Does Reliable, Patient-Oriented Evidence Support the Request?

David A. Frenz, M.D.

13 March 2025

United Healthcare | Optum Health Education™

Learning Objectives

At the end of this educational activity, participants should be able to:

1. Appraise evidence using the Oxford Centre for Evidence-Based Medicine's Levels of Evidence (LOE) Taxonomy
2. Distinguish between disease-oriented outcomes (DOO) and patient-oriented outcomes (POO)
3. Determine whether reliable, patient-oriented evidence supports a treatment option or coverage for a requested health care service

Collision

“ What are we to do when the irresistible force of the need to offer clinical advice meets the immovable object of flawed evidence? All we can do is our best: give the advice, but alert the advisees to the flaws in the evidence on which it is based. ”

[Evidence-Based On-Call: Acute Medicine. Edinburgh: Churchill Livingstone, 2001, p. 641]

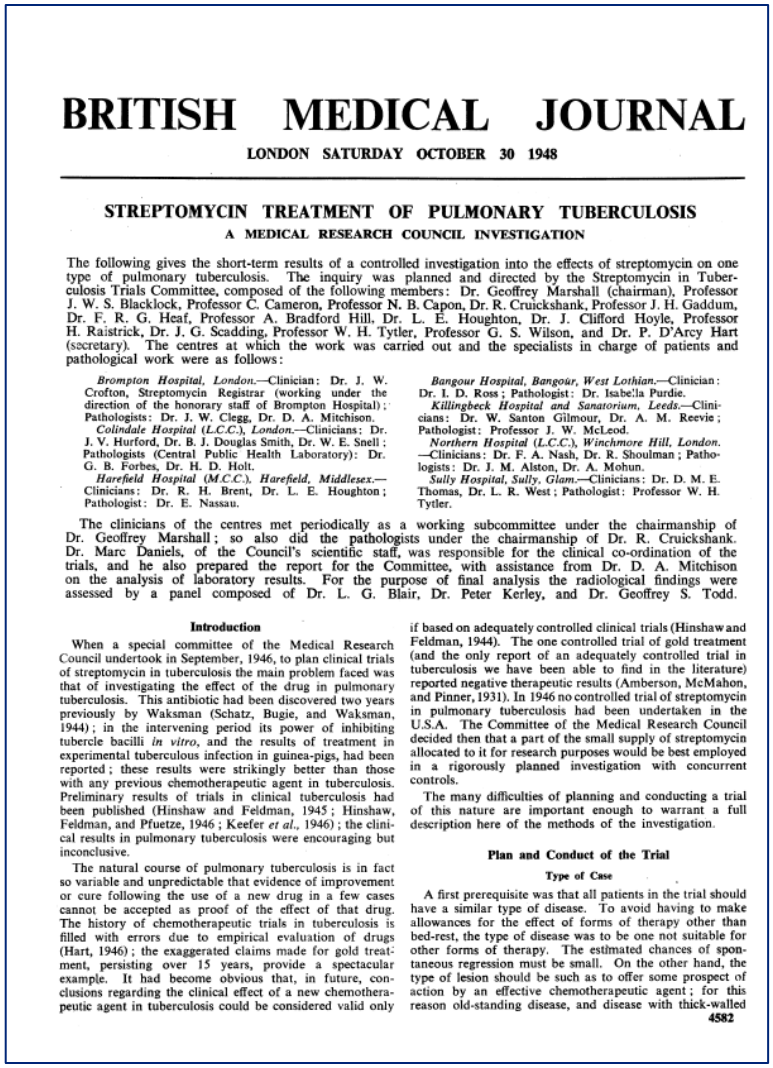
Reliable Evidence

Everyone Has Evidence

- The question is whether the evidence is sufficiently reliable for the clinical question or context

Evolution of Reliability

1. Mechanism-based reasoning
2. Observational evidence
3. Randomized controlled trials
4. Systematic reviews and meta-analyses



BRITISH MEDICAL JOURNAL

LONDON SATURDAY OCTOBER 30 1948

STREPTOMYCIN TREATMENT OF PULMONARY TUBERCULOSIS

A MEDICAL RESEARCH COUNCIL INVESTIGATION

The following gives the short-term results of a controlled investigation into the effects of streptomycin on one type of pulmonary tuberculosis. The inquiry was planned and directed by the Streptomycin in Tuberculosis Trials Committee, composed of the following members: Dr. Geoffrey Marshall (chairman), Professor J. W. S. Blacklock, Professor C. Cameron, Professor N. B. Capon, Dr. R. Cruickshank, Professor J. H. Gaddum, Dr. F. R. G. Heaf, Professor A. Bradford Hill, Dr. L. E. Houghton, Dr. J. Clifford Hoyle, Professor H. Raistrick, Dr. J. G. Scadding, Professor W. H. Tytler, Professor G. S. Wilson, and Dr. P. D'Arcy Hart (secretary). The centres at which the work was carried out and the specialists in charge of patients and pathological work were as follows:

Brompton Hospital, London.—Clinician: Dr. J. W. Crofton, Streptomycin Registrar (working under the direction of the honorary staff of Brompton Hospital): Pathologists: Dr. J. W. Clegg, Dr. D. A. Mitchison.

Colindale Hospital (L.C.C.), London.—Clinicians: Dr. J. V. Hurford, Dr. B. J. Douglas Smith, Dr. W. E. Snell; Pathologists (Central Public Health Laboratory): Dr. G. B. Forbes, Dr. H. D. Holt.

Harefield Hospital (M.C.C.), Harefield, Middlesex.—Clinicians: Dr. R. H. Brent, Dr. L. E. Houghton; Pathologist: Dr. E. Nassau.

Bangour Hospital, Bangour, West Lothian.—Clinician: Dr. I. D. Ross; Pathologist: Dr. Isabella Purdie.

Killingbeck Hospital and Sanatorium, Leeds.—Clinicians: Dr. W. Santon Gilmour, Dr. A. M. Reevie; Pathologist: Professor J. W. McLeod.

Northern Hospital (L.C.C.), Winchmore Hill, London.—Clinicians: Dr. F. A. Nash, Dr. R. Shoulman; Pathologists: Dr. J. M. Alston, Dr. A. Mohun.

Sully Hospital, Sully, Glam.—Clinicians: Dr. D. M. E. Thomas, Dr. L. R. West; Pathologist: Professor W. H. Tytler.

The clinicians of the centres met periodically as a working subcommittee under the chairmanship of Dr. Geoffrey Marshall; so also did the pathologists under the chairmanship of Dr. R. Cruickshank. Dr. Marc Daniels, of the Council's scientific staff, was responsible for the clinical co-ordination of the trials, and he also prepared the report for the Committee, with assistance from Dr. D. A. Mitchison on the analysis of laboratory results. For the purpose of final analysis the radiological findings were assessed by a panel composed of Dr. L. G. Blair, Dr. Peter Kerley, and Dr. Geoffrey S. Todd.

Introduction

When a special committee of the Medical Research Council undertook in September, 1946, to plan clinical trials of streptomycin in tuberculosis the main problem faced was that of investigating the effect of the drug in pulmonary tuberculosis. This antibiotic had been discovered two years previously by Waksman (Schatz, Bugie, and Waksman, 1944); in the intervening period its power of inhibiting tubercle bacilli *in vitro*, and the results of treatment in experimental tuberculous infection in guinea-pigs, had been reported; these results were strikingly better than those with any previous chemotherapeutic agent in tuberculosis. Preliminary results of trials in clinical tuberculosis had been published (Hinshaw and Feldman, 1945; Hinshaw, Feldman, and Pfuetze, 1946; Keefer *et al.*, 1946); the clinical results in pulmonary tuberculosis were encouraging but inconclusive.

The natural course of pulmonary tuberculosis is in fact so variable and unpredictable that evidence of improvement or cure following the use of a new drug in a few cases cannot be accepted as proof of the effect of that drug. The history of chemotherapeutic trials in tuberculosis is filled with errors due to empirical evaluation of drugs (Hart, 1946); the exaggerated claims made for gold treatment, persisting over 15 years, provide a spectacular example. It had become obvious that, in future, conclusions regarding the clinical effect of a new chemotherapeutic agent in tuberculosis could be considered valid only if based on adequately controlled clinical trials (Hinshaw and Feldman, 1944). The one controlled trial of gold treatment (and the only report of an adequately controlled trial in tuberculosis we have been able to find in the literature) reported negative therapeutic results (Amberson, McMahon, and Pinner, 1931). In 1946 no controlled trial of streptomycin in pulmonary tuberculosis had been undertaken in the U.S.A. The Committee of the Medical Research Council decided then that a part of the small supply of streptomycin allocated to it for research purposes would be best employed in a rigorously planned investigation with concurrent controls.

The many difficulties of planning and conducting a trial of this nature are important enough to warrant a full description here of the methods of the investigation.

Plan and Conduct of the Trial

Type of Case

A first prerequisite was that all patients in the trial should have a similar type of disease. To avoid having to make allowances for the effect of forms of therapy other than bed-rest, the type of disease was to be one not suitable for other forms of therapy. The estimated chances of spontaneous regression must be small. On the other hand, the type of lesion should be such as to offer some prospect of action by an effective chemotherapeutic agent; for this reason old-standing disease, and disease with thick-walled

[Br Med J 1948;2(4582):769-82. PMID: 18890300]

Smith vs. Jones

“ A common method of integrating several studies with inconsistent findings is to carp on the design or analysis deficiencies of all but a few studies—those remaining frequently being one’s own work or that of one’s students or friends—and then advance the one or two ‘acceptable’ studies as the truth of the matter. ”

[Educ Res 1976;5(10):3-8]

SPECIAL ARTICLE

EVIDENCE FAVORING THE USE OF ANTICOAGULANTS IN THE HOSPITAL PHASE OF ACUTE MYOCARDIAL INFARCTION

THOMAS C. CHALMERS, M.D., RAYMOND J. MATTA, M.D., HARRY SMITH, JR., PH.D., AND ANNE-MARIE KUNZLER, M.A.

Abstract Since the last comprehensive review of anticoagulation in acute myocardial infarction four additional randomized control trials have been reported. The overwhelming majority of all trials favored anticoagulation. Rates of thromboembolism were higher in the control, and hemorrhagic complications in the anticoagulated group. Pooling of all randomized control trials gives mean case fatality rates of 19.6 per cent for the control and 15.4 per cent for the anticoagulated group, a relative reduction of 21 per cent ($P<0.05$ or <0.001, depending on the analytic method). Five of six randomized control trials reported "no effect" because the difference favoring anticoagulation was not statistically significant. However, sample sizes in these "negative" papers were too small to protect against missing a 21 per cent reduction in true case fatality rate due to anticoagulation ($\beta>0.10$). All patients who present no specific contraindication should receive anticoagulants during hospitalization for infarction. (N Engl J Med 297:1091-1096, 1977)

SINCE the discovery of the first coumarin derivative by Link[1] in 1943, more clinical trials have been performed to evaluate the use of anticoagulants in acute myocardial infarction than in almost any other therapeutic situation.[2-33] Although the American Heart Association officially endorsed anticoagulation for all such patients in 1948,[34] there is no unanimity about their use at present. Currently, anticoagulant therapy in acute myocardial infarction is arbitrary, with most authorities generally recommending anticoagulants only for patients with large infarctions, heart failure or complications requiring prolonged bed rest.[35-39]

A detailed review published in 1969[40] concluded that methodologic difficulties in most trials precluded any definite decisions about efficacy. Since that was written four randomized control trials[30-33] have revealed a lower case fatality rate for the treated patients, although the difference was statistically significant in only one. Two large studies employing historical controls with attempted matching[18,19] have reported a highly significant decrease in case fatality rates by anticoagulants.

