Journal of Medical Statistics and Informatics
ISSN 2053-7662

HOAJ
Herbert Open Access Journals

Original Open Access

False negatives in evidence based medicine

David Kault

Correspondence: David.Kault@jcu.edu.au

Discipline of Mathematics, School of Engineering and Physical Sciences, James Cook University, Townsville, Australia.

Abstract

Evidence Based Medicine (EBM) is a term used for the current dominant methodology for deciding what medical treatments should be accepted as valid. It places great emphasis on Randomised Clinical Trials (RCTs) which are analysed according to a strict frequentist paradigm, with a rigid p-value ≤0.05 criterion but with little consideration of prior probabilities or the cost of errors. Accordingly, low cost, safe treatments where there is prior knowledge of at least slight effectiveness, may often be inappropriately discarded by EBM. The Cochrane Collaboration is an online central repository of RCTs and meta-analyses of RCTs. This paper uses statistical methods applied to a random sample of outcomes listed in the Cochrane Collaboration, to estimate the negative predictive value when treatments are declared ineffective as a result of positive outcomes which do not achieve the $p \leq 0.05$ criterion. The data were analysed using six different models in order to determine the proportion of genuinely ineffective treatments in the set of all positive outcomes where $p>0.05$. All six methods give point estimates substantially less than half for the negative predictive value when the decision rule is to declare treatments to be ineffective when their outcome is positive but $p>0.05$. Although confidence interval estimation indicates considerable uncertainty in these estimates, it seems reasonable to conclude that when a RCT gives a positive outcome but $p>0.05$, the conventional EBM decision to declare the treatment to be ineffective, is likely to be wrong more often than not.

Keywords: Evidence based medicine, false negatives, false non-discovery rate, low cost treatments, negative predictive value, statistical models, type II error

Introduction

Whilst statistics and its application to medicine is a relatively modern discipline, its use as the fundamental arbiter of truth in medicine is much more recent still. The term "Evidence Based Medicine" (EBM) was coined in the early 1990s and with it came a paradigm shift in which evidence from human experiments - randomised controlled clinical trials (RCTs) - came to be regarded as the final arbiter of effective medical treatments. The birth of EBM was perhaps stimulated by the discovery of some important counterintuitive results. In particular, a statistical analysis was published of the effects of anti-arrhythmic drugs which had been advocated during heart attack because it was known that they suppressed minor heart irregularities. It was found that these drugs actually increased the risk of fatal disturbances of heart rhythm [1]. The interest in EBM with its increased emphasis on RCTs, may have also been related to the increase in the output of new pharmaceuticals and the corresponding increase in the need for objective assessments of drugs.

Prior to the early 1990s, decisions about which treatments were effective were made as a result of a combination of an understanding of mechanism where possible, expert opinion and experience, and by occasional RCTs. Experiences such as those with anti-arrhythmic drugs, prompted a pendulum swing towards greater reliance on RCTs, under the banner of EBM. EBM is now the current medical orthodoxy and indeed, challenging it can be seen as almost like challenging motherhood.

However, the EBM pendulum may have swung too far in overrating statistical evidence at the expense of all other kinds of evidence. In particular, EBM labels many treatments as having "no effect" when this conclusion is simply not plausible. Such a label may be applied to trials of additional checks and precautions which must logically be at least of some benefit in a small proportion of cases and may involve negligible costs. There are also examples where a low cost drug is known to alter physiology in a way which is almost certainly of at least slight benefit in a disease, but EBM will label the drug as having "no effect" for that disease. The "no effect" label is also used in implausible statements of the sort claiming a benefit for men but "no effect" for women, or benefit for those under 50 years of age but "no effect" on those over 50, or a benefit 3 months after treatment but "no effect" 6 months after the treatment. The term "no effect" is not just being used as an unfortunate abbreviation for "no statistically significant effect", or, "no effect that could not be quite easily be explained by chance variation". In this author's reading of the medical literature, the term "no effect" is treated as a statement that is to be regarded as literally true, at least provisionally, pending any reassessment after some future RCT. A recent opinion piece in the Journal of the American Medical Association expressed similar concerns about the more nuanced expression "there is no evidence to suggest " [2]. It is this author's experience that most of the time in such situations where "no effect" is declared, reference to the original data will

doi: 10.7243/2053-7662-2-5

reveal a small effect in the direction that would be expected.

A good example of an ultra low cost treatment which almost certainly has a worthwhile benefit overall in further reducing a small risk, is provided by skin cleaning. EBM proponents have argued that there is no need to clean the skin prior to minor procedures such as injections [3], despite 150 years of scientific knowledge regarding hygiene. The EBM enthusiasts make this argument simply because in the case of injections there are no RCTs which have resulted in a p-value ≤ 0.05 for skin cleaning and some relatively small uncontrolled trials with zero incidence of infection, so indicating "no evidence" for any benefit of hygiene in this case. In response to many such EBM pronouncements, the suggestion has been made that EBM proponents need to take part in a RCT of the effectiveness of parachutes in preventing trauma on descent from aircraft - as they deserved "to come down to earth with a bump" [4].

The issues here stem from a rigid approach by EBM in applying frequentist statistics, so that the size of Type I and II errors are arbitrarily fixed with little account being taken of prior knowledge or the cost of errors [5]. This is a particular concern when EBM is used to assess low cost treatments thought to avert low risk, high cost events as in skin cleaning preventing infections after injections. The need to assess low cost precautions against very low risk, very high cost situations, arises commonly in anaesthesia and to avoid the limitations of EBM in such situations, an alternative decision making paradigm is emerging [6].

Although there is no general agreement on exactly what constitutes EBM, there is now a formal evidence rating system [7], with the strongest rated evidence (Level 1 evidence) applied to treatments that show "significant" results in multiple RCTs. "Significant" is defined in terms of obtaining a p-value ≤ 0.05. Prior knowledge such as mechanism-based reasoning does not rate a mention as a form of evidence about a treatment, except when there is no statistical evidence [7]. When physiological or mechanism-based reasoning or commonsense and expert opinion are the sole basis for a positive assessment of a treatment, the evidence is defined to be at the lowest level (Level 5 evidence), so the parachute parody [4] does not appear to have had much impact on EBM. Bayesian methods to allow for prior knowledge and decision theory to take account of the cost of errors, are in practice little used in the assessment of most treatments even though Bayesian methods in medicine is a field of active research. In EBM the term "significant" is virtually synonymous with obtaining a p-value ≤0.05. Obtaining $p>0.05$ for a treatment in an RCT is taken to mean that, at least for the time being, the treatment is to be regarded as ineffective.

