The reproducibility and sample size requirements of exercise-induced bronchoconstriction measurements

B. Dahlén*, P.M. O’Byrne#, R.M. Watson*, A. Roquet*, F. Larsen*, M.D. Inman#


ABSTRACT: Dry air exercise challenges are frequently used to screen medications that have potential utility in the management of exercise-induced bronchoconstriction (EIB). The purpose of this study was to determine the reproducibility of three outcome measurements made using such challenges, and sample size requirements for drug evaluation studies based on these outcomes.

Forty adult subjects with asthma, who tested positively on a screening exercise challenge, were subjected to two further identical challenges, separated by 1 to >35 days. Outcome measurements included the maximum per cent fall in forced expiratory volume in one second (FEV1), after exercise (% fallmax), and the area under the per cent fall in FEV1/time curve for 30 min (AUC30) and 60 min (AUC60) after exercise.

The reproducibility of these outcomes, as assessed by intraclass correlation coefficients was 0.72, 0.53 and 0.35 for % fallmax, AUC30 and AUC60 measurements, respectively. The sample size requirements to demonstrate an attenuation of EIB equivalent to a 50% reduction in % fallmax was 9, 14 and 19 subjects for the % fallmax, AUC30 and AUC60 responses, respectively (90% power).

It is concluded that the maximum percentage fall in forced expiratory volume in one second has greater reproducibility and results in greater power in clinical trials than area under the curve measurements. Sample size calculation curves are provided which may be used in study design and interpretation of published studies.


Exercise-induced bronchoconstriction (EIB) occurs to some degree in 70–80% of asthmatic patients. This bronchoconstriction is thought to be a manifestation of airway hyperresponsiveness and as such, does not cause an inflammatory response or worsening of the underlying asthma. Thus, the goal in the asthmatic patient should be to encourage exercise, while providing optimal anti-inflammatory treatment of the underlying asthma and any additional treatment required to minimise symptoms associated with exercise.

While treatment with inhaled steroids can reduce the magnitude of EIB by 50% or more [1–5], treatment with additional drugs is usually required to eliminate all symptoms. Inhalation of short and long-acting β2-agonists are effective in protecting against, or reversing EIB [6–8], but this effect can be reduced following periods of regularise [9–12]. More specific agents have been shown to provide partial attenuation of the response, including anti leukotrienes, anticholinergics and antihistamines [13–15]. Clearly, further studies are required in this area to identify agents, or more likely combinations that can be used on a regular basis to completely prevent asthmatic symptoms associated with exercise.

In many studies where the efficacy of treatment on the magnitude of EIB has been evaluated, the outcome measurement has been the maximum per cent fall in forced expiratory volume in one second (FEV1) after exercise (% fallmax) [3, 14, 16–19]. Occasionally, additional analyses are performed on the area under the per cent fall in FEV1/time curve (AUC) [13, 15, 17, 20, 21]. Surprisingly, there is little information available concerning the reproducibility of either of these outcome measurements, or on sample size requirements for assessing and comparing agents in their ability to protect against EIB. Furthermore, to the authors knowledge, there have been no reports on the clinical relationship between these outcome variables. Thus, there is no basis for assuming that a given per cent attenuation of the per cent fall in FEV1 is equivalent to the same per cent attenuation of the AUC.

The purpose of this study was to determine the reproducibility of both the % fallmax as well as the AUC for 30 (AUC30) and 60 min (AUC60) after exercise. A second aim was to provide guidelines for sample size determination in exercise challenge studies of treatment efficacy, similar to those provided previously for allergen and methacholine challenges [22, 23]. Finally, the per cent fall and AUC measurements were compared in terms of clinical equivalence.

Methods

Subjects

Forty asthmatic subjects (25 male, 15 female; table 1) with EIB participated voluntarily in the study, which
was approved by the Research Ethics Committees of The Karolinska Hospital and McMaster University Health Sciences Centre. Each subject gave written informed consent before taking part. All subjects were nonsmokers and had stable asthma controlled by short-acting inhaled β₂-agonists alone. Three subjects used a stable dose of inhaled budesonide daily (200, 400 and 800 μg, respectively) in addition to their β₂-agonist use. All subjects had FEV₁ >70% predicted [24]. Atopic subjects were not studied during periods associated with seasonal environmental allergen exposure. Subjects were not knowingly exposed to other environmental allergens (except house dust mite) for ≥ 2 weeks prior to any study visits.

