A new statistical trend in clinical research – Bayesian statistics

Article *in* Physical Therapy Reviews · October 2010

DOI: 10.1179/174328810X12786297204756

Arnold YL Wong
The Hong Kong Polytechnic University
126 PUBLICATIONS **1,584** CITATIONS
SEE PROFILE

Greg Kawchuk
University of Alberta
167 PUBLICATIONS **3,274** CITATIONS
SEE PROFILE

Narrative Review

A new statistical trend in clinical research – Bayesian statistics

Arnold Y. L. Wong, Sharon Warren, Gregory N. Kawchuk*

Department of Physical Therapy, Faculty of Rehabilitation Medicine, University of Alberta, Canada

Background: The emphasis on evidence-based practice in physical therapy has increased the number of clinicians who perform and interpret clinical research. Unfortunately, the traditional statistical analysis (frequentist approach) used most often in clinical research (except meta-analysis) has been criticized by biostatisticians for potential bias and misleading results if used with data from single studies. Alternatively, Bayesian inference can be used instead of the traditional frequentist approach although this trend has yet to be seen in rehabilitation research. Used for at least three decades, the Bayesian approach provides a formal framework for researchers to incorporate prior knowledge and current evidence to derive new probabilities for various hypotheses. Since the results are presented in terms of probability, clinicians can interpret and apply research findings to clinical practice directly.

Objectives: The objectives of this review are to discuss the common misconceptions among users of the frequentist approach, the inherent limitations of the frequentist approach, as well as to introduce the characteristics and limitations of the Bayesian approach using illustrated examples.

Conclusions: The Bayesian approach can be used as an alternative or adjunct to the frequentist method in future studies. This approach is also robust in situations that are unfavourable to traditional statistics such as sequential clinical trials. However, biostatisticians may have to be consulted for some sophisticated Bayesian analysis. As the Bayesian approach may gain popularity, a good understanding of this method will benefit clinicians in interpreting research papers and planning their future clinical studies.

Keywords: Bayesian approach, Clinical research, Hypothesis testing, Rehabilitation, Statistics, Review

Introduction

With the increasing demand for evidence-based clinical practice, rehabilitation clinicians are expected to be competent in both interpreting results of research papers and carrying out clinical research at their workplace.[1] While most clinicians have been trained in hypothesis testing using a traditional approach (also known as the frequentist method),[2] they may not be aware of some of the limitations of this traditional statistical method[3,4] or the presence of alternative statistical methods.

The purpose of inferential statistics is to generalize sample findings to a targeted population parameter (a characteristic that describes a population). There are two mainstream approaches to inferential statistical methods, namely, the frequentist and Bayesian approaches. The frequentist approach applies the concept of proof by contradiction.[5] Although there are different statistical tests under the frequentist approach, a typical frequentist test usually involves two hypotheses: a null hypothesis (H_o) and an alternative hypothesis (H_a). Frequentists presume H_o to be true before the start of the experiment in order to predict the outcome. If the empirical sample data do not support H_o, H_o is rejected.[5] This method is traditionally classified as an objective method. However, some biostatisticians criticize this method because conclusions are based on the results of a single study.[4]

In contrast to the frequentist approach, the Bayesian approach derives the probabilities ('posterior' probability) of population parameters from the empirical data of the current experiment and the corresponding probabilities ('prior' probability) of these parameters reported in previous studies.[6] Such deduction is achieved by the repeated use of Bayes' theorem, discussed below. The resulting inferences are expressed by the 'posterior' probabilities of the observations conditional on the parameters. The observation with the highest posterior probability is usually accepted as more likely to be true. However, some researchers criticize the 'subjective' determination of 'prior' probability.[6]

Despite the perception that the Bayesian approach is subjective, it has been proposed as a surrogate or

Correspondence to: Gregory N Kawchuk, Department of Physical Therapy, Faculty of Rehabilitation Medicine, University of Alberta, 3-44 Corbett Hall, Edmonton, Alberta T6G 2G4, Canada. Email: greg.kawchuk@ualberta.ca

DOI 10.1179/174328810X12786297204756

an adjunct to the frequentist approach because it incorporates different knowledge (like the physiological knowledge) and evidence from previous studies[4,7] into statistical analysis. These considerations can prevent the incorrect (although inadvertent) interpretation of the results and improve the credibility of studies. The Bayesian approach is now applied to many scientific areas.[7–11] Leading medical journals have also provided guidelines to users of the Bayesian method.[12,13] Unfortunately, there are few, if any, examples of the use of the Bayesian approach in rehabilitation research. The aims of this paper are to (1) highlight common misconceptions among users of the frequentist method, (2) identify limitations of the frequentist statistical approach, (3) introduce the concept and limitations of Bayesian inference and (4) give suggestions for future applications in rehabilitation research. By understanding these two statistical approaches, clinicians may be more confident in evaluating, designing and performing clinical research.

Historical Development of the Frequentist Approach

The contemporary frequentist approach originated from the concepts of *P*-value and hypothesis testing. The use of the *P*-value was first proposed by R. A. Fisher in the 1920s.[14] By definition, the *P*-value represents the probability of obtaining a result at least as extreme as the actual observed data if the null hypothesis is true.[6] Therefore, the *P*-value measures the evidence against a single H_o (e.g. null effect of treatment). Fisher did not provide specific guidelines for using the *P*-value. He simply suggested combining the *P*-value with background information in order to justify the H_o rejection. Although Fisher proposed a mathematical method to deduce *P*-value, his method makes no reference to any alternative hypothesis. It simply examines whether the observed data look extreme or not.[15] Neyman and Pearson proposed an alternative statistical method to make inferences using the concept of hypothesis testing, in which H_o (there is no difference or no effect), H_a (an opposite of H_o), alpha (α, probability of false positive error) and beta (β, probability of false negative error) value are predefined for the inference.[16] Two types of error may arise in hypothesis testing. Type I error (false positive error) occurs if a researcher rejects a true H_o,[5] while Type II error (false negative error) occurs if a researcher fails to reject a false H_o.[5] Neyman and Pearson suggested researchers make inferential judgments based on the relative importance of these errors.[16] This method emphasizes limiting Type I or Type II errors of a statement in the long-run at the expense of proving the truth or falsehood of each separate hypothesis.

