Systematic review and meta-analysis as a structured platform for teaching principles of experimentation

Land, Stephen C.; Booth, David

Published in:
Advances in Physiology Education

DOI:
10.1152/advan.00131.2019

Publication date:
2020

Document Version
Peer reviewed version

Systematic Review and Meta-analysis as a Structured Platform for Teaching Principles of Experimentation.

Stephen C. Land[1] and David Booth

Biological and Biomedical Science Education, School of Life Sciences,

University of Dundee, Dundee, DD1 4HN,

Scotland, United Kingdom.

Tables: 0

Figures: 8

[1]Address for Correspondence:

Dr Stephen Land
Biological and Biomedical Science Education,
School of Life Sciences, University of Dundee,
Dundee, DD1 4HN,
Scotland,
United Kingdom.
Tel: +44 1382 384760
Email: scland@dundee.ac.uk

Objectives & Overview

Access to knowledge has never been easier in the internet age and so it is important that students develop skills to discriminate undependable information from reliably investigated research. We have created an exercise which teaches good research practice by exploring the history, ethics and design of clinical trials. Students apply their understanding of these principles through an assessed systematic review and meta-analysis (SRMA) exercise. Here, a clinically themed hypothesis is tested using a structured literature search in conjunction with an eligibility matrix to map study design, ethics, subject selection, randomization & blinding, methodological standards, study power and other potential sources of inter-study heterogeneity. Data extracted from selected studies is used to produce a forest plot with an aggregated effect size, confidence range and measure of inter-study heterogeneity. A funnel plot is then used in conjunction with the eligibility matrix to evaluate study bias tendency and, in this way, students reflect upon the factors which promote disparate conclusion-making among studies with a common research focus. This exercise produced a normally distributed grade-profile across three academic year cohorts and comparison of individual exercise grade with year-long aggregated average suggested students who performed less well on conventional assignments engaged successfully with the systematic nature of this assessment. Those opting to use this format for their final year capstone project also performed above their grade point average from the preceding year. We suggest that SRMA offers a readily applied method for students to quantitatively explore how differences in experimental research practices influence study dependability.

Introduction

Scientific objectivity and critical thinking skills are often taught through small-group reflective activities such as journal clubs, critical writing workshops, mock grant committee-style peer review panels or seminar reviews (e.g. 5). These tend to focus on high-impact articles in a research field, published over a defined time-line, with compare-and-contrast discussion around assumptions, experimental design, analysis, conclusions and next steps. Pedagogic evaluation of these methods suggests that students make measurable gains in cognitive and critical thinking skills and that teaching methods tend to diversify to promote student engagement with the activity (25). These approaches may, however, tend to foster the impression that science advances purely by conclusive experimentation or fortunate discovery and that studies reporting neutral or negative results are less valuable or, in some way, flawed. This is a concern because publication bias in favour of positive research outcomes is believed to be fuelling a data reproducibility crisis in the life sciences and commonly employed literature-based teaching techniques may not give adequate attention to this issue (13-15).

Systematic review compliments these approaches by encouraging students to view research as a continuum, where positive, neutral and negative results of differing magnitudes are reported from different research locations over time. Rather than selecting studies based on concluding results or impact, a structured literature search is used to identify all primary research articles reporting data around a selected hypothesis, regardless of individual study outcome. These are screened for strengths of study design before outcomes are assessed using meta-analysis to obtain a measure of effect size based on the weighted contribution of each study (21). By aggregating effect sizes across several related studies, statistical power increases to the point where large or small biological effects can be discriminated, with factors that drive differences in reported outcomes (inter-study heterogeneity) evaluated retrospectively. In this way, emphasis is placed on the experimental principles which underpin each study rather than bottom-line results. Here, we describe an adaptation of this systematic

review and meta-analysis (SRMA) approach as an exercise for undergraduate biomedical students which encompasses teaching of research ethics, clinical trial regulation and bias management strategies alongside a structured approach to hypothesis testing using meta-analysis.

Learning Objectives

The objective is to review the modern history of human experimentation which has driven the development of ethical frameworks for clinical testing and the design of clinical trials. This establishes the background knowledge necessary to conduct an independent systematic review and meta-analysis exercise.

Specific Learning Outcomes

After completing this activity, the student should be able to:

- Describe the key historical events and developments in ethical reasoning which underpin present day regulation of human and animal experimentation.
- Use advanced search engine strategies to identify primary research literature which may be used to test a specific hypothesis
- Use the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow sheet to report the triage of studies for inclusion in a meta-analysis.
- Understand the forms of bias which can influence data collection and analysis and develop an eligibility matrix to screen articles for adherence to good research practices.
- Extract data from selected research articles which may be used to generate a forest plot using categorical or continuous data.
- Describe and interpret data patterns revealed by a forest plot.

- Apply this information to a funnel plot and use the output to interpret sources of variation which occurs between studies.
- Explore the effect of removing studies which contain identified sources of bias upon data distribution and intra-study heterogeneity on the forest plot
- Explain the physiological mechanisms which underpin the effects revealed by the forest plot.

Activity Level

We use this activity as a guided learning exercise for students in the third and fourth year of undergraduate study in the biomedical sciences (Scottish Credit Qualification Framework (SCQF) Levels 9 (equivalent to BSc Ordinary Degree) and 10 (equivalent to BSc Honours Degree), however, the clinical emphasis of the activity also carries relevance for medical students. This exercise provides appropriate training for quantitative literature-based research as part of an independent capstone project in the final year of undergraduate study paving the way for advanced postgraduate study. The exercise runs with a class size of around 100 students.

Prerequisite Student Knowledge

Students should have practical experience of experimental design gained from laboratory practical sessions as well as a fundamental grasp of physiology and pharmacology. We use R as a platform for teaching statistical analysis and so a basic understanding of R commands and grounding in statistical principles is an advantage.

Time and Resources Required

The exercise runs over 3 workshops each of 2hrs duration. The first session covers the history and ethics of clinical trials, the second explains the principles and approach to meta-analysis and the third is

computer-based session covering the process of meta-analysis in R (Fig 1.). Students spend a total of 45hrs in face-to-face teaching and completing their final SRMA report.

METHODS

Instructions

Workshop 1: History and Ethics of Clinical Trials. The first workshop provokes discussion about the purpose, history and ethics of clinical trials by explaining the timeline of events that led up to present-day regulation of human and animal experimentation. This begins with an introduction to James Lind's "Treatise on the Treatment of Scurvy", published in 1753 (19), which is the earliest documented use of systematic literature review in conjunction with a clinical trial (his successful, but misinterpreted attempt to identify a cure for scurvy (1)). This leads to a discussion of Bradford-Hill's streptomycin and tuberculosis study as the first case controlled randomised clinical trial design (20) and the relevance of Bradford Hill's Disease Causation Criteria to modern epidemiology (11). Development of the ethical framework governing clinical trials is presented through the events of World War II which led to the 1948 Nuremberg Code, Thalidomide testing and the 1962 Kefauver Amendments, the 1964 Declaration of Helsinki followed by the Tuskegee Syphilis Experiment and the principle of informed consent laid out in the 1979 Belmont Report. We explore what happens when clinical trials go wrong using examples from the University of Pennsylvania Ornithine Transcarbamylase (OTC) adenovirus gene therapy trial, the TGN1412 humanised monoclonal antibody trial as well as contemporary events reported in the media. The workshop ends with a discussion of the 3R Principle of Replacement, Reduction and Refinement as it relates to the use of animals in scientific procedures.

Workshop 2: How to Conduct a Systematic Review and Meta-analysis. This workshop establishes the principle that systematic review coupled with meta-analysis provides an overall estimate of effect size

and variance that is based upon the weighted outcomes of multiple studies testing a similar hypothesis. Students are taken stepwise through the meta-analysis process:

1. Establishing a single hypothesis.
2. Screening the literature for appropriate studies. Emphasis is placed on study design which must include steps which have been taken to minimize experimenter bias (eg randomization of treatments to subjects, concealment, blinding of treatments, full data collection).
3. The use of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) together with the PRISMA Screening Checklist.
4. Construction of a matrix to report inclusion criteria and reporting quality.
5. How to extract data from studies and dealing with categorical and continuous data
6. Forest plot interpretation
7. Funnel plot measurement of heterogeneity

Workshop 3: IT session with practise data sets. The purpose of the IT session is to familiarise students with the process of creating and interpreting forest and funnel plots before testing a hypothesis of their own. A basic level of competence with R is assumed but since commands, with explanatory notes, are provided, it is possible for those with no experience to complete the analysis. Part 1 of the workshop focuses on meta-analysis with count data using a systematic review examining if BCG vaccination reduces risk of tuberculosis (TB) in children (24). Part 2 analyses continuous data exploring the effectiveness of reducing unnecessary antibiotic use for hospital inpatients (9). By the end of this session, students have all the necessary information to test a novel hypothesis of their own. Teaching support information for this session is provided in Supplementary Material.S1.(https://doi.org/10.6084/m9.figshare.11674110).

