Quality control of DLco instruments in global clinical trials

R. Jensen\textsuperscript{1}, M. Leyk\textsuperscript{2}, R. Crapo\textsuperscript{1}, D. Muchmore\textsuperscript{2} and PY. Berclaz\textsuperscript{2}.

\textsuperscript{1}LDS Hospital, Pulmonary Division, Salt Lake City, UT, 84143
\textsuperscript{2}Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN, USA

Corresponding Author:
Dr. Pierre Berclaz
Eli Lilly and Company,
Lilly Corporate Center
Indianapolis, IN 46285
Phone: 317.433.8973 Fax: 317.277.6351 Email: berclaz.pi@lilly.com

Short Title: DLco simulator testing in multi-centre trials
Abstract:

High inter- and intra-laboratory variability exists for the single-breath diffusing
capacity of the lung for carbon monoxide (DLco) test. To detect small changes in
diffusing capacity in multi-centre clinical trials, accurate measurements are essential.
This study assessed whether regular DLco simulator testing maintained or improved
instrument accuracy and reduced variability in multi-centre trials.

The 125 pulmonary function testing (PFT) laboratories that participated in clinical
trials for AIR® Inhaled Insulin validated and monitored the accuracy of their DLco
measuring devices by using a DLco simulator, which creates known target values for any
device. Devices measuring a simulated DLco different from target by >3 mL·min⁻¹·mmHg⁻¹ failed testing and were serviced. Device accuracy was assessed over time and
with respect to differences in several variables.

Initially, 31 (25%) laboratories had a DLco device that failed simulator testing. After fixing or replacing devices, 124 (99%) laboratories had passing devices. The
percentage of failed tests significantly decreased over time. Differences in geographical
region, device type, breath-hold time, temperature and pressure were not associated with
meaningful differences in DLco device accuracy.

Regular DLco simulator testing allows PFT laboratories to maintain the accuracy
of their DLco measurements leading to reduced variability across laboratories in multi-
centre clinical trials.

Word Count: 200

Keywords: accuracy, inhaled insulin, multi-centre trials, quality control, single-breath
diffusing capacity of the lung for carbon monoxide (DLco), simulator
Pulmonary function tests (PFTs) such as spirometry (forced expired volume in 1 second [FEV1], forced vital capacity [FVC]), lung volume measurements (total lung capacity [TLC]) and the single-breath diffusing capacity of the lung for carbon monoxide (DLco) are essential for evaluating the efficacy of new therapies for asthma, chronic obstructive pulmonary disease and other lung diseases or to assess the safety of treatments administered by the pulmonary route [1, 2]. The DLco, which is already a valuable tool for disease diagnosis [3–6], is also an important measurement used to identify subtle adverse effects of inhaled therapies on the diffusing capacity in clinical trials.

Although the utility of the DLco is well established, an issue facing investigators in large-scale, multi-centre clinical trials is the high inter- and intra-laboratory variability for this measurement [7–9]. Measuring the DLco is a complex process [10] largely automated in modern devices, and sources of inter-laboratory variability include differences in devices, test gases, software [11] and testing procedures. Within a single laboratory, the accuracy of gas or volume measurements may drift over time and daily changes in temperature and barometric pressure need to be accounted for [8, 12]. Increases in variability reduce the sensitivity of DLco measurements to detect potential harmful or beneficial effects of an inhaled product in a test population.

To control the accuracy and reproducibility of DLco data in clinical trials, processes for quality control must be established. To this end, clinical trials for AIR® Inhaled Insulin (AIR® is a registered trademark of Alkermes, Inc.) have used a centralised quality-assurance program in which each laboratory performing PFTs was qualified and individually quality-controlled throughout the study period [13]. To reduce
inter- and intra-laboratory variability due to instrumentation, PFT laboratories were required to first certify and then monitor the accuracy of their DLco devices using a simulator (Hans Rudolph, Kansas City, MO). The DLco simulator creates a precisely known, repeatable DLco value in any device. Comparing measured DLco values with the target DLco value allows the devices’ accuracy and precision to be assessed. Using the DLco simulator on a regular basis facilitates early detection of technician or device issues that could lead to test failures.

The aim of the present study was to demonstrate that a quality-assurance program that includes DLco simulator testing can establish and maintain the accuracy of DLco devices in global multi-site clinical trials thereby reducing inter- and intra-laboratory variability. The results of simulator tests performed throughout clinical trials for human insulin inhalation powder are reported. Absolute differences between measured and target values as well as trends over time in these differences were analysed. In addition, the accuracy of DLco measurements was assessed with respect to variations in room temperature, barometric pressure, breath-hold time, region, machine type and target simulator values.

