Quality control of DL\textsubscript{L,CO} instruments in global clinical trials

R. Jensen*, M. Leyk*, R. Crapo*, D. Muchmore* and P.Y. Berclaz*

ABSTRACT: High inter- and intra-laboratory variability exists for the single-breath diffusing capacity of the lung for carbon monoxide (DL\textsubscript{L,CO}) test. To detect small changes in diffusing capacity in multicentre clinical trials, accurate measurements are essential. The present study assessed whether regular DL\textsubscript{L,CO} simulator testing maintained or improved instrument accuracy and reduced variability in multicentre trials.

The 125 pulmonary function testing laboratories that participated in clinical trials for AIR\textsuperscript{®} Inhaled Insulin validated and monitored the accuracy of their DL\textsubscript{L,CO} measuring devices using a DL\textsubscript{L,CO} simulator, which creates known target values for any device. Devices measuring a simulated DL\textsubscript{L,CO} different from target by $> 3 \text{ mL min}^{-1} \text{ mmHg}^{-1}$ 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 DL\textsubscript{L,CO} 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 DL\textsubscript{L,CO} device accuracy.

Regular diffusing capacity of the lung for carbon monoxide simulator testing allows pulmonary function testing laboratories to maintain the accuracy of their diffusing capacity measurements, leading to reduced variability across laboratories in multicentre clinical trials.

KEYWORDS: Accuracy, inhaled insulin, multicentre trials, quality control, simulator, single-breath diffusing capacity of the lung for carbon monoxide

Pulmonary function tests (PFTs), such as spirometry (forced expiratory volume in one second and forced vital capacity), lung volume measurements (total lung capacity) and the single-breath diffusing capacity of the lung for carbon monoxide (DL\textsubscript{L,CO}), 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 DL\textsubscript{L,CO}, 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 DL\textsubscript{L,CO} is well established, an issue facing investigators in large-scale, multicentre clinical trials is the high inter- and intra-laboratory variability for this measurement [7–9]. Measuring the DL\textsubscript{L,CO} 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 DL\textsubscript{L,CO} measurements to detect potential harmful or beneficial effects of an inhaled product in a test population.

To control the accuracy and reproducibility of DL\textsubscript{L,CO} data in clinical trials, processes for quality control must be established. To this end, clinical trials for AIR\textsuperscript{®} Inhaled Insulin (Alkermes, Inc., Cambridge, MA, USA) have used a centralised quality-assurance programme, 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 DL\textsubscript{L,CO} devices using a simulator (Hans Rudolph, Kansas City, MO). The DL\textsubscript{L,CO} simulator creates a precisely known, repeatable DL\textsubscript{L,CO} value in any device. Comparing measured DL\textsubscript{L,CO} values with
the target DL,CO value allows the devices’ accuracy and precision to be assessed. Using the DL,CO 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 programme that includes DL,CO simulator testing can establish and maintain the accuracy of DL,CO 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 were reported. Absolute differences between measured and target values, as well as trends over time in these differences, were analysed. In addition, the accuracy of DL,CO measurements was assessed with respect to variations in room temperature, barometric pressure, breath-hold time (BHT), region, machine type and target simulator values.

METHODS

Study design
In total, 125 PFT laboratories from North America (Canada and the USA), the European Union (EU; Belgium, Bulgaria, Croatia, Germany, Hungary, Italy, Poland and Portugal), Central/South America (Argentina, Brazil, Chile, Colombia, Mexico and Puerto Rico) and Asia (India, Philippines, Singapore and Taiwan) were recruited to clinical trials for AIR Insulin. DL,CO measurements were performed throughout the trials in study subjects as a safety evaluation.

During the trials, each laboratory used a DL,CO 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 re-tested before being used.

Simulator testing

Simulator testing procedures have been described previously [11]. The DL,CO simulator delivered known inspired volumes (Vi) and exhaled concentrations of CO and tracer gas (helium, methane or neon). Two CO concentrations (low 0.1% and high 0.13%) and two Vi (3.5 and 4.5 L) were used to generate four test gas/Vi combinations. Each simulation used one test gas/Vi combination. Target DL,CO values for this combination were calculated while accounting for the 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 DL,CO devices were also analysed.