Another review of the available studies is in order. This paper has the dual purpose of assessing the technics and validity of the various therapeutic trials carried out in the last 30 years, and drawing conclusions from the data about the efficacy of anticoagulation. Review of 32 trials employing controls has revealed a marked preponderance of positive results, an inverse correlation between the size of the therapeutic benefit and the reliability of the trial as an experiment, and a statistically significant effect in favor of anticoagulants in the pooled randomized control trials that is apparently too small to be revealed by any but the largest of individual trials.

From the departments of Medicine and Biostatistics and the Office of the Dean, Mount Sinai School of Medicine of the City University of New York (address reprint requests to Dr. Chalmers at Mount Sinai School of Medicine of the City University of New York, Fifth Avenue and 100th St., New York, NY 10029).

MATERIALS AND METHODS

A search was made of the English-language periodical literature from 1948 through 1976. A computer search of the indexed literature was combined with a manual bibliographic search of all references found in review articles. Of over 150 articles reviewed, 32 studies were found in which anticoagulants were used as part of the therapeutic regimen for the treatment of acute myocardial infarction in the hospital and in which two essential items were included: a control group that received minimal or no anticoagulants, and the in-hospital case fatality rates. Any series with less than a total of 30 patients was omitted from consideration. All therapeutic regimens, coumarin and indandione derivatives alone, coumarin and indandione derivatives in conjunction with heparin and heparin alone, were analyzed. Only the most recent or most explicit article by a given author or group was used.

Most of the early studies excluded case fatalities that occurred soon after the diagnosis of infarction — i.e., within 24 to 72 hours. They reasoned that these deaths might be related to factors other than the use or nonuse of anticoagulants. When authors made no distinction between total and late fatality rates, we assumed that the published rates were total case fatality rates, not excluding any early deaths. When both were given they are listed in the tables. When the time of starting anticoagulants was given it was almost always within a few hours of the diagnosis. Follow-up observation of all patients in the reported studies lasted until the end of hospitalization for the infarction — a total of at least 21 days in all studies.

The 32 studies were grouped into two main categories: trials in which the controls were selected by a nonrandom procedure, and those in which treatment was assigned at random. The nonrandom trials were further subdivided into those with historical controls and those employing alternate or similarly collected simultaneous controls. As is customary, randomized control trial refers to the studies in which treatment was assigned at random.

The nonrandom trials with historical controls, 18 in number,[2-19] included studies in which the authors surveyed retrospectively what they considered to be comparable groups of treated and untreated patients. Patients were selected for treatment or nontreatment with anticoagulants on an ad hoc basis by the attending physicians, who usually were not the authors of the surveys. Often, the control patients in this group of studies were taken from record reviews. Not infrequently the controls were taken from different institutions.

The nonrandom trials with alternate or similarly collected simultaneous controls, eight in number,[20-27] were prospective trials in which the patients were allocated into treatment and nontreatment groups on the basis of even or odd days or by alternation of patients to the medical services on which therapy differed. If the decision for therapy was at random, the study was included in the category of random control trials.[28-33]

Case fatality rates, incidence of thromboembolism and incidence

[N Engl J Med 1977;297(20):1091-6. PMID: 909566]

The Mantel–Haenszel method yields a normal deviate z. Here, z = 2.44 (P<0.01). The conclusion applicable to all patients similar to those included in the trials is that the preponderance of evidence supports anticoagulation for the reduction of mortality, even though the reduction is small.

One further aspect of these published trials should be noted. The case fatality rates of the control patients were higher and the differences between control and treated larger for the nonrandom trials with historical controls and those with alternate or similar controls (largely conducted before 1962) than for the random control trials, five of which were carried out after 1962. While the quality of the studies was improving over time there was also a significant decrease in case fatality rates of the patients studied.

Table 2. Case Fatality Rates (CFR) According to Whether or Not the Early Deaths Were Excluded.

Study	Controls		Anticoagulated	
	Total CFR	CFR excluding early deaths*	Total CFR	CFR excluding early deaths*
Surveys employing historical controls:				
Greisman[2]	NA†	35.0 (48)	NA	9.3
Smith[3]	25.6	NA	14.3	NA
Furman[4]	NA	32.2 (24)	NA	18.0
Loudon[5]	40.8	NA	25.3	NA
Schnur[6]	NA	33.3 (24)	NA	36.8
Burton[7]	NA	31.0 (48)	NA	13.7
Manson[8]	NA	30.0 (24)	NA	14.0
Eastman[9]	42.2	28.2 (24)	23.2	18.7
Honey[10]	47.9	35.4 (48)	22.8	16.3
Rosenberg[11]	47.0	30.0 (48)	47.0	39.6
Toohey[12]	40.0	28.3 (48)	20.6	13.7
Richards[13]	NA	23.5 (24)	NA	17.1
Blake[14]	NA	28.6 (24)	NA	11.0
Griffith[15]	NA	54.4 (48)	NA	27.6
Gumpert[16]	47.1	10.0 (72)	28.7	24.4
Meltzer[17]	NA	31.6 (48)	NA	14.5
Modan[18]	27.3	17.2 (48)	8.2	6.4
Tonascia[19]	26.7	15.6 (48)	10.8	6.5
Mean ±1 SE	38.3 ±3.1	29.0 ±2.5	22.3 ±3.8	18.0 ±2.4
Studies employing alternately assigned controls:				
Wright[20]	23.9	NA	15.0	NA
Bresnick[21]	NA	12.5 (72)	NA	18.9
Tullock[22]	NA	40.5 (6)	NA	22.9
Holten[23]	NA	35.9 (24)	NA	22.4
Feldman[24]	30.3	24.3 (48)	30.3	20.9
Rashkoff[25]	33.3	26.2 (24)	22.4	12.7
McCluskie[26]	NA	39.1 (24)	NA	18.9
Hilden[27]	NA	25.4 (48)	NA	22.9
Mean ±1 SE	29.2 ±2.8	29.1 ±3.8	22.6 ±4.4	19.9 ±1.4
Randomized control trials:				
Carleton[28]	38.3	27.5 (48)	28.9	22.0
Wasserman[29]	21.4	11.3 (48)	15.6	14.5
MRC Co-op[30]	18.0	NA	16.2	NA
Drapkin[31]	21.2	NA	14.9	NA
Handley[32]	7.7	NA	7.4	NA
VA Co-op[33]	11.2	NA	9.6	NA
Mean ±1 SE	19.6 ±4.4	19.4	15.4 ±3.1	18.3

*Because of assumed delays in onset of anticoagulation the papers used various intervals in including early deaths. The intervals (in hours) are included in the parentheses as they were employed.
†Data not available.

The thromboembolism rate (Table 3) was significantly lower in the anticoagulated group in all the 22 studies in which it was reported. The average absolute decrease was 12.1 per cent in the nonrandom trials with historical controls, 10.7 per cent in those with alternate or similar controls and 7.9 per cent in the random control trials. The incidence of both minor and major hemorrhagic complications increased with anticoagulation (Table 4). A total of 13 deaths were reported to be a direct result of anticoagulation therapy, a number that is included in calculation of the reduction in deaths achieved by anticoagulation. However, none of these deaths occurred among the 1748 patients treated with anticoagulants in the random control trials. The reported causes of death from hemorrhage among control and treated patients in the random control trials are given in Table 5. The distribution of causes was similar in the control and treated patient groups.

Discussion

In view of the impressive results of so many clinical trials, why are anticoagulants not employed in all patients with acute myocardial infarction who do not have a specific contraindication, such as a cerebral hemorrhage, liver disease or peptic ulcer?

The most likely explanation lies in the poor quality

Table 3. Thromboembolism Rates in Every Study Reporting the Data.

Study	Rate (%)	
	Controls	Anti-coagulated
Surveys employing historical controls:		
Greisman[2]	21.0	4.0
Smith[3]	22.7	3.7
Furman[4]	19.4	14.0
Loudon[5]	21.6	14.7
Burton[7]	22.0	12.7
Manson[8]	36.0	23.8
Eastman[9]	13.7	3.5
Richards[13]	13.2	2.9
Meltzer[17]	23.7	5.5
Mean ±1 SE	21.5 ±2.2	9.4 ±2.4
Studies employing alternately assigned controls:		
Wright[20]	36.0	14.0
Bresnick[21]	25.0	21.3
Tullock[22]	28.6	12.9
Holten[23]	14.1	4.0
Feldman[24]	7.9	5.3
Rashkoff[25]	26.2	14.1
McCluskie[26]	33.9	18.0
Hilden[27]	13.5	10.2
Mean ±1 SE	23.2	12.5
Randomized control trials:		
Carleton[28]	30.0	24.4
Wasserman[29]	3.3	1.3
MRC Co-op[30]	24.1	14.5
Drapkin[31]	24.6	18.8
Handley[32]	26.9	0.0
VA Co-op[33]	18.8	7.8
Mean ±1 SE	21.3 ±3.9	11.1 ±4.0

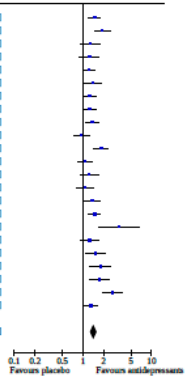
[N Engl J Med 1977;297(20):1091-6. PMID: 909566]

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1.9 Comparison 9 Vortioxetine versus placebo, Outcome 1 Rate of treatment response measured as a reduction of at least 50% on the HAM-A	1	610	Risk Ratio (IV, Random, 95% CI)	1.04 [0.85, 1.29]

Analysis 1.1. Comparison 1: Rate of treatment response measured as a reduction of at least 50% on the Hamilton Anxiety Scale (HAM-A), Outcome 1: Comparison 1 All antidepressants versus placebo, Outcome 1 Rate of treatment response measured as a reduction of at least 50% on the HAM-A

Study or Subgroup	Antidepressants Events	Total	Placebo Events	Total	Weight	Risk Ratio IV, Random, 95% CI
Allgulander 2001	273	399	60	130	5.4%	1.48 [1.22 , 1.81]
Allgulander 2004	102	182	55	188	4.6%	1.92 [1.48 , 2.48]
Bose 2008a	66	125	28	67	3.8%	1.26 [0.91 , 1.75]
Bose 2008b	65	125	28	67	3.8%	1.24 [0.90 , 1.73]
Bruenau-Mintzer 2006	97	164	78	162	5.3%	1.23 [1.00 , 1.51]
Davidson 1999	87	176	35	98	4.0%	1.38 [1.02 , 1.88]
Gommel 2015	103	198	82	197	5.2%	1.25 [1.01 , 1.55]
Hackett 2003	200	354	44	97	4.9%	1.25 [0.98 , 1.58]
Hartford 2007	165	326	60	161	5.0%	1.36 [1.08 , 1.70]
Kasper 2009	55	125	59	128	4.4%	0.95 [0.73 , 1.25]
Koponen 2007	193	338	54	175	4.8%	1.85 [1.46 , 2.35]
Lenox-Smith 2003	64	122	59	122	4.7%	1.08 [0.85 , 1.39]
Mahableshwarkar 2014b	76	149	32	77	4.0%	1.23 [0.90 , 1.67]
Mahableshwarkar 2014c	201	456	32	77	4.3%	1.06 [0.80 , 1.41]
Montgomery 2006	68	110	45	100	4.6%	1.37 [1.06 , 1.78]
Nicolini 2009	245	392	69	163	5.5%	1.48 [1.21 , 1.79]
Nimatoudis 2004	22	24	6	22	1.4%	3.36 [1.68 , 6.72]
Ryan 2008	67	168	51	159	4.2%	1.24 [0.93 , 1.67]
Stein 2008	45	63	27	58	3.9%	1.53 [1.12 , 2.11]
Stein 2014a	92	139	24	65	3.6%	1.79 [1.28 , 2.52]
Stein 2014c	89	139	24	65	3.6%	1.73 [1.23 , 2.44]
Stein 2017	164	268	32	140	3.9%	2.68 [1.95 , 3.68]
Wen-Yuan 2011	74	107	53	100	5.1%	1.30 [1.04 , 1.63]
Total (Wald4)		**4649**		**2618**	**100.0%**	**1.41 [1.29 , 1.55]**
Total events:	2613		1037			

Test for overall effect: Z = 7.35 (P < 0.00001)
Test for subgroup differences: Not applicable
Heterogeneity: Tau^2 (DL) = 0.03; Chi^2 = 58.91, df = 22 (P < 0.0001); I^2 = 63%



Footnotes

aStudy included twice as escitalopram and venlafaxine are different classes of antidepressants. Analyses vs placebo separated to avoid unit-of-analysis issues.
bStudy included twice as duloxetine and vortioxetine are different classes of antidepressants. Analyses vs placebo separated to avoid unit-of-analysis issues.
cStudy included twice as agomelatine and escitalopram are different classes of antidepressants. Analyses vs placebo separated to avoid unit-of-analysis issues.
dCI calculated by Wald-type method.
eTau² calculated by DerSimonian and Laird method.

