Comparing the ABL800 FLEX and the Roche/Hitachi cobas c system for estimating the GFR value by the CKD-EPI equation

Christoph Jäggli, Alexander Leichtle

November 28, 2019

Abstract

TODO: *Do abstract*

1 Introduction

TODO: *Write how to measure the GFR from the creatinine: the creatinine level is combined with the age, ethnicity, gender, height, and weight of the patient by the formula in Levey et al. [2009]*
TODO: *say that there are two different machines, that they take different time to execute etc. without providing too much details*
TODO: *from the ABL page: Creatinine is an endogenous waste product of muscle metabolism, derived from creatine, a molecule of major importance for energy production within muscle cells. Creatinine is removed from the body in urine and its concentration in blood reflects glomerular filtration and thereby kidney function.* TODO: *explain clearly and outwrite GFR: (Wikipedia) Glomerular filtration rate (GFR) describes the flow rate of filtered fluid through the kidney.*

2 Method

As mentioned in the introduction, the center for laboratory medicine in Bern hosts two different machines that measure the creatinine level of blood samples: the *ABL800 FLEX* blood gas analyzer and the *Roche/Hitachi cobas c* system. Henceforth we will simply call them ABL and COBAS, respectively. The ABL machine is an optical system that is trained for fast analysis of up to 18 parameters in blood samples. One measuring cycle takes 145-225 seconds and it is able to consider 16-24 samples per hour TODO: *https://www.radiometer.com/en/products/blood-gas-testing/abl800-flex-blood-gas-analyzer This means that it has only 1 sample position, is that true?*. COBAS on the other hand, is a photometric high performance analyzer with a throughput of up to 300 tests per hour. It is therefore designed to meet the requirements of clinical chemistry and immunochemistry assays in high-workload laboratories. However, the higher machine precision comes at the price of taking up to 22 minutes per measuring cycle TODO: *https://diagnostics.roche.com/global/en/products/instruments/cobas-c-311.html They write it has 108 sample positions, is that true?*.

The operation manuals of both machines report performance tests for assessing the accuracy of the creatinine measures. For a reference sample with a creatinine level of 87 μmol/L the inter-assay ($n = 45$) coefficient of variation for the ABL machine is $c_v = 3.6\%$. The COBAS method performed significantly better. A series of $n = 21$ repetitions with a reference sample of 190 μmol/L resulted in $c_v = 1.1\%$.

Statistical Analysis

The *correlation* between two variables X and Y (also called *Pearson's Correlation Coefficient (PCC)* [Pearson, 1895]) is defined as

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\operatorname{cov}(X,Y)$ is the covariance between X and Y and σ_X, σ_Y denote the standard deviation of X and Y, respectively. It measures the linear correlation between X and Y by a value ranging from -1 (total negative linear correlation) over 0 (no linear correlation) up to $+1$ (total positive linear correlation). Furthermore, Student [1908] presented a formula to compute a confidence interval as well as probability value (p-value) that uncorrelated and normally distributed samples give a correlation values at least as extreme as $\rho_{X,Y}$.

Linear regression is a widely used technique to measure and visualize the linear correlation between two samples. A number of different methods aim to find the most appropriate intercept β_0 and slope β_1 and fit a straight line to a set of points (x_i, y_i), where $i = 1, \ldots, n$. *Deming Regression* in particular, is used when both variables, X and Y, are measured with error [Deming, 1964]. As it incorporates measurement errors for both variables, it is often used for method comparison studies to look for systematic differences between two measurement techniques. Deming regression assumes that the ratio of the variance, denoted by λ is constant. In contrast to the PCC, here we consider the variance to be a point-wise quantification of the measurement error. The value of λ can either be required from the user or approximated by providing multiple observations from each subject. When $\lambda = 1$, Deming regression is equivalent to orthogonal regression. It is crucial to assume the ratio to be constant, so we use the described machine accuracies (c.f. coefficients of variation described in the previous section) and approximate λ by $(3.6\%/1.1\%)^2 = 3.2727^2$.

3 Results and Discussion

3.1 Data

The serum creatinine level of 766 patients (18-97 years, 332 female and 434 male) was measured by both of the two machines described above. After determining the creatinine level, all patients were consecutively referred for estimating the GFR by the CKD-EPI formula of Levey et al. [2009]. Table 1 shows basic statistical frequencies of the estimated GFR value.