Statistical analyses

Data from all sites were combined. Outliers were removed from the data set. These data points represented nonphysiological values or values that would be impossible to obtain in a simulator test and probably were the result of a data entry error or a leak in the connections between the simulator and the DL,CO measuring device. Outliers were defined as follows: measured and target DL,CO values <5 mL·min⁻¹·mmHg⁻¹ or >60 mL·min⁻¹·mmHg⁻¹; VA values <1 L or >7 L; Vi values less than the simulated Vi or >7 L; 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 DL,CO 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 the following covariates: years from initial certification test and tank gas/Vi combination. Statistical significance was concluded when the p-value for two-sided test was <0.05.
RESULTS

Initial DLCO device performance in pulmonary function laboratories

A total of 125 PFT laboratories were recruited to the present study. Initially, 31 (25%) sites had DLCO devices that measured a DLCO value $>$ 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 9,083 DLCO simulation tests and for up to 5 yrs after certification. After outliers were removed, data for 9,025 DLCO measurements, 7,271 VA measurements, 8,810 VI measurements, 7,770 expired CO concentration measurements and 8,300 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 fitted to data from all four tank gas/VI combinations (fig. 3). At first, the mean $\pm$ SE absolute difference from target for all tests was 1.54 $\pm$ 0.05 mL·min$^{-1}$·mmHg$^{-1}$ and decreased significantly by -0.13 mL·min$^{-1}$·mmHg$^{-1}$·yr$^{-1}$ ($p=0.0002$). 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 mean $\pm$ SE absolute difference between measured and target VA and VI of -13.03 $\pm$ 4.61 mL·yr$^{-1}$ ($p=0.0047$) and -16.14 $\pm$ 2.07 mL·yr$^{-1}$ ($p=0.0001$), 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 and 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, giving 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 (VI 4.5, high concentration of CO).

DLCO device accuracy over ranges of pressure, BHT and temperature

Variations in barometric pressure (range ~550–800 mmHg) and BHT (range ~1–15 s) 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 (VI 3.5, high concentration of CO, and VI 4.5, high concentration of CO; table 3), and a mean $\pm$ SE change of -0.03 $\pm$ 0.01 or -0.009 $\pm$ 0.01 mL·min$^{-1}$·mmHg$^{-1}$ per increase of 1°C 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 in the present study suggests that, in order to collect quality DLCO data in large multicentre 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 the current authors’ knowledge, the present study is the first and largest study of DL,CO simulator testing data collected during global, multicentre clinical studies. Data was collected from >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 the present study were not measuring the DL,CO within a targeted degree of accuracy prior to corrective intervention. This has significant implications for any multi-site clinical trial that requires DL,CO measurements, because it suggests that a large percentage of existing PFT laboratories have DL,CO devices that are either malfunctioning or are being used improperly. To reduce

### TABLE 1

<table>
<thead>
<tr>
<th>Tank gas/VI combination</th>
<th>Decrease in absolute difference per yr</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>VmL</td>
<td>-13.03 ± 4.61</td>
<td>0.005</td>
</tr>
<tr>
<td>Vl mL</td>
<td>-16.14 ± 2.07</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Expired CO ppm</td>
<td>6.58 ± 9.28</td>
<td>0.48</td>
</tr>
<tr>
<td>Expired tracer ppm</td>
<td>-0.50 ± 88.96</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Data are presented as least squared mean ± SE, unless otherwise stated. Vm: alveolar volume; Vl: inspired volume. *: calculated for difference from zero.

### TABLE 2

<table>
<thead>
<tr>
<th>Tank gas/VI combination</th>
<th>Average absolute difference*</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>VI 3.5 L, high CO</td>
<td>1.4 ± 0.05</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VI 3.5 L, low CO</td>
<td>1.2 ± 0.05</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VI 4.5 L, high CO</td>
<td>1.7 ± 0.05</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VI 4.5 L, low CO</td>
<td>1.5 ± 0.05</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Data are presented as least squared mean ± SE, unless otherwise stated. *: between measured and target value of diffusing capacity of the lung for carbon monoxide. *: calculated for difference from zero.
variability to an acceptable standard, simulator testing is necessary, in order 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 predefined 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 provided helped 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 programme 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 were highly comparable. On average, DLco measurements made in different regions and by different devices were <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.

