Using Bayes Factors for testing hypotheses about intervention effectiveness in addictions research

Article (Accepted Version)

Using Bayes Factors for testing hypotheses about intervention effectiveness in addictions research

Emma Beard[1,2], Zoltan Dienes[3], Colin Muirhead[4], Robert West[1]

[1] Research Department of Clinical, Educational and Health Psychology, University College London, London

[2] Department of Epidemiology and Public Health, University College London, London

[3] School of Psychology, University of Sussex, Brighton

[4] Institute of Health and Society, Newcastle University, Newcastle upon Tyne

Journal: Addiction

Word count: 1822

Correspondence to: Emma Beard, Cancer Research UK Health Behaviour Research Centre, University College London, WC1E 6BP. Email: e.beard@ucl.ac.uk. Tel: 02031083179

Abstract: (n=293)

Background and aims: It has been proposed that more use should be made of Bayes Factors in hypothesis testing in addiction research. Bayes Factors are the ratios of the likelihood of a specified hypothesis (e.g. an intervention effect within a given range) to another hypothesis (e.g. no effect). They are particularly important for differentiating lack of strong evidence for an effect and evidence for lack of an effect. This paper reviewed randomised trials reported in Addiction between January and June 2013 to assess how far Bayes Factors might improve the interpretation of the data.

Methods: Seventy five effect sizes and their standard errors were extracted from 12 trials. Seventy three per cent (n=55) of these were non-significant (i.e. p>0.05). For each non-significant finding a Bayes Factor was calculated using a population effect derived from previous research. In sensitivity analyses, a further two Bayes Factors were calculated assuming clinically meaningful and plausible ranges around this population effect.

Results: Twenty per cent (n=11) of the non-significant Bayes Factors were < 1/3rd and 3.6% (n=2) were > 3. The other 76.4% (n=42) of Bayes Factors were between 1/3rd and 3. Of these, 26 were in the direction of there being an effect (Bayes Factor >1 & <3); 12 tended to favour the hypothesis of no effect (Bayes Factor <1 & >1/3rd); and for 4 there was no evidence either way (Bayes Factor =1). In sensitivity analyses, 13.3% of Bayes Factors were <1/3rd (n=20), 62.7% (n=94) were between 1/3rd and 3 and 24.0% (n=36) were >3, showing good concordance with the main results.

Conclusions: Use of Bayes Factors when analysing data from randomised trials of interventions in addiction research can provide important information that would lead to more precise conclusions than are typically obtained using currently prevailing methods.

Introduction

Bayesian statistical analyses are being increasingly used in addictions research and it has been proposed that this trend should accelerate (1). One important component of Bayesian analysis is the calculation of Bayes Factors, which overcome many of the problems of traditional frequentist statistics (2). One of these is the misinterpretation that p-values can be used to make claims of 'no effect' (3-5). P-values signal the extremeness of the data under the assumption of the null hypothesis and so only tell us the probability of a test statistic at least as extreme as the one observed (6). Thus, a p>0.05 may reflect evidence for 'no effect' or data insensitivity i.e. a failure to distinguish the null hypothesis from the alternative because, for example, the standard error (SE) is high.

Bayes Factors are the ratio of the (average) likelihood of two hypotheses being correct given a set of data. When evaluating interventions, the two hypotheses are typically H_1: that the intervention had a desired effect (for a given range of plausible sizes), or within a certain range, versus H_0: that it had no effect. Thus a Bayes Factor is equivalent to a likelihood ratio (7) (averaged over different plausible effect sizes) and thus is often denoted as:

$$Bayes\ Factor = \frac{Average\ likelihood\ of\ data\ given\ H_1}{likelihood\ of\ data\ given\ \ H_0} = \frac{P(D|H_1)}{P(D|H_0)}$$

which simply represents the probability of the data (D) given the alternative hypothesis divided by the probability of the data given the null hypothesis.

Use of Bayes Factors has become more feasible in recent years following the development of online calculators (8) and R code (9, 10). Conventional cut-offs for the interpretation of Bayes Factors typically depend on those set by Jeffreys (2) in the 1930s, with a Bayes factor greater than 3, or else less than 1/3, representing sufficient evidence to be taken note of for the experimental and null hypotheses respectively; while values between roughly 1/3 and 3 indicate that the data are insensitive (see Table 1).

Table 1: Jeffreys' Bayes Factor cut-offs

Bayes Factor	**Interpretation**
>100	Extreme evidence for the experimental hypothesis
30-100	Very strong evidence for the experimental hypothesis
10-30	Strong evidence for the experimental hypothesis
3-10	Moderate evidence for the experimental hypothesis
1-3	Anecdotal evidence for the experimental hypothesis
1	No evidence
1/3-1	Anecdotal evidence for the null hypothesis
1/3-1/10	Moderate evidence for the null hypothesis
1/10-1/30	Strong evidence for the null hypothesis
1/30-1/100	Very strong evidence for the null hypothesis
<1/100	Extreme evidence for the null hypothesis

Note: The original label for 3< Bayes Factor<10 was "substantial evidence". Lee & Wagenmakers changed it to moderate as they thought the original label sounded too decisive (3, 11).

This paper uses a set of randomised trials in the field of addiction to examine whether, and in what way, the conclusions may have been different had the authors calculated Bayes Factors in their analyses. This should be useful in future research to assess whether and when to use this form of analysis.

Calculating Bayes Factors

Several software packages are available including an online calculator developed by Zoltan Dienes (http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm) and a modified version by John Christie using R code which allows one to adjust the quality of the estimation (9, 10).

Both approaches require the specification of an expected effect size (i.e. a plausible range of predicted values based on previous studies, judgement or clinical significance), the published effect size (e.g. mean difference or log odds ratio) and standard error of this parameter. They also both assume that the sampling distribution of the parameter estimate is normally distributed (hence the need to use the natural logs of odds ratios). The natural log of the odds ratio is approximately normally distributed with known standard error given by $\sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}$, where A is the number of individuals in the experimental condition with the outcome of interest, B is the number of individuals in the experimental condition without the outcome of interest, and C and D reflect the number of individuals with and without the outcome of interest in the control condition respectively (i.e. Odds Ratio = (A/B)/(C/D)), provided that these numbers are not very small. For adjusted Odds Ratios, and/or where standard errors are not reported, 95% confidence intervals can be used to derive the standard error (i.e. [LN(upper confidence interval)-LN(lower confidence interval)]/3.92).

In instances where the primary outcome measure is a continuous variable, standard errors can be derived for mean differences or regression coefficients (β) either using the standard formula for the SE of mean difference, i.e. $[(SD^2_{control}/N_{Control})+(SD^2_{experimental}/N_{experimental})]$; or t-test values using [Mean difference (or β)/t-test value]; or 3) 95% Confidence Intervals: [LN(upper confidence interval)-LN(lower confidence interval)]/3.92).

A worked example, using the calculator associated with Dienes, can be found in supplementary Appendix 1.

Others have advocated alternative methods of computing Bayes Factors, including the Jeffreys-Zellner-Siow (JZS) t-test (4, 12) which can be implemented in R (13, 14) (see Dienes & McLatchie, submitted, for comparison). Moves have also been made towards full Bayesian modelling, which requires a much more advanced knowledge of R or specialist software packages, and is beyond the scope of the current paper (e.g. WinBUGS) (3, 11).

