

[5] Ha Yeong Choi and Ji-Eun Chang. Targeted Therapy for Cancers: From Ongoing Clinical Trials to FDA-Approved Drugs. *International Journal of Molecular Sciences*, 24(17):13618, September 2023.

[6] Steven M. Corsello, Rohith T. Nagari, Ryan D. Spangler, Jordan Rossen, Mustafa Kocak, Jordan G. Bryan, Ranad Humeidi, David Peck, Xiaoyun Wu, Andrew A. Tang, Vickie M. Wang, Samantha A. Bender, Evan Lemire, Rajiv Narayan, Philip Montgomery, Uri Ben-David, Colin W. Garvie, Yejia Chen, Matthew G. Rees, Nicholas J. Lyons, James M. McFarland, Bang T. Wong, Li Wang, Nancy Dumont, Patrick J. O'Hearn, Eric Stefan, John G. Doench, Caitlin N. Harrington, Heidi Greulich, Matthew Meyerson, Francisca Vazquez, Aravind Subramanian, Jennifer A. Roth, Joshua A. Bittker, Jesse S. Boehm, Christopher C. Mader, Aviad Tsherniak, and Todd R. Golub. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2):235–248, January 2020.

[7] Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *PNAS*, 108(17):7265–7270, 2011.

[8] Li Li, Wei-Yi Cheng, Benjamin S. Glicksberg, Omri Gottesman, Ronald Tamler, Rong Chen, Erwin P. Bottinger, and Joel T. Dudley. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med.*, 7(311), 2015.

[9] Wojciech Chachólski and Henri Riihimäki. A topological data analysis based classification method for multiple measurements. *BMC Bioinformatics.*, 21(336), 2020.

[10] Wojciech Chachólski and Henri Riihimäki. Metrics and stabilization in one parameter persistence. *SIAM J. Appl. Algebra Geom.*, 4(1):69–98, 2020.

[11] Oliver Gäfvert and Wojciech Chachólski. Stable invariants for multiparameter persistence, 2021.

[12] Jens Agerberg, Ryan Ramanujam, Martina Scolamiero, and Wojciech Chachólski. Supervised learning using homology stable rank kernels. *Frontiers in Applied Mathematics and Statistics*, 7, 2021.

[13] Gloria Colombo, Ryan J. Cubero, Lida Kanari, and et al. A tool for mapping microglial morphology, morphomics, reveals brain-region and sex-dependent phenotypes. *Nature Neuroscience*, 25:1379—-1393, 2022.

[14] Robert J. Motzer, Paul B. Robbins, Thomas Powles, Laurence Albiges, John B. Haanen, James Larkin, Xin-meng Jasmine Mu, Keith A. Ching, Motohide Uemura, Sumanta K. Pal, Boris Alekseev, Gwenaelle Gravis, Matthew T. Campbell, Konstantin Penkov, Jae Lyun Lee, Subramanian Hariharan, Xiao Wang, Weidong Zhang, Jing Wang, Aleksander Chudnovsky, Alessandra di Pietro, Amber C. Donahue, and Toni K. Choueiri. Avelumab

medRxiv preprint doi: https://doi.org/10.1101/2024.07.01.24309803; this version posted July 3, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

plus axitinib versus sunitinib in advanced renal cell carcinoma: Biomarker analysis of the phase 3 JAVELIN Renal 101 trial. *Nature medicine*, 26(11):1733–1741, November 2020.