Study design

The subjects attended the laboratory for ≥ 2 screening sessions and then on 2 days, separated by a period of 1–21 days (33 subjects) or ≥ 35 days (7 subjects). Subjects were instructed not to use bronchodilating drugs for ≥ 8 h and caffeine for ≥ 24 h prior to all laboratory visits. Exercise challenges performed were administered at the same time of day (+/- 30 min) for each subject.

Methods

During the initial screening visit, subject characteristics and history were documented, and an incremental cycle ergometer exercise test was performed until subjective exhaustion [25]. During the second screening visit, a dry air exercise challenge (described later) was profaned at a work rate equal to 80% of the maximum achieved during the incremental test. If this resulted in a fall in FEV₁ from baseline of 15–45%, then this work rate was chosen for subsequent exercise challenges. If the per cent fall in FEV₁ was outside this range, then further screening challenges were performed at appropriately lesser or greater work rates. At least 24 h separated all screening visits.

Subjects in whom a fall in FEV₁ of ≥ 15% was measured during a screening exercise challenge then performed on separate days, two further exercise challenges at the same work rate as that used during screening.

Dry-air exercise challenge. Subjects exercised on a stationary cycle ergometer (Karolinska site: SECA Cardiotest 100, Vogel and Halke, Hamburg, Germany. McMaster site: Ergomed 740, Siemens Mississauga, Canada) for 5 min at a constant work rate. Subjects wore nose clips and breathed dry room air (<10% relative humidity) at room temperature (21.5 °C) from a Douglas bag reservoir connected via the inspiratory port of a 3-way Hans-Rudolph (model 2700, Kansas City, MO, USA) valve to a mouthpiece. The FEV₁ was measured (Karolinska site: Vitalograph MDI Compact, Förbundsmaterial, Stockholm, Sweden. McMaster site: Collins 10 L water spirometer, Collins Inc., Braintree, MA, USA) immediately prior to and immediately following exercise, as well as at 1, 3, 5, 8, 10, 15, 20, 30, 40, 50 and 60 min postexercise. The pre-exercise (baseline) FEV₁ was established as the greatest of three measurements, separated by ≥ 30 s from each other.

Analysis

The bronchoconstrictor response following exercise was measured as % fallmax, AUC30 and AUC60. Specifically, % fallmax represented the lowest FEV₁ measured in the hour following exercise, expressed as a percentage of the pre-exercise FEV₁ measurement. The AUC30 measurement was the area under the FEV₁ (expressed as a percentage of the pre-exercise value) vs. time curve (up to 30 min postexercise) calculated using linear trapezoidal integration. On occasions when FEV₁ was greater than baseline values, area was subtracted from the AUC measurement. The AUC60 was calculated similarly, but for the 60 min postexercise period.

An index of the reproducibility of responses, as measured using each of these analysis techniques was obtained using the intraclass correlation coefficient [26]. This gives a number usually varying ± 1–1, where 1 indicates perfect agreement, 0 indicates no relationship and a negative number indicates disagreement. The magnitude of the number gives the fraction of the total variability in the measurement that can be accounted for by between-subject factors.

To obtain an indication of the utility of an exercise challenge for evaluating antiasthma treatments, the sample size required to demonstrate a given percentage attenuation of the response was calculated. Sample size was calculated based on measured SD and anticipated differences in mean responses using the following equation:

