HAUCA Curves for the Evaluation of Biomarker Pilot Studies with Small Sample Sizes and Large Numbers of Features

Frank Klawonn[1,2(✉)], Junxi Wang[1], Ina Koch[3], Jörg Eberhard[4], and Mohamed Omar[5]

[1] Biostatistics, Helmholtz Centre for Infection Research, Inhoffenstr. 7, 38124 Braunschweig, Germany
frank.klawonn@helmholtz-hzi.de
[2] Department of Computer Science, Ostfalia University of Applied Sciences, Salzdahlumer Str. 46/48, 38302 Wolfenbüettel, Germany
[3] Institute for Molecular Bioinformatics, Johann Wolfgang Goethe-University, Robert-Mayer-Str. 11-15, 60325 Frankfurt, Germany
[4] Department of Prosthetic Dentistry and Biomedical Materials Science, Hannover Medical School, Carl-Neuberg-Str. 1, Hannover, Germany
[5] Trauma Department, Hannover Medical School, Carl-Neuberg-Str. 1, Hannover, Germany

Abstract. Biomarker studies often try to identify a combination of measured attributes to support the diagnosis of a specific disease. Measured values are commonly gained from high-throughput technologies like next generation sequencing leading to an abundance of biomarker candidates compared to the often very small sample size. Here we use an example with more than 50,000 biomarker candidates that we want to evaluate based on a sample of only 24 patients. This seems to be an impossible task and finding purely random-based correlations is guaranteed. Although we cannot identify specific biomarkers in such small pilot studies with purely statistical methods, one can still derive whether there are more biomarkers showing a high correlation with the disease under consideration than one would expect in a setting where correlations are purely random. We propose a method based on area under the ROC curve (AUC) values that indicates how much correlations of the biomarkers with the disease of interest exceed pure random effects. We also provide estimations of sample sizes for follow-up studies to actually identify concrete biomarkers and build classifiers for the disease. We also describe how our method can be extended to other performance measures than AUC.

1 Introduction

A biomarker is a measurable value that is an indicator for a biological state. In recent years, the search for biomarkers for diseases has gained high interest in medicine. A well-known biomarker is the so-called prostate-specific antigen (PSA) which was or is sometimes still used as a biomarker for prostate cancer

H. Boström et al. (Eds.): IDA 2016, LNCS 9897, pp. 356–367, 2016.
DOI: 10.1007/978-3-319-46349-0_31

although its reliability and usefulness is sometimes doubted [1]. Along with the advancement of high-throughput technologies like microarrays, next generation sequencing and mass spectrometry, that allow to measure the whole or large parts of the genome, transcriptome, proteome or metabolome, came a strong hope to find a single biomarker for each disease or state of a disease to be diagnosed with very high certainty. However, this dream did not come true and it seems to be unrealistic from today's point of view. Biological systems are probably too complex for simple single-cause single-effect associations. Nevertheless, there are biomarker candidates that show a high correlation with specific diseases but are not reliable enough to function as predictors for the presence of a specific disease alone. Therefore, instead of relying on a single biomarker, the idea is to combine biomarkers that are not good enough for the diagnosis of a specific alone but can jointly provide a diagnosis with high certainty. An example of such a combination is the Enhanced Liver Fibrosis (ELF) score [2] that uses a linear combination of (log-)values of three single biomarkers to predict fibrosis stages in chronic liver disease patients. The EFL score was derived from a quite limited number of standard blood values of altogether 479 patients. No high-throughput technology was involved.

The use of high-throughput technologies for finding reliable combinations poses new challenges. First of all, in contrast to standard blood values, patient data based on high-throughput technologies are not commonly available in hospitals. This means, they have to be generated separately. Secondly, although the prices for generating data from high-throughput technologies are constantly decreasing, it is still quite expensive and also time-consuming to carry out these experiments. This implies high costs for such data. The advantage and the curse at the same time is that such experiments easily yield thousands or even far beyond 10,000 possible candidates for biomarkers. The sample size in expensive pilot studies is usually very limited, sometimes less than 20. From a machine learning or classification point of view one then tries to derive a classifier from a data set with more than 1000 attributes (biomarker candidates) and perhaps only around 20 or 30 instances. Finding random associations and overfitting is therefore hard to avoid.

