Appendix 2 (as supplied by the authors): Statistical considerations

Table 1. Notation for the analysis of a parallel group randomised trial where the same measurement scale outcome is evaluated by both blind and nonblind assessors.

Experimental treatment (*T*)		Blind average (X) and standard deviation (S_x)		**Control treatment** (*t*)		Blind average (X) and standard deviation (S_x)
	(X_i, Y_i), $i = 1, \ldots, M$	X ; S_X			(x_j, y_j), $j = 1, \ldots, m$	x ; s_x
Nonblind average (Y) and standard deviation (S_y)	Y ; S_Y	Number of observations = M Degrees of freedom (d.f.) = $M - 1$		Nonblind average and standard deviation (S_y)	y ; s_y	Number of observations = m Degrees of freedom (d.f.) = $m - 1$

A standardised mean difference (SMD) is an effect estimate, such as $X - x$, divided by a purposely selected standardisation standard deviation (SD), generically named S. Likewise, a dSMD is a differential effect estimate (here, a bias estimate), $B = (Y - y) - (X - x)$, divided by S. The S's used in this study are

$$\text{blind control } S = s_x, \quad m - 1 \text{ d.f.}, \tag{A1}$$

$$\text{blind pooled } S = (S_X \text{ pooled with } s_x), \quad M + m - 2 \text{ d.f.}, \tag{A2}$$

with pooling in terms of the usual d.f.-weighted mean of the variances, and:

$$\text{nonblind pooled } S = (S_Y \text{ pooled with } s_y), \quad M + m - 2 \text{ d.f.} \tag{A3}$$

Whenever, as here, S is an observed SD and a Gaussian error distribution is pragmatically assumed, the impact of its inherent randomness is completely determined by the associated degrees of freedom (d.f.), generically called f. Notably, the resulting SMD or dSMD is beset with a statistical bias. It arises because $1/S$ is a biased standardisation factor (the usual SD formula ensures that S^2 is unbiased; other powers of S are not). In Gaussian theory it is possible to devise as corrected standardisation factor, $(1/S)c$, where the correction $c = c(f) \sim 1 - 3/(4f - 1)$; in practice, it is close to 1.0. The resulting dSMD thus becomes:

$$\text{dSMD} = (B/S)c\,. \tag{A4}$$

By the theory of Gaussian models, B and S will be stochastically independent, no matter how S is chosen, so the squared standard error of this dSMD will consist of a term reflecting the conventional standard error of B and one reflecting the variability of S. Approximately

$$\text{SE}^2\{\text{dSMD}\} = [\text{SE}^2\{B\} \cdot S^{-2} + B^2 \cdot [S^{-2}\, h]]c^2\,, \tag{A5}$$

where $h = h(f) \sim 1/(2f - 7)$ represents the variability (relative variance) of a reciprocal SD estimate on f degrees of freedom. With the pertinent Student-t for the bias test available, *viz.* $t = B/\text{SE}\{B\}$, the formula simplified into:

$$\text{SE}^2\{\text{dSMD}\} = \text{SE}^2\{B\} \cdot S^{-2}\,[1 + t^2\, h]c^2\,. \tag{A6}$$

When the further assumption is made that the SDs of the variables X, Y, x and y are identical and estimated by S, then the conventional $\text{SE}^2\{B\}$ estimate is

Appendix to: Hróbjartsson A, Thomsen A, Emanuelsson F, et al. Observer bias in randomized clinical trials with blinded and nonblinded assessors of measurement scale outcomes: a systematic review. *CMAJ* 2013. DOI:10.1503/cmaj121505.

$$S^2[(2/M + 2/m)(1 - r)] , \quad \text{(A7)}$$

where r represents the estimated blind-nonblind correlation, assumed to be the same in the two treatment groups, i.e., $\rho(X_i, Y_i) = \rho(x_j, y_j)$. Consequently, one obtains a still simpler SE formula,

$$SE^2\{dSMD\} = (2/M + 2/m)(1 - r)[1 + t^2 \cdot h]c^2 , \quad \text{(A8)}$$

in which the two right-most factors are close to 1.0 in most cases.

The approach described above was adjusted in two situations. First, in two trials, Meltzer 2003 (1,2) and Landsman 2010 (3) there was a discrepancy between the number of patients assessed blind and nonblind in the experimental group (M), or between the number of patients assessed blind and nonblind in the control group (m). We derived at an approximate M and m for both trials by choosing the lowest number of patients. Second, in one sensitivity analysis we standardised by different factors in the SMD based on the blind and nonblind assessments: the pooled blind SD was used to calculate SMD_{blind} and the pooled nonblind SD to calculate $SMD_{nonblind}$.

