tween the cancer patients and controls. The frequency of the *5A* allele in the Italian patients with breast cancer (60.5%) was higher than the frequency of this allele in our cases (Table 1), but this may be attributable to small numbers of cases with mammary tumors (n = 43) in the Italian study. Clearly, although MMP3 was suggested to play a role in tumor initiation in studies on transgenic mice *(5)* and mouse mammary tumor cell lines *(6)*, more data will be needed to support the previous conclusion *(8)*.

References

1. McCawley LJ, Matrisian LM. Matrix metalloproteinases: multifunctional contributors to tumor progression. Mol Med Today 2000;6:149–56.
2. Sternlicht MD, Bissell MJ, Werb Z. The matrix metalloproteinase stromelysin-1 acts as a natural mammary tumor promoter. Oncogene 2000; 19:1102–13.
3. Ioachim EE, Athanassiadou SE, Kamina S, Carassavoglou K, Agnantis NJ. Matrix metalloproteinase expression in human breast cancer: an immunohistochemical study including correlation with cathepsin D, type IV collagen, laminin, fibronectin, EGFR, c-*erbB*-2 oncoprotein, p53, steroid receptors status and proliferative indices. Anticancer Res 1998;18:1665–70.
4. Johansson N, Ahonen M, Kahari VM. Matrix metalloproteinases in tumor invasion. Cell Mol Life Sci 2000;57:5–15.
5. Lochter A, Galosy S, Muschler J, Freedman N, Werb Z, Bissell MJ. Matrix metalloproteinase stromelysin-1 triggers a cascade of molecular alterations that leads to stable epithelial-to-mesenchymal conversion and a premalignant phenotype in mammary epithelial cells. J Cell Biol 1997;139:1861–72.
6. Sternlicht MD, Lochter A, Sympson CJ, Huey B, Rougier JP, Gray JW, et al. The stromal proteinase MMP3/stromelysin-1 promotes mammary carcinogenesis. Cell 1999;98:137–46.
7. Ye S, Eriksson P, Hamsten A, Kurkinen M, Humphries SE, Henney AM. Progression of coronary atherosclerosis is associated with a common genetic variant of the human stromelysin-1 promoter which results in reduced gene expression. J Biol Chem 1996;271:13055–60.
8. Biondi ML, Turri O, Leviti S, Seminati R, Cecchini F, Bernini M, et al. MMP1 and MMP3 polymorphisms in promoter regions and cancer. Clin Chem 2000;46:2023–4.
9. Lei HX, Sjoberg-Margolin S, Salahshor S, Werelius B, Jandakova E, Hemminki K, et al. CDH1 mutations are present in both ductal and lobular breast cancer, but promoter allelic variants show no detectable breast cancer risk. Int J Cancer 2002; in press.

Haixin Lei[1]
Jan Zaloudik[2]
Igor Vorechovsky[1*]

[1] *Karolinska Institute*
Department of Biosciences at Novum
S-14157 Huddinge, Sweden

[2] *Masaryk Memorial Cancer Institute*
Department of Surgery
600 00 Brno, Czech Republic

*Author for correspondence. Fax 46-8-6089269; e-mail igvo@cbt.ki.se.

Application of the Bland–Altman Plot for Interpretation of Method-Comparison Studies: A Critical Investigation of Its Practice

To the Editor:

Current guidelines for the combined graphical/statistical interpretation of method-comparison studies *(1)* include a scatter plot combined with correlation and regression analysis *(2)* and/or a difference plot combined with calculation of the 2s limits of the differences between the methods (the so-called 95% limits of agreement) *(3, 4)*. The former approach has a long tradition in clinical chemistry, and its advantages and pitfalls are well known *(5)*. The latter approach, however, which was deemed "simple both to do and to interpret" and was propagated as a substitute for regression analysis *(4, 5)*, became available only in recent years and has increased in popularity. The general features of the Bland–Altman plot have been well described *(4)* (see also Fig. 1A). The *x* axis shows the mean of the results of the two methods ([A + B]/2), whereas the *y* axis represents the absolute difference between the two methods ([B − A]). When the standard deviation increases with concentration, Bland and Altman recommend a logarithmic *y* scale, whereas others propose a percent *y* scale *(6)*. Although generally there is not much difference in effect between using percentages and using a log transformation of the data, we prefer the percent plot (except when data extend over several orders of magnitude) because numbers can be read directly from the plot without the need for back-transformation. Additionally, the plot includes the line for the mean difference and the experimentally observed 2s limits of the differences between the methods. Often forgotten, the Bland–Altman approach consists of a comparison of the 2s limits with a clinically acceptable difference between the two methods.

