A summary of statistical analyses of bioimpedance method for detection of osteoporosis using Bone Vitae company's device

Andrzej Giniewicz*

3.11.2014

Abstract

The purpose of this document is to create a short summary of statistical analyses of bioimpedance method for detection of osteoporosis using Bone Vitae company's device. A clinical in vivo trials have been carried out at SYNEXUS Medical Centre in Warsaw. Synexus Warsaw is a part of the world's largest multi-national company entirely focused on the recruitment and running of clinical trials at its own Dedicated Research Centres with HQ in Manchester. The director of clinical trials was Dr. Andrzej Sawicki. The permission no. KB/802/11 to carry out the clinical trial has been issued by Bioethical Commission of Regional Medical Chamber in Warsaw.

This summary is is based on a series of reports by Przemysław Biecek, PhD. The reports showed, that it is not possible to classify well into three classes: healthy, ostopenia and osteoporosis, but one can obtain good results by classifying patients into two groups: "healthy" and "not healthy". Those classes correspond to T-Score value of more than -2 and less than -2 respectively, so all cases of osteoporosis are considered "not healthy" and all cases of patients without osteoporosis or osteopenia are considered "healthy". Patients with mild osteopenia are considered "healthy" and with severe case of osteopenia are considered "not healthy". Final examination was conducted on group of 27 patients, whose T-Score was measured using DEXA to find true state of patient. Based on those observations, a random forest model was fitted, using all 44 variables measured by device, age, weight and height of patient — even though current version of device does not allow inputting those additional patient data.

Additional analysis was carried in this summary report, to validate the earlier model. It has been shown, that we should expect accuracy of around 77%, unless device or measuring method will be improved. If device included an interface to input weight, height and age, accuracy will go up by few percent, but real improvement could be obtained by multiple measurements or decrease in their variability, resulting in over 80% of accuracy, or nearly 90% of sensitivity while keeping 74% of specificity.

1 A general description of the conducted experiments

Statistical analysis was conducted in two stages using data from two examination sessions. Both sessions were made using different versions of device, improved based on previous experiences. A clinical in vivo trials have been carried out at SYNEXUS

[*] Andrzej Giniewicz (Andrzej.Giniewicz@pwr.edu.pl) — independent expert, got Masters of Engineering in Mathematical Statistics from Wrocław University of Technology in 2009, since 2013 is working as an Assistant in Mathematics Department at the same university. In 2014 submitted the Ph.D. thesis in Mathematical Statistics, titled "Optimal Sequential Procedures in the Life-testing Problems."

Medical Centre in Warsaw. Synexus Warsaw is a part of the world's largest multinational company entirely focused on the recruitment and running of clinical trials at its own Dedicated Research Centres with HQ in Manchester. The director of clinical trials was Dr. Andrzej Sawicki. The permission no. KB/802/11 to carry out the clinical trial has been issued by Bioethical Commission of Regional Medical Chamber in Warsaw.

The device uses bioimpedance method for detection of osteoporosis, by applying alternating current with known constant amplitude and one of eleven frequencies, then reading a complex value describing relation of voltage and current. The device performs measurements using a number of electrodes attached to an arm of patient. The number varied between examination sessions — first version of analysed device had ability to measure using two, four and six (two variants, A and B) electrodes, second version of analysed device used two, four and five electrodes. The eleven frequencies were: 254Hz, 509Hz, 1017Hz, 2035Hz, 3815Hz, 7884Hz, 15513Hz, 31281Hz, 62561Hz, 124868Hz and 249990Hz, where the highest one is the highest value of band 1 frequency range. For each frequency a complex impedance-like measure was calculated and converted into four real valued parameters: Z', Z'', $|Z|$ and ϕ. The Z' and Z'' are real and imaginary part of measured value, so they can be seen as related to resistance and reactance, while $|Z|$ and ϕ are absolute value and argument of same complex value, so they can be seen as related to absolute impedance and phase shift. Altogether, a single measurement consists of 44 real valued readouts, 4 values for each of 11 frequencies. In both examination sessions a series of measures was conducted, and recorded together with age, weight and height of patient, making a total of 47 predictive variables.

