Title

Plasma-derived cell-free DNA methylomes might not enable detection and discrimination of intracranial tumors.

Running title

Matters Arising from Nassiri *et al*. Nat. Med. 2020

Authors

Kristoffer Vitting-Seerup[1]

Affiliations

[1]The Bioinformatics Section, Department of Health Technology, The Technical University of Denmark (DTU), Denmark.

Contact

k.vitting.seerup@gmail.com

Abstract

Intracranial tumors are both hard to detect and diagnose, resulting in poor patient outcomes. For that reason, non-invasive methods enabling detection and discrimination of intracranial tumors have significant clinical potential. Recently Nassiri *et al*.[1] propose plasma-derived cell-free DNA methylomes as such a method. Here I show that the results have been misinterpreted, and for many comparisons, no evidence supporting the conclusions are actually presented. While my analysis highlights the potential of plasma-derived cell-free DNA methylomes, the evidence provided by Nassiri *et al*.[1] is simply currently insufficient.

Introduction

Intracranial tumors are notoriously difficult to detect since most symptoms, such as headaches, are nonspecific and frequently occur in healthy people[2]. These tumors, therefore, often go unnoticed, resulting in late discovery and thereby worse patient outcomes[2]. When intracranial tumors are discovered, highly invasive intracranial surgery is needed to confirm and elaborate on the diagnosis[1]. Non-invasive diagnosis approaches capable of detecting and discriminating intracranial tumors could be vital for both screening and diagnosis purposes, ultimately improving patient outcomes.

Results

To this end, Nassiri *et al.*[1] recently reported they could detect and discriminate intracranial tumors using plasma-derived cell-free DNA methylomes. In their study, they created plasma-derived methylome profiles for 220 patients with various types of intracranial tumors. These methylome profiles were then combined with previously published methylome profiles from other cancer types and healthy controls (n=447). Using this dataset, Nassiri *et al*. test whether gliomas can be distinguished from extracranial samples and whether intracranial tumors can be distinguished from each other. In both cases, the analysis is framed as a one-vs-rest classification problem (e.g., glioma vs. all extracranial samples). Performance was measured on held-out validation data and reported as Area Under receiver operator Curve (AUC).

Combining a one-vs-rest framing (e.g., 60 gliomas vs. 447 extracranial samples) and measuring performance with AUCs is, however, potentially problematic. The problem is that AUCs can be misleading (or rather, misinterpreted) when applied to a classification problem where the numbers of samples in one class (e.g., 60 gliomas) is severely outnumbered by the number of samples in the other class (e.g., 447 extracranial)[3]. For such an imbalanced dataset, the problem arises because a model that predicts all samples as the major class (e.g., extracranial) can appear appealing when only using performance metrics such as accuracy and AUCs. The overall performance will simply seem adequate as the error caused by the misclassification of the minor group is negligible. But such a model is naturally not desirable. In the case of the glioma vs. extracranial classification, a model that predicts all samples as extracranial would have no clinical value. A common approach to avoiding this problem is to evaluate performance by measuring sensitivity and specificity[3]. Sensitivity and specificity, respectively, measure how large a fraction of each class was correctly predicted. It follows that one of them will be zero for the scenario where no minor class was predicted, whereby these problematic models can be avoided.

To investigate if the analysis done by Nassiri *et al.*[1] was robust to the problems caused by their imbalanced dataset, I re-analyze the predictions made (and provided) by Nassiri *et al*. using additional performance metrics. From this, it is clear that the analysis done by Nassiri *et al.* is currently not robust to the imbalanced nature of the datasets.

The problem with imbalanced datasets is most prominent for the analysis of intracranial samples. Here Nassiri *et al.*[1] analyze 161 intracranial cancer samples originate from 6 different cancer types: IDH mutant gliomas (n=41), IDH wild-type gliomas (n=22), 60 meningiomas, 9 hemangiopericytomas, 14 low-grade glial–neuronal tumors, and 15 brain metastases. To determine if one cancer type can be distinguished from the other cancer types, they use the one-vs-rest approach. This means they train a model to distinguish the 14 low-grade glial–neuronal tumors from all the 147 other cancer samples (8.7% vs. 91.3%). As reported in Nassiri *et al.*[1], the average resulting AUC for this comparison is 0.93. When inspected more thoroughly it the high AUC originate from classifying all samples as 'other'. In other words, no samples were actually classified as low-grade glial–neuronal tumors (Figure 1, leftmost plot). I find this exact pattern for four of the six cancer types (Figure 1), meaning these results have no clinical potential. The only cancer type with some performance was Meningioma, where 68% of Meningioma samples were correctly classified. While this result is probably still too low for any clinical applications, it indicates the potential of the data. Unfortunately, Nassiri *et al.*[1] currently does not provide the evidence to conclude that plasma-derived cell-free DNA methylomes enable discrimination of intracranial tumors.