METHODS:

Study design

One hundred-twenty five PFT laboratories from North America (Canada and USA), the European Union (Belgium, Bulgaria, Croatia, Germany, Hungary, Italy, Poland and Portugal), Central/South America (Argentina, Brazil, Chile, Colombia, Mexico, Puerto Rico) and Asia (India, Philippines, Singapore and Taiwan) were recruited
to clinical trials for AIR® Inhaled Insulin. DLco measurements were performed throughout the trials in study subjects as a safety evaluation.

During the trials, each laboratory used a DLco simulator to demonstrate that their device’s measurements were accurate to within 3 mL·min⁻¹·mmHg⁻¹. Before patient testing, all laboratories were required to pass an initial certification test meeting this standard. After passing the initial certification, laboratories performed simulator testing weekly for 3 months, and then bi-weekly for the duration of the trials (fig. 1). Devices not meeting this standard at any time were fixed or replaced and retested before being used.

**Simulator testing**

Simulator testing procedures have been described previously [11]. The DLco simulator delivers known inspiratory volumes (VI) and exhaled concentrations of carbon monoxide (CO) and tracer gas (helium, methane or neon). Two CO concentrations (low: 0.1% and high: 0.13%) and two VI (3.5 L and 4.5 L) were used to generate four test gas/VI combinations. Each simulation used one test gas/VI combination. Target DLco values for this combination were calculated while accounting for the breath-hold time (BHT), temperature and barometric pressure. Target values for alveolar volume (VA) were also calculated. The absolute differences between all measured and target values were used as measures of accuracy. Changes in accuracy were assessed over time. The effects of different regions, device type, BHT, temperature and pressure on the accuracy of DLco devices were also analysed.

**Statistical analyses**

SAS (Versions 8 and 9, SAS Institute, Raleigh, N.C.) was used for analyses. Data from all sites were combined. Outliers were removed from the data set. These data points
represented non-physiologic values or values that would be impossible to obtain in a simulator test and likely were the result of a data entry error or a leak in the connections between the simulator and the DLco measuring device. Outliers were defined as: measured and target DLco values <5 ml·min⁻¹·mmHg⁻¹ or >60 mL·min⁻¹·mmHg⁻¹; VA values <1L or >7L; VI values less than the simulated inspired volume or >7L; simulated expired CO concentration >100%; measured expired CO concentration <5% or >120% of the simulated CO concentration; simulated expired tracer gas concentrations <0.10%; measured expired tracer gas concentrations <50% or >150% of the simulated value.

The Cochran-Armitage trend test was used to detect trends over time in the percentage of failed simulation tests. To assess changes in accuracy over time or the effects of regions, device type, BHT, temperature and pressure on DLco accuracy, a repeated measures model was fitted using sites as the random effect and the following covariates: years from initial certification test and tank gas/VI combination; region and the region and tank gas/VI combination interaction (to assess effect of regions); device type and the device type and tank gas/VI combination interaction (to assess effect of devices); BHT and the BHT and tank gas/VI combination interaction (to assess effect of BHT); temperature, pressure and the temperature and tank gas/VI combination interaction (to assess the effects of temperature and pressure). Compound symmetry was used as the covariance structure. Interaction terms that were not significant at the 0.05 level were removed from the model. A similar model was used to assess changes in accuracy for VA, VI and CO and tracer gas concentrations with covariates of: years from initial certification test and tank gas/VI combination. Statistical significance was concluded when the p-value for two-sided test was less than 0.05.
RESULTS:

**Initial DLco device performance in pulmonary function laboratories**

A total of 125 pulmonary function testing laboratories were recruited to this study. Initially, 31 (25%) sites had DLco devices that measured a DLco value more than 3 mL·min⁻¹·mmHg⁻¹ different from the target and were considered to have failed. Of the sites that failed the initial certification, 24 (19.8%) passed after replacing parts and/or repairing the equipment. Devices at six (4.8%) other sites passed after new equipment was purchased. One (0.8%) site was unable to certify its DLco device and was disqualified from the study. After correcting device issues or purchasing new equipment, 124 sites (99%) had devices that were measuring the DLco within 3 mL·min⁻¹·mmHg⁻¹ of the simulated value.