$$n = \left(\frac{T_1 - \alpha/2 + T_1 \cdot \beta}{\mu_2 - \mu_1}\right)^2$$

where n is the predicted sample size, T is the t-test t-score corresponding to the desired α (probability of a type I error) and β (probability of a type 2 error), σ is SD and μ₁-μ₂ is the minimum clinically important difference that should be detected [27]. This equation is appropriate for two-tailed tests. Using an iterative process, all sample size predictions were made using t-distributions with the degrees of freedom set at a level corresponding to the resulting sample size calculation. Sample size predictions were made for % fallmax, AUC30 and AUC60 measurements. The SD entered into the equation was the difference between the measurement on the first and second challenges. Thus,
degree of EIB was assessed using ≥2 of these techniques, both under placebo and treatment conditions [10, 13, 15, 21, 28]. The resulting relationships between per cent attenuation of the % fallmax measurement and per cent attenuation of both the AUC30 and AUC60 techniques are shown in figure 1. The per cent attenuation was calculated as the percentage difference between an outcome under placebo and treatment conditions. These data sets were analysed using linear regression, resulting in the two following equations:

\[ \text{attenuation (\% fallmax)} = 0.99 \times \text{attenuation(AUC30)} - 13.33 \ (r^2 = 0.96) \] (2)

\[ \text{attenuation (\% fallmax)} = 0.99 \times \text{attenuation(AUC60)} - 25.77 \ (r^2 = 0.95) \] (3)

These equations were used to compare sample size calculations for the three analysis techniques in terms of clinically equivalent degrees of attenuation.

Results

Baseline FEV₁ was not systematically different between the first and second challenges (3.42±0.10 L versus 3.40±0.10 L, respectively; \(p>0.05\)). The mean absolute difference between the baseline FEV₁ values on the two challenge days was 4.25±0.70%.

The mean magnitudes of the % fallmax, AUC30, and AUC60 responses are included in table 2. The intra-class correlation coefficient for the % fallmax measurement was 0.72, for the AUC30 measurement was 0.59 and for the AUC60 measurement was 0.35 (fig. 2). Thus, within-subject variability accounted for 28%, 41% and 65% of the total variance in the % fallmax, AUC30, and AUC60 measurements, respectively. Also in figure 2, the difference between the first and second measurements has been plotted against the mean of the two measurements for % fallmax, AUC30 and AUC60. These plots illustrate that the magnitude of the difference between the first and second measurements remains essentially the same over the range of measurements made from these subjects.

To determine whether the degree of reproducibility was affected by the time interval allowed between the challenges, the absolute value of the difference between the % fallmax measurement on the first and second challenges was plotted against the interval (fig. 3). The difference in this measurement between the first and second challenge did not increase as the time interval

<table>
<thead>
<tr>
<th>Challenge</th>
<th>% fallmax</th>
<th>AUC30</th>
<th>AUC60</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean±SD</td>
<td>23.3±111.9</td>
<td>21.4±12.5</td>
<td>-1.9±9.0</td>
</tr>
<tr>
<td></td>
<td>384±257</td>
<td>354±281</td>
<td>-30±244</td>
</tr>
<tr>
<td></td>
<td>561±453</td>
<td>557±477</td>
<td>-4±531</td>
</tr>
</tbody>
</table>

Table 2. – Data for maximum per cent fall in forced expiratory volume in one second (FEV₁) after exercise (% fallmax), and area under the per cent fall in FEV₁/time curve for 30 (AUC30) and 60 mins (AUC60), during two challenges, and the difference between the two (\(\Delta\))
increased from 1–>35 days; in fact it appeared to decrease.

The estimated sample size requirements to illustrate statistically significant attenuation of the % fallmax, AUC30 and AUC60 responses, based on the mean of the two responses and the sd of the delta values in table 2 are illustrated in figure 4. Thus, the curves in figure 4a are calculated, based on attenuation of a 22.3% fallmax with sd of the difference between the treated and untreated responses of 9.02% fallmax units.

Included in these figures is an illustration of the estimated sample size requirements to demonstrate, with 90% power, a 50% attenuation of each response; this being 9, 21 and 40 subjects for the % fallmax, AUC30 and AUC60 measurements, respectively. In figure 5, separate x axes have been included for % fallmax, AUC30, and AUC60 curves and these have been aligned relative to each other, according to the relationships in equations (2) and (3), so that there is a clinically equivalent degree of attenuation on all three curves at
any point where a vertical line is placed. After making this correction, the number of subjects required to demonstrate an attenuation equivalent to a 50% reduction in % fallmax was 9, 14 and 19 subjects using the % fallmax, AUC30, and AUC60, measurements, respectively.