Apart from the intracranial analysis, Nassiri *et al.*[1] also seek to determine if intracranial tumors can be distinguished from extracranial samples. Here a one-vs-rest approach is used to compare glioma samples to a dataset containing both other cancer types and healthy controls. This corresponds to comparing 60 gliomas samples (11.8%) to 447 non-gliomas samples (88.2%), meaning it is also an imbalanced dataset. As reported in Nassiri *et al.*[1], this comparison yields an average AUROC is 0.99, which considering a score of 1 indicates perfect classification, could seem very impressive. Such a conclusion would, however, be a misinterpretation of the AUROC values. When analyzed with additional performance metrics, the analysis has a median sensitivity of 0.818 (Figure 2A). The corresponding False Negative Rate (FNR) indicates that approximately 1 in 5 glioma patients would not be identified (Figure 2A). While the sensitivity indicates a clear potential, it still means Nassiri *et al.*[1] currently provide insufficient evidence to conclude plasma-derived cell-free DNA methylomes enable detection of intracranial tumors.

Another potential problem with the analysis of gliomas vs. extracranial samples is that Nassiri *et al.*[1] combine three distinct datasets originating from different laboratories (and distinct points in time). The joining of the datasets potentially results in batch effects where systematic variation between

samples is introduced because of differences in the laboratory conditions (e.g., temperature), personnel, etc.[4]. We know batch effects exist in the data analyzed by Nassiri *et al*. as they point to batch effects as the reason the data is analyzed as two individual cohorts (instead of one combined)[1]. It is, however, currently unclear how Nassiri *et al*. have corrected other potential batch effects as this is not described in the method section. It does, however, appear that batch effects have not been adequately handled as the glioma samples appear as two distinct clusters in a PCA analysis (Figure 2B) (also seen from Figure 1G in Nassiri *et al.*[1]). This clustering could result from a batch effect - a hypothesis supported by the clusters not being explained by the available clinical meta-data (Figure 2C).

Conclusion

In summary, due to both poor classification performance and potential batch effects, Nassiri *et al.*[1] currently provide insufficient evidence to conclude that plasma-derived cell-free DNA methylomes enable detection and discrimination of intracranial tumors.

Methods

The processed data, the scripts created, and the classification predictions made by Nassiri *et al*. were downloaded from their Zenodo repository (www.doi.org/10.5281/zenodo.3715312). A Rmarkdown document reproducing the analysis presented here can be found on Figshare (http://doi.org/10.6084/m9.figshare.14406866.v1). Performance metrics were calculated as defined defined in Saito *et al.*[3].

References

1. Nassiri, F. *et al.* Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nat. Med.* **26**, 1044–1047 (2020).
2. Silantyev, A. S. *et al.* Current and Future Trends on Diagnosis and Prognosis of Glioblastoma: From Molecular Biology to Proteomics. *Cells* **8**, (2019).
3. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, 1–21 (2015).
4. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-

throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).