We reviewed difference plots published in this journal and discuss here the key aspects associated with their use. We screened all articles in this journal, starting from the first issue of 1995 up to May 2001. We observed increasing use of the Bland–Altman plot over the years, from 8% in 1995 to 14% in 1996, and 31–36% in more recent years. In addition to the Bland–Altman method, method comparisons were performed using correlation and regression analysis and the concordance plot. In total, we found 96 uses of difference plots [listed in the Data Supplement that accompanies the online version of this letter at *Clinical Chemistry Online* (clinchem.org/content/vol48/issue5)]. Most authors also used correlation and regression analysis, suggesting that difference plots are viewed as complementary to, rather than substitutes for, regression analysis. Among 96 references (in total, 98 plots) with Bland–Altman plots, 75 used the absolute difference plot, 20 applied a percent *y*-scale version, and 3 a logarithmic version of the plot. In total, 50 presented the results in an additional scatter plot.

The following general problems were observed. In 13 cases, the *x* axis was constructed using only the values of the comparison method (see Data Supplement, Addendum 2, for listing). By doing so, however, the plot may falsely show a concentration-dependent difference even when there is none *(7)*. The 2s limits were presented in only 67 cases, and most importantly, only 2 authors compared the 2s limits with a clinically acceptable difference between the two methods. The 2s limits were more generally used in absolute (59) and logarithmic (3) difference plots, but rarely in percent (5) difference plots.

