Appendix 3 (as supplied by the authors): Supplementary results

Table 1: Risk of confounding and characteristics of the outcome assessment

Trial	No risk of confounding[1]					Characteristics of the outcome assessment[2]		
	Short time between assessments	Exactly same patients	Same type of assessors	Same type of assessment procedure	Effective blinding	Outcome subjectivity	Trial involvement of nonblind assessor	Vulnerability of outcome to nonblind patients
Cohen 2004 (1,2)	Yes	No	No	No	Yes	5	5	1
Oesterle 2000 (3)	Yes	Yes	Yes	No	No	5	4	5
Powell 2001 (4)	Yes	No	Yes	No	Yes	5	1	1
Burkhoff 1999 (5)	Yes	Yes	No	No	No	5	4	5
Wedekind 2006 (6)	Yes	No	Yes	Yes	No	5	5	5
Weaver 2009 (7)	Yes	Yes	Yes	Yes	Yes	5	5	4
Noseworthy 1994 (8)	Yes	Yes	Yes	Yes	No	5	5	5
Narins 2010 (9)	Yes	No	No	No	Yes	3	5	1
Ulm 1999 (10)	Yes	Yes	Yes	No	Yes	5	4	1
Meltzer 2003 (11,12)	Yes	No	Yes	Yes	No	5	5	5
Miller 2003 (13)	Yes	No	Yes	No	Yes	5	5	2
Taber 1983 (14)	Yes	No	Yes	Yes	Yes	3	5	3
Hylaform 2004 (15)	Yes	No	No	No	Yes	5	5	1
Landsman 2010 (16)	Yes	No	No	No	Yes	5	3	1
Iglesia 2010 (17)	Yes	Yes	No	Yes	Yes	3	5	1
Reddihough 2004 (18)	Yes	No	Yes	No	Yes	4	4	3

[1]A "Yes" indicates that the trial is considered free from the potential confounding factor in question. Evaluations were conducted by authors masked to the comparison between blind vs. nonblind trial results (except for the evaluation of whether the patients in both type of assessments were exactly the same). [2]Masked evaluation on 1-5 Likert scales (5 signify a high degree of, e.g. outcome subjectivity).

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.

Table 2: Correlation and inter-observer agreement

Trials with blind & nonblind assessment of a measurement scale outcome			Studies validating the scale used in the trials[1]		
		Correlation blind/nonblind	Inter-observer agreement		Characteristics of the scale validation study
Trial	Outcome	rho[2]	Weighted Kappa[3]	ICC[4]	
Cohen 2004 (1,2)	Facial Fold Assessment Scale (0-5)	0.50	na[5]	0.89[6]	8 observers assessed 80 photographs of persons with facial wrinkles (19)
Oesterle 2000 (3)	CCSA[7] class (I-IV)	0.58	0.60	na	2 physicians assessed 75 patients referred for treadmill exercise tolerance testing (20)
Powell 2001 (4)	Nasal obstruction (10 cm VAS[8])	na	na		
Burkhoff 1999 (5)	CCSA[6] class (I-IV)	0.54	0.60	na	2 physicians assessed 75 patients referred for treadmill exercise tolerance testing (20)
Wedekind 2006 (6)	PAS[9] (0-52)	0.91	(0.78)[10]		2 psychiatrists assessed 23 patients with panic disorder in out-patient unit (21)
Weaver 2009 (7)	UPDRS III[11] (0-108)	na	na	0.82	2 neurologists assessed 24 patients with Parkinson's in community-based study (22)
Nosworthy 1994 (8)	EDSS[12] (0-10)	0.89	na	0.78	2 neurologists assessed 125 patients with multiple sclerosis (23)
Narins 2010 (9)	Wrinkle Severity Rating Scale (0-4)	na	0.75	na	5 observers assessed 30 photographs of persons with facial wrinkles (24)
Ulm 1999 (10)	UPDRS III[11] (0-108)	na	na	0.82	2 neurologists assessed 24 patients with Parkinson's in community-based study (22)
Meltzer 2003 (11,12)	CGI-SS[13]	na	na		
Taber 1983 (14)	Illness severity score (0-3), day 1	na			
Miller 2003 (13)	Synechia (0-3), last follow-up	na			
Hylaform 2004 (15)	Severity Grading Scale (0-5)	na	0.85	na	13 observers assessed 32 persons with nasolabial folds (25)
Landsman 2010 (16)	Clinical assessment scale (0-3)	na	na		
Iglesia 2010 (17)	POP-Q[14] (0-IV)	0.67	0.64	na	2 observers assessed 45 women attending outpatient (uro)gynaecology clinics (26)
Reddihough 2004 (18)	GMFM[15] (0-264)	0.89[16]	na	0.99	2 observers assessed 26 children with cerebral palsy (27)