There has been some move to go beyond simple inspection of p-values to make a decision about a treatment. In particular, there is encouragement to define a clinically significant benefit and fix a reasonable level of Type II error. Additionally, the CONSORT statement [8] encourages the use of 95% confidence intervals as well as p-values in summarising the result of a statistical analysis so that studies with narrower confidence intervals can be favoured in terms of level of evidence. The CONSORT statement also suggests that it may be appropriate to consider the confidence interval as a range of values compatible with the size of the treatment effect, even when the confidence interval includes the zero effect value. This author would have preferred to see CONSORT replace the word "compatible" in the preceding sentence by a term such as "easily compatible", but nevertheless, it is positive to see acknowledgement within EBM that a confidence interval which includes zero, does not necessarily equate with "no effect". However, in most medical literature, if the 95% confidence interval includes zero treatment effect or equivalently the p-value is >0.05 the treatment is assessed, at least provisionally, as worthless. Writing ten years ago in the first issue of the general statistical interest magazine "Significance", a consultant medical statistician commented that medical researchers were only interested in "p-values ... less than 0.05" [9]. Unfortunately it is this author's perception working both as a part time general practitioner and part time statistical consultant, that the enthusiasm for p-values as the sole criterion in medical decision making has not changed. There is a myriad of disease management summaries that this author, working as a general practitioner, receives to assist in continuing professional development, where almost all advice as to which treatments should be assumed to work seems based solely on p-values.

The cost of errors are not explicitly taken into account by EBM. However, the cost of Type II errors is to some degree, implicitly considered when the size of this error is set. Whilst the setting of the Type II error rate is less ossified than is the case with the $p\leq0.05$ criterion, a common requirement is that the Type II error rate be=0.2. However, specification of the Type II error rate also implies a requirement to specify a minimum size of effect that is clinically significant [8] though this is not the only approach to balancing errors and effect size [10,11]. Unfortunately, much of the time it is not possible to objectively specify a size of effect that is just clinically significant. Often the size of the effect that is clinically significant in a real sense, the minimum clinically important difference, will imply a requirement for trials that are too large to be feasible. For example, most cancer patients given the choice of two almost equal cost chemotherapy regimes, one of which has a 50% chance of saving their life and the other which has a 51% chance, will have a definite preference for the latter. For them, a 1% extra chance of life is "clinically significant". A trial to detect this 1% difference for a particular cancer at a particular stage, with the Type I error rate set at 0.05 and the Type II error rate set at 0.20, would require 78,000 participants and certainly not be feasible. Likewise, whilst patients may accept their doctor's bland reassurance based on EBM, that cleaning the skin prior to an injection has "no effect" on the chance of a serious infection, they might baulk at the idea of a doctor declining to use an alcohol wipe, if the risk that could

doi: 10.7243/2053-7662-2-5

be averted by use of an alcohol wipe was explained as the risk of a month or so commuting on the roads. (A risk of the order of 1 in 100,000 was suggested in the discussion following the paper on EBM and the issue of skin cleaning for injections [3]). A patient might then ask "How many alcohol wipes are worth my life?". However, a RCT to prove the risk of not using an alcohol wipe was appreciably smaller than the risk of a month's commuting, would require even greater numbers of participants than in the previous cancer trial scenario. The EBM solution to this difficulty amounts to redefining the term "clinically significant". In a paper giving instructions on the statistics of so called non-inferiority trials, it is clearly indicated that in the cancer survival scenario above, a 10% decline in survival or even a halving of the cure rate could be regarded as "non-inferior", simply by reinterpreting the term "clinically significant" to mean little more than "detectable by a trial of feasible size" [12].

The discussion above suggests that EBM is too ready to dismiss low cost treatments that are likely to have a worthwhile, albeit less than dramatic, effect. RCTs of many such treatments may be declared to be negative, when a small positive effect may be apparent in the outcomes of such RCTs. This may occur too often for one to reasonably believe the excess of trials with weakly positive outcomes are due to chance. This paper assesses the false non-discovery rate of EBM using the methods of EBM, frequentist statistics, together with six different models or sets of assumptions, about the distribution of the effectiveness of interventions. In particular, for each model, estimates are obtained of the negative predictive value of declaring a treatment ineffective conditional on a weakly positive outcome. More precisely, this paper is concerned with:-

P(treatment ineffective | outcome of trial positive but $p>0.05$)

In the remainder of this paper the term "(conditional) negative predictive value" may be used for brevity, without specifying the conditionality.

The selection of a random list of RCTs is facilitated by the existence of an online encyclopaedic compilation of EBM analyses known as the Cochrane Collaboration. Considerable care is taken in the Cochrane database to include all relevant data and so minimize bias towards trials which give particular outcomes, such as positive outcomes. A random selection of point estimates with confidence intervals was chosen from topics across the Cochrane Collaboration. These estimates, regarded as data points, are then analysed by six different models.

Methods

Randomisation

Data was obtained over the internet from the Cochrane Collaboration [13] during the first quarter of 2013. The Cochrane Collaboration is organised in a tree like structure with headings or main branches, subheadings or secondary branches and sub-subheadings etc, leading to links to the abstracts that may contain summary statistics either from meta-analyses or analyses of individual RCTs. These abstracts constitute a major reference database for EBM. The Cochrane Collaboration's decision about whether a particular treatment is effective is based largely on the statistical point estimates and confidence intervals in these abstracts. To ensure an unbiased sample of the statistical results in the Cochrane Collaboration, a portable random number generator [14] was used to steer a path through to an individual quantitative result in an abstract. The random number generator is seeded by an initial integer which then determines a sequence of pseudo-random numbers between 0 and 1. In each search for an individual quantitative result, consecutive numbers from this sequence were used at each level of the Cochrane Collaboration branching tree to determine which of the branches to follow. For example, if at the k^{th} level there were 10 branches, the branch chosen would be the $n^{th}+1$ of the branches listed at that level where n was the first decimal of the k^{th} random number in the sequence. If on arrival at an abstract there was more than one point estimate with a confidence interval, the next number in the sequence was used to choose one of the estimates. A particular random path to a single quantitative result is then determined by providing the random number generator with a particular seed. Some pathways led to abstracts which did not contain suitable quantitative results. In those cases the random number calling program was restarted with a new seed. Ultimately seeding integers 1 to 248 were used to obtain 100 point estimates together with their confidence intervals.