Assessment: Learning outcomes were assessed through a SRMA exercise which tested the following hypothesis:

H_1: Sperm concentration is lower in smokers compared to non-smokers in human males of reproductive age.

This topic was selected because, i) the subject focus is concise but is reported over an extended time-line and from different geographical locations, ii) standardised measurements of semen quality (eg concentration, volume, motility) are widely reported and simple for students to identify in the literature, iii) the number of articles students would be expected to screen is not excessive and, iv) the subject matter promotes understanding of a wide range of biological processes.

Students reported their results using a proforma which dispersed marks across 6 steps of the SRMA process (Supplementary Table.S2;(https://doi.org/10.6084/m9.figshare.11674113). An example response to each step, together with a commentary, is provided in RESULTS. A grading rubric which explains each category and unit of assessment was made available to students in advance of the exercise and served to guide markers through the assessment (Supplementary Table.S3 (https://doi.org/10.6084/m9.figshare.11674122).

RESULTS

Example responses to the 6 components of the assessed exercise are provided as follows:

1. How did you search and screen for selected research articles? (15% of marks).

Students provide a breakdown of their search strategy using the PRISMA Flow Diagram. An advanced search in Pubmed using ((smoke[Title] OR smoking[Title])) AND (sperm[Title] OR semen[Title]) yields 140 articles; with an additional 13 articles from other sources, the total number of articles for the initial

screen was 153 which proved to be a manageable volume for our students. The screen yields several recent meta-analyses on the topic of smoking, tobacco and fertility (eg 3) which should be noted for comparison in later analysis but not included in the primary literature search. Search engine filters should be used to eliminate irrelevant articles and, for practical purposes, those which are not free-to-view (instructor should emphasize that this is not normal SRMA practise). An example PRISMA triage is shown in Fig. 2; the aim is to identify a short-list of publications that contain suitable data for a forest plot. The primary review follows a standard SRMA reporting process with the exception that no article is eliminated during eligibility screening process since this will be used by the student to evaluate sources of inter-study heterogeneity later in the exercise. 29 articles were identified through this process and are referenced in Supplementary Material.S4 (https://doi.org/10.6084/m9.figshare.11698473) together with data to be used in the meta-analysis. Each study is ordered by year of publication and identified by a letter of the alphabet for ease of interpretation in figures.

2. Assess your Articles for Eligibility and Create a Matrix (10 % of marks).

The instructor-led workshops include discussion of the bias containment strategies which are key to clinical trial design and owe their origins in Bradford Hill's original randomised case-controlled study of streptomycin and tuberculosis (20). Students demonstrate their understanding of this by creating an eligibility matrix which lists features of experimental design which they identify as important for each article in the meta-analysis (Fig 3). Their matrix is used in later stages of the assignment to identify causes of inter-study heterogeneity that affect their forest and funnel plot analysis.

3. Create and interpret a Forest Plot (25% of marks).

The forest plot was developed by Lewis and Ellis, 1982 (17, see 16 for a historical perspective) to determine an estimated effect size based on the proportionate contribution of studies testing the same hypothesis and has since been adopted as the standard method of evaluating effects across multiple studies, conducted at different times and in different geographical locations. It facilitates interpretation of a pooled point estimate and overall variation at a glance and, with assessment of intra-study heterogeneity (10,18), provides a powerful, structured approach for evaluating multi-study tests of a common hypothesis.

Example analysis: Students are instructed to extract mean, standard deviation and n for smokers and non-smoker sperm concentration from their selected articles. The continuous data protocol explored in Workshop 3 is used to generate a forest plot using data from studies identified in their literature search. An example forest plot is shown in fig 4, where mean sperm concentration (10^6 cells.ml^{-1}), standard deviation (SD) and subject numbers (Total columns) are shown for the 29 studies investigating smoking effects on sperm concentration listed in Supplementary Material.S4. Mean differences (MD) in sperm concentration between smokers and non-smokers are given on the right of the graph together with 95% confidence intervals and study weighting for fixed and random effects models. A summary aggregate analysis is given in the bottom two lines of the plot. Smoking and non-smoking subjects total 6159 and 11517 respectively. The fixed effect model assumes near identical study designs across all articles such that inter-study differences arise from chance sampling variation alone. It reports that that smoking reduces sperm concentration by 3.7x10^6 cells.ml^{-1} and that there is 95% confidence that the true value will lie within the range of (-4.9 to -2.6) x10^6 cells.ml^{-1}. The random effects model assumes that differences in the approach used by each study (eg cross sectional, prospective, retrospective experimental design and sampling differences) will introduce additional causes of variance above the natural pattern assumed by the fixed effects model (23). This model reports that smoking reduces sperm concentration by 4.45x10^6 cells.ml^{-1} with 95% confidence limits from (-8.7 to -0.2) x10^6 cells.ml^{-1}.

An analysis of Inter-study heterogeneity is provided in the lower left of the plot. I^2 reports the proportion (%) of the X^2 statistic which is not explained by the variation within the studies (<50=moderate-to-low heterogeneity; >51= high heterogeneity). τ^2 reports the variance of the true effect sizes based on the random effects model and the probability value reports likelihood of variation between studies. In this example, $p<0.01$ indicates high probability that each study reports an outcome which differs from the mean result so the fixed and random effect models should be treated with caution. Further investigation of inter-study heterogeneity should be performed using Funnel Plot analysis.

4. **Create and Interpret a Funnel Plot (25% of marks).**

Funnel plots provide a visual evaluation of the precision of studies in the forest plot. By plotting the standard error against mean difference, a distribution is obtained where high-powered studies cluster either side of the mean result near the plot apex and low-powered studies occur towards the base. Since 95% confidence limits vary inversely with study precision, a funnel-shaped confidence interval boundary is created which enables studies deviating outside these limits to be identified and investigated (18).

Example Analysis: The data used to generate the forest plot produces the funnel plot shown in Fig. 5. Most studies cluster close to the apex around the mean difference, however some lie outside the 95% confidence intervals suggesting that the overall data set is heterogeneous and could include extremes of bias. Students are instructed to use their eligibility matrix to evaluate application of bias containment strategies across their selected articles. Fig 3 identifies numerous possible causes of inter-study heterogeneity in this data set, however, recent comment in the field has highlighted poor compliance with internationally agreed semen analysis protocols as a problem (2,4). Students are encouraged to explore this in their own data by testing if inter-study heterogeneity decreases among studies which cite

257 the 5th (2010) edition of the "WHO Laboratory Manual for the Examination and Processing of Human
258 Semen" (26). A refined forest plot constructed from 8 articles which follow these criteria shows that
259 inter-study heterogeneity remains high (I^2 90%, P<0.01) but the funnel plot now reveals three studies
260 which lie beyond the 95% confidence limits (B,C,J in Fig. 6). In common with most of the studies used in
261 this analysis, few report adequate steps to contain bias and, so, given that there are no unique reasons
262 to exclude any one study, the final part of the analysis tested the effect of removing Study B as the most
263 distant outlier to the aggregated mean distance (Fig 7). This had the following effects: 1) mean
264 difference values for the fixed and random effect models now closely agree and increase from those
265 given in Figs. 4 & 5. 2) I^2 is < 50% and the likelihood of studies deviating from the mean effect value is
266 now 38%. Taking the random effects model as the most conservative estimate, it is now safe to
267 conclude that smoking reduces sperm concentration by $8.4x10^6$ cells.ml^{-1}. WHO lower reference limits
268 for normal sperm concentration are 15 (12-16) $x10^6$ cells.ml^{-1} (5^{th} centile, 95% confidence limits) (7) and
269 so it can be concluded that smoking induces a 44% reduction in fertility below this lower reference
270 value.

