Statistical Methods in Medical Research 1999; **8**: 135–160

Measuring agreement in method comparison studies

J Martin Bland Department of Public Health Sciences, St George's Hospital Medical School, London, UK and **Douglas G Altman** ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford, UK

Agreement between two methods of clinical measurement can be quantified using the differences between observations made using the two methods on the same subjects. The 95% limits of agreement, estimated by mean difference ± 1.96 standard deviation of the differences, provide an interval within which 95% of differences between measurements by the two methods are expected to lie. We describe how graphical methods can be used to investigate the assumptions of the method and we also give confidence intervals. We extend the basic approach to data where there is a relationship between difference and magnitude, both with a simple logarithmic transformation approach and a new, more general, regression approach. We discuss the importance of the repeatability of each method separately and compare an estimate of this to the limits of agreement. We extend the limits of agreement approach to data with repeated measurements, proposing new estimates for equal numbers of replicates by each method on each subject, for unequal numbers of replicates, and for replicated data collected in pairs, where the underlying value of the quantity being measured is changing. Finally, we describe a nonparametric approach to comparing methods.

1 The method comparison problem

In clinical medicine we often wish to measure quantities in the living body, such as cardiac stroke volume or blood pressure. These can be extremely difficult or impossible to measure directly without adverse effects on the subject and so their true values remain unknown. Instead we have indirect methods of measurement, and when a new method is proposed we can assess its value by comparison only with other established techniques rather than with the true quantity being measured. We cannot be certain that either method gives us an unequivocally correct measurement and we try to assess the degree of agreement between them. The standard method is sometimes known as the 'gold standard', but this does not – or should not – imply that it is measured without error.

Some lack of agreement between different methods of measurement is inevitable. What matters is the amount by which methods disagree. We want to know by how much the new method is likely to differ from the old, so that if this is not enough to cause problems in clinical interpretation we can replace the old method by the new, or even use the two interchangeably. For example, if a new machine for measuring blood pressure were unlikely to give readings for a subject which differed by more than, say, 10 mmHg from those obtained using a sphygmomanometer, we could rely on measurements made by the new machine, as differences smaller than this would not materially

Address for correspondence: JM Bland, Department of Public Health Sciences, St George's Hospital Medical School, Cranmer Terrace, London SW17 0RE, UK.

 0962-2802(99)SM181RA

affect decisions as to management. On the other hand, differences of 30 mmHg or more would not be satisfactory as an error of this magnitude could easily lead to a change in patient management. How far apart measurements can be without leading to problems will depend on the use to which the result is put, and is a question of clinical judgement. Statistical methods cannot answer such a question. Methods which agree well enough for one purpose may not agree well enough for another. Ideally, we should define satisfactory agreement in advance.

In this paper we describe an approach to analysing such data, using simple graphical techniques and elementary statistical calculations.[1,2] Our approach is based on quantifying the variation in between-method differences for individual patients. We provide a method which is simple for medical researchers to use, requiring only basic statistical software. It gives estimates which are easy to interpret and in the same units as the original observations. We concentrate on the interpretation of the individual measurement on the individual patient.

We extend the approach to the case where the between-method differences vary with the size of the measurement, and we show how to compare methods when replicated measurements are available. We also describe a nonparametric approach for use when there may be occasional extreme deviations between the methods.

2 Limits of agreement

We want a measure of agreement which is easy to estimate and to interpret for a measurement on an individual patient. An obvious starting point is the difference between measurements by the two methods on the same subject. There may be a consistent tendency for one method to exceed the other. We shall call this the *bias* and estimate it by the mean difference. There will also be variation about this mean, which we can estimate by the standard deviation of the differences. These estimates are meaningful only if we can assume bias and variability are uniform throughout the range of measurement, assumptions which can be checked graphically (Section 2.1).

Table 1 shows a set of systolic blood pressure data from a study in which simultaneous measurements were made by each of two experienced observers (denoted J and R) using a sphygmomanometer and by a semi-automatic blood pressure monitor (denoted S). Three sets of readings were made in quick succession. We shall start by considering only the first measurement by observer J and the machine (i.e. J1 and S1) to illustrate the analysis of unreplicated data (Figure 1).

The mean difference (observer minus machine) is $\bar{d} = -16.29$ mmHg and the standard deviation of the differences is $s_d = 19.61$ mmHg. If the differences are normally distributed, we would expect 95% of the differences to lie between $\bar{d} - 1.96s_d$ and $\bar{d} + 1.96s_d$ (we can use the approximation $\bar{d} \pm 2s_d$ with minimal loss of accuracy). We can then say that nearly all pairs of measurements by the two methods will be closer together than these extreme values, which we call 95% *limits of agreement*. These values define the range within which most differences between measurements by the two methods will lie. For the blood pressure data these values are

Table 1 Systolic blood pressure measurements made simultaneously by two observers (J and R) and an automatic blood pressure measuring machine (S), each making three observations in quick succession (data supplied by Dr E O'Brien, see Altman and Bland[20])

Subject	J1	J2	J3	R1	R2	R3	S1	S2	S3
1	100	106	107	98	98	111	122	128	124
2	108	110	108	108	112	110	121	127	128
3	76	84	82	76	88	82	95	94	98
4	108	104	104	110	100	106	127	127	135
5	124	112	112	128	112	114	140	131	124
6	122	140	124	124	140	126	139	142	136
7	116	108	102	118	110	102	122	112	112
8	114	110	112	112	108	112	130	129	135
9	100	108	112	100	106	112	119	122	122
10	108	92	100	108	98	100	126	113	111
11	100	106	104	102	108	106	107	113	111
12	108	112	122	108	116	120	123	125	125
13	112	112	110	114	112	110	131	129	122
14	104	108	104	104	108	104	123	126	114
15	106	108	102	104	106	102	127	119	126
16	122	122	114	118	122	114	142	133	137
17	100	102	102	102	102	100	104	116	115
18	118	118	120	116	118	118	117	113	112
19	140	134	138	138	136	134	139	127	113
20	150	148	144	148	146	144	143	155	133
21	166	154	154	164	154	148	181	170	166
22	148	156	134	136	154	132	149	156	140
23	174	172	166	170	170	164	173	170	154
24	174	166	150	174	166	154	160	155	170
25	140	144	144	140	144	144	158	152	154
26	128	134	130	128	134	130	139	144	141
27	146	138	140	146	138	138	153	150	154
28	146	152	148	146	152	148	138	144	131
29	220	218	220	220	218	220	228	228	226
30	208	200	192	204	200	190	190	183	184
31	94	84	86	94	84	88	103	99	106
32	114	124	116	112	126	118	131	131	124
33	126	120	122	124	120	120	131	123	124
34	124	124	132	126	126	120	126	129	125
35	110	120	128	110	122	126	121	114	125
36	90	90	94	88	88	94	97	94	96
37	106	106	110	106	108	110	116	121	127
38	218	202	208	218	200	206	215	201	207
39	130	128	130	128	126	128	141	133	146
40	136	136	130	136	138	128	153	143	138
41	100	96	88	100	96	86	113	107	102
42	100	98	88	100	98	88	109	105	97
43	124	116	122	126	116	122	145	102	137
44	164	168	154	164	168	154	192	178	171
45	100	102	100	100	104	102	112	116	116
46	136	126	122	136	124	122	152	144	147
47	114	108	122	114	108	122	141	141	137
48	148	120	132	146	130	132	206	188	166
49	160	150	148	160	152	146	151	147	136
50	84	92	98	86	92	98	112	125	124
51	156	162	152	156	158	152	162	165	189
52	110	98	98	108	100	98	117	118	109
53	100	106	106	100	108	108	119	131	124
54	100	102	94	100	102	96	136	116	113
55	86	74	76	88	76	76	112	115	104

Table 1 Continued

Subject	J1	J2	J3	R1	R2	R3	S1	S2	S3
56	106	100	110	106	100	108	120	118	132
57	108	110	106	106	118	106	117	118	115
58	168	188	178	170	188	182	194	191	196
59	166	150	154	164	150	154	167	160	161
60	146	142	132	144	142	130	173	161	154
61	204	198	188	206	198	188	228	218	189
62	96	94	86	96	94	84	77	89	101
63	134	126	124	132	126	124	154	156	141
64	138	144	140	140	142	138	154	155	148
65	134	136	142	136	134	140	145	154	166
66	156	160	154	156	162	156	200	180	179
67	124	138	138	122	140	136	188	147	139
68	114	110	114	112	114	114	149	217	192
69	112	116	122	112	114	124	136	132	133
70	112	116	134	114	114	136	128	125	142
71	202	220	228	200	220	226	204	222	224
72	132	136	134	134	136	132	184	187	192
73	158	162	152	158	164	150	163	160	152
74	88	76	88	90	76	86	93	88	88
75	170	174	176	172	174	178	178	181	181
76	182	176	180	184	174	178	202	199	195
77	112	114	124	112	112	126	162	166	148
78	120	118	120	118	116	120	227	227	219
79	110	108	106	110	108	106	133	127	126
80	112	112	106	112	110	106	202	190	213
81	154	134	130	156	136	132	158	121	134
82	116	112	94	118	114	96	124	149	137
83	108	110	114	106	110	114	114	118	126
84	106	98	100	104	100	100	137	135	134
85	122	112	112	122	114	114	121	123	128

