Comparing Imperfect Measurements with the Bland-Altman Technique: Application in Gene Expression Analysis

Lucila Ohno-Machado, MD, PhD[1,2], Staal Vinterbo, PhD[1], Stephan Dreiseitl, PhD[1,3], Tor-Kristian Jenssen, MSc [4], and Winston Kuo, DMD, MS[1,2,5]

[1]Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School
[2]Division of Health Sciences and Technology, Massachusetts Institute of Technology
[3]Dept. of Software Engineering for Medicine, Polytechnic University of Upper Austria
[4]Dept. of Computer Science, Norwegian University of Science and Technology
[5]Harvard School of Dental Medicine

ABSTRACT

Several problems in medicine and biology involve the comparison of two measurements made on the same set of cases. The problem differs from a calibration problem because no gold standard can be identified. Testing the null hypothesis of no relationship using measures of association is not optimal since the measurements are made on the same cases, and therefore correlation coefficients will tend to be significant. The descriptive Bland-Altman method can be used in exploratory analysis of this problem, allowing the visualization of gross systematic differences between the two sets of measurements. We utilize the method on three sets of matched observations and demonstrate its usefulness in detecting systematic variations between two measurement technologies to assess gene expression.

INTRODUCTION

In both clinical and research medical settings, the problem of comparing two different measurements of the same entity is not uncommon. Experiments may be replicated and results compared using the same instrument. Instruments or technologies for the same type of measurement may exist, and there may not be a gold-standard. Frequently used techniques to perform this comparison are measures of association such as kappa or correlation coefficients, and regression. For example, Steinke et al. [1] compared measurements of carotid artery stenosis obtained with color Doppler flow imaging and angiography using linear regression; Morrin et al [2] compared ultrasonography and tomography for detecting unresectable periampullary cancer using kappa coefficients; and Kuo et al. [3] compared mRNA levels obtained from different microarray platforms using linear and rank correlation coefficients. Simple regression can be used as it also provides a measure of correlation. However, a slope of 1 does not necessarily correspond to good levels of agreement, as data may not fit the regression assumptions. Furthermore, there may be systematic differences that are difficult to capture by the inspection of the resulting model. Usual tests of association are based on the null hypothesis of no association, which is less useful to know whether results from different instruments can be interchangeable (e.g. correlation coefficients may be significant but relatively low). Correlation may depend on the range of data in the sample. Note that the problem of comparing measurements made by two imperfect instruments is different from that of calibration, in which a gold-standard measurement can be obtained.

Bland and Altman have proposed a simple technique to assess the agreement between two sets of observations derived from the same cases [4,5,6]. The method has been successfully used in certain clinical [7] and laboratory assessments [8], but has not yet been utilized in the comparison of gene expression levels originating from different platforms. In these types of experiments, there may not exist a gold standard, and the data are inherently noisy. A simple visual method to verify whether there seems to be systematic differences between the measurements can be helpful.

MATERIALS AND METHODS

We used matched data from cDNA and oligonucleotide microarrays obtained from three cell lines of the standard panel of 60 from the National Cancer Institute, to verify whether there seemed to be systematic differences in the mRNA measurements. These data were matched by a procedure described in detail in [3]. In short, the GenBank accession numbers provided by each laboratory for their microarray platform were used to obtain the corresponding sequence data. For each cDNA probe sequence, BLAST was used to find the best matching probe set on the oligonucleotide arrays. From all possible pairs of one cDNA probe and one sequence represented by a probe set in the oligonucleotide microarray, matches with a score less significant than e-50 were removed. We selected three different cell lines whose correlation coefficients fell in three distinct ranges. This small sample was used to illustrate the use of the Bland-Altman technique, and not to make definitive conclusions on the quality of each microarray technology. A larger data set with several replicates would be necessary for that purpose.

Methods

The Bland-Altman technique consists of analysis and visualization of the differences between two sets of observations. Using the data set described above, the differences between each pair of standardized observations were computed, as well as the overall mean and standard deviation of the differences. The estimate of the "true value" of the measured gene is the mean of the two measurements (i.e., neither type of measurement is assumed to be superior). We plotted the differences of each observation against the estimated true value in search for systematic variation. We calculated the descriptive statistics and measures of association in SAS [9].