271 5. **Linking the constituents of cigarette smoke to spermatogenesis (25% of marks).**

272 Students are asked to summarise the molecular mechanisms which could link smoking behaviour to
273 spermatogenesis. We impose a strict limit of 150 words to encourage a focussed, abstract-style
274 paragraph. Students are required to describe one or more mechanisms which make a clear link between
275 smoking behaviour and the molecular regulation of spermatogenesis with informative graphical
276 abstracts encouraged. For information about the effects of cigarette constituents on spermatogenesis,
277 the reader is referred to refs 6,8 & 22.

278 **6. References (5% of marks)**

References are cited according to instructions from a leading journal in the field. For this exercise, this was the journal, Human Reproduction (Oxford Academic, Oxford, UK)

Evaluation of Student Work

The pro-forma report and marking rubric guide the student through the task, award a weighted grade for different skill components and facilitate feedback. We do not require students to perform an exhaustive literature search but they should aim to demonstrate appropriate use of advanced search engines, triage reporting methodology and develop a suitable eligibility screen. The analysis shown here identified a total of 29 relevant publications from 1992-2016, however, students typically based their reports on approximately half this number, with 5 stipulated as the lowest acceptable number of articles. The literature search process is time-consuming and so students were encouraged to use discussion boards to share search strategies, though exchange of reference material was not permitted.

In our hands, this exercise produces a normal grade distribution with a median score of 60-63% from a pooled cohort of 298 students from 3 consecutive years (Fig 8A). The normal distribution suggests effective discrimination of student ability across the range of grades available with no evidence of kurtosis that might arise from variable student engagement or differences in assessment approaches. Individual performance on this exercise was examined by plotting the grade difference for the SRMA assignment (assignment grade point minus year-long running average grade point) against year-long performance for each student. Regression analysis suggests a modest relationship (r^2=0.31) whereby individual attainment tended to be greater among students whose overall year-long attainment was otherwise low (Fig 8B). We did not evaluate the reasons behind this but note that active learning exercises, of the type described here, increase performance among several metrics of learning attainment (12). It may be that SRMA encompasses a structured approach to literature review which facilitates engagement across the diverse learning abilities and styles. Finally, we examined retention of

SRMA learning outcomes by following the performance of students who opted to complete a SRMA capstone project in a subsequent year of study. Here, individual performance was assessed as the difference between capstone project grade and personal aggregate performance in the previous year of study. Individual performance was found to be consistent between SRMA and other project formats (wet laboratory, science communication or bioinformatics) (Figure 8C) suggesting that students were able to retain their understanding of SRMA from one year to the next and apply this to varied questions in the biosciences to a standard that matched other capstone project formats.

Common issues and errors.

1. A set of frequently asked questions (FAQ) has been collated from our SRMA on-line discussion board which addresses most issues encountered by students on this exercise (see Supplementary Material.S5 (https://doi.org/10.6084/m9.figshare.11702067).
2. Data conversions. Studies may report different measures of variance, requiring conversion to standard deviation. Similarly, studies may categorise smoking intensity of subjects in different ways and so, for simplicity, we advise students to determine an aggregated mean and SD of all smoking intensities. Conversion advice is provided in the FAQ's.
3. Selection of criteria for eligibility matrix. Students are referred to supporting lecture material on this topic and are encouraged to consider the Bradford-Hill selection criteria (11). Some may, however, become aware of the PRISMA checklist which runs to over 27 selection criteria. We advise students to avoid being too prescriptive in their selection of eligibility criteria but that they should focus on use of strategies to contain bias, adherence to measurement standards if relevant to the field, study size and use of techniques to assess study power.

4. Interpretation of forest plots. Students tend to gloss over the detail of forest plots to focus on the bottom-line result. This is compounded by confusion between appropriate use of fixed and random effects models and interpretation of heterogeneity information. Most commonly, students mis-interpret significant intra-study heterogeneity P values as an indication of effect size significance. We suggest careful guidance in the IT workshop together with advice offered on the discussion board as the best approach to address these issues.
5. Interpretation of Funnel Plots. As with Forest Plots, students tend to focus on the basic interpretation of the plot without attempting a deeper analysis of the data. It is important to emphasize that no paper may be removed from the analysis without justified cause, as reflected in the eligibility matrix. Students should recognise that there are limits to this analysis and that inter study heterogeneity may be an issue which affects the wider field.

Limitations/Adaptations

The following should be considered when adapting this format to other topics: 1. Subject relevance should complement wider teaching goals. In the example described in this article, smoking and semen quality facilitated discussion of the modern history of epidemiology as well as the toxicology of tobacco smoke constituents. 2. The hypothesis should be precise and encourage focus on a single measured parameter or outcome. 3. The literature base for the topic should be manageable in size, readily identified using advanced search methods and freely accessible. The PRISMA report (Fig 2) indicates the volume of literature analysis and screening expected of students in the present exercise. 4. Consider if data will be extracted from tables or graphs. Is the data continuous or categorical? Our exercise required extraction of continuous data that is commonly reported among standard measurements of semen quality. Categorical data may be converted to an odds ratio or similar as described in the

Supplementary Material.S1. 5. Are there known causes of inter-study heterogeneity which may provide the opportunity for critical evaluation of experimental approaches in the field? For our exercise, heterogeneity arose primarily from varied adherence to a standard methodology over time, however, there are some notable geographical differences in the reporting of smoking as a positive or negative influence upon sperm concentration.

By running this exercise in the penultimate year of study, our intention was to provide students with the skills to conduct an independently researched SRMA capstone project in the subsequent final year. To date, 19 SRMA project topics in neuroscience, pharmacology and physiology have been completed at our institution suggesting that this format may be readily applied to a range of topics. Capstone project titles which are the basis for the data set in figure 8C are given in Supplementary Material.S6 (https://doi.org/10.6084/m9.figshare.11674107)

Conclusion

The SRMA exercise described here provides an opportunity for structured, quantitative evaluation of a focussed question using the peer reviewed scientific literature. The process requires students to consider the factors which underpin reliable study design, management of bias and data reporting. Our analysis of student performance reveals that the active learning attributes of the exercise may benefit students who tend to perform less well other forms of assessment. The format provides a suitable grounding for in-depth exploration of diverse topics through a capstone project in the advanced years of the undergraduate curriculum.

Additional Resources

Students are referred to Cochrane.org for access to articles explaining the SRMA process and to the Cochrane Library, a searchable database of evidence-based clinical studies. Students and instructors may also find the Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES) website helpful which can be accessed at www.dcn.ed.ac.uk/camarades/contact.html. The CAMARADES collaboration provides support for SRMA of data from experimental animal studies.

Acknowledgements

We thank undergraduate Biomedical Science, Neuroscience, Pharmacology and Physiological Sciences students at Dundee University whose participation and feedback on this exercise has informed its development.

References.

1. **Bartholomew M.** James Lind's *Treatise of the Scurvy (1753)*. Postgrad Med J 78:695–696, 2002.
2. **Bjorndahl L, Barratt CL, Mortimer D & Jouannet P.** "How to count sperm properly": checklist for acceptabilities of studies based on human semen analysis. Hum Reprod 31, 227–232, 2016.
3. **Bundhun PK, Janoo G, Bhurtu A, Teeluck AR, Soogund MZS, Pursun M, Huang F.** Tobacco smoking and semen quality in infertile males: a SRMA. BMC Public Health. J19(1):36, 2019.
4. **Carrell, DT and De Jonge CJ.** The troubling state of the semen analysis. Andrology, 2016, 4, 761–762.
5. **Clark IE, Romero-Caldero R. Olson J, Jaworski L, Lopatto D, Banerjee U. '**Deconstructing' Scientific Research: A Practical and Scalable Pedagogical Tool to Provide Evidence-Based Science Instruction. PLoS Biol 7(12): e1000264, 2009.

