ARTICLE

The failure of evidence-based medicine?

D. Stephen Hickey BA PhD MSB CBiol[a], Andrew Hickey Dip Comp (Oxon)[b] and Leonardo A. Noriega BA MSc PhD LLB(CPE) MBCS[c]

a Head of Newlyn Research Group, Newlyn, Penzance, UK
b Senior Researcher, Newlyn Research Group, Newlyn, Penzance, UK
c Senior Lecturer, Faculty of Computing, Engineering and Technology, The Octagon Staffordshire University, Beaconside, Stafford, UK

Abstract

Evidence-based medicine (EBM) claims to provide gold-standard methods based on group and population statistics. However, the main issues in clinical medicine concern classification and prediction. During diagnosis, a patient's illness is classified; then it is predicted that a specific treatment will be successful with that particular patient. Most scientific disciplines concerned with classification and prediction have rejected group and population statistics as being misleading, inadequate and inaccurate for such applications. This paper lists some of the critical findings from the decision sciences that bring the utility, application and validity of EBM into question. One of the foundations of EBM is that large clinical trials provide the best evidence. However, EBM misapplies the law of large numbers and best evidence really means selected data. EBM is inconsistent with modern science, theoretically unsound, impractical and erroneous in its application.

Keywords

Clinical trials, cybernetics, EBM, evidence-based medicine, individual patients induction systems, person-centered medicine, scientific method, statistics, statistical populations

Correspondence address

Dr Len Noriega, Faculty of Computing, Engineering and Technology, The Octagon, Staffordshire University, Beaconside, Stafford, ST18 0AD, UK. Email: l.a.noriega@staffs.ac.uk

Accepted for publication: 10 April 2012

Introduction

Most physicians agree that randomised controlled trials are considered a gold standard, but interpretation of the term evidence-based medicine (EBM) differs [1]. EBM is medicine based on aggregate statistical results from large clinical trials. Recently, however, a series of technical challenges to the validity of EBM have emerged. The challenges include philosophical, statistical and practical issues. Hickey and Roberts [2] suggest that EBM does not conform to the requirements of good science, rational decision making or information theory. This paper explains some key elements of these limitations and a detailed account of misinterpretation of statistics, particularly those from large clinical trials.

We take the concept of a rational doctor-patient combination as axiomatic [2]. Informed choice is based on the premise that medical decisions should be rational, as to suggest that such decisions be irrational would be absurd. In order for a patient to make a rational decision, the doctor needs to present unbiased information and have the authority to use suitable treatments [3]. Ross Ashby, the psychiatrist and cybernetician, provided a concise description of rational decision-making: use what you know to narrow the field as far as possible and after that do as you please [4]. Thus, a requirement for rational patient choice is that the patient and doctor have the knowledge, ability and freedom to consider the data.

Medicine increasingly relies on group statistics from clinical trials as a guide to treatment. This is exemplified in the use of evidence-based medicine and its use as a guide for the treatment of individual patients. The nominal guidelines for EBM allow for the inclusion of a doctor's expertise and a patient's unique circumstances when making statistical evidence-based decisions about patients [5]. The introduction of EBM has occurred along with an increase in the size of published clinical trials and the use of systematic reviews to enlarge the population considered. Bland found that in 1972 the median size of trials in the *Lancet* and *British Medical Journal* (BMJ) were 33 and 37 subjects respectively [6]. For the same month in 2007, 35 years later, the corresponding figures for the *Lancet* and BMJ were 3,116 and 3,104 subjects, respectively. This increase of almost 2 orders of magnitude corresponds to a widespread belief that larger trials provide more reliable evidence. The law of large numbers states that the precision of estimates of the mean value increases with escalating sample size [7,8]. The EBM paradigm is built around application of the law of large numbers and group statistics in clinical trials.

Despite claims that EBM introduces more scientific methods into medicine, it remains controversial. Here we describe some critical issues with the use of EBM for the treatment of patients. We show that the use of the law of large numbers to suggest benefits for large clinical trials can be misleading. Furthermore, we explain how most quantitative scientific disciplines, involved with classification and prediction, have rejected group statistics. Importantly, day-to-day practical medicine, such as finding a diagnosis and selecting a suitable treatment for a patient, falls into the category of classification and prediction.

Populations or people?

The analytic divergence of populations and individuals has been generally recognised. Aggregate statistics such as the mean and standard deviation describe the normal and similar distributions, but model the population not the individual data points. In economics, Friedrich Hayek described how the flow of information needed to pass from the individuals to the controllers [9]. Aggregate statistics do not provide the fine grained data resolution required for economic analysis. Correspondingly, economic modelling often works on the principle of emergence from independent, autonomous, intelligent agents [10]. Similar wisdom of crowd concepts provide the foundation of numerous disciplines including evolution [11], complexity theory [12], physics of systems [13], swarm intelligence [14], prediction markets [15], Delphi methods [16], crowd simulation [17] and computer graphic simulation [18]. A key feature of this approach is the emergence of order out of complexity that highlights the potential deficiencies of the linear statistics used in EBM and by related organisations [19].

Starting with cybernetics [20], a number of disciplines have addressed the inappropriate use of group statistics for individual prediction. Pattern recognition [21], machine intelligence [22], computer vision [23], neuroscience [24], artificial intelligence [25], induction systems [26], computational intelligence [27] and related disciplines that are explicitly concerned with classification and prediction generally do not depend on group statistics. It is an accepted finding that group statistics are not accurate or reliable, in individual prediction or classification. Group statistics describe or are used to analyse population data, or samples, but are not directly applicable to individuals.

Population and group statistics used in EBM can be viewed as a method of data compression [28] providing aggregate population data and sample estimates. However, this data compression is 'lossy', dropping information about the individual datum. The induced information loss means collective statistics describing the normal and similar distributions are computationally efficient. Notably, the primary statistical methods employed in EBM were developed in the pre-computer society. Least squares arose from Legendre and Gauss starting in the early 1800s [29], meta-analysis by Pearson in 1904 [30], Students t-test in 1908 [31], Fisher's F-ratio in the 1920s [32], analysis of variance in the 1930s and so on. However, it is now possible to process large datasets and make specific predictions for individual people. In the world of decision science, EBM methods appear increasingly old and antiquated.