We produced an artificial set of observations by adding random normally distributed noise to the measurements of the oligonucleotide technique for comparison. The Bland-Altman plot for this set represents the expected plot for replications of the experiment using the same platform and serves as a visual guide towards what could be expected from observations that had very good agreement. We also used one set of replicates from each platform to illustrate what would be expected if the measurements were reasonably repeatable.

Results

Figure 1 shows the plots of the paired measurements. It is easy to see that the correlation is not high (although it is significantly different from zero), but it is hard to determine whether different levels of expression are associated with higher differences in the two measurements, given the high number of cases with low expression levels.

Table 1 shows the mean and standard deviation of the differences between standardized measurements, as well as the correlation coefficients between the two sets. Several related measures of association are presented to illustrate the fact that the correlations are different from zero, but differ substantially from each other and do not provide much guidance into the repeatability issue.

Table 1. Mean differences and measures of association.

	Line 1	Line 2	Line 3	Noisy	Repl. CHD2	Repl. Oligo
Mean Dif	0	0	0	0	0	0
Stdev	1.22	1.17	1.08	0.10	0.41	0.53
Pearson	0.26	0.32	0.42	0.99	0.92	0.86
Spearm.	0.41	0.33	0.48	0.84	0.96	0.65
Tau B	0.28	0.23	0.34	0.67	0.82	0.49
Hoeffd.	0.06	0.04	0.09	0.35	0.59	0.17

(all *p*-values for measures of association were < .0001)

Figure 2 shows the Bland-Altman plots with bars corresponding to two standard deviations from the mean. From the picture it is possible to inspect in which range the differences higher than 2 standard deviations from the mean lie. The correlation coefficient between the absolute difference and the means indicates that there are systematic biases (ranges for which the differences are more pronounced).

Figure 3 shows the plots of the paired measurements from the artificially created noisy data, replicates for Cell Line 1 using cDNA and oligonucleotide microarrays. Figure 4 shows the corresponding Bland-Altman plots with bars two standard deviations above and below the difference means.

In order to utilize the standard deviations to construct the confidence intervals, it is necessary to verify whether the assumption of normality holds for the distribution of differences [10]. Figure 5 shows how the differences fit a normal distribution for paired measurements on Cell Line 1 and the cDNA replicates on the same cell line.

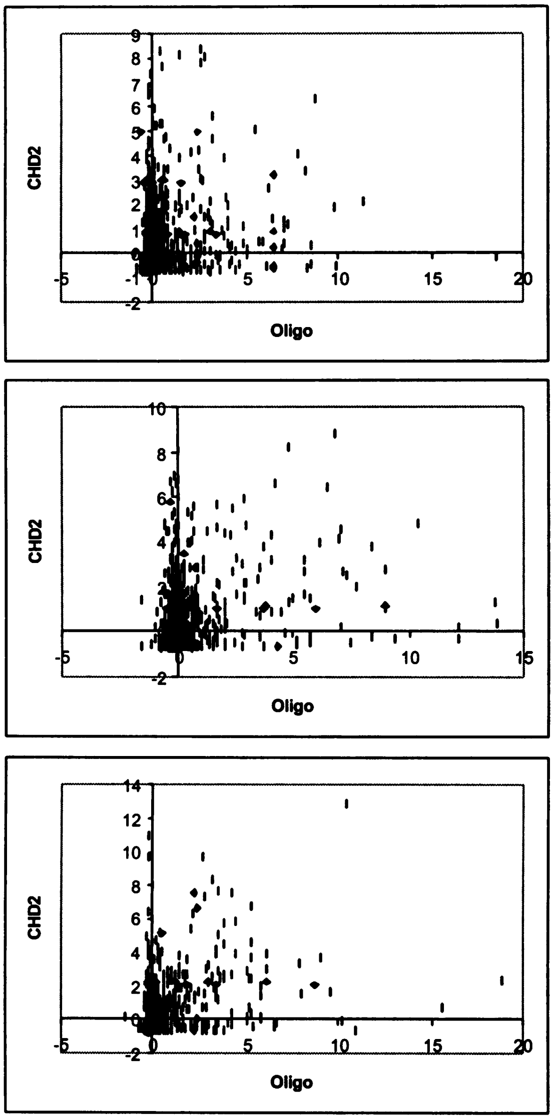

Figure 1. From top to bottom: Plot of standardized measurements from the two different microarrays for cell lines 1, 2, and 3.