All patients were examined using Dual-energy X-ray absorptiometry (DXA or DEXA), to calculate T-Score for various points of their body, including vertebral column, hip and arm. The T-Score below -2.5 indicates osteoporosis and T-Score below -1 osteopenia. Also, BDM (bone density) was gathered and Z-Score — a comparison with age-matched normal, measured by the number of standard deviations of BDM from the average BDM for given age. Those variables were used to determine true state of patient for analysis purposes.

2 First analysis

First analysis was carried for data of 20 patients. All versions of device configuration were considered — with 2, 4 and 6 electrodes. All bioimpedance measurements were repeated three times.

A problem of classification to three classes: osteoporosis, osteopenia and healthy was considered. Models based on linear and quadratic discriminant analysis, or LDA and QDA, were fitted using data for only one frequency and up to four classification variables. Two configurations were considered:

1. maximum accuracy in three classes, where a model is considered to be "best", if the probability of sickness level being correctly classified is highest,

2. maximum sensitivity, where a model is considered to be "best", if it reaches maximum specificity, while finding all cases of osteoporosis or osteopenia — without distinguishing between the two.

Analysis began from verifying if there are patterns in response with growth of frequency. They were analysed in chapters 2, 3 and 4, for 2, 4 and 6 electrodes respectively. No such patters were observed for 2 electrode method. Analysis for 6 electrode method showed, that there are issues with gathered data, i.e. that the measurements are illegible in many cases. In received reports there is no data if this was for variant A or B.

A descriptive analysis of data, present in chapters 11–14, showed a strong correlation between T-Score, Z-Score and BDM between various points in body. Also, this analysis confirmed big differences between repeated examinations for 6 electrodes variant, especially between first and later measurements. In received reports there is no data if this was for variant A or B.

After simple validation of data, they were analysed using quadratic discriminant analysis or QDA method. They were considered in chapters 5, 6 and 7 of report, for 2, 4 and 6 electrodes respectively. Aim of those chapters was to confirm if three measurements for single patient are close to each other. In all cases frequencies were considered separately, fitting a model with 4 variables. Conducted analysis confirmed, that predictions between repeated examinations for same patient are highly variable.

The most important part of this analysis is present in chapter 8–10, where quality of obtained models were considered. When classifying into three classes following results were obtained:

2 electrodes accuracy of 65%, for QDA method on Z' and Z'' variables for frequencies 509, 1017 and 2035Hz,

4 electrodes accuracy of 75%, for QDA method on Z' and Z'' variables for frequencies 254, 15513, 31281 and 62561Hz,

6 electrodes, variant A accuracy 70%, for QDA method on Z' and Z'' variables for frequencies 31281, 62561 and 249990Hz,

6 electrodes, variant B accuracy 85%, for LDA method on Z' variable for frequency 62561Hz, or accuracy 80%, for QDA method on Z' and Z'' variables for frequency 249990Hz.

When algorithm is tuned for 100%, sensitivity reaches

2 electrodes 20% specificity for frequency 254 and 509Hz,

4 electrodes 20% specificity for all frequencies,

6 electrodes, variant A 20% specificity for frequency 7884Hz,

6 electrodes, variant B 80% specificity for frequency 62561Hz.

The very large specificity when considering 100% sensitivity for case of 6 electrode device in variant B is alarming, considering that for this type of device large variances of results have been observed. Nonetheless, for this case 100% of osteoporosis cases, 75% cases of osteopenia and 80% of patients without illness were correctly classified.

The author suggested to increase the number of measurements per patient. Author also suggested trying out other frequencies. He also notices, that number of patients was too small to make a decision for such high variance in readings.

3 Second analysis

A second report was created based on new results, obtained with improved device, designed to decrease big variability observed in previous analysis. This time there were six repeated examinations, three for single day for 27 patients. Also, set of DEXA variables was reduced — it included T-Score and Z-Score in two points on arm, labelled U and 1/3.