Two separate concepts of *P*-value and hypothesis testing are, however, combined in current frequentist statistics. In a typical approach, H_o and H_a are determined after the identification of a clear research question. Since the test statistic (a characteristic that describes a sample, such as the mean) of random samples drawn from a population will vary from one another, the resulting sampling distribution of test statistic will form a specific pattern (e.g. normal distribution).[5] Sampling distribution of a test statistic under H_o is the theoretical distribution of the test statistic if H_o is true. Such distribution together with a predetermined alpha-value can verify whether H_o can correctly predict the sample test statistic. An alpha-value is conventionally set at either 0.01 or 0.05. The extremely unlikely sample outcomes, as defined by the alpha-value, make up the critical regions in the ‘tails’ of the sampling distribution of a test statistic under H_o. If the test statistic (evidence) falls into the critical region, H_o is rejected. As a result, one will conclude that the observed result is unlikely to occur by chance given H_o is true.[5] Therefore, the frequentist method uses an alpha-value to confine the long-run probability of Type I error and makes an inference solely based on the sample data. Although it is a logical concept, there are some common misconceptions among some who use the frequentist approach.

Common Misconceptions of the Frequentist Approach

One common misconception of the frequentist approach is the interpretation of the *P*-value. The *P*-value is commonly misinterpreted as the Type I error of sample statistics.[17] Although both *P* and alpha-values represent the tailed-area probability of the sampling distribution of a test statistic under H_o, the *P*-value is not a measure of sample Type I error.[4,5] The *P*-value only indicates the probability of finding the observed value or more extreme value if H_o is true. This concept is illustrated by the six-minute walk test (6 MWT) results of a resistance exercise (RE) group in a study examining the efficacy of home-based exercise programmes among breast cancer survivors.[18] The 6 MWT was carried out at baseline and 12-weeks post-exercise programme on subjects in the RE group. H_o stated that the difference in the walking distance of 6 MWT between pre- and post-12-week exercise of subjects in RE group was less than or equal to zero. The Wilcoxin-Signed Rank Test showed a significant improvement in their walking distance ($Z=2.366$, one-tailed $P=0.009$) at the end of 12-week period. The alpha-value of the test was set at 0.01. Figure 1 displays the test result. One should note that the curve representing the sampling distribution of differences in

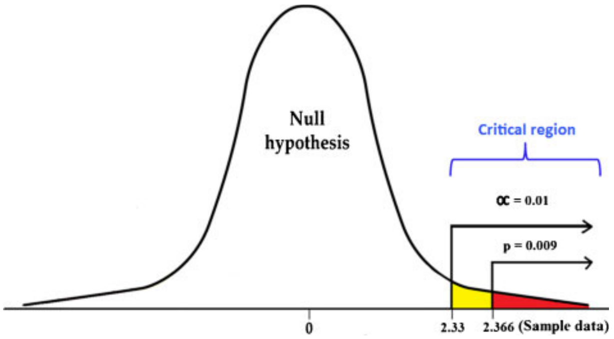

Figure 1 A sampling distribution of difference in walking distance of 6 MWT of pre- and post-resistance exercise programme in a study investigating the efficacy of home-based exercise on breast cancer survivors. It represents the probability of every possible outcome under null hypothesis (H_o). With α=0.01, *Z*=2.33 defines the boundary that separates the extreme 1% from the rest of 99% possible value of sampling distribution under H_o. The critical region, determined by alpha-value, is composed of extreme sample values that are very unlikely to be observed if H_o is true. The *P*-value is the probability of finding results equal to or more extreme than the actual observed sample data given H_o is true. The *P*-value is obtained after data collection (*P*=0.009 when the empirical data, *Z*=2.366). Both alpha and *P*-value are found at the tail region of this sampling distribution. (The above sampling distribution may not reflect the actual distribution from sample data and the critical region should cover the infinite value on the right hand side of the curve).

pre- and post-walking distance of 6 MWT under H_o is not in the exact scale and the actual distribution of data may not be normally distributed. The *P*-value in this case was very small (*P*=0.009) since the observed value was found at the extreme end of this sampling distribution under H_o. However, the *P*-value can sometimes be very large if the observed value is close to the centre of the distribution or has a sample mean close to the value specified in H_o. Unlike the *P*-value, the alpha-value is adopted by researchers before the commencement of an experiment. The alpha-value is independent of the experimental results while the *P*-value depends on the results. In other words, the *P*-value in a single study can only serve as evidence against H_o. It is not equivalent to the alpha-value, which represents the maximum Type I error that researchers will tolerate if a true H_o is rejected.

Another common error made when using the frequentist approach is treating the *P*-value as either the probability of having the observed data due to chance, or the probability of a true H_o.[19] The distinction between these errors and the true definition is the absence of the phrase 'if H_o is true'.[19] Since H_o is assumed to be true in the hypothesis test, the *P*-value can only quantify the probability of obtaining a result equal to or more extreme than the actual observation, given H_o is true. It cannot calculate the actual probability of a true H_o. The frequentist method is not intended to measure the probability of any hypothesis.

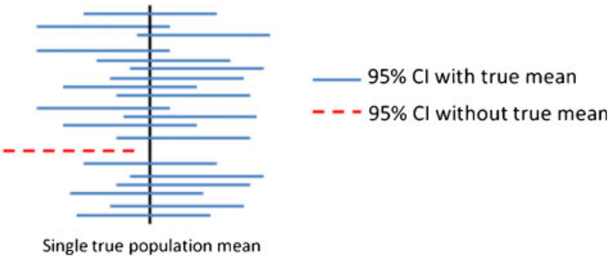

Figure 2 Graphical representation of 95% confidence intervals (CI) of a population mean. Each horizontal solid line represents a 95% CI that contains the true population mean. The horizontal dotted line represents a 95% CI that does not contain the true population mean. The vertical line represents the true population mean. A 95% CI will either contain the true population mean or it will not. It is an all-or-none phenomenon

Besides the common misunderstanding of the *P*-value, some clinicians may misinterpret the meaning of a 95% confidence interval (CI), therefore affecting their application of research findings to clinical practice. In the frequentist approach, every population parameter (such as a true treatment effect) is usually assumed to be a single unknown value that will vary over time. A 95% CI is interpreted as the long-run probability of this interval including the true population parameter. It only denotes that if random samples of the identical sample size are repetitively drawn from a population, 95% of the resulting sample mean intervals will contain the true population mean while 5% will not (Fig. 2). Hence, a deduced 95% CI of sample mean from a single sample may or may not contain the true population mean. Instead of interpreting CI in terms of probability, one should only state that there is 95% confidence that the actual population mean can be found within this interval.