Exclusions

There were 148 abstracts that were not used. These abstracts were excluded for the following reasons:

1. Most of the excluded abstracts simply did not contain any quantitative data. Indeed, many of these abstracts did no more than state that a meta-analysis had been attempted, but the attempt failed because there were no suitable RCTs on the topic to be reviewed.
2. Abstracts that gave point estimates but not confidence intervals were also excluded.
3. Abstracts that reviewed new drugs under patent were excluded, partly because of concern that commercial pressures through publication bias or other means might lead to distortions that would differ from those where commercial pressures were not involved [15]. Excluding costly new drugs, also allows this paper to give more weight to the question of how often EBM erroneously rejects low cost treatments. However, abstracts evaluating new drugs make up only a small proportion of the Cochrane database.
4. Generally an RCT will compare the effects of a standard treatment against the effects of a standard treatment plus an extra treatment which it is hoped may be beneficial. Whilst there is an implied suggestion that the extra

doi: 10.7243/2053-7662-2-5

treatment may be beneficial, it may turn out to be useless or (hopefully rarely) counterproductive. However, some RCTs are of a different form and compare mutually exclusive treatments where there is no suggestion of superiority of one or the other. Data were excluded if there was no implied suggestion that a treatment was being assessed that may be superior to an existing standard treatment. Whether a suggestion was implied about the possible superiority of a treatment, was determined by the author's commonsense informed by the background material in the abstract and by the author's general knowledge as a medical doctor. However, it is likely that commonsense alone, without specialised medical knowledge, would nearly always have led to the same conclusion. In a few cases the implied suggestion was that the new treatment might be inferior. For example, there were trials to see if non specialist surgeons could be as successful in some areas as specialist surgeons, where the existing standard of treatment was treatment by a specialist surgeon. In such cases the outcome was treated as though there was an implied suggestion of possible superiority for the existing standard.

5. Abstracts were excluded if there were mistakes or "typos" which made the quantitative information indecipherable. For example, there was an instance where the quoted confidence interval did not contain the point estimate and the correct values remained unresolved after examination of the full Cochrane review article [**16**].
6. It was assumed that in the case of differences in continuous quantities, the confidence interval would be symmetrically distributed about the point estimate. In the case of odds ratios, relative risks and hazard rates it was assumed that the log of the confidence interval end points would be symmetrically distributed about the log of the point estimate. This was checked and the data was not included if, to within rounding error, the (log) point estimate was not in the middle of the (log) confidence interval. Only one data point, involving a Peto odds ratio, was rejected on this basis. However, for the purposes of the first, second, third and fifth models, data was included in two cases where an odds ratio or the end point of a confidence interval of an odds ratio, included the value zero. In these two cases, information was obtained from the body of the Cochrane review to enable the point estimate to be matched to a value from a standard normal distribution in terms of equal p-value.
7. Where abstracts contained more than one point estimate with confidence intervals, the data was excluded from randomisation if it referred to issues such as side effects of a treatment rather than to the main purpose of a treatment.
8. Finally abstracts were rejected from further consideration if a previous random path had led to the same abstract. In other words, the random sampling of the Cochrane Collaboration was performed in a non-replacement mode.

It may be noted that although the process used to select a statistic is random with virtually no possibility of subjective judgement being allowed to distort the selection process, the statistics selected, whilst unbiased by this author's personal preference, may not be fully representative of the statistics recorded in the Cochrane Collaboration. In particular, if we regard the individual statistics as leaves on a tree, those leaves whose connection to the trunk bear fewer leaves, have more chance of being selected.

Data preprocessing

In the case of continuous data, the amount of effect was rearranged if necessary so that the implied a priori suggestion about the effect of the treatment was assigned a positive number if a desirable effect was found and vice versa. For example, if a weight loss treatment resulted in an average change in weight of -2kg for the treatment group in comparison to the controls, this value would be recorded as a "+2".

In the case of odds ratios and relative risks and rates, data was rearranged if necessary, so that the implied a priori suggestion about the effect of the treatment was assigned a value greater than 1 if the effect was desirable. For example, if a treatment for a relapsing disease reduced the rate of relapse by 50%, the relative rate change would be recorded as 2.00 with of course the corresponding treatment for the limits of the confidence intervals. Log transforms were taken of all odds ratios, relative risks and rates prior to analysis.

The 100 values obtained along with their confidence intervals, give information about the size of an effect and the accuracy with which it has been assessed. However, across these 100 values it is not possible to compare size of effects because continuous data is dependent on the unit of measurement. On the other hand, odds ratios, relative risks and rates can be compared on an absolute scale, so for example a treatment which halves the mortality of newborns, can be regarded as equally efficacious in quantitative terms as a treatment which doubles the odds of someone successfully giving up smoking.

To enable an analysis which could include together both continuous data and data on relative rates and ratios in a way that is not dependent on unit of measurement, it is necessary to sacrifice the information on accuracy and retain only information on size of effect, by expressing all data in terms of standard deviations from zero effect. This is done by dividing the size of the effect by the standard error as determined by the width of the confidence interval to give a t-statistic. For simplicity, throughout this paper, it is assumed that the statistics that are being sampled, have emerged from RCTs or meta-analyses where the numbers involved are sufficient for an approximation by the normal distribution. The data used in models 1, 2, 3 and 5 are the 100 sizes of effect as measured in standard deviations. Models 1 and 2 use all data points, but models 3 and 5 which assume an underlying normal distribution, neglect a single outlier.

In these models, information on precision is combined with

doi: 10.7243/2053-7662-2-5

amount of effect, into a single measure of "size" as measured in standard deviations. In a sense, this amounts to throwing away some of the information collected. In addition, the "size" that is used, is not the size of an effect, but the size of how convincingly an effect has been assessed. There may be disparate motives in such an assessment. In particular, those researching treatments in orthodox medicine may prefer to economise in terms of the number of subjects employed and so not demonstrate the size of the effect so convincingly. On the other hand researchers studying unlikely treatments that happen to be effective, may want to demonstrate their effectiveness to a much higher degree of certainty.

The issues here can be avoided if attention is focussed on relative risks and odds and hazard ratios where there is an absolute scale. In these situations, information about how big the effect is can be used separately from the measure of how accurately it has been measured. Models 4 and 6 are then confined to these types of measures omitting 28 cases involving continuous data. Two data points concerning odds ratios that involved the value 0 are also omitted, as such values were incompatible with normality assumptions required of the log of the data values. In effect, two outliers are being discarded. There were then 70 data points with accompanying confidence intervals for models 4 and 6.

Calculations

All computer calculations were performed using Fortran with the assistance of standard algorithms [**14**]. More complex calculations were checked by alternative methods where practical. In particular, some calculations using the EM algorithm [**17**] were checked using Nelder and Mead's simplex method to find maximum likelihood estimates. Some results were also checked by simulation methods.