6. **Condorelli RA, La Vignera S, Giacone F, Iacoviello L, Mongioì LM, Li Volti G, Barbagallo I, Avola R, Calogero AE.** Nicotine Effects and Receptor Expression on Human Spermatozoa: Possible Neuroendocrine Mechanism. Front Physiol. 28; 8:177, 2017.
7. **Cooper TG, Noonan E, von Eckardstein S, Auger J, Baker HW, Behre HM, Haugen TB, Kruger T, Wang C, Mbizvo MT, Vogelsong KM**. World Health Organization reference values for human semen characteristics. Hum Reprod Update. 16(3):231-45, 2010.
8. **Dai JB, Wang ZX, Qiao ZD.** The hazardous effects of tobacco smoking on male fertility. Asian J Androl. 17(6):954-60, 2015.
9. **Davey P, Marwick CA, Scott CL, Charani E, McNeil K, Brown E, Gould IM, Ramsay CR, Michie S.** Interventions to improve antibiotic prescribing practices for hospital inpatients. Cochrane Database Syst Rev. 2:CD003543, 2017.
10. **Egger M, Davey Smith G, Schneider M, Minder C.** Bias in meta-analysis detected by a simple, graphical test. BMJ. 315(7109):629-34, 1997.
11. **Fedak, K. M., Bernal, A., Capshaw, Z. A., & Gross, S.** Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerging themes in epidemiology*, *12*, 14, 2015.
12. **Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP.** Active learning increases student performance in science, engineering and mathematics. Proc. Natl. Acad. Sci. USA 111(23) 8410-8415, 2014
13. **Goodman S, Greenland S.** Why most published research findings are false: Problems in the analysis. PLoS Med 4: e168, 2007
14. **Ioannidis JPA** Why most published research findings are false. PLoS Med 2: e124, 2005.
15. **Ioannidis JPA** Why Most Published Research Findings Are False: Author's Reply to Goodman and Greenland. PLoS Med 4(6): e215. 2007.

16. **Lewis S, Clarke M.** Forest Plots: Trying to see the wood for the trees. BMJ 322: 1479-1480, 2001.
17. **Lewis JA, Ellis SH.** A statistical appraisal of post-infarction beta-blocker trials. Prim Cardiol suppl 1: 31–37, 1982.
18. **Light RJ, Pillemer, DB.** Summing up: The Science of Reviewing Research. Cambridge, Massachusetts.: Harvard University Press. 1984.
19. **Lind J.** A Treatise of the Scurvy in 3 Parts. Second Edition, A. Millar, London, 450pp, 1753.
20. **Medical Research Council Streptomycin in Tuberculosis Trials Committee**. Streptomycin Treatment of Pulmonary Tuberculosis. Br Med J 2:769, 1948
21. **Møller, AM & Myles PS**. What makes a good SRMA? Brit. J Anaesthesia 117(4): 428-430, 2016.
22. **Nesseim WH, Haroun HS, Mostafa E, Youakim MF, Mostafa T.** Effect of nicotine on spermatogenesis in adult albino rats. Andrologia. 43(6):398-404, 2011.
23. **Riley RD, Higgins JPT**. Interpretation of random effects meta-analyses. BMJ 342: d549, 2011.
24. **Roy A, Eisenhut M, Harris RJ, Rodrigues LC, Sridhar S, Habermann S, Snell L, Mangtani P, Adetifa I, Lalvani A, Abubakar I.** Effect of BCG vaccination against Mycobacterium tuberculosis infection in children: SRMA. BMJ. 349: g4643, 2014.
25. **Stevens LM, Hoskins SG.** The CREATE strategy for intensive analysis of primary literature can be used effectively by newly trained faculty to produce multiple gains in diverse students. CBE Life Sci Educ.13:224–242, 2014
26. **World Health Organization.** WHO laboratory manual for the examination and processing of human semen. 5th ed. Geneva: World Health Organization, 2010.

Figure Legends

Figure 1. Timeline of topics covered in the SRMA exercise.

Figure 2. Preferred Reporting Items for SRMA (PRISMA) flow diagram illustrating initial triage of articles to produce the final selection of articles to be used for the meta-analysis.

Figure 3. Eligibility matrix showing map of experimental design, bias management, reporting of ethics and adherence to methodological standards for 29 studies to be incorporated into the meta-analysis.

Figure 4. Forest plot showing study size, mean, standard deviation (SD), mean difference (MD), 95% confidence intervals (CI) and fixed or random model weighting from 29 studies reporting effects of smoking on sperm concentration. Mean and SD are reported as 10^6 cells.ml^{-1}. Pooled values were used from studies which reported sperm parameters values for mild, moderate and heavy smoking habits. Fixed and random effect summary outcomes are shown.

Figure 5. Funnel plot showing the relationship between individual study standard error and mean difference. Letters denote studies listed in Table 2. Long dash, fixed effect model median; short dash, random effect model median; angled dashed lines, region within which 95% of studies would be expected to lie in absence of intra-study variability and publication bias (calculated as the fixed effect summary log mean difference ± 1.96 × standard error of summary log mean difference).

458 **Figure 6.** Reduced forest (**A**) and funnel (**B**) plots based on 8 studies which cite the 5th edition of the
459 WHO methodology (2010) (26). Labelling details are as indicated in figures 2 & 3.

460

461 **Figure 7.** A minimum forest (**A**) and funnel (**B**) model which produces a statistically insignificant level of
462 heterogeneity between studies, achieved by removal of study [B].

463

464 **Figure 8**. **A.** Frequency histogram showing distribution of grades as percentage categories for an
465 identical SRMA exercise conducted over two consecutive years. N=298. Grade distribution is indistinct
466 from a normal distribution (Kruskal-Wallis Rank Sum test $X^2 = 18$, d.f= 18, $p = 0.46$) Note that grade bin
467 categories are non-linear at range extremities. **B.** Individual performance on this meta-analysis exercise.
468 Δ grade point average (ΔGPA) was determined as the difference between exercise grade point and year-
469 long aggregated average grade point as determined on a 23-point scale. Regression is linear ($y = -0.57x$
470 $+ 8.7$; $R^2 = 0.31$, $F(1,179)=77.6$, $p<0.001$, N=181). Long dash line indicates zero intercept; Data points
471 above this line indicate an exercise performance which is above the individual's running average for the
472 year. **C.** Individual performance of SRMA capstone projects (N=19) compared to all other project formats
473 (N= 205). Δ grade point average (ΔGPA) was determined as the difference between capstone project
474 grade point and year-long aggregated average grade point during the penultimate year as determined
475 on a 23-point scale. Dash line indicates zero intercept; data points above this line indicate exercise
476 performance above the individual's running average for the penultimate year of study. An independent-
477 samples t-test was conducted to compare ΔGPA between those taking the SMRA capstone project
478 versus other project formats. There was not a significant difference in the scores for SMRA capstone
479 project (M=1.11, SD=2.38) and other project (M=1.50, SD=2.50) conditions; $t(222)= -0.68$, $p = 0.49$.

