Reproducibility of non-specific bronchial challenge in adults: implications for design, analysis and interpretation of clinical and epidemiological studies

S Chinn, J P Schouten

Background: Poor reproducibility of an outcome measure reduces power and, in an independent variable, biases results. The intraclass correlation coefficient measures loss of power and degree of bias. Information is lacking on the intraclass correlation coefficient for bronchial responsiveness and factors affecting reproducibility.

Methods: Papers containing information on reproducibility of bronchial responsiveness were identified using a Medline search and citations. Within and between person components of variance of PD_{20} or PC_{20} were expressed in doubling dose or concentration units, and the intraclass correlation coefficient calculated when not reported.

Results: Results were extracted from 32 papers. Intraclass correlation coefficients were over 0.9 in short term studies of highly selected asthmatic patients, but larger and most long term studies had lower intraclass correlation coefficients, less than 0.5 in some cases, due to greater within person or lower between person variation. Reproducibility of dose or concentration-response slope was generally higher, but still less than that of forced expiratory volume in 1 second.

Conclusions: Information is available to calculate sample size for studies with bronchial responsiveness as the outcome, but results when bronchial responsiveness is an explanatory variable may be misleading.
limited to histamine and methacholine—that is, a small number of papers on reproducibility of BHR to carbachol, cold air, exercise or hypertonic saline were excluded. Measurements repeated on the same day were not included, and where methods of administration were compared in the same subjects only the preferred method was included.

Repeatability data were extracted and within and between subject components of variation were expressed in doubling dose or concentration standard deviations. Papers which did not report the within subject standard deviation or the ICC or allow either to be estimated were omitted. Unless otherwise stated, a published 95% range for a single value was assumed to be calculated as ±2 within subject standard deviations. Where limits were stated to be a “confidence interval”, statistics were derived only if it was clear from the text whether the limits were calculated from a standard deviation (that is, a 95% range) or from a standard error (that is, a true confidence interval). When data were presented only graphically they were measured from the graph, taking account of differing scales on the axes where necessary. Raw data were used if given and analysed by one way analysis of variance of dose or concentration in doubling dose units by subject, and components of variance calculated,7 from which the ICC was derived. Based on the distribution of length of follow up in the papers, an arbitrary division into short term and long term follow up was made at a cut off of 4 months.

RESULTS

The Medline search produced 101 abstracts, of which 37 potentially met the inclusion criteria and 23 were found to have useful repeatability data.4 10–31 Of the 14 exclusions, eight were found not to meet the inclusion criteria on reading the full paper, one gave data for a subset of data reported in eight were found not to meet the inclusion criteria on reading the full paper, one gave data for a subset of data reported in another paper, and five did not report results in a form that allowed derivation of components of variance or the ICC. A further eight papers were identified from citations32–39 and one study primarily of other measures was included.40 Where only a measure of within person variation, or only the ICC, was stated but data were represented graphically, there was good agreement between the stated estimate and the corresponding value calculated from the measured data except in one case mentioned below.

Short term repeatability

Table 1 gives short term estimates of repeatability of PD20 or PC20 from eight studies published before 1987. These were each carried out on a small number of asthmatic patients. ICCs were above 0.9 when the within person standard deviation was less than 0.5 doubling doses, and 0.97 or more when combined with a between person standard deviation of at least 2.0 doubling doses. In one study the difference between the stated within person variation (1.0) and that derived from the graphical data (0.7) was noteworthy.38

Table 2 shows corresponding estimates from nine studies published from 1987 to 1991. Each of these gave a measure of within person variability, and most the ICC as well. Two of the studies were on population samples,13 14 but the larger study mostly comprised participants who had a measurable PD20 at the first occasion,15 and the smaller recruited participants with wheeze or asthma.16 Three studies achieved low within person variation15 16 40 but most studies had greater within person variation than the earlier studies, and hence lower ICCs. The study of hospital personnel had low between person variation and hence a low ICC. Table 3 shows estimates published from 1993 to 2001. All but one of these studies was carried out in asthmatic patients.