Pilot studies with a small sample size and a large number of biomarker candidates are – as the name already points out – not intended to finally mark down a biomarker combination for clinical use but to check whether there are potential biomarkers that make a more expensive follow-up study with a larger sample size worthwhile. In this paper, we propose methods to assess biomarker pilot studies with respect to their potential to yield promising results when they are extended by follow-up studies. Section 2 gives a formal definition of our problem and provides an illustrating example. Section 3 describes how a pilot study can be analysed with respect to the potential to find reliable biomarker combinations in a follow-up study. A rough estimation of the required sample size is provided in Sect. 4. The ideas of Sect. 3 based on simple area under the ROC curve (AUC) values are extended to other measures for classifiers in Sect. 5.

The final conclusions address the problem of possible high correlations between biomarker candidates.

2 Problem Formalisation and an Example

From a formal point of view, we face the following problem. We have n instances – usually patients – from which we have measured m attributes (biomarker candidates). The patients are assigned to c different classes, i.e. different diagnoses or different states of a certain disease. The number of classes c is usually small, in many cases even $c = 2$ where we only want to distinguish patients suffering from a certain disease from patients who do not have this disease. Typically, we have $m \ggg n$. Our ultimate goal is to find a classifier that can predict the class (diagnosis) based on the values of the m attributes. For reasons of simplicity, we assume we do not have to deal with missing values[1]. Due to the fact that we have to face $m \ggg n$, we cannot directly build a standard classifier based on the given data set. A feature selection technique is required to reduce the number of attributes drastically. As a possible way to evaluate the predictive power of a classifier, we could apply cross-validation and because of the small sample size we would prefer to use leave-one-out cross-validation (the jackknife method). It must be emphasised that when we want to evaluate a classifier, we must not separate feature selection from the classifier. It was already noted in [3,4] that first applying feature selection on the whole data set and then evaluate a classifier using only the selected features based on cross-validation can lead to a strong model selection bias, since the actual model consists of the classifier *and* the (pre-)selected features. To illustrate this problem, we have carried out the following simulation. We have generated $m = 1000$ random attributes following a standard normal distribution for different sample sizes n. The we have randomly assigned the n instances to two classes, $n/2$ instances to each class, i.e. if we consider the two classes as healthy vs. sick, we have a prevalence of 50 %. This means that correlations between the 1000 attributes and the classes are purely random. We have then carried out the following two experimental settings.

(a) We have first selected the best 20 attributes from the whole data set and then trained classifiers and evaluated them based on the leave-one-out method (LOO). This is how it should not be done!
(b) Within the leave-one-out method, i.e. when the test sample had already been removed from the training data set, we have selected the best 20 attributes and trained the classifiers without the sample that was left out for testing.

The selection of the "best" attributes was based on a very simple strategy. We chose 20 attributes with the highest area under the ROC curve (AUC) values. As classifiers we used support vector machines (SVM), random forests (RF)

[1] This is a more or less realistic assumption for microarray and next generation sequencing data but not for data from mass spectrometry.

Table 1. Percentage of correctly classified instances in a completely random data set with 1000 features when feature selection is applied to the whole data set before (before LOO) and within leave-one-out cross-validation (during LOO). The sample size with two classes is given by n. Prevalence is 50 %. Classifiers are support vector machines (SVM), random forests (RF) and linear discriminant analysis (LDA).

n	before LOO			during LOO		
	SVM	LDA	RF	SVM	LDA	RF
20	100	80	95	65	65	55
30	97	70	97	43	43	67
40	95	83	85	53	55	45
50	84	80	82	26	20	22
100	82	80	76	47	46	48
150	79	79	79	60	61	60
200	72	70	70	44	48	44

and linear discriminant analysis (LDA). Table 1 shows the percentage of correctly classified instances for the leave-one-out evaluation. Since the data set is completely random with a prevalence of 50 %, we would expect to classify about half of the instances correctly. One can see easily see that this is not true for the (inappropriate) method explained in (a) where the feature selection is carried out on the whole data set before leave-one-out cross-validation. Even for a sample size of 200, around 70 % of the instances are still correctly classified, a value that is never achieved for any sample size n and any classifier with the correct method (b). It is noteworthy that for $n = 50$ the correct method by chance performs even far worse than random guessing.