Despite the aforementioned misconceptions, one should note that the frequentist approach is a very useful statistical tool as long as users are aware of common misconceptions.

Limitations of the Frequentist Approach

Like any type of statistical method, the frequentist approach has its limitations. Since the frequentist method does not intend to estimate the probability of different hypotheses, it may not provide the most pertinent information to clinicians. Clinicians always want to know the probability of a true H_o or H_a given the observed evidence. In a hypothetical experiment investigating the effect of ultrasound therapy (US) and cryotherapy (Fig. 3), the frequentist approach can provide a *P*-value or CI of an observed experimental result. Although it provides useful inference, clinicians may prefer to directly obtain the probability of different hypotheses, which allows them to choose the most effective treatment modalities.

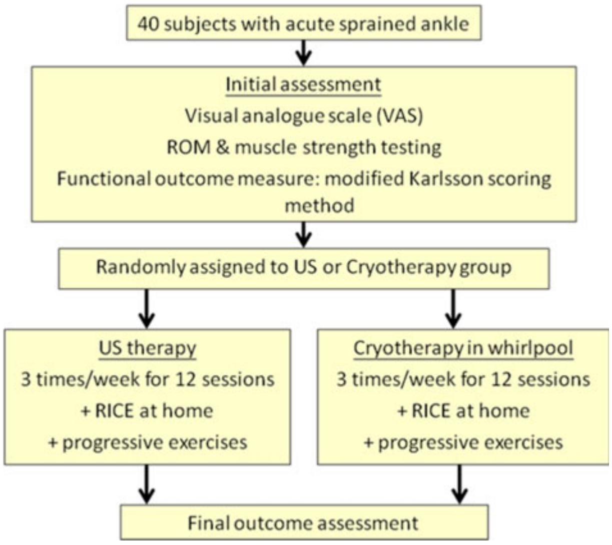
40 subjects with acute sprained ankle
Initial assessment
Visual analogue scale (VAS)
ROM & muscle strength testing
Functional outcome measure: modified Karlsson scoring method
Randomly assigned to US or Cryotherapy group
US therapy
3 times/week for 12 sessions
+ RICE at home
+ progressive exercises
Cryotherapy in whirlpool
3 times/week for 12 sessions
+ RICE at home
+ progressive exercises
Final outcome assessment

Figure 3 Experiment comparing the treatment effect of ultrasound therapy (US) and cryotherapy in whirlpool for patients with an acute sprained ankle. RICE stands for rest, ice, compression and elevation of leg above the heart level. In the frequentist approach, null hypothesis (H_o) is that there is no difference in treatment effect between US and cryotherapy when treating patients with a sprained ankle. Alternative hypothesis (H_a) is that there is a difference in treatment effect between US and cryotherapy in the treatment of sprained ankles. alpha=0.05.

Besides the aforementioned limitation, the *P*-value is highly vulnerable to artifacts as a result of sample size. Theoretically, even the smallest treatment difference can become statistically significant if the sample size is sufficiently large, as is the case in large scale clinical trials.[3,20]

The frequentist approach may also cause inflexibility on some research designs. In order to control the Type I error in the long run, sophisticated designs and calculations are required in studies such as sequential clinical trials. In sequential trials, interim analyses and multiple endpoint comparisons are performed. The result of each interim analysis will affect the subsequent progression of the studies. Since each interim analysis is associated with a risk of Type I error, the potential cumulative errors resulting from multiple interim inspections may inflate the overall false-positive findings in the research. Strict control over the study design and adjustment of nominal significance level for each interim test are needed prior to the experiment.[21] This restriction inevitably prohibits the amendment of study protocols when an unforeseeable situation occurs during the experimental period.[6] For example, it is impossible to eliminate part of the interim data collections even if there are tight financial constraints during the research period. Such inflexibility of the frequentist approach can be circumvented by the Bayesian approach, which does not require the calculation of the alpha-value.[3,22]

Although the frequentist approach is supposed to be an 'objective' statistical approach, the choice of alpha-value and interpretation of the *P*-value are implicitly subjective. The arbitrary choice of alpha-value has never been justified although it is generally accepted as 0.01 or 0.05.[23] Furthermore, the *P*-value may be used inconsistently for inference.[4,17] When an unexpectedly large *P*-value is obtained from a study, some investigators may justify their preconceived notions by imputing the insignificant result to small sample size or describing the findings as 'trend' or 'very likely to have effect'. However, if the *P*-value in a study is small, researchers may blindly accept the alternative hypothesis without the same critical thought, and rationalize their findings with new self created theories. Even though it may indicate a genuine new discovery, it may also imply an inconsistent interpretation of *P*-value.[4] To improve the interpretability of results, the inclusion of 95%CI in research reports has been suggested.[5]

The Bayesian Approach

An alternative to the frequentist approach is the Bayesian inference. Bayes' theorem (equation (1)) was first proposed 200 years ago by the Reverend Thomas Bayes.[3,6] It is a mathematical formula using 'prior' probability ('prior') obtained from previous studies and evidence from a current study to calculate the 'posterior' probability ('posterior') of different hypotheses.[6] It is a conditional probability that takes into account the 'prior' and observed probability of a particular hypothesis and its rival alternative hypotheses. Based on the external knowledge and evidence from available literature, investigators formulate the 'prior' of all hypotheses before the start of an experiment. On completion of the experiment, new supportive or non-supportive evidence for a particular hypothesis is obtained. This new evidence will update the 'prior' to derive the 'posterior' of a particular hypothesis

$$P(H_i|\mathrm{Data}) = \frac{P(\mathrm{Data}|H_i)P(H_i)}{\sum_{j=1}^{2} P(\mathrm{Data}|H_j)P(H_j)} (i=1, 2) \qquad (1)$$