[1]The identification of the validation studies was based on a literature search on PubMed, and was not the result of a systematic review. [2]Spearman's correlation coefficient. [3]Weigted Kappa or mean Weighted Kappa (confidence intervals not accessible); [4]Intra-class correlation coefficient (95% confidence intervals only accessible in Reddihough 2004: (0.97-0.99); [5]not accessible; [6]Blind outcome assessors in Cohen 2004 were reported to have ICC > 0.80; [7]Canadian cardiovascular society (grading of) angina pectoris; [8]Visual analogue scale; [9]Panic and agoraphobia scale;[10]Unclear what measure was used; [11]Unified Parkinson's Disease Rating Scale;[12]Expanded disability status scale;[13]Clinical global impression on suicide severity scale (7 point version); [14]Pelvic organ prolapse quantification exam; [15]Gross Motor Function Measure;[8]Concordance correlation coefficient ranged from 0.78 to 0.99. We assumed that the nearly identical Spearman's correlation coefficient was 0.89.

Qualitative summary of results in trials with incomplete or unclear data:
Eight included trials had incomplete or unclear outcome data. Qualitative information, or results from other similar trials, indicated notable observer bias in three of these trials. Ash 1998 (28) was a split-body trial comparing the effect of tretinoin vs. L-ascorbic acid on 10 patients with striae alba. Assessment of percentage improvement was conducted by both blind and nonblind assessor. The means of the improvements were reported, and showed a clear difference favouring the nonblind assessment, but SD was not reported. We obtained a SD from a somewhat related trial (29), but it was debatable whether the trials were truly comparable. However, based on the SD, we derived at a tentative estimate dSMD of -0.81, indicating a substantial degree of observer bias.

Purdue 1997 (30) was a split-body trial comparing biosynthetic skin replacement vs. cadaver skin for burn wounds in 66 patients. Wound healing (percentage "take of autograft") was assessed by blind and nonblind assessors. The mean of the percentage take was reported inconsistently for the nonblind assessment, and SDs were not reported, so we decided to regard the data as too uncertain for inclusion in our main analysis. Based on the most conservative assessment and SD from a similar trial (31), we derived at a tentative dSMD of -0.48, indicating a substantial degree of observer bias.

Kadish 2011 (32) was a parallel group trial comparing cardiac contractility modulation vs. standard treatment in 428 patients with advanced heart failure. Assessment of NYHA class was conducted both by blind and nonblind assessors. The trial publication did not report the mean values of blind vs. nonblind assessments, and the company was unwilling to share the data. However, one of the authors, Dr. Daniel Burkhoff, informed us that there was a marked difference between effect estimates based on blind and nonblind assessment, similar to that of the two other cardiological trials included (3,5), with a pooled dSMD -0.56, indicating a substantial degree of observer bias.

There was indication of no or little bias in two trials. Bauman 2007 (33) was a large parallel group trial comparing three types of hyaluronic acid dermal fillers vs. collagen in 439 patients. Assessments of nasiolabel folds on the wrinkle assessment scale were conducted by blind and nonblind observers. Incomplete data based on two of the three intervention groups (277 patients) provided a tentative dSMD of -0.06, indicating low degree of observer bias.

Herberger 2011 (34) was a parallel group trial comparing the effect of ultrasound-assisted wound treatment vs. surgical debridement in 67 patients. Assessment of four dimensions of wound status was conducted by blind and nonblind assessors on a five point response scale. The authors provided us with individual patient data. We were unable to reproduce the published blinded results and number of patients was not identical. Further contact to the authors did not resolve the issue, so we decided to regard the data as too uncertain for inclusion in our main analysis. Our tentative estimate of dSMD was 0.12, indicating bias in the reversed direction.

In the remaining three trials there was no information on how disagreements between blind and nonblind assessors affected estimated intervention effects. Realmuto 1984 (35) was a parallel group trial comparing the effect of thiothixene vs. thioridazine in 21 schizophrenic adolescents. The blind and nonblind ratings on the Brief Psychiatric Rating Scale and Clinical Global Impression Scale "did not show statistically significant differences". Havel 1999 (36) was a parallel group trial comparing the effect of propofol vs. midazolam for procedural sedation in 91 patients. The weighted kappa value between the blind and nonblind ratings of Ramsey Sedation scores was reported as 0.74. Alam 2006 (37) was a split-body trial comparing the effect of single pulse dye laser treatment v no-treatment for scar appearance in 20 patients. The ratings of the blind and nonblind assessors of overall scar appearance were "highly correlated".