Methods

Bayes Factors were calculated for 12 randomised controlled trials published in the first six issues of Addiction in 2013 (between January and June). Effect sizes, standard errors, p–values, and the main conclusions drawn by the authors, were extracted from the papers for both primary and main secondary outcomes. Studies are generally only powered to detect estimated differences between experimental and control groups for the primary outcome, and thus Bayes Factors may be particularly useful for secondary analyses (15, 16). Concerns have previously been raised regarding the interpretation of non-significant findings for sensitivity analyses (15, 16).

Adjusted effect sizes (where available) and those reported at the longest point of follow-up were used. Bayes Factors were calculated using the online calculator provided by Dienes (8) and the modified version using R code by Christie (9, 10). Predicted values for the effect size or population SD were based on previous studies (see Table 2). Additional sensitivity analyses were run to assess the effect of using higher

and lower values. The chosen range was based either on the reported confidence interval of the predicted effect size selected from previous publications, or when not available, the opinion of the lead author as to what would be a plausible effect.

When specifying the predicted effect, we used a 'half normal distribution' whose peak was at 0 (no effect) and extending upwards with a standard deviation equal to the expected effect size. This represents a hypothesis that the intervention had at least some positive effect with the effect being more likely to be smaller than larger. This is a conservative approach to prediction. Another approach would be to specify the hypothesis as a uniform distribution between 0 (or a minimally clinically significant value) and a plausible upper bound. Given that none of the authors of the studies reviewed indicated what they considered to be a clinically meaningful effect or a plausible upper bound for the effect size, we took the conservative approach.

Results

Out of the 12 studies, 55 non-significant effects and 20 significant effects were reported. For each of these, three Bayes Factors were calculated: one based on an expected population SD (identified from previous studies) and two based on a range of values around the expected population SD (identified from previous studies or based on expert opinion). Thus a total of 75 Bayes Factors were calculated in the main analysis and 150 Bayes Factors were derived in the sensitivity analysis (see Table 2).

Fifty-six per cent (n=42) of the Bayes Factors were between 1/3rd and 3; 14.7% (n=11) were < 1/3rd and 29.3% (n=22) were > 3. When considering only the non-significant findings (n=55), 20.0% (n=11) of Bayes Factors were < 1/3rd and 3.6% (n=2) were > 3. The other 76.4% (n=42) of Bayes Factors were between 1/3rd and 3. Of these, 26 were in the direction of there being an effect (Bayes Factor >1 & <3); 12 tended to favour the hypothesis of no effect (Bayes Factor <1 & >1/3rd); and for 4 there was no evidence either way (Bayes Factor =1).

In sensitivity analyses, 13.3% of Bayes Factors were <1/3rd (n=20), 62.7% (n=94) were between 1/3rd and 3 and 24.0% (n=36) were >3, showing good consistency with the main results.

Authors either decided not to discuss results where p>0.05, to report them as non-significant, and/or to state that no association was found. Good concordance was noted between the online calculator (8) and the adapted R code (9), except for those Bayes Factors that indicated extreme evidence for the experimental hypothesis.

Discussion

Only 1/5th of all non-significant findings provided support for the hypothesis of no effect; while nearly 2/3rds of the Bayes Factors indicated data insensitivity. Thus, reporting 'no difference' between conditions or lack of associations was only appropriate for a small number of papers. A minority of Bayes Factors for the non-significant effects also supported the experimental hypothesis; this tended to occur with p-values close to statistical significance.

The development of online calculators and R code (9, 10) means that researchers in the addiction field can easily calculate Bayes Factors to include as an adjunct to traditional frequentist results. The requirement to specify the experimental hypothesis means that scientific judgment is needed. This is a common criticism of Bayesian type methods (17), but it can also be a potential strength because it forces researchers to be specific about what it is they are testing. Moreover, if there are differences of view about what may be plausible values of the effect size, it is a simple matter to conduct sensitivity analyses to assess what difference this makes if any. As a rule of thumb, if one is interested in a clinically relevant range then the uniform distribution can be specified; alternatively one can use a half-normal distribution with the peak at 0 if one is interested in any effect at all and has little confidence in the likely value. To prevent researcher

bias, pre-specified analysis plans may be published which detail the method which will be used to calculate Bayes Factors, the cut-off values for interpretation and the plausible effect size which is expected.

The findings of this review show that researchers should avoid the use of terms such as 'no difference' or 'lack of associations' for p-values >0.05, unless a Bayes Factor <0.3 is also found. Otherwise null findings should be framed as 'the findings were inconclusive as to whether or not a difference/association was present' or some similar wording. This is now encouraged practice by the journal, Addiction (1). Researchers may also wish to use Bayes Factors in order to quantify the evidence for the experimental hypothesis (i.e. moderate, strong, very strong and extreme) and/or use such a calculation as a stopping rule for data collection (18). For ethical and perhaps financial reasons interim analyses are often planned for randomised trials, with early stopping occurring if there is demonstrated efficacy, the intervention is harmful, or there is no beneficial effect. P-values cannot inform about us about the latter; in contrast a Bayes Factor indicating data insensitivity would suggest further recruitment, while a Bayes Factor indicating evidence for the null hypothesis may point towards early termination.

Note that the methods used to derive Bayes Factors in this paper did not cover all the possibilities. More advanced Bayesian hierarchical modelling (BHM) (11), implemented in R and winBUGS, allows a wider range of distributions e.g. gamma, Poisson, binomial and negative binomial.

Acknowledgements

RW salary is funded by Cancer Research UK (CRUK). EB is funded by CRUK and by the National institute for Health Research (NIHR)'s School for Public Health Research (SPHR). The views are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. SPHR is a partnership between the Universities of Sheffield; Bristol; Cambridge; Exeter; UCL; The London School for Hygiene and Tropical Medicine; the LiLaC collaboration between the Universities of Liverpool and Lancaster and Fuse; The Centre for Translational Research in Public Health, a collaboration between Newcastle, Durham, Northumbria, Sunderland and Teesside Universities.

Conflicts of interest

EB has received unrestricted funding from Pfizer. RW undertakes consultancy and research for and receives travel funds and hospitality from manufacturers of smoking cessation medications but does not, and will not take funds from EC manufacturers or the tobacco industry. RW is an honorary co-director of the National Centre for Smoking Cessation and Training and a Trustee of the stop-smoking charity, QUIT. ZD has no conflicts of interest to declare.

Table 2: *Results, conclusions and corresponding Bayes Factors for RCTs published in the Journal of Addiction in the first 6 issues of 2013*