First model

Model description

The first model assumes that treatments are of two sorts - effective treatments and treatments that are genuinely useless and whose effect sizes will therefore be scattered around zero. The number of genuinely useless treatments giving weakly negative results will tend to approximate the number of genuinely useless treatments giving weakly positive results, so if all effective treatments gave strongly positive results, the ratio of the weakly negative to weakly positive results will be around 1. Here it is linguistically convenient to use the term "weakly positive" to refer to a positive result with a p-value > 0.05 with "weakly negative" defined likewise. However, in practice the weakly positive results will also include some of the outcomes of genuinely effective treatments. The number of weakly negative outcomes can then be used as an estimate of the number of genuinely useless treatments in the set of all weakly positive outcomes. The ratio of the weakly negative outcomes to the weakly positive outcomes is then an estimate of the genuinely useless treatments as a proportion of all the weakly positive treatments. In other words this ratio gives an estimate of the (conditional) negative predictive value. This first model has an advantage in terms of simplicity. It has disadvantages in that it ignores the possibility of perverse chance resulting in a few of the genuinely effective treatments actually giving weakly negative outcomes. It also ignores the possibility of genuinely counterproductive treatments. It is also noted that the estimate here of the negative predictive value involves estimating a ratio of random quantities, and estimation methods can do better than taking the crude ratio of the point estimates [**18**]. However, preliminary calculations show that such corrections here are orders of magnitude smaller than other sources of error, so such refinements are not pursued.

Model 1 results

Of the 100 outcomes there were 2 outcomes that seemed convincingly counterproductive (outcomes < -1.96 standard deviations), there were 8 values in the range (-1.96, 0), there was 1 value of exactly 0 and 23 values in the range (0, +1.96). The overall pattern is illustrated in **Figure 1**. In forming the ratio of weakly negatives to weakly positives, the single value 0 was counted as an additional 0.5 of a value in both numerator and denominator. The ratio of weakly negative to weakly positive treatments then gives the negative predictive value to be $\frac{8.5}{23.5} \approx 36.2\%$. Confidence intervals for the (conditional) negative predictive value of declaring treatments that yield outcomes in the (0, +1.96) standard deviation range to be ineffective, are obtained by bootstrapping. All confidence intervals quoted in this paper are bootstrap confidence intervals based on resampling from the data to produce 10,000 sets of the relevant data points. The result is a 95% confidence interval of (13.7% , 71.4%).

Second Model

It is common to require clinical trials to be designed to have a Type II error rate of 20%. This specification requires some preliminary estimate of the variability in the quantity to be measured, some decision about the size of the effect that is to be detected and then adjustment of the numbers in the trial accordingly. In practice, the actual Type II error rate will differ from this specification as the variability in the system and the actual effect size will likely differ from the values used in planning. Furthermore, not all clinical trials will specify a Type II error rate of 20%. However, for the purposes of this second model, it is assumed that all outcomes of effective treatments are normally distributed and subject to a Type II error rate of 20%, so effective treatments, measured in standard deviations, are distributed as $N(1.96+0.842, 1^2)$. It is further assumed that all other treatments are entirely ineffective and their outcomes are distributed as a standard normal random variable. If q is the proportion of ineffective treatments then the proportion of treatments above 1.96 standard deviations

doi: 10.7243/2053-7662-2-5

=0.8(1-*q*)+0.025*q*. Between 0 and 1.96 standard deviations the proportion will be 0.1974(1-*q*)+0.475*q* and 0.0026(1-*q*)+0.5*q* will be less than zero. *q* is then chosen to minimize the chi square sum difference between the expected and observed proportions. The (conditional) negative predictive value is then

$$\frac{0.475q}{0.475q + 0.1974(1-q)}$$

The minimum chi-squared estimate of *q* is 19.4%. With this estimate, the model gives a good fit to the data with a χ_1^2 p-value of 0.71:

	<0	0 to 1.96	>1.96
Observed	10.5	23.5	66
Expected	9.92	25.13	64.95

The negative predictive value point estimate is 36.7% with 95% bootstrap confidence interval of (21.5%, 50.4%).

Third Model

The third model assumes that each treatment has an intrinsic effectiveness that is chosen from some fixed unknown distribution and that chance factors from the clinical trial are added to this intrinsic effectiveness to result in the actual outcome. A model of the distribution of outcomes is required that is a reasonable match to the data depicted in **Figure 1**. One value was clearly an outlier (This value related to the effectiveness of a Chinese herb on viral myocarditis and the effect of this herb had been demonstrated, far, far more convincingly than any other of the 99 effects examined). This outlier was dropped for the purposes of the curve fitting. The remaining 99 values appear in **Figure 1** to be compatible with the assumption of a normal distribution. We then assume that each treatment has an underlying effectiveness chosen from the distribution $X \sim N(\mu, \hat{\sigma}^2)$ and that to this intrinsic effectiveness, normally distributed random effects are added. The result is then normally distributed. Since the sizes of the effects are measured in standard deviations, the uncertainties added by the clinical trials are normally distributed with a standard deviation of 1 and are denoted by $Y \sim N(0,1^2)$ The outcomes of trials of disparate treatments is then distributed as $X + Y \sim N(\mu, \hat{\sigma}^2 + 1^2)$. The parameters μ and $\sigma = \sqrt{\hat{\sigma}^2 + 1}$ are then estimated by curve fitting.

Of most interest is the negative predictive value given that the outcome is in the range (0, 1.96) standard deviations. Since in this model we make the (unrealistic) assumption that treatments with precisely no effectiveness have zero probability, then the only treatments which would be correctly rejected are those having negative intrinsic effectiveness. Since our negative predictive value calculation is conditional on weakly positive outcomes, what we are interested in is:-

$$P(\text{intrinsic effectiveness} \leq 0 \mid \text{outcome} \in (0,1.96))$$

To evaluate this we require the numerator

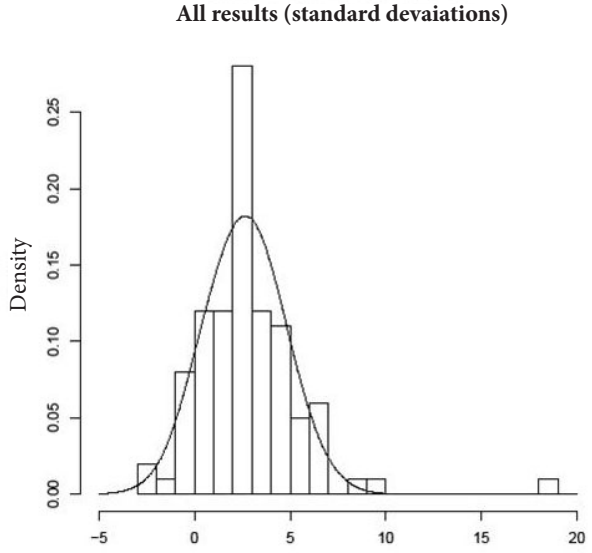

Figure 1. Histogram of all results as measured in standard deviations, together with a curve giving the density of the mixture distribution $y=0.02\ \phi(x,0,1^2) + 0.98\ \phi(x,2.66,\ 2.15^2)$. The numerical values used here are described in the results of the fifth model.