Conversion of dSMD to ratio of odds ratios (ROR)

To facilitate the comparison between the degree of observer bias in trials with measurement scale outcomes and binary outcomes we converted all SMDs to ORs, and subsequently all dSMDs to ROR. This conversion is based on assumptions of equal variance and logistic distributions of the measurements in each group (38).

The pooled result of all 16 trials produced an I^2 of 46% (P = 0.02), and a ratio of odds ratio of 0.66 (0.48 to 0.90).

An ROR < 1 indicates that nonblind outcome assessors generate a more optimistic estimate of the treatment effect than blind outcome assessors.

The result is coherent with that of our previous analyses of observer bias in 21 trials with binary outcomes, providing a pooled ratio of odds ratios of 0.64 (0.43 to 0.96) (39).

References

1. 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.

2. FDA Summary of safety and effectiveness data; www.accessdata.fda.gov/cdrh_docs/pdf2/P020012b.pdf. Accessed August 5th, 2011.

3. Oesterle SN, Sanborn TA, Ali N, Resar J, Ramee SR, Heuser R, et al. Percutaneous transmyocardial laser revascularisation for severe angina: the PACIFIC randomised trial. Potential Class Improvement From Intramyocardial Channels. Lancet 2000;356:1705-10.

4. Powell NB, Zonato AI, Weaver EM, Li K, Troell R, Riley RW, Guilleminault C. Radiofrequency treatment of turbinate hypertrophy in subjects using continuous positive airway pressure: a randomized, double-blind, placebo-controlled clinical pilot trial. Laryngoscope. 2001 Oct;111(10):1783-90.

5. Burkhoff D, Schmidt S, Schulman SP, Myers J, Resar J, Becker LC, et al. Transmyocardial laser revascularisation compared with continued medical therapy for treatment of refractory angina pectoris: a prospective randomised trial. ATLANTIC Investigators. Angina Treatments-Lasers and Normal Therapies in Comparison. Lancet 1999;354:885-90.

6. Wedekind D, Broocks A, Weiss N, Engel K, Neubert K, Bandelow B. A randomized, controlled trial of aerobic exercise in combination with paroxetine in the treatment of panic disorder. World J Biol Psychiatry. 2010 Oct;11(7):904-13.

7. Weaver FM, Follett K, Stern M, Hur K, Harris C, Marks WJ Jr, Rothlind J, Sagher O, Reda D, Moy CS, Pahwa R, Burchiel K, Hogarth P, Lai EC, Duda JE, Holloway K, Samii A, Horn S, Bronstein J, Stoner G, Heemskerk J, Huang GD; CSP468 Study Group. Bilateral deep brain stimulation vs best medical therapy for patients with advanced Parkinson disease: a randomized controlled trial. JAMA.2009 Jan 7;301(1):63-73.

8. 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.

9. 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

10. Ulm G, Schüler P. Cabergolin versus pergolid: a video-blinded, randomised multicenter cross-over study. Akt Neurolgie 1999;25:360-65.

11. 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.

12. 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.)

13. 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.

14. 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.

15. Hylaform 2004. FDA summary of safety and effectiveness data, premarket approval application P030032. http://www.accessdata.fda.gov/cdrh_docs/pdf3/P030032b.pdf (accessed August 5th 2011).

16. 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.

17. Iglesia CB, Sokol AI, Sokol ER, Kudish BI, Gutman RE, Peterson JL, Shott S.Vaginal mesh for prolapse: a randomized controlled trial. Obstet Gynecol. 2010 Aug;116(2 Pt 1):293-303.

18. Reddihough DS, King JA, Coleman GJ, Fosang A, McCoy AT, Thomason P, Graham HK. Functional outcome of botulinum toxin A injections to the lower limbs in cerebral palsy. Dev Med Child Neurol. 2002 Dec;44(12):820-7.

19. Lemperle G, Holmes RE, Cohen SR, Lemperle SM. A classification of facial wrinkles. Plast Reconstr Surg. 2001 Nov;108(6):1735-50; discussion 1751-2.

20. Goldman L, Hashimoto B, Cook EF, Loscalzo A. Comparative reproducibility and validity of systems for assessing cardiovascular functional class: advantages of a new specific activity scale. Circulation. 1981 Dec;64(6):1227-34.

21. Bandelow B. Assessing the efficacy of treatments for panic disorder and agoraphobia. II. The Panic and Agoraphobia Scale. Int Clin Psychopharmacol. 1995 Jun;10(2):73-81.