As an example for illustration purposes of our approach we use a data set from $n = 24$ patients[2] who had undergone a surgery for a hip prosthesis which later on caused problems. The final goal is to classify whether the problems are caused by a low-grade periprosthetic hip infection or by aseptic hip prosthesis failure, i.e. to see whether the problems come from an infection or not. A microarray kit was used to obtain $m = 50,416$ biomarker candidates based on measured genes and RNA values [5]. When we apply the above mentioned method (b) to this data set, we obtain rates of correctly classified instances of around 50 % which corresponds to random guessing. Even changing the number of selected biomarker candidates – for instance choosing only the top 10 or 4 instead of 20 – in the leave-one-out cross-validation loop does not lead to an improvement. Should we draw the conclusion that this pilot study has failed and it is not worthwhile to consider a follow-up study? An answer to this question will be provided in the following section.

[2] The data set is currently submitted to a medical journal.

Table 2. Top 10 AUC values and their p-values for the hip prosthesis infection data set.

Biomarker	AUC	p-value (raw)	p-value (corrected)
1	0.951388889	0.0000333	1
2	0.944444444	0.0000496	1
3	0.944444444	0.0000496	1
4	0.930555556	0.0001028	1
5	0.930555556	0.0001028	1
6	0.930555556	0.0001028	1
7	0.930555556	0.0001028	1
8	0.930555556	0.0001028	1
9	0.923611111	0.0001442	1
10	0.923611111	0.0001442	1

3 HAUCA Curves

One might argue that our feature selection method using only the AUC values is too simple. Indeed, there are more sophisticated techniques, especially those that do not simply rank the single features but look directly for combinations of features. However, it is extremely difficult to choose among feature subsets from more than 50,000 features. We do not want to dive into advanced feature selection methods here. In any case, it would be unrealistic to build a classifier based on a data set with 24 instances (patients).

Although the concept of AUC is sometimes criticised [6], especially because it does not take the prevalence into account, and it is restricted to two-class problems, it is still a meaningful approach [7] and we will take a closer look at the AUC values in our data set. The second column of Table 2 shows the 10 highest AUC values for our example data set – computed on the whole data set.

AUC values over 0.9 are definitely interesting although such a value might not be sufficient for a medical test. Nevertheless, a combination of biomarkers with such high AUC values might lead to a classifier with sufficient predictive power. However, we have seen in the previous section that this does not really apply in the case of our data set. The classifiers we had mentioned using the biomarker candidates with the highest AUC values – based on the feature selection and cross-validation strategy (b) – were not better than random guessing. So can we conclude that the high AUC values occur just by chance?

Fortunately, there is a method to compute the probability that an attribute with n values randomly assigned to two classes with a given prevalence exceeds a given AUC value [8]. The computation is mainly based on the same statistic that is used for the Wilcoxon-Mann-Whitney-U test. We just need to be able to compute the quantile of the statistic of the Mann-Whitney-U test. For a sample of size n with an absolute prevalence of n_+ of the disease in the sample, the

probability that a single biomarker candidate with randomly assigned values exceeds an AUC value of a is

$$F_U((n - n_+) \cdot n_+ \cdot (1 - a), n - n_+, n_+) \tag{1}$$

where F_U is the cumulative distribution function of the Wilcoxon-Mann-Whitney-U statistic [8]. Actually this probability should be multiplied by two (and cut off at 1 in case it exceeds 1) because an AUC value close to 0 is also of high interest because it indicates a high, but negative correlation between the value of the biomarker candidate and the disease.

The corresponding probabilities for the top 10 AUC values in hip prosthesis infection data set are given in the third column "p-value (raw)" of Table 2. These probabilities can be interpreted as p-values for the hypothesis test with the null hypothesis that the classes are randomly assigned to the instances or that all biomarker candidates have random values that are not correlated with infection. Although the raw p-values in the third column of Table 2 seem to be small enough to reject the null hypothesis, we must take into account that we have applied multiple testing, i.e. we have applied the test to all biomarker candidates, so that the test was repeated $m = 50,416$. Therefore, a correction for multiple testing is needed. No matter which correction method we choose – here we have applied Bonferroni-Holm correction [9] – all p-values are changed to 1, losing their significance as can be seen from the last column of Table 2. So does this support what we have already observed when we constructed the classifiers, i.e. that the data set does not seem to indicate any non-random correlation between the biomarker candidates and infection?

