Longitudinal validity of spirometers – a challenge in longitudinal studies

Nino Künzli[a, b], Birgit Kuna-Dibbert[b*], Dirk Keidel[b], Roland Keller[c], Otto Brändli[d], Christian Schindler[b], Karl-Manfred Schweinzer[e], Philippe Leuenberger[f], Ursula Ackermann-Liebrich[b], on behalf of the SAPALDIA team*

[a] University of Southern California, Los Angeles, USA
[b] Institute of Social and Preventive Medicine, University of Basel, Switzerland
[c] Pulmonology Clinic, Barmelweid, Aarau, Switzerland
[d] Züricher Höhenklinik Wald, Switzerland
[e] Federal Office and Research Centre for Forests, Vienna, Austria
[f] Service of Pulmonology, University Hospital Lausanne, Switzerland

Summary

Question under study: Pulmonary function testing (PFT) in longitudinal studies involves the repeated use of spirometers over long time periods. We assess the comparability of PFT results taken under biologic field conditions using thirteen certified devices of various technology and age. Comparability of measurements across devices and over time is relevant both in clinical and epidemiological research.

Methods: Forced Vital Capacity (FVC), Forced Expiratory Volume in the first second (FEV$_1$) and Forced Expiratory Flow 50% (FEF50) were compared before and after the data collection of the Swiss Study on Air Pollution and Lung Diseases in Adults (SAPALDIA) and the European Community Respiratory Health Survey (ECRHS) cohort studies. Three test series were conducted with 46, 50 and 56 volunteers using various combinations of spirometers to compare the eight flow-sensing spirometers (Sensormedics 2200) used in the SAPALDIA cross-sectional and follow-up, two new flow-sensing instruments (Sensormedics V$_{max}$) and three volume displacement spirometers (two Biomedin/Baires and one Sensormedics 2400).

Results: The initial comparison (1999/2000) of eight Sensormedics 2200 and the follow-up comparison (2003) of the same devices revealed a maximal variation of up to 2.6% for FVC, 2.4% for FEV$_1$ and 2.8% for FEF50 across devices with no indication of systematic differences between spirometers. Results were also reproducible between Biomedin, Sensormedics 2200 and 2400. The new generation of Sensormedics (V$_{max}$) gave systematically lower results.

Conclusions: The study demonstrates the need to conduct spirometer comparison tests with humans. For follow-up studies we strongly recommend the use of the same spirometers.

Key words: spirometer comparison; quality control; cohort studies; lung function measurements

Introduction

Forced expiratory lung function measurements are an objective indicator of cardio-respiratory health, being related to both acute and long-term morbidity and mortality [1–4] To investigate environmental determinants of lung function growth in children [5] and decline among adults such as in the Swiss Study on Air Pollution and Lung Diseases (SAPALDIA) [6] and the European Community Respiratory Health Survey (ECRHS) [7], repeated measurements of lung function are required across longer time periods [8, 9]. In the first quality control studies we assessed the comparability of different devices, teams and field-workers during the cross-sectional SAPALDIA study to ensure unbiased comparisons across communities and sufficient precision to detect small differences [10, 11]. To investigate determinants of lung function change over time, measurement quality and standard procedures need to remain constant not only across fieldworkers, teams, and centres [11] but also over time. We now address key challenges faced by multi-centre follow-up studies that repeat spirometries after several years. Similar to clinical settings, follow-up studies have to consider revisions or replacement of equipment and/or updates of software. Systematic differences

of a few percent between devices may not influence clinical decisions but could have devastating effects in research. Therefore, we conducted a total of three comparison tests. The goals of the first two test series were to compare the performance of the eight formerly used SAPALDIA spirometers (Sensormedics, Yorba Linda, USA) with each other, with the same reference instrument used in the 1992 comparison test [11], with a new generation of Sensormedics devices (V_{max} 22), and with the spirometers used in most of the ECRHS centres (Biomedin Baires, Padova, Italy) since a subgroup of SAPALDIA belongs to ECRHS. The results of these tests led to the conclusion to again use the instruments from SAPALDIA 1 in SAPALDIA 2 [8]. The goal of the third test series was a comparison across the eight spirometers used in SAPALDIA 2 to assess device comparability at completion of the follow-up. We will discuss the implications of these comparison tests on long-term research projects.