These issues are often described with reference to the ecological fallacy, also called the fallacy of division, which states that you should not apply population statistics to the individual [33]. Classically, you can determine the mean shoe size with a little arithmetic; but do not give everyone the average shoe, because most will be disappointed with the fit. The ecological fallacy is related to the fallacies of composition and division; what is true of a sample is not necessarily true of the whole and, conversely, findings from the whole do not necessarily describe the sample [34]. Group statistics do not apply to an individual patient, in the same way as an EBM statistician would not make a prediction for the whole population from a single case report.

Rational patient expectations

We take a rational patient and a rational doctor to be the primary decision centre. A rational patient has an expectation that the treatment received has a reasonable probability of providing him or her with overall benefit. Correspondingly, a rational doctor would provide only those treatments with a reasonable probability of benefiting the patient. At a minimum, the expected gain should exceed the risk of harm. Similarly, game theory, economics and related disciplines, suggest that a rational patient would not accept risk of side effects, or other harm, with no dominating expected benefit [35]. We exclude from the analysis the special case of medical altruism, where a patient might undergo the risk of treatment to provide benefit for others.

A reasonable probability that a treatment is successful is somewhat vague. Moreover, risk is defined as cost (or benefit) multiplied by probability. To be definite, we suggest a rational patient might expect to receive the most effective treatment for a typical illness and have a sizable chance of benefiting from it. The critical values vary with the patient, illness and specific conditions and may be modelled by the probability of the individual's successful treatment, $p_{s.}$ This probability of successful treatment is often approximated in clinical trials using the number needed to treat (NNT).

The first law of cybernetics

It is an axiom of EBM that well-designed, large-scale, placebo-controlled, randomised, clinical trials (RCTs) provide the best data for treating patients. This assumption is based on the statistical techniques employed and has not been tested against the competing methodologies [2]. Nor has it been adequately explained how RCTs avoid the ecological fallacy.

The ecological fallacy is a direct consequence of *Ashby's law of requisite variety* also known as the first law

of cybernetics. Ashby's requisite variety provides a minimum limit to the information required to solve, or even fully describe, a problem [36]. This is a general finding, in information theory, for example, Shannon's 10th theorem is considered a special case of Ashby's law [37]. An effective solution needs an equal amount of information, or variety, to the problem itself. If there is insufficient information, the solution will be ineffective. Ashby gave the example of a bacterial infection. A person must have a sufficient number of varied antibodies to protect against the array of possible invading organisms [38]. To demonstrate the validity of EBM, it is a minimum requirement that, despite the inherent data compression, RCT statistics conform to Ashby's law when applied to the individual patient.

Good regulator theorem

While Ashby's law has widespread acceptance, there is a secondary restriction on the regulation of systems. Ashby and his student Roger Conant provided the *good regulator* theorem [39]. An effective solution, or medical treatment, needs to be a specific model of the problem. To extend Ashby's infection example, the patient must have an effective antibody that explicitly and selectively pattern matches the invading organism. That is the shape and structure of the antibody's active site is a close match, or model, of part of a bacterial protein but does not fit the host proteins. Note, prediction and classification are implied in this process of *pattern recognition*. Similarly, an enzyme's active site matches the substrate molecule as the acetabulum models the head of femur [40].

Current large clinical trials are often based on a multivariate risk factor model. Diseases are assumed to be complex and are addressed by determining the associated risk factors and their linearly independent contributions or correlations. This has been found universally in other disciplines to invite the curse of dimensionality which degrades the utility of the analysis [41,42]. Corresponding to Ashby's law and the good regulator theorem, practical systems have an intrinsic dimensionality, a maximum number of independent factors required for full description. Moreover, the solution must be isomorphic with the system being regulated [39]. The curse of dimensionality is often described mathematically in terms of an exponentially increasing segmentation of higher dimensional spaces [43], or the introduction of high dimensional random noise which cannot be separated from the embedded information [2].

For medical decisions we require Goldilocks solutions that are not too simple, do not contain too many risk factors and are not over complicated [2]. Good regulators are models that map onto the intrinsic dimensionality and thus there are a specific and usually small number of factors required for a practical solution. Large trials and related solutions can be over-determined. The inherent danger is that over-determined models over-fit the data [44], producing deceptively accurate results which fail when used for real world classification and prediction.

Unfortunately, the multiple risk factor model of a disease in statistical medicine forms an adaptive system [45,46]. As a result, scientific refutation of such adaptive models may be impractical within the EBM framework; multiple risk factor models simply adjust to incorporate new data. The classic case of external refutation is the direct experimentation by Marshall and Warren showing peptic ulcers to result from *Helicobacter pylori* infection [47]. With ulcers, direct experiment provided a good regulator model comprising a parsimonious solution and an effective treatment, which dominated the earlier risk factor explanations. There are many other examples of over-complicated risk factors descriptions from the history of medicine being replaced by the germ theory of disease; mycobacteria as the cause of tuberculosis is a specific case [2].

Good model making is compulsory rather than optional [39]. With certain minimal restrictions, any attempt to find or validate a new treatment needs to model the clinical situation. The clinical target is a rational patient and his or her doctor. For medicine, the system is based on specific doctor-patient decision-making and the unique circumstances and biology of the individual patient and the disease. The good regulator theorem implies that to be effective the methods employed in clinical science must model discrete decision-making in the individual patient-doctor system. A form of pattern recognition is the implied approach. Group statistics emerge from the analysis, but are not good regulators.