22. Richards M, Marder K, Cote L, Mayeux R. Interrater reliability of the Unified Parkinson's Disease Rating Scale motor examination. Mov Disord. 1994 Jan;9(1):89-91.

23. Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain. 2000 May;123 (Pt 5):1027-40.

24. Day DJ, Littler CM, Swift RW, Gottlieb S. The wrinkle severity rating scale: a validation study. Am J Clin Dermatol. 2004;5(1):49-52.

25. Monheit GD, Gendler EC, Poff B, Fleming L, Bachtell N, Garcia E, Burkholder D. Development and validation of a 6-point grading scale in patients undergoing correction of nasolabial folds with a collagen implant. Dermatol Surg. 2010 Nov;36 Suppl 3:1809-16.

26. Stark D, Dall P, Abdel-Fattah M, Hagen S. Feasibility, inter- and intra-rater reliability of physiotherapists measuring prolapse using the pelvic organ prolapse quantification system. Int Urogynecol J. 2010 Jun;21(6):651-6.

27. Brunton LK, Bartlett DJ. Validity and reliability of two abbreviated versions of the Gross Motor Function Measure. Phys Ther. 2011 Apr;91(4):577-88.

28. Ash K, Lord J, Zukowski M, McDaniel DH. Comparison of topical therapy for striae alba (20% glycolic acid/0.05% tretinoin versus 20% glycolic acid/10% L-ascorbic acid). Dermatol Surg. 1998 Aug;24(8):849-56.

29. Sadick NS, Magro C, Hoenig A. Prospective clinical and histological study to evaluate the efficacy and safety of a targeted high-intensity narrow band UVB/UVA1 therapy for striae alba. J Cosmet Laser Ther. 2007 Jun;9(2):79-83.

30. Purdue GF, Hunt JL, Still JM Jr, Law EJ, Herndon DN, Goldfarb IW, Schiller WR, Hansbrough JF, Hickerson WL, Himel HN, Kealey GP, Twomey J, Missavage AE, Solem LD, Davis M, Totoritis M, Gentzkow GD. A multicenter clinical trial of a biosynthetic skin replacement, Dermagraft-TC, compared with cryopreserved human cadaver skin for temporary coverage of excised burn wounds. J Burn Care Rehabil. 1997 Jan-Feb;18(1 Pt 1):52-7.

31. Schurr MJ, Foster KN, Centanni JM, Comer AR, Wicks A, Gibson AL, Thomas-Virnig CL, Schlosser SJ, Faucher LD, Lokuta MA, Allen-Hoffmann BL. Phase I/II clinical evaluation of StrataGraft: a consistent, pathogen-free human skin substitute. J Trauma. 2009 Mar;66(3):866-73; discussion 873-4.

32. Kadish A, Nademanee K, Volosin K, Krueger S, Neelagaru S, Raval N, Obel O, Weiner S, Wish M, Carson P, Ellenbogen K, Bourge R, Parides M, Chiacchierini RP, Goldsmith R, Goldstein S, Mika Y, Burkhoff D, Abraham

WT. A randomized controlled trial evaluating the safety and efficacy of cardiac contractility modulation in advanced heart failure. Am Heart J. 2011 Feb;161(2):329-337.e1-2.

33. Baumann LS, Shamban AT, Lupo MP, Monheit GD, Thomas JA, Murphy DK, Walker PS; JUVEDERM vs. ZYPLAST Nasolabial Fold Study Group. Comparison of smooth-gelhyaluronic acid dermal fillers with cross-linked bovine collagen: a multicenter, double-masked, randomized, within-subject study. Dermatol Surg. 2007 Dec;33 Suppl 2:S128-35.

34. 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.

35. Realmuto GM, Erickson WD, Yellin AM, Hopwood JH, Greenberg LM. Clinical comparison of thiothixene and thioridazine in schizophrenic adolescents. Am J Psychiatry. 1984 Mar;141(3):440-2.

36. Havel CJ Jr, Strait RT, Hennes H. A clinical trial of propofol vs midazolam for procedural sedation in a pediatric emergency department. Acad Emerg Med. 1999 Oct;6(10):989-97.

37. Alam M, Pon K, Van Laborde S, Kaminer MS, Arndt KA, Dover JS. Clinical effect of a single pulsed dye laser treatment of fresh surgical scars randomized controlled trial. Dermatol Surg. 2006 Jan;32(1):21-5.

38. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Stat Med. 2000 Nov 30;19(22):3127-31.

39. Hróbjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, Ravaud P, Brorson S. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ. 2012 Feb 27;344:e1119.