DISCUSSION

This simple analysis shows that the Bland-Altman technique is useful in determining the form of the systematic biases in the two measurements. From the plots, there seem to be important systematic biases, even between replicates using the same type of microarray. The problems appear especially important when the mean measurement is high.

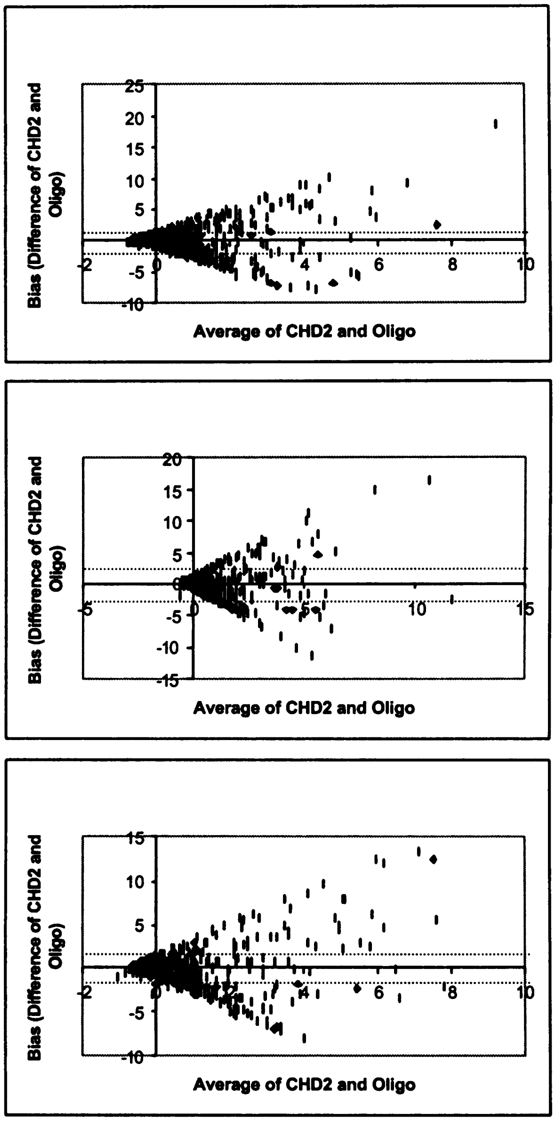

Figure 2. From top to bottom: Plot of differences between the two paired measurements and the estimated "true" value (i.e., the average measurement) for cell lines 1, 2, and 3. Horizontal lines represent the mean difference plus or minus 2 standard deviations. Pearson correlation coefficients were 0.84, 0.80. and 0.78, respectively. The p values for the correlations were all < 0.0001.

The analysis was not exhaustive, and this report shows a few examples of the problems of repeatability between two specific types of microarrays. We should exercise caution when interpreting these results, since despite the fact that the cell lines were originally the same, the measurements were performed in different laboratories.