In equation (1), H_i represents mutually exclusive rival hypotheses where $i=1$ and 2. Although i and j can be any number, they are limited to 2 for this example. $P(H_i)$ represents the 'prior' probabilities of two hypotheses. The probabilities, $P(\mathrm{Data}|H_i)$ ($i=1,2$), of observing supportive evidence (Data) for a particular H_i given H_i is true, are also known as the likelihoods of the sample data. The 'posterior' probabilities, $P(H_i|\mathrm{Data})$ ($i=1,2$), of two hypotheses being true are updated based on the relative weight of newly observed evidence (Data) and previous knowledge, $P(H_i)$. The denominator of the right-hand side of

'Posterior' Probability		Current Data ('Evidence')		Subjective 'Prior' Probability
Posterior (final) odds of null hypothesis is true	=	Bayes Factor	X	Prior odds of null hypothesis is true (from previous evidence)

$$\frac{P(H_0|\text{Data})}{P(H_a|\text{Data})} = \frac{P(\text{Data}|H_0)}{P(\text{Data}|H_a)} \times \frac{P(H_0)}{P(H_a)}$$

$$\text{where Bayes factor} = \frac{\text{Probability (Evidence given null hypothesis)}}{\text{Probability (Evidence given alternative hypothesis)}}$$

Figure 4 'Odds' form of Bayes' theorem. H_o is the null hypothesis, H_a is the alternative hypothesis. $P(H_o|Data)$ is the 'posterior' probability of H_o given the observed data. $P(H_a|Data)$ is the 'posterior' probability of H_a given the observed data. The ratio of these 'posterior' probabilities constitutes the 'posterior' odds ratio. $P(Data|H_o)$ is a conditional probability meaning probability of Data (empirical outcome) to be observed given that H_o is true. $P(Data|H_a)$ is a conditional probability meaning probability of Data to be observed given H_a. Bayes factor (BF) is a likelihood ratio between these two conditional probabilities. P (H_o) and P (H_a) represent the 'prior' probability of H_o and H_a respectively. If H_o is the complement of H_a, P (H_o) plus P (H_a) must be equal to 1.

equation (1) is simply a normalizing constant independent of *i*.[6]

Since the original Bayes' theorem requires the calculation of probabilities of all the possible hypotheses in the denominator, a likelihood ratio (also called Bayes factor, BF) is used as an alternative to simplify the calculation (Fig. 4). BF is the ratio of the conditional probability of observed data given H_o and the conditional probability of observed data given H_a. Similar to hypothesis testing in the frequentist approach, the Bayesian approach uses BF to compare two hypotheses at a time. It compares how well each hypothesis predicts the observed data of a given study.[6] The hypothesis that predicts data more accurately will be supported by more new evidence.[17] If more evidence supports H_o, the BF will increase and vice versa. The combination of BF and 'prior' modifies the previous 'belief' and leads to new 'posterior' for both hypotheses.[24,25]

To demonstrate the calculation of 'posterior' using the Bayesian approach, we revert to the hypothetical US and cryotherapy experiment. A successful treatment (μ) is defined as $\geqslant$80% recovery of a sprained ankle. From the total number of μ in each treatment group (within the first 4 weeks of treatment), the likelihood for μ to occur within each treatment group at a different period can be calculated. Table 1 shows the hypothetical probability distributions (likelihoods) of observed μ in two treatment groups. The likelihoods of μ in US and cryotherapy group at different weeks are denoted by discrete f(μ|US) and f(μ|Ice) respectively. These probability distributions show that most of μ are observed within the first 2 weeks of US treatment while most of μ in the cryotherapy group are found in the third and fourth week of treatment. However, the 'posterior' distributions of μ may be changed by the incorporation of 'prior'. Table 2 shows the 'posterior' for all values of μ under the influence of three sets of 'prior': (1) each treatment modality has the same 'prior' for μ (equal to 0.50), (2) cryotherapy is nine times more probable than US for μ, and (3) US is four times more probable than cryotherapy to obtain μ. If equal probability (0.50) is assigned to both treatments, the resulting 'posterior' distributions will be similar to that of the observed results. If the 'prior' distributions are similar to that of the observed data, the 'posterior' will further support the 'prior'. An example of

Table 1 A hypothetical observed probability distribution of successful treatment from patients with a sprained ankle receiving ultrasound therapy (US) and cryotherapy (Ice) over first 4 weeks

Successful treatment (μ) after	First week	Second week	Third week	Fourth week
f(μ|US)	0.72	0.18	0.07	0.03
f(μ|Ice)	0.10	0.006	0.47	0.424

Note: successful treatment is defined as $\geqslant$80% recovery for a sprained ankle. The probability distributions of successful treatment from US and cryotherapy group are defined by their discrete densities f(μ|US) and f(μ|Ice) respectively.

Table 2 Calculation of 'posterior' probability of sprained ankle treatment in ultrasound (US) & cryotherapy (Ice) group using Bayes' theorem based on three different 'prior' distributions x, y and z

'Prior' probability	'Posterior' probability	First week	Second week	Third week	Fourth week
P_x(US)=0.50	P_x(US|μ)	0.88	0.97	0.01	0.07
P_x(Ice)=0.50	P_x(Ice|μ)	0.12	0.03	0.99	0.93
P_y(US)=0.10	P_y(US|μ)	0.44	0.77	0.02	0.008
P_y(Ice)=0.90	P_y(Ice|μ)	0.56	0.23	0.98	0.992
P_z(US)=0.80	P_z(US|μ)	0.97	0.99	0.37	0.22
P_z(Ice)=0.20	P_z(Ice|μ)	0.03	0.01	0.63	0.78

Notes: P_x(US) & P_x(Ice); P_y(US) & P_y(Ice); and P_z (US) & P_z(Ice) are different sets of 'prior' probabilities distributions for the ultrasound therapy and cryotherapy; P_x(US|μ) & P_x(Ice|μ); P_y(US|μ) & P_y(Ice|μ); and P_z(US|μ) & P_z(Ice|μ) are the corresponding pairs of 'posterior' probabilities of ultrasound therapy and cryotherapy. The 'posterior' probabilities are calculated from Bayes' theorem by combining the empirical data in Table 1 and different set of 'prior' probabilities in Table 2.