The answer is no. There is another way of looking at the AUC values. Because of the large number of biomarker candidates, we would expect high AUC values just by chance. But how many high AUC values could we expect if the data were completely random? According to Eq. (1) we can compute the probability that an attribute with n values randomly assigned to two classes with a given prevalence exceeds a given AUC value, we can also calculate the number of expected biomarker candidates exceeding a given AUC value in a random data set. It is simply the probability for the AUC times the number of biomarker candidates m, here $m = 50,416$.

The bottom black curve in Fig. 1 shows on the y-axis how many biomarker candidates one could expect to exceed a given AUC value marked on the x-axis in a random data set that contains the same number of biomarker candidates as our hip prosthesis infection data set. The top blue curve in the figure shows how many biomarker candidates exceed the corresponding AUC threshold in our real data set on infection after hip prosthesis surgery. One can clearly see that there are many more biomarker candidates in the AUC range from 0.85 to over 0.9 than one would expect in a random data set.

Of course, the expected number of high AUC values in a random data set is just a bottom line for comparison with the AUC values found in the real data set. We also provide a 95 % upper confidence band for the number of high AUC values in a random data set. This confidence band is indicated by the middle red

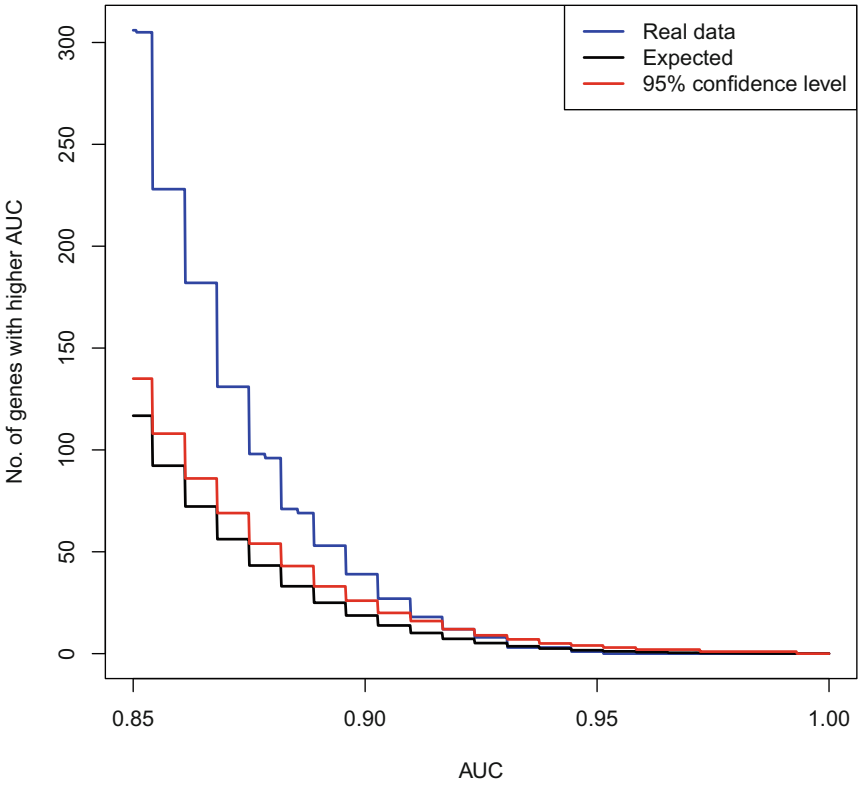

Fig. 1. HAUCA curves for the hip prosthesis infection data set showing on the y-axis how many biomarker candidates exceed a given AUC threshold indicated on the x-axis. Top blue curve: real data. Bottom black curve: Expected number in a random data set. Middle red curve: 95 % upper confidence band for a random data set. (Color figure online)

line in Fig. 1. Again, one can observe that the number of biomarker candidates in the range between 0.85 and 0.9 exceeds even this 95 % upper confidence band.