Normally, the ideal situation would be to have access to correlation coefficients between blind and nonblind assessments, thus improving precision. However, correlations between blind and non-blind assessments may differ in trials with bias compared with those without bias, as a high degree of bias will theoretically induce a low degree of correlation. Thus, for a trial with no observer bias to have the same correlation coefficient as an otherwise identical trial with a high degree of observer bias the latter needs to have a compensatory decrease in the degree of random disagreement. Therefore, if the correlations were included in the main analysis, this could inappropriately impact on the meta-analytic estimate with trials with no bias receiving too much weight.

Cross-over or split body designed trials

A SE can be calculated that takes account of the paired trial design when information on the SD for the paired difference ($d_i = X_i - x_i$) is available (SD_d), either based on the blind assessor or the nonblind assessor, and under the assumption that the SD_d is identical between control group and experimental group, and between the blind and nonblind assessors:

$$SE\{B\} = sqrt(2SD_d^2/M) \quad \text{(A9)}$$

$$SE\{dSMD\} = SE\{B\}^2 (1 + t^2 h)/S^2 \quad \text{(A10)}$$

The additional correlation between blind and nonblind assessments can be accounted for when individual patient data is available.

Conversion of data for calculation of standardised mean difference

In 10 parallel group trials we had access to complete data: post-treatment means, corresponding standard deviations and number of patients included in each group, both for the blind and for the nonblind assessments.

We had no access to the original standard deviations in Taber 1983 (4); however, standard deviations were reported in another similar study that had used a comparable scale (5).

Cohen 2004 (6,7) reported change from baseline, but not baseline values or post-treatment values. The transcripts of the FDA's General and Plastic Surgery Devices Panel of the Medical Devices Advisory Committee included the baseline means of the blind assessments and an exact P-value for assessment of nasolabial folds (7). The transcripts also contained information on the difference between the medians of the blind and nonblind baseline assessments. Based on this information we calculated approximate baseline means for the nonblind assessment. Furthermore, we assumed that the standard deviation at baseline and post-treatment for both blind and nonblind assessments were approximately the same. The difference in standardized mean difference, when based on approximated post-treatment values of nasolabial folds, was –1.10 (-1.30 to -0.90). The difference in standardized mean difference, when based on reported change from baseline (overall assessment) was similar, –1.02 (–1.28 to –0.78).

Appendix to: Hróbjartsson A, Thomsen A, Emanuelsson F, et al. Observer bias in randomized clinical trials with blinded and nonblinded assessors of measurement scale outcomes: a systematic review. *CMAJ* 2013. DOI:10.1503/cmaj121505.

Narins 2010 (8) reported the post-treatment mean and the baseline means for the blinded assessment, along with complete data for change from baseline. We calculated the post-treatment values of the nonblind assessment, based on the assumption that the standard deviation for the post-treatment nonblind assessment was approximately the same as for the blind assessment. The difference in standardized mean difference, based on approximated post-treatment values, was –0.11 (–0.29 to 0.07). When based on the reported change from baseline values the result was very similar, –0.07 (–0.33 to 0.19).

Handling individual patient data
We received individual patient data from two trials: Noseworthy 1994 (9) and Herberger 2011 (10). In Noseworthy 1994 all randomised patients were accounted for. We chose as our preferred time of assessment the first post-treatment time point for which all randomised patients could have been assessed (disregarding drop-outs), and analysed post-treatment assessments of "Expanded Disability Status Scale (EDSS) at 12 months". In Herberger 2011, we encountered discrepancies between the results in the published paper and the file the authors sent us. We excluded the trial from our main analysis (see below).

Cross-over trials and split-body trials
We included three trials that were not of parallel group design. Narins 2010 (8) and Miller 2003 (11) were split-body trials. Ulm 1999 was a cross-over trial (12).

Such designs usually imply a more precise estimate of treatment effect, as patients serve as their own controls, and hence a more precise estimate of the difference between treatment effects based on blind vs. nonblind outcome assessors. This however, requires the reporting of the SD of the paired difference (SD_d), which is often missing. Our predefined analysis plan was to treat the data as deriving from parallel group trials, acknowledging that the trials would have less weight in the meta-analysis. The rationale for this was partly pragmatic as we had not expected to have SD_d, and partly for comparative reasons, because we wanted our approach to be similar to that of our previous analysis of binary data.