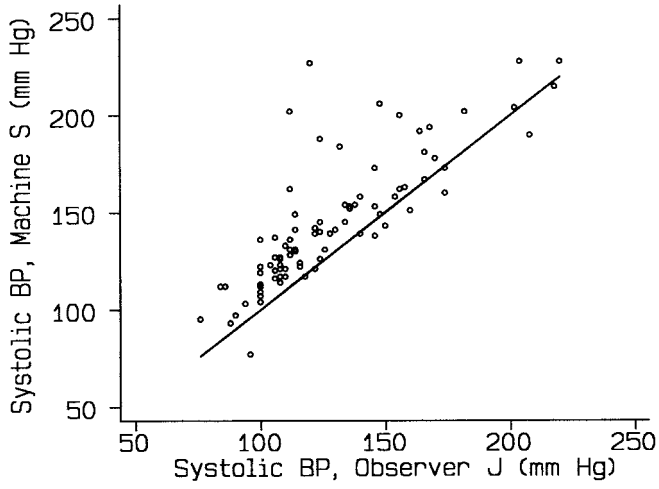

Figure 1 Systolic blood pressure measured by observer J using a sphygmomanometer and by machine S, with the line of equality

$-16.29 - 1.96 \times 19.61$ and $-16.29 + 1.96 \times 19.61$ mmHg or -54.7 and $+22.1$ mmHg. Provided differences within the observed limits of agreement would not be clinically important we could use the two measurement methods interchangeably.

Note that despite the superficial similarity these are not the same thing as confidence limits, but like a reference interval. Of course, we could use some percentage other than 95% for the limits of agreement, but we find it convenient to stick with this customary choice.

As we indicated, the calculation of the 95% limits of agreement is based on the assumption that the differences are normally distributed. Such differences are, in fact, quite likely to follow a normal distribution. We have removed a lot of the variation between subjects and are left with the measurement error, which is likely to be normal anyway. We then added two such errors together which will increase the tendency towards normality. We can check the distribution of the differences by drawing a histogram or a normal plot. If the distribution is skewed or has very long tails the assumption of normality may not be valid. This is perhaps most likely to happen when the difference and mean are related, in which case corrective action can be taken as described in Section 3. Further, a non-normal distribution of differences may not be as serious here as in other statistical contexts. Non-normal distributions are still likely to have about 5% of observations within about two standard deviations of the mean, although most of the values outside the limits may be differences in the same direction. We note that is not necessary for the measurements themselves to follow a normal distribution. Indeed, they often will not do so as subjects are chosen to give a wide and uniform distribution of the quantity measured rather than being a random sample.

If there is a consistent bias it is a simple matter to adjust for it, should it be necessary, by subtracting the mean difference from the measurements by the new method. In general, a large s_d and hence widely spaced limits of agreement is a much more serious problem.

In the blood pressure example we can see that there are some very large differences where the machine gave readings considerably above the sphygmomanometer. There are in fact still about 5% of values outside the limits of agreement (4/85) but they all lie below the lower limit. We can evaluate the impact of the two largest, apparently outlying values (from subjects 78 and 80) by recalculating the limits excluding them. The mean difference becomes -14.9 mmHg and the 95% limits of agreement are -43.6 to $+15.0$ mmHg. The span has reduced from 77 to 59 mmHg, a noticeable but not particularly large reduction. We do not recommend excluding outliers from analyses, but it may be useful to assess their influence on the results in this way. We usually find that this method of analysis is not too sensitive to one or two large outlying differences.

From Table 1 we can see that these large discrepancies were not due to single odd readings as the difference was present for all three readings by each method. In the case of automatic blood pressure measuring machines this phenomenon is quite common. For this reason a nonparametric approach was developed to handle such data – we describe this method in Section 6.

As we remarked earlier (Section 1) the decision about what is acceptable agreement is a clinical one; statistics alone cannot answer the question. From the above

calculations we can see that the blood pressure machine may give values between 55 mmHg above the sphygmomanometer reading to 22 mmHg below it. Such differences would be unacceptable for clinical purposes.

2.1 Graphical presentation of agreement

Graphical techniques are especially useful in method comparison studies. Figure 1 shows the measurement by the observer using a sphygmomanometer plotted against that by the machine. The graph also shows the line of equality, the line all points would lie on if the two meters always gave exactly the same reading. We do not calculate or plot a regression line here as we are not concerned with the estimated prediction of one method by another but with the theoretical relationship of equality and deviations from it. It is helpful if the horizontal and vertical scales are the same so that the line of equality will make an angle of 45° to both axes. This makes it easier to assess visually how well the methods agree. However, when the range of variation of the measurements is large in comparison with the differences between the methods this plot may obscure useful information.

A better way of displaying the data is to plot the difference between the measurements by the two methods for each subject against their mean. This plot for the blood pressure data (Figure 2) shows explicitly the lack of agreement that is less obvious in Figure 1. The plot of difference against mean also allows us to investigate any possible relationship between the discrepancies and the true value. We can examine the possible relation formally by calculating the rank correlation between the absolute differences and the average; here Spearman's rank correlation coefficient is $r_s = 0.07$. The plot will also show clearly any extreme or outlying observations. It is often helpful to use the same scale for both axes when plotting differences against mean values (as in Figure 2). This feature helps to show the discrepancies in relation

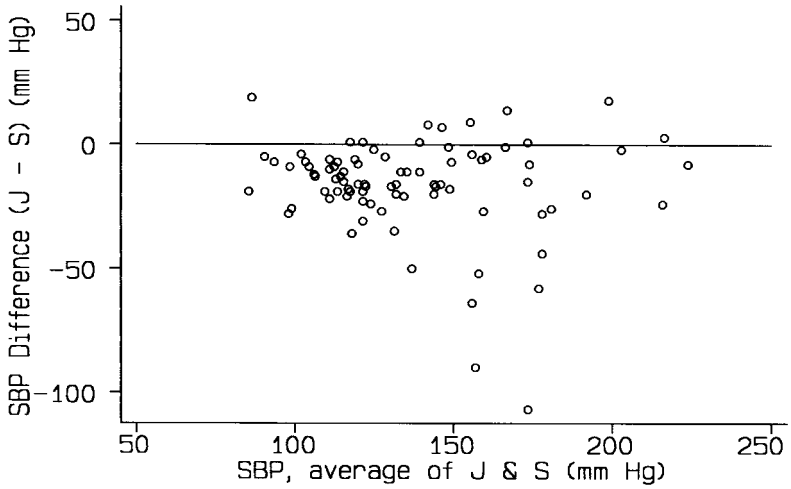

Figure 2 Systolic blood pressure: difference (J−S) versus average of values measured by observer J and machine S

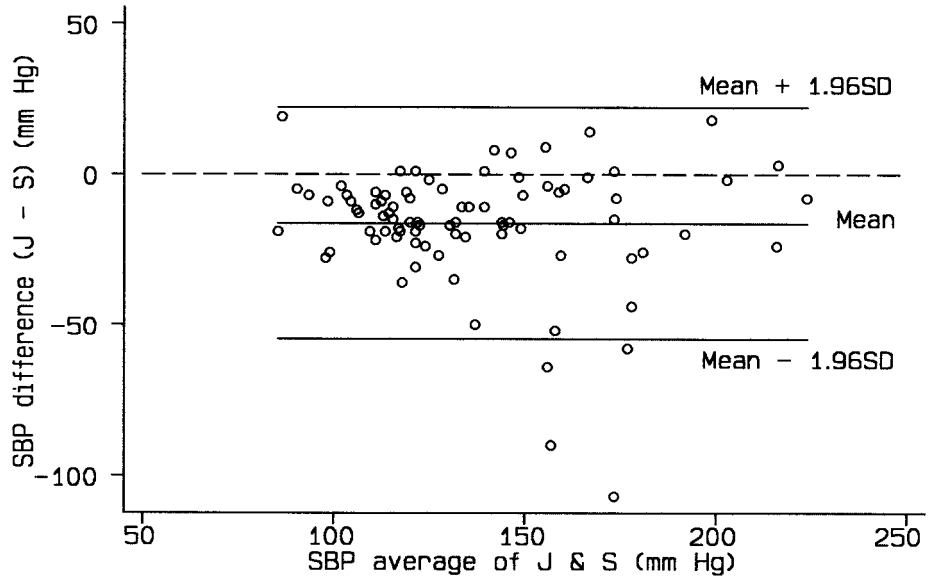

Figure 3 Systolic blood pressure: difference (J−S) versus average of values measured by observer J and machine S with 95% limits of agreement

to the size of the measurement. When methods agree quite closely, however, equal scaling may be impracticable and useful information will again be obscured. We can add the 95% limits of agreement to this plot (Figure 3) to provide a good summary picture.