### FIGURE 4.

Accuracy was assessed for all 12 device types for measuring diffusing capacity of the lung for carbon monoxide (DLco) used by the sites in the present 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/inspired volume (Vi) combinations used. a) Vi 4.5 L, low concentration of CO; b) Vi 4.5 L, high concentration of CO; c) Vi 3.5 L, low concentration of CO; d) Vi 3.5 L, high concentration of CO. Manufacturers of the devices were as follows. Zan300: nSpire Health, Longmont, CO, USA. SpiroTech: VASYS Healthcare, Yorba Linda, CA, USA. SensorMedic: Sensor Medics, Yorba Linda. SandMKeystone: S&M Instruments Co., Doylestown, PA, USA. Medisoft: Medisoft, Dinant, Belgium. MedGraph: Medical Graphics Corp., St Paul, MN, USA. Lung Function: XXXXXXXXX. Jaeger: Jaeger, Wurzburg, Germany. GansHorn: Ganshorn Medizin Electronic, Niederlauer, Germany. CosMed: Cosmed, Rome, Italy. Collins: Ferraris Respiratory, Louisville, KY, USA.

### TABLE 3

<table>
<thead>
<tr>
<th>Tank gas/Vi combination</th>
<th>Change in absolute difference³ mL·min⁻¹·mmHg⁻¹ per 1 °C</th>
<th>p-value ⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vi 3.5 L, high CO</td>
<td>-0.03 ± 0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Vi 3.5 L, low CO</td>
<td>-0.002 ± 0.01</td>
<td>0.87</td>
</tr>
<tr>
<td>Vi 4.5 L, high CO</td>
<td>-0.05 ± 0.01</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Vi 4.5 L, low CO</td>
<td>-0.009 ± 0.01</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Data are presented as least squared mean ± se, unless otherwise stated. *: for measuring diffusing capacity of the lung for carbon monoxide (DLco); #: calculated for difference from zero.
combinations were similar is consistent with what has been seen in clinical testing. Punjabi et al. [14] analysed 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, multicentre 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 12–24 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 manoeuvre 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 multicentre 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 instrumentatation and a demonstrated competency of technicians, can reduce the variability of DLCO tests in patients in multicentre clinical trials [19]. A root-mean-squared coefficient of variation (RMSCV) of 6.0% was found, compared with 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 the present 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 multicentre clinical trials and in clinical testing.

The results of the present study demonstrate that, in order to obtain high quality diffusing capacity of the lung for carbon monoxide data in multicentre, global clinical trials, appropriate monitoring of diffusing capacity devices is essential. Using a diffusing capacity simulator allows pulmonary function testing laboratories to maintain accuracy, thus reducing inter- and intra-laboratory variability of the measurement of diffusing capacity of the lung for carbon monoxide 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.

REFERENCES
1 Muchmore DB, Gates JR. Inhaled insulin delivery – where are we now? Diabetes Obes Metab 2006; 8: 634–642.
8 Crapo RO, Jensen RL. Standards and interpretive issues in lung function testing. Respir Care 2003; 48: 764–772.
16 Quattrin T, Bélanger A, Bohannon NJ, Schwartz SL, Exubera Phase III Study Group, Efficacy and safety of
inhaled insulin (Exubera) compared with subcutaneous
insulin therapy in patients with type 1 diabetes: results of a
6-month, randomized, comparative trial. Diabetes Care

17 Skyler JS, Weinstock RS, Raskin P, et al. Use of inhaled
insulin in a basal/bolus insulin regimen in type 1 diabetic

18 Jensen RL, Teeter JG, England RD, et al. Sources of long-

19 Wise RA, Teeter JG, Jensen RL, et al. Standardization of the
single-breath diffusing capacity in a multicenter clinical