Methods

Subjects and sessions

Subjects in all tests were healthy non-smoking male and female volunteers recruited for each test series separately from the student body of the University of Basel. We chose the sample sizes of the different comparison tests in such a way as to guarantee that a true 3%-range in the average volume measurements of the spirometers under comparison would be detected at the 5%-significance level with a probability of 80%. Power was estimated using the SAS-Macro fpower from M. Friendly (http://www.math.yorku.ca/SCS/sasmac/fpower.html).

Volunteers performed forced expiratory pulmonary function testing (PFT) using each spirometer in the test series. A Latin square design was used to determine the order of devices. These sessions consisted of at least three and up to eight manoeuvres at each spirometer (seated position) until ATS acceptability and repeatability criteria were fulfilled. Quality control, acceptability and repeatability criteria, and manoeuvre selection procedures were identical to those employed in SAPALDIA 1 and ECRHS, and met the American Thoracic Society (ATS) standards [6, 9, 12]. The session at the first spirometer was considered a practice session, thus, the data were barred from the analysis and students had to perform again at this instrument later on in their circuit.

Spirometers

The spirometers used were Sensormedics 2200, Sensormedics 2400, Sensormedics V_{max} 22 (all from Yorba Linda, USA, distributor Gambro AG/Pilger-AG, Switzerland) and Biomedin Baires (Biomedin, Padova, Italy). Sensormedics 2200 and Sensormedics V_{max} are flow-sensing spirometers (heated wire), whereas Sensormedics 2400 (dry rolling seal) and Biomedin Baires (water seal) are volume-displacement spirometers. In our first device comparison [11] we compared the flow sensing devices with the traditional dry rolling seal instruments (Sensormedics 2400); thus, we considered the latter as our "reference device".

All instruments measured with conventional body temperature, pressure and saturation (BTPS) corrections.

Calibration of spirometers

Devices were calibrated with 3-litre syringes at least twice a day and when instruments were switched on. Since the Biomedin syringe tube did not fit on Sensormedics spirometers we had to use two syringes. Digital measurements of room temperature and air pressure were entered into the spirometer software before calibration was performed. The calibration was conducted at varying speeds and considered successful when the inspired and expired manoeuvres measured between 99% and 101% of the volume. All calibrations performed during all test series met these requirements prior to starting the tests.

Three test series

Table 1 presents the time period, sample size and devices relevant to each test series. All series were conducted in the same two adjacent rooms with stable room temperature. Test series 1 included the eight SAPALDIA instruments used in SAPALDIA 1 (Sensormedics 2200; Yorba Linda, USA) and one device (V_{max} #1) of the new generation of mass flow meter (Sensormedics, V_{max} 22 Yorba Linda, USA). Test series 2 compared two new Biomedin/Baires instruments (Biomedin; Padova, Italy), two

Table 1

Description of three spirometer comparison test series.

Factor	test series 1	test series 2	test series 3
Test dates (each series took 3 days)	Nov. 1999	Jan. 2000	June 2003
Subjects (N)	46	50	56
Females (in %)	49.3	52.3	51.7
Mean (SD) age (years)	24.1 (6.5)	24.3 (5.2)	25.4 (5.6)
Mean (SD) height (cm)	173.4 (9.3)	172.8 (8.6)	173.2 (8.3)
Number of devices (Number of tests)[a]	9 (414)	6 (300)	8 (448)
Devices tested			
Sensormedics (Yorba Linda, CA, USA)			
– 2200 hot wire flow meter[b]	device S1 to S8	S2 , S3	device S1 to S8
– 2400 dry seal spirometer	–	S-ref	–
– Vmax hot wire flow sensor	V_{max} #1	V_{max} #2	–
Biomedin Baires (Padova, Italy)			
– Biomedin water seal spirometer	–	B1, B2	–

[a] Number of tests = number of subjects × number of devices included in the test series
[b] These eight devices have been used in SAPALDIA 1 and SAPALDIA 2

of the above mentioned eight SAPALDIA 1 devices (Sensormedics 2200), one Sensormedics 2400 (Yorba Linda, USA), already used as a "reference" in the 1992 comparison [11] and a second new Sensormedics (V_{max} #2). Test series 3 compared the eight SAPALDIA 1 instruments (Sensormedics 2200) at completion of SAPALDIA 2.