Only analysis for 4 and 5 electrodes was made. It was noted, that T-Score in point U can indicate illness, while in point 1/3 it can classify patient as healthy. Reverse is true as well. At the same a strong correlation between Z-Score and T-Score for same point was found. This is why analysis was conducted using T-Score in both points, leaving Z-Score out. Later, it was assumed, that one can use average of those two T-Scores. Finally, single variable resulting from DEXA measurements was used to classify true state of patient.

The analysis for 4 and 5 electrodes showed, that for improved device there are no big differences between measurements for same day, although there still is noticeable difference for measurements between days. No cause for such behaviour been found in data, including analysis of age, weight and height of patient, although extra variables (age, weight and height) were shown to be statistically important for the problem of classification.

An analysis of explained variance followed, then a set of single and multidimensional models has been tested. It has been noticed, that patients with osteopenia are hard to classify, because for some variables they behave like osteoporosis, and for some like healthy patients. This is why classification to only two groups was further analysed. The groups are "healthy" and "not healthy", where all cases of osteoporosis (T-Score below -2.5) should be classified as "not healthy", and all cases of T-Score above -1 should be classified as healthy. It means, that there is a value of T-Score between -2.5 and -1 that should be considered as boundary value between those two cases, or in other words, severe cases of osteopenia are classified as "not healthy" and mild cases of osteopenia as "healthy". Three values of limiting T-Score were considered: -2, -1.5 and -1.

A final model using Random Forests was proposed, using all 47 predictive variables (44 measurements, weight, height and age). For considered T-Score values, following models were obtained (based on averaged measurements per day, i.e. 54 total measurements, two for each of 27 patients). The parameter were estimated using "out-of-bag" method, which is similar to cross-validation, where part of data is removed during model fitting state, to later test model performance on those previously removed data. Results are summarized in table below. Note, that "not healthy" and "healthy" columns represent population state, not test results, so they depend only on selected limiting value of T-Score.

T-Score	not healthy	healthy	sensitivity	specificity	accuracy
-2	14	40	0.71	0.85	0.81
-1.5	24	30	0.79	0.80	0.79
-1	30	24	0.73	0.58	0.66

It is worth to note, that while model with limiting T-Score of -2 gives the best accuracy. By looking at ROC curve included in last report appendix (dated 2nd Oct. 2013) it seems, that it isn't calculated correctly — the point selected in table is far from the curve. This is why the last part of analysis is repeated in this summary report, to validate this result.

4 Model validation

To validate the results, a data set of 166 measurements was provided, with data for 28 patients. By quick comparison, in original analysis, patient with surname encoded as 1022 and measurement with number 10925 were removed from data. It leaves data for 162 measurements and 27 patients, 6 measurements for each patient. Each patient was examined three times a day, at two different dates.

We will analyse four cases. First, we will recreate original model, which seems was built using following conditions:

1. three measurements from given day were averaged,
2. measurements at different dates were considered to be from different patients,
3. model included all measured variables, weight, height and age.

Such model exactly matches the contingency table of results present in report. We will then analyse similar model, but without weight, height and age. To further validate applied model, no averaging of measurements will be assumed and measurements from same patients will be distinguished, even if at different dates. We will consider two such models — with and without extra variables (weight, height, age).

For each models a ROC curve will be plotted, and placed on same image for comparison. To stabilize the ROC curves they were calculated using following methodology: for each patient all his measurements were removed from data and new random forest was found. Then, using this data fitted without removed patient data, his state was predicted using generated random forest. A fraction of trees voting for "healthy" was noted. This procedure was repeated for each patient, generating a vector of probabilities, that each patient is healthy. Ten such vectors were generated and averaged, to decrease variability and improve stability of results. This procedure generates a vector of expected probability returned by random forest calculated for this patient. Then, for each cut-off point between zero and one, which we call parameter, we classify patients, whose probability of being healthy is higher than selected value, as "healthy". This allows us to calculate sensitivity (fraction of "non healthy" patients correctly classified) and specificity (fraction of "healthy" patients correctly classified) for each parameters. Plotted ROC curve are all those values for different values of parameters. Such technique can be called averaged "out-of-bag" method. That way instead of fit measurement, we get estimation of predictive power of method. This nullified the potential overfitting of random forests models.