$$P(\text{intrinsic effectiveness} \leq 0 \cap \text{outcome} \in (0,1.96))$$

and the denominator, $P(\text{outcome} \in (0,1.96))$ In particular, for the numerator we want $P(X \leq 0 \cap X+Y \in (0,1.96))$. This is then

$$= \int_{u=0}^{1.96} P(X \leq 0 \cap X+Y \in (u, u+du))$$
$$= \int_{u=0}^{1.96} P(X \leq 0 \mid X+Y \in (u, u+du)) \times P(X+Y \in (u, u+du))$$

To proceed we note that in general if $X \sim N(\mu, \sigma_x^2)$ and independently $Y \sim N(0, \sigma_y^2)$ then the deconvolution distribution of *X* given *X*+*Y* can be obtained by making the change of variable U=*X*+*Y* in the joint density function f_{XY} and then dividing by f_U. We obtain the distribution of

$$X \mid (X+Y=u) \sim N\left(\frac{\sigma_y^2\mu + \sigma_x^2 u}{\sigma_x^2 + \sigma_y^2}, \frac{\sigma_x^2\sigma_y^2}{\sigma_x^2 + \sigma_y^2}\right)$$

Here $\sigma_x = \hat{\sigma}$, $\sigma_y = 1$ and $\sqrt{\sigma_x^2 + \sigma_y^2} = \sqrt{\hat{\sigma}^2 + 1} = \sigma$

so here $P(X \leq 0 \mid X+Y=u) = \Phi\left(-\frac{\frac{\mu}{\hat{\sigma}} + \hat{\sigma} u}{\sigma}\right)$ where Φ is the cumulative distribution function of a standard normal variable.

The proportion of outcomes giving values in the 0 to 1.96 range, which are appropriately rejected as negative because they are counterproductive then is

 doi: 10.7243/2053-7662-2-5

$$\int_0^{1.96}\Phi\left(-\frac{\frac{\mu}{\hat{\sigma}}+\hat{\sigma}u}{\sigma}\right)\times\varphi(u,\mu,\sigma^2)du$$

where $\phi(u, \mu, \sigma^2)$ is the normal density function of $X + Y$. This is then the numerator of the negative predictive value. The denominator is

$$\Phi\left(\frac{1.96-\mu}{\sigma}\right)-\Phi\left(\frac{0.0-\mu}{\sigma}\right)$$

The parameter estimates and results of the negative predictive value calculation are given in the table below:

parameter	Point estimate	95% confidence interval	
μ	2.61	2.18	3.04
$\sigma=\sqrt{\hat{\sigma}^2+1}$	2.17	1.83	2.50
Neg pred val	7.81%	4.50%	10.35%

The calculations here are checked by simulation. A million values of $X \sim N(\mu,\hat{\sigma}^2)$ and $Y \sim N(0,1^2)$ are generated using the parameter point estimates above. A count is then kept of the instances in which $X < 0$ with $X+Y \in (0, 1.96)$. This gives the numerator of the (conditional) negative predictive value with the denominator being counted similarly. The negative predictive value by simulation is 7.75%.

Fourth Model

Model description

This model uses the subset of the original data where outcomes are relative risks and odds and hazard ratios. The distribution of the 70 values that comprise this data is displayed in **Figure 2**. Again it is assumed that the underlying effectiveness of treatments is distributed as $X \sim N(\mu,\hat{\sigma}^2)$, though this assumption is modified later. To this underlying distribution of effect size, additional variability is added due to the limited accuracy of any clinical trial, so if the i^{th} data point has a confidence interval indicating a variance of σ_i^2, then we consider that the outcome is a point from a $N(\mu,\hat{\sigma}^2+\sigma_i^2)$ distribution. The collection of all such outcomes would be values from a mixture distribution with 70 components. For curve fitting purposes, we appeal to the central limit theorem and assume that the distribution of outcomes can be reasonably approximated by an $N(\mu,\hat{\sigma}^2+\overline{\sigma}^2)$ distribution where $\overline{\sigma}^2$ is the average of the σ_i^2. An estimate of μ and $\hat{\sigma}^2$ is then obtained by fitting this distribution to the data.

To obtain the negative predictive value when treatments are declared ineffective, though their results are in the range (0, 1.96) standard deviations, we return to regarding the outcomes as coming from 70 separate distributions $X+Y_i \sim N(\mu,\hat{\sigma}^2+\sigma_i^2)$. Using reasoning analogous to that used in the third model, we require $P(X \le 0 \bigcap X+Y_i \in (0,1.96\sigma_i))$ which in the model can be calculated as

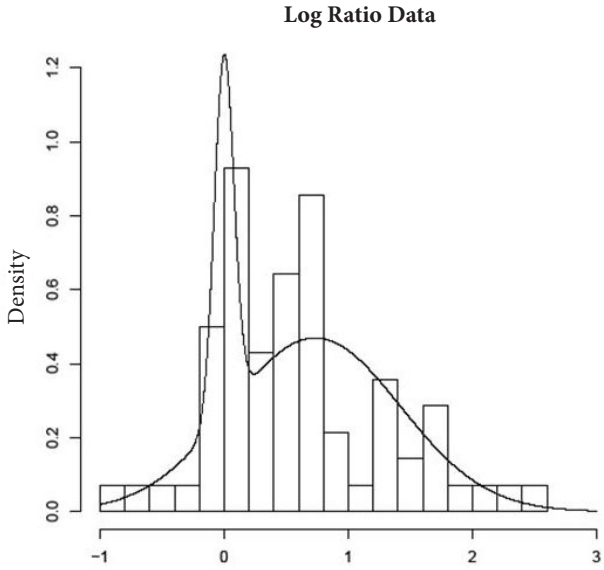

Figure 2. Histogram of logarithms of the relative risk, rate and ratio data, together with the mixture distribution $y=0.192\phi(x,0,\ 0.0789^2)+\ 0.808\phi(x,0.734,\ 0.686^2)$ where the numerical values used here are calculated as described in the sixth model.

$$\int_{u=0}^{1.96\sigma_i} P(X\le 0\mid X+Y_i\in(u,u+du))\times P(X+Y_i\in(u,u+du))$$

$$=\int_{u=0}^{1.96\sigma_i}\Phi\left(-\left(\frac{\mu\sigma_i}{\hat{\sigma}}+\frac{u\hat{\sigma}}{\sigma_i}\right)\Big/\sqrt{\hat{\sigma}^2+\sigma_i^2}\right)\varphi\left(u,\mu,\hat{\sigma}^2+\sigma_i^2\right)\mathrm{d}u=N_i \qquad (1)$$

The estimated value of the numerator is the sum of 70 such integrals, each modelling an outcome to do with one value u_i.