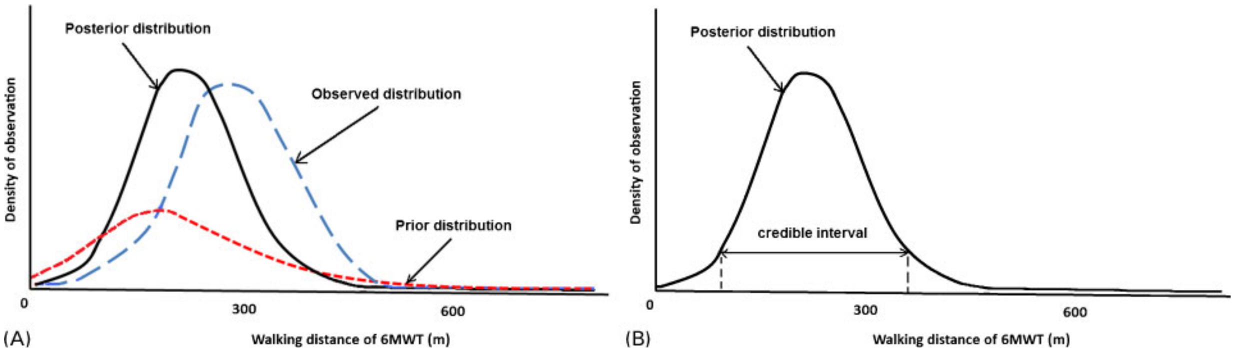

Figure 5 (A) A simulated 'prior', 'posterior' probability distribution and the distribution of observed walking distance in 6-minute walk test (6 MWT) of male post-cardiac surgery patients (⩾71 age with left ventricle ejection fraction ⩾50%) in a study investigating the walking distance of post-cardiac surgery patients before an in-hospital rehabilitation programme.[27] The simulated data are used because the original study did not use the Bayesian approach. The 'prior' distribution is modified by the observed data to obtain a 'posterior' distribution. (B) A simulated posterior probability distribution of the walking distance of 6 MWT of male post-cardiac surgery patients in a study investigating the walking distance of 6 MWT before an in-hospital rehabilitation programme.[27] A 95% credible interval is determined from the 'posterior' probability distribution.

this is when 'prior' favours US [P_z(US)=0.80; P_z(Ice)=0.20], 'posterior' further supports that US is more likely to have a successful treatment in the first week, than cryotherapy [P_z(US|μ)=0.97; P_z(Ice|μ)=0.03]. It implies that if a patient is successfully treated within the first week, there is a 97% chance that this patient received US given US and cryotherapy as the treatment options. Under another situation, 'prior' may wash out the effect of observed data. For instance, when the 'prior' strongly favours cryotherapy [P_y(US)=0.10; P_y(Ice)=0.90], the resulting 'posterior' will favour cryotherapy [P_y(US|μ)=0.44; P_y(Ice|μ)=0.56] although the distribution of observed data favours US [f(μ|US)=0.72; f(μ|Ice)=0.10]. The resulting 'posterior' makes the cryotherapy group more likely to obtain successful treatment in the first week. However, if the empirical evidence is strong, the existence of contradicting 'prior' may not change the conclusions. For example, the 'posterior' still favours cryotherapy in the fourth week [P_z(US|μ)=0.22; P_z(Ice|μ)=0.78] although 'prior' favours US [P_z(US)=0.80; P_z(Ice|μ)=0.20]. It implies that if a patient requires 4 weeks to be successfully treated, there is a 78% chance that this patient was treated by cryotherapy given the treatment option of either US or cryotherapy. Regardless of the effect of 'prior' on the observed data, the 'posterior' derived by Bayes' theorem can be used as 'prior' for future studies.

Although the previous example illustrates the calculation of 'posterior' for a discrete random variable (successful treatment), the Bayesian method can be used to infer the probability distributions of continuous population parameters (such as walking distance) based on that of the sample test statistics.[6] In the Bayesian approach, a continuous population parameter (theta) is considered to be a random variable that has different values and corresponding probabilities. The set of all possible unobserved values of theta is called the parameter space. The Bayesian approach assumes that our knowledge of the true value of theta can be expressed by a probability distribution over the parameter space.[6,26] The prior knowledge of a parameter is expressed as a 'prior' distribution of theta. Empirical data update the 'prior', and the resulting information of theta is described by its 'posterior' distribution.[6,26] To illustrate this concept graphically, the 6 MWT result from a study investigating the walking distance of post-cardiac surgery patients before an in-hospital rehabilitation programme is used.[27] Figure 5A shows the simulated probability distributions of walking distance of male post-cardiac surgery patients aged ⩾71 with left ventricle ejection fraction ⩾50%.[27] Since the original study did not use the Bayesian approach, we simulate the 'prior', 'posterior' distribution and the distribution of observed walking distance in this figure. The 'prior' distribution is modified by the observed data to obtain a 'posterior' distribution. Although the 'posterior' distribution constitutes the complete inferential statement about theta, sometimes a certain summary measure of this 'posterior' distribution may suffice. Therefore, the 95% credible interval, in which there is 95% probability that the true theta lies (Fig. 5B), is commonly used.[6] This credible interval is similar to the frequentist 95% CI except it reports in terms of probability. In general, the resulting 'posterior' can be reported as a single value with the highest posterior probability density (Bayesian point estimate) or as a more informative summary (Bayesian credible interval).[6,19] Table 3 shows the similarities and differences between the frequentist and the Bayesian approaches.