Probability of successful treatment

We can use the binomial distribution to model a series of independent, individual, clinical trial experiments. Here, the experiment is the treatment of a single patient. The group mean, m_s, is the probability of successful treatment, p_s, multiplied by the number of trials or patients, n, that is:

$$m_s = np_s \quad (1)$$

the variance, σ^2, is given by

$$\sigma_s^2 = np_s(1 - p_s) \quad (2)$$

The binomial distribution is closely related to the normal distribution which is often used as an approximation to it, with a minor continuity correction [48]. The normal approximation increases in accuracy with n but decreases with more extreme values of p_s, that is, probabilities close to 0 or 1. This approximation is robust and a minimum heuristic for the approximation to be useful is n>5 [49]. Larger values of n (>20) are normally considered adequate depending on the application. The primary limitation with the binomial distribution is that we are dealing with a binary (e.g., sick, healthy) criterion.

Clinical trials

Standard EBM clinical trials typically are built around the assumption of normality and tests of the difference between mean values. Here, we use the normal distribution and the corresponding z-test of the difference between means, as an exemplar of a statistical test in a clinical trial:

$$z = (m - \mu)/s \qquad (3)$$

where m is the sample mean and μ the population mean. The standard error, SE, is $\sigma/\sqrt{n}$ where σ^2 is the variance. The use of the normal distribution and z-test as exemplar is not a critical assumption for the discussion; however, it facilitates a definite and straightforward explanation.

In the z-test, we have 2 samples one controls and the other treated patients, which we can approximate using the binomial distribution, with sample means:

$$m_t = n_t p_{ts} \text{ and } m_c = n_c p_{cs} \qquad (4)$$

and sample variances, s^2,

$$s_t^2 = n_t p_{ts}(1 - p_{ts}) \text{ and } s_c^2 = n_c p_{cs}(1 - P_{cs}) \quad (5)$$

where the subscripts t and c indicate treated and controls respectively and subscript s is a reminder that we are dealing with the probability of successful treatment. Substituting equations 4 and 5 into equation 3 and combining the sample variances gives:

$$z = n_t p_{ts} - n_c p_{cs}/\sqrt{[n_t p_{ts}(1 - p_{ts})/n_t + n_c p_{cs}(1 - p_{cs})/n_c]}$$

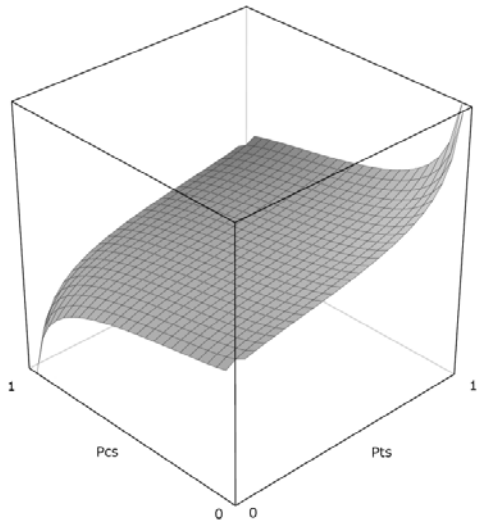
For simplicity, assume the samples have equal size, $n=n_t=n_c$, then:

$$z = n(p_{ts} - p_{cs})/\sqrt{[p_{ts}(1 - p_{ts}) + p_{cs}(1 - p_{cs})]} \qquad (6)$$

The value of z is thus proportional to the sample size n and a term containing p_{ts} and p_{cs}. However, the value of the (p_{ts}, p_{cs}) term is constrained for practical values. This constraint is the case except where p_{ts} approaches 1 while p_{cs} approaches 0, or conversely p_{cs} approaches 1 but p_{ts} approaches 0, as shown graphically in Figure 1. These extreme conditions apply when a treatment gives near perfect results in an otherwise hopeless disease or almost all untreated patients recover while the treatment causes disease continuation in effectively all patients. Such conditions are clinically rare.

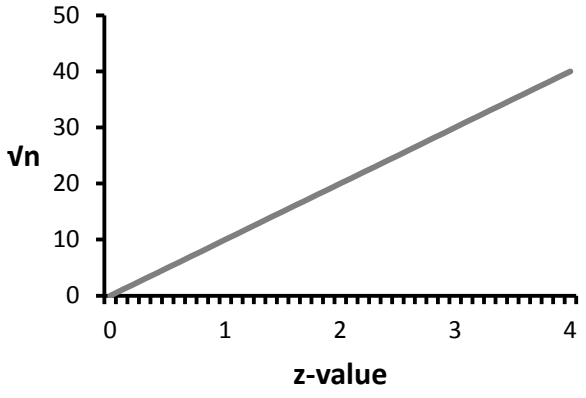
We should be clear about the meaning. We have specified the z-test in terms of the probability that an individual patient will benefit from the treatment, p_{ts}, compared with being untreated, p_{cs}. These probabilities are not the p-values normally quoted in clinical trials, which relate to the difference between the groups and are often obtained from the value of z using statistical tables. This measure of the treatment benefiting the individual patient (p_{ts}-p_{tc}) is related to the aggregate effect size. Equation 6 tells a rational patient that the z-test value for the normal distribution is proportional to and dominated by, the sample size, n.

Figure 1 The term $(p_{ts}-p_{cs})/\sqrt{[p_{ts}(1-p_{ts})+p_{cs}(1-p_{cs})]}$ is shown for the values of p_{tc} and p_{cs} between 0 and 1. The range of the computed values shown is ±4.9.



A rational patient is concerned with the chance that the treatment will work. A clinical study with large n, can obscure the probability difference (p_{ts}-p_{cs}). Thus, there is a danger that statistical testing of the difference between groups may represent the size of the study, rather than the ability to predict the benefit for a patient.

Figure 2 The linear relationship between the **square root of the study size, √n and the** calculated z-value is shown. This established relationship arises from the law of large numbers.