Technicians

All measurements in test series 1 and 2 were coached by the same technician. A second fieldworker conducted test series 3. Both technicians had worked for SAPALDIA and in lung function laboratories for at least one year doing daily routine measurements.

Analysis of spirometric data

We present results for forced vital capacity (FVC), forced expiratory volume in one second (FEV_1) and the mid-expiratory flow, FEF50, taken from the best manoeuvre. The personal means across all devices in the test series were calculated for each parameter. The deviation from this personal mean was derived for each person and spirometer. We compare the device-specific means of these deviations.

We used three-way-ANOVA to detect differences due to spirometers, subjects and order of testing, since each subject was tested on each spirometer once and the order of spirometers was changed between subjects according to a Latin square design. Since each subject provided a series of measurements it could not be assumed that a classical three-way-ANOVA, requiring independent statistical errors within subjects, would be appropriate. We therefore also considered two mixed linear models, one with random individual temporal trends and the other one assuming a first order autoregressive covariance structure of errors within each subject. These two analyses yielded very similar point estimates, standard errors and p-values as the classical three-way-ANOVA.

Since device was a significant factor in all series, we conducted pair wise comparisons of adjusted device-specific means using the simulation test supplied in the GLM-procedure of SAS. All analyses were conducted with SAS V8.2 and STATA 8.0.

Results

Characteristics for all test series are shown in table 1.

Figure 1

Device-specific mean (and 95% CI) deviation in percent from personal mean FVC calculated across results taken at all devices in test series 1.

Figure 2

Device-specific mean (and 95% CI) deviation in percent from personal mean FVC calculated across results taken at all devices in test series 2.
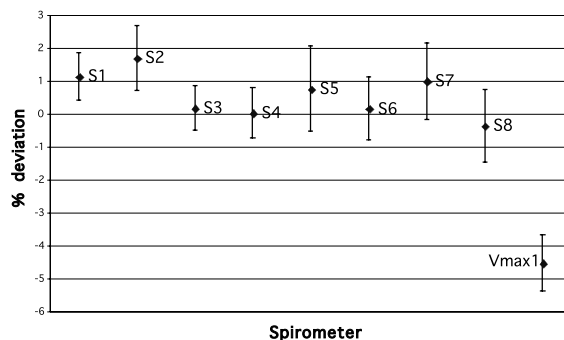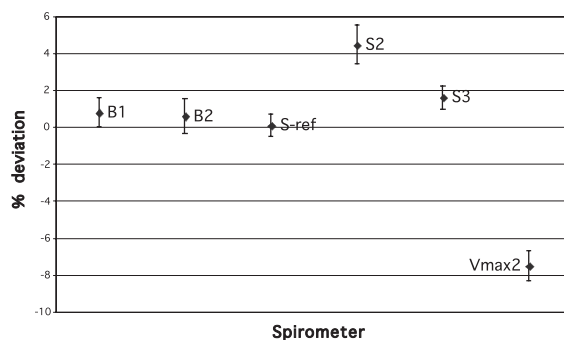

Figure 3

Device-specific mean (and 95% CI) deviation in percent from personal mean FVC calculated across results taken at all devices in test series 3.

Test series 1

Figure 1 displays the mean (and 95% confidence interval) of the percent deviation from the personal mean FVC across all instruments for each device of test series 1. Results for FEV_1 and FEF50 are not shown, as these are practically identical. The average deviations from the personal means observed for V_{max} #1 were all statistically significant and reached –4.5%, –4.5%, and –7.6% for FVC, FEV_1 and FEF50 respectively. Comparison of results after exclusion of the V_{max} spirometer resulted in mean deviations within ±1.1%, ±1.7%, and ±2.1%, respectively. The difference between the devices with the smallest and largest mean was statistically significant only in the case of FEV_1 (deviation from mean –1.2% versus +1.7%).