Given the key role of 'prior' in the Bayesian approach, it is essential to choose an appropriate 'prior'. In general, 'prior' or 'prior' distribution is chosen based on researchers' knowledge, experts' opinions and available literatures. Different mathematical methods and 'prior' distribution models have been proposed for situations encompassing prior ignorance, vague prior knowledge and substantial prior knowledge.[6,26,28] Given the complexity of this topic, readers are referred to appropriate statistics books for detail.[6,26,28]

Since the philosophy of the Bayesian approach is different from that of the frequentist approach, it allows calculation of the probability of the data under a true H_o. Goodman calculated different Bayes factors corresponding to different *P*-values found in the frequentist method (Table 4). He showed that if the *P*-value in an experiment calculated by the frequentist method was equal to 0.05, but the results from previous studies strongly supported H_o (75% 'prior'), the 'posterior' of a true H_o calculated from Bayes' theorem using BF=1/6.8 would be as high as 31%. The calculation is shown below

$$\frac{P(H_o|\text{Data})}{1-P(H_o|\text{Data})}=\frac{1}{6{\cdot}8}\times\frac{0{\cdot}75}{0{\cdot}25}=0{\cdot}45;$$

$$\text{where } 1-P(H_o|\text{Data})=P(H_a|\text{Data})$$

$$P(H_0|\text{Data})=\frac{0{\cdot}45}{1+0{\cdot}45}=0{\cdot}3103 \text{ (i.e. } 31{\cdot}03\%)$$

Characteristics of the Bayesian Approach

Therapeutic effect exploration

The Bayesian approach can be as effective as the frequentist approach in assessing the magnitude of therapeutic effect. It can precisely estimate all possible differences between two treatments by defining multiple hypotheses.[22] The strength of evidence for each hypothesis is proportional to the probability of observed data under that hypothesis in the form of BF. The idea of multiple comparisons can be once again explained by the US and cryotherapy experiment. Given that the treatment effects of two therapies may vary among individuals, the actual differences in treatment effect in the population may be distributed over a range. To identify the most probable treatment effect, researchers can establish multiple hypotheses ranging from those favouring cryotherapy to those favouring US (e.g. 0 to 100% difference). From the observed results, researchers can subsequently compare BFs of all hypotheses and identify the one that has the highest 'posterior'.

Table 3 Comparison between the frequentist and the Bayesian approach

	Frequentist	Bayesian
'Prior' probability determination	Unnecessary	Essential
'Posterior' probability	Cannot be calculated	Official outcome
Sample size	Predetermined	Unrestricted
Hypothesis testing	Predetermined	Unrestricted
Truth of null hypothesis	Assumed	Not assumed
Point estimate	*P*-value	'Posterior' probability
Interval estimate	Confidence interval	Credible interval
Alpha and beta value	Predetermined	Unnecessary
Interim analysis	Predetermined	Unrestricted

Table 4 Relation between two-sided, fixed sample size *P*-values under Gaussian distribution and corresponding Bayes factor and the effect of such evidence on the probability of the null hypothesis (H_o)

Frequentist statistics	Current sample results	Bayesian statistics	
		Change in probability of H_o	
P-value (Z-score)	Corresponding Bayes factor (strength of evidence against H_o)	Prior probability of H_o	Posterior probability of H_o
0.10 (1.64)	1/3.8 (weak)	75 (strong support)	44
		50	21
		17 (weak support)	5
0.05 (1.96)	1/6.8 (moderate)	75 (strong support)	31
		50	13
		26 (weak support)	5
0.01 (2.58)	1/28 (moderate to strong)	75 (strong support)	10
		60	5
		50 (neutral support)	3.5

Notes: This table shows the relation between the *P*-values (from frequentist approach), Bayes factor (from empirical data using the Bayesian approach) and different 'posterior' probabilities of true H_o (by the Bayesian approach). For example, when *P*=0.05, corresponding Bayes factor=1/6.8 and the 'prior' probability for H_o calculated from previous studies is 75%, the 'posterior' probability of a true H_0 calculated by the Bayesian approach is still as high as 31%. (Table is modified from Goodman[17] article with permission).

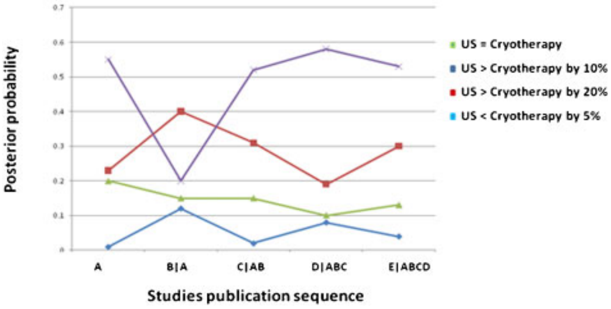

Figure 6 A simulated Bayesian meta-analysis of ultrasound therapy against cryotherapy in treating sprain ankles. A to E represents five journal articles that were published chronologically. A: first trial without 'prior' information, B|A: B result given A result; C|AB: C result given A and B results, D|ABC: D result given ABC results; E|ABCD: E result given ABCD results. The y-axis represents the 'posterior' probability of different hypotheses based on different study results. Different hypotheses: same ultrasound and cryotherapy treatment effect (US=Cryotherapy), treatment effect of ultrasound is better than cryotherapy by 10% (US>Cryotherapy by 10%), treatment effect of ultrasound is better than cryotherapy by 20% (US>Cryotherapy by 20%), treatment effect of ultrasound is inferior to cryotherapy by 5% (US<Cryotherapy by 5%).

Meta-analysis

The Bayesian model is well suited to meta-analysis because it provides a clear framework to analyze information from different sources.[19,29] For example, a researcher investigating the effect of US and cryotherapy in treating a sprained ankle by Bayesian meta-analysis would attribute equal probability weight to the 'prior' of each hypothesis in the first (A) of five chronologically published studies given there was no similar study prior to A (Fig. 6). The 'posterior' of each study would then be used as the 'prior' for the subsequent trial. The 'posterior' obtained from the second trial, B, may be quite different from that of A. However, as more evidence is accumulated from prior studies, the relative probability of each hypothesis becomes more distinct and constant. Diamond and Kaul used the Bayesian and frequentist meta-analysis approach to compare the effect of conservative and aggressive cardiovascular intervention for uncomplicated post-infarction patients.[3] Based on the same set of data from five randomized controlled trials involving 9000 patients, both methods found statistically significant reduction of death by aggressive intervention. Although both the frequentist meta-analysis and the Bayesian meta-analysis can assess the magnitude of therapeutic response from pooled data, the Bayesian approach provides a direct probability interpretation which may be preferred by clinicians for determining the effectiveness of treatments and making clinical judgment.