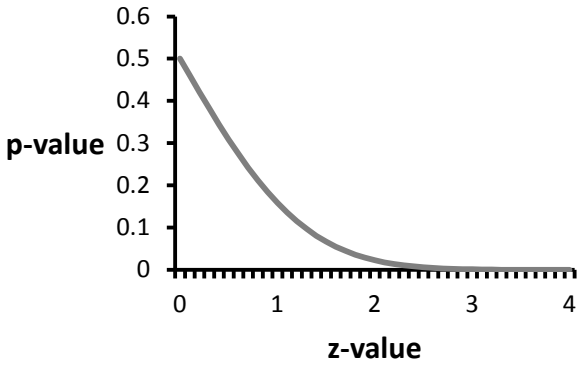


Normal probabilities

Here, we describe the probability of group difference, the p-value. The variation of z-value is linearly proportional to

√n and this well established relationship is shown in Figure 2. This linear relationship with √n arises from the reduction in the standard error on the mean, SE=σ/√n, in equation 3 and is the basis of claims for increased precision in large trials. We reproduce this chart for direct visual contrast to the rapid decay of the corresponding p-values. The variation of the group probability (p-value) with the z-value is non-linear, as shown in Figure 3. The decline in p-value with increasing z-value is fast. That is the decay in p-value is far more rapid and in practice dominates the linear increase in z-value or the corresponding linear decrease in standard error.

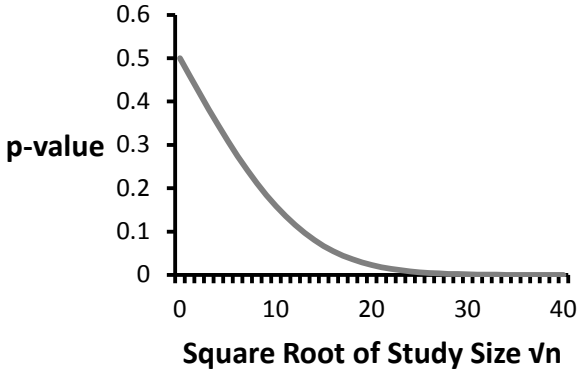
Figure 3 The variation of p-value with the value of the z-test is shown. These are standard tabular results displayed graphically to illustrate the rapid decline in derived probability with increasing z.



Thus, the effects of the increase in precision with increasing n may be misunderstood. The corresponding increase in the z-value (and decrease in standard error) with study size provides benefit only if there is a large (dominating) correction to the acceptable p-value (confidence limit). The variation in p-value with √n shown in Figure 4. For this figure, we assumed fixed difference between means (1) and constant variance (10). This assumption is used merely to facilitate graphical representation and other values produce similar results. To clarify, $p<0.05$ in the year 1975 (with median n about 30) is not the same as $p<0.05$ in 2007 (with median n>3,000). In effect, the rules have changed.

Recent successful trials would be rejected had the criterion for acceptance not changed. The introduction of EBM has weakened statistical requirements for clinical trials. It is immediately apparent, from Figure 4, that a rapid non-linear decrease in the computed probability accompanies an increase in study size. Importantly, this rapid non-linear decline in p-value dominates the linear relationship shown in Figure 2 for moderate values of n. Consequently, the p-value is a decreasing measure of a rational patients expected benefit (p_{ts}-p_{cs} in Equation 6) as the study size increases. In practical terms, for the same p-value a larger study is *less* useful as a predictor of a useful treatment. This finding contradicts the conventional supposition of an increased utility of large clinical trials, which arises from the law of large numbers and its effect on reducing the standard error.

Figure 4 This figure shows the non-linear decline in p-value with (the square root of) study size. This rapid decline dominates the linear increase in standard error (Figure 3).



Simulation method

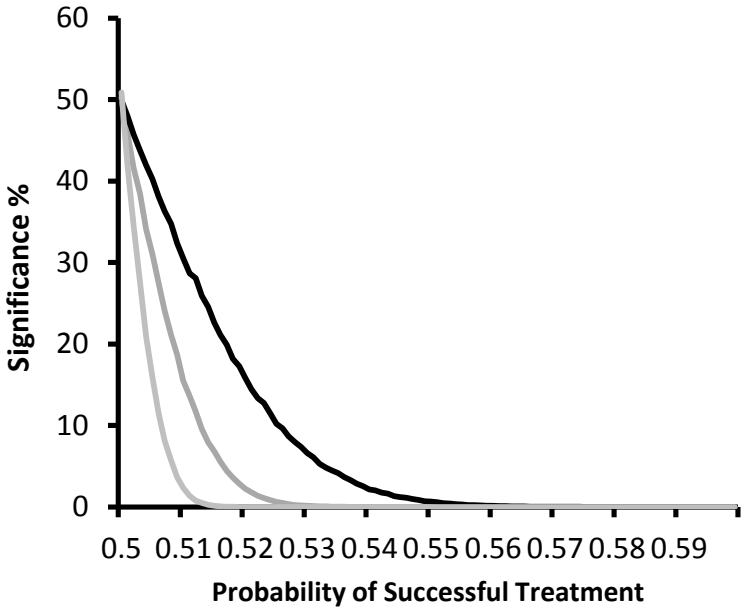
We ran Monte-Carlo simulations in Java using the fast FastMersenne Twister algorithm based on version MT199937 of the algorithm for raw pseudorandom number generation [50] using a standard Java pseudorandom seed. These numbers were employed to produce a series of binary strings to represent positive (return to health) and negative (remain sick) outcomes in a clinical trial. The results naturally form a binomial distribution. The binomial distribution was used as a model of the normal distribution, the inverse of the common functional mapping. Then the z-value was calculated from the computed sample means and variances. Finally, the normal probability was computed from the z-value using a standard numerical integration method implemented in Java [51].

For most runs, we used a standard sample size of 100 patients for ease of representation as percentage treatment success and averaged over up to 1,000,000 experimental repeats for consistency. We varied the sample size from 10 to 10,000 patients in trial experiments. For investigative simulation, we employed trial runs of 10,000 experimental repeats and these are the results displayed here.

Results of simulation

Typical simulation results for the relationship between the probability of successful treatment and the group p-value for a study size of 50, 100 and 200 controls and equal number of treated patients are shown in Figure 5. This figure assumes that half the patients will recover naturally, $p_{cs}=0.5$. Note that Figure 5 indicates that a 5% significance ($p = 0.05$) corresponds to a marginally effective benefit to

Figure 5 Results of Monte-Carlo simulation of the percentage significance (or p-value) in terms of the probability of patient benefit, p_{ts}. A value for p_{ts} of 0.55 means that a patient has a 1 in 20 chance of benefit. The curves represent study sizes of 50 (black line), 100 (dark grey line), and 200 (light grey line).



an individual patient that decreases with increasing study size.