In such studies we do not know the true values of the quantity we are measuring, so we use the mean of the measurements by the two methods as our best estimate. It is a mistake to plot the difference against either value separately, as the difference will be related to each, a well-known statistical phenomenon.[3]

2.2 Precision of the estimated limits of agreement

We can calculate standard errors and confidence intervals for the limits of agreement if we can assume that the differences follow a distribution which is approximately normal. The variance of $\bar{d}$ is estimated by s_d^2/n, where n is the sample size. Provided the differences are normally distributed, the variance of s_d is approximately estimated by $s_d^2/2(n-1)$. (This because s_d^2 is distributed as $\chi^2 \times \sigma_d^2/(n-1)$ and $\sqrt{\chi^2}$ has approximate variance 1/2.) The mean difference $\bar{d}$ and s_d are independent. Hence the variance of the limits of agreement is estimated by

$$\begin{aligned}\mathrm{Var}(\bar{d} \pm 1.96 s_d) &= \mathrm{Var}(\bar{d}) + 1.96^2 \mathrm{Var}(s_d) \\ &= \frac{s_d^2}{n} + 1.96^2 \frac{s_d^2}{2(n-1)} \\ &= \left(\frac{1}{n} + \frac{1.96^2}{2(n-1)}\right) s_d^2\end{aligned}$$

Unless n is small, this can be approximated closely by

$$\begin{aligned}\mathrm{Var}(\bar{d} \pm 1.96 s_d) &= \left(1 + \frac{1.96^2}{2}\right)\frac{s_d^2}{n} \\ &= 2.92\frac{s_d^2}{n} \\ &= 1.71^2\frac{s_d^2}{n}\end{aligned}$$

Hence, the standard errors of $\bar{d} - 1.96 s_d$ and $\bar{d} + 1.96 s_d$ are approximately $1.71 s_d/\sqrt{n} = 1.71\mathrm{SE}(\bar{d})$. Ninety-five per cent confidence intervals can be calculated by finding the appropriate point of the t distribution with $n - 1$ degrees of freedom. The confidence interval will be t standard errors either side of the observed value.

For the blood pressure data $s_d = 19.61$ mmHg, so the standard error of the bias $\bar{d}$ is $s_d/\sqrt{n} = 19.61/\sqrt{85} = 2.13$ mmHg. For the 95% confidence interval, we have 84 degrees of freedom and t=1.99. Hence the 95% confidence interval for the bias is $-16.29 - 1.99 \times 2.13$ to $-16.29 + 1.99 \times 2.13$, giving -20.5 to -12.1 mmHg. The standard error of the 95% limits of agreement is $1.71\mathrm{SE}(\bar{d}) = 3.64$ mmHg. The 95% confidence interval for the lower limit of agreement is $-54.7 - 1.99 \times 3.64$ to $-54.7 + 1.99 \times 3.64$, giving -61.9 to -47.5 mmHg. Similarly the 95% confidence interval for the upper limit of agreement is $22.1 - 1.99 \times 3.64$ to $22.1 + 1.99 \times 3.64$, giving 14.9 to 29.3 mmHg. These intervals are reasonably narrow, reflecting the quite large sample size. They show, however, that even on the most optimistic interpretation there can be considerable discrepancies between the two methods of measurement and we would conclude that the degree of agreement was not acceptable.

These confidence limits are based on considering only uncertainty due to sampling error. There is the implicit assumption that the sample of subjects is a representative one. Further, all readings with the sphygmomanometer were made by a single (skilled) observer. The calculated uncertainty associated with the limits of agreement is thus likely to be somewhat optimistic.

3 Relationship between difference and magnitude

In Section 2 we assumed that the mean and standard deviation of the differences are the same throughout the range of measurement. The most common departure from the assumptions is an increase in variability of the differences as the magnitude of the measurement increases. In such cases a plot of one method against the other shows a spreading out of the data for larger measurements. The mean difference ($\bar{d}$) may also be approximately proportional to the magnitude of the measurement. These effects are seen even more clearly in the difference versus mean plot. For example, Table 2 shows measurements of plasma volume expressed as a percentage of the expected value for normal individuals. The data are plotted in Figure 4(a). It can be seen immediately that the two methods give systematically different readings, and that all the

Table 2 Measurements of plasma volume expressed as a percentage of normal in 99 subjects, using two alternative sets of normal values due to Nadler and Hurley (data supplied by C Doré, see Cotes *et al.*[21])

Sub	Nadler	Hurley	Sub	Nadler	Hurley	Sub	Nadler	Hurley
1	56.9	52.9	34	93.5	86.0	67	104.8	97.1
2	63.2	59.2	35	94.5	84.3	68	105.1	97.3
3	65.5	63.0	36	94.6	87.6	69	105.5	95.1
4	73.6	66.2	37	95.0	84.0	70	105.7	95.8
5	74.1	64.8	38	95.2	85.9	71	106.1	95.5
6	77.1	69.0	39	95.3	84.4	72	106.8	95.9
7	77.3	67.1	40	95.6	85.2	73	107.2	95.4
8	77.5	70.1	41	95.9	85.2	74	107.4	97.3
9	77.8	69.2	42	96.4	89.2	75	107.5	97.7
10	78.9	73.8	43	97.2	87.8	76	107.5	93.0
11	79.5	71.8	44	97.5	88.0	77	108.0	97.6
12	80.8	73.3	45	97.9	88.7	78	108.2	96.1
13	81.2	73.1	46	98.2	91.2	79	108.6	96.2
14	81.9	74.7	47	98.5	91.8	80	109.1	99.5
15	82.2	74.1	48	98.8	92.5	81	110.1	99.8
16	83.1	74.1	49	98.9	88.0	82	111.2	105.3
17	84.4	76.0	50	99.0	93.5	83	111.7	103.6
18	84.9	75.4	51	99.3	89.0	84	111.7	100.2
19	86.0	74.6	52	99.3	89.4	85	112.0	100.0
20	86.3	79.2	53	99.9	89.2	86	113.1	98.8
21	86.3	77.8	54	100.1	91.3	87	116.0	110.0
22	86.6	80.8	55	101.0	90.4	88	116.7	103.5
23	86.6	77.6	56	101.0	91.2	89	118.8	109.4
24	86.6	77.5	57	101.5	91.4	90	119.7	112.1
25	87.1	78.6	58	101.5	93.0	91	120.7	111.3
26	87.5	78.7	59	101.5	91.2	92	122.8	108.6
27	87.8	81.5	60	101.8	92.0	93	124.7	112.4
28	88.6	79.3	61	101.8	91.8	94	126.4	113.8
29	89.3	78.9	62	102.8	96.8	95	127.6	115.6
30	89.6	85.9	63	102.9	92.8	96	128.2	118.1
31	90.3	80.7	64	103.2	94.0	97	129.6	116.8
32	91.1	80.6	65	103.8	93.5	98	130.4	121.6
33	92.1	82.8	66	104.4	95.8	99	133.2	115.8

observations lie above the line of equality. It is less easy to see that the differences increase as the plasma volume rises, but a plot of difference versus mean shows such an effect very clearly (Figure 4b).

We could ignore the relationship between difference and magnitude and proceed as in Section 2. The analysis will still give limits of agreement which will include most differences, but they will be wider apart than necessary for small plasma volumes, and rather narrower than they should be for large plasma volumes. It is better to try to remove this relationship, either by transformation of the measurements, or, if this fails, by a more general method.

3.1 Logarithmic transformation

Under these circumstances, logarithmic (log) transformation of both measurements before analysis will enable the standard approach to be used. The limits of agreement derived from log transformed data can be back-transformed to give limits for the ratio

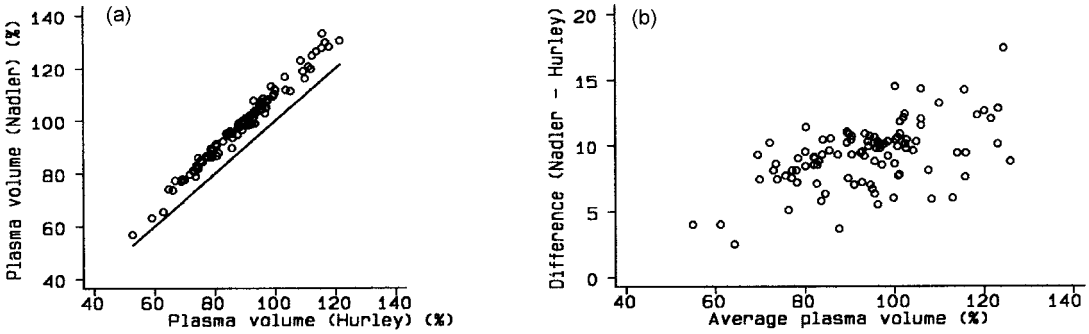

Figure 4 (a) Measurements of plasma volume as listed in Table 2; (b) plot of differences versus average with 95% limits of agreement

of the actual measurements.[2] While other transformations could in principle be used (such as taking square roots or reciprocals), only the log transformation allows the results to be interpreted in relation to the original data. As we think that there is a clear need in such studies to be able to express the results in relation to the actual measurements we do not think that any other transformation should be used in this context.