[Cochrane Database Syst Rev 2025;1(1):CD012942. PMID: 39880377]

Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence

Question	Step 1 (Level 1*)	Step 2 (Level 2*)	Step 3 (Level 3*)	Step 4 (Level 4*)	Step 5 (Level 5)
How common is the problem?	Local and current random sample surveys (or censuses)	Systematic review of surveys that allow matching to local circumstances**	Local non-random sample**	Case-series**	n/a
Is this diagnostic or monitoring test accurate? (Diagnosis)	Systematic review of cross sectional studies with consistently applied reference standard and blinding	Individual cross sectional studies with consistently applied reference standard and blinding	Non-consecutive studies, or studies without consistently applied reference standards**	Case-control studies, or "poor or non-independent reference standard**	Mechanism-based reasoning
What will happen if we do not add a therapy? (Prognosis)	Systematic review of inception cohort studies	Inception cohort studies	Cohort study or control arm of randomized trial*	Case-series or case-control studies, or poor quality prognostic cohort study**	n/a
Does this intervention help? (Treatment Benefits)	Systematic review of randomized trials or *n*-of-1 trials	Randomized trial or observational study with dramatic effect	Non-randomized controlled cohort/follow-up study**	Case-series, case-control studies, or historically controlled studies**	Mechanism-based reasoning
What are the COMMON harms? (Treatment Harms)	Systematic review of randomized trials, systematic review of nested case-control studies, *n*-of-1 trial with the patient you are raising the question about, or observational study with dramatic effect	Individual randomized trial or (exceptionally) observational study with dramatic effect	Non-randomized controlled cohort/follow-up study (post-marketing surveillance) provided there are sufficient numbers to rule out a common harm. (For long-term harms the duration of follow-up must be sufficient.)**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning
What are the RARE harms? (Treatment Harms)	Systematic review of randomized trials or *n*-of-1 trial	Randomized trial or (exceptionally) observational study with dramatic effect			
Is this (early detection) test worthwhile? (Screening)	Systematic review of randomized trials	Randomized trial	Non -randomized controlled cohort/follow-up study**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning

*** Level may be graded down on the basis of study quality, imprecision, indirectness (study PICO does not match questions PICO), because of inconsistency between studies, or because the absolute effect size is very small; Level may be graded up if there is a large or very large effect size.**

**** As always, a systematic review is generally better than an individual study.**

How to cite the Levels of Evidence Table

OCEBM Levels of Evidence Working Group*. "The Oxford 2011 Levels of Evidence".
Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653

** OCEBM Table of Evidence Working Group = Jeremy Howick, Iain Chalmers (James Lind Library), Paul Glasziou, Trish Greenhalgh, Carl Heneghan, Alessandro Liberati, Ivan Moschetti, Bob Phillips, Hazel Thornton, Olive Goddard and Mary Hodgkinson*

Optum

InterQual®

Clinical Development Process 2024

InterQual Integrity Charter

Thousands of people in hospitals, health plans, and government agencies trust InterQual® evidence-based clinical decision support content to provide recommendations about the appropriateness and management of care and resource use including helping to facilitate equitable access to care. During the last four decades, InterQual has helped define and advance the disciplines of utilization and care management, providing medical directors, utilization management leaders, and other hospital and health plan professionals support in making the type of objective, evidence-based decisions that define top-quality, safe, equitable, and efficient care. This leads to greater transparency and collaboration between payers and providers.

The InterQual suite has expanded to 30 modules providing industry-leading objective clinical evidence and expert technology to help payers and providers collaborate for better healthcare outcomes at lower cost. While individual solutions meet key needs in a time of rapid industry change, the breadth of our portfolio allows healthcare organizations to combine our solutions in innovative ways to turn challenge into opportunity.

Our history of growth and innovation highlights our commitment to enhance InterQual through unbiased clinical content development and technology solutions. With over 4,000 InterQual customers relying on the guidance that InterQual provides, we take our responsibility to provide accurate, objective, evidence-based content very seriously. For over 40 years, our commitment is reflected in our rigorous evidence-based development process validated by our external peer reviewers, designed to protect against bias.

InterQual development process

InterQual is produced using a rigorous development process based on the principles of evidence-based medicine (EBM). InterQual clinical content is created by the Optum research and content development staff of over 55 research and clinical decision support specialists including physicians, registered nurses, physician assistant, nurse practitioners, social workers, physical and occupational therapists, and other healthcare professionals, including a medical librarian. The physicians' backgrounds include experience or specialization in internal medicine, neurology, psychiatry, substance use disorders, hospital medicine, pulmonary medicine, and critical care medicine. Most of the clinical staff hold advanced degrees (e.g., MD, DO, Masters, Ph.D.), certifications (e.g., nurse practitioner, physician assistant), and/or case management certification. All InterQual research and content development staff receive comprehensive, ongoing training at least quarterly in the concepts and methods of EBM and value-based clinical improvement to ensure that InterQual uses the best available evidence to support improved clinical decision-making, outcomes, quality, and value. New content staff receive comprehensive training in the principles of evidence-based medicine including the completion of 9 modules. Additionally, all staff participate in annual refresher training regarding mental health parity to reinforce that the processes, strategies, factors, and evidentiary standards are applied consistently and no more stringently between

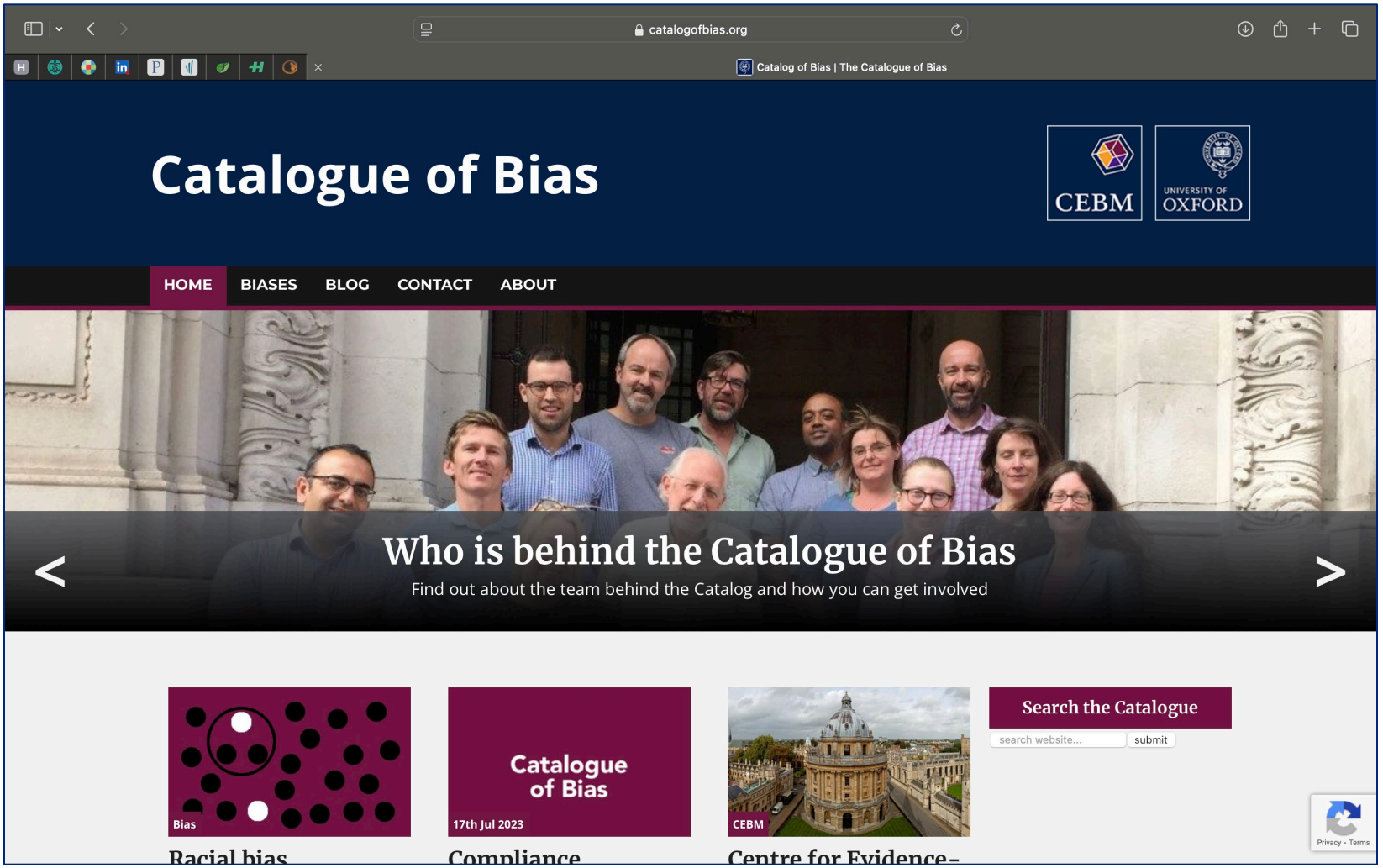
Classification	Type of Evidence
Class I	Meta-analysis, technology assessment, or systematic review
Class II	Randomized controlled clinical trial
Class III	Observational or epidemiologic study
Class IV	Evidence-based guideline
Class V	Expert opinion, panel consensus, literature review, text or reference book, descriptive study, case report, or case series

[InterQual, ibid.]

Free of Bias

“ Bias is defined as the systematic tendency of any factors associated with the design, conduct, analysis, evaluation and interpretation of the results of a study to make the estimate of the effect of a treatment or intervention deviate from its true value. ”

[Acta Obstet Gynecol Scand 2018;97(4):380-387. PMID: 29377058]

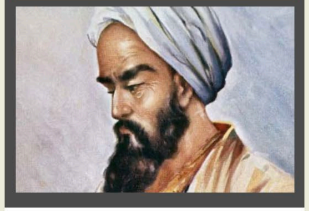


jameslindlibrary.org

The James Lind Library - Illustrating the development of fair tests of treatments in health care The James Lind Library

BROWSE THE LIBRARY

FAIR TESTS

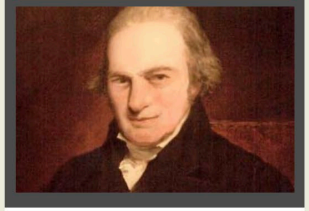


Despite acting with the best of intentions, health professionals have sometimes done more harm than good to the patients who have looked to them for help. Some of this suffering can be reduced by ensuring that *fair tests* are done to address uncertainties about the effects of treatments.