From Figure 5, a treatment only needs to help 3 patients in 100 (NNT≈33) on average to be significant with n=50. Moreover, a highly significant group probability ($p<0.01$) corresponds to helping an average of about 5 patients in 100 (NNT≈20) with this small sample size. Very high significance ($p<0.001$) occurs when about 8% of patients receive benefit (and NNT≈12). Taking a sample size of 200, $p<0.05$, $p<0.01$ and $p<0.001$ all occur when a rational individual's expectation of benefit is approximately 1%. Large trials are not simply bigger versions of small studies. The results of Monte Carlo simulation support the contention that, for rational patients, the results of large significant trials are less important than the findings of smaller studies.

Selection of data

A central claim of EBM is its use of the best evidence. The best evidence is selected according to varied rules based around the assumed gold standard of large-scale placebo-controlled randomised clinical trials. However, all selection other than random choice is based on information. Use of the best evidence implies knowledge of the underlying reality and cannot be based on a statistical or other methodology. Selection bias results from non-random selection of data [52]. The inadvisability of such selection was known to the Greek philosophers, notably Epicurus.

EBM's scientific consistency is superficial and the approach is technically inadequate. Use of the best evidence implies EBM is a local rather than global search strategy [53]. The implication is that EBM has a limited application, specifically to social medicine. We note that EBM has been criticised for positivism in that it emphases verification and observation, but is less strong in terms of explanation [54]. A good inductive system is one that approximates to Solomonoff Induction [55] in that it uses all available data and updates current beliefs with additional information [56]. Science is a practical example of such induction [2,57]. A general rule in science is that it is invalid to preferentially select "good" data, for example, outliers are not removed when plotting a graph [58]. The data selection in EBM's use of the "best evidence" is inconsistent with the scientific method and with a good inductive system.

Meta-analysis is a form of systematic review in which data is non-randomly selected from the literature and combined using linear statistics. Some EBM organisations recognise meta-analysis as the ultimate form of evidence [59,60]. We examined series of meta-analyses in both the Cochrane database and the *Journal of the American Medical Association* (JAMA) and every review inspected contained major statistical and methodological errors [61].

In JAMA, we examined 38 papers containing the term meta-analysis in the title or abstract, published in the years 2005-2006, but not one was statistically sound. The reviewers were free to select the studies. Narrative fallacy could have explained biased selection events after they have occurred [62]. The reviewers had information about

the results of the clinical trials they were choosing. Eight of the 38 published reviews were not actually meta-analyses. One of the reviews of observational studies had poor blinding in study selection, but included all reasonable studies [63]. A second review of refugees had some blind data extraction, but studies were openly selected [64]. The other 28 reviews selected their clinical trials with no blinding.

In total, 30 out of 30 JAMA meta-analyses were subjective, with potentially biased study selection. The authors of the reviews and the people selecting the data had access to the names of the original study authors, the full text of the studies and the study results. Of the meta-analyses, 6 did not select independently and 4 did not make the selection method clear. The selection involved "independent" researchers in the remaining 19 reviews. However, it was not clear in what way they were considered independent. People who know the data cannot be expected to provide an independent selection.

Selection bias dominated the meta-analyses. In the JAMA reviews, almost all the available data were excluded. Of 39,894 studies, only 962 (2.4%) were included, while the remaining 38,932 studies were ignored. The reviewers handpicked particular study authors and contacted them, to provide additional information. In half (15/30) of the reviews, there was communication with study authors. However, in only 3 cases was it declared that authors or representatives from all the studies were contacted. The reviewers obtained non-peer reviewed data for chosen studies of particular interest.

Fourteen JAMA reviews used additional unpublished or non-reviewed data. The reviewers apparently found it appropriate to ask selected scientists for additional unpublished information. Six included unspecified data that could not be independently verified, from study authors that they had chosen to contact. Consistent with the suggestion of narrative fallacy, inconsistent reasons were given for excluding studies. Only 14 of the 30 meta-analyses in JAMA restricted their data to papers from peer-reviewed journals. Five reviews provided incomplete criteria on inclusion and exclusion. Only 3 reported that they used pre-prepared selection criteria before gathering the data, which should have been an essential if inadequate requirement, as the reviewers were familiar with the data.

Even pre-prepared data selection is not reassuring. If the authors knew the literature then, consciously or unconsciously, they could choose the selection criteria to achieve a desired result. Thus, they could have biased the criteria. Conversely, if they were not familiar with the literature, it might be argued that they should not be attempting the review. Authors of 25 of the meta-analyses could have chosen their criteria after the selection was made. Finally, in 19 of the reviews, the outcome measures were not identical from one study to another; the review was comparing apples with oranges [65]. All the meta-analyses published in the *Journal of the American Medical Association*, in the period 2005 to 2006 were flawed, as summarised in Table 1.

The Cochrane Foundation reviews also failed to meet minimum decision science requirements. Cochrane reviews included only a small fraction of the total available information.

Table 1 Summary of major statistical errors in JAMA meta-analyses.

JAMA meta-analyses examined (2005-2006)	
Total number	38
Actual meta-analyses	30
Number of studies considered	39,894
Number of studies selected	962 (2.4%)
Independent selection	19 (63%)
Inappropriate selection	30 (100%)
Authors contacted	15 (50%)
Used only peer reviewed studies	14 (47%)
Additional unpublished data used	6 (20%)
Selection criteria incompletely specified	5 (17%)
Outcome measures invalid	19 (63%)
Post hoc criteria possible	25 (83%)

The Cochrane Foundation reviews also failed to meet minimum decision science requirements. Cochrane reviews included only a small fraction of the total available information. In Cochrane reviews, data selection is validated by suggesting that at least 2, preferably independent, people were expected to assess the eligibility of studies [66], using methods that are transparent, minimize bias and human error. They do not specify how 2 people could be independent, minimize bias and reduce human error. Cochrane described at length the potential for errors that arise from non-blinding and lack of randomisation in the original clinical trials, but ignored the selection bias in their own reviews.