Table 1: Basic statistics for the estimated GFR values by ABL and COBAS

	ABL	**COBAS**
Min	5	5
Max	170	165
Mean	88.81	88.41
Std	27.95	27.83
CV	0.3146	0.3149

By using the correlation measure described in 2, the GFR values described above produced correlation coefficient of $\rho_{X,Y} = 0.9906\pm$ with 95%-confidence interval between 0.9891 and 0.9918 and p-value $< 10^{-10}$.

3.2 Relation Between ABL and COBAS

For the following analysis we assumed the error ratio to be equal to 1 in which case Deming regression gives the same result as orthogonal regression. The left figure of 1 shows the regression line (dashed red) and its uncertainty region (magenta) between the GFR values of ABL and COBAS. With a confidence interval of 95%, the regression line is defined by the intercept $\beta_0 = -0.7727 \pm$

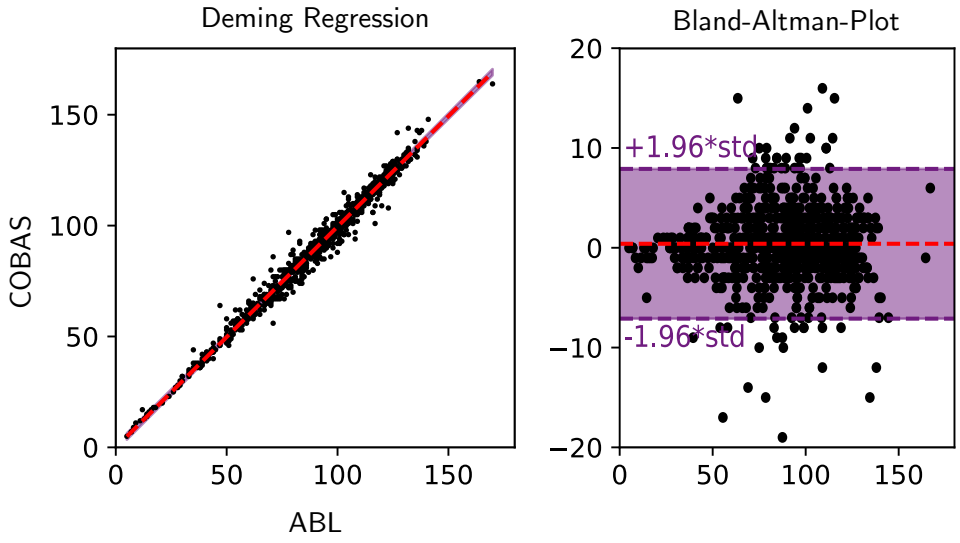

Figure 1: Deming regression line and the corresponding Bland-Altman-Plot of the two data sets.

0.7613 and the slope $\beta_1 = 1.0041 \pm 0.0084$ and produces a relatively large coefficient of determination reading $R^2 = 0.9809$. The Bland-Altman-Plot on the right hand side of figure 1 shows that the differences in the estimated GFL values between ABL and COBAS are well centered around zero and do not show a significant value dependence. In other words, no clear trend can be observed.

3.3 Predict GFL Categories

For determining the degree of kidney failure, the estimated GFR values are categorized into six groups as shown in table 2 [Group, 2013]. As mentioned above, the last section suggests to avoid

Table 2: GFR-categorization of the chronically kidney diseases.

Category	**Description**	**Range** ($\mathrm{ml}/\min/1.73\mathrm{m}^2$)
G1	Normal or high	≥ 90
G2	Mildly decreased	60-89
G3a	Mildly to moderately decreased	45-59
G3b	Moderately to severely decreased	30-44
G4	Severely decreased	15-29
G5	Kidney failure	< 15

the COBAS measure procedure and directly use the GFR values issued from the ABL method in order to predict the kidney failure category of the patient. Figure 2 shows the confusion matrix when comparing the two techniques. The x-axis and the y-axis represent the GFR categories from the ABL and the COBAS method, respectively. In total, there are 719 ($p = 0.9386$) subjects that