Long term repeatability

Long term repeatability over a period of 4 months or more was estimated in seven studies (table 4). Three of these studies were general population studies which gave lower ICCs than other studies. The largest study found an ICC for PD20 of 0.32 for asymptomatic and 0.42 for symptomatic subjects.4 New results including an extra follow up survey gave an overall ICC of 0.37, with a within person standard deviation of 1.0 doubling concentrations that was comparable to other studies, but lower between person variation. An ICC of 0.45 for PD20 was obtained by Beckett et al.29 with the largest within person standard deviation of any study. The third study, carried out in general practice, found an ICC of 0.48 for 27 subjects with complete data and measurable PD20 on six occasions, but higher ICCs (0.56 and 0.68) when all first year and all second year pairs were analysed.44 The three long term studies on selected asthmatic subjects had lower within person standard deviations and higher ICCs.10 28 37 The study of aluminium smelter workers included data only for 36 people, those with a 20% fall in FEV1 by the maximum dose of 6.14 μmol at each occasion, in the calculation of ICC for PD20.39 New results from the Vlagtwedden/Vlaardingen study showed increasing within person variation, and hence decreasing ICC, with increasing length of follow up (not shown).

Repeatability of dose-response slope

Table 5 shows estimates of short term repeatability of the FEV1-dose response slope from two studies and long term

---

### Table 1

**Short term (less than 4 months) repeatability of PD20 or PC20 in early studies on asthmatic patients and normal controls. Measurements in duplicate except where otherwise stated.**

<table>
<thead>
<tr>
<th>Year and source</th>
<th>No of participants</th>
<th>Age range (years)</th>
<th>Provocation agent and maximum concentration/cumulative dose</th>
<th>Time interval</th>
<th>Summary statistic</th>
<th>Within person SD (doubling doses or concentrations)</th>
<th>Between person SD (doubling doses or concentrations)</th>
<th>Intraclass correlation coefficient (ICC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978*</td>
<td>11 asthmatic, 3 normal</td>
<td>Unknown</td>
<td>Histamine, 16 mg/ml</td>
<td>Within 1 week</td>
<td>PC20 FEV1</td>
<td>0.2*</td>
<td>2.5*</td>
<td>0.994</td>
</tr>
<tr>
<td>1978*</td>
<td>11 asthmatic, 2 normal</td>
<td>Unknown</td>
<td>Methacholine, 16 mg/ml</td>
<td>Within 1 week</td>
<td>PC20 FEV1</td>
<td>0.3*</td>
<td>2.6*</td>
<td>0.990</td>
</tr>
<tr>
<td>1981*</td>
<td>10 asthmatic</td>
<td>16–65</td>
<td>Histamine, 16 mg/ml</td>
<td>Within 1 week</td>
<td>PC20 FEV1</td>
<td>0.3</td>
<td>2.0</td>
<td>0.97</td>
</tr>
<tr>
<td>1981*</td>
<td>12 asthmatic, studied</td>
<td>25–63</td>
<td>Histamine, 11.5 mg/ml</td>
<td>1–12 days</td>
<td>PC20 FEV1</td>
<td>0.4</td>
<td>1.5</td>
<td>0.94</td>
</tr>
<tr>
<td>1981*</td>
<td>6 times</td>
<td>Unknown</td>
<td>Histamine, 7.8 μmol</td>
<td>Within 10 days</td>
<td>PD20 FEV1</td>
<td>0.4*</td>
<td>1.5*</td>
<td>0.93*</td>
</tr>
<tr>
<td>1981*</td>
<td>18 asthmatic</td>
<td>19–55</td>
<td>Histamine, 8 mg/ml</td>
<td>Within 5 days</td>
<td>PD20 FEV1</td>
<td>0.3*</td>
<td>2.0*</td>
<td>0.98*</td>
</tr>
<tr>
<td>1984*</td>
<td>18 asthmatic</td>
<td>19–55</td>
<td>Histamine, 32 mg/ml</td>
<td>Within 2 weeks</td>
<td>PD20 FEV1</td>
<td>0.8</td>
<td>2.2, calculated from ICC, and within person SD</td>
<td>0.88</td>
</tr>
<tr>
<td>1985*</td>
<td>27 mixed</td>
<td>17–49</td>
<td>Histamine, 16 mg/ml</td>
<td>65 days</td>
<td>PC20 FEV1</td>
<td>1.0</td>
<td>2.3</td>
<td>0.91*</td>
</tr>
<tr>
<td>1986*</td>
<td>24 asthmatic</td>
<td>18–55</td>
<td>Histamine, 7.8 μmol</td>
<td>1–7 days</td>
<td>PD20 FEV1</td>
<td>1.1</td>
<td>1.7</td>
<td>0.72*</td>
</tr>
</tbody>
</table>

*Estimated from graphical information. SD, standard deviation. 

www.thoraxjnl.com
ICC from four. Dose or concentration-response slope was calculated from two data points, except for one study which used regression of percentage decline in FEV₁ on dose. This study reported data from 104 participants which included 90 whose repeatability of PD₂₀ was given in table 2. The latter only included people with a measurable PD₂₀ on at least one occasion, while the FEV₁-dose response slope was calculated for each participant who received two or more doses of histamine. The ICCs were 0.89 for slope and 0.81 for PD₂₀. The study of aluminum smelter workers, which included data only for persons with two measurable PD₂₀ values in the PD₂₀-ICC, found a much higher ICC for log dose-response slope (0.73 compared with 0.28). Trigg et al found a higher ICC for the dose-response slope (0.75) than for PD₂₀ (0.48), and Beckett et al slightly higher (0.54 compared with 0.45).