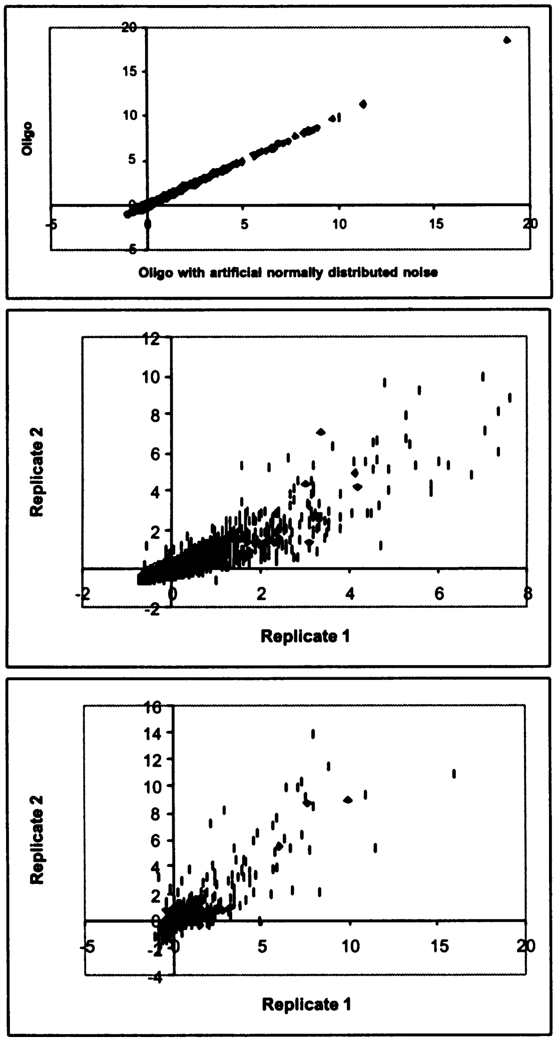

Figure 3. Plot of standardized measurements for cell line 1. The top graph plots measurements from oligonucleotide arrays and noisy measurements derived artificially. Random normally distributed noise was added to the oligo measurement to illustrate the appearance of a Bland-Altman plot without systematic variation. The other two graphs represent replicates as measured by cDNA and oligonucleotide microarrays, respectively.

We investigated biases in three different cell lines, and the format of the Bland-Altman plot was very similar for the comparisons, except for the artificial set against its original set, for which we verified our expectation of a higher percentage of measurements differences falling within two standard deviations from the mean. The same analysis was performed on log transformed data, and systematic differences were also verified. In the experiments described here, which used differences in standardized measurements, we were making an implicit assumption that the scales of both techniques were of the same nature, which may be incorrect.

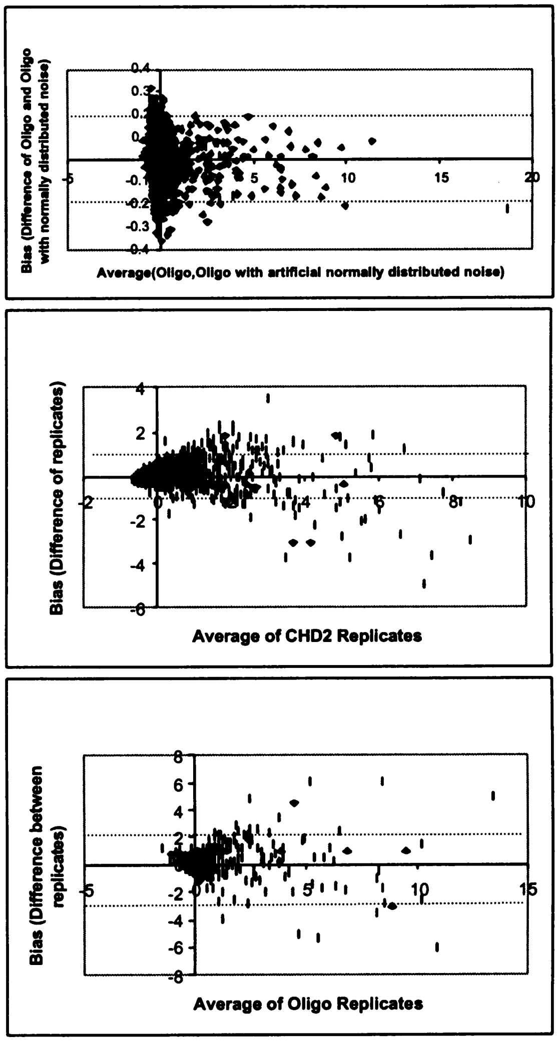

Figure 4. Plot of differences between the two paired measurements and the estimated "true" value (i.e., the average measurement). Pearson correlation coefficients were 0.43, 0.68, and -0.37. P values for the correlation of replicates were below 0.01. The p value for the noisy set was 0.036.

The fact that the Spearman (rank) correlation was higher than the Pearson (linear) correlation for the pairs of measurements that exclude the noisy set gives us a hint about a possible violation of this assumption. Our preliminary results indicate that at least one of the scales may be highly non-linear and that differences in probe features may be associated with different types of measurement errors. This fact does not invalidate the results presented here, as there has been no large systematic experiment comparing both measurements to a verifiable "gold standard". In this report, we verify that the agreement between the two measurements is not ideal, illustrate the differences according to mean measurements, and try to determine whether the differences are systematic.