For the data of Table 2, log transformation is highly successful in producing differences unrelated to the mean. Figure 5 shows the log transformed data and the difference versus mean plot with superimposed 95% limits of agreement. The data clearly meet the requirements of the statistical method very well. The mean difference (log Nadler − log Hurley) is 0.099 with 95% limits of agreement 0.056 and 0.141. Because of the high correlation and reasonably large sample size the confidence intervals for the limits of agreement are narrow. For example, the 95% confidence interval for the lower limit is from 0.049 to 0.064.

These results relate to differences between log percentages and are not easy to interpret. As noted above, we can back-transform (antilog) the results to get values

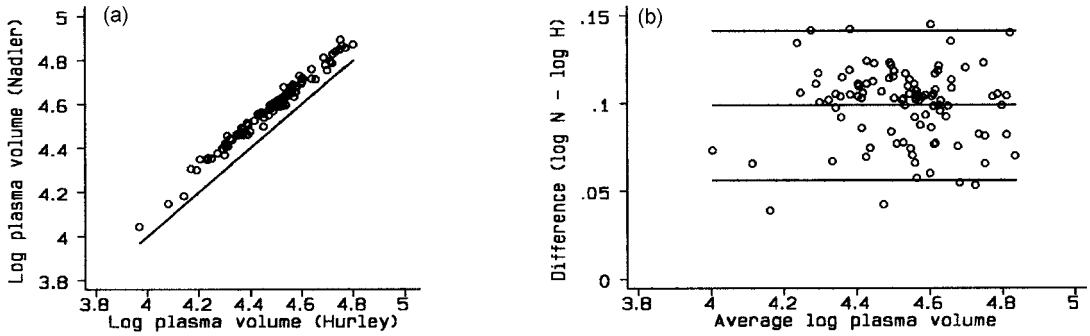

Figure 5 (a) Measurements of plasma volume after log transformation; (b) Difference between plasma volume measurements plotted against their average after $\log_e$ transformation with 95% limits of agreement

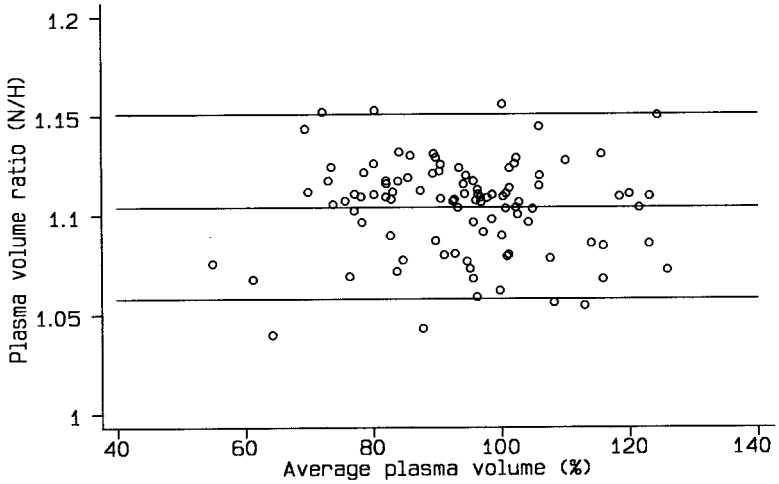

Figure 6 Ratio of plasma volume measurements plotted against their average

relating to the ratios of measurements by the two methods. The geometric mean ratio of values by the Nadler and Hurley methods was 1.11 with 95% limits of agreement 1.06 to 1.15. Thus the Nadler method exceeds the Hurley method by between 1.06 and 1.15 times, i.e. by between 6% and 15%, for most measurements.

This example illustrates the importance of both the bias ($\bar{d}$) and the standard deviation of between-method differences (s_d). Here there is a clear bias with all observations lying to one side of the line of equality (Figure 4a). However, because the scatter around the average difference is rather small we could get excellent agreement between the two methods if we first applied a conversion factor, multiplying the Hurley method or dividing the Nadler method by 1.11.

We can make the transformation process more transparent by working directly with the ratios. Instead of taking logs and calculating differences we can simply calculate the ratio of the two values for each subject and calculate limits of agreement based on the mean and SD of these. Figure 6 shows the plot corresponding to this approach, which is almost identical to Figure 5(b).

The type of plot shown in Figure 6 was suggested as a general purpose approach for method comparison studies,[4] although without any suggestion for quantifying the differences between the methods. Another variation is to plot on the vertical axis the difference between the methods as a percentage of their average.[5]

3.2 A regression approach for nonuniform differences

Sometimes the relationship between difference and mean is complicated and log transformation may not solve the problem. For example, the differences may tend to be in one direction for low values of the quantity being measured and in the other direction for high values. For data sets for which log transformation does not remove the relationship between the differences and the size of the measurement, a plot in the style of Figure 5(b) is still enormously helpful in comparing the methods. In such

cases, formal analysis as described in Section 2 will tend to give limits of agreement which are too far apart rather than too close, and so should not lead to the acceptance of poor methods of measurement. Nevertheless, it is useful to have better approaches to deal with such data.

We propose modelling the variability in the SD of the d_i directly as a function of the level of the measurement, using a method based on absolute residuals from a fitted regression line. If the mean also changes as a function of level we can first model that relation and then model the SD of the residuals. This approach is based on one used to derive age-related reference intervals.[6]

We first consider the mean difference between the methods in relation to the size of the measurement. We suggested in our first paper on method comparison studies[1] that when the agreement between the methods varies as the measurement varies, we can regress the difference between the methods (D) on the average of the two methods (A). However, we did not give a worked example, and did not consider there the possibility that the standard deviation of the differences, s_d, might also vary with A. A similar idea was put forward by Marshall *et al.*,[7] who use a more complex method than that proposed here.

Unless the plot of the data shows clear curvature (which is very unlikely in our experience) simple linear regression is all that is needed, giving

$$\hat{D} = b_0 + b_1 A \tag{3.1}$$

If the slope b_1 is not significant then $\hat{D} = \bar{d}$, the mean difference. We have deliberately not defined what we mean by significant here as we feel that this may require clinical judgement as well as statistical considerations. If b_1 is significantly different from zero we obtain the estimated difference between the methods from equation (3.1) for any true value of the measurement, estimated by A.

Table 3 shows the estimated fat content of human milk (g/100 ml) determined by the measurement of glycerol released by enzymic hydrolysis of triglycerides and

Table 3 Fat content of human milk determined by enzymic procedure for the determination of triglycerides and measured by the Standard Gerber method (g/100 ml)[8]

Trig.	Gerber	Trig.	Gerber	Trig.	Gerber
0.96	0.85	2.28	2.17	3.19	3.15
1.16	1.00	2.15	2.20	3.12	3.15
0.97	1.00	2.29	2.28	3.33	3.40
1.01	1.00	2.45	2.43	3.51	3.42
1.25	1.20	2.40	2.55	3.66	3.62
1.22	1.20	2.79	2.60	3.95	3.95
1.46	1.38	2.77	2.65	4.20	4.27
1.66	1.65	2.64	2.67	4.05	4.30
1.75	1.68	2.73	2.70	4.30	4.35
1.72	1.70	2.67	2.70	4.74	4.75
1.67	1.70	2.61	2.70	4.71	4.79
1.67	1.70	3.01	3.00	4.71	4.80
1.93	1.88	2.93	3.02	4.74	4.80
1.99	2.00	3.18	3.03	5.23	5.42
2.01	2.05	3.18	3.11	6.21	6.20

measurements by the standard Gerber method.[8] Figure 7(a) shows that the two methods agree closely, but from Figure 7(b) we can see a tendency for the differences to be in opposite directions for low and high fat content. The variation (SD) of the differences seems much the same for all levels of fat content. These impressions are confirmed by regression analyses. The regression of D on A gives

$$\hat{D} = 0.079 - 0.0283A \text{ g/100 ml}$$

as noted by Lucas *et al.*[8] The SD of the residuals is 0.08033.

In the second stage of the analysis we consider variation around the line of best agreement (equation (3.1)). We need to model the scatter of the residuals from model (3.1) as a function of the size of the measurement (estimated by A). Modelling is considerably simplified by the assumption that these residuals have a normal distribution whatever the size of the measurement, which is a fairly natural extension of the assumption we make already in such analyses. We then regress the absolute values of the residuals, which we will call R, on A to get

$$\hat{R} = c_0 + c_1 A \tag{3.2}$$

If we take a normal distribution with mean zero and variance σ^2, it is easy to show that the mean of the absolute values, which follow a half-normal distribution, is $\sigma\sqrt{2/\pi}$. The standard deviation of the residuals is thus obtained by multiplying the fitted values by $\sqrt{\pi/2}$. The limits of agreement are obtained by combining the two regression equations (see Altman[6]).

Although in principle any form of regression model might be fitted here, it is very likely that if the SD is not constant then linear regression will be adequate to describe the relationship. If there is no 'significant' relation between R and A the estimated standard deviation is simply the standard deviation of the adjusted differences, the residuals of equation (3.1).