480

481

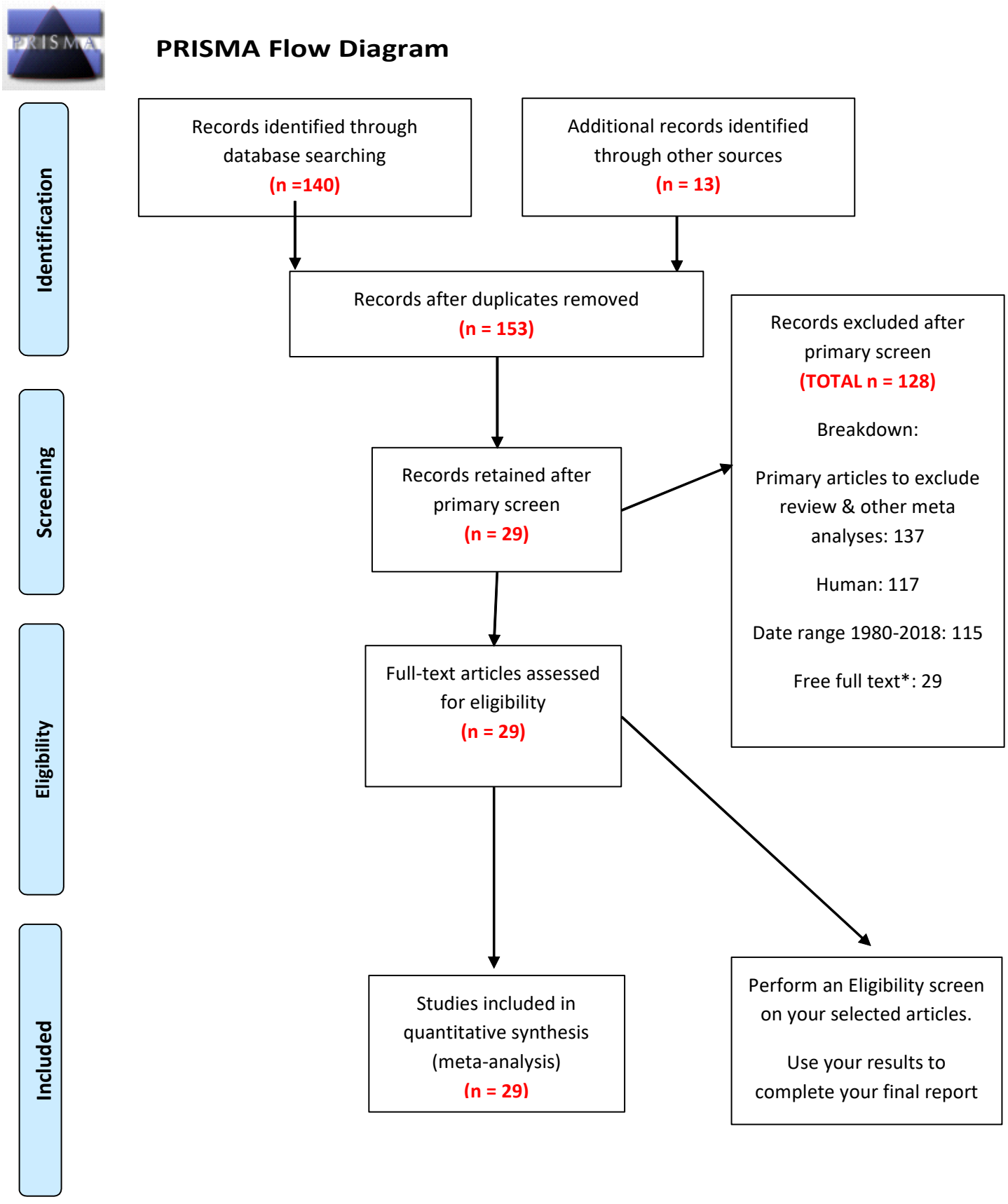
- **Clinical Trial Historical References** King Nebuchadnezzar II and the Royal Diet (605-562 BC) James Lind Treatise on the Scurvy (1753).
- **Ethics & The Modern History of Clinical Trials** Nuremberg – Thalidomide & the Kefauver Amendments Declaration of Helsinki – Tuskegee Syphilis Trial - Belmont report.
- **Bradford-Hill Criteria** – limiting bias in clinical trials.
- **When Clinical Trials go wrong** – Pennsylvania Adenovirus OTC Trial, TGN1412 Northwick Park Hospital.
- **Ethical developments for the next generation** Open discussion Animals in experimentation and the 3Rs.

- **An introduction to Meta Analysis** Purpose and origins.
- **Statistical Power** & evaluating effects across studies.
- **Use of PRISMA** & reporting of eligibility criteria when evaluating studies for inclusion in meta analysis.
- **Extracting the data.**
- **Interpreting Forest plots.**
- **Dealing with intra-study heterogeneity & interpreting funnel plots.**

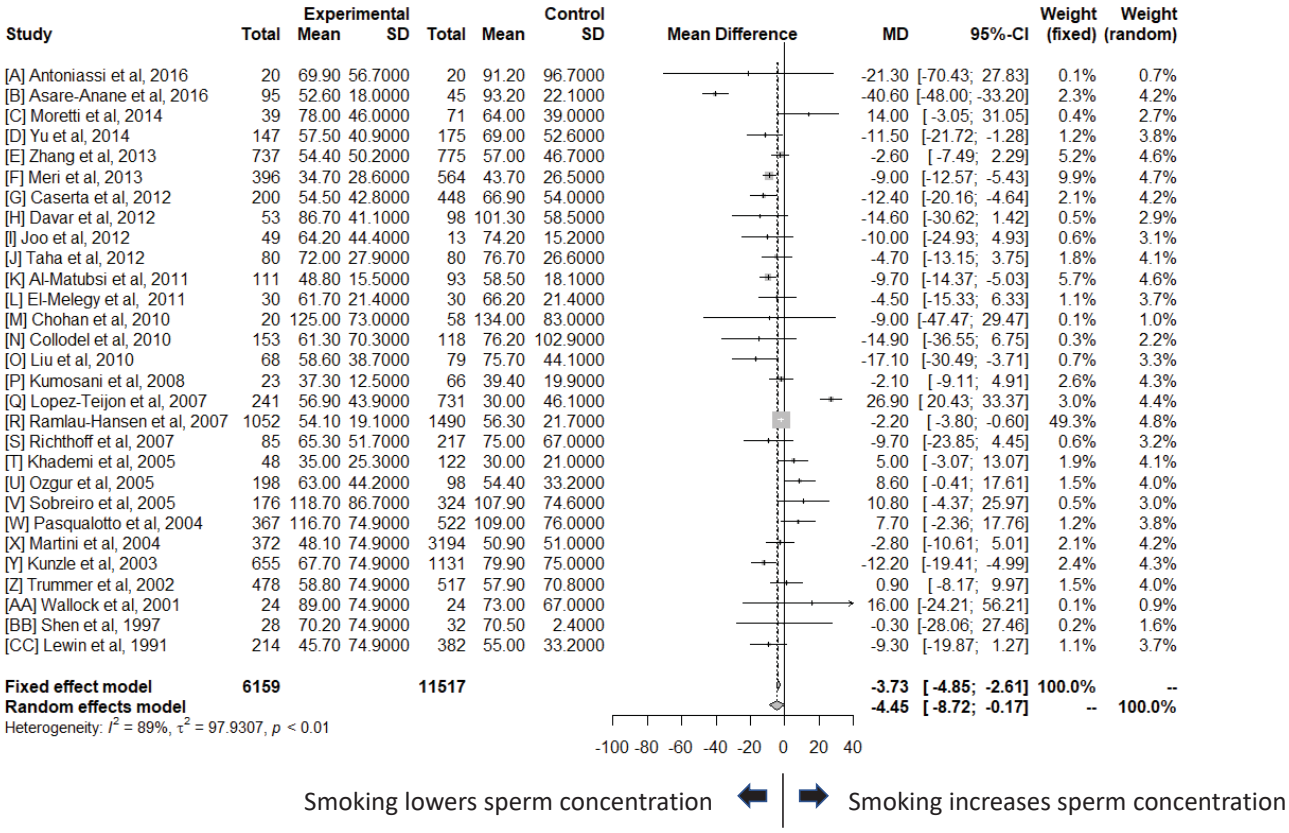
- **Creation of Forest & Funnel plots** using categorical and continuous data.
- **Meta Analysis with Categorical Data:** BCG vaccination reduces risk of tuberculosis (TB) in children. Focus on Roy *et al*, 2014 (24).
- **Meta Analysis with Continuous Data:** Do antibiotic prescribing interventions improve patient recovery times? Focus on Davey *et al*, 2017 (9).
- **Designing your study question.**