[15] Jens Agerberg, Wojciech Chacholski, and Ryan Ramanujam. Global and relative topological features from homological invariants of subsampled datasets. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 302–312. PMLR, 2023.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.01.24309803; this version posted July 3, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

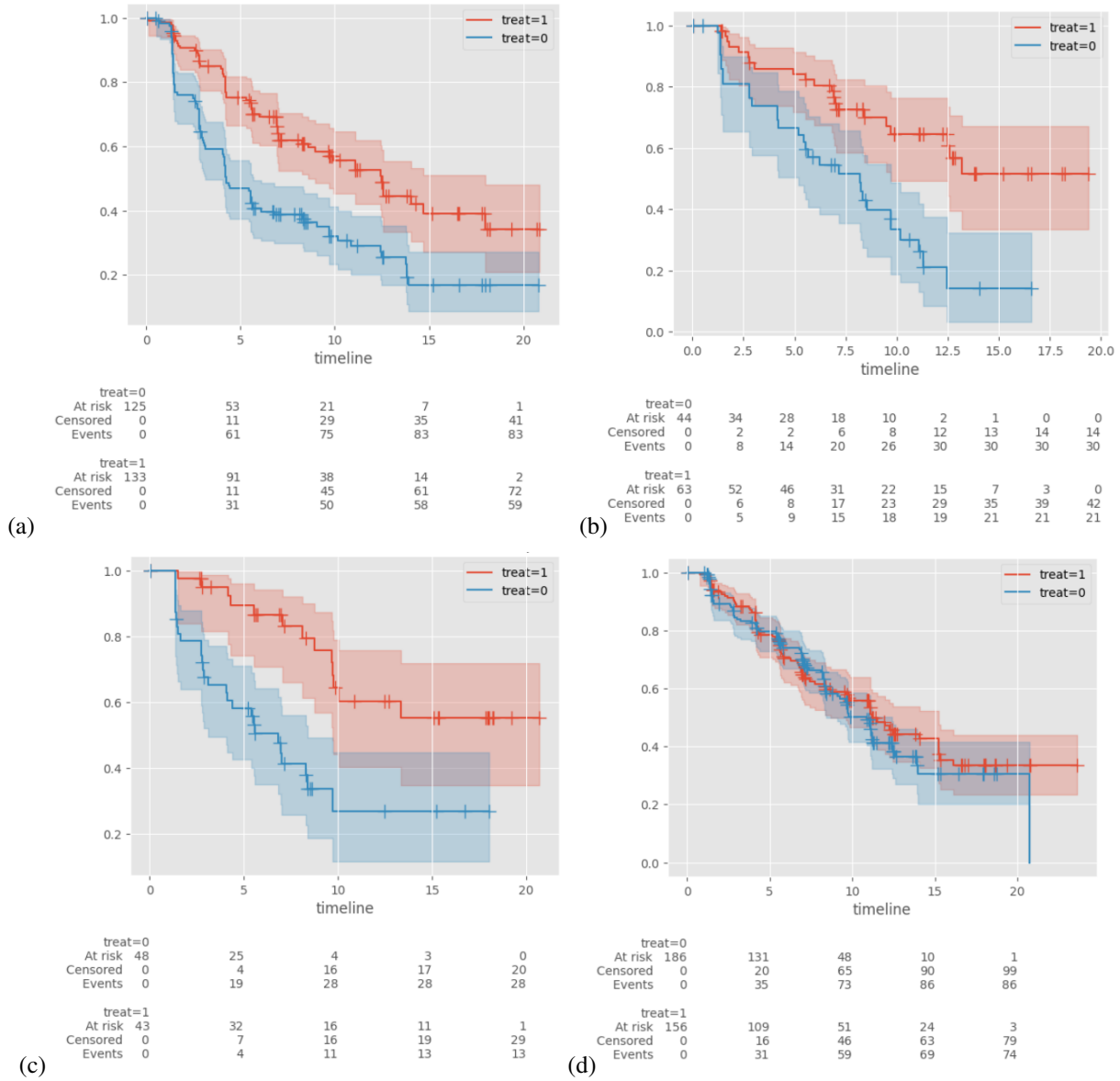

Figure 1: Kaplan-Meier estimator of subgroups created with geometrical gene expression signatures methodology. Three subgroups were detected, which when combined contained 384 patients (52.9% of the cohort). The p-value after permutation testing of these findings was 0.012. Treatment arm is indicated by (treat=1) while the comparator is (treat=0).