Sequential clinical trials

As discussed in a previous section regarding sequential clinical trials using the frequentist approach, adjustment of the alpha-value in interim analysis is essential for confining the overall Type I error within the preset limit.[30] Interim analyses will not affect the BF of Bayesian analysis. If there is strong evidence against H_o (small BF), the BF will remain robust, irrespective of the number of interim analyses.[31] This is because BF depends on the probability of observed data alone; the termination of an experiment will not affect BF. Furthermore, the inflation of probability of Type I error by performing multiple tests is not important in the Bayesian approach because its philosophy is unrelated to Type I error.[3] Hence, this approach can provide accurate results without jeopardizing the flexibility of sequential clinical trials.[32] This property enables frequent interim analyses and timely termination of trials if accumulative results significantly support a superior treatment or disprove an ineffective treatment. Lewis and Berry compared the frequentist and the Bayesian approach in a theoretic clinical sequential trial. They concluded that the Bayesian approach could reduce both the sample size and cost of experiment without affecting the credibility of the results.[32] Given these advantages, the Bayesian approach has been used in different pharmaceutical sequential trials.[33,34]

Standardized interpretation of Bayes factors and conclusion

The interpretation of 'posterior' and BF is straightforward and standardized. The results of different hypotheses are expressed in terms of probability. The explicit definition of hypotheses also enhances the specificity of each hypothesis and facilitates readers to understand the relative strength of each hypothesis directly.[22] Although the Bayesian approach cannot guarantee that each researcher will derive the same 'prior' from the same previous data, it ensures that they will have the same conclusion if they choose the same magnitude of 'prior'.[3]

Prediction

Since the Bayesian approach calculates the 'posterior' for each hypothesis, it provides valuable predictive probability for a future event. This property is beneficial to clinicians. For example, a clinician can anticipate the recovery rate of a patient with a sprained ankle given that the pain score of that patient has dropped by 20% in the last two weeks. Hence, the clinician can not only estimate the discharge time more accurately, but also make timely adjustments to the treatment regime.

Limitations of the Bayesian Approach

Despite the advantages of the Bayesian approach, a few factors have limited its use. Firstly, the Bayesian approach requires researchers to apply sophisticated mathematical calculations for most situations, so they may need assistance from biostatisticians. The formula shown in Fig. 4 is simplified to facilitate understanding of the basic Bayesian concept. The complexity of calculation increases as the number of outcome variables increases.

With the advancement of computer technology, this limitation has been overcome. Researchers can now perform complex Bayesian analysis by using Bayesian inference software available on the internet.[35] Well-developed databases and search engines also facilitate searches of relevant articles and books for 'prior' determination.

A second factor limiting the popularity of the Bayesian approach is the requirement of explicit 'prior' determination. There is no consensus regarding the method of determination for 'prior' or 'prior' distribution.[6] Theoretically, 'prior' should be derived from results of meta-analyses including all available randomized clinical trials. However, since most research studies aim to explore new findings or relationships, it is difficult to find comparable previous studies. Even if there are related prior studies, their quality will affect the results of 'prior' determination. In addition, different 'posterior' could be derived from the same set of data due to the disparities in researchers' 'beliefs'. By virtue of the subjective nature of 'prior' justification, the Bayesian method is less preferred by many researchers.[4,25]

To minimize the contention regarding subjectivity and biases of researchers toward this issue, Bayesian researchers can perform a sensitivity analysis in which a range of 'prior' distributions can be assigned to H_o or other hypotheses based on evidence in the literature. For example, the 'prior' for H_o can range from the most supportive (e.g. 99% 'prior' to support H_o) to the most skeptical (e.g. 0% 'prior' to support H_o). The corresponding 'posterior' probabilities derived from each of the 'priors' can be tabulated and explained thoroughly to improve the interpretability of findings by allowing readers to draw their own (possibly different) conclusions.

In the case of insufficient prior information, 'non-informative prior' or 'objective prior' can be adopted.[6] This means that no supporting evidence is given to any of the proposed hypotheses. In an experiment with only two complementary hypotheses, 50% 'prior' is assigned to each hypothesis based on this principle.[36] Although the idea of 'objective prior' appears to be applicable, its usage remains controversial even among Bayesian statisticians.[37] If 'objective prior' is applied, the conclusions from the Bayesian method will not be very distinct from that of the frequentist approach. Another solution to address this 'prior' ignorance is to increase the sample size of the current study so that the influence of current findings can outreach the relative contribution of 'prior'. With the increase in sample size, the power of the statistical test will increase. The result will be a better reflection of truth rather than subjective 'belief'.

Suggestions for Future Research Statistics

Given the inherent limitations of the frequentist approach, future studies should report results in terms of *P*-value, CI and effect size. This can provide more information for readers to determine the clinical implications of study results. Moreover, the study findings should be compared with similar studies in the discussion section to improve the credibility of the results.[16]

The Bayesian approach can be used as an alternative or adjunct to the frequentist method in future scientific and clinical studies. Since the Bayesian approach unites external evidence from previous studies with those from a current experiment, it minimizes the risk of drawing wrong conclusions from a single study and improves the overall strengths of the conclusions. The Bayesian approach is an ideal analysis for sequential clinical trials as BFs are unaffected by interim inspections. However, biostatisticians may have to be consulted for some sophisticated Bayesian analysis. Researchers should also justify their choice of 'prior' as well as acknowledge differences between reference studies and the current study.

Regardless of types of statistical approach, one should be aware that a statistically significant result is not equivalent to clinical significance.[38] In order to discern between these two significances, readers should interpret the reported results in research papers with other statistical parameters, such as *P*-value, CI, effect size, posterior probabilities or credible interval. Researchers should also report the clinical significance of results in clinical trials or perform hypothesis testing based on the minimal clinically important difference, which represents the smallest change in outcome measures that signifies a meaningful change in an individual patient's symptom.[39]

Conclusions

As evidence-based practice becomes commonplace in clinical settings, clinicians will more frequently incorporate research findings into their practice as well as critically interpret the results from research papers. As the Bayesian approach may become more common in future research studies, a good understanding of this method will benefit clinicians in interpreting research papers and planning their future clinical studies.