In the figure 1, which was made using described technique, the black line shows ROC curve for model present in earlier analyses. It is easy to see big differences between this curve, and one present in report from 2nd Oct. 2013. Blue line shows model without weight, age and height. Green is version without averaging of measurements but with weight, age and height, and red is made without averaging and without extra variables.

The AUC or "Area Under Curve" statistics for those models is 0.90, 0.87, 0.85 and 0.83 respectively. It is obvious, that model with averaging and with extra variables gives

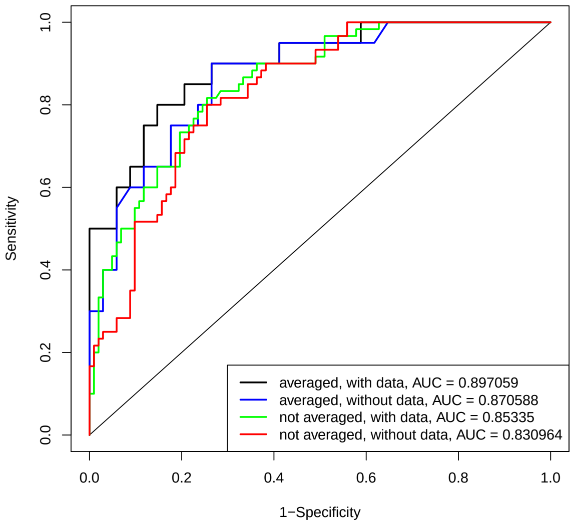


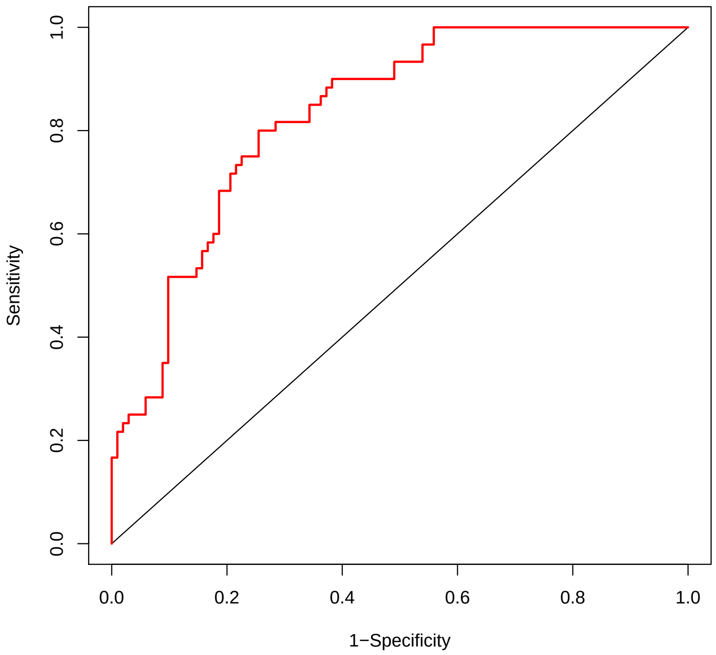
Figure 1: ROC curves for different model assumptions

best results (black line, AUC 0.90), but for single measurement without weight, height and age, we have to follow red curve (AUC 0.83). Analysis of AUC also confirmed, that limiting values of T-Score between -2 and -1.9 results in the best predictive power.

Notice, that any of those models can be tuned differently — any value of sensitivity can be reached, but at the cost of specificity. Similarly, any value of specificity can be reached, at the cost of sensitivity.

Below we present some values of sensitivity and specificity for all four cases. We present parameter, that is a fraction of trees that should vote for “healthy”, to conclude, that T-Score is greater than -2. We also list sensitivity, specificity and accuracy. There are three rows in each table: first corresponds to highest sensitivity for specificity of 0.90 or more, second corresponds to highest balanced accuracy, and the last one corresponds to highest specificity for sensitivity of 0.90 or more.