We examined 100 reviews in the Cochrane archives, chosen by searching on the term "meta-analysis". Five of the reviews were protocols or experimental designs. Of the remaining 95 reviews, only 3 had any blinding in selecting the clinical trials. The first of these 3 had some blinding of the results; however, the selectors examined the abstracts to determine eligibility [67]. The selectors had a summary of the trials, the results and the conclusions. Since the abstract summarises the paper, this approach would not prevent bias. The second of these 3 reviews was partially blinded to the names of authors, institution and funding sources. However, the selectors knew the study results [68]. In the third review, a third party removed the title, authors and results, but the review authors would presumably be competent in the field, familiar with the literature and the blinding ineffective [69]. Thus, selection bias could not be avoided, even in these 3 exceptional reviews, which had at least acknowledged the problem and tried to minimize the resulting error. Most of the reviews were completely deficient. The remaining 92 of the 95 reviews had no blind study selection. In 90 reviews, both the names of the study authors and their conclusions were available for study selection. The full study text was used in study selection for 91 reviews. Additional non-peer

reviewed and unpublished material was included in 65 of the reviews. The reviews selected evidence based on published study quality and then included extra unpublished and unspecified data from preferred authors.

Non-peer reviewed clinical trials were included in 67 of the reviews. Furthermore, 6 reviews included unpublished results from unnamed "experts". Only 7 reviews made it explicit that the selection criteria were chosen before examining the data. Unconsciously or otherwise, 88 of 95 reviewers may have decided how to select the data after they knew the results.

All of the 95 reviews had inadequate study selection. Only about half of the reviews (47) stated the numbers of studies considered. In these, a mere 1.1% of total clinical trials were selected. Finally, only 71 of the reviews measured the same outcomes. Once again, they were comparing apples with oranges: one trial might report a change in blood pressure, while another mentioned increased cholesterol. The results from Cochrane are summarized in Table 2.

Table 2 Results from a brief analysis of Cochrane reviews in 2007 is presented here. None of the reviews were statistically sound.

Cochrane Meta-analyses	
Total number	100
Actual meta-analyses	95
Any partial blinding at all in selection	3 (3%)
No blinding in selection	92 (97%)
Authors names used in selection	90 (95%)
Full text used for selection	91 (96%)
Used additional unpublished results	65 (68%)
Used non peer reviewed trials	67 (70%)
Opinion/data from "experts" used	6 (6%)
Specific non post hoc criteria	7 (7%)
Clearly defined outcome measures	71 (75%)

Discussion and Conclusion

A rational patient looking at the data from EBM clinical trials could conclude that they have little relevance to his or her medical treatment. The trials do not provide the necessary classification and prediction data necessary for treating an individual. EBM trials are subject to the ecological fallacy and larger trials provide less adequate data [2]. The multiple risk factors can produce over-determined adaptive systems that over-fit the data and provide results that are deceptively robust, but inaccurate in the real world. EBM's risk factor models also invite the curse of dimensionality. The assumption that EBM results provide useful predictive data for treatment or disease prevention appears not to be explicitly addressed within the paradigm. Perhaps most importantly, EBMs methods are not good regulators, as they model groups and populations, rather than the treatment of individual patients. Good modelling is essential for an effective methodology.

EBM's statistical methods are particularly unsuited to classification and prediction in clinical medicine. The results are not valid pattern recognition and should not be applied to the treatment of individual patients as each patient has dominating specific characteristics. The primary utility of EBM appears to be providing data for government and large organisations. Aggregate statistics are essential when determining the provision of medical resources to a large population in a city or country. EBM data could however be used as background information or provide an indication of prior probabilities in a Bayesian analysis. In addition, EBM data might provide information on disease incidence useful for a clinical test. In general, the benefits of EBM are peripheral to the central issues of classification and prediction in treating patients or preventing illness.

Miller and Miller suggest that in promoting statistics-based research, EBM has divorced itself from real-world common sense and scientific causation [70]. The rational patient approach described here reaches the same conclusion using largely cybernetic methods. Information theory, game theory, and the fundamentals of inductive systems provide another view of EBM. There is more to information and decision-making than the use of aggregate statistics and the selective evidence-base in EBM may be illusory [2].

Our simple analysis of the utility of normal statistics used for predicting actual patient benefit suggested that large trials are particularly inappropriate. In statistical terms, the larger the trial the smaller is the significant effect size. Large trials lack relevance to a rational patient. Since the 1970s, the increase in study size has made the results of clinical trials less important to evaluating the utility of a possible treatment. To compare current large trials to earlier results requires a non-linear reduction of the p-values to correct for sample size. From the viewpoint of a rational patient, a statistical power calculation is appropriate only when, p_{ts}, the chance of actual patient benefit is large and held constant.

It would appear that current EBM methods are aimed at showing that a marginally effective drug provides a statistically significant benefit, when averaged over a large population. It is clearly easier to produce a slightly effective drug than a blockbuster and the costs and restrictions on large-scale trials help produce a monopoly for multinational companies. The increasing size of clinical trials could be covering the failure of EBM as a useful predictive or classification system for patients. Fortunately, we do not need to speculate. Stafford Beer, the management scientist, provided a cybernetic rule for evaluating systems and their behaviour, POSIWID or *The Purpose of a System Is What It Does* [71]. POSIWID is intended for systems in general and is rational unless the system is broken or malfunctioning. Beer suggested POSIWID to be the default cybernetic position unless a specific coherent and convincing explanation is forthcoming. In this case, POSIWID indicates that EBM is not intended to help rational patients.

The problems we describe for meta-analysis are easily replicated or refuted, by direct perusal of published

reviews. In 2010, Shamliyan *et al* has reported similar statistical problems with meta-analysis reviews [72]. Despite the methodological limitations, meta-analyses have a place in providing some supporting data for a hypothesis, particularly for social medicine and large populations. However, they have limitations when compared with human reviews while retaining their inherent selectivity and subjectivity [2]. From the viewpoint of a rational patient, the primary concern is the position of meta-analyses or other systematic reviews, at the peak of the EBM evidence hierarchy.