Unexpectedly, we did access SD_d for all three trials. For Ulm 1999 (12), we used SD_d from a roughly comparable trial using the same scale (13). For Narins 2010 (8) a confidence interval for the paired difference was reported, and for Miller 2003 (11) the exact *P*-value from a paired t-test was reported (though for a different time point), from which we derived at SD_d. We used this information in a sensitivity analyses where we adjusted the SE(dSMD) in the three non-parallel group trials.

References

1) Meltzer HY, Alphs L, Green AI, Altamura AC, Anand R, Bertoldi A, et al. Clozapine treatment for suicidality in schizophrenia: International Suicide Prevention Trial (InterSePT). Arch Gen Psychiatry 2003;60:82-91.
2) Also data from FDA Statistical review and evaluation [InterSePT].www.fda.gov/ohrms/dockets/ac/02/briefing/3908B1_02_E- %20Statistical%20Review.pdf. Accessed October 25 2010.)
3) Landsman AS, Robbins AH, Angelini PF, Wu CC, Cook J, Oster M, et al. Treatment of mild, moderate, and severe onychomycosis using 870- and 930-nm light exposure. J Am Podiatr Med Assoc 2010;100:166-77.
4) Taber LH, Knight V, Gilbert BE, McClung HW, Wilson SZ, Norton HJ, Thurson JM, Gordon WH, Atmar RL, Schlaudt WR. Ribavirin aerosol treatment of bronchiolitis associated with respiratory syncytial virus infection in infants. Pediatrics. 1983 Nov;72(5):613-8.
5) Taylor CE, Webb MS, Milner AD, Milner PD, Morgan LA, Scott R, Stokes GM, Swarbrick AS, Toms GL. Interferon alfa, infectious virus, and virus antigen secretion in respiratory syncytial virus infections of graded severity. Arch Dis Child. 1989 Dec;64(12):1656-60.
6) Cohen SR, Holmes RE. Artecoll: a long-lasting injectable wrinkle filler material: Report of a controlled, randomized, multicenter clinical trial of 251 subjects. Plast Reconstr Surg. 2004 Sep 15;114(4):964-76; discussion 977-9.
7) FDA Summary of safety and effectiveness data; www.accessdata.fda.gov/cdrh_docs/pdf2/P020012b.pdf. Accessed August 5th, 2011.
8) Narins RS, Coleman W, Donofrio L, Jones DH, Maas C, Monheit G, et al. Nonanimal sourced hyaluronic acid–based dermal filler using a cohesive polydensified matrix technology is superior to bovine collagen in the

correction of moderate to severe nasolabial folds: results from a 6-month, randomized, blinded, controlled, multicenter study. Dermatol Surg 2010;36:730-40
9) Noseworthy JH, Vandervoort MK, Penman M, Ebers G, Shumak K, Seland TP, et al. Cyclophosphamide and plasma exchange in multiple sclerosis. Lancet 1991;337:1540-1.
10) Herberger K, Franzke N, Blome C, Kirsten N, Augustin M. Efficacy, Tolerability and Patient Benefit of Ultrasound-Assisted Wound Treatment versus Surgical Debridement: A Randomized Clinical Study. Dermatology. 2011;222(3):244-9.
11) Miller RS, Steward DL, Tami TA, Sillars MJ, Seiden AM, Shete M, et al. The clinical effects of hyaluronic acid ester nasal dressing (Merogel) on intranasal wound healing after functional endoscopic sinus surgery. Otolaryngol Head Neck Surg 2003;128:862-9.
12) Ulm G, Schüler P. Cabergolin versus pergolid: a video-blinded, randomised multicenter cross-over study. Akt Neurolgie 1999;25:360-65.
13) Navan P, Findley LJ, Jeffs JA, Pearce RK, Bain PG. Double-blind, single-dose, cross-over study of the effects of pramipexole, pergolide, and placebo on rest tremor and UPDRS part III in Parkinson's disease. Mov Disord. 2003 Feb;18(2):176-80.

Appendix to: Hróbjartsson A, Thomsen A, Emanuelsson F, et al. Observer bias in randomized clinical trials with blinded and nonblinded assessors of measurement scale outcomes: a systematic review. *CMAJ* 2013. DOI:10.1503/cmaj121505.