The estimated value of the denominator is the sum of 70 values of the probability $P(X+Y_i \in (0,\ 1.96\ \sigma_i))$. Each of these can be evaluated as

$$\Phi\left(\frac{-\mu+1.96\sigma_i}{\sqrt{\hat{\sigma}^2+\sigma_i^2}}\right)-\Phi\left(\frac{-\mu}{\sqrt{\hat{\sigma}^2+\sigma_i^2}}\right)=D_i \qquad (2)$$

The negative predictive value (conditional on weakly positive results) is then the estimated numerator over this estimated denominator.

Model 4 results

The parameter estimates and results of the negative predictive value calculation are given in the table below:

parameter	Point estimate	95% confidence interval	
μ	0.596	0.441	0.759
$\hat{\sigma}$	0.477	0.230	0.630
Neg pred val	7.96%	0.63%	12.32%

doi: 10.7243/2053-7662-2-5

The calculations here were checked by simulating a data set of 1,000,000 values where each value was determined by random numbers with the distributions assumed in the model and parameters estimated from the real data set. In this data set, as the simulated values are generated a count is made of the instances for which $X + Y_i \epsilon (0,\ 1.96\sigma_i)$ with $X < 0$ and compared with the numerator in the previously described parametric calculation. A check of the denominator is performed similarly. This simulation then gives a (conditional) negative predictive value of 7.85%.

Fifth model

Model description

An obvious oversimplification in the third and fourth model is the assumption that all treatments have some intrinsic effectiveness, mostly positive but occasionally negative. It would seem to be much more realistic to assume that a proportion, say *q*, of the treatments are in fact entirely useless whereas the remaining proportion 1-*q* have a distribution of effectiveness, with the distribution being mostly positive although occasionally treatments will be counterproductive. We again assume normality for the effective treatments. The overall distribution of the effectiveness of the treatments is then a mixture distribution $q \times \delta(0) + (1-q) \times N(\mu, \hat{\sigma}^2)$ where $\delta(0)$ is the Dirac delta function. To this underlying distribution, extra noise of the form $N(0, 1^2)$ is added, accounting for the random outcome about the true value given by a clinical trial, when outcomes are measured in standard deviations. The resulting distribution is then $q \times N(0, 1^2) + (1-q) \times N(\mu, \hat{\sigma}^2 + 1^2)$ The EM algorithm was then used on the data to estimate the three parameters of this mixture distribution - the proportion of the data corresponding to entirely ineffective treatments, and the mean and standard deviation $\sigma = \sqrt{\hat{\sigma}^2 + 1}$ of the outcomes which had a definite effect (mostly positive, but occasionally negative).

Again estimates of these parameters were used to calculate the negative predictive value given that the outcome was in the 0 to 1.96 standard deviation range. Using the same reasoning as in model 3, but incorporating the extra assumption of the mixture distribution, the expected value of the numerator of the negative predictive value calculation in this model becomes

$$0.475 \times q + (1-q) \times \int_0^{1.96} \Phi\left(-\frac{\frac{\mu}{\hat{\sigma}} + u}{\sigma} \right) \times \varphi(u, \mu, \sigma^2) du$$

The number 0.475 here refers to the proportion from a standard normal that is in the range (0, 1.96).
Similarly the denominator is

$$0.475q + (1-q)\left(\Phi\left(\frac{1.96 - \mu}{\sigma} \right) - \Phi\left(\frac{0.0 - \mu}{\sigma} \right) \right)$$

Model 5 results

The data for all 100 points is displayed in **Figure 1** together with the mixture distribution fitted by the EM algorithm to the 99 points excluding the outlier. A Shapiro Wilks test shows that the overall distribution could quite easily be regarded as a single Gaussian normal (*p*=0.321). However, the best fitting mixture distribution suggests a slight bulge in the histogram near zero.

The parameter estimates and results of the (conditional) negative predictive value calculation are given in the table below:

parameter	Point estimate	95% confidence interval	
q	0.0205	0.000	0.187
μ	2.66	2.27	3.26
σ	2.15	1.74	2.44
Neg pred val	10.7%	4.8%	31.9%

The fact that the lower end of the confidence interval for *q* includes zero, reflects the fact that an appreciable number of the bootstrap selections give data that the EM algorithm fits with a single Gaussian distribution, a result compatible with the previously quoted Shapiro-Wilks test showing that the data can quite easily be fitted with a single normal distribution.

A simulation of a million values for X and Y gave a simulated point estimate of the negative predictive value of 10.9%

Sixth model

This model extends the fourth model. It is again confined to the 70 outcomes that involved relative rates and ratios. However we now use the extra assumption that we are dealing with a mixture distribution representing treatments which are entirely useless together with treatments whose effectiveness is drawn from some underlying normal distribution. The underlying distribution of the treatments is then $q \times \delta(0) + (1-q) \times N(\mu, \hat{\sigma}^2)$. For each treatment, randomness in the outcome of the form $N(0, \sigma_i^2)$ is added.

The overall distribution is the average of 70 mixture distributions each with one normal component centred on 0 and the other on μ. We again appeal to the central limit theorem and assume that the outcomes as a whole are drawn from the distribution $q \times N(0, \sigma^2_{\text{ineffective}}) + (1-q) \times N(\mu, \sigma^2_{\text{effective}})$ This is used for curve fitting purposes.

The EM algorithm gives *q* as an estimate of the overall proportion of the first component, but also gives an individual weight w_i for the probability of a given data point coming from the first component. The distribution for a single outcome is then $w_i \times N(0, \sigma_i^2) + (1-w_i) \times N(\mu, \hat{\sigma}^2 + \sigma_i^2)$ and the overall distribution is the average of 70 such distributions.

It can then be seen that

$$\sigma^2_{\text{ineffective}} = \frac{\sum w_i \sigma_i^2}{\sum w_i}$$

doi: 10.7243/2053-7662-2-5

and

$$\sigma^2_{\text{effective}} = \hat{\sigma}^2 + \frac{\sum(1-w_i)\sigma_i^2}{\sum(1-w_i)}$$

allowing $\hat{\sigma}$ to be estimated.

Using the reasoning given in model 5 and model 4, we have the negative predictive value given an outcome in the range $(0, 1.96\sigma_i)$ is

$$\frac{\sum_{i=1}^{70} 0.475w_i + (1-w_i)N_i}{\sum_{i=1}^{70} 0.475w_i + (1-w_i)D_i}$$

where N_i and D_i are defined in equations (1) and (2) of model 4.