(a) Subgroup 1 ($n = 258$). Median time to PFS was 4.24 in the comparator arm and 12.48 in the treatment arm.

(b) Subgroup 2 ($n = 107$). Median time to PFS was 8.21 in the comparator arm and 50% line not crossed in the treatment arm.

(c) Subgroup 3 ($n = 91$). Median time to PFS was 6.80 in the comparator arm and 50% line not crossed in the treatment arm.

(d) Non-subgroup patients ($n = 342$). Median time to PFS was 9.72 in the comparator arm and 11.1 in the treatment arm.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.01.24309803; this version posted July 3, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

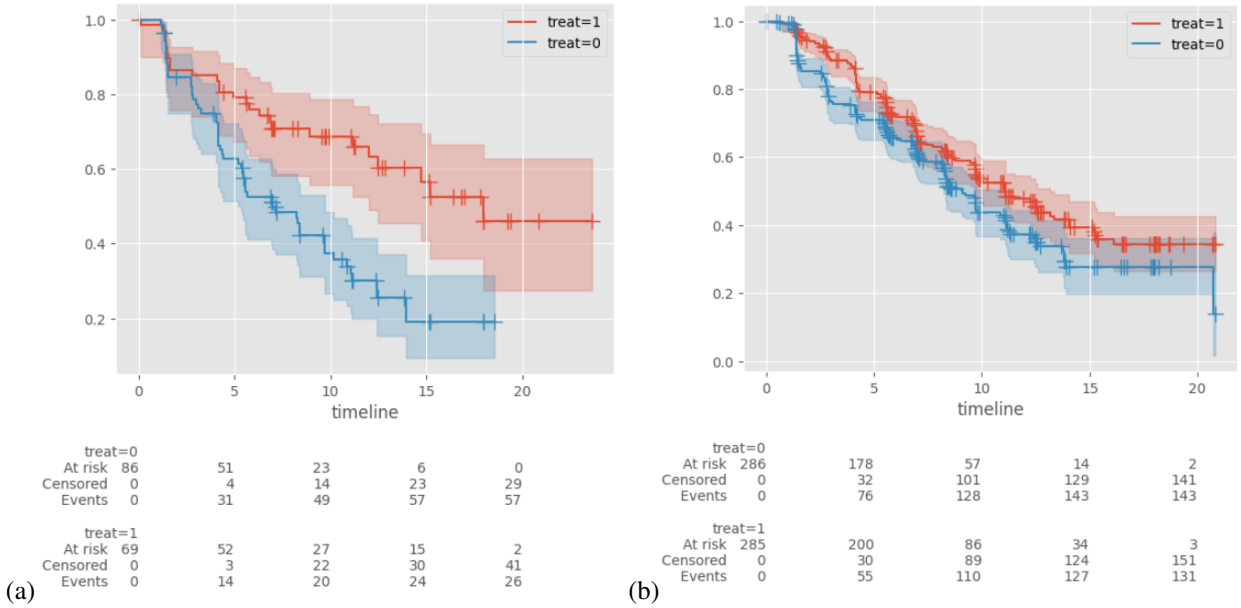

Figure 2: Kaplan-Meier estimator of subgroup and non-subgroup created with standard methodology

(a) Kaplan-Meier estimator of the significant subgroup created using t-SNE to reduce to two dimensions and DB Scan to produce clusters (n=155). Median time to PFS was 7.0 in the comparator arm (treat=0) and 17.97 in the treatment arm (treat=1).

(b) Kaplan-Meier estimator of patients not in the subgroup created using t-SNE to reduce to two dimensions and DB Scan to produce clusters (n=571). Median time to PFS was 9.1 in the comparator arm (treat=0) and 11.07 in the treatment arm (treat=1).