In the general case where both models (3.1) and (3.2) are used, the expected value of the difference between methods is given by $\hat{D} = b_0 + b_1 A$ and the 95% limits of

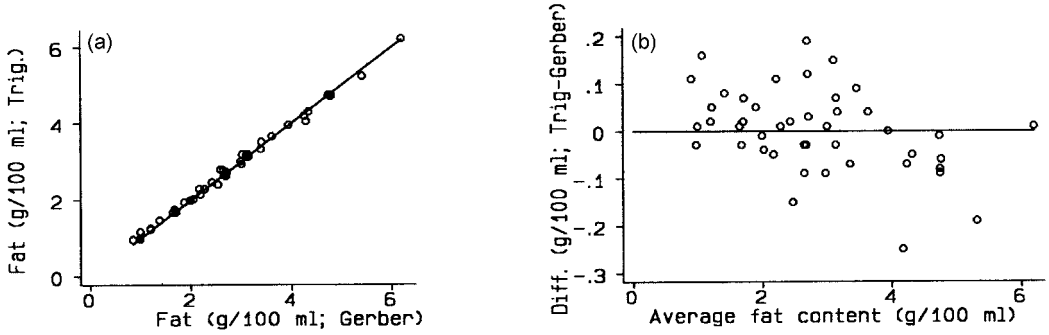

Figure 7 (a) Fat content of human milk determined by enzymic procedure for the determination of triglycerides and measured by the standard Gerber method (g/100 ml); (b) plot of difference (Triglyceride–Gerber) against the average

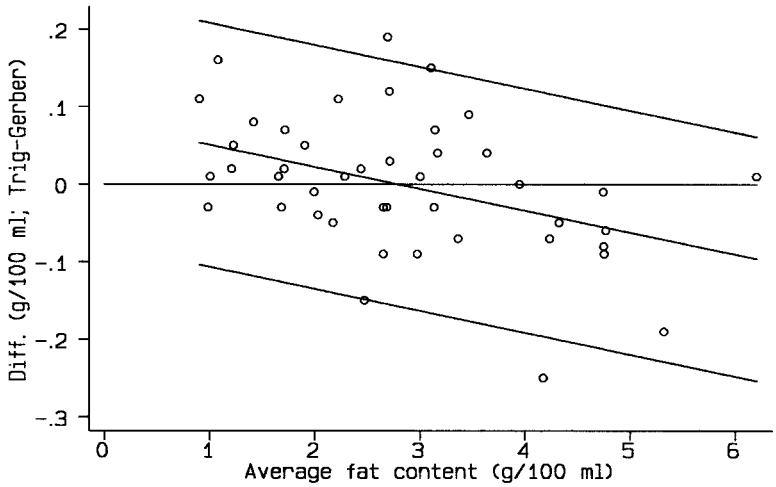

Figure 8 Regression based limits of agreement for difference in fat content of human milk determined by Triglyceride and Gerber methods (g/100 ml)

agreement are obtained as

$$\hat{D} \pm 1.96\sqrt{\pi/2}\,\hat{R} = \hat{D} \pm 2.46\hat{R}$$

or as

$$b_0 + b_1 A \pm 2.46\{c_0 + c_1 A\} \tag{3.3}$$

Returning to the example, there was no relation between the residuals from the first regression model and A and so the SD of the adjusted differences is simply the residual SD from the regression, so that $s_d = 0.08033$. We can thus calculate the regression based 95% limits of agreement as $0.079 - 0.0283a \pm 1.96 \times 0.08033$ g/100 ml, where a is the magnitude (average of methods) of the fat content. These values are shown in Figure 8. Of course, in clinical practice, when only one method is being used, the observed value by that method provides the value of a.

4 The importance of repeatability

The comparison of the repeatability of each method is relevant to method comparison because the repeatabilities of two methods of measurement limit the amount of agreement which is possible. Curiously, replicate measurements are rarely made in method comparison studies, so that an important aspect of comparability is often overlooked. If we have only one measurement using each method on each subject we cannot tell which method is more repeatable (precise). Lack of repeatability can interfere with the comparison of two methods because if one method has poor repeatability, in the sense that there is considerable variation in repeated measure-

ments on the same subject, the agreement between the two methods is bound to be poor. Even if the measurements by the two methods agreed very closely on average, poor repeatability of one method would lead to poor agreement between the methods for individuals. When the old method has poor repeatability even a new method which was perfect would not agree with it. Lack of agreement in unreplicated studies may suggest that the new method cannot be used, but it might be caused by poor repeatability of the standard method. If both methods have poor repeatability, then poor agreement is highly likely. For this reason we strongly recommend the simultaneous estimation of repeatability and agreement by collecting replicated data.

It is important first to clarify exactly what we mean when we refer to replicate observations. By replicates we mean two or more measurements on the same individual taken in identical conditions. In general this requirement means that the measurements are taken in quick succession.

One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice.

4.1 Estimating repeatability

A very similar analysis to the limits of agreement approach can be applied to quantify the repeatability of a method from replicated measurements obtained by the same method. Using one-way analysis of variance, with subject as the factor, we can estimate the within-subject standard deviation, s_w, from the square root of the residual mean square. We can compare the standard deviations of different methods to see which is more repeatable. Each standard deviation can also be used to calculate limits within which we expect the differences between two measurements by the same method to lie. As well as being informative in its own right, the repeatability indicates a baseline against which to judge between-method variability.

The analysis is simple because we expect the mean difference between replicates to be zero – we do not usually expect second measurements of the same samples to differ systematically from first measurements. Indeed, such a systematic difference would indicate that the values were not true replicates. A plot should show whether the assumption is reasonable, and also whether the differences are independent of the mean. If repeatability gets worse as the measurements increase we may need first to log transform the data in the same way as for comparing methods.

Returning to the blood pressure data shown in Table 1, we can estimate the repeatability of each method. For observer J using the sphygmomanometer the within-subject variance is 37.408. Likewise for observer R we have $s_w^2 = 37.980$ and for the machine (S) we have $s_w^2 = 83.141$. We can see that both observers have much better repeatability than the machine and that their performance is almost identical.

Two readings by the same method will be within $1.96\sqrt{2}\,s_w$ or $2.77 s_w$ for 95% of subjects. This value is called the *repeatability coefficient*. For observer J using the sphygmomanometer $s_w = \sqrt{37.408} = 6.116$ and so the repeatability coefficient is $2.77 \times 6.116 = 16.95$ mmHg. For the machine S, $s_w = \sqrt{83.141} = 9.118$ and the repeatability coefficient is $2.77 \times 9.118 = 25.27$ mmHg. Thus, the repeatability of the machine is 50% greater than that of the observer. We can compare these 95%

repeatability coefficients to the 95% limits of agreement. The 95% limits of agreement correspond to the interval $-2.77s_w$ to $2.77s_w$. If these are similar, then the lack of agreement between the methods is explained by lack of repeatability. If the limits of agreement are considerably wider than the repeatability would indicate, then there must be some other factor lowering the agreement between methods.

The use of the within-subject standard deviation does not imply that other approaches to repeatability, such as intraclass correlation, are not appropriate. The use of s_w, however, facilitates the comparison with the limits of agreement. It also helps in the interpretation of the individual measurement, being in the same units.

5 Measuring agreement using repeated measurements

When we have repeated measurements by two methods on the same subjects it is clearly desirable to use all the data when comparing methods. A sensible first step is to calculate the mean of the replicate measurements by each method on each subject. We can then use these pairs of means to compare the two methods using the limits of agreement method. The estimate of bias will be unaffected by the averaging, but the estimate of the standard deviation of the differences will be too small, because some of the effect of measurement error has been removed. We want the standard deviation of differences between single measurements, not between means of several repeats. Here we describe some methods for handling such data, first for the case with equal replication and then allowing unequal numbers of replicates.

We assume that even though multiple readings are available the standard clinical measurement is a single value. Where it is customary to use the average of two or more measurements in clinical practice (e.g. with peak expiratory flow) the approach described below would not be used. Rather the limits of agreement method would be applied directly to the means.

5.1 Equal numbers of replicates

When we make repeated measurements of the same subject by each of two methods, the measurements by each method will be distributed about the expected measurement by that method for that subject. These means will not necessarily be the same for the two methods. The difference between method means may vary from subject to subject. This variability constitutes method times subject interaction. Denote the measurements on the two methods by X and Y. We are interested in the variance of the difference between single measurements by each method, $D = X - Y$. If we partition the variance for each method we get

$$\mathrm{Var}(X) = \sigma_t^2 + \sigma_{xI}^2 + \sigma_{xw}^2$$

$$\mathrm{Var}(Y) = \sigma_t^2 + \sigma_{yI}^2 + \sigma_{yw}^2$$

where σ_t^2 is the variance of the true values, σ_{xI}^2 and σ_{yI}^2 are method times subject interaction terms, and σ_{xw}^2 and σ_{yw}^2 are the within-subject variances from measure-

ments by the same method, for X and Y, respectively. It follows that the variance of the between-subject differences for single measurements by each method is

$$\mathrm{Var}(X - Y) = \sigma_D^2 = \sigma_{xI}^2 + \sigma_{yI}^2 + \sigma_{xw}^2 + \sigma_{yw}^2 \tag{5.1}$$