Sub-topics:

- *The need to address treatment uncertainties*
- *Treatment comparisons are essential*
- *Treatment comparisons must be fair*

BIASES

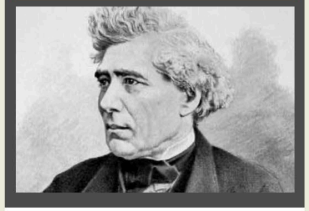


Biases in tests of treatments are those influences and factors that can lead to conclusions about treatment effects that are systematically different from the truth.

Sub-topics:

- *Design bias*
- *Allocation bias*
- *Co-intervention bias*
- *Observer bias*
- *Analysis bias*
- *Biases in judging unanticipated possible effects*
- *Reporting bias*
- *Biases in systematic reviews*
- *Researcher/sponsor bias and fraud*

PLAY OF CHANCE



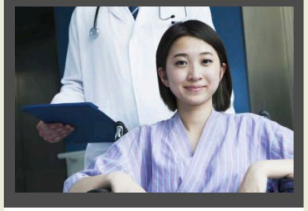
When treatments are compared, any differences in outcome events may simply reflect *the play of chance*.

Increasing the number of events studied in research reduces the likelihood of being misled in this way.

Sub-topics:

- *Recording and interpreting numbers*
- *Quantifying uncertainty*
- *Using meta-analysis*

SERVING PATIENTS



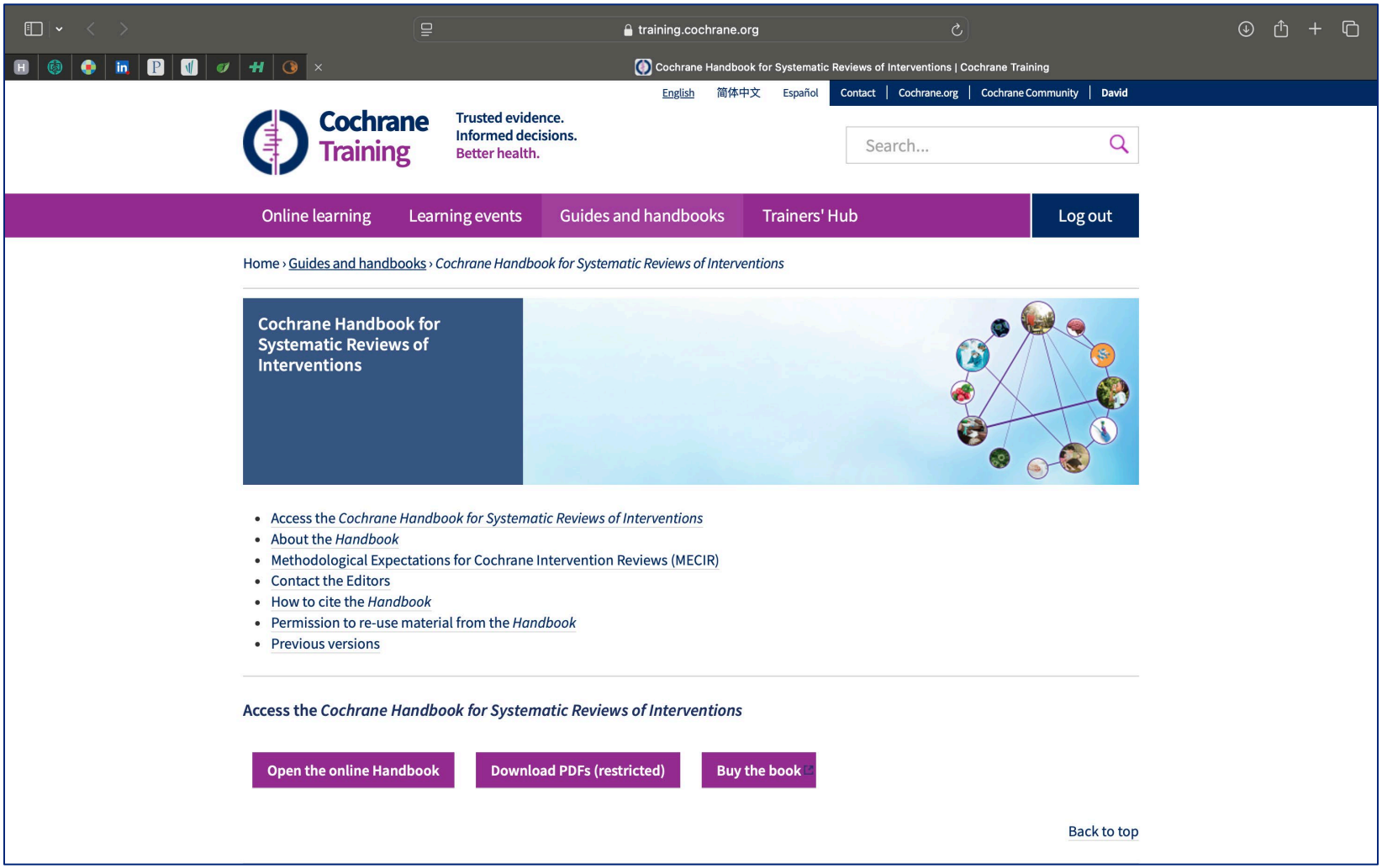
The interests of patients can be served by: improving reports of research, preparing and updating systematic reviews of reliable studies, and using these to inform decisions about treatment.

Sub-topics:

- *Improving reports of research*
- *Preparing and maintaining systematic reviews*
- *Using the results of systematic reviews*

SPECIALIST COLLECTIONS

You can also browse the Library by choosing topics of interest from the drop-down menus.



Five Major Sources of Bias

1. Bias arising from the randomization process
2. Bias due to deviations from intended interventions
3. Bias due to missing outcome data
4. Bias in measurement of the outcome
5. Bias in selection of the reported result

[Cochrane Handbook for Systematic Reviews of Interventions, v. 6.4 (2024), Table 8.2.a]

Random Allocation

“ Generation of random sequence should be done by some independent personnel, usually a statistician, who is not going to be involved in the conduct of the RCT. The access to this sequence should be restricted to only a few individuals who absolutely need to have access (such as the pharmacist who will be preparing the medication) and not the investigators or personnel involved in ascertaining outcome. ”

[Acta Obstet Gynecol Scand 2018;97(4):380-387. PMID: 29377058]

Allocation Concealment

“ This means that neither front-line care providers, investigators or participants are aware of whether the next eligible participant will be receiving treatment or control intervention. ”

[Acta Obstet Gynecol Scand 2018;97(4):380-387. PMID: 29377058]

Blinding

“ Unconscious information bias may be introduced if the investigators or participants are aware of who is getting the intervention and who is not. The procedure of blinding the participants (single blind) or both investigators and participants (double blind) helps to eliminate this unconscious information bias. ”

[Acta Obstet Gynecol Scand 2018;97(4):380-387. PMID: 29377058]

Study Conduct

“ The main premise of conducting an RCT is that the participants should be treated exactly the same way in both arms except for the intervention/control treatment. All other procedures of treatment, diagnosis, investigations, alterations etc. should follow the routine process and no undue advantage or testing should be performed on patients in the trial. These data should be collected to identify issues of contaminations, crossover of intervention and co-interventions. ”

[Acta Obstet Gynecol Scand 2018;97(4):380-387. PMID: 29377058]

Outcome Ascertainment

“ The prespecified primary and secondary outcomes should be collected by independent observers who are unaware of the allocation and treatment arms of participants. As far as possible, it is advisable that objective measures are used for ascertaining outcome so that personal bias on the part of the collector does not come into play. ”

[Acta Obstet Gynecol Scand 2018;97(4):380-387. PMID: 29377058]

Missing Data

“ It is also important that the outcome is collected in all randomized patients. The number of patients with missing outcome data should be minimized as far as possible. A high rate of attrition will lead to reduced confidence in the results and may lead to biased estimates. ”

[Acta Obstet Gynecol Scand 2018;97(4):380-387. PMID: 29377058]

Sample Size & Power

“ One would always like to conduct a study that has adequate sample size and power so that the conclusions generated from the experiment can be applied to the broader population with ample confidence. The required sample size to test a hypothesis is governed by the effect size. ”

[Acta Obstet Gynecol Scand 2018;97(4):380-387. PMID: 29377058]

Figure 2. Review authors' judgements about each risk of bias item presented as unclear (yellow), low (green), and high (red) across all included studies

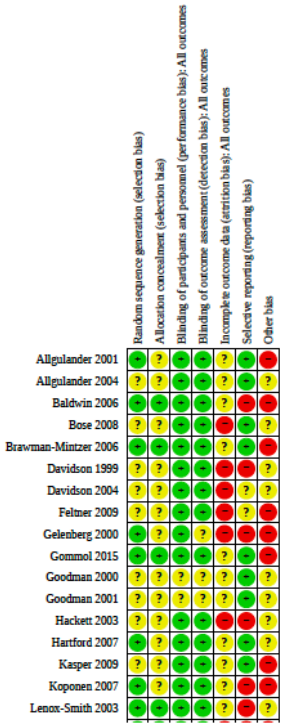


Figure 2. (Continued)

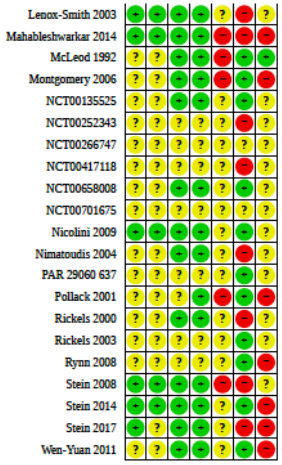
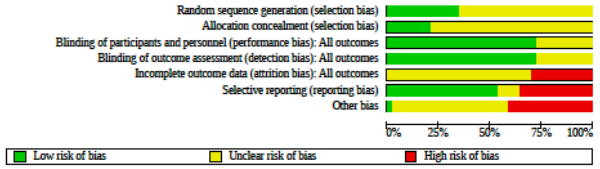


Figure 3. Review authors' judgements about each risk of bias item presented as percentages across all included studies



Allocation

Thirteen studies specified how their randomisation sequence was generated and were at low risk of bias, while random sequence generation was unclear for 24 studies. Eight studies specified the methods for allocation concealment and were at low risk of bias, while it was unclear in the remaining 29 studies.

[Cochrane Database Syst Rev 2025;1(1):CD012942. PMID: 39880377]

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Liraglutide for Children 6 to <12 Years of Age with Obesity — A Randomized Trial

Claudia K. Fox, M.D.,[1] Margarita Barrientos-Pérez, M.D.,[2] Eric M. Bomberg, M.D.,[1,3] John Dcruz, M.D.,[4] Inge Gies, Ph.D.,[5] Nina M. Harder-Lauridsen, Ph.D.,[6] Muhammad Yazid Jalaludin, M.D.,[7] Kushal Sahu, M.Sc.,[4] Petra Weimers, Ph.D.,[6] Thomas Zueger, M.D.,[8,9] and Silva Arslanian, M.D.,[10] for the SCALE Kids Trial Group*

ABSTRACT

BACKGROUND

No medications are currently approved for the treatment of nonmonogenic, nonsyndromic obesity in children younger than 12 years of age. Although the use of liraglutide has been shown to induce weight loss in adults and adolescents with obesity, its safety and efficacy have not been established in children.

METHODS

In this phase 3a trial, which consisted of a 56-week treatment period and a 26-week follow-up period, we randomly assigned children (6 to <12 years of age) with obesity, in a 2:1 ratio, to receive either once-daily subcutaneous liraglutide at a dose of 3.0 mg (or the maximum tolerated dose) or placebo, plus lifestyle interventions. The primary end point was the percentage change in the body-mass index (BMI; the weight in kilograms divided by the square of the height in meters). The confirmatory secondary end points were the percentage change in body weight and a reduction in BMI of at least 5%.