Direct decision science methods could provide a rational doctor-patient based approach. An early medical system MYCIN for suggesting antimicrobial treatment outperformed specialists [73] but was never used in practice. Doctors may have felt that their authority and autonomy were threatened by the technology. Alternatively, there was resistance to a machine making medical decisions. However, a medicine based on science, data and information would naturally place rational decision-making with the individual doctor-patient unit. In a scientific medicine, the individual doctor would have medical autonomy and the patient a rational choice.

References

[1] Raman, R. (2011). Evidence-based medicine and patient-centered care: cross-disciplinary challenges and healthcare information technology enabled solutions. *International Journal of Person Centered Medicine* 1 (2) 279-294.

[2] Hickey, S. & Roberts, H. (2011). Tarnished Gold: The Sickness of Evidence-based Medicine. CreateSpace.

[3] Miles, A. & Mezzich, J.E. (2011). The care of the patient and the soul of the clinic: person centered medicine as an emergent model of modern clinical practice. *International Journal of Person Centered Medicine* 1 (2) 207-222.

[4] Ashby, W.A. (1960). Computers and decision making. *New Scientist*, March 24, 746.

[5] Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B. & Richardson, W.S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal* 312 (7023) 71-72.

[6] Bland, M. (2009). The tyranny of power: is there a better way to calculate sample size? *British Medical Journal* 339, 1133-1135.

[7] Feller, W. (1971). Law of large numbers for identically distributed variables. In: An Introduction to Probability Theory and Its Applications, vol 2, 3rd ed, pp. 231-234. Oxford: Wiley.

[8] Feller, W. (1968). The Strong Law of Large Numbers. In: An Introduction to Probability Theory and Its Applications, vol. 1, 3rd ed, pp. 243-245. Oxford: Wiley.

[9] Hayek, F. (1945). The use of knowledge in society. *American Economic Review* XXXV, 4, 519-530.

[10] Beinhocker, E.D. (2006). Origin of Wealth: Evolution, Complexity, and the Radical Remaking of Economics. Boston: Harvard Business School Press.

[11] Weibull, J.W. (1997). Evolutionary Game Theory. Cambridge MA: The MIT Press.

[12] Casti, J.L. (1979). Connectivity, Complexity and Catastrophe in Large-scale Systems. Oxford: John Wiley & Sons.

[13] Albert, R. & Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics* 74, 47–97.

[14] Dorigo, M. & Birattari, M. (2007). Swarm intelligence. *Scholarpedia* 2 (9) 1462.

[15] Wolfers, J. & Zitzewitz, E. (2004). Prediction Markets. *Journal of Economic Perspectives* 18 (2) 107-126.

[16] Tayler, W.J. (2005). Preliminary identification of core domains for outcome studies in psoriatic arthritis using Delphi methods. Developing assessment methodology in psoriatic arthritis. *Annals of the Rheumatic Diseases* 64 (Supplement 2) ii110-ii112.

[17] Shendarkar, A., Vasudevan, K., Lee, S. & Son Y. (2006). Crowd simulation for emergency response using BDI agent based on virtual reality. In: Proceedings of the 38th conference on Winter simulation (WSC '06). L. Felipe Perrone, Barry G. Lawson, Jason Liu & Frederick P. Wieland (Eds.) pp. 545-553. Winter Simulation Conference.

[18] Terzopoulos, D. (1999). Artificial life for computer graphics. *Communications of the ACM* 42 (8) 32-42.

[19] McKelvey, B. & Andriani, P. (2005). Why Gaussian statistics are mostly wrong for strategic organization. *Strategic Organization* 3 (2) 219-228.

[20] Wiener, N. (1948). Cybernetics or Control and Communication in the Animal and the Machine. Oxford: John Wiley & Sons.

[21] Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition, Second Edition (Computer Science & Scientific Computing). Amsterdam: Academic Press.

[22] Pal, S.K. & Biswas, S. (2005). Pattern Recognition and Machine Intelligence, First International Conference, *PReMI 2005*, Kolkata, India, December, Proceedings, Springer.

[23] Davies, E.R. (2005). Machine Vision, Third Edition: Theory, Algorithms, Practicalities (Signal Processing and its Applications). Amsterdam: Morgan Kaufmann.

[24] Ballard, D. (1999). An Introduction to Natural Computation (Complex Adaptive Systems). Cambridge MA: The MIT Press.

[25] Bender, E.A. (1996). Mathematical Methods in Artificial Intelligence. Oxford: Wiley-Blackwell.

[26] Solomonoff, R. (1978). Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory* 24 (4) 422–432.

[27] Engelbrecht, A.P. (2007). Computational intelligence: an introduction. Oxford: Wiley-Blackwell.

[28] Hankerson, D.C., Harris, G.A. & Johnson, P.D. (2003). Introduction to Information Theory and Data Compression, 2nd ed. London: Chapman and Hall/CRC.

[29] Stigler, S.M. (1990). The History of Statistics: The Measurement of Uncertainty before 1900. Boston: Harvard University Press.