Study	Intervention	Control	Participants	Outcome	Sample mean	Sample standard error	Significance *p*	Study conclusions Results conclusions for non-significant findings	Expected effect size	Bayes factor: Dienes (Christie) (8-10)	Interpretation of Bayes Factor using Dienes (8)	Interpretation of Bayes Factors using Jeffreys (2)
Kypri (19)	Web based alcohol screening and brief intervention for reducing hazardous drinking among Maori university students	Screening only	6,697 students aged 17-24	P: Frequency of alcohol consumption	RaR 0.89	0.04	0.01**	**"Web-based screening and brief intervention reduced hazardous and harmful drinking among non-help-seeking Maori students"** No mention of results >0.05	RaR 0.91[a] RaR 0.85[Ω] RaR 0.97[Ω]	17.5 (17.5) 16.0 (16.0) 5.3 (5.3)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect)	Strong evidence for experimental hypothesis Strong evidence for experimental hypothesis Moderate evidence for experimental hypothesis
				P: Quantity of alcohol	RaR 0.92	0.04	0.04*		RaR 0.96[a] RaR 0.91[Ω] RaR 0.99[Ω]	3.0 (3.0) 3.4 (3.4) 1.4 (1.4)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence is insensitive	Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				P: Volume of alcohol	RaR 0.78	0.06	<0.001***		RaR 0.89[a] RaR 0.82[Ω] RaR 0.96[Ω]	261.6 (261.3) 475.0 (466.2) 13.2 (13.2)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect)	Extreme evidence for experimental hypothesis Extreme evidence for experimental hypothesis Moderate evidence for experimental hypothesis
				P: Academic Role Expectation and Alcohol Scale (AREAS)	RaR 0.81	0.08	0.01*		RaR 0.95[a] RaR 0.82[Ω] RaR 0.99[Ω]	3.9 (3.9) 13.1 (13.1) 1.3 (1.3)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence is insensitive	Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: Binge drinking	OR 0.80	0.12	0.06		OR 0.89[a] OR 0.65 [Ω] OR 0.99 [Ω]	3.2 (3.2) 2.8 (2.8) 1.1 (1.1)	Evidence for experimental hypothesis (i.e. an effect) Evidence is insensitive Evidence is insensitive	Moderate evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: Heavy drinking	OR 0.65	0.15	<0.001***		OR 0.55[a] OR 0.38 [Ω] OR 0.80 [Ω]	19.0 (19.0) 13.9 (13.9) 15.5 (15.5)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect)	Strong evidence for experimental hypothesis Strong evidence for experimental hypothesis Strong evidence for experimental hypothesis
Li (20)	Methadone maintenance therapy (MMT) care intervention (with motivational interviewing)	Standard care	41 providers and 179 clients from six clinics	P: Provider client interaction	MD 4.82	2.23	0.033*	**"The MMT CARE intervention targeting providers in methadone maintenance clinics can improve providers' treatment knowledge and their interaction with clients. The intervention can also reduce clients' drug-using behaviour through motivational interviewing sessions conducted by trained providers . . . It is difficult to explain the unexpected findings in provider MMT knowledge and client drug avoidance self-efficacy [long term]; this may be a result of the small sample size and the pilot nature of the study"** No mention of results >0.05	MD 4.65[b] MD 2.18 [Ω] MD 7.01 [Ω]	5.6 (5.6) 4.2 (4.2) 4.9 (4.9)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect)Evidence for experimental hypothesis (i.e. an effect)	Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis
				P: MMT knowledge	MD 1.00	0.56	0.544		MD 4.65[b] MD 2.18 [Ω] MD 7.01 [Ω]	1.1 (1.1) 2.1 (2.1) 0.7 (0.7)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for null hypothesis
				P: Perceived stigma	MD -1.87	2.31	0.421		MD -5.1[c] MD -1.2 [Ω] MD -9.0 [Ω]	0.8 (0.8) 1.2 (1.2) 0.5 (0.5)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for null hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for null hypothesis
				P: Perceived client support	MD 1.82	0.65	0.006**		MD 4.65[b] MD 2.18 [Ω] MD 7.01 [Ω]	12.9 (12.9) 20.8 (20.8) 8.9 (8.9)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect)	Strong evidence for experimental hypothesis Strong evidence for experimental hypothesis Moderate evidence for experimental hypothesis
				P: Drug avoidance self-efficacy	MD 1.25	1.24	0.312		MD 0.9[d] MD 0.3 [Ω] MD 1.5 [Ω]	1.4 (1.4) 1.2 (1.2) 1.4 (1.3)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				P: Concurrent drug use	OR 0.36	0.59	0.084		OR 0.66[e] OR 0.56[Ω] OR 0.78 [Ω]	2.3 (2.3) 2.7 (2.7) 1.7 (1.7)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
Ward (21)	Behavioural support and nicotine replacement therapy (NRT)	Behavioural support	269 adults in four primary care clinics	P: 12 month prolonged abstinence	OR 0.51	0.50	0.182	**"Nicotine patches may not be effective in helping smokers in low-income countries to stop when given as an adjunct to behavioural support . . . Our results do not support the incremental value of providing NRT in addition to behavioural counselling"** "Between-group differences [for 12 month prolonged abstinence] were not statistically significant at follow-up . . . No significant between-group differences were found for seven-day point prevalence abstinence"	OR 1.51[f] OR 1.35 [Ω] OR 1.70 [Ω]	1.8 (1.8) 1.6 (1.6) 1.1 (1.1)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: 7-day point prevalence abstinence	OR 0.69	0.32	>0.05		OR 1.78[f] OR 1.49 [Ω] OR 2.12 [Ω]	1.4 (1.4) 1.5 (1.5) 1.2 (1.2)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
Borland (22)	OnQ: An interactive text messaging program	Minimal intervention	3530 smokers interested in quitting	P: 6-months sustained abstinence	OR 1.44	0.24	>0.05	**"Smokers interested in quitting who were assigned randomly to an offer of wither the internet-based support program and/or the intervention automated text-messaging program had a non-significantly greater odds of quitting for at least 6 months than those randomised to an offer of a single website . . . we failed to find clear significant effects between the intervention and the control"** "There were no differences in the proportion who reported making a quit attempt by the 1-month follow-up . . . At the 7-month follow up, 8.5% of the sample achieved 6-month sustained abstinence. No significant differences were found by condition, but the control condition was numerically least successful".	OR 1.50[g] OR 1.20 [α] OR 1.80 [α]	2.2 (2.2) 2.0 (2.0) 1.9 (1.9)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: 7-day point prevalence abstinence	OR 1.20	0.15	>0.05		OR 1.50[g] OR 1.20 [α] OR 1.80 [α]	1.2 (1.2) 1.6 (1.6) 0.9 (0.9)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for null hypothesis
				S: Quit attempt	OR 1.11	0.12	>0.05		OR 1.50[g] OR 1.20 [α] OR 1.80 [α]	0.6 (0.6) 1.1 (1.1) 0.4 (0.4)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for null hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for null hypothesis
	QuitCoach: Personalised tailored internet-delivered advice program	Minimal intervention		P: 6-months sustained abstinence	OR 1.40	0.24	>0.05		OR 1.50[g] OR 1.20 [α] OR 1.80 [α]	1.9 (1.9) 1.8 (1.8) 1.6 (1.6)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: 7-day point prevalence abstinence	OR 1.03	0.15	>0.05		OR 1.50[g] OR 1.20 [α] OR 1.80 [α]	0.4 (0.4) 0.7 (0.7) 0.3 (0.3)	Evidence is insensitive Evidence is insensitive Evidence for null hypothesis (i.e. no effect)	Anecdotal evidence for null hypothesis Anecdotal evidence for null hypothesis Moderate evidence for null hypothesis
				S: Quit attempt	OR 0.91	0.12	>0.05		OR 1.50[g] OR 1.20 [α] OR 1.80 [α]	0.6 (0.6) 1.0 (1.0) 0.4 (0.4)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for null hypothesis No evidence Anecdotal evidence for null hypothesis
	Integration of onQ and QuitCoach	Minimal intervention		P: 6-months sustained abstinence	OR 1.06	0.15	>0.05		OR 1.92[g] OR 1.40 [α] OR 2.40 [α]	0.3 (0.3) 0.6 (0.6) 0.2 (0.2)	Evidence for null hypothesis (i.e. no effect) Evidence is insensitive Evidence for null hypothesis (i.e. no effect)	Moderate evidence for null hypothesis Anecdotal evidence for null hypothesis Moderate evidence for null hypothesis
				S: 7-day point prevalence abstinence	OR 1.45	0.24	>0.05		OR 1.92[g] OR 1.40 [α] OR 2.40 [α]	1.8 (1.8) 2.3 (2.3) 1.5 (1.5)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: Quit attempt	OR 1.03	0.12	>0.05		OR 1.92[g] OR 1.40 [α] OR 2.40 [α]	0.2 (0.2) 0.4 (0.4) 0.2 (0.2)	Evidence for null hypothesis (i.e. no effect) Evidence is insensitive Evidence for null hypothesis (i.e. no effect)	Moderate evidence for null hypothesis Anecdotal evidence for null hypothesis Moderate evidence for null hypothesis
	Choice of either alone or combined program	Minimal intervention		P: 6-months sustained abstinence	OR 1.47	0.24	>0.05		OR 1.92[g] OR 1.40 [α] OR 2.40 [α]	2.0 (2.0) 2.5 (2.5) 1.6 (1.6)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: 7-day point prevalence abstinence	OR 1.07	0.15	>0.05		OR 1.92[g]	0.3 (0.3)	Evidence for null hypothesis (i.e. no effect)	Moderate evidence for null hypothesis