Discussion

In this study, the reproducibility and power of three techniques for quantifying the degree of exercise-induced bronchoconstriction has been compared. The greatest reproducibility has been observed for the maximum % fall in FEV1 and the least reproducibility for the area under the 60 min FEV1/time curve. Furthermore, the power of these techniques to detect the attenuating effects of treatment are such that small degrees of attenuation are more likely to be detected using the % fallmax technique with the same sample size.

The intraclass correlation coefficient (ICC) for the AUC30 technique measured in this study was 0.59, and was similar to the value of 0.67 observed by HOFSTRA et al. [29]. The observation of a higher ICC for the % fallmax technique (0.72) differs from HOFSTRA et al. [29], who observed a lower ICC for this outcome (0.57). Generally, however, the two studies are in agreement that less than half of the variability of the EIB response is within-subject or error variance. The present observations also suggest that this reproducibility extends to time intervals between challenges of greater than 35 days.

Fig. 3. – The absolute value of the difference between the two maximum per cent fall in forced expiratory volume in one second (FEV1) after exercise (% fallmax) measurements, plotted against the time interval between the two challenges.

Fig. 4. – Sample size requirements to demonstrate a range of expected attenuations of a) maximum per cent fall in forced expiratory volume in one second (FEV1) after exercise (% fallmax); b) area under the per cent fall in FEV1/time curve (AUC) for 30 min after exercise (AUC30); c) and for 60 min after exercise (AUC60) response with 90% power. The x axes are calibrated and positioned based on equations (2) and (3), so that the degree of attenuation of each response is clinically equivalent at any point.

Fig. 5. – Sample size requirements to demonstrate a range of expected attenuations of maximum per cent fall in forced expiratory volume in one second (FEV1) after exercise (% fallmax; --); area under the per cent fall in FEV1/time curve (AUC) for 30 min after exercise (AUC30; - - -) and for 60 min after exercise (AUC60; - - - -) response with 90% power. The x axes are calibrated and positioned based on equations (2) and (3), so that the degree of attenuation of each response is clinically equivalent at any point.

Abs (% fall max1 - % fall max2)

Fig. 3. – The absolute value of the difference between the two maximum per cent fall in forced expiratory volume in one second (FEV1) after exercise (% fallmax) measurements, plotted against the time interval between the two challenges.
Although it has been observed that there is good reproducibility of the % fallmax measurement, the authors argue that this is not sufficient to infer a high degree of utility in all research applications. In crossover studies where the change in a measurement is the important outcome variable, it is the sample size, variability in the change, and clinically important magnitude of the change that determine the power of the study. For this reason the sd of the change in % fallmax, AUC30, and AUC60 between the first and second challenges have been measured and used, along with a range of degrees of attenuation and desired power levels, to predict sample sizes that should be used in clinical trials. The authors feel that this information gives much more information on the utility of each technique than the ICC does.

It has been estimated that a sample size of ≥ 9 subjects should be sufficient to demonstrate, with 90% power, a 50% attenuation in the % fallmax measurement in response to treatment. It must be stressed that if a degree of attenuation of > or <50% was thought to be the minimal clinically important difference, then correspondingly less or more subjects would be required in the study design, as indicated in figure 4. The present estimation of nine subjects being required to demonstrate 50% attenuation of the response is greater than the sample size of six estimated to be required to demonstrate the same effect in children as reported by Hofstra et al. [29]. In that study, sample sizes were estimated using a Z distribution, while in the present study, a t-distribution has been used, given that sds were estimated from a relatively small sample of the population (n=40). If a Z-distribution had been used, then a sample size of 7 would have been estimated, similar to that of Hofstra et al. [29]. Unlike Hofstra et al. [29], the present results suggest that greater sample sizes would be required to demonstrate a similar attenuation of the AUC30 and AUC60 outcomes. Reasons for the disagreement between the present findings and those of Hofstra et al. [29] are not clear, but may be due to their calculations being based on paediatric subjects and a treadmill based protocol.

