Supplemental material

1. Bootstrapping for adjusting AUC estimate. To study the quality of the bootstrap adjusted AUC estimate described by Harrell et al.[1], we followed the bootstrap procedure and performed experiments on non-signal data sets of size 30 using 10 and 1000 features with Ridge regression (Ridge) and k-nearest neighbors (KNN) as classification methods.The number of bootstrap sampled with replacement from the original data set was 200. Each experiment was repeated 10 000 times to compute the mean and variance of the difference between the adjusted AUC estimate and the true AUC, formally defined as $\Delta\hat{A} = \hat{A} - 0.5$, where 0.5 is the true AUC for the data with no signal. Figure S1(a) presents the mean of $\Delta\hat{A}$ for the bootstrap adjusted AUC estimate (BOOTS) and for each of the cross-validation methods (i.e. LOO, LPO and TLPO) estimates as the positive class fraction varies. In all our experiments with non-signal data BOOTS shows an optimistic bias whereas the bias of the cross-validation methods is close to zero or is pessimistic as in the case of LOO. The variance of $\Delta\hat{A}$ for each of our experiments is shown in Figure S1(b), from these results we observe that BOOTS has variance close to zero and lower than the variance of the cross-validations methods. The high AUC value obtained by bootstrap even on data with no signal indicates that the experimenter can not distinguish between the cases in which the learning algorithm has overfit to the data or it has actually learned a useful classifier. Due to its low variance bootstrap can still be a valuable method when using classical statistical methods and low-dimensional data. It can become quite biased when working with more expressive methods that can always fit themselves to their training data to such extent that they can predict it perfectly. This is true for example of the KNN method, modern machine learning methods such as (deep) neural networks, and even linear models when having more features than sample units.

2. Level of inconsistency. In a tournament with m sample units the level of inconsistency can be measured, as explained in[2–4], by counting the number of circular triads (c) in the corresponding graph

$$c = m(m-1)(2m-1)/12 - \frac{1}{2}\sum_{i=1}^{m} S(i)^2 \ ,$$

where $S(i)$ is the score for unit i. The maximum number of circular triads in a tournament graph is defined by

$$c_{max}(m) = \begin{cases} \frac{m^3-m}{24} & when\ m\ is\ odd, and \\ \frac{m^3-4m}{24} & when\ m\ is\ even. \end{cases}$$

Kendall and Babington Smith[2] coefficient of consistency (ξ) is then

$$\xi = 1 - \frac{c}{c_{max}(m)} = \begin{cases} 1 - \frac{24c}{m^3 - m} & when\ m\ is\ odd, and \\ 1 - \frac{24c}{m^3 - 4m} & when\ m\ is\ even. \end{cases}$$

3. Tournament inconsistency when a random learning algorithm is used for training. A random learning algorithm ignores the training set and infers a prediction function, say with values in range $[-1, 1]$, that is randomly drawn from an uniform distribution over the set of functions $[-1, 1]^{\mathbb{N}}$. When a random learning algorithm is used with TLPO, the prediction functions inferred during different rounds of the cross-validation are independent of each other, and hence the number of cycles follow the distribution of cycle amounts of a tournament graph with random edge directions. In the case of TLPO with a random learning algorithm and a sample size of 30, the expected value of ξ is 0.1 indicating a high level of inconsistency.

4. Coefficient of consistency ξ of TLPO in the experiments with synthetic data and real medical data. Mean of ξ for each of the experiments in our study are presented in Figures S2, S3, S4. In most of our synthetic experiments, TLPO with KNN is less consistent than Ridge regression. Figures S2 and S3 show that with KNN ξ is 0.97 or below, while Ridge regression ξ is 0.96 or above. This may be the reason for the small negative bias observed on TLPO AUC estimate with KNN.

In our experiments with the real medical data set, TLPO with both learning algorithms, Ridge regression and KNN, has ξ above 0.97 (figure S4). In these experiments, the small negative bias previously observed in TLPO AUC estimate with KNN disappears.