Test series 2

Figure 2 presents the device specific deviations from the personal mean FVC. Values from V_{max} #2 spirometer were significantly lower than those on any other device with deviations of –7.5% for FVC, –8.4% for FEV_1 and –11.3% for FEF50. Exclusion of this device revealed similar deviations as in test series 1 after exclusion of the V_{max} #1 spirometer. The deviations were within ±2.3%, ±1.9, and ±2.2%, respectively. The differences between the lowest and highest deviation were statistically significant in all three parameters.

Test series 3

Figure 3 presents the mean percent deviation of FVC for all eight SAPALDIA devices as observed at the completion of SAPALDIA 2. Deviations were within ±2.6%, ±2.4%, and ±2.8% for FVC, FEV, and FEF50, respectively and reached statistical significance between the extremes.
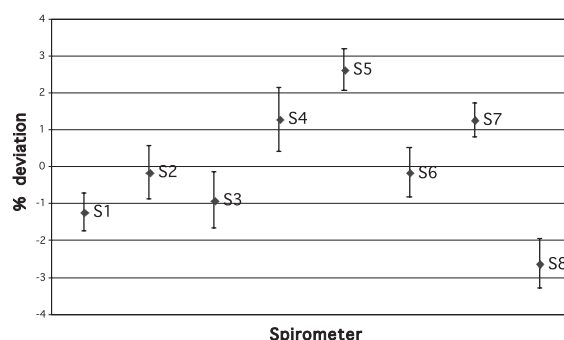
Discussion

Our study demonstrates that comparison of spirometers in the field, ie with human subjects under biologic conditions, is of paramount importance in the planning, conducting and quality assurance of epidemiological studies. This is particularly true for multi-centre studies and long-term follow-up investigations. We have shown that although all devices comply with the ATS standards of accurate instruments, and all calibrations being within the required precision, lung function test results taken under biologic conditions did differ significantly between instruments. Environmental conditions such as temperature, air pressure, humidity and fieldworker were the same for the tested instruments within each test series, thus do not explain our results.

The first two test series led to the informed decision about the reuse of the SAPALDIA 1 spirometers in the follow up study [8]. We first discuss the very important finding of systematic differences observed between two generations of the same device.

We then interpret the random variation of means measured across SAPALDIA devices, Sensormedics 2400 and the Biomedin spirometers that were widely used in ECRHS.

Systematic differences

To our surprise, the two Vmax devices provided systematically and substantially lower measurements in all parameters. Both Sensormedics 2200 and V_{max} use flow-sensing technology (heated wire). The systematic differences may be due to both hardware and software. Computer software has been described as a major source of discrepancies and may be one reason for our findings [13]. Details about changes in technological principles and software algorithms are not published or accessible; thus, the causes of the differences observed in these comparisons taken among biological testing conditions are not easily revealed.

Calibrations were all within the required 99% to 101% of the volume and only one syringe has been used for these Sensormedics devices. Linn et al. showed that a 1% difference in air volume readouts of a syringe can be a source of appreciable error in spirometric data [14]. Yet this should not lead to systematic differences given that each test series involved several calibrations. Moreover no systematic difference was observed between SAPALDIA devices and the Biomedin instruments although the latter were calibrated with another 3-litre-syringe. Thus, it is highly unlikely that the lower measurements in the V_{max} spirometers can be explained by calibration problems.

For the older spirometers (Sensormedics 2200, 2400), temperature and barometric pressure are measured in the ambient air and put into the system manually by the technician. Sensormedics V_{max} devices contained inbuilt temperature and barometric pressure sensors with the respective software. To our knowledge, there is no published report available about the influence of a changed technique of temperature and barometric pressure assessment in flow-sensing spirometers. Nevertheless, for volume measurements, deviations of up to 6% in FVC and FEV_1 have been found due to the altered assessment of temperature [15]. Linn et al. describe for electronic rolling-seal spirometers that the temperature reading in the computer software was updated whenever temperature changed by more than 0.2 °C [14]. In contrast, for spirometers with manually entered temperature, the update of temperature is done at calibration checks and appreciable measurement variation may occur because of imprecision in temperature measurements. Given our repeated calibrations and the rather stable temperature during all test series, temperature is a plausible source of random variation across devices but unlikely to explain the systematic differences. Software algorithms cannot be investigated in full detail by the user of devices; thus, undisclosed differences in BTPS corrections would not be detected.