								OR 1.40[α] OR 2.40[α]	0.6 (0.6) 0.3 (0.3)	Evidence is insensitive Evidence for null hypothesis (i.e. no effect)	Anecdotal evidence for null hypothesis Moderate evidence for null hypothesis
				S: Quit attempt	OR 1.15	0.12	>0.05	OR 1.92[ε] OR 1.40[α] OR 2.40[α]	0.6 (0.6) 1.0 (1.0) 0.4 (0.4)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for null hypothesis No evidence Anecdotal evidence for null hypothesis
Study	**Intervention**	**Control**	**Participants**	**Outcome**	**Sample mean**	**Sample standard error**	**Significance *p***	**Study conclusions** Results conclusions for non-significant findings	**Expected effect size**	**Bayes factor: Dienes (Christie) (8-10)**	**Interpretation of Bayes Factor using Dienes (8)**	**Interpretation of Bayes Factors using Jeffreys (2)**
Rendall-Mkosi (23)	Motivational Interviewing	Minimal intervention	165 women aged 18-44 years at risk of alcohol exposed pregnancy	P: Alcohol exposed pregnancy	OR 0.46	0.35	0.024*	**"A five session motivational interviewing intervention was found to be effective with women at risk of an alcohol-exposed pregnancy . . . it is noteworthy that the reduction in risk for AEP in this study was mainly due to the improved contraceptive rather than a reduction in risky alcohol use"** "At the 12-month follow-up, the reduction [in risky drinking] in the MI group (14.75%) was modestly larger when compared to the control group (10.94%), but this difference was also not statistically significant . . . the reduction in the proportion of participants who were using ineffective contraception at 12 months was no longer statically significant"	OR 1.90[h] OR 1.36[Ω] OR 2.66[Ω]	6.5 (6.5) 4.2 (4.2) 6.2 (6.2)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect)	Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis
				S: Risky drinking	OR 0.75	0.53	0.580		OR 0.84[i] OR 0.70[α] OR 0.90[α]	1.1 (1.1) 1.1 (1.1) 1.1 (1.1)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: Ineffective contraception	OR 0.51	0.37	0.067		OR 0.63[i] OR 0.54[α] OR 0.74[α]	3.0 (3.0) 3.2 (3.2) 2.6 (2.6)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence is insensitive	Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
Coffin (24)	Aripiprazole	Placebo	90 methamphetamine dependent, sexually active adults from the community	P: Methamphetamine use	RR 0.88	0.15	0.410	**"Compared with placebo, apripiprazole did not reduce methamphetamine use significantly among actively dependent adults . . . notwithstanding the promising pre-clinical results suggesting that apripiprazole might be effective at decreasing craving for methamphetamine and reducing it rewarding properties, we found no effect of this medication on methamphetamine use, severity of craving. We also did not evidence that apripiprazole was associated with increased methamphetamine use or rewards, as suggested by some investigators."** "In the intention-to-treat GEE analysis, the risk of testing positive for methamphetamine was similar in the aripiprazole arm compared to the placebo arm . . . difference between arms over follow-up was not significant [in severity of dependence . . . After controlling for imbalanced baseline characteristics, sexual risk behaviors declined similarly in the aripiprazole and placebo arms."	RR 1.12[j] RR 1.02[α] RR 1.22[α]	1.3 (1.3) 1.1 (1.1) 1.1 (1.1)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: Adherence – medication event monitoring systems	RR 1.33	0.43	0.310		RR 0.99[k] RR 0.80[α] RR 1.00	1.0 (1.0) 1.2 (1.2) 0.7 (0.7)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	No evidence Anecdotal evidence for experimental hypothesis Anecdotal evidence for null hypothesis
				S: Adherence – self-reported	RR 0.59	0.49	0.170		RR 1.03[k] RR 1.01[α] RR 1.10[α]	1.1 (1.1) 1.0 (1.0) 1.2 (1.2)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis No evidence Anecdotal evidence for experimental hypothesis
				S: Number of partners with whom methamphetamines were used	RR 0.38	0.86	0.254		RR 0.45[k] RR 0.24[Ω] RR 0.82[Ω]	1.5 (1.5) 1.4 (1.4) 1.2 (1.2)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: Number of sexual partners	RR 0.69	0.46	0.418		RR 0.20[k] RR 0.04[Ω] RR 0.93[Ω]	0.2 (0.2) 0.1 (0.1) 0.9 (0.9)	Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect) Evidence is insensitive	Strong evidence for null hypothesis Strong evidence for null hypothesis Anecdotal evidence for null hypothesis
				S: Episodes of anal and/or vaginal sex with sero-discordant partners	RR 0.42	0.65	0.190		RR 0.31[k] RR 0.14[Ω] RR 0.66[Ω]	1.7 (1.7) 1.3 (1.3) 1.7 (1.7)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: Episodes of unprotected anal and/or vaginal sex with sero-discordant partners	RR 0.61	0.98	0.612		RR 0.34[k] RR 0.17[Ω] RR 0.70[Ω]	0.9 (0.9) 0.7 (0.7) 1.1 (1.1)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for null hypothesis Anecdotal evidence for null hypothesis Anecdotal evidence for experimental hypothesis
				S: Episodes of insertive unprotected anal sex with sero-discordant partners	RR 0.54	0.72	0.385		RR 0.29[k] RR 0.14[Ω] RR 0.58[Ω]	1.0 (1.0) 0.8 (0.8) 1.3 (1.3)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for null hypothesis Anecdotal evidence for experimental hypothesis
				S: Episodes of receptive unprotected anal and/or vaginal sex with sero-discordant partners	RR 0.02	1.32	0.007**		RR 0.27[k] RR 0.05[Ω] RR 0.49[Ω]	12.0 (12.0) 30.9 (30.9) 4.4 (4.4)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect)	Strong evidence for experimental hypothesis Very strong evidence for experimental hypothesis Moderate evidence for experimental hypothesis
				S: Methamphetamine craving	MD 6.8	7.65	0.380		MD 35[k] MD 8[Ω] MD 62[Ω]	0.5 (0.5) 1.3 (1.3) 0.3 (0.3)	Evidence is insensitive Evidence is insensitive Evidence for null hypothesis (i.e. no effect)	Anecdotal evidence for null hypothesis Anecdotal evidence for experimental hypothesis Strong evidence for null hypothesis
				S: Severity of dependence	MD -0.04	0.85	0.960		MD 2.00[j] MD 1.00[α] MD 3.00[α]	0.4 (0.4) 0.7 (0.7) 0.3 (0.3)	Evidence is insensitive Evidence is insensitive Evidence for null hypothesis (i.e. no effect)	Anecdotal evidence for null hypothesis Anecdotal evidence for null hypothesis Strong evidence for null hypothesis
				S: Depression	MD 1.47	2.19	0.500		MD 2.00[j] MD 1.00[α] MD 3.00[α]	1.1 (1.1) 1.2 (1.2) 1.0 (1.0)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
Gilbert (25)	Tailored cessation on advice reports, including levels of reading ability	Generic self-help booklet	58,66 current cigarette smokers aged 18-65 years, identified from general practitioner records	P: Prolonged abstinence for 3 months	OR 1.18	0.13	0.184	**"ESCAPE . . . appears to increase the rate at which smokers try to stop, but if there is an effect on prolonged abstinence it is small… Quit rates for the primary outcome of three months of prolonged abstinence were not significantly different between study groups. Thus, the intervention showed no effect. Quit rates in a number of different outcome measures of abstinence also showed no significant effect. However, all outcome measures showed a non-significant trend towards more abstinence in the intervention group"** "The difference [in 3 month prolonged abstinence] was not significant. . . No significant differences were found between the intervention and control groups on shorter periods or on point-prevalence measures of abstinence".	OR 1.42[m] OR 1.21[Ω] OR 1.68[Ω]	1.3 (1.3) 1.7 (1.7) 0.9 (0.9)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for null hypothesis
				S: Prolonged abstinence for 1 month	OR 1.17	0.11	0.130		OR 1.42[m] OR 1.21[Ω] OR 1.68[Ω]	1.5 (1.5) 2.0 (2.0) 1.1 (1.1)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: 7-day point prevalence abstinence	OR 1.11	0.10	0.307		OR 1.42[m] OR 1.21[Ω] OR 1.68[Ω]	0.8 (0.8) 1.1 (1.1) 0.5 (0.5)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for null hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for null hypothesis
				S: 24-hour point prevalence abstinence	OR 1.15	0.09	0.131		OR 1.42[m] OR 1.21[Ω] OR 1.68[Ω]	1.4 (1.4) 2.1 (2.1) 1.0 (1.0)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: Quit attempt	OR 1.11	0.06	0.074		OR 1.42[m] OR 1.21[Ω] OR 1.68[Ω]	1.4 (1.4) 2.3 (2.3) 1.0 (1.0)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
Alessi (26)	Compensation for video recording alcohol breath tests using a cell phone and contingency management with escalating vouchers for on-	Compensation for video recording alcohol breath tests using a cell phone	30 adults who drank frequently but were not physiologically dependent	P: Negative breath sample	MD 20.20	5.74	<0.001***	**"Cellphone technology may be useful for extending contingency management to treatment for alcohol problems"** No mention of results >0.05	MD 8.00[n] MD 5.00[α] MD 12.00[α]	69.8 (69.9) 21.7 (21.7) 134.1 (134.2)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect)	Very strong evidence for experimental hypothesis Strong evidence for experimental hypothesis Extreme evidence for experimental hypothesis
				S: Longest duration of negative samples	MD 10.90	3.52	<0.001***		MD 2.00[n] MD 1.00[α] MD 3.00[α]	5.3 (5.3) 2.2 (2.2) 11.2 (11.2)	Evidence for experimental hypothesis (i.e. an effect) Evidence is insensitive Evidence for experimental hypothesis (i.e. an effect)	Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis Strong evidence for experimental hypothesis
				S: Days of drinking	MD -11.00	3.48	<0.001***		MD 3.71[o] MD 1.00[α] MD 7.00[α]	19.5 (19.5) 2.3 (2.3) 49.4 (49.4)	Evidence for experimental hypothesis (i.e. an effect) Evidence is insensitive Evidence for experimental hypothesis (i.e. an effect)	Strong evidence for experimental hypothesis Moderate evidence for experimental hypothesis Very strong evidence for experimental hypothesis