First table corresponds to red curve, which best describes the future predictive power of device without any improvements — as long as it cannot utilise data like weight, height and age:

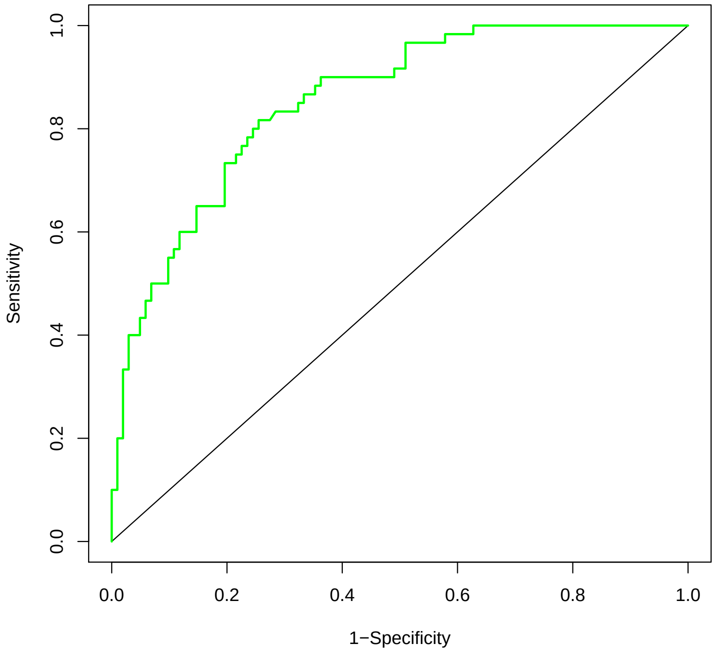


parameter	sensitivity	specificity	accuracy
0.373	0.52	0.90	0.76
0.653	0.80	0.75	0.77
0.738	0.90	0.62	0.72

For received data this results in following contingency tables:

parameter	0.373		0.653		0.738	
test result	positive	negative	positive	negative	positive	negative
not healthy	31	29	48	12	54	6
healthy	10	92	26	76	39	63

If device gains ability to input age, weight and height, those values will change to:

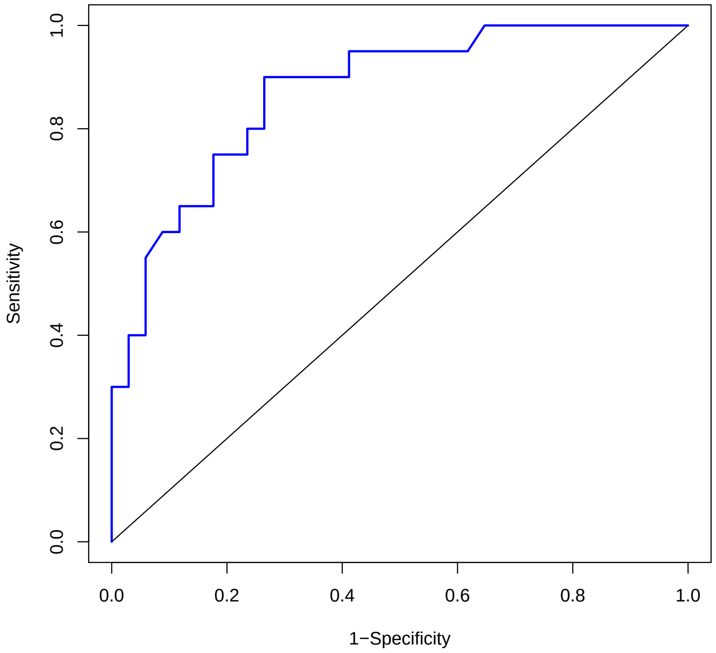


parameter	sensitivity	specificity	accuracy
0.399	0.55	0.90	0.77
0.637	0.82	0.75	0.77
0.750	0.90	0.64	0.73

For received data this results in following contingency tables:

parameter	0.399		0.637		0.750	
test result	positive	negative	positive	negative	positive	negative
not healthy	33	27	49	11	54	6
healthy	10	92	26	76	37	65