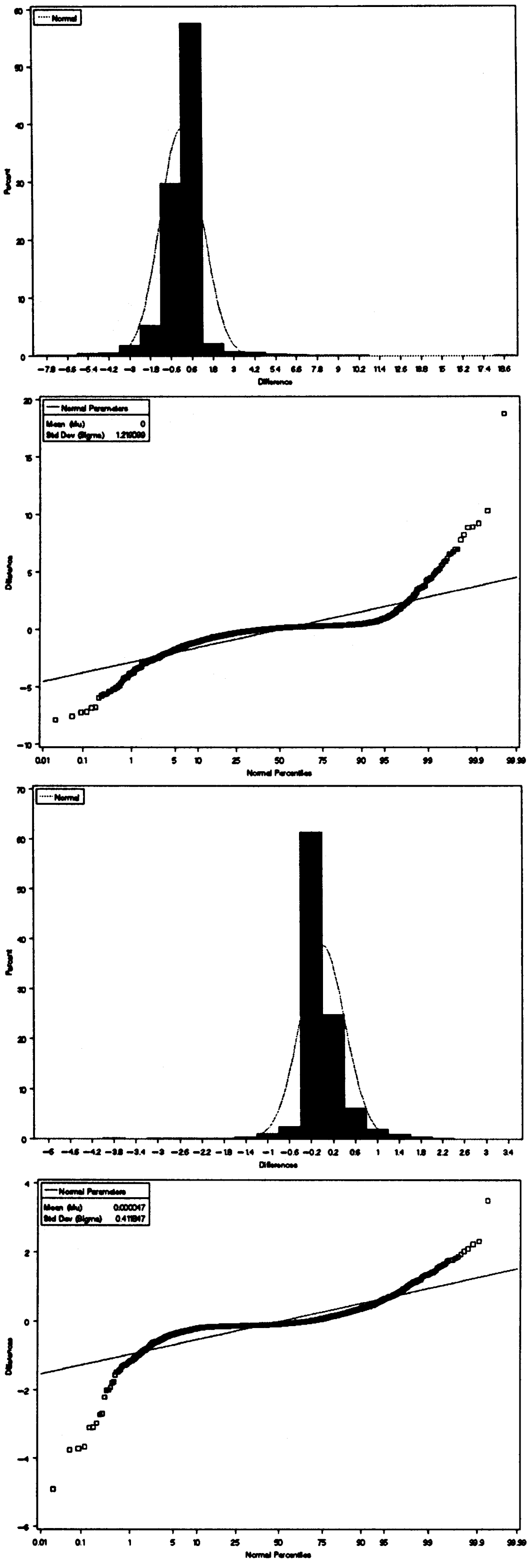

Figure 5. Histogram and normal plots showing fit to the normal distribution for the different technologies (top two graphs, Cell Line 1) and cDNA replicates (top bottom).

This report was intended to demonstrate the usefulness of the Bland-Altman method in gene expression analysis. No conclusions should be derived from the relative quality of the microarrays presented here. Further experiments are necessary to investigate this important question. The main advantages of using the Bland-Altman method derive from (a) its simplicity, (b) its focus on the differences between the measurements in the absence of a gold standard, and (c) the ease with which results can be visualized and interpreted.

Acknowledgments

This work was funded in part by the HST Taplin fellowship award. The authors thank the insightful comments of two anonymous reviewers.

References

[1] Morrin MM, Kruskal JB, Raptopoulos V, Weisinger K, Farrell RJ, Steer ML, Kane RA. State-of-the-art ultrasonography is as accurate as helical computed tomography and computed tomographic angiography for detecting unresectable periampullary cancer. J Ultrasound Med 2001 May;20(5):481-90.

[2] Steinke W, Ries S, Artemis N, Schwartz A, Hennerici M. Power Doppler imaging of carotid artery stenosis. Comparison with color Doppler flow imaging and angiography. Stroke 1997 Oct;28(10):1981-7.

[3] Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS. Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 2002; 18(3):405-12.

[4] Altman DG and Bland JM, Measurement in medicine: the analysis of method comparison studies. The Statistician 1983, 32, 307-317.

[5] Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med. 1990;20(5):337-40.

[6] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986 Feb 8;1(8476):307-10.

[7] Edmonds ZV, Mower WR, Lovato LM, Lomeli R. The reliability of vital sign measurements. Ann Emerg Med 2002 Mar;39(3):233-7.

[8] King TW, Brey EM, Youssef AA, Johnston C, Patrick CW Jr. Quantification of vascular density using a semiautomated technique for immunostained specimens. Anal Quant Cytol Histol 2002 Feb;24(1):39-48.

[9] SAS Institute. SAS/STAT Version 8.1.

[10] Bland M. An introduction to medical statistics. Oxford, 2000.