[30] O'Rourke, K. (2007). An historical perspective on meta-analysis: dealing quantitatively with varying study

results. *Journal of the Royal Society of Medicine* 100 (12) 579-582.
[31] Mankiewicz, R. (2004). The Story of Mathematics. Princeton NJ: Princeton University Press.
[32] Fisher, R.A. (1925). Statistical Methods For Research Workers. Edinburgh: Oliver and Boyd.
[33] Robinson, W.S. (1935). Ecological correlations and the behavior of individuals. *Journal of the American Statistical Association* 30, 517-536.
[34] Vogt, W.P. (2011). Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences. London: Sage Publications.
[35] von Neumann, J. & Morgenstern, O. (2007). Theory of Games and Economic Behavior. Princeton NJ: Princeton University Press.
[36] Ashby, W.R. (1956). An Introduction to Cybernetics. London: Chapman & Hall.
[37] Krippendorff, K. (2009). Ross Ashby's information theory: a bit of history, some solutions to problems, and what we face today. *International Journal of General Systems* 38 (2) 189-212; correction 38 (6) 667-668.
[38] Ashby, W.R. (1958). Requisite variety and its implications for the control of complex systems, *Cybernetica* 1 (2) 83-99.
[39] Conant, R. C. & Ashby, W.R. (1970). Every good regulator of a system, *International Journal of Systems Science* 1 (2) 89-97.
[40] Scholten, D.L. (2010). Every good key must be a model of the lock it opens. www.goodregulatorproject.org accessed 4 Nov 2011.
[41] Powell.W.B. (2007). Approximate Dynamic Programming: Solving the Curses of Dimensionality. Oxford: Wiley-Interscience.
[42] Marsland, S. (2009). Machine Learning: An Algorithmic Perspective. London: Chapman and Hall/CRC.
[43] Duda, R.O., Hart, P.E. & Stork D.G. (2000). Pattern Classification. Oxford: Wiley-Interscience.
[44] Hawkins, D.M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Science* 44 (1) 1-12.
[45] Miller, J.H. & Page, S.E. (2007). Complex Adaptive Systems: An Introduction to Computational Models of Social Life. Princeton NJ: Princeton University Press.
[46] Mareels, I. & Polderman, J.W. (1996). Adaptive Systems: An Introduction. Basel: Birkhäuser.
[47] Marshall, B.J. & Warren, J.R. (1984). Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* 161 (8390) 1311-1315.
[48] Feller, W. (1945). On the normal approximation to the binomial distribution. *Annals of Mathematical Statistics* 16 (4) 319-329.
[49] Box, G.E.P., Hunter, J.S. & Hunter, W.G. (2005). Statistics for Experimenters: Design, Innovation, and Discovery. Oxford: Wiley Series in Probability and Statistics.
[50] Matsumoto, M. & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation* 8 (1) 3–30.
[51] Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1992). Numerical Recipes in C: The Art of Scientific Computing, Second Edition. Cambridge: Cambridge University Press.
[52] Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica* 47 (1) 153-162.
[53] Migdalas, A., Pardalos, P.M. & Värbrand, P. (2001). From Local to Global Optimization. Amsterdam: Springer.
[54] Walsh, B. & Gillett, G. (2011). A post-structuralist view of evidence-based medicine (EBM): what EBM contributes to philosophy. *International Journal of Person Centered Medicine* 1 (2) 223-231.
[55] Solomonoff, R.J. (1964). A formal theory of inductive inference: parts 1 and 2. *Information and Control* 7, 1-22 & 224-254.
[56] Hutter, M. (2007). Algorithmic information theory. *Scholarpedia* 2 (3) 2519.
[57] Rathmanner, S. & Hutter, M. (2011). A philosophical treatise of universal induction. *Entropy* 13, 1076-1136.
[58] Huff, D. (1991). How to Lie with Statistics. London: Penguin.
[59] Merlin, T., Weston, A. & Tooher, R. (2009). Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. *BMC Medical Research Methodology* 9, 34.
[60] Hemingway, P. & Brereton N. (2009). What is a systematic review? What is Series. *Bandolier*, April.
[61] Hickey, S., Hickey, A. & Noriega, L.A. (2009). Implications and insights for human adaptive mechatronics from developments in algebraic probability theory. Presented at the EPSRC UK Postgraduate Workshop on Human Adaptive Mechatronics (HAM), Staffordshire University, 15–16 January.
[62] Taleb, N.N. (2007). The Black Swan: The Impact of the Highly Improbably. New York: Random House.
[63] Juffer, F. & van IJzendoorn, M.H. (2005). Behavior Problems and mental health referrals of international adoptees, a meta-analysis. *Journal of the American Medical Association* 293, 2501-2515.
[64] Porter, M. & Haslam, N. (2005). Predisplacement and postdisplacement factors associated with mental health of refugees and internally displaced persons, a meta-analysis. *Journal of the American Medical Association* 294, 602-612.
[65] Barone, J.E. (2000). Comparing apples and oranges: a randomised prospective study. *British Medical Journal* 321 (7276) 1569-1570.
[66] Higgins, J.P.T. & Green, S. (2008). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 February, *The Cochrane Collaboration*.
[67] Kendrick, D., Coupland, C., Mulvaney, C., Simpson, J., Smith, S.J., Sutton, A., Watson, M. & Woods, A. (2007). Home safety education and provision of safety equipment for injury prevention. *Cochrane Database of Systematic Reviews*, Issue 1, CD005014.
[68] Adams, N., Lasserson, T.J., Cates, C.J. & Jones, P.W. (2007). Fluticasone versus beclomethasone or budesonide for chronic asthma in adults and children. *Cochrane Database of Systematic Reviews*, Issue 4, CD002310.
[69] Ferguson, T., Wilcken, N., Vagg, R., Ghersi, D. & Nowak, A.K. (2007). Taxanes for adjuvant treatment of

early breast cancer. *Cochrane Database of Systematic Reviews*, Issue 4, CD004421.
[70] Miller, C.G. & Miller, D.W. (2011). The real world failure of evidence-based medicine. *International Journal of Person Centered Medicine* 1 (2) 295-300.
[71] Beer, S. (2004). What is cybernetics? *Kybernetes: The International Journal of Systems & Cybernetics* 33 (3-4) 853-863.
[72] Shamliyan, T., Kane, R.L. & Jansen, S. (2010).Quality of systematic reviews of observational nontherapeutic studies. *Preventing Chronic Disease* 7 (6) A133.
[73] Yu, V.L., Fagan, L.M., Wraith, S.M., Clancey, W.J., Scott, A.C., Hannigan, J., Blum, R.L., Buchanan, B.G. & Cohen, S.N. (1979). Antimicrobial selection by a computer. A blinded evaluation by infectious diseases experts. *Journal of the American Medical Association* 242 (12) 1279-1282.