Most importantly, our results demonstrate that technological changes and improvements can lead to systematically different readings even if all instruments fulfil the required technical quality criteria. New devices are usually tested in the laboratory only with the ATS 24 waveforms. The major advantage of our approach is the use of human subjects rather than machines to compare devices. This is a more realistic and relevant scenario with several potentially important differences. First, the gas concentrations and thermal conductivity of exhaled air differs from the ambient air; second, turbulence during exhalation is not simulated with the machines; third, the temperature profile of exhaled air is not taken into account in these tests. Effects of changes in the temperature of the spirometer or the humidity of the exhaled air are not tested in the mechanical waveform tests. Gilliland et al. showed significant associations between BTPS corrected FEV_1 and spirometer temperature (–0.24% per 1 degree Celsius) [16]. All these factors may contribute to systematic differences of test results conducted in humans, under field conditions. We did not conduct wave form comparisons, thus we do not know to what extent laboratory tests would come to the same conclusions.

Precision

After exclusion of the V_{max} devices, our tests showed much smaller differences between the mean measurements. We interpret these deviations as random rather than systematic device-specific effects; test series 1 and 3 that included all of the 8 SAPALDIA Sensormedics devices give no indication of systematic errors. For example, instru-

ment S2 ranked highest in series 2 but not in series 3. Moreover, the ranking of devices in these series was not associated with the ranking observed across the same devices in 1992 [11]. Therefore, our comparisons indicate that the combined sources of variability encountered in repeated testing of lung function under biologic conditions may lead to differences within a narrow range.

All deviations lie within ±3% for FVC, FEV_1 and even for FEF50. Other studies reported deviations of up to 6.2% for FVC and 5.8% for FEV_1 compared to a reference [17, 18]. American Thoracic Society (ATS) standards [12] allow individual spirometers' volume measurements to differ by ±3% from a reference instrument. The minimal recommendations for diagnostic spirometry for the range of 0.5 to 8 litres are ±3% of the readings for FVC as well as for FEV_1. For the FEF50 the accuracy should lie within ±5% for a range of 7 litres [12].

The fact that these small differences between devices reached statistical significance is a consequence of the sample size. With one exception due to a technical problem, differences in the 1992 test were not statistically significant but of similar size. Künzli et al. reported deviations between the eight SAPALDIA devices of up to 3.9% for FVC and up to 2.8% for FEV_1. In our study, we found a maximal variation in average deviations from personal means across the spirometers (V_{max} device excluded) in test series 1–3 of up to 2.6% for FVC, 2.4% for FEV_1 and 2.8% for FEF50 (figure 1; table available upon request).

It was reassuring to see that the eight Sensormedics 2200 generated comparable results, without evidence for systematic errors even though the instruments had been used for more than ten years. All devices compared well to the dry seal spirometer, ie our "reference device"; thus, a general "ageing" effect among the flow sensing Sensormedics 2200 is unlikely unless "ageing" had identical effects on both types of spirometers [11]. We conclude that device validity and accuracy are not a major concern in SAPALDIA 2.

The high agreement between Sensormedics 2200, Sensormedics 2400 and the Biomedin spirometers is particularly important for ECRHS II where the vast majority of centres used either Biomedin or Sensormedics. Therefore, device-specific deviations appear an unlikely source of errors among those centres [9].

In studies investigating potentially "small effects" such as in air pollution research the observed differences across the SAPALDIA devices may still be a source of bias or noise. Long-term multi-centre studies may thus adopt various strategies to minimise errors. The use of a single device and team, if feasible, strongly reduces the risk of systematic errors. Studies involving several devices and teams may rotate the instruments to prevent systematic errors.

Both long-term cohort studies as well as clinical laboratories need to take potential systematic and random errors into account in the planning of studies and the transition to newer technologies. Quality control programs ought to conduct comparisons under biological conditions. Test cycles ought to include a well maintained certified device that has not undergone hard- or software modifications as this can serve as a point of reference in times of technological advances.