	time alcohol-negative tests.			S: Drinks per drinking day	MD -0.80	0.83	0.350		MD 1.20[o] MD 0.5 [α] MD 1.90 [α]	1.2 (1.2) 1.3 (1.3) 1.0 (1.0)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis No evidence
				S: Addiction Severity Index	MD -0.09	0.03	0.010**		MD 0.10[n] MD 0.01 [α] MD 0.20 [α]	41.3 (41.3) 2.6 (2.6) 28.0 (28.0)	Evidence for experimental hypothesis (i.e. an effect) Evidence is insensitive Evidence for experimental hypothesis (i.e. an effect)	Very strong evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Very strong evidence for experimental hypothesis
				S: Drinker Inventory of Consequences	MD -0.80	0.23	<0.001***		MD 1.00[p] MD 0.2[Ω] MD 1.8 [Ω]	120.0 (120.0) 18.1 (18.1) 83.4 (83.4)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect)	Extreme evidence for experimental hypothesis Strong evidence for experimental hypothesis Very strong evidence for experimental hypothesis
Study	**Intervention**	**Control**	**Participants**	**Outcome**	**Sample mean**	**Sample standard error**	**Significance *p***	**Study conclusions** Results conclusions for non-significant findings	**Expected effect size**	**Bayes factor: Dienes (Christie) (8-10)**	**Interpretation of Bayes Factor using Dienes (8)**	**Interpretation of Bayes Factors using Jeffreys (2)**
Richmond (27)	Nortriptyline added to multi-component smoking cessation intervention (included nicotine replacement therapy and cognitive behavioural therapy)	Placebo added to multi-component smoking cessation intervention (included nicotine replacement therapy and cognitive behavioural therapy)	425 male prisoners	P: Continuous abstinence	OR 0.98	0.30	>0.05	**"Adding nortriptyline to a smoking cessation treatment package consisting of behavioural support and nicotine replacement therapy does not appear to improve long-term abstinence rates in male prisoners . . . In this study, we found no significant difference in an intention-to-treat analysis between the two study groups, suggesting that the additional use of NOR does not enhance quit rates for tobacco in the longer term"** "Based on an intention-to-treat analysis and cut-off point for CO of ≤ 10 p.p.m, continuous abstinence between the treatment and comparison groups were not statistically different at 3 months . . . point-prevalence abstinence, using the ≤ 5 p. p. m. cut-off between the treatment and control groups, was also not statistically significant different at three months".	OR 1.21[q] OR 1.01 [Ω] OR 1.55 [Ω]	0.9 (0.9) 1.0 (1.0) 0.6 (0.6)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Moderate evidence for null hypothesis No evidence Moderate evidence for null hypothesis
				P: Point prevalence abstinence	OR 0.81	0.29	>0.05		OR 1.21[q] OR 1.01 [Ω] OR 1.55 [Ω]	1.1 (1.1) 1.0 (1.0) 1.0 (1.0)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis No evidence No evidence
				S: Smoking reduction (>50% reduction in cigarette consumption)	OR 0.75	0.26	>0.05		OR 0.43[q] OR 0.12 [Ω] OR 0.99 [Ω]	0.9 (0.9) 0.4 (0.4) 1.0 (1.0)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Moderate evidence for null hypothesis Moderate evidence for null hypothesis No evidence
Levin (28)	Venlafaxine-extended release	Placebo	103 cannabis dependent adults	P: Two week abstinence	OR 0.23	0.52	<0.001***	**"For depressed, cannabis-dependent patients, venlafaxine-extended release does not appear to be effective at reducing depression and may lead to an increase in cannabis use"** "No significant effect of treatment and no significant effect of baseline HAMD on 50% reduction of HAMD".	OR 0.80[r] OR 0.70 [α] OR 0.90 [α]	2.9 (2.9) 5.5 (5.5) 1.6 (1.6)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence is insensitive	Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				P: 50% reduction in depressive symptoms (Hamilton Depression rating scale)	OR 0.75	0.42	0.510		OR 1.43[s] OR 1.20 [α] OR 1.60 [α]	1.1 (1.1) 1.1 (1.1) 1.1 (1.1)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: THC urine levels	MD 964	320.27	<0.001***		MD 137.3[t] MD 100 [α] MD 300 [α]	3.3 (3.3) 2.3 (2.3) 11.9 (11.9)	Evidence for experimental hypothesis (i.e. an effect) Evidence is insensitive Evidence for experimental hypothesis (i.e. an effect)	Moderate evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Strong evidence for experimental hypothesis
				S: Use in grams	MD 2.67	4.72	0.320		MD 0.45[u] MD 0.02 [α] MD 0.88 [α]	1.0 (1.0) 1.0 (1.0) 1.1 (1.1)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	No evidence No evidence Anecdotal evidence for experimental hypothesis
Okuyemi (29)	Motivational interviewing and nicotine patch	Nicotine patch and brief advice to quit	430 homeless smokers	P: 7-day point prevalence abstinence	OR 1.33	0.21	0.170	**"Adding motivation interviewing counselling for nicotine patch did not increase smoking rate significantly at 26-week follow-up for homeless smokers . . . MI did not improve adherence measures among participants who received MI."** "Motivation for adherence scores at week 6 were marginally higher for participants in the intervention group than those in the control group . . . There were no differences between study groups in the proportion of participants who had their nicotine patches on at various study visits".	OR 1.35[v] OR 1.02 [Ω] OR 1.78 [Ω]	1.8 (1.8) 1.1 (1.1) 1.4 (1.4)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis
				S: Motivation to adhere	MD 1.4	0.49	0.080		MD 4.97 [w] MD 1.19 [Ω] MD 8.75 [Ω]	11.2 (11.2) 25.0 (25.0) 6.6 (6.6)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect)	Strong evidence for experimental hypothesis Strong evidence for experimental hypothesis Moderate evidence for experimental hypothesis
				S:Self-efficacy to adhere	MD 2.5	3.12	0.220		MD 4.97 [w] MD 1.19 [Ω] MD 8.75 [Ω]	1.0 (1.0) 1.2 (1.2) 0.7 (0.7)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for experimental hypothesis Anecdotal evidence for null hypothesis
				S: Nicotine patch use	OR 1.0	0.20	0.970		OR 1.14[x] OR 1.02 [α] OR 1.28 [α]	0.8 (0.8) 1.0 (1.0) 0.6 (0.6)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Moderate evidence for null hypothesis No evidence Moderate evidence for null hypothesis
Gustafson (30)	Interest circle calls	No intervention	201 clinics	P: Waiting-time (mean days between first contact and first treatment)	MD -0.24	2.12	0.911	**"When trying to improve the effectiveness of addiction treatment services, clinic-level coaching appears to help improve waiting-time and number of new patients while other components of improvement collaboratives (interest circle calls and learning sessions) do not seem to add further value"** "Learning sessions had a modest waiting time reduction while interest circle calls had a slight increase, but these two groups' changes were not statistically significant . . . None of the groups showed significant improvement in retention for the 6-month intervention period (Table 3a), or the entire intervention and sustainability period (Table 3b), and there were no significant differences between groups"	MD 10.6[y] MD 15 [α] MD 5 [α]	0.2 (0.2) 0.2 (0.2) 0.4 (0.4)	Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect) Evidence is insensitive	Strong evidence for null hypothesis Strong evidence for null hypothesis Moderate evidence for null hypothesis
				P: Retention (percentage of patients retained from first to fourth treatment session)	MD -0.003	0.03	0..912		MD 7.5[y] MD 10 [α] MD 5 [α]	0.01 (0.01) 0.00 (0.00) 0.01 (0.01)	Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect)	Very strong evidence for null hypothesis Extreme evidence for null hypothesis Very strong evidence for null hypothesis
				P: Annual number of new patients	MD -0.04	0.04	0.369		MD 14.2[y] MD 20 [α] MD 10 [α]	0.01 (0.01) 0.00 (0.00) 0.01 (0.00)	Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect)	Very strong evidence for null hypothesis Extreme evidence for null hypothesis Very strong evidence for null hypothesis
	Coaching	No intervention		P: Waiting-time (mean days between first contact and first treatment)	MD 4.86	1.95	0.013*		MD 10.6[y] MD 15 [α] MD 5 [α]	7.2 (7.2) 5.4 (5.4) 10.7 (10.7)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect	Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis Strong evidence for experimental hypothesis
				P: Retention (percentage of patients retained from first to fourth treatment session)	MD 0.035	0.02	0.118		MD 7.5[y] MD 10 [α] MD 5 [α]	0.0 (0.0) 0.0 (0.0) 0.0 (0.0)	Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect)	Extreme evidence for null hypothesis Extreme evidence for null hypothesis Extreme evidence for null hypothesis
				P: Annual number of new patients	MD 0.20	0.09	0.028*		MD 0.14[y] MD 0.20 [α] MD 0.10 [α]	6.0 (6.0) 6.3 (6.3) 5.0 (5.0)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect	Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis
	Learning sessions	No intervention		P: Waiting-time (mean days between first contact and first treatment)	MD 3.14	1.93	0.103		MD 10.6[y] MD 15 [α] MD 5 [α]	1.2 (1.2) 0.9 (0.9) 2.1 (2.1)	Evidence is insensitive Evidence is insensitive Evidence is insensitive	Anecdotal evidence for experimental hypothesis Anecdotal evidence for null hypothesis Anecdotal evidence for experimental hypothesis