The computation of the confidence band is based on the following considerations. For any given AUC value a we know the probability p_a that a biomarker candidate with random values would obtain an AUC value larger than a. If we have m independent random biomarker candidates the probability that exactly k random biomarker candidates have a higher AUC value than a follows a binomial distribution $B(m, p_a)$, so that we simply have to compute the 95 % quantile of this binomial distribution to obtain the value of the 95 % upper confidence band at the AUC value a. Of course, one could choose other values than 95 % for the confidence band and just replace the values in the graph by the corresponding quantile of the binomial distribution.

We call the curves shown in Fig. 1 high AUC abundance (HAUCA) curves. The HAUCA curves in Fig. 1 clearly indicate that there is more than just a random correlation between the biomarker candidates and infection with the hip prosthesis infection data set. Taking a closer look at the AUC value of 0.85 in the HAUCA curves, we can see that the real data set contains over 300 biomarker candidates with an AUC value higher than 0.85, whereas one would expect in a random data set clearly less than 150. Even the 95 % upper confidence bound at an AUC value of 0.85 does not reach the value 150. This means that the high correlation with infection of about half of the biomarker candidates in the real data set with AUC value greater than 0.85 cannot be explained by pure random effects. We cannot identify which biomarker candidates are the right ones. But there should be some valid biomarkers that once – once they are identified in a follow-up study – can be used to build a classifier.

So from this pilot study on infection after hip prosthesis surgery we cannot confirm any specific biomarker candidates. But we can nevertheless say that there must be very good candidates and it is worthwhile to extend the pilot study to a larger sample. Of course, one could also look at the functional annotations of the genes (biomarker candidates) with high AUC values and see which ones are associated with infection processes to make a pre-selection of promising biomarker candidates for an extended study. But this is out of the scope of our purely statistics oriented discussion here.

Figure 2 shows another example of HAUCA curves for data from a biomarker study published in [10][3]. The data set contains information about the microbiome in the mouth of $n = 19$ patients of which 9 suffered from periodontitis. The microbiome was characterised by the abundance of $m = 242$ operational taxonomic units (OTUs). Here again the HAUCA curves clearly indicate that there is more than just random correlation between OTUs and periodontitis. In this case, the classifiers based on leave-one-out cross-validation and feature selection within the cross-validation loop could even provide for 14 out of 19 patients the correct diagnosis which is far away from being of clinical use but also indicates a correlation between OTUs and periodontitis.

4 Sample Size Estimation for Follow-Up Studies

In the previous section, we have seen that based on very small pilot studies we can at least find good indicators for a connection between the set of biomarker candidates and the disease under consideration. But our approach is neither capable to identify specific biomarkers nor to construct a reliable classifier that supports the diagnosis which is no surprise given the small sample size compared to the number of biomarker candidates. The question that arises here is how to determine the sample size of a follow-up study that should either confirm the validity of biomarkers with high AUC values or even construct a classifier based on a combination of good biomarker candidates.

[3] The HAUCA curves were neither available nor discussed in the paper [10].