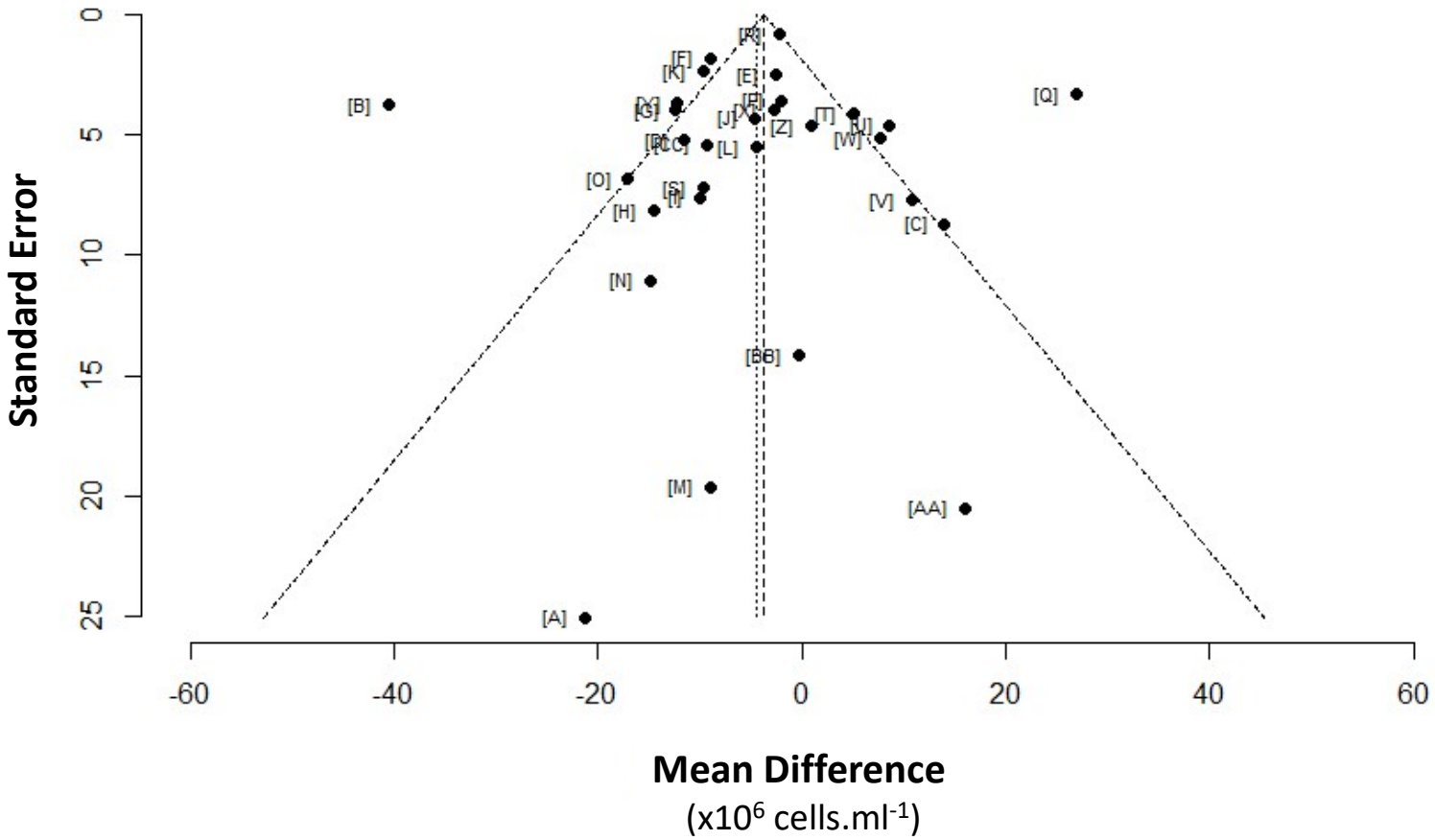
- **Independent systematic review & meta analysis** using clinical trial literature.
- **Proforma Style Report (Meta Analysis) –** 4.5 credits (45 hours of effort).
- **CMA - ethical questions arising from clinical trials** (incorporated into wider assessment).

PRISMA Flow Diagram



Reference	Cross-sectional Study	Prospective Study	Retrospective Study	Blinded Participant Selection	Randomised Study Design	Blinded Analysis	Inter-observer coefficient of variation reported	Independent Statistical Analysis	Power Analysis Performed	Biological Gradient (Exposure in pack years)	Excludes Subjects with Pathologies Related to Infertility	Ethical Statement	Cites Standardised Methodologies before 2010	Cites Standarised Methodology (WHO 5th Edition, 2010)
[A] Antoniassi et al, 2016	✓			×	×	×	×	×	×	×	✓	✓	×	✓
[B] Asare-Anane et al, 2016	✓			×	×	×	×	×	✓	✓	✓	✓	×	✓
[C] Moretti et al, 2014		✓		×	×	×	×	×	×	✓	✓	✓	×	✓
[D] Yu et al, 2014		✓		?	?	×	×	×	×	✓	✓	✓	×	✓
[E] Zhang et al, 2013			✓	×	×	×	×	×	×	✓	✓	✓	✓	×
[F] Meri et al, 2013			✓	×	×	?	×	×	×	✓	✓	×	✓	×
[G] Caserta et al, 2012		✓		×	×	×	×	×	×	×	✓	✓	×	✓
[H] Davar et al, 2012	✓			×	×	×	×	×	×	×	✓	×	×	✓
[I] Joo et al, 2012	✓			✓	✓	✓	×	✓	×	✓	✓	✓	✓	×
[J] Taha et al, 2012		✓		×	×	×	×	×	×	✓	✓	✓	×	✓
[K] Al-Matubsi et al, 2011	✓			✓	×	×	×	✓	×	×	✓	✓	×	✓
[L] El-Melegy et al, 2011			✓	×	×	×	×	×	×	✓	✓	✓	✓	×
[M] Chohan et al, 2010	✓			×	×	×	×	×	×	×	×	✓	✓	×
[N] Collodel et al, 2010			✓	×	×	×	×	×	×	✓	✓	✓	✓	×
[O] Liu et al, 2010		✓		×	×	×	×	×	×	✓	✓	✓	✓	×
[P] Kumosani et al, 2008	✓			✓	×	×	×	×	×	×	✓	×	✓	×
[Q] Lopez-Teijon et al, 2007		✓		✓	×	?	✓	×	×	✓	?	✓	✓	×
[R] Ramlau-Hansen et al, 2007		✓		✓	×	×	×	×	×	✓	✓	✓	✓	×
[S] Richthoff et al, 2007	✓			×	×	×	✓	×	×	✓	✓	✓	✓	×
[T] Khademi et al, 2005		✓		×	×	✓	×	×	×	✓	✓	✓	×	×
[U] Ozgur et al, 2005			✓	×	×	×	×	×	×	×	✓	✓	×	×
[V] Sobreiro et al, 2005		✓		×	×	×	×	×	×	✓	✓	×	✓	×
[W] Pasqualotto et al, 2004			✓	×	×	×	×	×	×	✓	✓	✓	?	?
[X] Martini et al, 2004			✓	×	×	×	×	×	×	✓	✓	×	✓	×
[Y] Kunzle et al, 2003		✓		×	×	×	×	×	×	×	✓	✓	✓	×
[Z] Trummer et al, 2002		✓		×	×	×	×	✓	×	✓	✓	✓	✓	×
[AA] Wallock et al, 2001	✓			×	×	×	×	×	×	✓	✓	✓	✓	×
[BB] Shen et al, 1997		✓		×	×	×	×	×	×	×	✓	×	✓	×
[CC] Lewin et al, 1991	✓			×	×	×	×	×	×	×	✓	×	×	×