				P: Retention (percentage of patients retained from first to fourth treatment session)	MD -0.003	0.02	0.899		MD 7.5[y] MD 10[α] MD 5[α]	0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect)	Extreme evidence for null hypothesis Extreme evidence for null hypothesis Extreme evidence for null hypothesis
				P: Annual number of new patients	MD -0.001	0.07	0.982		MD 14.2[y] MD 20[α] MD 10[α]	0.00 (0.00) 0.00 (0.00) 0.01 (0.01)	Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect)	Extreme evidence for null hypothesis Extreme evidence for null hypothesis Very strong evidence for null hypothesis
	Combination	No intervention		P: Waiting-time (mean days between first contact and first treatment)	MD 6.16	1.97	0.002**		MD 10.6[y] MD 15[α] MD 5[α]	41.2 (41.2) 31.8 (31.8) 50.4 (50.4)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect	Very strong evidence for experimental hypothesis Very strong evidence for experimental hypothesis Very strong evidence for experimental hypothesis
				P: Retention (percentage of patients retained from first to fourth treatment session)	MD -0.003	0.02	0.891		MD 7.5[y] MD 10[α] MD 5[α]	0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect) Evidence for null hypothesis (i.e. no effect)	Extreme evidence for null hypothesis Extreme evidence for null hypothesis Extreme evidence for null hypothesis
				P: Annual number of new patients	MD 0.09	0.04	0.029*		MD 0.14[y] MD 0.20[α] MD 0.10[α]	5.6 (5.6) 4.4 (4.4) 6.5 (6.5)	Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect) Evidence for experimental hypothesis (i.e. an effect	Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis Moderate evidence for experimental hypothesis