**DLco device performance over time**

After the initial certification test, these 124 PFT laboratories tested their equipment using the simulator weekly for 3 months and then bi-weekly for the duration of the trials. When devices were found to be malfunctioning, they were serviced. Data was collected from 9083 DLco simulation tests and for up to 3 years after certification. After outliers were removed, data for 9025 DLco measurements, 7271 VA measurements, 8810 VI measurements, 7770 expired CO concentration measurements and 8300 expired tracer gas concentration measurements were analysed.

In Figure 2 the percentage of all simulator tests in which a device failed is displayed in 2-month intervals from the date of the initial certification test. After a transient increase from 5.0% in the first interval (0–2 months) to 7.1% in the second interval (2–4 months), the proportion of test failures significantly decreased to 3.1% for
all tests conducted between 12 months and up to 36 months from the initial certification test (p<0.001).

To more closely examine how well sites maintained the accuracy of their device, the results of all simulator tests were analysed over time. A regression line relating the absolute difference from the target DLco value for each test and the time since the initial certification test for each site was fit to data from all four tank gas/VI combinations (fig. 3). At first the average absolute difference from target for all tests was 1.54 ± 0.05 mL·min⁻¹·mmHg⁻¹ (Mean ± SE) and decreased significantly by –0.13 mL·min⁻¹·mmHg⁻¹/yr (p=0.0002). A similar analysis was used to examine trends in the accuracy of the measured values for VI, VA, expired CO and tracer gas concentrations. A decrease in the absolute difference between measured and target VA and VI of –13.03 ± 4.61 mL/yr (Mean ± SE), p=0.0047 and –16.14 ± 2.07 mL/yr (Mean ± SE), p<0.0001, respectively was observed (table 1). There was no significant change in the accuracy of measured expired CO concentration or tracer gas concentration.

**DLco device accuracy associated with each region, tank gas/VI combination and device type**

The results of the simulator testing were also used to examine the effects of variations in several different conditions on the accuracy of DLco devices. For geographical regions, the average absolute differences between measured and target simulator values ranged from 1.1 to 2.0 mL·min⁻¹·mmHg⁻¹ (North America: 1.2–1.7, Central/South America: 1.2–1.8, EU: 1.2–1.8, Asia: 1.1–2.0). For each tank gas/VI combination that was used, the
average absolute difference between the measured and target ranged from 1.2 to 1.7 mL·min⁻¹·mmHg⁻¹ (table 2).

Each site used a DLco measuring device of their choice for the duration of the trial, for a total of 12 device types used across all of the sites. For most of the DLco device types, the average absolute difference between the measured and target DLco value ranged from 0.6 to 2.1 mL·min⁻¹·mmHg⁻¹ (fig. 4). Only one of the devices made measurements that were on average 3.7 mL·min⁻¹·mmHg⁻¹ different from the target at one of the four tank gas/VI combinations (Vol. 4.5 [CO] high).

**DLco device accuracy over ranges of pressure, BHT and temperature**

Variations in barometric pressure (range: ~550–800 mmHg) and BHT (range: ~1–15 seconds) did not significantly affect the accuracy of DLco devices (p=0.18 and p=0.24, respectively). Although there was a statistically significant effect of variations in temperature (range: ~10–37°C) on DLco accuracy, this was only true for two of the simulated target values (Vol. 3.5 [CO] high and Vol. 4.5 [CO] high, table 3), and a change of −0.03 ± 0.01 (Mean ± SE) or −0.009 ± 0.01 mL·min⁻¹·mmHg⁻¹ (Mean ± SE) per increase of 1 degree Celsius might not be considered a meaningful effect on the accuracy of DLco devices.

**DISCUSSION:**

A reported high inter- and intra-laboratory variability [7–9] has limited the effective use of the DLco measurement in clinical trials. The analysis reported here suggests that in order to collect quality DLco data in large multi-centre trials, PFT laboratories need to be monitored and that regular DLco simulator testing is an effective
method for maintaining device accuracy and precision for the duration of a clinical trial. By maintaining accurate devices, variability of DLco data is reduced facilitating detection of small changes in diffusing capacity in a test population.

To our knowledge, this is the first and largest study of DLco simulator testing data collected during global, multi-centre clinical studies. Data was collected from over 9,000 simulator tests performed in 124 PFT laboratories, in several different countries and using several different machines over several years.

Of note, 25% of PFT laboratories included in this study were not measuring the DLco within a targeted degree of accuracy prior to corrective intervention. This has significant implications for any multi-site clinical trial that requires DLco measurements because it suggests that a large percentage of existing PFT laboratories have DLco devices that are either malfunctioning or are being used improperly. To reduce variability to an acceptable standard, simulator testing is necessary to identify these laboratories and mitigate their deficiencies before inclusion into clinical trials.