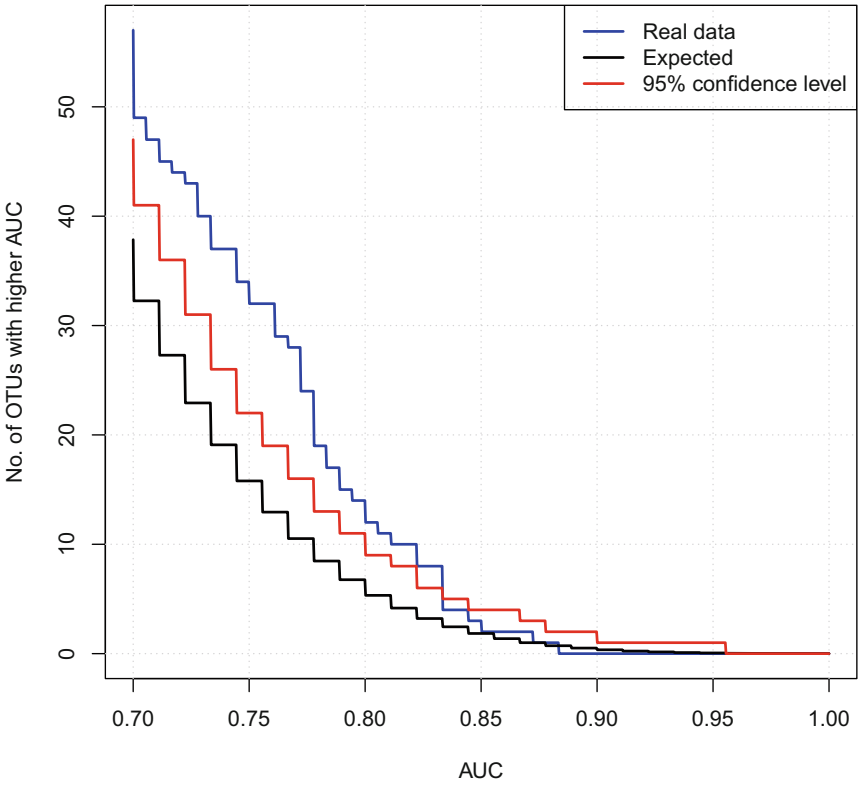

Fig. 2. HAUCA curves for the OTU data set.

It is quite simple to specify the sample size for confirming the validity of a high AUC value for a biomarker. For instance, the top biomarker candidate in Table 2 has an AUC value of 0.95 whose low raw p-value turns into a non-significant p-value after correction for multiple testing. Assuming that the AUC value of this biomarker would remain at 0.95 in a follow-up study with a larger sample size, we can compute the resulting p-value in this study after correction for multiple testing using Eq. (1) (multiplied by two to account for biomarkers that show a negative correlation with infection). We simply need to increase n and n_+ proportionally – assuming the prevalence in the follow-up study remains the same – until the resulting probability is small enough, so that it remains significant after correction for multiple testing, in our example after multiplication with $m = 50,416$. Already at $n = 30$, the p-value after correction for multiple testing drops below 5 %. In the same way, the two last candidates in Table 2 with an AUC value of 0.92 would need a sample size of at least $n = 36$ to be confirmed given the AUC and the prevalence remain the same for the larger sample size in the follow-up study.

Apart from confirming high AUC values of single biomarker candidates we are also interested in how many of such potentially valid biomarkers we need to combine for a classifier to obtain sufficiently reliable predictions. This, of course, highly depends on the correlation between the good biomarker candidates. In the worst case, their correlation equals 1 and their combination does not lead to any improvement compared to the single biomarkers. In order to get a rough idea of how well a classifier based on a combination of good biomarkers could perform, one could exploit the ideas from [11] although the underlying assumptions are quite restrictive. There it is assumed that the values for the biomarkers follow normal distributions and the normal distributions for the two classes differ. The paper [11] provides a method how to compute the AUC value of linear discriminant analysis based on the AUC values of the single biomarkers and their correlations within the two classes. Of course, the estimation of the correlation based on the pilot study is not very reliable. But the proposed procedure of estimating the AUC values of the biomarker combination still provides a rough judgment how well the biomarker combination could perform for later prediction.

5 Alternatives to AUC

As mentioned already in the beginning of Sect. 3, AUC values are neither the only nor the best performance measure of scores used for classification. In principle, one could replace AUC by other performance measures, for instance entropy, accuracy or the area above cost curves [12]. These performance measures also have the advantage that they are not restricted to binary classification problems but are also applicable in the context of multiclass classification problems. Of course, for multiclass classification problems one could also use extensions of AUC to more than two classes as described in [13–17]. However, for all these measures it is no longer obvious how the corresponding p-values can be computed that are needed for the equivalent to the HAUCA curves.

A possible solution is an estimation of these p-values based on Monte-Carlo or permutation tests. For a Monte-Carlo test, one would generate a large number of biomarkers with random values and compute the values of the corresponding performance measure. Then the p-value of a biomarker candidate in the real data set is the proportion of random biomarkers with a better value for the performance measure than the considered biomarker in the real data set. For a permutation test, one would randomly permute the classes while fixing the values of the biomarker candidates to compute values of the performance measures for random biomarkers. The p-value of a biomarker candidate in the real data set is then computed in the same way as for the Monte-Carlo test. Estimating the p-values based on Monte-Carlo or permutation tests requires a large number of simulations implying high computational costs. For instance, in the example of the hip prosthesis infection data set the best AUC value has a p-value of approximately $3 \cdot 10^{-5}$. If we had not been able to compute this p-value based on Eq. (1) and had to rely on Monte-Carlo or permutation tests, we would need

at least 10^6, better even more than 10^7 simulations to get a rough estimation of this small probability. And the above mentioned performance measures already need a little bit of computation time for a single biomarker.