A similar search was performed in

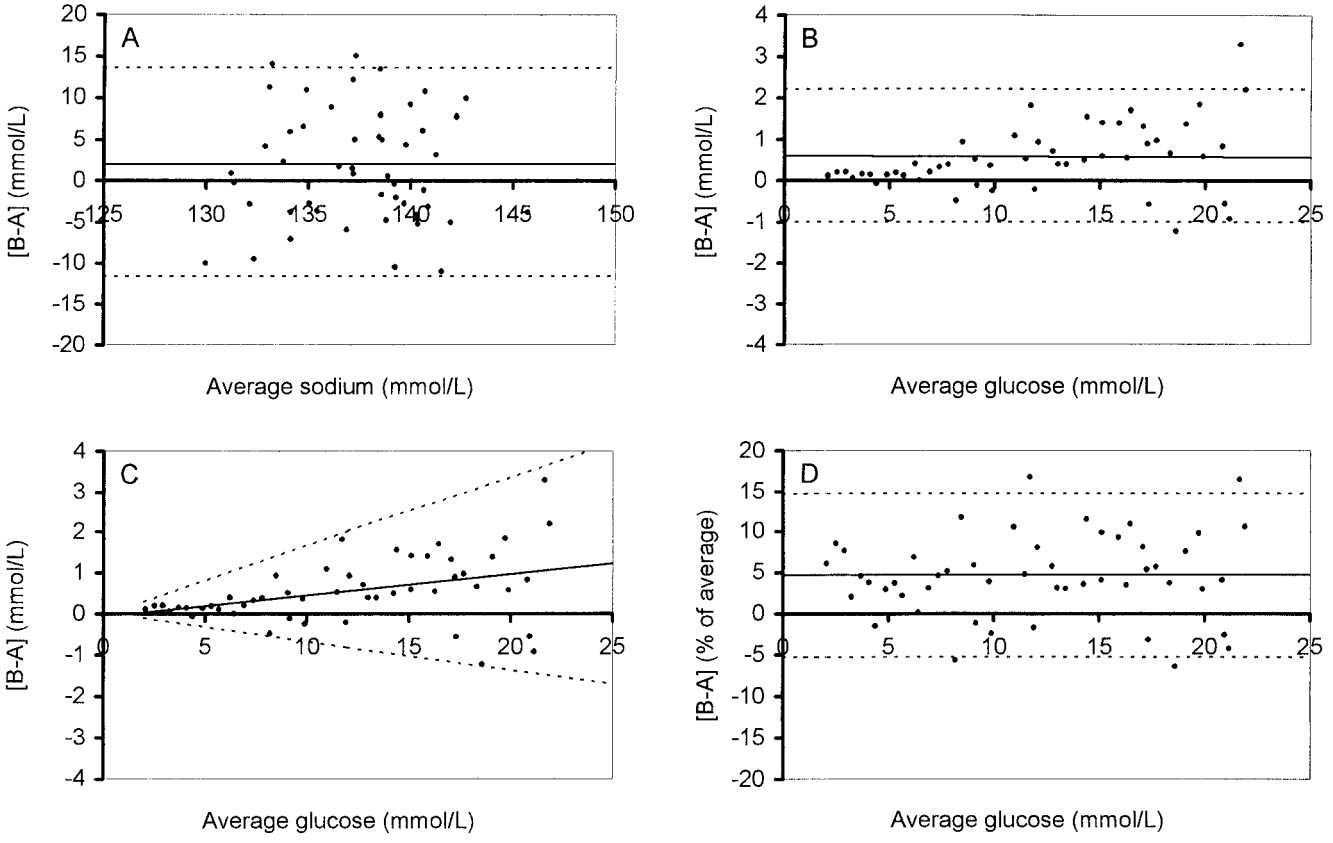

Fig. 1. Overview of difference plots with mean differences (*solid lines*) and 2s limits (*dashed lines*).
Shown are a classical absolute difference plot (*A*) and absolute difference (*B* and *C*) and percent difference (*D*) plots of two data sets with a proportional difference.

two other laboratory medicine journals for the period 1996–2001. We found in *Clinical Chemistry and Laboratory Medicine* and *Annals of Clinical Biochemistry*, respectively, 29 and 43 difference plots (17 and 34 absolute, 10 and 7 percent, and 2 and 2 logarithmic difference plots). We found that the characteristics of the plots in *Clinical Chemistry and Laboratory Medicine* were similar to those reported for this journal (see Data Supplement, Addenda 3a, 3b, and 3c). However, in *Annals of Clinical Biochemistry*, additional scatter plots were very seldom presented. This apparently results from the fact that the "Instructions for Authors" deprecate the use of regression analysis, which traditionally is accompanied by a scatter plot.

Bland and Altman *(4)* show method comparisons that cover a small concentration range and data sets without proportional differences between the methods. In this situation, a constant standard deviation may be assumed, and parallel 2s limits and a mean bias are justified (Fig. 1A). However, this case is rather unusual in clinical chemistry. In the 75 examined references with absolute difference plots (showing 103 figures), we found, by eye, 57 data sets with a standard deviation increasing with concentration and/or with a proportional difference (see Fig. 1B). In these cases, Bland and Altman recommend the use of a log transformation of the data points. Neither a mean bias (in Fig. 1B suggested by the horizontal line at 0.6 mmol/L) nor constant and parallel 2s limits are justified. Rather, the 2s limits should be "V-shaped" around the regression line of the differences *(8, 9)* (see Fig. 1C). Alternatively, to use parallel 2s limits, a percent difference plot can be used (Fig. 1D). Overall, we found that 87% of plots had technical flaws, similar to data reported by Mantha et al. *(10)*, who made an analogous survey in the field of anesthesia. Most striking, in both surveys, interpretation of the data by comparison of the actually observed limits of agreement with a priori ones was missing in >90% of the cases.

In summary, difference plots are useful for the presentation and interpretation of method-comparison studies, but most authors in this journal use them as supplements to regression analysis and the scatter plot, a practice that is also recommended by the NCCLS *(1)*. Unfortunately, many authors uncritically apply the classical absolute difference plot in method-comparison studies that cover a wider concentration range, where they would better use a percent (or log) difference plot. Last but not least, the main objective of the Bland–Altman approach, namely, comparison of the experimentally observed deviations with a preset clinical acceptance limit, is seldom followed despite recommendations for doing so that were given earlier in this journal *(11)*.

To emphasize, the key aspects of

the appropriate construction and use of the Bland–Altman plot are the following. The *x* axis should be constructed by the mean of the methods and the *y* axis in a way that is most sensible to the concentration range of the *x* data (absolute: small range; percentage: medium range; log-scale: large range). The 95% limits of agreement should reflect the actually observed nature of the differences (whether or not there is a relationship between difference and magnitude) *(9)*. Most important, interpretation of the data should be done by comparison of the observed limits of agreement with a priori ones.