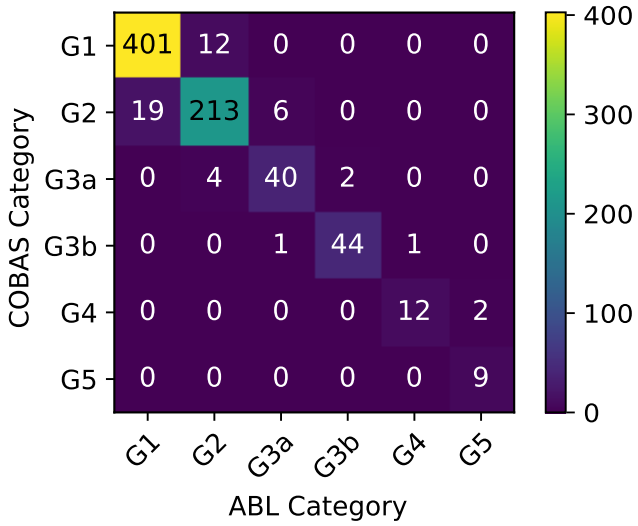

Figure 2: Confusion matrix when comparing the GFR categories of the ABL and the COBAS method.

are classified similarly what leaves 47 ($p = 0.0614$) conflicts. Among the defective classifications, all fall into the immediate neighbor class. This means that the difference never exceeds one class. Considering ABL as the classifier and COBAS as the truth, we conclude that there are 23 objects that are assessed too severe and 24 objects that are assessed too mildly. Furthermore, both prior distributions highly favor the first two categories ($n = 649$ (ABL), 651 (COBAS)) over the remaining ones ($n = 117$ (ABL), 115 (COBAS)). This becomes even more clear in table 3, where assignment frequencies for both distributions are presented.

Table 3: Assignment frequencies for the ABL and the COBAS categories.

Category	**ABL**		**COBAS**	
G1	420	($p = 0.5483$)	413	($p = 0.5392$)
G2	229	($p = 0.299$)	238	($p = 0.3107$)
G3a	47	($p = 0.0614$)	46	($p = 0.0601$)
G3b	46	($p = 0.0601$)	46	($p = 0.0601$)
G4	13	($p = 0.017$)	14	($p = 0.0183$)
G5	11	($p = 0.0144$)	9	($p = 0.0117$)

3.4 Inter-Rater Reliability

The *inter-rater reliability* expresses the degree of agreement between two (or more) ratings of categorical items. An obvious technique to do so is to simply compute the percent of agreement between the raters. However, this is often thought to be not very robust, as it does not consider the baseline distribution of the categorical fractions. In other words, a highly skewed prior distribution over the categories would enhance the degree of agreement by pure chance. More robust measures are the Cohen's κ [Cohen, 1960], the Fleiss' κ [Fleiss, 1971], and Krippendorff's α [Hayes and Krippendorff, 2007]. All measures values are bounded from above by 1 and it is commonly assumed that a value ≥ 0.8 expresses a very high degree of reliability. The agreement measure values between

the ABL and the COBAS GFR categories are shown in table 4. We observe a relatively high baseline

Table 4: Inter-rater reliability of the ABL and the COBAS GFR categories.

Agreement	0.9386
Baseline	0.3963
Cohen's κ	0.8984
Fleiss' κ	0.8984
Krippendorff's α	0.9732

agreement. This is due to the skewed prior distribution mentioned above. All the other statistical measures describe a very strong degree of agreement between the two categorization strategies.

3.5 COBAS Measurement Uncertainty

The measurement accuracy of the COBAS machine can be illustrated by using the error rate of the measurement device. This means that we may take a synthetic set of creatinine level values, perturb them according to the error distribution of the machine and compare the resulting GFR categories against the original ones. In order to make the confusion matrix comparable to the one in figure 2 we take the synthetic set to be the creatinine levels observed by the COBAS machine. In other words, we start from the 766 creatinine observations by the COBAS machine, add a Gaussian error with zero mean and 1.1% variation and then compute the perturbed GFR estimation. Therefore, the creatinine levels are computed as

$$\hat{c} = c(1 + 0.011\epsilon),$$

where c and $\hat{c}$ are the true and the perturbed value and ϵ follows a standard normal distribution $\mathcal{N}(0, 1)$. After using estimating the GFR value, we compared the obtained kidney disease categories to the original ones. This results in the confusion matrix in figure 3 and their corresponding probability measures in table 5

Table 5: Assignment frequencies for the true and the predicted categories for the perturbed COBAS measurements.