6 Conclusions

In this paper, we have presented an approach how to judge biomarker pilot studies with small sample sizes and large numbers of possible biomarker candidates. For binary classification problems we can use the AUC as a measure of performance for the single biomarkers, leading to closed form solutions of the required calculations and therefore to fast computation. Other performance measures could also be applied for the price of high computational costs due to the need of simulations instead of closed form solutions. Efficient algorithms or new solutions will be a topic of further research.

Another question concerns the correlation between the biomarker candidates. The computation of the p-values in the context of AUC values and for the Monte Carlo test assumes independent biomarker candidates. This is definitely an unrealistic assumption because at least subsets of the biomarker candidates – no matter whether they are associated with the disease or not – will show high correlations because these measured values interact in a highly complicated biological system and cannot function independently. In a certain way, this would be taken into account by a permutation test because the correlation among the biomarker candidates is not changed, only the distribution of the classes is rearranged. This aspect will need further investigations.

References

1. De Angelis, G., Rittenhouse, H., Mikolajczyk, S., Blair, S., Semjonow, A.: Twenty years of PSA: from prostate antigen to tumor marker. Rev. Urol. **9**(3), 113–123 (2007)
2. Lichtinghagen, R., Pietsch, D., Bantel, H., Manns, M., Brand, K., Bahr, M.: The enhanced liver fibrosis (ELF) score: normal values, influence factors and proposed cut-off values. J. Hepatol. **59**(2), 236–242 (2013)
3. Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. **99**(10), 6562–6566 (2002)
4. Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC Bioinform. **7**(91), 1 (2006). doi:10.1186/1471-2105-7-91
5. Omar, M., Klawonn, F., Brand, S., Stiesch, M., Krettek, C., Eberhard, J.: Transcriptome-wide high-density microarray analysis reveals differential gene transcription in periprosthetic tissue from hips with low-grade infection versus aseptic loosening. J. Arthroplasty (2016, to appear). doi:10.1016/j.arth.2016.06.036
6. Hand, D.: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach. Learn. **77**, 103–123 (2009)
7. Flach, P., Hernández-Orallo, J., Ferri, C.: A coherent interpretation of AUC as a measure of aggregated classification performance. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 657–664 (2011)

8. Mason, S.J., Graham, N.E.: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q. J. Royal Meteorol. Soc. **128**(584), 2145–2166 (2002)
9. Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**, 65–70 (1979)
10. Szafranski, S., Wos-Oxley, M., Vilchez-Vargas, R., Jáuregui, R., Plumeier, I., Klawonn, F., Tomasch, J., Meisinger, C., Kühnisch, J., Sztajer, H., Pieper, D., Wagner-Döbler, I.: High-resolution taxonomic profiling of the subgingival microbiome for biomarker discovery and periodontitis diagnosis. Appl. Environ. Microbiol. **81**, 1047–1058 (2015)
11. Demler, O., Pencina, M., D'Agostino, R.S.: Impact of correlation on predictive ability of biomarkers. Stat. Med. **32**, 4196–421 (2013)
12. Montvida, O., Klawonn, F.: Relative cost curves: An alternative to AUC and an extension to 3-class problems. Kybernetika **50**, 647–660 (2014)
13. Hand, D., Till, R.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach. Learn. **45**, 171–186 (2001)
14. Li, J., Fine, J.: ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. Biostatistics **9**, 566–576 (2008)
15. Li, J., Fine, J.: Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. J. Stat. Plan. Infer. **139**, 4133–4142 (2009)
16. Hernández-Orallo, J.: Pattern Recogn. ROC curves for regression **46**(12), 3395–3411 (2013)
17. Novoselova, N., Della Beffa, C., Wang, J., Li, J., Pessler, F., Klawonn, F.: HUM calculator and HUM package for R: easy-to-use software tools for multicategory receiver operating characteristic analysis. Bioinformatics **30**, 1635–1636 (2014)