As a final note, we want to remark that in this journal, already in 1981, a similar plot (with the *y* axis constructed as a ratio) was proposed for the evaluation of method-comparison data *(12)*. Strange to say, this report has been overlooked.

This work was supported by the Research Fund of the University Ghent (Grant BOF 011109000).

References

1. National Committee for Clinical Laboratory Standards. Method comparison and bias estimation using patient samples, approved guideline. NCCLS publication EP9-A. Villanova, PA: NCCLS, 1995.
2. Westgard JO, Hunt MR. Use and interpretation of common statistical tests in method comparison studies. Clin Chem 1973;19:49–57.
3. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician 1983;32:307–17.
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
5. Hollis S. Analysis of method comparison studies [Guest Editorial]. Ann Clin Biochem 1996; 33:1–4.
6. Pollock MA, Jefferson SG, Kane JW, Lomax K, MacKinnon G, Winnard CB. Method comparison—a different approach. Ann Clin Biochem 1992;29:556–60.
7. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995; 346:1085–7.
8. Thienpont LM, Van Nuwenborg JE, Stöckl D. Intrinsic and routine quality of serum total potassium measurement as investigated by split-sample measurement with ion chromatography candidate reference method. Clin Chem 1998;44:849–57.
9. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135–60.
10. Mantha S, Roizen MF, Fleisher LA, Thisted R, Foss J. Comparing methods of clinical measurement: reporting standards for Bland and Altman analysis. Anesth Analg 2000;90:593–602.
11. Petersen PH, Stöckl D, Blaabjerg O, Pedersen B, Birkemose E, Thienpont L, et al. Graphical interpretation of analytical data from comparison of a field method with a reference method by use of difference plots. Clin Chem 1997;43: 2039–46.
12. Eksborg S. Evaluation of method-comparison data [Letter]. Clin Chem 1981;27:1311–2.

Katy Dewitte
Colette Fierens
Dietmar Stöckl
Linda M. Thienpont*

Laboratorium voor Analytische Chemie
Universiteit Gent
Harelbekestraat 72
9000 Gent, Belgium

*Author for correspondence. Fax 32-9-264-81-98; e-mail Linda.thienpont@rug.ac.be.

In the following commentary, Drs. Altman and Bland elaborate on the issues raised in the above letter:

Commentary on Quantifying Agreement between Two Methods of Measurement

Our interest in the analysis of method comparison studies stemmed from discussions about consulting problems we were independently working on in the late 1970s. Examination of published papers showed that, at that time, most authors were using the Pearson correlation coefficient. It was obvious to us that this method did not assess agreement, but association, and that a high correlation was no guarantee of good agreement.

We felt that a comparison of two methods of measurement, such as different assays, should attempt to quantify the differences and that *P* values were largely irrelevant. The question is not whether the two methods agree, but how closely they agree. Our statistical approach was based on investigation of the distribution of the between-method differences. We suggested summarizing the data by the mean and 95% range of the differences, which we called the 95% limits of agreement. The graph, which many think is the whole of our method, was intended as a visual check that the approach was reasonable and that the data were "well-behaved". Thus the graph shows whether the variability of differences between methods is roughly constant across the range of measurement, but the key element of the approach is to examine and summarize the individual differences between the two methods. Indeed, in our original paper we included histograms of these differences. This distribution should be approximately normal, and (apart from occasional outliers) this is usually what we see.

Our first two papers outlined the basic ideas *(1, 2)*, but a recent report contained the fullest exposition of our method, including various extensions to deal with replicated observations and complex relationships between the between-method difference and the magnitude of the measurement *(3)*.

Our original work related to clinical rather than laboratory measurements, but it was soon obvious that, broadly speaking, the same issues arose. Concerns about the use of the Pearson correlation coefficient had been expressed in this Journal as long ago as 1973 *(4)*, but the method remained widespread for decades (and it still is in the wider medical literature).

The idea of plotting difference vs mean was not new *(5)*, but as far as we knew its use had not been proposed in this context. The same type of plot was suggested as a general purpose approach for method comparison studies at around the same time *(6)*, although without any suggestion for quantifying the differences between the methods.

A particular issue that we were aware of from the start is that there are some measurements where the between-method (and within-method) variability increases as the measurement increases. We found that the SD of the differences tended to be proportional to the size of the measurement, so that log transformation of the original data led to