We wish to estimate this variance from an analysis of the means of the measurement for each subject, $\bar{D} = \bar{X} - \bar{Y}$, that is from $\mathrm{Var}(\bar{X} - \bar{Y})$. With this model, the use of the mean of replicates will reduce the within-subject variance but it will not affect the interaction terms, which represent patient-specific differences. We thus have

$$\mathrm{Var}(\bar{X}) = \sigma_t^2 + \sigma_{xI}^2 + \frac{\sigma_{xw}^2}{m_x}$$

where m_x is the number of observations on each subject by method X, because only the within-subject within-method error is being averaged. Similarly

$$\mathrm{Var}(\bar{Y}) = \sigma_t^2 + \sigma_{yI}^2 + \frac{\sigma_{yw}^2}{m_y}$$

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{D}) = \sigma_{xI}^2 + \frac{\sigma_{xw}^2}{m_x} + \sigma_{Iy}^2 + \frac{\sigma_{yw}^2}{m_y}$$

The distribution of $\bar{D}$ depends only on the errors and interactions, because the true value is included in both X and Y, which are differenced. It follows from equation (5.1) that

$$\mathrm{Var}(X - Y) = \mathrm{Var}(\bar{D}) + \left(1 - \frac{1}{m_x}\right)\sigma_{xw}^2 + \left(1 - \frac{1}{m_y}\right)\sigma_{yw}^2 \tag{5.2}$$

If $s_{\bar{d}}^2$ is the observed variance of the differences between the within-subject means, $\mathrm{Var}(X - Y) = \sigma_d^2$ is estimated by

$$\hat{\sigma}_d^2 = s_{\bar{d}}^2 + \left(1 - \frac{1}{m_x}\right)s_{xw}^2 + \left(1 - \frac{1}{m_y}\right)s_{yw}^2 \tag{5.3}$$

In the common case with two replicates of each method we have

$$\hat{\sigma}_d^2 = s_{\bar{d}}^2 + \frac{s_{xw}^2}{2} + \frac{s_{yw}^2}{2}$$

as given by Bland and Altman.[2] It is easy to see that the method still works when one method is replicated and the other is not.

For an example, we compare observer J and machine S from Table 1. From Section 4.1 the within-subject variances of the two methods of measuring blood pressure are 37.408 for the sphygmomanometer (observer J) and 83.141 for the machine (S). The mean difference is −15.62 mmHg. The variance of the differences between-subject means is $s_{\bar{d}}^2 = 358.493$ and we have $m_J = m_S = 3$ observations by each method on each subject. From equation (5.3) the adjusted variance of differences is given by

$$\hat{\sigma}_d^2 = 358.493 + \left(1 - \frac{1}{3}\right) \times 37.408 + \left(1 - \frac{1}{3}\right) \times 83.141 = 438.859$$

so $s_d = \sqrt{438.859} = 20.95$ mmHg and the 95% limits of agreement are $-15.62 - 1.96 \times 20.95 = -56.68$ and $-15.62 + 1.96 \times 20.95 = 25.44$. This estimate is very similar to that obtained using a single replicate (Section 2).

An approximate standard error and confidence interval for these limits of agreement can be found as follows. Provided the measurement errors are normal and independent, for n subjects $n(m_x - 1)s_{xw}^2/\sigma_{xw}^2$ follows a chi-squared distribution with $n(m_x - 1)$ degrees of freedom, and so has variance $2n(m_x - 1)$. Hence

$$\mathrm{Var}\left(s_{xw}^2\right) = \frac{2\sigma_{xw}^4}{n(m_x - 1)} \quad \text{and} \quad \mathrm{Var}\left(s_{yw}^2\right) = \frac{2\sigma_{yw}^4}{n(m_y - 1)} \tag{5.4}$$

and the variance of the correction term of equation (5.3) is given by

$$\mathrm{Var}\left(\left(1 - \frac{1}{m_x}\right)s_{xw}^2 + \left(1 - \frac{1}{m_y}\right)s_{yw}^2\right) =$$

$$\left(1 - \frac{1}{m_x}\right)^2 \frac{2\sigma_{xw}^4}{n(m_x - 1)} + \left(1 - \frac{1}{m_y}\right)^2 \frac{2\sigma_{yw}^4}{n(m_y - 1)} \tag{5.5}$$

$$= \frac{2(m_x - 1)\sigma_{xw}^4}{nm_x^2} + \frac{2(m_y - 1)\sigma_{yw}^4}{nm_y^2} \tag{5.6}$$

Similarly, the variance of $s_{\bar{d}}^2$ is given by

$$\mathrm{Var}(s_{\bar{d}}^2) = \frac{2\sigma_d^4}{n - 1} \tag{5.7}$$

Applying equations (5.6) and (5.7) to equation (5.3) we have

$$\mathrm{Var}(\hat{\sigma}_d^2) = \frac{2\sigma_d^4}{n - 1} + \frac{2(m_x - 1)\sigma_{xw}^4}{nm_x^2} + \frac{2(m_y - 1)\sigma_{yw}^4}{nm_y^2} \tag{5.8}$$

Using the well-known results that

$$\mathrm{Var}(f(z)) \approx \left(\frac{df(z)}{dz}\right)_{z=\mathrm{E}(z)}^2 \mathrm{Var}(z)$$

and hence that

$$\mathrm{Var}(\sqrt{z}) \approx \left(\frac{1}{2\sqrt{z}}\right)_{z=\mathrm{E}(z)}^2 \mathrm{Var}(z) = \frac{1}{4\mathrm{E}(z)}\mathrm{Var}(z)$$

we have for $\mathrm{Var}(\hat{\sigma}_d)$

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}_d) &= \frac{1}{4\sigma_d^2}\left(\frac{2\sigma_{\bar{d}}^4}{n-1} + \frac{2(m_x-1)\sigma_{xw}^4}{nm_x^2} + \frac{2(m_y-1)\sigma_{yw}^4}{nm_y^2}\right) \\ &= \frac{1}{2\sigma_d^2}\left(\frac{\sigma_{\bar{d}}^4}{n-1} + \frac{(m_x-1)\sigma_{xw}^4}{nm_x^2} + \frac{(m_y-1)\sigma_{yw}^4}{nm_y^2}\right) \end{aligned} \tag{5.9}$$

The variance of the mean difference $\bar{d}$ is estimated by $\hat{\sigma}_d^2/n$, and mean and standard deviation of the differences are independent. Substituting the estimates of the variances in equation (5.9), the variance of the limits of agreement $\bar{d} \pm 1.96\hat{\sigma}_d$ is estimated by

$$\mathrm{Var}(\bar{d} \pm 1.96\hat{\sigma}_d) = \frac{\hat{\sigma}_d^2}{n} + \frac{1.96^2}{2\hat{\sigma}_d^2}\left(\frac{s_{\bar{d}}^4}{n-1} + \frac{(m_x-1)s_{xw}^4}{nm_x^2} + \frac{(m_y-1)s_{yw}^4}{nm_y^2}\right) \tag{5.10}$$

For $m_x = m_y = 2$ this equation becomes

$$\mathrm{Var}(\bar{d} \pm 1.96\hat{\sigma}_d) = \frac{\hat{\sigma}_d^2}{n} + \frac{1.96^2}{2\hat{\sigma}_d^2}\left(\frac{s_{\bar{d}}^4}{n-1} + \frac{s_{xw}^4}{4n} + \frac{s_{yw}^4}{4n}\right)$$

and for unreplicated observations with $m_x = m_y = 1$, $\hat{\sigma}_d$ is replaced by the direct estimate s_d and we have

$$\begin{aligned}\mathrm{Var}(\bar{d} \pm 1.96s_d) &= \frac{s_d^2}{n} + \frac{1.96^2}{2s_d^2}\frac{s_d^4}{n-1} \\ &= s_d^2\left(\frac{1}{n} + \frac{1.96^2}{2(n-1)}\right)\end{aligned}$$

as in Section 2.2. These values can be used to estimate 95% confidence intervals for the limits of agreement.

For the blood pressure data, the variance of the limits of agreement is

$$\begin{aligned}\mathrm{Var}(\bar{D} \pm 1.96s_D) = &\\ &\frac{438.859}{85} + \frac{1.96^2}{2 \times 438.859}\left(\frac{358.493^2}{84} + \frac{2 \times 37.4078^2}{9 \times 85} + \frac{2 \times 83.1412^2}{9 \times 85}\right) \\ &= 11.9941\end{aligned}$$

Hence the standard error is $\sqrt{11.9941} = 3.463$ mmHg. The 95% confidence interval for the lower limit of agreement is $-56.68 - 1.96 \times 3.463$ to $-56.68 + 1.96 \times 3.463$, giving -63.5 to -49.9 and for the upper limit of agreement $25.44 - 1.96 \times 3.463$ to $25.44 + 1.96 \times 3.463$, giving 18.70 to 32.2 mmHg.

The standard error here is very similar to that found for only one replicate (3.64 mmHg) in Section 2.2. The use of replicates only reduces that part of the variation due each method's lack of precision, and the method times subject

interaction component remains. If this is large, e.g. if large discrepancies for a subject exist in all replicates (as is often the case in our experience) replication does not improve the precision of the limits of agreement much. We still advocate two replicates, however, so that method repeatability can be investigated.

This method is different from that in our original paper[1] which ignored the subject times method interaction. We now think that approximation was unreasonable and that the method given here is clearly superior.