By reviewing studies where both % fallmax and AUC30 or AUC60 measurements were made, estimates of equations 1 and 2 were made, with which an attempt has been made to convert attenuation of % fallmax measurements to clinically equivalent attenuation of AUC30 and AUC60 measurements. For both equations, it is clear that a given degree of attenuation of % fallmax (when the treated response is expressed as a percentage of untreated response) is equivalent to greater attenuation of AUC measurements. Even after applying these equations, it was observed that clinically equivalent degrees of attenuation would be detected with smaller sample sizes using the % fallmax measurement.

While sample size requirements for crossover designed studies have been presented, calculations can be made to determine sample sizes for parallel group studies. Equation 1 would still apply, but σ should be the sd of the measurement itself, rather than the sd of the difference between two repeated measurements. This equation could then be used to calculate the subjects required in each treatment arm. Performing this calculation to determine the number of subjects required to demonstrate a 50% attenuation of % fallmax these equations calculated that 15 would be required in each treatment arm (assuming an untreated response of 22.32% fall and an sd of the treated and untreated responses of 12.2% fall units). This number is similar to the recommendation of 12 subjects in each treatment arm made by Hofstra et al. [29] in their study of asthmatic children. This calculation assumes that the outcome measurement in the parallel group design would be the response to a single exercise challenge. If, in fact, the outcome was the change in the response from pre- to post-treatment conditions, then the sd of the difference between two challenges (9.02 % fallmax units as measured here) should be used to estimate sample size. In this case, the curves in figure 3 could be used to estimate the sample size requirements for each arm of the study.

In this study, it has been recommended that fewer than 10 subjects are required to demonstrate a 50% attenuation of % fallmax. Clearly, more subjects are required to illustrate a smaller degree of attenuation. Furthermore, more subjects may be required to demonstrate that the protective effects of two treatments are different. For example, if one drug blocked 75% of the % fallmax and another drug blocked 50% of the % fallmax, both of these effects should be detected at least 90% of the time with nine subjects. However, the difference in the degree of attenuation between the two drugs is only 25% (75%-50%), a difference that would require 30 subjects to demonstrate using the % fallmax outcome (fig. 3). Thus, when designing studies to compare two drugs, it is important to decide what difference between the two drugs is thought to be clinically relevant. This difference should then be used in equation 1 or plotted on figure 3a to calculate the appropriate sample size.

A further use of these sample size requirements is to evaluate the power of published studies. Usually, studies of the efficacy of various drugs in the prevention of EIB have ≥10 subjects [10, 14, 16, 20, 30–33] and are thus adequately powered, assuming that a degree of protection equivalent to <40–50% attenuation of the % fallmax is considered not to be clinically important. Often, studies are performed to compare two drugs or doses with respect to their ability to attenuate EIB. Frequently these studies contained no more subjects than studies evaluating single compounds [13, 18, 19, 21, 34] and were therefore probably underpowered to detect clinically important differences between drugs or doses.

The mean % fallmax in this study was ~22%. In some studies a greater fall in FEV1 is achieved under placebo treatment conditions. It is likely that the sample sizes estimated in this study will overestimate those required in studies where the untreated % fallmax is larger than that reached by subjects in this study.

It is interesting to note that the % fallmax value did not exceed 15 in seven of the subjects in response to the first exercise challenge, and 13 of the subjects in response to the second exercise challenge. There were six subjects for whom the % fallmax value was <15 on both exercise challenges. While it could be argued that these subjects represented screening failures, the
MEASURING EXERCISE-INDUCED BRONCHOCONSTRICTION

587

authors elected to include them in the analysis, as these subjects, having met screening criteria, would be included in clinical trials for which the sample size estimates will be used.

In summary, it has been shown that the maximum per cent fall of forced expiratory volume in one second after exercise has greater reproducibility than area under the curve measurements. More importantly, information on sample size requirements has been provided that can be used both in the design of well powered studies and in the evaluation of already published studies. These results show that for drug efficacy and drug comparison studies, maximum per cent fall in forced expiratory volume in one second is more powerful than area under the curve measurements.

Acknowledgements. Statistical advice from C. Goldsmith was greatly appreciated in the preparation of this manuscript.

References