RESULTS

A total of 82 participants underwent randomization; 56 were assigned to the liraglutide group and 26 to the placebo group. At week 56, the mean percentage change from baseline in BMI was −5.8% with liraglutide and 1.6% with placebo, representing an estimated difference of −7.4 percentage points (95% confidence interval [CI], −11.6 to −3.2; P<0.001). The mean percentage change in body weight was 1.6% with liraglutide and 10.0% with placebo, representing an estimated difference of −8.4 percentage points (95% CI, −13.4 to −3.3; P=0.001), and a reduction in BMI of at least 5% occurred in 46% of participants in the liraglutide group and in 9% of participants in the placebo group (adjusted odds ratio, 6.3 [95% CI, 1.4 to 28.8]; P=0.02). Adverse events occurred in 89% and 88% of participants in the liraglutide and placebo groups, respectively. Gastrointestinal adverse events were more common in the liraglutide group (80% vs. 54%); serious adverse events were reported in 12% and 8% of participants in the liraglutide and placebo groups, respectively.

CONCLUSIONS

Among children (6 to <12 years of age) with obesity, treatment with liraglutide for 56 weeks plus lifestyle interventions resulted in a greater reduction in BMI than placebo plus lifestyle interventions. (Funded by Novo Nordisk; SCALE Kids ClinicalTrials.gov number, NCT04775082.)

The authors' affiliations are listed at the end of the article. Dr. Fox can be contacted at lusc0001@umn.edu or at the Center for Pediatric Obesity Medicine, Department of Pediatrics, University of Minnesota Medical School, 717 Delaware St. SE, Rm. 370 G, Minneapolis, MN 55414.

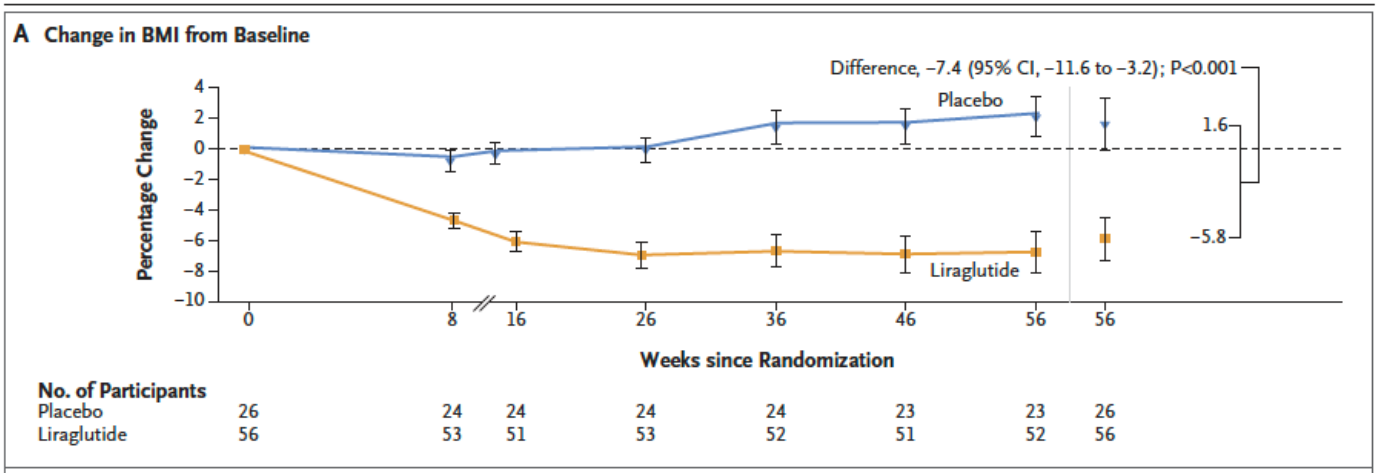
*A list of the investigators in the SCALE Kids trial is provided in the Supplementary Appendix, available at NEJM.org.

This article was published on September 10, 2024, and updated on January 7, 2025, at NEJM.org.

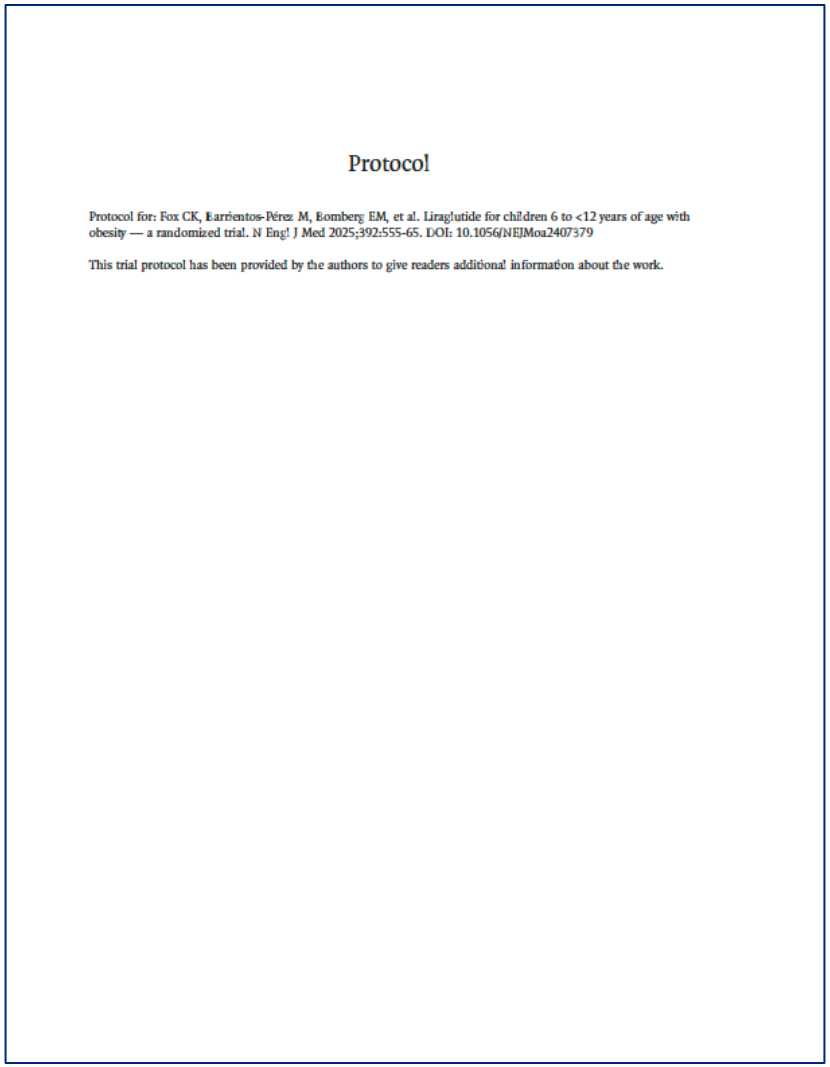
N Engl J Med 2025;392:555-65.
DOI: 10.1056/NEJMoa2407379

CME

[N Engl J Med 2025;392(6):555-565. PMID: 39258838]



[N Engl J Med 2025;392(6):555-565. PMID: 39258838]



Protocol

Protocol for: Fox CK, Barrientos-Pérez M, Bomberg EM, et al. Liraglutide for children 6 to <12 years of age with obesity — a randomized trial. N Engl J Med 2025;392:555-65. DOI: 10.1056/NEJMoa2407379

This trial protocol has been provided by the authors to give readers additional information about the work.

[N Engl J Med 2025;392(6):555-565. PMID: 39258838]

From 1.8 mg to 1.2 mg	From 1.8 mg to 1.2 mg
From 1.2 mg to 0.6 mg	From 1.2 mg to 0.6 mg
	From 0.6 mg to 0.3 mg

The reason for lowering the trial product dose must be documented in the subject's medical record.

If the cause of the subject's tolerability issues is intermittent illness or otherwise transient, as judged by the investigator, the subject can return to MTD again.

If doses are completely missed, refer to Section 7.6.1 for actions.

7.3 Method of treatment assignment

All subjects will be centrally randomised using an IWRS and assigned to the next available treatment according to randomisation schedule. Trial product will be dispensed/allocated at the trial visits summarised in the flowchart.

Randomisation will be stratified according to the following Tanner stages of pubertal development:

- Tanner stage 1 pre-pubertal (premature adrenarche permitted)
- Tanner stage 2-3
- Tanner stage 4-5

7.4 Blinding

The active drug and placebo drug are visually identical for the following trial products:

- Liraglutide 6.0 mg/mL, 3.0 mL, for s.c. injection with FlexPen®.
- Liraglutide placebo, 3.0 mL, for s.c. injection with FlexPen®.

The IWRS is used for blind-breaking. The blind may be broken in a medical emergency if knowing the actual treatment would influence the treatment of the subject. Novo Nordisk will be notified immediately after breaking the blind. The date when and reason why the blind was broken must be recorded in the source documentation.

Whenever the blind is broken, the person breaking the blind must print the "code break confirmation" notification generated by the IWRS, record the reason and sign and date the document.

When the blind is broken, the treatment allocation will be accessible to the investigator and the Novo Nordisk Global Safety department. If IWRS is not accessible at the time of blind break, the IWRS helpdesk should be contacted. Contact details are listed in Attachment I.

[N Engl J Med 2025;392(6):555-565. PMID: 39258838]

Table 9-1 Body measurements and instructions

Measurement	Measurement instructions	Equipment and equipment instructions	Unit
Body weight	Measured at all site visits without shoes, with an empty bladder and only wearing light clothing	Digital scale with graduation increments of 0.1 kg/0.1 lb The same scale should be used throughout the trial The scale must be calibrated yearly as a minimum	Kilograms or pounds (one decimal)
Height	Measured without shoes as two individual measurements performed by a single observer using identical technique The subject should be repositioned between the two measurements	Harpenden or another wall mounted stadiometer, with graduation increments of 0.1 cm/0.1 in The same stadiometer should be used throughout the trial	Centimetres or inches (one decimal)
Waist circumference (defined as abdominal circumference located midway between the lower rib margin and the iliac crest)	Measures must be obtained in standing position with a non-stretchable measuring tape The tape should touch the skin but not compress soft tissue and twists in the tape should be avoided. The subject should be asked to breathe normally	Non-stretchable measuring tape The same measuring tape should be used throughout the trial The measuring tape will be provided by Novo Nordisk to ensure standardisation	Nearest centimetre or inch

BMI will be calculated in the CRF every time the weight and height are measured.

9.1.2 Clinical efficacy laboratory assessments

All protocol-required laboratory assessments, as defined in Appendix 2 must be conducted in accordance with the flowchart and the laboratory manual.

9.2 Adverse events

The definitions of AEs and SAEs can be found in Appendix 4, along with a description of AEs requiring additional data collection.

[N Engl J Med 2025;392(6):555-565. PMID: 39258838]

10 Statistical considerations

Taxonomy of week 56 assessments

For each subject a given assessment at week 56 may be available or missing and Table 10-1 describes the taxonomy for this. Note, this is done per assessment and per subject; subjects may be a different type for different assessments (a subject may have "available on randomised treatment (AT)" for body weight but "missing on randomised treatment (MT)" for waist circumference).