**DISCUSSION**

**Variation in ICC**

The early studies on short term repeatability in selected asthmatic patients achieved good repeatability, as indicated by the within person standard deviation of less than 0.5 doubling doses or concentrations and hence a high ICC. Early enthusiastic exponents of bronchial challenge may have taken greater care over procedures or selected highly cooperative patients. Many later studies, particularly the larger population studies, had a within person standard deviation of around 1.0 doubling doses or concentrations. Selection of subjects determines the between subject variation. A population study has a large majority of “non-responsive” individuals whose values are clustered at the maximum dose or concentration; even when this is high, the use of a logarithmic scale reduces the apparent variation at the upper end of the scale.

Variation is expressed on the doubling dose or concentration scale as this is the most appropriate for PD₂₀ or PC₂₀ but ICCs are independent of linear transformation—that is, they are the same on any logarithmic scale. Repeatability of histamine and methacholine BHR appears similar on a logarithmic scale. There is no agreement over scale for the dose-response slope (table 5) but correlation with log PD₂₀ has been shown to be high when the dose-response slope is reciprocally transformed or log-transformed.

The Pearson correlation coefficient measures the degree of any linear relation between two variables. It is therefore inappropriate for repeatability studies as, for example, change in mean BHR over time would not affect it but does lower the ICC. The unsuitability of the Pearson correlation coefficient for method comparison and repeatability was made clear in 1986 but, despite this, several later papers reported it, although within person variation was generally also reported.

There were too few long term studies with fixed follow up time to relate within person variation to length of follow up. Unpublished results from the Vlagtwedden/Vlaarding study suggest an increase in within person variation with length of follow up. It is unclear whether the lower ICCs in

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Short term (less than 4 months) repeatability of PD₂₀ or PC₂₀ in studies published from 1987 to 1991</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year and source</td>
<td>Participants</td>
</tr>
<tr>
<td>1987&lt;sup&gt;23&lt;/sup&gt;</td>
<td>20 asthmatic patients</td>
</tr>
<tr>
<td>1988&lt;sup&gt;24&lt;/sup&gt;</td>
<td>27 with wheeze or asthma, population sample</td>
</tr>
<tr>
<td>1988&lt;sup&gt;24&lt;/sup&gt;</td>
<td>20 asthmatics</td>
</tr>
<tr>
<td>1991&lt;sup&gt;29&lt;/sup&gt;</td>
<td>14 healthy, “responsive”</td>
</tr>
<tr>
<td>1991&lt;sup&gt;29&lt;/sup&gt;</td>
<td>20 asthmatic, inexperienced</td>
</tr>
</tbody>
</table>

*Estimated from graphical information. SD, standard deviation.
In carefully controlled studies with selected participants an ICC of 0.99 can be achieved, as high as that for FEV₁. However, such a high ICC is unlikely to be achieved in larger studies. In the studies which assessed repeatability of FEV₁ and BHR in the same subjects, the ICC for BHR was lower than that for FEV₁, so that studies of BHR generally require more participants than those on FEV₁ to detect an equivalent size of effect. The standard deviation that should be used in a sample size calculation is the total short term variation in a study with similar participants; this can be calculated by adding the squares of the within and between standard deviation and taking the square root of the result.

### Change in BHR as outcome in short term follow up studies

The standard deviation of change in any continuous outcome is calculated by multiplying the within person standard deviation by the square root of two. In randomised controlled trials the recommended analysis is of final outcome with the baseline value as a covariate, as this increases power and is unbiased as baseline mean values will be equal on average. However, this method is inadvisable in an observational study. The regression coefficient of final on initial value is biased towards zero. It is used to adjust the estimated means at follow up of groups that differ in mean initial value, and so will affect the comparison of interest.