If instead measurement accuracy improves, we can assume, that device will behave like for case with averaged measurements:

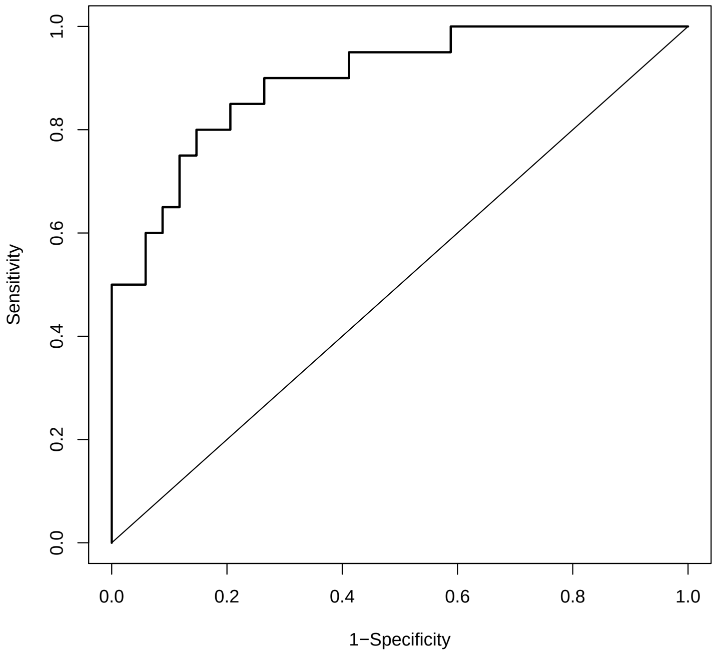


parameter	sensitivity	specificity	accuracy
0.412	0.60	0.91	0.80
0.658	0.90	0.74	0.80
0.643	0.90	0.74	0.80

For received data this results in following contingency tables:

parameter	0.412		0.658		0.643	
test result	positive	negative	positive	negative	positive	negative
not healthy	12	8	18	2	18	2
healthy	3	31	9	25	9	25

If additionally it would be possible to use age, weight and height of patient, we will get:



parameter	sensitivity	specificity	accuracy
0.472	0.65	0.90	0.81
0.575	0.80	0.85	0.83
0.613	0.90	0.74	0.80

For received data this results in following contingency tables:

parameter	0.472		0.575		0.613	
test result	positive	negative	positive	negative	positive	negative
not healthy	13	7	16	4	18	2
healthy	3	31	5	29	9	25

I suggest selecting the most conservative approach, using model fitted without weight, height and age, and without averaging measurements. For this model, selecting parameter 0.653 resulted in the highest expected accuracy based on received data set. If examination method with less variance between measurements will be created or there will be possibility to input weight, height and age of patient during examination, parameters from other tables will give more accurate results.

5 Summary and selected model

Based on analysis, a following model was considered: all 44 measurement variables, height, weight and age were considered. For those variables a model of random forest has been fitted, which reaches accuracy around 77%, i.e. correctly classifies more than 77% of cases of T-Score of above or below -2. Patients classified as having T-Score of below -2 should be examined using other methods to distinguish between osteoporosis and osteopenia.

The proposed method can be tuned to reach different properties, for example to detect 90% of cases with T-Score below -2, but it happens at cost of more healthy patients classified as ones with illness and overall decrease in accuracy. Nonetheless, depending on needs, such results can be obtained.

Additional analysis was carried in this summary report, to validate the earlier model. It has been shown, that we should expect accuracy of around 77%, unless device or measuring method will be improved. If device included an interface to input weight, height and age, accuracy will go up by few percent, but real improvement could be obtained by multiple measurements or decrease in variability, resulting in over 80% of accuracy, or nearly 90% of sensitivity while keeping 74% of specificity. Similar conclusions were present in earlier reports.

Overall the method seems to give good results. While the set of parameters might be not final, i.e. the algorithm might need further tuning, it seems that results cannot be an effect of pure chance and the device already successfully determines state of most patients.