Model 6 results

The logarithm of the data for the 70 odds ratios, relative risks and rates is displayed in **Figure 2** together with the mixture distribution fitted by the EM algorithm to these 70 points. A Shapiro Wilks test shows that the overall distribution could not easily be regarded as a single Gaussian normal (p=0.028) and the mixture distribution described above is fitted.
The parameter estimates and results of the negative predictive value calculation are given in the table below:

parameter	Point estimate	95% confidence interval	
q	0.192	0.003	0.441
μ	0.734	0.554	1.02
$\hat{\sigma}$	0.471	0.000	0.618
Neg pred val	30.8%	5.2%	58.6%

A simulation with the parameter values above, was used to check the point estimate for the negative predictive value. Using 70 million simulated values obtained from the set of 70 equations $w_i \times N(0, \sigma_i^2) + (1-w_i) \times N(\mu, \hat{\sigma}^2 + \sigma_i^2)$ the simulated negative predictive value is 30.9%.

Result summary

It is helpful to summarise the various estimates we have for the negative predictive value of a treatment being correctly labelled as ineffective conditional on there being a positive outcome but with a p-value >0.05. In assessing these figures, it should be noted that models 1 and 2 are based on 100 data points, models 3 and 5 are based on 99 data points and models 4 and 6 are based on 70 pairs of data points, with each pair giving a point estimate and a measure of variability. It should also be noted that models 1 and 2 do not exclude any outlier, models 3 and 5 exclude one outlier and models 4 and 6 implicitly exclude two outliers. Models 3 and 4 exclude the possibility of treatments being purely useless.

Model	Point estimate	95% confidence interval	
1	36.2%	13.7%	71.4%
2	36.7%	21.5%	50.4%
3	7.8%	4.5%	10.4%
4	8.0%	0.6%	12.3%
5	10.7%	4.8%	31.9%
6	30.8%	5.2%	58.6%

It is important to note that all the point estimates of the negative predictive value are considerably less than 50%. This tells us that when EBM declares a treatment to be ineffective despite an outcome in an RCT being positive (albeit not convincingly so in terms of $p \leq 0.05$ then EBM is likely to be wrong more often than not. The bootstrap confidence intervals however show there is considerable uncertainty about the estimates.

Discussion

Alternative interpretations of the data

The models here basically assume that the difference between the proportion of weak positive values and weak negative values is explained by effective treatments giving unconvincing results. A possible alternative explanation is a form of publication bias that applies within the subset of all RCTs which show weak results and differentiates between RCTs that show weak positive and weak negative outcomes. There are various estimates of the extent of publication bias when the contrast is between results that are "statistically significant" and those which are not and where commercial pressures may not be excluded. A median estimate is a ratio of 2.3 [**19**]. One would expect the extent of publication bias to be much smaller between weakly positive or negative studies for which the conclusion either way is "no effect" and so it seems unlikely that publication bias would account for the almost 3-fold difference between weak positive studies and weak negative studies seen here. Indeed one could speculate on whether any tendency to publish results which are slightly perverse would balance out a tendency to publish weak results in the other direction. It seems more likely that most of the 3-fold difference is accounted for by Type II error. If so, provisionally labelling treatments as having "no effect", when outcomes are weakly positive would seem to be an inappropriate practice, at least when the treatments are low cost and when there are prior reasons to expect them to be effective.

Limitations of the models and their implications for the results

Model 1 implicitly assumes that chance effects will result in treatments which are truly positive, contaminating the set of weakly negative outcomes to the same extent as treatments which are truly negative contaminate the set of weakly positive outcomes. Since truly positive treatments almost

doi: 10.7243/2053-7662-2-5

certainly outnumber the truly negative, it can be seen that this will lead to the model overestimating the (conditional) negative predictive value.

In model 2, if we assumed that the effective treatments were somewhat less effective than the 80% power assumption implies, it can be seen that the three compartments into which results are classified (outcomes <0, outcomes ∈ (0,1.96 standard deviations) and outcomes >1.96, could then be best fitted by taking q=0. The implication is that if power is consistently overestimated then this model will overestimate the (conditional) negative predictive value. The converse is also true. However, the introductory comments in this paper are relevant here - the desirability of detecting small effects may be ignored because that would require trials that are infeasibly large. Since this suggests that power tends to be overestimated, the implication for model 2 is that it is more likely that the q here is overestimated so that the (conditional) negative predictive value is in turn overestimated.

Models 3 to 6 simplify by assuming normal distributions. It is reasonable to believe that there will be some treatments trialled that will be precisely useless and where chance effects will yield results normally distributed about zero. On the other hand, there is no compelling reason to believe that the central limit theorem will apply to disparate treatments that do have some effect, so as to make their outcomes fit a normal distribution and there is no reason to assume a normal distribution for the totality of treatment effects. However, the errors caused by such simplifying assumptions may be reduced because it is only the region around the weak positive portion of the distribution of outcomes that is most relevant to the calculations. Models 3 and 4 on the one hand and models 5 and 6 on the other, can be seen as representing extremes of a reasonable range of options in this region. In models 3 and 4 smoothness is assumed around the zero effectiveness mark whereas in models 5 and 6 this part of the distribution is interrupted with a delta spike degraded by uncertainties in measurement. Models 3 and 4 give lower estimates of the proportion of useless treatments than models 5 and 6 because the latter allow for a treatment to be useless in two ways - either because they are counterproductive or because they are purely ineffective whereas models 3 and 4 reject treatments only to the extent that they are assessed as genuinely counterproductive. Since it may be conjectured that there will be an appreciable proportion of treatments which have just a minimal positive effect (which perhaps would still be worthwhile if these treatments were very low cost), the true negative predictive value may lie between the two extremes represented by models 3 and 4 on the one hand and models 5 and 6 on the other hand. In other words, it may be reasonable to expect that the true answer lies between the values suggested by these two sets of models.

We see then, that for models 1, 2, 5 and 6, the estimates of the (conditional) negative predictive value may be too high, whereas for models 3 and 4 the estimates may be too low.

Model extensions

This work could be extended to avoid the parametric assumption of a Gaussian mixture model and instead use numerical methods applied to the actual distribution of the outcomes. We have in effect a deconvolution problem, where we have a "blurring" function added to the intrinsic effectiveness of each treatment to give the pattern of outcomes. A unique solution to the underlying distribution of the effectiveness of the treatments could perhaps be obtained by constraining this distribution to be smooth in some way. This approach is likely to be involved and is likely to require much more data for reasonable accuracy. It will not be pursued here.

Bayes' theorem together with the models, could be used to calculate the negative predictive value as a function of the outcome in terms of the standard deviations from zero. Given the wide confidence intervals already obtained, such an approach does not seem warranted without more data.

Implications for EBM

Considerable attention has been paid to the problem of false positives in EBM, the inappropriate acceptance of treatments that later turn out to be ineffective. Publication bias due to a preference for positive results is always an issue, but is a particular problem when there are commercial pressures concerning new drugs under patent [15]. Publication bias is also relevant together with a regression to the mean effect, when out of a set of many novel treatments, a few are unduly accepted because of promising initial results which are not backed up by later assessments [20]. There has been considerable research into publication bias as a source of undue type I errors and there are quantitative methods based on analysis of funnel plots to assess this [21]. There is even a Cochrane Collaboration study on publication bias in studies of publication bias [19]. However, there seems to have been less attention paid to the converse problem of false negatives - in particular, the inappropriate rejection of low cost treatments that external considerations indicate must have some, albeit slight, effect, but which are labelled by EBM as having "no effect". Whilst concerns about this converse problem have been expressed [2,4,5], this author is not aware of any previous quantitative analysis of this problem.