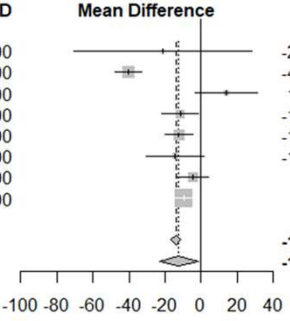

Study
Experimental Total Mean SD
Control Total Mean SD
Mean Difference
MD 95%-CI
Weight (fixed)
Weight (random)
[A] Antoniassi et al, 2016 20 69.90 56.7000 20 91.20 96.7000 -21.30 [-70.43; 27.83] 0.1% 0.7%
[B] Asare-Anane et al, 2016 95 52.60 18.0000 45 93.20 22.1000 -40.60 [-48.00; -33.20] 2.3% 4.2%
[C] Moretti et al, 2014 39 78.00 46.0000 71 64.00 39.0000 14.00 [-3.05; 31.05] 0.4% 2.7%
[D] Yu et al, 2014 147 57.50 40.9000 175 69.00 52.6000 -11.50 [-21.72; -1.28] 1.2% 3.8%
[E] Zhang et al, 2013 737 54.40 50.2000 775 57.00 46.7000 -2.60 [-7.49; 2.29] 5.2% 4.6%
[F] Meri et al, 2013 396 34.70 28.6000 564 43.70 26.5000 -9.00 [-12.57; -5.43] 9.9% 4.7%
[G] Caserta et al, 2012 200 54.50 42.8000 448 66.90 54.0000 -12.40 [-20.16; -4.64] 2.1% 4.2%
[H] Davar et al, 2012 53 86.70 41.1000 98 101.30 58.5000 -14.60 [-30.62; 1.42] 0.5% 2.9%
[I] Joo et al, 2012 49 64.20 44.4000 13 74.20 15.2000 -10.00 [-24.93; 4.93] 0.6% 3.1%
[J] Taha et al, 2012 80 72.00 27.9000 80 76.70 26.6000 -4.70 [-13.15; 3.75] 1.8% 4.1%
[K] Al-Matubsi et al, 2011 111 48.80 15.5000 93 58.50 18.1000 -9.70 [-14.37; -5.03] 5.7% 4.6%
[L] El-Melegy et al, 2011 30 61.70 21.4000 30 66.20 21.4000 -4.50 [-15.33; 6.33] 1.1% 3.7%
[M] Chohan et al, 2010 20 125.00 73.0000 58 134.00 83.0000 -9.00 [-47.47; 29.47] 0.1% 1.0%
[N] Collodel et al, 2010 153 61.30 70.3000 118 76.20 102.9000 -14.90 [-36.55; 6.75] 0.3% 2.2%
[O] Liu et al, 2010 68 58.60 38.7000 79 75.70 44.1000 -17.10 [-30.49; -3.71] 0.7% 3.3%
[P] Kumosani et al, 2008 23 37.30 12.5000 66 39.40 19.9000 -2.10 [-9.11; 4.91] 2.6% 4.3%
[Q] Lopez-Teijon et al, 2007 241 56.90 43.9000 731 30.00 46.1000 26.90 [20.43; 33.37] 3.0% 4.4%
[R] Ramlau-Hansen et al, 2007 1052 54.10 19.1000 1490 56.30 21.7000 -2.20 [-3.80; -0.60] 49.3% 4.8%
[S] Richthoff et al, 2007 85 65.30 51.7000 217 75.00 67.0000 -9.70 [-23.85; 4.45] 0.6% 3.2%
[T] Khademi et al, 2005 48 35.00 25.3000 122 30.00 21.0000 5.00 [-3.07; 13.07] 1.9% 4.1%
[U] Ozgur et al, 2005 198 63.00 44.2000 98 54.40 33.2000 8.60 [-0.41; 17.61] 1.5% 4.0%
[V] Sobreiro et al, 2005 176 118.70 86.7000 324 107.90 74.6000 10.80 [-4.37; 25.97] 0.5% 3.0%
[W] Pasqualotto et al, 2004 367 116.70 74.9000 522 109.00 76.0000 7.70 [-2.36; 17.76] 1.2% 3.8%
[X] Martini et al, 2004 372 48.10 74.9000 3194 50.90 51.0000 -2.80 [-10.61; 5.01] 2.1% 4.2%
[Y] Kunzle et al, 2003 655 67.70 74.9000 1131 79.90 75.0000 -12.20 [-19.41; -4.99] 2.4% 4.3%
[Z] Trummer et al, 2002 478 58.80 74.9000 517 57.90 70.8000 0.90 [-8.17; 9.97] 1.5% 4.0%
[AA] Wallock et al, 2001 24 89.00 74.9000 24 73.00 67.0000 16.00 [-24.21; 56.21] 0.1% 0.9%
[BB] Shen et al, 1997 28 70.20 74.9000 32 70.50 2.4000 -0.30 [-28.06; 27.46] 0.2% 1.6%
[CC] Lewin et al, 1991 214 45.70 74.9000 382 55.00 33.2000 -9.30 [-19.87; 1.27] 1.1% 3.7%
Fixed effect model 6159 11517 -3.73 [-4.85; -2.61] 100.0% --
Random effects model -4.45 [-8.72; -0.17] -- 100.0%
Heterogeneity: I^2 = 89%, τ^2 = 97.9307, p < 0.01
-100 -80 -60 -40 -20 0 20 40
Smoking lowers sperm concentration
Smoking increases sperm concentration


Standard Error
0
5
10
15
20
25
[A]
[B]
[C]
[E]
[F]
[H]
[K]
[L]
[M]
[N]
[O]
[Q]
[T]
[V]
[W]
[Z]
[AA]
-60
-40
-20
0
20
40
60
Mean Difference
(x10^6 cells.ml^-1)

A

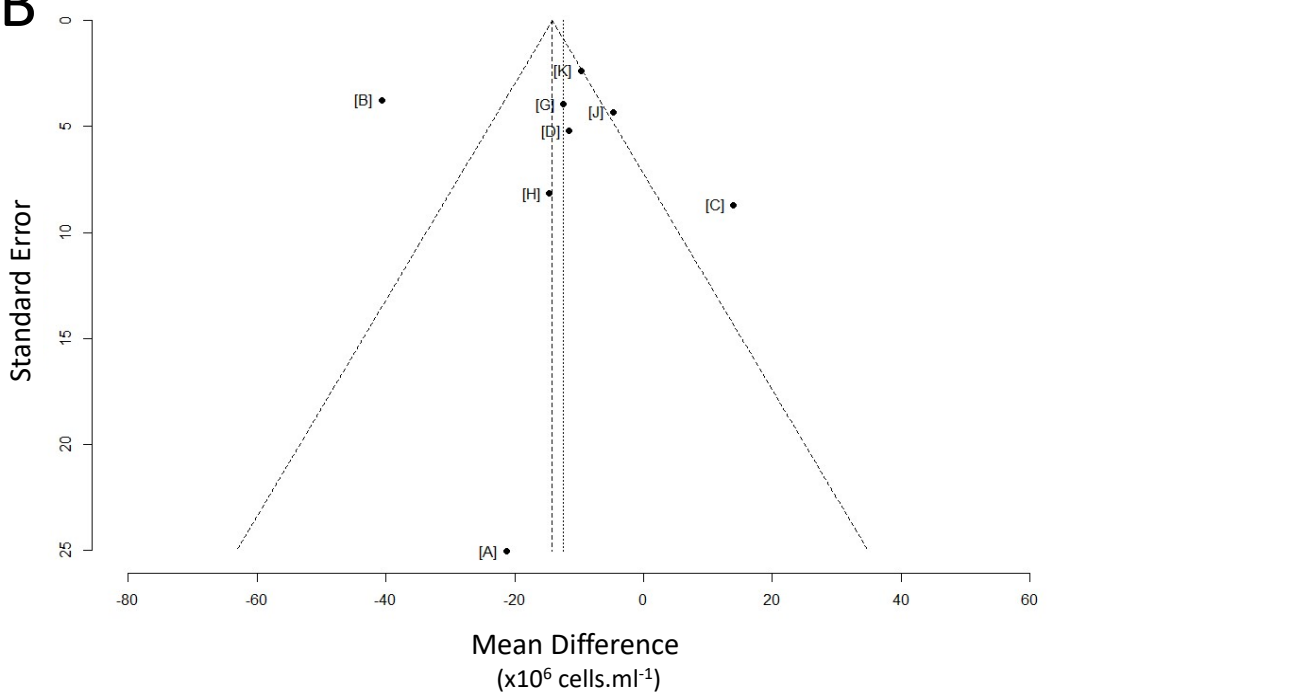
Study	Experimental Total	Mean	SD	Control Total	Mean	SD	MD	95%-CI	Weight (fixed)	Weight (random)
[A] Antoniassi et al, 2016	20	69.90	56.7000	20	91.20	96.7000	-21.30	[-70.43; 27.83]	0.4%	3.7%
[B] Asare-Anane et al, 2016	95	52.57	18.0209	45	93.17	22.1000	-40.60	[-48.00; -33.19]	16.3%	14.8%
[C] Moretti et al, 2014	39	78.00	46.0000	71	64.00	39.0000	14.00	[-3.05; 31.05]	3.1%	11.3%
[D] Yu et al, 2014	147	57.50	40.9000	175	69.00	52.6000	-11.50	[-21.72; -1.28]	8.6%	13.9%
[G] Caserta et al, 2012	200	54.50	42.8000	448	66.90	54.0000	-12.40	[-20.16; -4.64]	14.8%	14.7%
[H] Davar et al, 2012	53	86.70	41.0600	98	101.30	58.5000	-14.60	[-30.61; 1.41]	3.5%	11.7%
[J] Taha et al, 2012	80	72.00	27.9000	80	76.70	26.6000	-4.70	[-13.15; 3.75]	12.5%	14.5%
[K] Al-Matubsi et al, 2011	111	48.80	15.5000	93	58.50	18.1000	-9.70	[-14.37; -5.03]	40.9%	15.4%
Fixed effect model	**745**			**1030**			**-14.15**	**[-17.14; -11.16]**	**100.0%**	**--**
Random effects model							**-12.50**	**[-23.21; -1.79]**	**--**	**100.0%**