Note: P=primary outcome; S=secondary outcome; * significant at $p<0.05$; ** significant at $p<0.01$; *** significant at $p<0.001$; RaR=Rate ratio; RR= Relative risk; OR=Odds ratio; MD=mean difference; [α]range of population SD reflects the CI of the expected effect size; [α] range of population SD based on opinion on a viable effect; a one directional relationship was assumed in all instances; Based on: [a] (31); [b] (32) [c] (33) ; [d] (34); [e] (35); [f] (36); [g] values specified in the sample size calculation; [h] (37); [i] (38); [j] (39); [k] (40); [l] (41); [m] (42); [n] (43); [o] (44); [p] (45); [q] (46); [r] (47); [s] (48); [t] (49); [u] (50); [v] (51); [w] (52); [x] (53); [y] values specified in the sample size calculation

References

1. West R. Using Bayesian analysis for hypothesis testing in addiction science. Addiction. 2016;111(1):3-4.
2. Jeffreys H. Theory ofprobability. Oxford: Clarendon Press; 1961.
3. Wagenmakers E-J. A practical solution to the pervasive problems ofp values. Psychonomic bulletin & review. 2007;14(5):779-804.
4. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic bulletin & review. 2009;16(2):225-37.
5. Gallistel C. The importance of proving the null. Psychological review. 2009;116(2):439.
6. Schervish MJ. P values: what they are and what they are not. The American Statistician. 1996;50(3):203-6.
7. Berger JO. Statistical decision theory and Bayesian analysis: Springer Science & Business Media; 2013.
8. Dienes Z. Using Bayes to get the most out of non-significant results. Frontiers in psychology. 2014;5.
9. Christie J. Bayes Factor Calculator. R code. 2011; Available from: http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/bayesFactorCalc2.R.
10. Baguley T, Kaye WS. Review of Understanding psychology as a science: An introduction to scientific and statistical inference. British Journal of Mathematical & Statistical Psychology. 2010;63:695-8.
11. Lee M, Wagenmakers E. Bayesian modeling for cognitive science: A practical course. Cambridge UP. 2013.
12. Rouder JN, Morey RD, Speckman PL, Province JM. Default Bayes factors for ANOVA designs. Journal of Mathematical Psychology. 2012;56(5):356-74.
13. Baguley T. Serious stats: A guide to advanced statistics for the behavioral sciences: Palgrave Macmillan; 2012.
14. Morey R, Rouder J, Jamil T. BayesFactor: Computation of Bayes factors for common designs. R package version 09. 2014;8.
15. Koch M, Riss P, Umek W, Hanzal E. The primary outcomes and power calculations in clinical RCTs in urogynecology - need for improvement? Trials. [journal article]. 2015;16(1):1-.
16. Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? BMJ : British Medical Journal. 2001 12/07/accepted;322(7292):989-91.
17. Sprenger J. The Objectivity of Subjective Bayesian Inference. 2015; Available from: http://philsci-archive.pitt.edu/11936/1/ObjectiveBayesianStatistics_v3.pdf.
18. Rouder JN. Optional stopping: No problem for Bayesians. Psychonomic bulletin & review. 2014;21(2):301-8.
19. Kypri K, McCambridge J, Vater T, Bowe SJ, Saunders JB, Cunningham JA, et al. Web-based alcohol intervention for Māori university students: double-blind, multi-site randomized controlled trial. Addiction. 2013;108(2):331-8.
20. Li L, Wu Z, Liang LJ, Lin C, Zhang L, Guo S, et al. An intervention targeting service providers and clients for methadone maintenance treatment in China: a cluster-randomized trial. Addiction. 2013;108(2):356-66.
21. Ward KD, Asfar T, Al Ali R, Rastam S, Weg MWV, Eissenberg T, et al. Randomized trial of the effectiveness of combined behavioral/pharmacological smoking cessation treatment in Syrian primary care clinics. Addiction. 2013;108(2):394-403.
22. Borland R, Balmford J, Benda P. Population-level effects of automated smoking cessation help programs: a randomized controlled trial. Addiction. 2013;108(3):618-28.
23. Rendall-Mkosi K, Morojele N, London L, Moodley S, Singh C, Girdler-Brown B. A randomized controlled trial of motivational interviewing to prevent risk for an alcohol-exposed pregnancy in the Western Cape, South Africa. Addiction. 2013;108(4):725-32.
24. Coffin PO, Santos GM, Das M, Santos DM, Huffaker S, Matheson T, et al. Aripiprazole for the treatment of methamphetamine dependence: a randomized, double-blind, placebo-controlled trial. Addiction. 2013;108(4):751-61.

25. Gilbert HM, Leurent B, Sutton S, Alexis-Garsee C, Morris RW, Nazareth I. ESCAPE: a randomised controlled trial of computer-tailored smoking cessation advice in primary care. Addiction. 2013;108(4):811-9.
26. Alessi SM, Petry NM. A randomized study of cellphone technology to reinforce alcohol abstinence in the natural environment. Addiction. 2013;108(5):900-9.
27. Richmond R, Indig D, Butler T, Wilhelm K, Archer V, Wodak A. A randomized controlled trial of a smoking cessation intervention conducted among prisoners. Addiction. 2013;108(5):966-74.
28. Levin FR, Mariani J, Brooks DJ, Pavlicova M, Nunes EV, Agosti V, et al. A randomized double-blind, placebo-controlled trial of venlafaxine-extended release for co-occurring cannabis dependence and depressive disorders. Addiction. 2013;108(6):1084-94.
29. Okuyemi KS, Goldade K, Whembolua GL, Thomas JL, Eischen S, Sewali B, et al. Motivational interviewing to enhance nicotine patch treatment for smoking cessation among homeless smokers: a randomized controlled trial. Addiction. 2013;108(6):1136-44.
30. Gustafson DH, Quanbeck AR, Robinson JM, Ford JH, Pulvermacher A, French MT, et al. Which elements of improvement collaboratives are most effective? A cluster-randomized trial. Addiction. 2013;108(6):1145-57.
31. Kypri K, Hallett J, Howat P, McManus A, Maycock B, Bowe S, et al. Randomized controlled trial of proactive web-based alcohol screening and brief intervention for university students. Archives of Internal Medicine. 2009;169(16):1508-14.
32. Andrews S, Sorensen JL, Guydish J, Delucchi K, Greenberg B. Knowledge and Attitudes About Methadone Maintenance Among Staff Working in a Therapeutic Community. Journal of maintenance in the addictions. 2005 10/24;3(1):47-59.
33. Livingston JD, Milne T, Fang ML, Amari E. The effectiveness of interventions for reducing stigma related to substance use disorders: a systematic review. Addiction. 2012;107(1):39-50.
34. Hser Y-I. Predicting long-term stable recovery from heroin addiction: Findings from a 33-year follow-up study. Journal of addictive diseases. 2007;26(1):51-60.
35. Mattick RP, Breen C, Kimber J, Davoli M. Methadone maintenance therapy versus no opioid replacement therapy for opioid dependence. Cochrane Database Syst Rev. 2009;3.
36. Stead LF, Perera R, Bullen C, Mant D, Lancaster T. Nicotine replacement therapy for smoking cessation. Cochrane Database Syst Rev. 2008;1(1).
37. Floyd RL, Sobell M, Velasquez MM, Ingersoll K, Nettleman M, Sobell L, et al. Preventing alcohol-exposed pregnancies: a randomized controlled trial. American journal of preventive medicine. 2007;32(1):1-10.
38. Ingersoll KS, Ceperich SD, Nettleman MD, Karanda K, Brocksen S, Johnson BA. Reducing alcohol-exposed pregnancy risk in college women: Initial outcomes of a clinical trial of a motivational intervention. Journal of substance abuse treatment. 2005;29(3):173-80.
39. Tiihonen J, Kuoppasalmi K, Föhr J, Tuomola P, Kuikanmäki O, Vorma H, et al. A comparison of aripiprazole, methylphenidate, and placebo for amphetamine dependence. American Journal of Psychiatry. 2007;164(1):160-2.
40. Colfax GN, Santos G-M, Das M, Santos DM, Matheson T, Gasper J, et al. Mirtazapine to reduce methamphetamine use: a randomized controlled trial. Archives of general psychiatry. 2011;68(11):1168-75.
41. Meini M, Moncini M, Cecconi D, Cellesi V, Biasci L, Simoni G, et al. Aripiprazole and ropinirole treatment for cocaine dependence: evidence from a pilot study. Current pharmaceutical design. 2011;17(14):1376-83.
42. Lancaster T, Stead LF. Self-help interventions for smoking cessation. Cochrane Database Syst Rev. 2005;3(3).
43. Petry NM, Martin B, Cooney JL, Kranzler HR. Give them prizes and they will come: Contingency management for treatment of alcohol dependence. Journal of consulting and clinical psychology. 2000;68(2):250.
44. Barnett NP, Tidey J, Murphy JG, Swift R, Colby SM. Contingency management for alcohol use reduction: A pilot study using a transdermal alcohol sensor. Drug and alcohol dependence. 2011;118(2):391-9.
45. Litt MD, Kadden RM, Kabela-Cormier E, Petry NM. Changing network support for drinking: network support project 2-year follow-up. Journal of consulting and clinical psychology. 2009;77(2):229.
46. Hughes JR, Stead LF, Hartmann-Boyce J, Cahill K, Lancaster T. Antidepressants for smoking cessation. Cochrane Database Syst Rev. 2014;1:CD000031.