Table 10-1 Taxonomy for subjects based on week 56 assessments

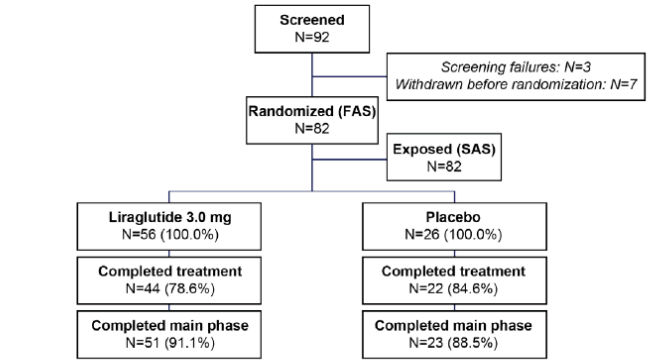
Assessment at week 56	Subjects on randomised treatment at week 56	Type description	Type Abbreviation
Available	Yes	**Available on randomised treatment**: Subjects who complete the trial on randomised treatment with an assessment at week 56: Includes those that stop and restart trial product.	AT
	No	**Available but discontinued:** Subjects who discontinued randomised treatment prematurely but returned to have an assessment at week 56. These are also called retrieved subjects	AD
Missing	Yes	**Missing on randomised treatment**: Subjects who complete the trial on randomised treatment without an assessment at week 56: Includes those that stop and restart trial product.	MT
	No	**Missing and discontinued**: Subjects who discontinued randomised treatment prematurely and did not return to have an assessment at week 56. These are also called non-retrieved subjects	MD

10.1 Sample size determination

As no long-term clinical trial of liraglutide 3.0 mg on weight management in children has yet been completed, the estimate of effect is based on observations from adult trials. For the power calculation it is assumed that the effect of liraglutide 3.0 mg on weight management in children will be similar to the effect observed in adults despite dosing flexibility. This assumption is based on the fact that body weight is the primary determinant for exposure; although the individual maximum tolerated dose may be lower in a proportion of the subjects with lower weight, the exposure is expected to be within the same range. The anticipated treatment difference for change in BMI (%) as -4.5% and -5.0% were investigated (based on phase III 56-week trials 1839, 1922 and 1923; 3,037 subjects in pooled analysis). The range of standard deviations assessed (SD) is based on the two more recent trials NN8022-4272 and NN8022-4274, where retention was optimized.

In agreement with EMA/PDCO, the minimum number of randomised subjects should be 78. In a fixed sample size design with a total of 78 subjects, the power calculation is performed using the following scenarios with SD of 4%, 5.0%, 5.5%, 6.0%, 6.5%, 7.0% and 7.5%.

[N Engl J Med 2025;392(6):555-565. PMID: 39258838]



148

149 **Figure S2. Participant Disposition**

150 Participants who discontinued treatment could remain in the study for subsequent assessments until

151 the end of the study. FAS, full analysis set; SAS, safety analysis set.

27

[N Engl J Med 2025;392(6):555-565. PMID: 39258838]

Meta-Analytic Bias

- Search bias
  - Not enough databases
  - Language restrictions
  - Insufficient attention to grey literature

Grey Literature

" One valid caveat was pointed out by an anonymous peer reviewer of this paper. It is possible that published random control trials are a biased sample of all completed trials. Trials wholly negative or favoring the placebo may have never been submitted or accepted for publication. Data suggesting publication bias have been reported by one of us […]. The only conceivable way to handle this problem us to make a plea that anyone having such unpublished data send them to the authors of this review. "

[N Engl J Med 1977;297(20):1091-6. PMID: 909566]

Grey Literature

“ […] on average, published trials showed a 9% greater treatment effect than grey trials […]. Overall there were more published trials included in the meta-analyses than grey trials […]. Published trials had more participants on average. The most common types of grey literature were abstracts (55%) and unpublished data (30%). […] This has important implications for reviewers who need to ensure they identify grey trials, in order to minimise the risk of introducing bias into their review. ”

[Cochrane Database Syst Rev 2007;2007(2):MR000010. PMID: 17443631]

High Risk of Bias

“ Our results indicate that systematic reviews published in some of the most influential journals in the field do not implement enough measures in their search strategies to reduce the risk of PB [publication bias], nor do they assess the risk of its presence or take the risk of its presence into consideration when inferring their results. ”

[Syst Rev 2024;13(1):11. PMID: 38169404]

ANNALS OF SCIENCE

THE TRUTH WEARS OFF

Is there something wrong with the scientific method?

BY JONAH LEHRER

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects' psychiatric symptoms. As a result, second-generation antipsychotics had become one of the fastest-growing and most profitable pharmaceutical classes. By 2001, Eli Lilly's Zyprexa was generating more revenue than Prozac. It remains the company's top-selling drug.

But the data presented at the Brussels meeting made it clear that something strange was happening: the therapeutic power of the drugs appeared to be steadily waning. A recent study showed an effect that was less than half of that documented in the first trials, in the early nineteen-nineties. Many researchers began to argue that the expensive pharmaceuticals weren't any better than first-generation antipsychotics, which have been in use since the fifties. "In fact, sometimes they now look even worse," John Davis, a professor of psychiatry at the University of Illinois at Chicago, told me.

Before the effectiveness of a drug can be confirmed, it must be tested and tested again. Different scientists in different labs need to repeat the protocols and publish their results. The test of replicability, as it's known, is the foundation of modern research. Replicability is how the community enforces itself. It's a safeguard for the creep of subjectivity. Most of the time, scientists know what results they want, and that can influence the results they get. The premise of replicability is that the scientific community can correct for these flaws.

But now all sorts of well-established, multiply confirmed findings have started to look increasingly uncertain. It's as if our facts were losing their truth: claims that have been enshrined in textbooks are suddenly unprovable. This phenomenon doesn't yet have an official name, but it's occurring across a wide range of fields, from psychology to ecology. In the field of medicine, the phenomenon seems extremely widespread, affecting not only antipsychotics but also therapies ranging from cardiac stents to Vitamin E and antidepressants: Davis has a forthcoming analysis demonstrating that the efficacy of antidepressants has gone down as much as threefold in recent decades.

For many scientists, the effect is especially troubling because of what it exposes about the scientific process. If replication is what separates the rigor of science from the squishiness of pseudoscience, where do we put all these rigorously validated findings that can no longer be proved? Which results should we believe? Francis Bacon, the early-modern philosopher and pioneer of the scientific method, once declared that experiments were essential, because they allowed us to "put nature to the question." But it appears that nature often gives us different answers.

Jonathan Schooler was a young graduate student at the University of Washington in the nineteen-eighties when he discovered a surprising new fact about language and memory. At the time, it was widely believed that the act of describing our memories improved them. But, in a series of clever experiments, Schooler demonstrated that subjects shown a face and asked to describe it were much less likely to recognize the face when shown it later than those who had simply looked at it. Schooler called the phenomenon "verbal overshadowing."

52 THE NEW YORKER, DECEMBER 13, 2010

[New Yorker. 13 Dec 2010, pp. 52ff]

Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence

Question	Step 1 (Level 1*)	Step 2 (Level 2*)	Step 3 (Level 3*)	Step 4 (Level 4*)	Step 5 (Level 5)
How common is the problem?	Local and current random sample surveys (or censuses)	Systematic review of surveys that allow matching to local circumstances**	Local non-random sample**	Case-series**	n/a
Is this diagnostic or monitoring test accurate? (Diagnosis)	Systematic review of cross sectional studies with consistently applied reference standard and blinding	Individual cross sectional studies with consistently applied reference standard and blinding	Non-consecutive studies, or studies without consistently applied reference standards**	Case-control studies, or "poor or non-independent reference standard**	Mechanism-based reasoning
What will happen if we do not add a therapy? (Prognosis)	Systematic review of inception cohort studies	Inception cohort studies	Cohort study or control arm of randomized trial*	Case-series or case-control studies, or poor quality prognostic cohort study**	n/a
Does this intervention help? (Treatment Benefits)	Systematic review of randomized trials or *n*-of-1 trials	Randomized trial or observational study with dramatic effect	Non-randomized controlled cohort/follow-up study**	Case-series, case-control studies, or historically controlled studies**	Mechanism-based reasoning
What are the COMMON harms? (Treatment Harms)	Systematic review of randomized trials, systematic review of nested case-control studies, *n*-of-1 trial with the patient you are raising the question about, or observational study with dramatic effect	Individual randomized trial or (exceptionally) observational study with dramatic effect	Non-randomized controlled cohort/follow-up study (post-marketing surveillance) provided there are sufficient numbers to rule out a common harm. (For long-term harms the duration of follow-up must be sufficient.)**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning
What are the RARE harms? (Treatment Harms)	Systematic review of randomized trials or *n*-of-1 trial	Randomized trial or (exceptionally) observational study with dramatic effect			
Is this (early detection) test worthwhile? (Screening)	Systematic review of randomized trials	Randomized trial	Non -randomized controlled cohort/follow-up study**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning

*** Level may be graded down on the basis of study quality, imprecision, indirectness (study PICO does not match questions PICO), because of inconsistency between studies, or because the absolute effect size is very small; Level may be graded up if there is a large or very large effect size.**

**** As always, a systematic review is generally better than an individual study.**

How to cite the Levels of Evidence Table

OCEBM Levels of Evidence Working Group*. "The Oxford 2011 Levels of Evidence".
Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653

** OCEBM Table of Evidence Working Group = Jeremy Howick, Iain Chalmers (James Lind Library), Paul Glasziou, Trish Greenhalgh, Carl Heneghan, Alessandro Liberati, Ivan Moschetti, Bob Phillips, Hazel Thornton, Olive Goddard and Mary Hodgkinson*

Patient-Oriented Evidence

A POEM a week for the BMJ

A POEM is Patient-Oriented Evidence that Matters

From now the *BMJ* will publish every week a POEM—a summary of a valid piece of research that carries information that is important to patients and so to their doctors. Unfortunately most research does not provide information that matters to patients. The POEMs will be published beside Editor's Choice. POEM stands for Patient-Oriented Evidence that Matters, and the concept was developed by David Slawson and Allen Shaughnessy, academics in family practice from University of Virginia in the United States.[1,2]

The concept has its origins in a formula developed by Slawson and Shaughnessy:

$$U = \frac{R \times V}{W}$$

where U = usefulness of the information to doctors, R = relevance of the information to doctors, V = validity of the information, and W = work to access the information. In words, the most useful information for doctors is information that is relevant to their practice, valid, and does not take too much work to access. After listening to a presentation by Maria Musoke, a researcher from Uganda, on the usefulness of information to rural health workers in Uganda I added "interactivity" to the top line of the equation.[3] The information is still more useful if you can interact with the source and interrogate it.

The formula provides a test of the ways in which doctors look for information they need. Traditional journal articles, although usually valid, are rarely directly relevant to a practitioner and are hard work to read—and they cannot be interrogated, although rapid responses (electronic letters to the editor) provide a possible means of getting answers from authors. The usefulness of original articles might thus be categorised as low. Textbooks should be relevant, although it's disturbing how often they fail to provide an answer to a direct question, and are comparatively easy to access. Their validity is questionable because they are rarely based on a systematic review of the literature and are often out of date, and they cannot be interrogated. They are thus of medium usefulness. In contrast, expert colleagues will give a direct and relevant answer to a question, should be little work to access, and can be interrogated. They are thus a highly useful source of information, although sometimes the validity of their answers may be low—"the blind leading the blind." The formula thus explains why doctors use colleagues most commonly to answer questions and journals least often.[4]

Doctors suffer from what Muir Gray, director of the National Electronic Library of Health, calls "the information paradox": they are overwhelmed with information, many receiving their own weight in journals and newspapers every month, and yet cannot find the information they need when they need it. At least two questions arise during the average consultation between a doctor and patient.[4] Most of those questions can be answered but few are. When I asked a sample of doctors to give me the one adjective they associate with their information supply, 90% gave a negative answer—overwhelmed, crushed, despairing. More than half of doctors feel guilty that they don't read more. Information has negative connotations for doctors.

What is a POEM?

POEM stands for Patient-Oriented Evidence that Matters.
POEMs have to meet three criteria:
- They address a question that doctors encounter
- They measure outcomes that doctors and their patients care about: symptoms, morbidity, quality of life, and mortality
- They have the potential to change the way doctors practise.