Category	**True**		**Predicted**	
G1	427	($p = 0.5474$)	418	($p = 0.5359$)
G2	238	($p = 0.3051$)	243	($p = 0.3115$)
G3a	46	($p = 0.059$)	50	($p = 0.0641$)
G3b	46	($p = 0.059$)	43	($p = 0.0551$)
G4	14	($p = 0.0179$)	17	($p = 0.0218$)
G5	9	($p = 0.0115$)	9	($p = 0.0115$)

The x-axis shows the perturbed (predicted) categories and the y-axis the synthetic (true) categories. In total, there are 745 ($p = 0.9726$) subjects that are classified correctly what leaves 21 ($p = 0.0274$) misassignments. Comparing the two confusion matrices in figures 2 and 3 we clearly observe that the misassignments that are due to measurement uncertainties are significantly less frequent than the ones from the prediction method. This finding however, is not surprising as the COBAS measure is supposed to deliver more precise creatinine measurements (see also remark in

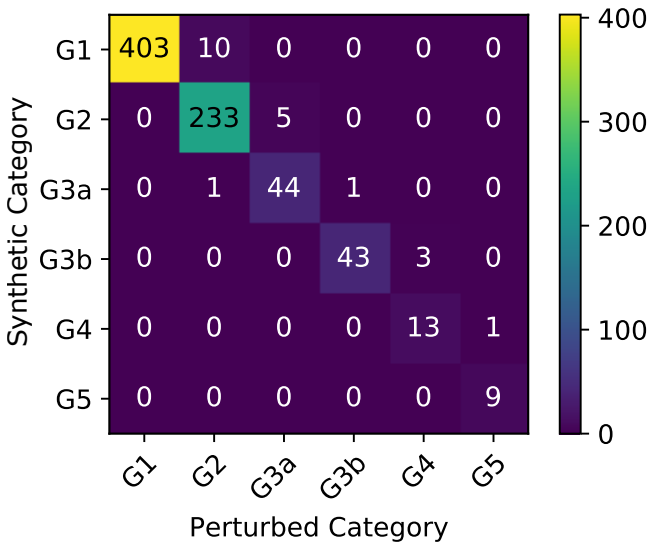

Figure 3: Confusion matrix of the estimated GFR values that are computed from the perturbed synthetic creatinine levels.

4 Conclusion

Based on the dataset of 766 patients with large age spectrum and of both genders, we observed a significant linear correlation and a very strong degree of agreement between the estimated GFR values issued from the ABL800 FLEX and the Roche/Hitachi cobas c system. It is clear that the faster measuring cycle of the ABL device compromises its precision. Consequently, any technique that aims to omit an exacter but slower examination needs to accept some compromises. However, for the underlying dataset we were able to show that the probability of the issues to become problematic is very low. Let us consider this statement in more detail.

Deming regression produced a negative intercept ($\beta_0 = -0.7727 \pm 0.7613$) what suggests that the ABL slightly undervalues the estimated GFR levels with respect to the COBAS method. This can potentially be problematic when the GFR value is used to determine the degree of kidney failure by the categories in table 2. However, considering the confusion matrix in figure 2, we observe that the underestimations mainly fell into categories G1 and G2. Among medical specialists, it is not uncommon to assume that the critical categories are G3a-G5. Considering all under-categorized subjects that COBAS rated within G3a-G5, we only find 5 problematic cases out of which 4 fall into the COBAS category G3a and 1 into G3b. This means that based on the underlying dataset, the potentially severe misclassifications (when using ABL over COBAS) are very rare (i.e. 0.65%).

References

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL `https://doi.org/10.1177/001316446002000104`.

W.E. Deming. *Statistical adjustment of data*. Dover Books on Mathematics Series. Dover Publications, 1964. URL `https://books.google.ch/books?id=KH8pAQAAMAAJ`.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Improving Global Outcomes (KDIGO) CKD Work Group. Kdigo 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int Suppl*, 3:1–150, 2013.

Andrew F. Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007. doi: 10.1080/19312450709336664. URL https://doi.org/10.1080/19312450709336664.

Andrew S. Levey, Lesley A. Stevens, Christopher H. Schmid, Yaping (Lucy) Zhang, III Castro, Alejandro F., Harold I. Feldman, John W. Kusek, Paul Eggers, Frederick Van Lente, Tom Greene, Josef Coresh, and for the CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A New Equation to Estimate Glomerular Filtration Rate. *Annals of Internal Medicine*, 150(9):604–612, 05 2009. ISSN 0003-4819. doi: 10.7326/0003-4819-150-9-200905050-00006. URL https://doi.org/10.7326/0003-4819-150-9-200905050-00006.

Karl Pearson. *Proceedings of the Royal Society of London*. Taylor & Francis, 1895. URL https://books.google.ch/books?id=60aL0zlT-90C.

Student. Probable error of a correlation coefficient. *Biometrika*, pages 302–310, 1908.