Correspondence:
Nino Künzli, M.D. Ph.D.
Associate Professor,
Keck School of Medicine
University of Southern California,
1540 Alcazar Street CHP 236,
Los Angeles, CA 90033–9013
E-Mail: kuenzli@usc.edu

References

1 Sorlie PD, K.W., O'Connor G. Mortality associated with respiratory function and symptoms in advanced age. The Framingham Study. Am Rev Respir Dis 1989;140:379–84.

2 Chambers DC, B.D., Ayres JG. Chronic respiratory symptoms, von Willebrand factor and longitudinal decline in FEV_1. Respir Med 1999;93:726–33.

3 Beeckman LA, W.M., Petsonk EL, Wagner GR. Rapid declines in FEV_1 and subsequent respiratory symptoms, illnesses, and mortality in coal miners in the United States. Am J Respir Crit Care Med 2001;163:633–9.

4 Pelkonen M, N.I., Tukiainen H, Tervahauta M, Tuomilehto J, Nissinen A. Smoking cessation, decline in pulmonary function and total mortality: a 30 year follow up study among the Finnish cohorts of the Seven Countries Study. Thorax 2001;56:703–7.

5 Gauderman WJ, et al. The effect of air pollution on lung development from 10 to 18 years of age. N Engl J Med 2004;351:1057–67.

6 Martin BW, A.-L.U., Leuenberger P, Kunzli N, Stutz EZ, Keller R, Zellweger JP, et al. SAPALDIA: methods and participation in the cross-sectional part of the Swiss Study on Air Pollution and Lung Diseases in Adults. Soz Praventivmed 1997;42:67–84.

7 Burney PG, L.C., Chinn S, Jarvis D. The European Community Respiratory Health Survey. Eur Respir J 1994;7:954–60.

8 Ackermann-Liebrich U, K.-D.B., Probst-Hensch NM, et al. Follow-up of the Swiss Cohort Study on Air Pollution and Lung Diseases in Adults (SAPALDIA 2) 1991–2003: methods and characterization of participants. Soz Präventivmed 2005;50:1–19.

9 The ECRHS II Steering Committee, The European Community Respiratory Health Survey II. Eur Respir J 2002;20:1071–9.

10 Ackermann-Liebrich U, L.P., Schwartz J, Schindler C, Monn C, Bolognini G, Bongard JP, et al. Lung function and long term exposure to air pollutants in Switzerland. Study on Air Pollution and Lung Diseases in Adults (SAPALDIA) Team. Am J Respir Crit Care Med 1997;155:122–9.

11 Künzli N, A.-L.U., Keller R, Perruchoud AP, Schindler C. Variability of FVC and FEV_1 due to technician, team, device and subject in an eight centre study: three quality control studies in SAPALDIA. Swiss Study on Air Pollution and Lung Disease in Adults. Eur Respir J 1995;8;371–6.

12 American Thoracic Society, Standardization of spirometry: 1994 update. Am J Respir Crit Care Med 1995;152):1107–36.

13 Nelson SB, G.R., Crapo RO, Jensen RL. Performance evaluation of contemporary spirometers. Chest 1990;97:288–97.

14 Linn WS, S.J., Gong H Jr, Avol EL, Navidi WC, Peters JM. Standardization of multiple spirometers at widely separated times and places. Am J Respir Crit Care Med 1996;153:1309–13.

15 Johnson LR, E.P., Voelker HT, Tashkin DP. Volume spirometers need automated internal temperature sensors. Am J Respir Crit Care Med 1994;150:1575–80.

16 Gilliland FD, L.W., Rappaport E, Avol E, Gong H, Peters J. Effect of spirometer temperature on FEV_1 in a longitudinal epidemiological study. Occup Environ Med 1999;56:718–20.

17 Viegi G, S.M., Pistelli F, Englert N, Salonen R, Niepsuj G, Haluszk J, Roeme W. Inter-laboratory comparison of flow-volume curve measurements as quality control procedure in the framework of an international epidemiological study (PEACE project). Respir Med 2000;94:194–203.

18 Enright PL, J.L., Connett JE, Voelker H, Buist AS. Spirometry in the Lung Health Study. 1. Methods and quality control. Am Rev Respir Dis 1991;143:1215–23.