### BHR as outcome variable in a cross sectional study

In such a study, the ICC is calculated by multiplying the between person standard deviation by the square root of two. Appropriate within person standard deviations in tables 1–3 can therefore be used to calculate sample size or power. In randomised controlled trials the recommended analysis is of final outcome with baseline as a covariate, as this increases power and is unbiased as baseline mean values will be equal on average. However, this method is inadvisable in an observational study. The regression coefficient of final on initial value is biased towards zero. It is used to adjust the estimated means at follow up of groups that differ in mean value at baseline and so will affect the comparison of interest and can even reverse the sign of the difference. In addition, Schouten and Tager have explained why adjusting for baseline may give misleading results. The analysis of final outcome with baseline as covariate has little to recommend it.

### Table 4 Long term (more than 4 months) repeatability of BHR

<table>
<thead>
<tr>
<th>Year and source</th>
<th>Participants</th>
<th>Age range (years)</th>
<th>Provocation agent and maximum concentration/cumulative dose</th>
<th>Time interval</th>
<th>Summary statistic</th>
<th>Within person SD (doubling doses or concentrations)</th>
<th>Between person SD (doubling doses or concentrations)</th>
<th>Intraclass correlation coefficient (ICC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990⁷</td>
<td>35 asthmatic</td>
<td>Not stated</td>
<td>Histamine, 1.6 mg/ml</td>
<td>10–30 months</td>
<td>PC₂₀/FEV₁</td>
<td>1.0⁷</td>
<td>1.7⁷</td>
<td>0.74⁷</td>
</tr>
<tr>
<td>1993³</td>
<td>19 healthy ''non-responsive'' workers</td>
<td>15–54 at baseline</td>
<td>Histamine, 32 mg/ml</td>
<td>3–22 years</td>
<td>PC₂₀/FEV₁</td>
<td>0.8</td>
<td>1.4</td>
<td>0.87</td>
</tr>
<tr>
<td>1994³</td>
<td>2173 population sample</td>
<td>15–54 at baseline</td>
<td>Histamine, 32 mg/ml</td>
<td>3–22 years</td>
<td>PC₂₀/FEV₁</td>
<td>1.0</td>
<td>0.8</td>
<td>0.37</td>
</tr>
<tr>
<td>1997³</td>
<td>10 asthmatic, PC₂₀&lt;9 mg/ml</td>
<td>25–82</td>
<td>Histamine, 16 mg/ml</td>
<td>6 months</td>
<td>PC₂₀/FEV₁</td>
<td>0.7</td>
<td>1.4</td>
<td>0.80</td>
</tr>
<tr>
<td>1997³</td>
<td>88 healthy working adults</td>
<td>Not stated</td>
<td>Methacholine, 13.06 μmol</td>
<td>1–3 years</td>
<td>PC₂₀/FEV₁</td>
<td>1.8</td>
<td>1.6</td>
<td>0.45</td>
</tr>
</tbody>
</table>

*Estimated from graphical information.

SD, standard deviation.

V/V unpublished results from Vlagtwedden/Vlaardingen study; published results also included.

---

### Table 5 Intraclass correlation coefficients for measures of FEV₁-dose-response slope

<table>
<thead>
<tr>
<th>Year and source</th>
<th>Participants</th>
<th>Age range (years)</th>
<th>Provocation agent and maximum concentration/cumulative dose</th>
<th>Time interval</th>
<th>Summary statistic</th>
<th>Intraclass correlation coefficient (ICC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990³</td>
<td>726 workers</td>
<td>Not stated</td>
<td>Methacholine, 6.14 μmol</td>
<td>1 year</td>
<td>Log dose-response slope</td>
<td>0.73</td>
</tr>
<tr>
<td>1991³</td>
<td>16 healthy ''non-responsive''</td>
<td>Mean (SD) 22.2 [3.2]</td>
<td>Methacholine, 8.5 mg/ml</td>
<td>1–14 days</td>
<td>Log dose-response slope</td>
<td>0.30</td>
</tr>
<tr>
<td>1993³</td>
<td>104 population sample</td>
<td>18–64</td>
<td>Methacholine, 16 mg/ml</td>
<td>1–14 days</td>
<td>Log dose-response slope</td>
<td>0.30</td>
</tr>
<tr>
<td>1994³</td>
<td>67 general practice on 6 occasions</td>
<td>18–75</td>
<td>Methacholine, 247 μmol</td>
<td>1/4 dose response slope</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>1997³</td>
<td>88 healthy working adults</td>
<td>Mean (SD) 27.4 [3.4]</td>
<td>Methacholine, 13.06 μmol (in repeatability data)</td>
<td>1–3 years</td>
<td>Log dose-response slope</td>
<td>0.97</td>
</tr>
</tbody>
</table>

FEV₁, forced expiratory volume in 1 second.
in non-randomised studies and, for an outcome with an ICC that may be as low as 0.5 in some circumstances, it is definitely to be avoided in such studies.