Figure 1

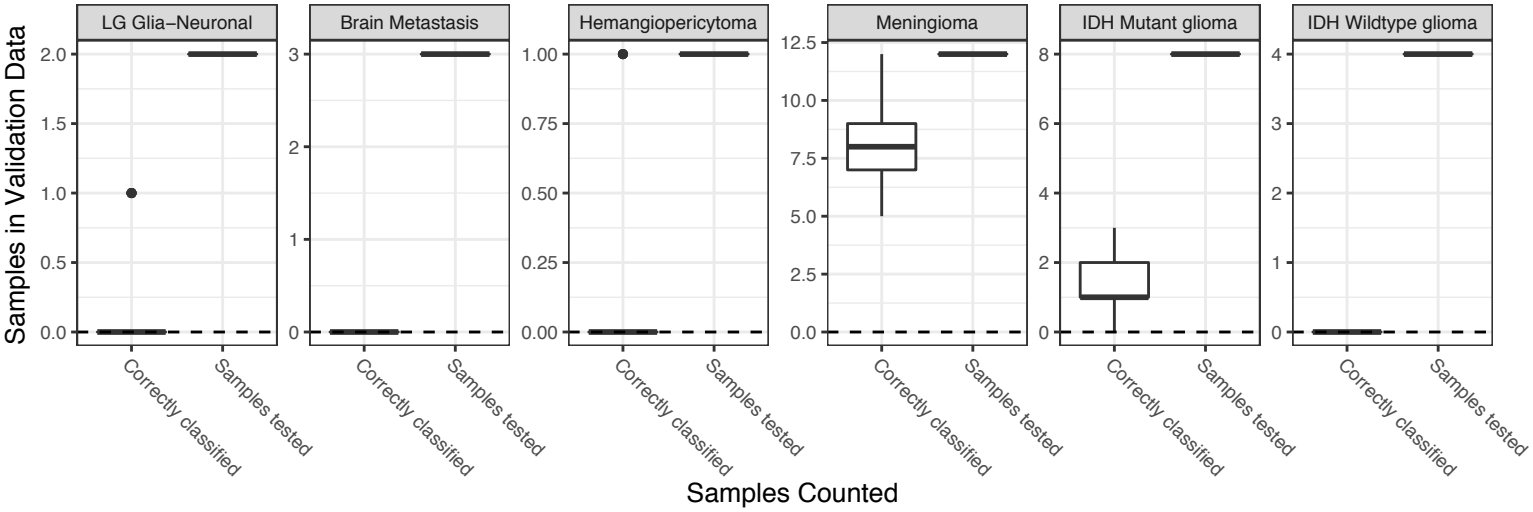

Figure 1: Distinguishing intracranial tumors. For each tumor type (sub-plots), the number of samples (y-axis) tested and the number of samples correctly classified (x-axis) for the 50 iterations of prediction on held-out data provided by Nassiri *et al.*

Figure 2

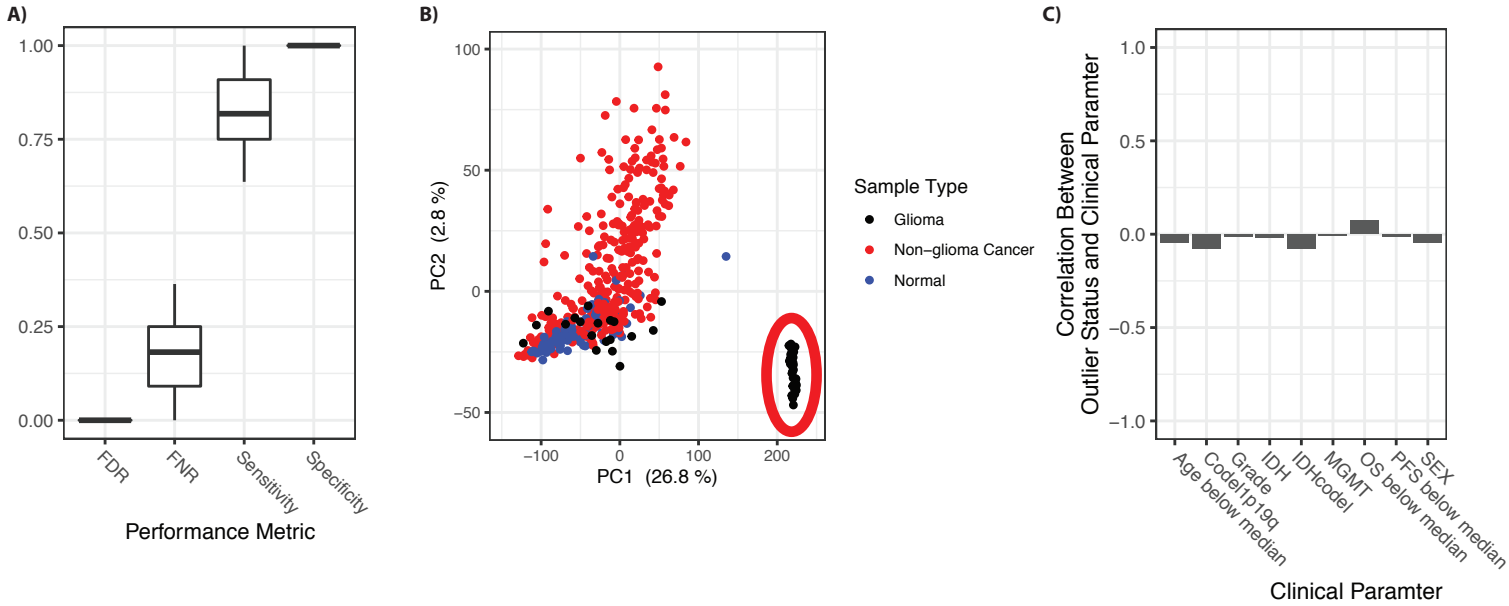

Figure 2: Detecting intracranial tumors. **A)** The x-axis indicates the different performance metrics calculated directly from the predictions provided in Nassiri et al. Y-axis indicates the performance for the 50 iterations of prediction on held-out data provided by Nassiri *et al*.. **B)** PCA plots of the 25,000 most informative methylated regions show the clustering of glioma and extracranial samples. Samples are colored by type, and a red ellipse highlights the glioma outliers. **C)** Correlation between glioma outlier status (as defined in B) and clinical metadata. OS: Overall Survival. PFS: Progression Free Survival.