Doctors are in a "knowledge business" and yet have severe information problems. The electronic age allows the possibility of a solution,[4] but it hasn't been found yet. POEMs are a step forward. The box shows the three criteria that POEMs have to meet. Very importantly they have to provide information that will matter to patients. Will they live or die? Will they feel sick? Will they have pain? Will they be able to do what they want to do? A great many studies in medical journals give information on mechanisms of disease, aetiology, prevalence, pathophysiology, and pharmacology—studies that may be important but don't matter to patients. Faced with far more material than they can ever hope to master doctors might find it useful to concentrate on the studies that provide evidence that will matter to patients. They will discover that it is a minority of studies.

POEMs are selected by searching the current issues of 100 journals looking for relevant studies, potential POEMs, which are then evaluated for validity. The valid POEMs are summarised, and the summary is then reviewed and revised. The service is provided by InfoRetriever, who have kindly allowed us to publish a POEM each week. Those who would like to subscribe to their full service should access their site at www.infopoems.com/index.cfm

Richard Smith *editor, BMJ*

Competing interest: RS is the editor of the *BMJ* and the chief executive of the BMJ Publishing Group. He will not benefit financially from the arrangement with InfoRetriever. The BMJ Publishing Group might.

1 Shaughnessy AF, Slawson DC, Bennett JH. Becoming an information master: a guidebook to the medical information jungle. *J Fam Pract* 1994;39:489-99.
2 Slawson DC, Shaughnessy AF. Obtaining useful information from expert based sources. *BMJ* 1997;314:947-9.
3 http://www.infopoems.com/newslet/may01.html#health (accessed 28 October 2002).
4 Smith R. What clinical information do doctors need? *BMJ* 1996;313:1062-8.

BMJ 2002;325:983

[BMJ 2002;325(7371):983. PMID: 12411333]

SORT

The Strength-of-Recommendation Taxonomy

AFP uses the Strength-of-Recommendation Taxonomy (SORT),[1] to label key recommendations in clinical review articles. In general, only key recommendations are given a Strength-of-Recommendation grade. Grades are assigned on the basis of the quality and consistency of available evidence. *Table 1* shows the three grades recognized.

As the table indicates, the strength-of-recommendation grade depends on the quality and consistency of the evidence for the recommendation. Quality and consistency of evidence are determined as indicated in *Table 2* and *Table 3*.

An alternative way to understand the significance of a strength-of-recommendation grade is through the algorithm generally followed by authors and editors in assigning grades based on a body of evidence *(Figure 1)*. While this algorithm provides a general guideline, authors and

TABLE 1

Strength-of-Recommendation Grades

Strength of recommendation	Basis for recommendation
A	Consistent, good-quality patient-oriented evidence*
B	Inconsistent or limited-quality patient-oriented evidence*
C	Consensus, disease-oriented evidence,** usual practice, expert opinion, or case series for studies of diagnosis, treatment, prevention, or screening

*—Patient-oriented evidence measures outcomes that matter to patients: morbidity, mortality, symptom improvement, cost reduction, and quality of life.
**—Disease-oriented evidence measures intermediate, physiologic, or surrogate end points that may or may not reflect improvements in patient outcomes (e.g., blood pressure, blood chemistry, physiologic function, pathologic findings).

TABLE 2

Assessing Quality of Evidence

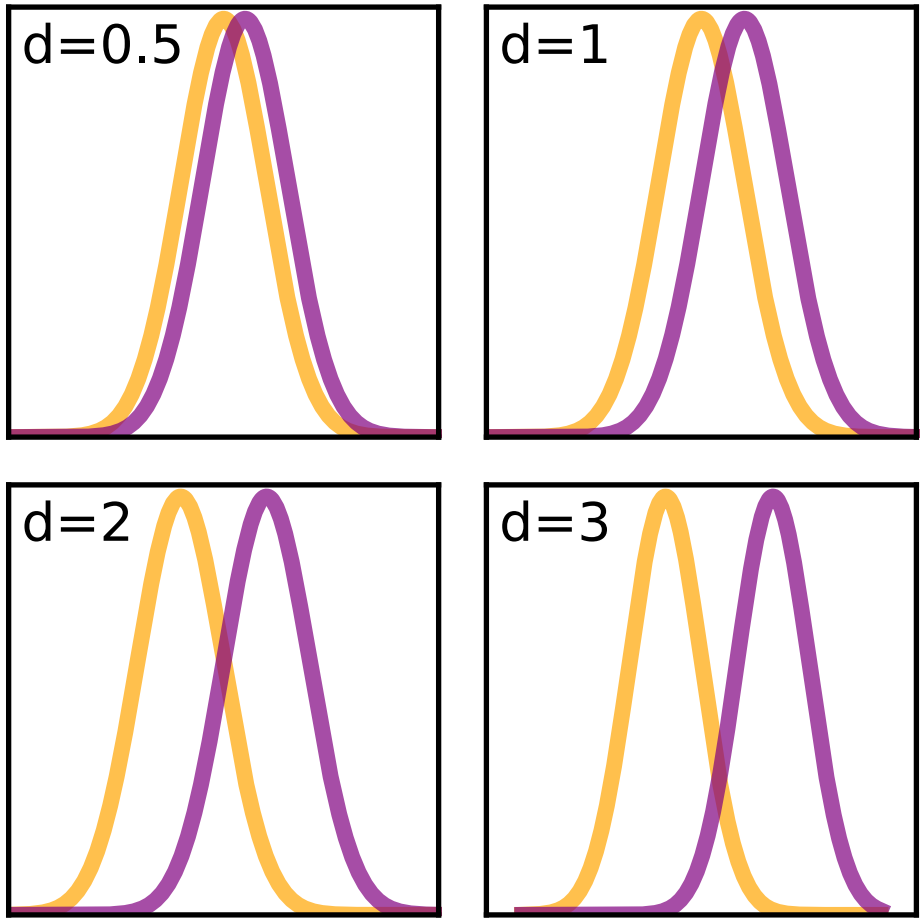
Study quality	Diagnosis	Treatment/prevention/screening	Prognosis
Level 1: good-quality, patient-oriented evidence	Validated clinical decision rule Systematic review/meta-analysis of high-quality studies High-quality diagnostic cohort study*	Systematic review/meta-analysis of RCTs with consistent findings High-quality individual RCT† All-or-none study‡	Systematic review/meta-analysis of good-quality cohort studies Prospective cohort study with good follow-up
Level 2: limited-quality patient-oriented evidence	Unvalidated clinical decision rule Systematic review/meta-analysis of lower quality studies or studies with inconsistent findings Lower quality diagnostic cohort study or diagnostic case-control study	Systematic review/meta-analysis of lower quality clinical trials or of studies with inconsistent findings Lower quality clinical trial Cohort study Case-control study	Systematic review/meta-analysis of lower quality cohort studies or with inconsistent results Retrospective cohort study or prospective cohort study with poor follow-up Case-control study Case series
Level 3: other evidence	Consensus guidelines, extrapolations from bench research, usual practice, opinion, disease-oriented evidence (intermediate or physiologic outcomes only), or case series for studies of diagnosis, treatment, prevention, or screening		

RCT = randomized controlled trial.
*—High-quality diagnostic cohort study: cohort design, adequate size, adequate spectrum of patients, blinding, and a consistent, well-defined reference standard.
†—High-quality RCT: allocation concealed, blinding if possible, intention-to-treat analysis, adequate statistical power, adequate follow-up (greater than 80 percent).
‡—In an all-or-none study, the treatment causes a dramatic change in outcomes, such as antibiotics for meningitis or surgery for appendicitis, which precludes study in a controlled trial.

[Am Fam Physician 2018;98(11):660G-660H]

Would a Patient Notice?

- Effect size
- Dichotomized outcomes
- Minimal clinically important difference (MCID)



[Wikipedia. Effect Size. Accessed 03/03/2025]

Dichotomized Data

- Clinical trials for major depressive disorder (MDD)
  - ≥ 50% reduction in symptoms compared to baseline
  - Remission

DESIGN

The ESCAPE-TRD trial was an open-label, single-blind (with raters unaware of trial-group assignments), randomized, active-controlled trial that was conducted across 171 sites comprising hospitals, inpatient and outpatient clinics, and research centers in 24 countries. The goal of the trial was to evaluate the efficacy, safety, and side-effect profile of esketamine nasal spray as compared with extended-release quetiapine, both in combination with a continuing SSRI or SNRI, in patients with treatment-resistant depression.

The trial consisted of a screening phase of up to 14 days, an initial treatment phase of 8 weeks, a maintenance phase of 24 weeks, and a safety follow-up through 2 weeks after the last dose of trial treatment (Fig. S1). After the screening phase, patients were randomly assigned, in a 1:1 ratio, to receive esketamine nasal spray plus an SSRI or SNRI (esketamine group) or extended-release quetiapine plus an SSRI or SNRI (quetiapine group). Randomization was performed with the use of a computer-generated schedule prepared before the trial, in randomly permuted blocks and with stratification according to age (18 to ≤64 years vs. 65 to ≤74 years) and the total number of past treatments that failed (2 vs. ≥3). Patients who discontinued the trial treatment remained in the trial and were invited to attend all visits through week 32. The doses of esketamine nasal spray and extended-release quetiapine were flexible and accorded with the summary of product characteristics for each agent.[12,17] Details about the dosing and administration of the trial treatments are provided in the Supplementary Material S1 section in the Supplementary Appendix.

EFFICACY

The efficacy analyses included all the patients who underwent randomization (intention-to-treat approach). The primary and key secondary end points were assessed according to the score on the Montgomery–Åsberg Depression Rating Scale (MADRS; scores range from 0 to 60, with higher scores indicating more severe depression); the clinical interview to determine the score was performed on site by independent raters who were unaware of the trial-group assignments. The primary end point was remission — defined as a score of 10 or less on the MADRS[23] — at week 8 after randomization (short-term efficacy). The key secondary end point was no relapse through week 32 after remission at week 8 (long-term efficacy). Relapse was defined as a MADRS score that worsened to 22 or higher at two consecutive assessments within 5 to 15 days of each other; hospitalization for worsening depression, suicide prevention, or suicide attempt; suicide attempt; completed suicide; or any other event assessed by the investigator to be indicative of relapse.

Analyses of the rates of remission (defined as a MADRS score of ≤10) and response (defined as an improvement of ≥50% in the MADRS score from baseline or as a MADRS score of ≤10) and analysis of the change in the MADRS score from baseline over time are also reported. In addition, we analyzed remission at week 8 and freedom from relapse through week 32 after remission at week 8 using a MADRS score of 12 or less as the threshold for remission — a threshold that was used in the registrational trials of esketamine nasal spray — to facilitate the contextualization of our trial with the previous phase 3 trials in the clinical development program of esketamine nasal spray (see the Supplementary Material S2 section).[18,20,24-26]

SAFETY

The safety analysis included all the patients who received at least one dose of the trial treatment. Adverse events (classified according to the preferred terms in the *Medical Dictionary for Regulatory Activities*, versions 23 to 25) were considered to have occurred during the treatment period if they occurred between the first dose and the safety follow-up visit (14 days after the last dose) or, in the case of serious adverse events, if they occurred between the first dose and 30 days or less after last dose. Safety evaluations were performed throughout the trial.