**Change in BHR as outcome in long term follow up studies**

It is likely that the change in BHR over several months or years will be more variable than in the short term. Although the number of studies is small with most of the information from population based studies, lower ICCs are unlikely to be due wholly to differences in participants. Firstly, the short term population study found a relatively high ICC due to low within person variation and, secondly, the large long term study found even lower ICCs on adjustment for individual explanatory variables, as between person variation was reduced proportionally more than within person variation. The within person standard deviations in table 4 can be used in sample size calculations, although they will be conservative as some of the within person variation will be explained by changes in explanatory variables. On the other hand, the use of standard deviations in tables 1–3 may result in too small a sample size. The recommendation to analyse absolute change, and not final adjusted for initial value, applies even more strongly to long term than to short term observational studies.

**BHR as an independent variable**

A number of authors have used BHR as an independent variable, particularly as a predictor of decline in FEV1, dividing participants into “responders” and “non-responders”. Few authors have reported a kappa statistic for repeatability of dichotomised BHR, but it can be expected to be similar in value to the ICC. BHR has a unimodal continuous distribution in the general population and is not a fixed state, as many authors seem to assume. The problem—whether BHR is dichotomised or not—is the same as that of using baseline BHR as a covariate when final BHR is the outcome in a longitudinal study, that there will be bias in the regression coefficient of outcome on BHR and also of the other regression coefficients in a multiple regression. Correction for bias requires estimates of the ICC for variances and covariances of the explanatory variables which can only be determined from a repeatability study of all covariates subject to within person variability carried out on all, or a substantial random sample, of the participants unless certain assumptions are met.

**Conclusion**

The analysis of BHR as an outcome variable is straightforward and there is considerable information to allow studies to be planned with adequate sample size to take account of the inherent variation. PD20 or PC20 are known only to be above the maximum dose or concentration (that is, “censored”) when a 20% fall in FEV1 has not occurred when the challenge is stopped. This has often led authors to express BHR as “responsive” or “not responsive” and to use logistic regression to analyse the data, but greater power is achieved if regression methods for censored data are used or a dose-response slope or other continuous outcome analysed. This is reinforced by ICCs for the dose-response slope being at least as high and probably greater than those for PD20.

In contrast, analysis of BHR as an explanatory variable is liable to give biased and possibly misleading results. This is true of any explanatory variable for which the short term ICC may be as low as 0.5. BHR contrasts with FEV1 as the short term ICC for FEV1 can be presumed to be over 0.91 and has been reported to be 0.89 over 1–3 years. Lung function has been shown to be strongly associated with BHR in cross sectional studies, part of which may due to inherent dependence of BHR summary statistics on FEV1. Analyses of BHR as the outcome therefore need to adjust for lung function even if a causal role is not assumed.

Rijcken and Weiss posed the question of whether a lower level of FEV1 is a cause or a result of increased airway responsiveness and stated that longitudinal analyses are necessary to answer the question. We can add to this that either multiple measurements of BHR should be made to increase precision or the regression coefficients should be adjusted for lack of repeatability. Unfortunately, if ICCs of variances are highly variable, those of covariances may be even harder to estimate and extrapolation from another study is unlikely to be sound. Unless researchers take steps to increase precision, the inclusion of BHR as an explanatory variable may be misleading.

**Authors’ affiliations**

S Chinn, Department of Public Health Sciences, King’s College London, London SE1 3QD, UK

J P Schouten, Department of Epidemiology and Bioinformatics, University Medical Center, 9700 RB Groningen, The Netherlands

No competing interests declared.

**REFERENCES**


37 Loon O
35 Cockcroft DW
29 Hedman J
24 Seppälä O-P
23 Trigg CJ
36 Ruffin RE
33 Ryan G
30 Beckett WS
32 Juniper EF

Reproducibility of non-specific bronchial challenge in adults: implications for design, analysis and interpretation of clinical and epidemiological studies
S Chinn and J P Schouten

Thorax 2005 60: 395-400
doi: 10.1136/thx.2004.039230

Updated information and services can be found at: http://thorax.bmj.com/content/60/5/395

These include:

References
This article cites 47 articles, 19 of which you can access for free at: http://thorax.bmj.com/content/60/5/395#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections
- Airway biology (1100)
- Asthma (1782)
- Epidemiologic studies (1829)
- Internet (104)
- Lung function (773)

Notes

To request permissions go to: http://group.bmj.com/group/rights-licensing/permissions
To order reprints go to: http://journals.bmj.com/cgi/reprintform
To subscribe to BMJ go to: http://group.bmj.com/subscribe/