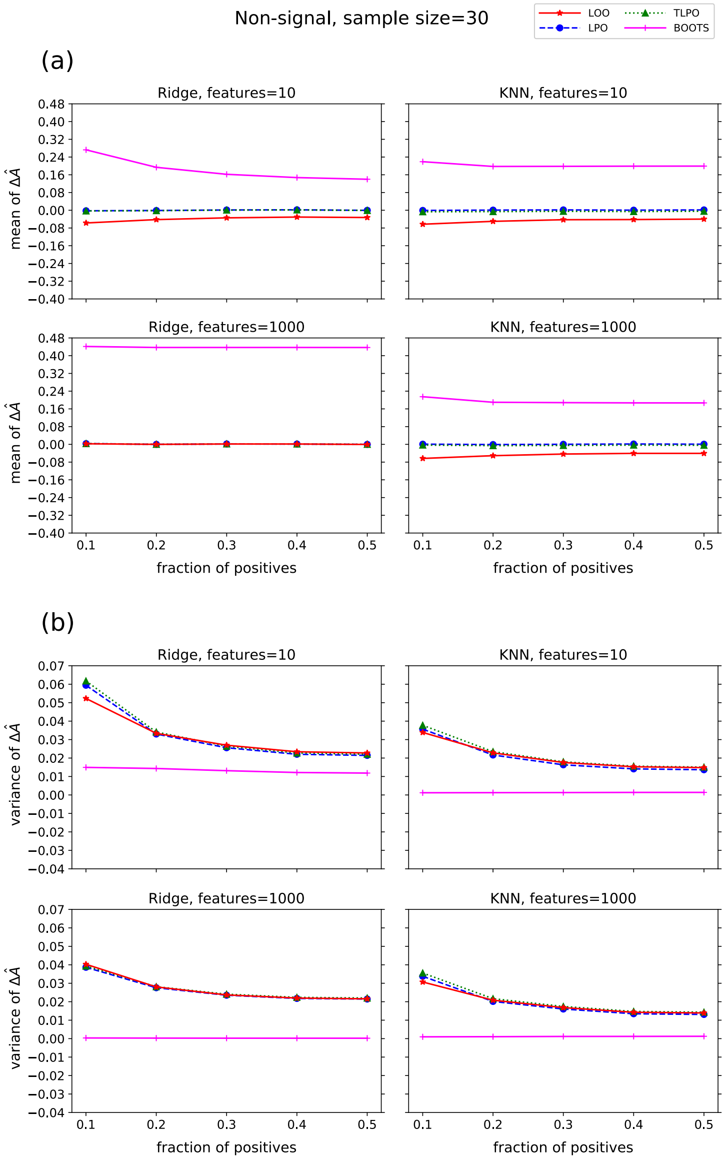

Figure S1. (a) Mean $\Delta\hat{A}_{CV}$ and (b) $\Delta\hat{A}_{CV}$ variance of LOO, LPO, TLPO, and BOOTS over 10 000 repetitions for all our experiments in non-signal data as class-fraction balanced. $\Delta\hat{A}_{CV}$: difference between estimated and true AUC; LOO: leave-one-out; LPO: leave-pair-out; TLPO: tournament leave-pair-out; BOOTS: method for adjusting optimistic bias present in the resubstitution estimate using 200 bootstrap samples; Ridge: ridge regression; KNN: k-nearest neighbors.

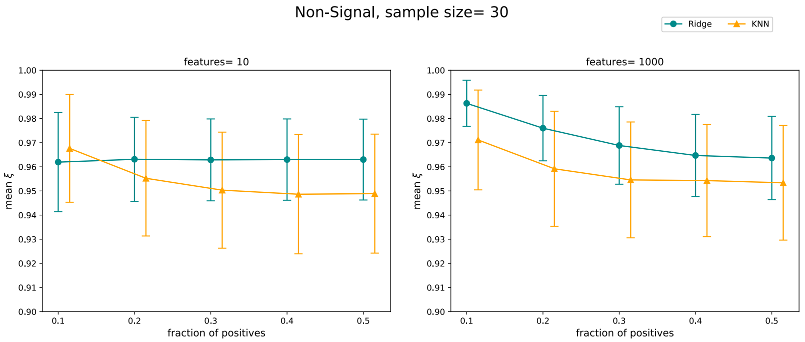

Figure S2. Mean of the coefficient of consistency ξ for TLPO on non-signal data, using Ridge and KNN as learning algorithms and varying the fraction of positives. Each experiment was repeated 10 000 times.