The undervaluing by EBM of prior knowledge from biological science or commonsense and lack of accounting for the cost of errors, particularly Type II errors, makes one very suspicious of the "no effect" label. In nearly all cases the bland statement that the treatment showed "no effect", will be false if taken in a precise literal sense. It is noted that for only 1 out of 100 trials selected as data here, was it literally true that the outcome showed no effect. As well as being false in a precise literal sense, this paper shows that the pronouncement "no effect" is false in its practical implications. The calculations here show that more often than not, it is likely to be wrong if a declaration is made that a treatment is ineffective when the outcome is positive, but does not achieve $p \leq 0.05$. This

conclusion is reached using only statistical analysis. No account is taken of the fact that in all cases there was enough prior information from science or commonsense to suspect the treatment would be effective, and so warrant an RCT. As in much of EBM, the analysis here comes to its conclusion using frequentist statistics, free of any a priori reasoning. Unfortunately, with the limited amount of data collected, the different approaches used in this paper indicate considerable uncertainty about the estimate of the (conditional) negative predictive value. However, looking at the estimates collated in result summary section and the discussion above on model limitations, it would not be unreasonable to summarise by stating that as a best guess, when outcomes are positive but $p>0.05$, there is a probability of perhaps 2/3 that the treatment is indeed effective.

The implication is, that leaving aside the costs of errors (or assuming the cost of Type I and II errors are equal [**10**]), EBM seems too conservative in its assessment of RCTs. A conservative approach may be reasonable to limit over-enthusiastic acceptance of new treatments where there are commercial pressures and a likelihood of publication bias together with a high financial cost and possibly a high cost in terms of side effects [**15**]. The same approach seems inappropriate when it is used to reject established treatments supported by conventional wisdom and with clear mechanisms of action. The case of the rejection of basic principles of hygiene for injections discussed in the introduction [**3**], is in this author's opinion, just one of numerous examples of very low cost treatments where the occasional benefit is certainly worthwhile, but where the treatment has been inappropriately rejected by EBM.

Whether it is reasonable to decide to act as though a treatment is ineffective, should depend not only on p-value calculations but also on prior knowledge, some estimate of the relative costs of Type I and Type II errors and a synthesis of all these factors. The purpose of this paper is to argue for a wiser use of statistics in medicine. In the absence of a full decision theory analysis of each treatment in medicine, this paper is an argument for EBM to be modified so that, particularly for treatments where commercial pressures are not relevant, wisdom and mechanism is accorded more status.

Competing interests

The author declares that he has no competing interests.

Acknowledgement

I thank my son, Dr. Sam Kault and anonymous reviewers for their helpful comments.

Publication history

Editor: Zhongxue Chen, Indiana University Bloomington, USA.
EIC: Jimmy Efird, East Carolina University, USA.
Received: 30-Apr-2014 Final Revised: 17-May-2014
Accepted: 21-May-2014 Published: 17-Jun-2014

References

1. Hine LK, Laird N, Hewitt P and Chalmers TC. **Meta-analytic evidence against prophylactic use of lidocaine in acute myocardial infarction**. *Arch Intern Med.* 1989; **149**:2694-8. | Article | PubMed
2. Braithwaite RS. **A piece of my mind. EBM's six dangerous words**. *JAMA.* 2013; **310**:2149-50. | Article | PubMed
3. Del Mar CB, Glasziou PP, Spinks AB and Sanders SL. **Is isopropyl alcohol swabbing before injection really necessary?** *Med J Aust.* 2001; **174**:306. | PubMed
4. Smith GCS and Pell JP. **Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials**. *BMJ.* 2003; **227**:1459-1461. | Article
5. Westover MB, Westover KD and Bianchi MT. **Significance testing as perverse probabilistic reasoning**. *BMC Med.* 2011; **9**:20. | Article | PubMed Abstract | PubMed Full Text
6. Toff NJ. **Human factors in anaesthesia: lessons from aviation**. *Br J Anaesth.* 2010; **105**:21-5. | Article | PubMed
7. **Centre for Evidence Based Medicine**, *Oxford University*. 2013. | Website
8. Schulz KF, Altman DG and Moher D. **CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials**. *Ann Int Med*. 2010:152.
9. Fisher. **Dr Fisher's casebook**. *Significance*. 2004; **1**:1:26. | Article
10. Lilford RJ and Johnson N. **The alpha and beta errors in randomized trials**. *N Engl J Med.* 1990; **322**:780-1. | Article | PubMed
11. Gibbons JD, Olkin I and Sobel M. **Selecting and ordering poppulations - a new statistical methodology**. *John Wiley & sons.* 1977.
12. D'Agostino RB, Sr., Massaro JM and Sullivan LM. **Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics**. *Stat Med.* 2003; **22**:169-86. | Article | PubMed
13. **The Cochrane library**. | Website
14. Press WH, Flannery BP, Teukolsky SA and Vetterling WT. **Numerical Recipes: the art of scientific computing (fortran version)**. *Cambridge University Press*. 1989.
15. Goldacre B. **Bad Pharma: How drug companies mislead doctors and harm patients**. Fourth Estate, 2012 (UK).
16. Lonergan E, Britton AM, Luxenberg J and Wyller T. **Antipsychotics for delirium**. *Cochrane Database Syst Rev.* 2007; CD005594. | Article | PubMed
17. Dempster AP, Laird NM and Rubin DB. **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *J Roy Stats Soc Series*. 1977; **39**:1-38. | Pdf
18. Rice JA. **Mathematical statistics and data analysis**. *Duxbury Press.* 1995:p153.
19. Dubben H and Beck-Bornholdt H. **Systematic review of publication bias in studies on publication bias**. *BMJ.* 2005; **331**:433. | Article
20. Prasad V, Vandross A, Toomey C, Cheung M, Rho J, Quinn S, Chacko SJ, Borkar D, Gall V, Selvaraj S, Ho N and Cifu A. **A decade of reversal: an analysis of 146 contradicted medical practices**. *Mayo Clin Proc.* 2013; **88**:790-8. | Article | PubMed
21. **Cochrane Handbook for Systematic Reviews of Interventions**. | Website

Citation:
Kault D. **False negatives in evidence based medicine**. *J Med Stat Inform*. 2014; **2**:5.
http://dx.doi.org/10.7243/2053-7662-2-5