5.2 Unequal numbers of replicates

We now consider the case where there are unequal numbers of observations per subject, m_{xi} and m_{yi} by methods X and Y on subject i. Such data can arise, for example, if patients are measured at regular intervals during a procedure of variable length, such as surgery. For example, Table 4 shows measurements of cardiac output by two methods, impedance cardiography (IC) and radionuclide ventriculography (RV), on 12 subjects.[9] The solution for equally replicated data (equation (5.2)) depends on the well known result that the variance of the mean of n independent random variables with the same mean and variance σ^2 is σ^2/n. If W_i is the mean of m_i observations with mean

Table 4 Cardiac output by two methods, RV and IS, for 12 subjects (data provided by Dr LS Bowling[9])

Sub	RV	IC	Sub	RV	IC	Sub	RV	IC
1	7.83	6.57	5	3.13	3.03	9	4.48	3.17
1	7.42	5.62	5	2.98	2.86	9	4.92	3.12
1	7.89	6.90	5	2.85	2.77	9	3.97	2.96
1	7.12	6.57	5	3.17	2.46	10	4.22	4.35
1	7.88	6.35	5	3.09	2.32	10	4.65	4.62
2	6.16	4.06	5	3.12	2.43	10	4.74	3.16
2	7.26	4.29	6	5.92	5.90	10	4.44	3.53
2	6.71	4.26	6	6.42	5.81	10	4.50	3.53
2	6.54	4.09	6	5.92	5.70	11	6.78	7.20
3	4.75	4.71	6	6.27	5.76	11	6.07	6.09
3	5.24	5.50	7	7.13	5.09	11	6.52	7.00
3	4.86	5.08	7	6.62	4.63	11	6.42	7.10
3	4.78	5.02	7	6.58	4.61	11	6.41	7.40
3	6.05	6.01	7	6.93	5.09	11	5.76	6.80
3	5.42	5.67	8	4.54	4.72	12	5.06	4.50
4	4.21	4.14	8	4.81	4.61	12	4.72	4.20
4	3.61	4.20	8	5.11	4.36	12	4.90	3.80
4	3.72	4.61	8	5.29	4.20	12	4.80	3.80
4	3.87	4.68	8	5.39	4.36	12	4.90	4.20
4	3.92	5.04	8	5.57	4.20	12	5.10	4.50

μ and variance σ^2, so having variance σ^2/m_i, then the expected variance of the means will be

$$\mathrm{Var}(W_i) = \frac{1}{n}\left(\sum \frac{1}{m_i}\right)\sigma^2 \tag{5.11}$$

For subject i we have m_{xi} observations by method X and m_{yi} observations by method Y. For each subject, we calculate the differences between means of measurements by the two methods and then calculate the variance of these differences. The expected value of this variance estimate is thus

$$\mathrm{Var}(\bar{D}) = \sigma^2_{xI} + \frac{1}{n}\left(\sum \frac{1}{m_{xi}}\right)\sigma^2_{xw} + \sigma^2_{yI} + \frac{1}{n}\left(\sum \frac{1}{m_{yi}}\right)\sigma^2_{yw} \tag{5.12}$$

Thus we have

$$\mathrm{Var}(D) = \mathrm{Var}(\bar{D}) + \left(1 - \frac{1}{n}\left(\sum \frac{1}{m_{xi}}\right)\right)\sigma^2_{xw} + \left(1 - \frac{1}{n}\left(\sum \frac{1}{m_{yi}}\right)\right)\sigma^2_{yw} \tag{5.13}$$

which reduces to equation (5.2) when $m_{xi} = m_x$ and $m_{yi} = m_y$.

For the data of Table 4, we must first check the assumption that the variances are independent of the subject means. For each method separately, we can plot the within-subject standard deviation against the subject mean. As Figure 9 shows, the assumption of independence is reasonable for these data. Then, for each subject, we plot the difference between the means for the two methods against their average (Figure 10). Again the assumption of independence is reasonable. We next estimate σ^2_{xw} and σ^2_{yw} by one-way analyses of variance for RV and IC separately, giving $s^2_{xw} = 0.1072$ and $s^2_{yw} = 0.1379$.

In this case

$$\frac{1}{n}\left(\sum \frac{1}{m_{xi}}\right) \quad \text{and} \quad \frac{1}{n}\left(\sum \frac{1}{m_{yi}}\right)$$

are the same, equal to 0.2097, because the data for each subject are balanced (this need

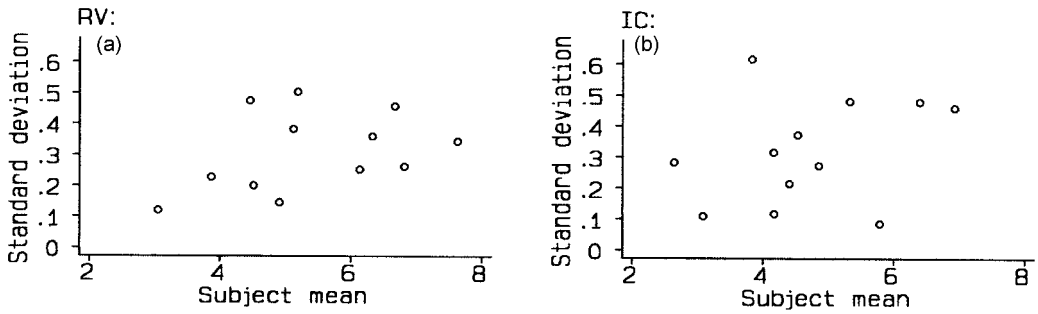

Figure 9 Subject standard deviation against subject mean for each method of measurement of cardiac output

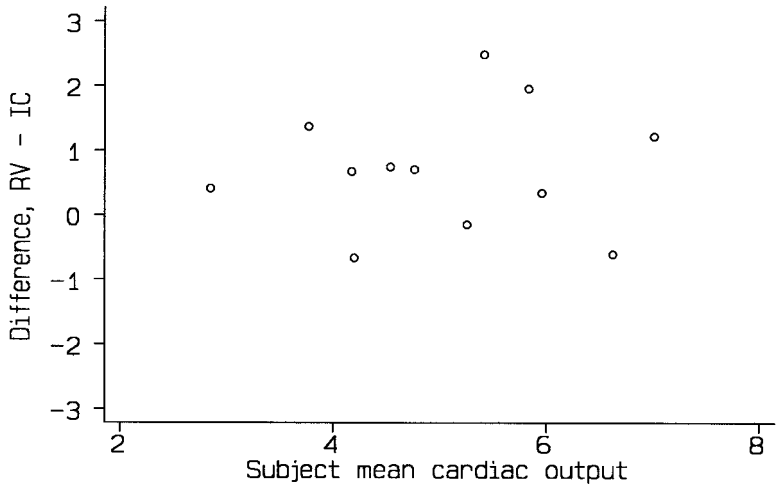

Figure 10 Subject difference, RV − IC, against mean

not be the case). The variance of the mean difference between methods for each subject is 0.9123. To calculate the variance of the difference between single observations by the two methods, we use equation (5.13). We have

$$\text{Var}(\text{IC} - \text{RV}) = 0.9123 + (1 - 0.2097) \times 0.1072 + (1 - 0.2097) \times 0.1379 = 1.1060$$

The standard deviation of differences between single observations by the two methods is estimated by $\hat{\sigma}_d = \sqrt{1.1060} = 1.0517$. The mean difference was 0.7092, so the 95% limits of agreement for RV − IC are $0.7092 - 1.96 \times 1.0517 = -1.3521$ and $0.7092 + 1.96 \times 1.0517 = 2.7705$. Thus, a measurement by RV is unlikely to exceed a measurement by IC by more than 2.77, or be more than 1.35 below.

5.3 Replicated data in pairs

The methods of Section 5.1 and Section 5.2 assume that the subject's true value does not change between repeated measurements. Sometimes, we are interested in measuring the instantaneous value of a continually changing quantity. We might make several pairs of measurements by two methods on each subject, where the underlying true value changes from pair to pair. We can estimate the limits of agreement by a components of variance technique. We use the differences for each pair of measurements. The difference for pair of measurements j on subject i may be modelled as

$$D_{ij} = B + I_i + E_{ij}$$

where B is the constant bias, I_i the subjects times methods interaction term, and X_{ij} the random error within the subject for that pair of observations. The variance of D_{ij} is thus

$$\sigma_d^2 = \sigma_{dI}^2 + \sigma_{dw}^2$$

We can estimate σ_{dI}^2 and σ_{dw}^2 by components of variance estimation from one way analysis of variance.[10] Suppose that for subject i we have m_i pairs of observations and there are n subjects. We have the within-subject or error mean square MS_w and the between-subjects mean square MS_b. Then the components of variance can be estimated by $\hat{\sigma}_{dw}^2 = MS_w$ and

$$\hat{\sigma}_{dI}^2 = \frac{(\sum m_i)^2 - \sum m_i^2}{(n-1)\sum m_i}(MS_b - MS_w)$$