The 124 sites that passed the certification improved their ability to maintain the accuracy of their devices within the pre-defined guidelines (<3 mL·min⁻¹·mmHg⁻¹ difference from target). Despite a slight initial increase, the percentage of failed simulation tests significantly decreased over time. In addition, sites improved the overall accuracy of their DLco devices over time. The observed initial increase in simulation test failures may reflect level of technician inexperience with respect to device maintenance in the first few months after certification. It is likely that the positive and negative feedback that the simulator testing provides helps the technicians become more aware
and capable with respect to maintaining and calibrating their DLco devices resulting in a decrease in simulation test failures over time.

The simulator-based quality control program was able to limit the effects of environmental and operational conditions that vary between laboratories and within a laboratory over time, and thus data collected in different regions and different laboratories was highly comparable. On average DLco measurements made in different regions and by different devices were lower than 2 and 2.1 mL·min⁻¹·mmHg⁻¹, respectively, and DLco device accuracy was stable over a range of barometric pressure, temperature and BHT.

DLco device accuracy was also stable across the test gas / VI combinations that were used. The test gas / VI combinations were used to create different DLco target values which could be calculated after accounting for the local temperature, barometric pressure and the BHT of the device. That the averages of the absolute differences calculated between the measured and the simulated targets for each of the tank gas/ VI combinations were similar is consistent with what has been seen in clinical testing. Punjabi et al. [14] analyzed repeated DLco measurements made in patients with a range of obstructive and restrictive ventilatory impairment. They showed that the absolute difference between repeated DLco measurements in individual patients remained stable over a wide range of DLco values.

With respect to patient testing in large, multi-centre clinical trials, maintaining accurate DLco devices and diminishing inter- and intra-laboratory variability can lead to detection of smaller changes in DLco over time. This improvement in sensitivity is particularly important because the effects of some inhaled therapeutic agents are likely to
be small. It has been reported that inhaled insulin products (insulin human [rDNA origin] inhalation powder and human insulin inhalation powder) on average caused reversible reductions in DLco of 0.5 – 1.0 mL·min⁻¹·mmHg⁻¹ after 12 to 24 weeks of treatment [13, 15-17], an amount difficult to reliably assess without optimum performance of DLco testing procedures across all study sites.

Although simulator testing reduces the variability of DLco measurements due to instrumentation, it does not address all sources of variability. The DLco maneuver is complex for study subjects and requires substantial effort and cooperation [10]. Each subject is an additional source of variability [8, 12], and it has been shown that the patient can account for between 30 and 60% of intra-subject variation in DLco measurements [18].

To minimise this source of variability in multi-centre clinical trials, additional standardisation is necessary for patient testing procedures. Another study has shown that a centralised review of patient test results along with standardised instrumentation and a demonstrated competency of technicians can reduce the variability of DLco tests in patients in multi-centre clinical trials [19]. A root mean-squared coefficient of variation (RMSCV) of 6.0% was found compared to a previously reported RMSCV >9.0% in trials that did not use these methods [15, 16]. In clinical trials for AIR® Inhaled Insulin, only highly experienced PFT laboratories were used to collect the data used in this study. In addition, patient test data underwent independent reviews throughout the trials. Combining these types of standards for patient testing with DLco simulator testing may further decrease the variability of DLco measurements in multi-centre clinical trials and in clinical testing.
The results of this study demonstrate that to obtain high quality DLco data in multi-centre, global clinical trials, appropriate monitoring of DLco devices is essential. Using a DLco simulator allows PFT laboratories to maintain accuracy thus reducing inter- and intra-laboratory variability of the DLco measurement throughout the study.