47. Findling RL, Pagano ME, McNamara NK, Stansbrey RJ, Faber JE, Lingler J, et al. The short-term safety and efficacy of fluoxetine in depressed adolescents with alcohol and cannabis use disorders: a pilot randomized placebo-controlled trial. Child and Adolescent Psychiatry and Mental Health. 2009 03/19

11/04/received

03/19/accepted;3:11-.
48. Keller MB, Trivedi MH, Thase ME, Shelton RC, Kornstein SG, Nemeroff CB, et al. The Prevention of Recurrent Episodes of Depression with Venlafaxine for Two Years (PREVENT) Study: Outcomes from the 2-year and combined maintenance phases. The Journal of clinical psychiatry. 2007;68(8):1246-56.
49. Bonnet U, Specka M, Stratmann U, Ochwadt R, Scherbaum N. Abstinence phenomena of chronic cannabis-addicts prospectively monitored during controlled inpatient detoxification: Cannabis withdrawal syndrome and its correlation with delta-9-tetrahydrocannabinol and-metabolites in serum. Drug and alcohol dependence. 2014;143:189-97.
50. Smeerdijk M, Keet R, Dekker N, van Raaij B, Krikke M, Koeter M, et al. Motivational interviewing and interaction skills training for parents to change cannabis use in young adults with recent-onset schizophrenia: a randomized controlled trial.
51. Hettema JE, Hendricks PS. Motivational interviewing for smoking cessation: a meta-analytic review. Journal of consulting and clinical psychology. 2010;78(6):868.
52. Alterman AI, Gariti P, Cook TG, Cnaan A. Nicodermal patch adherence and its correlates. Drug and alcohol dependence. 1999;53(2):159-65.
53. Hollands GJ, McDermott MS, Lindson-Hawley N, Vogt F, Farley A, Aveyard P. Interventions to increase adherence to medications for tobacco dependence. Cochrane Database Syst Rev. 2015;2:CD009164.

Supplementary Appendix 1

We present here a worked example to show how to use Dienes' online calculator to obtain Bayes Factors.

http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/bayes_factor.swf

Okuyemi et al (29) were interested in the effects of adding motivation interviewing (MI) counselling to the nicotine patch for smoking cessation among homeless smokers. They conducted a randomised controlled trial, whereby 430 participants were randomised to the intervention group or a control group. Verified seven-day abstinence rates at week 26 for the intervention group was non-significantly higher than for the control group (OR 1.33; 95% CI=0.88, 2.02; p= 0.17). They concluded that "Adding motivational interviewing counselling to nicotine patch did not significantly increase smoking rate at 26-week follow-up for homeless smokers".

To calculate the Bayes Factor the odds ratio first needs to be transformed using a natural logarithmic transformation:

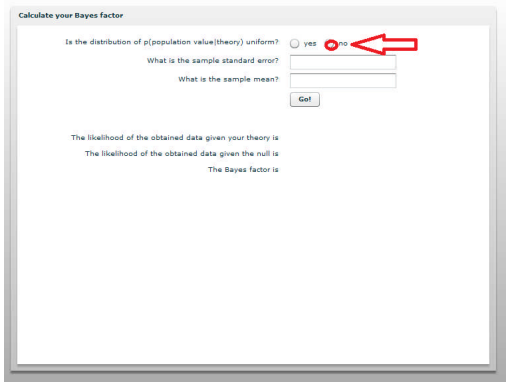
$$LN(1.33)= 0.29 \text{ (2 dp)}$$

and the standard error calculated as:

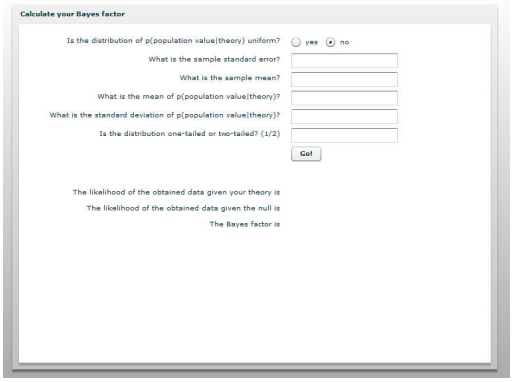
$$[LN(2.02)-LN(0.88)]/3.92= 0.21 \text{ (2 dp)}$$

We must then decide between three possible distributions to represent the predictions of the theory: uniform, normal or half-normal. If we can only specify a plausible maximum effect we should use the uniform distribution. In contrast, if a plausible predicted effect size can be specified we should opt for a normal or half-normal distribution. The choice between these depends on whether a directional hypothesis can be made, with the latter assuming a one-tailed test. In our example, a half-normal distribution is used as we hypothesize a positive impact of the intervention and can easily derive a predicted value from (51), which was a comprehensive meta-analysis of the use of MI for smoking cessation. This identified an OR for long-term follow-up of 1.35 (95%CI 1.02 to 1.78), which translates to a log odds ratio of 0.30.

We can now calculate our Bayes Factor. First mark the box ‘no’ next to ‘Is the distribution of p(population value|theory) uniform?’



You will then see the following screen with three new boxes. These ask for the mean, standard deviation and number of tails (of a normal). As we are using a half-normal we set mean to 0 (Note: half-normal distribution has a mode of 0), SD to our plausible expected value (Note: this scales the half-normal distribution’s rate of drop) and tails to 1. We must also enter the standard error and mean of our sample. Then click "Go".



This gives us a Bayes Factor of 1.82, indicating that the data favour the experimental hypothesis but not to a sufficient degree and are thus ‘insensitive’.