The sum of these estimates provides $\hat{\sigma}_d^2$. The mean bias is estimated by $(\sum m_i \bar{d}_i)/(\sum m_i)$, where $\bar{d}_i$ is the mean difference for subject i. Hence we estimate the 95% limits of agreement. Methods for the calculation of confidence intervals for combinations of components of variance are given by Burdick and Graybill.[11]

6 Nonparametric approach to comparing methods

The between-method differences do not always have a normal distribution. As we noted in Section 2, in general this will not have a great impact on the limits of agreement. Nevertheless, if there are one or more extreme discrepancies between the methods a nonparametric approach may be felt preferable. Such a situation arguably arises in the evaluation of (semi-) automatic blood pressure recording machines, as was illustrated by the blood pressure data shown in Figure 1. For this reason, the British Hypertension Society (BHS) protocol for evaluating these machines recommended a simple nonparametric method.[12]

We can retain the basic approach outlined in Section 2.1 up to and including the plot of the differences versus mean values of the two methods. There are then two similar ways of describing such data without assuming a normal distribution of differences. First, we can calculate the proportion of differences greater than some reference values (such as ± 10 mmHg). The reference values can be indicated on the scatter diagram showing the difference versus the mean. Second, we can calculate the values outside which a certain proportion (say 10%) of the observations fell. To do this we simply order the observations and take the range of values remaining after a percentage (say 5%) of the sample is removed from each end. The centiles can also be superimposed on the scatter diagram. This second method is effectively a nonparametric form of the limits of agreement method. The two nonparametric methods are generally less reliable than those obtained using normal distribution theory, especially in small samples. Confidence intervals can be constructed using the standard method for binomial proportions or the standard error of a centile.

The BHS protocol for evaluating blood pressure measuring devices suggested the use of the first of the above ideas, using the percentage of differences within certain limits.[12] Three such assessments are made, relating to the percentage of differences

Table 5 Grading of blood pressure devices based on differences between measurements (in mmHg) by device and those by sphygmomanometer[12]

Grade	Difference (mmHg) ≤5	≤10	≤15
A	60	85	95
B	50	75	90
C	40	65	85
D	fails to achieve C		

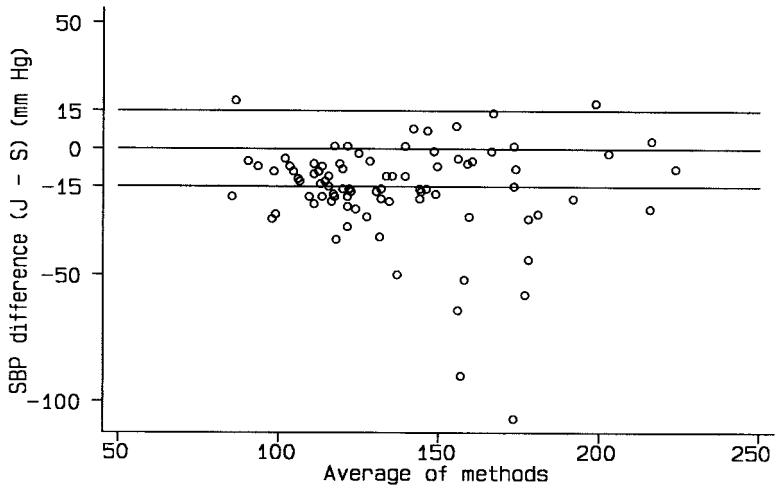

Figure 11 Data from Figure 2 showing differences between methods of ±15 mmHg

within 5, 10 and 15 mmHg. Table 5 shows the conditions which the data must meet to receive a grade of A, B, or C, which were based on what could be achieved using a sphygmomanometer.

An example is shown in Figure 11, using the same data as are shown in Figure 1. Only the values of ±15 are shown as the spread of differences was so large. For these data the percentages of between-methods differences within 5, 10 and 15 mmHg were 16%, 35% and 49%, so that the device clearly gets a grade D.

The nonparametric method is disarmingly simple yet provides readily interpreted results. It has been used before[13,14] but apparently only rarely. Perhaps its simplicity has led to the belief that it is not a proper analysis of the data.

7 Discussion

Previously, we have described the limits of agreement approach[1,2] and the nonparametric variant.[12] In this paper we have extended the method in several ways. We have also described a powerful method for dealing with data where the agreement varies in a complex way across the range of the measurement and we have described several approaches for replicated data.

Our approach is based on the philosophy that the key to method comparison studies is to quantify disagreements between individual measurements. It follows that we do not see a place for methods of analysis based on hypothesis testing. Agreement is not something which is present or absent, but something which must be quantified. Nor do we see a role for methods which lead to global indices, such as correlation coefficients. They do not help the clinician interpret a measurement, though they have a place in the study of associated questions such as the validity of measurement methods. Widely used statistical approaches which we think are misleading include correlation,[1,17,18,19] regression,[1] and the comparison of means.[1] Other methods which we think inappropriate are structural equations[15] and intraclass correlation.[16]

We advocate the collection of replicated data in method comparison studies, because this enables us to compare the agreement between the two methods with the agreement each method has to itself, its repeatability (Section 4.1). Such a study should be designed to have equal numbers of measurements by each method for each subject and the methods of Section 5.1 can be used for its analysis. We are rather disappointed that so few of the studies which cite our work have adopted the use of replicates. Other studies using replicates have usually adopted this as a convenience because subjects are hard to find, often because the measurements are very invasive. For these studies we offer two methods of analysis, one for use when the underlying value is assumed to remain constant and the other for when it assumed to vary.

We think that any method for analysing such studies should produce numbers which are useful to and easily understood by the users of measurement methods. For rapid adoption by the research community they should be applicable using existing software. The basic method which we have proposed and the extensions to cover the most frequent situations meets both these criteria. Only the more unusual designs and relationships between agreement and magnitude described in this paper should require the intervention of a statistician.

Acknowledgements

We thank the researchers who have provided the data included in this paper. The paper is drawn largely from work in progress for our forthcoming book *Statistical approaches to medical measurement*, Oxford University Press.

References

1 Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; **32**: 307–17.
2 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i**: 307–10.
3 Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995; **346**: 1085–87
4 Eksborg S. Evaluation of method-comparison data. *Clinical Chemistry* 1981; **27**: 1311–12.
5 Linnet K, Bruunshuus I. HPLC with enzymatic detection as a candidate reference method for serum creatinine. *Clinical Chemistry* 1991; **37**: 1669–75.
6 Altman DG. Calculating age-related reference centiles using absolute residuals. *Statistics in Medicine* 1993; **12**: 917–24.
7 Marshall GN, Hays RD, Nicholas R. Evaluating agreement between clinical assessment methods. *International Journal of Methods in Psychiatric Research* 1994; **4**: 249–57.
8 Lucas A, Hudson GJ, Simpson P, Cole TJ, Baker BA. An automated enzymic micromethod for the measurement of fat in human milk. *Journal of Dairy Research* 1987; **54**: 487–92.
9 Bowling LS, Sageman WS, O'Connor SM, Cole R, Amundson DE. Lack of agreement between measurement of ejection fraction by impedance cardiography versus radionuclide ventriculography. *Critical Care Medicine* 1993; **21**: 1523–27
10 Searle SR, Cassela G, McCulloch CE. *Variance components*. New York: New York, 1992.
11 Burdick RK, Graybill FA. *Confidence intervals on variance components*. New York: Dekker, 1992.
12 O'Brien E, Petrie J, Littler W, de Swiet M, Padfield PL, Altman DG, Bland M, Coats A, Atkins N. The British Hypertension Society protocol for the evaluation of blood pressure measuring devices. *Journal of Hypertension* 1993; **11**(suppl. 2): S43–S62.
13 Polk BF, Rosner B, Feudo F, Vandenburgh M. An evaluation of the Vita-Stat automatic blood pressure measuring device. *Hypertension* 1980; **2**: 221–27.
14 Latis GO, Simionato L, Ferraris G. Clinical assessment of gestational age in the newborn infant. *Early Human Development* 1981; **5**: 29–37.
15 Altman DG, Bland JM. Comparing methods of measurement. *Applied Statistics* 1987; **36**: 224–25.
16 Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in Biology and Medicine* 1991; **20**: 337–40.
17 Schoolman HM, Becktel JM, Best WR, Johnson AF. Statistics in medical research: principles versus practices. *Journal of Laboratory and Clinical Medicine* 1968; **71**: 357–67.
18 Westgard JO, Hunt MR. Use and interpretation of common statistical tests in method-comparison studies. *Clinical Chemistry* 1973; **19**: 49–57.
19 Feinstein AR. Clinical biostatistics. XXXVII. Demeaned errors, confidence games, nonplussed minuses, inefficient coefficients, and other statistical disruptions of scientific communication. *Clinical Pharmacology & Therapeutics* 1976; **20**: 617–31.
20 Altman DG, Bland JM. The analysis of blood pressure data. In O'Brien E, O'Malley K eds. *Blood pressure measurement*. Amsterdam: Elsevier, 1991: 287–314.
21 Cotes PM, Doré CJ, Liu Yin JA, Lewis SM, Messinezy M, Pearson TC, Reid C. Determination of serum immunoreactive erythropoietin in the investigation of erythrocytosis. *New England Journal of Medicine* 1986; **315**: 283–87.