Heterogeneity: I^2 = 90%, τ^2 = 187.7321, p < 0.01

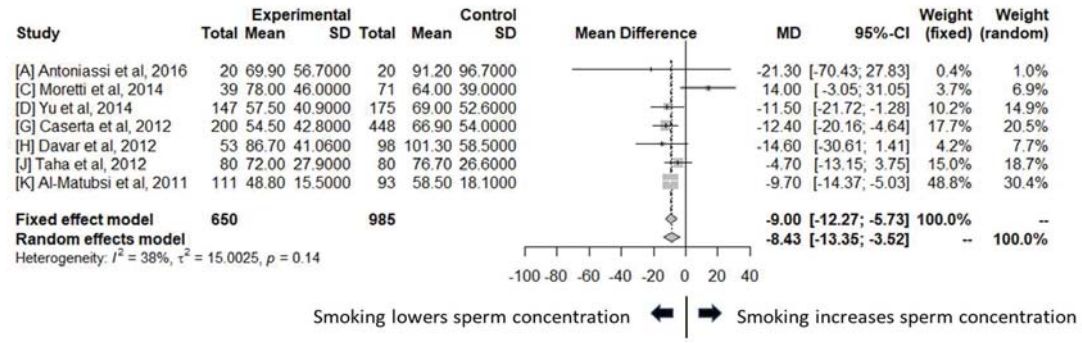


Smoking lowers sperm concentration ⬅ | ➡ Smoking increases sperm concentration

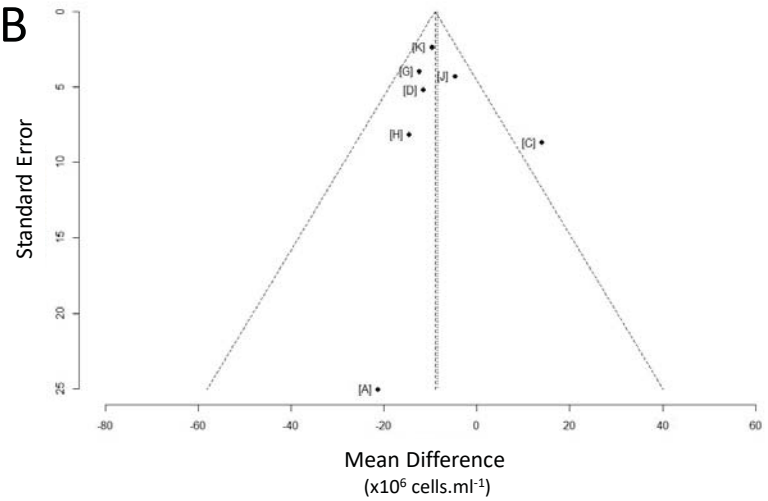
B

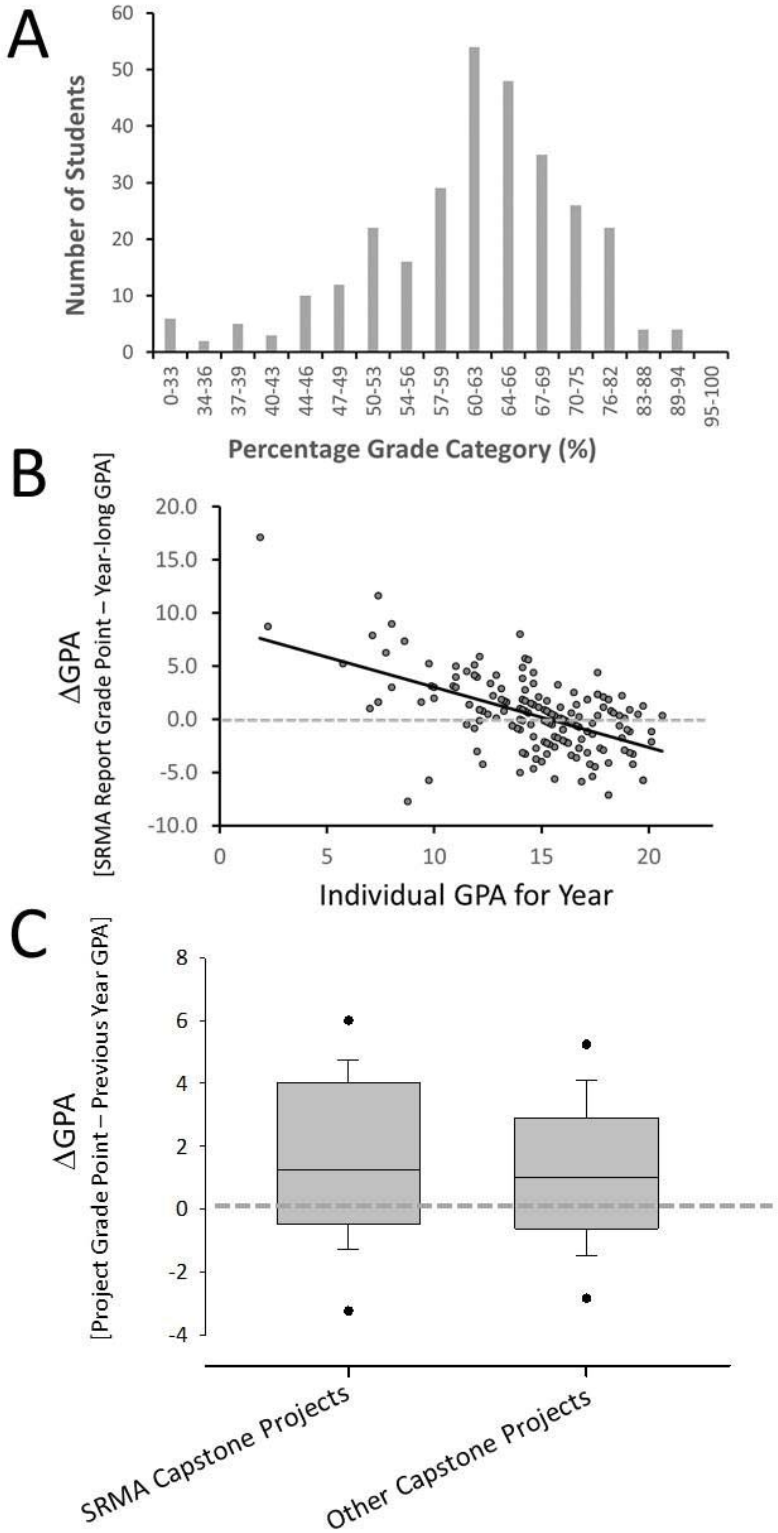


A


Study
Experimental
Total Mean SD
Control
Total Mean SD
Mean Difference
MD 95%-CI
Weight (fixed)
Weight (random)
[A] Antoniassi et al, 2016 20 69.90 56.7000 20 91.20 96.7000 -21.30 [-70.43; 27.83] 0.4% 1.0%
[C] Moretti et al, 2014 39 78.00 46.0000 71 64.00 39.0000 14.00 [-3.05; 31.05] 3.7% 6.9%
[D] Yu et al, 2014 147 57.50 40.9000 175 69.00 52.6000 -11.50 [-21.72; -1.28] 10.2% 14.9%
[G] Caserta et al, 2012 200 54.50 42.8000 448 66.90 54.0000 -12.40 [-20.16; -4.64] 17.7% 20.5%
[H] Davar et al, 2012 53 86.70 41.0600 98 101.30 58.5000 -14.60 [-30.61; 1.41] 4.2% 7.7%
[J] Taha et al, 2012 80 72.00 27.9000 80 76.70 26.6000 -4.70 [-13.15; 3.75] 15.0% 18.7%
[K] Al-Matubsi et al, 2011 111 48.80 15.5000 93 58.50 18.1000 -9.70 [-14.37; -5.03] 48.8% 30.4%
Fixed effect model 650 985 -9.00 [-12.27; -5.73] 100.0% --
Random effects model -8.43 [-13.35; -3.52] -- 100.0%
Heterogeneity: I^2 = 38%, τ^2 = 15.0025, p = 0.14
-100 -80 -60 -40 -20 0 20 40
Smoking lowers sperm concentration
Smoking increases sperm concentration

B


Standard Error
0
5
10
15
20
25
[K]
[G]
[J]
[D]
[H]
[C]
[A]
-80
-60
-40
-20
0
20
40
60
Mean Difference
(x10^6 cells.ml^{-1})


A
Number of Students
60
50
40
30
20
10
0
0-33
34-36
37-39
40-43
44-46
47-49
50-53
54-56
57-59
60-63
64-66
67-69
70-75
76-82
83-88
89-94
95-100
Percentage Grade Category (%)
B
ΔGPA
[SRMA Report Grade Point – Year-long GPA]
20.0
15.0
10.0
5.0
0.0
-5.0
-10.0
0
5
10
15
20
Individual GPA for Year
C
ΔGPA
[Project Grade Point – Previous Year GPA]
8
6
4
2
0
-2
-4
SRMA Capstone Projects
Other Capstone Projects