STATISTICAL ANALYSIS

The trial was designed to have 90% power for assessment of the primary end point and 80% power for assessment of the key secondary end point. Using the nonresponder imputation approach, we estimated that 41.25% of patients in the esketamine group and 28.88% of patients in the quetiapine group would have remission at week 8 and that 25.99% and 16.17% of patients,

[N Engl J Med 2023;389(14):1298-1309. PMID: 37792613]

MCID

- Smallest change in a treatment outcome that a patient would identify as important

Optum

Determination of Maximum Therapeutic Benefit

Optum Health Solutions Musculoskeletal (MSK)
Utilization Management Policy
Policy Number: 84

Effective Date: 04/25/2024

reference standard, and there is wide inter-person variability in self-reports of symptoms and function. Therefore, it is important to interpret PROMs using their MCID, which can be used as a criterion for assessing the beneficial effects of a therapy. (Salaffi et al., 2004).

The minimum clinically important difference (MCID) was first defined in 1989 as "the smallest difference in score in the domain of interest which patients perceive as beneficial" (Jaeschke et al., 1989). While others have described similar terms (e.g., minimal clinically important change) and definitions, the fundamental idea has remained the same: MCID is a calculated threshold value in an outcome of interest that patients and clinicians perceive as clinically meaningful, i.e., a value that demonstrates an appreciable change in outcome (Chung et al., 2017). According to Kirwan (2001), the basis for quantification and standardization of MCID is to minimize the variability in clinician judgment of patient 'change' following treatment. The inaccurate assessment of 'change' has been shown to mitigate the quality of clinician decision-making (Saintonge et al., 1988).

An international panel of experts has stated that 30% change from baseline may be considered a clinically meaningful improvement when comparing before and after patient-reported outcomes scores. The minimal [clinically] important change values adopted by the VII International Forum on Primary Care Research on Low Back Pain (Amsterdam, June 2006) are: 15/100 for the Visual Analogue Scale (VAS), 2/10 for the Numerical Rating Scale (NRS), 5/24 for the Roland-Morris Disability Questionnaire (RMDQ), 10/100 for the Oswestry Disability Index (ODI), and 20/100 for the Quebec Back Pain Disability Questionnaire QBPDQ) (Ostelo et al., 2008). MCID for the most common outcome assessment tool are shown in Table 4.

Table 4
MCID for Commonly Employed Outcome Measures

Author (date)	Domain	Outcome Tool	Study Characteristics	MCID	Follow Up Period
Beaton (2001)	Function/disability	DASH	*N*=200 Diverse subject group with either wrist/hand or shoulder problems	15 scale points	3 months
Binkley (1999)	Function/disability	LEFS	*N*=107 Convenience sample from 12 PT outpatient clinics All LE conditions included Correlated with SF-36 physical function score	9 scale points	1-2 days following baseline then weekly x 4 weeks
Farrar (2001)	Pain	NRS	*N*=2724 Retrospective analysis of controlled trials for diabetic neuropathy, post-herpetic neuralgia, chronic LBP, fibromyalgia, osteoarthritis	30% change from baseline	5-12 weeks
Fritz (2001)	Function/disability	ODI	*N*=67 Work-related LBP with and without LE pain	6%	4 weeks

[Optum, ibid.]

Mash-Up

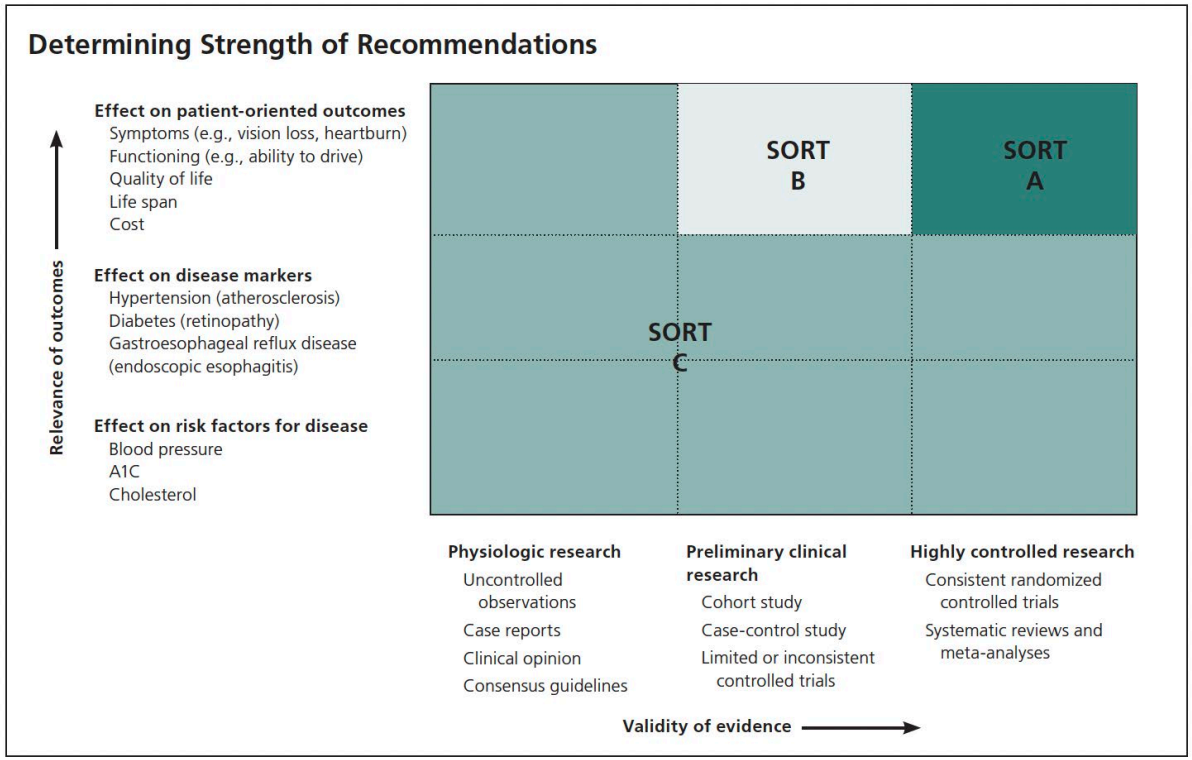


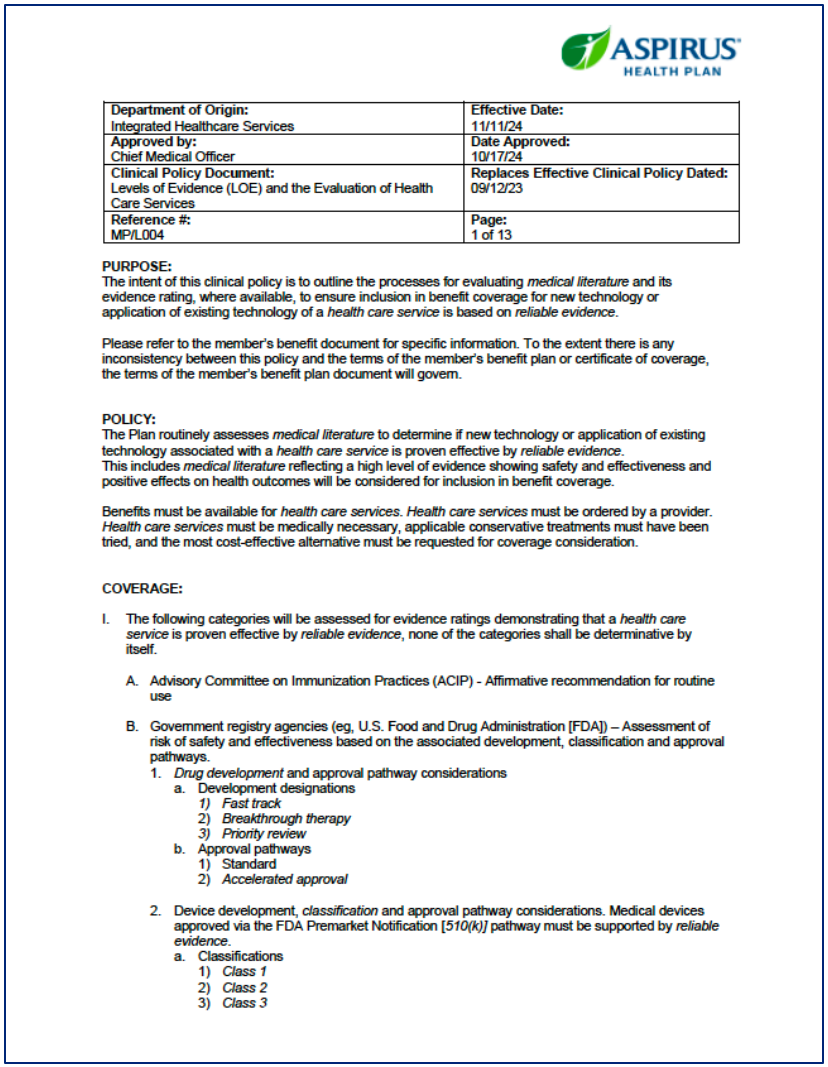
Figure 1. Clinical recommendations that receive higher Strength of Recommendation Taxonomy (SORT) ratings are based on more reliable information that specifically addresses patient-oriented outcomes. Only studies of patient-oriented outcomes can receive an A (good and consistent patient-oriented evidence) or B (limited or inconsistent patient-oriented evidence) rating. Recommendations from studies using disease-oriented evidence, and those based on clinical opinion or uncontrolled observations, receive a C rating.

[Am Fam Physician 2009;79(10):875-7. PMID: 19496387]

The End

Presenter's Qualifications

- I've been teaching evidence-based medicine (EBM) at the University of Minnesota since 2004
- I worked for Optum (and a predecessor company) from 2018 to 2025
  - I served on a medical policy team that developed coverage determination guidelines
    - The following slide is a high-level policy that integrates many of the concepts considered in this presentation



ASPIRUS HEALTH PLAN

Department of Origin: Integrated Healthcare Services	**Effective Date:** 11/11/24
Approved by: Chief Medical Officer	**Date Approved:** 10/17/24
Clinical Policy Document: Levels of Evidence (LOE) and the Evaluation of Health Care Services	**Replaces Effective Clinical Policy Dated:** 09/12/23
Reference #: MP/L004	**Page:** 1 of 13

PURPOSE:
The intent of this clinical policy is to outline the processes for evaluating *medical literature* and its evidence rating, where available, to ensure inclusion in benefit coverage for new technology or application of existing technology of a *health care service* is based on *reliable evidence*.

Please refer to the member's benefit document for specific information. To the extent there is any inconsistency between this policy and the terms of the member's benefit plan or certificate of coverage, the terms of the member's benefit plan document will govern.

POLICY:
The Plan routinely assesses *medical literature* to determine if new technology or application of existing technology associated with a *health care service* is proven effective by *reliable evidence*.
This includes *medical literature* reflecting a high level of evidence showing safety and effectiveness and positive effects on health outcomes will be considered for inclusion in benefit coverage.

Benefits must be available for *health care services*. *Health care services* must be ordered by a provider. *Health care services* must be medically necessary, applicable conservative treatments must have been tried, and the most cost-effective alternative must be requested for coverage consideration.

COVERAGE:

I. The following categories will be assessed for evidence ratings demonstrating that a *health care service* is proven effective by *reliable evidence*, none of the categories shall be determinative by itself.

   A. Advisory Committee on Immunization Practices (ACIP) - Affirmative recommendation for routine use

   B. Government registry agencies (eg, U.S. Food and Drug Administration [FDA]) – Assessment of risk of safety and effectiveness based on the associated development, classification and approval pathways.
      1. *Drug development* and approval pathway considerations
         a. Development designations
            1) *Fast track*
            2) *Breakthrough therapy*
            3) *Priority review*
         b. Approval pathways
            1) Standard
            2) *Accelerated approval*

      2. Device development, *classification* and approval pathway considerations. Medical devices approved via the FDA Premarket Notification [*510(k)]* pathway must be supported by *reliable evidence*.
         a. Classifications
            1) *Class 1*
            2) *Class 2*
            3) *Class 3*

[Optum serves as the TPA for AHP]