ACKNOWLEDGEMENTS

The authors would like to acknowledge D. Shrom (Eli Lilly and Company, Indianapolis, IN, USA) for support in the preparation of this manuscript.
Reference List

1. Muchmore DB, Gates JR. Inhaled insulin delivery--where are we now? *Diabetes Obes Metab* 2006; 8: 634-642.


<table>
<thead>
<tr>
<th></th>
<th>Decrease in absolute difference per year</th>
<th>P value&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>VA (mL)</td>
<td>$-13.03 \pm 4.61$</td>
<td>0.005</td>
</tr>
<tr>
<td>VI (mL)</td>
<td>$-16.14 \pm 2.07$</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Exp CO (ppm)</td>
<td>$6.58 \pm 9.28$</td>
<td>0.48</td>
</tr>
<tr>
<td>Exp Tracer (ppm)</td>
<td>$-0.50 \pm 88.96$</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Exp: expired; VA: alveolar volume; VI: inspiratory volume

<sup>a</sup> Least squared means ± SE

<sup>b</sup> p value calculated for difference from 0
TABLE 2  Average absolute difference between measured and simulated DLco values for each tank gas/VI combination

<table>
<thead>
<tr>
<th>Tank gas/VI combination</th>
<th>Average Absolute difference between measured and target value (mL·min⁻¹·mmHg⁻¹)ᵃ</th>
<th>P valueᵇ</th>
</tr>
</thead>
<tbody>
<tr>
<td>VI=  3.5 L [CO] high</td>
<td>1.4 ± 0.05</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VI=  3.5 L [CO] low</td>
<td>1.2 ± 0.05</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VI=  4.5 L [CO] high</td>
<td>1.7 ± 0.05</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VI=  4.5 L [CO] low</td>
<td>1.5 ± 0.05</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

VI: inspiratory volume
ᵃ Least squared means ± SE
ᵇ p value calculated for differences from 0
**TABLE 3**  The effect of temperature on DLco device accuracy with respect to four tank gas/VI combinations

<table>
<thead>
<tr>
<th>Tank gas/VI combination</th>
<th>$\Delta$ absolute difference between measured and target DLco value per $C^\circ b$</th>
<th>P value$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VI= 3.5 L [CO] high</td>
<td>-0.03 ± 0.01 (mL·min$^{-1}$·mmHg$^{-1}$)</td>
<td>0.03</td>
</tr>
<tr>
<td>VI= 3.5 L [CO] low</td>
<td>-0.002 ± 0.01 (mL·min$^{-1}$·mmHg$^{-1}$)</td>
<td>0.87</td>
</tr>
<tr>
<td>VI= 4.5 L [CO] high</td>
<td>-0.05 ± 0.01 (mL·min$^{-1}$·mmHg$^{-1}$)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VI= 4.5 L [CO] low</td>
<td>-0.009 ± 0.01 (mL·min$^{-1}$·mmHg$^{-1}$)</td>
<td>0.44</td>
</tr>
</tbody>
</table>

VI: inspiratory volume

$^a$ p value calculated for difference from 0

$^b$LS Means ± SE
FIGURE 1. Each pulmonary function laboratory was required to certify the accuracy of their DLco measuring device using a simulator. Each test used one of four tank gas/VI combinations 4.5 or 3.5 L inspiratory volumes and high or low [CO]. If a laboratory found its device to measure a DLco different from the target by greater than 3 mL·min⁻¹·mmHg⁻¹, it used the results of the simulator test and additional testing to isolate the source of the malfunction. After certifying a device, patient testing could begin. Regular simulator testing continued for the duration of the trials.
FIGURE 2. The percentage of simulated measurements that were different from the target by more than 3 mL·min$^{-1}$·mmHg$^{-1}$ is plotted in two month time intervals after the initial certification test. The percentage of failed tests decreased significantly over time. Cochran-Armitage trend test, two-sided probability: $p<0.001$.

![Diagram showing percent of failed simulation tests in two month intervals after the initial certification test.](image)

FIGURE 3. The absolute difference between the target DLco value and the measured value decreased on average over time for all four tank gas/VI combinations (a. VI=3.5 L [CO] high; b. VI=3.5 L [CO] low; c. VI=4.5 L [CO] high; d. VI=4.5 L [CO] low).

Regression lines are shown for change in absolute difference over time (years since initial certification test) on each graph. The slope of the line is ($-0.13$ mL CO·min$^{-1}$·mmHg$^{-1}$) / year.

![Diagram showing regression lines for change in absolute difference over time for different tank gas/VI combinations.](image)
FIGURE 4. Accuracy was assessed for all 12 DLco measuring device types used by the sites in this study by calculating the mean of the absolute values of the difference between the measured and target DLco values. Mean absolute differences plus 95% confidence intervals are presented for each machine at each of the tank gas/VI combinations used.
Mean absolute difference between measurement and target for each device

Vi = 4.5 L
[CO] low

Vi = 4.5 L
[CO] high

Vi = 3.5 L
[CO] low

Vi = 3.5 L
[CO] high

Abs (measured - target) mL min⁻¹ mmHg⁻¹