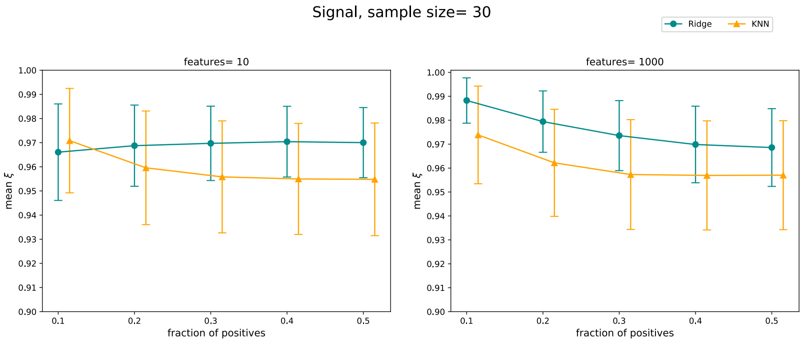

Figure S3. Mean of the coefficient of consistency ξ for TLPO on signal data, using Ridge and KNN as learning algorithms and varying the fraction of positives. Each experiment was repeated 10 000 times.

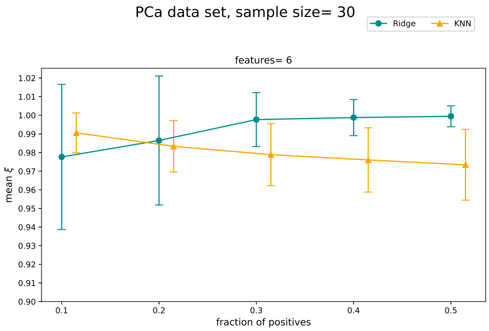

Figure S4. Mean of the coefficient of consistency ξ for TLPO on real medical prostate cancer (PCa) data set, using Ridge and KNN as learning algorithms and varying the fraction of positives. Each experiment was repeated 617 times.

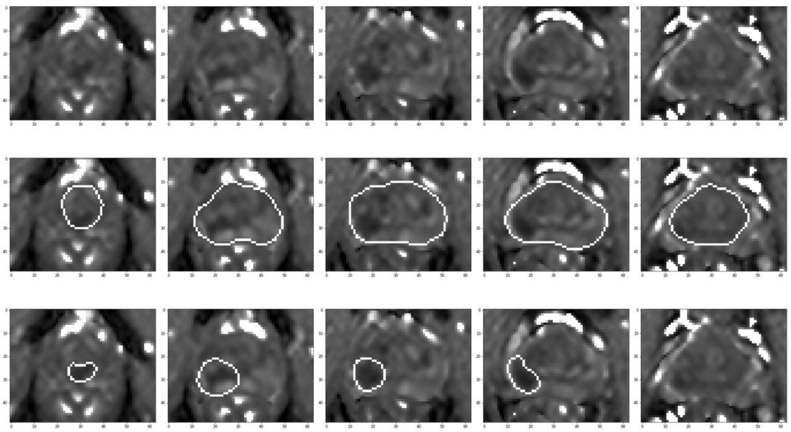

Figure S5. Images from patient no. 1. First Row: ADCm map of prostate. Second Row: prostate delineated. Third Row: tumor delineated

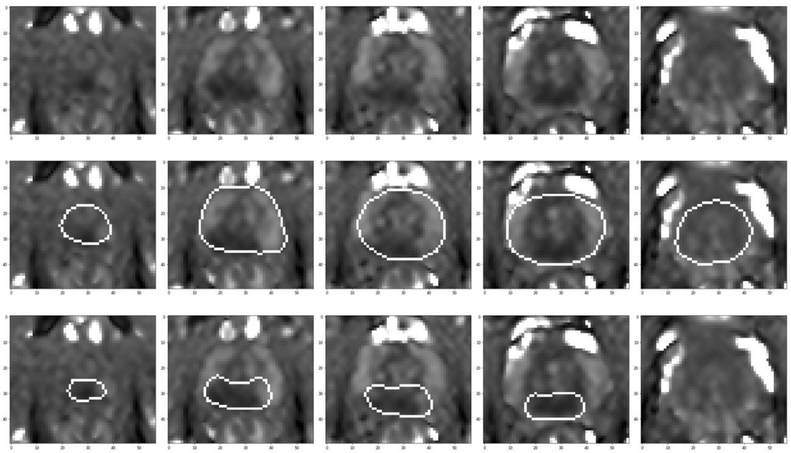

Figure S6. Images from patient no. 4. First Row: ADCm map of prostate. Second Row: prostate delineated. Third Row: tumor delineated

References

1. Harrell FE, Lee KL and Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* 1996; 15(4): 361–387.

2. Kendall MG and Smith BB. On the method of paired comparisons. *Biometrika* 1940; 31(3/4): 324–345.
3. Harary F and Moser L. The theory of round robin tournaments. *The American Mathematical Monthly* 1966; 73(3): 231–246.
4. Gass S. Tournaments, transitivity and pairwise comparison matrices. *Journal of the Operational Research Society* 1998; 49(6): 616–624.