Acknowledgements

The authors sincerely thank Monica Moreno, PhD candidate, at the Department of Mathematical and Statistical Sciences in University of Alberta, and Dr

Susan Armijo Olivo, at the Department of Physical Therapy in University of Alberta for their constructive advice.

References

1 Cooke J, Nancarrow S, Dyas J, Williams M. An evaluation of the 'Designated Research Team' approach to building research capacity in primary care. BMC Fam Pract 2008;**9**:37.

2* Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. 2nd ed. Washington DC: Chapman & Hall/CRC; 2004.

3* Diamond GA, Kaul S. Prior convictions Bayesian approaches to the analysis and interpretation of clinical megatrials. J Am Coll Cardiol 2004;**43**:1929–39.

4* Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. Ann Intern Med 1999;**130**:995–1004.

5* Gravetter FJ, Wallnau LB. Statistics for the behavioral sciences. 8th ed. Belmont: Wadsworth; 2009.

6* Barnett V. Comparative statistical inference. 3rd ed. Chichester: John Wiley & Sons Ltd; 1999.

7 Sawka AM, Boulos P, Beattie K, Papaioannou A, Gafni A, Cranney A, *et al.* Hip protectors decrease hip fracture risk in elderly nursing home residents: a Bayesian meta-analysis. J Clin Epidemiol 2007;**60**:336–44.

8 Resnik L, Feng Z, Hart DL. State regulation and the delivery of physical therapy services. Health Serv Res 2006;**41**:1296–316.

9 Bekele BN, Ji Y, Shen Y, Thall PF. Monitoring late-onset toxicities in phase I trials using predicted risks. Biostatistics 2008;**9**:442–57.

10 Zaretzki RL, Gilchrist M, Briggs WM, Armagan A. Bias correction and bayesian analysis of aggregate counts in sage libraries. BMC Bioinformatics 2010;**11**:72.

11 Trieu HT, Nguyen HT, Willey K. Shared control strategies for obstacle avoidance tasks in an intelligent wheelchair. Proceedings of the 30th Annual International IEEE EMBS Conference; 2008 Aug 20–25; Vancouver, Canada. IEE Xplore; 2008.

12 Appendix: information for authors. Ann Intern Med 2002;**136**:A1–5.

13 Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, *et al.* the Evidence-Based Medicine Working Group. User's guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. JAMA 2000;**284**:1290–6.

14 Hacking I. The logic of statistical inference. Cambridge: University Press; 1965.

15 Christensen R. Testing fisher, neyman, pearson, and bayes. Am Stat 2005;**59**:121–6.

16 Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 1933;**231**:289–337.

17* Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. Ann Intern Med 1999;**130**:1005–13.

18 Yuen HK, Sword D. Home-based exercise to alleviate fatigue and improve functional capacity among breast cancer survivors. J Allied Health 2007;**36**:e257–75.

19 Julious SA, Tan SB, Machin D. An introduction to statistics in early phrase trials. Chichester: John Wiley & Sons Ltd; 2010.

20 DeMets DL, Califf RM. Lessons learned from recent cardiovascular clinical trials: part II. Circulation 2002;**106**:880–6.

21* Pocock SJ. Clinical trials: a practical approach. New York: Wiley; 1984.

22* Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. Clin Trials 2005;**2**:282–90.

23 Fisher RA. Statistical methods for research workers. 13th ed. New York: Hafner; 1958.

24 Jennison C, Turnbull BW. Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. Technometrics 1983;**25**:49–58.

25 Carpenter J, Gajewski B, Teel C, Aaronson LS. Bayesian data analysis: estimating the efficacy of t'ai chi as a case study. Nurs Res 2008;**57**:214–9.

26 Bolstad WM. Introduction to Bayesian statistics. 2nd ed. New York: John Wiley & Son Inc; 2007.

27 Opasich C, De Feo S, Pinna GD, Furgi G, Pedretti R, Scrutinio D, *et al.* Distance walked in the 6-minute test soon after cardiac surgery toward an efficient use in the individual patient. Chest 2004;**126**:1796–801.

28* Press SJ. Subjective and objective bayesian statistics: principles, models, and applications. 2nd ed. Hoboken: NJ: John Wiley and Sons Inc; 2002.

29 Salpeter SR, Cheng J, Thabane L, Buckley NS, Salpeter EE. Bayesian meta-analysis of hormone therapy and mortality in younger postmenopausal women. Am J Med 2009;**122**:1016–22.e1.

30 Pocock SJ. Group-sequential methods in the design and analysis of clinical trials. Biometrika 1977;**64**:191–9.

31 Blume JD. Likelihood methods for measuring statistical evidence. Stat Med 2002;**21**:2563–99.

32 Lewis RJ, Berry DA. Group sequential clinical trials: a classical evaluation of bayesian decision-theoretic designs. J Am Stat Assoc 1994;**89**:1528–34.

33 Morita S, Baba H, Tsuburaya A, Takiuchi H, Matsui T, Maehara Y, *et al.* A randomized phase II selection trial in patients with advanced/recurrent gastric cancer: trial for advanced stomach cancer (TASC). Jpn J Clin Oncol 2007;**37**:469–72.

34 Tinmouth A, Tannock IF, Crump M, Tomlinson G, Brandwein J, Minden M, *et al.* Low-dose prophylactic platelet transfusions in recipients of an autologous peripheral blood progenitor cell transplant and patients with acute leukemia: a randomized controlled trial with a sequential bayesian design. Transfusion 2004;**44**:1711–9.

35 The BUGS Project [Internet]. Cambridge: MRC Biostatistics Unit; c1996–2008 [cited 2010 Apr 21]. Available from: http://www.mrc-bsu.cam.ac.uk/bugs

36* Neyman J. Two breakthroughs in the theory of statistical decision making. Rev Int Stat Inst 1962;**30**:11–27.

37 Vaurio J. Objective prior distributions and Bayesian updating. Reliab Eng Syst Safe 1992;**35**:55–9.

38 Musselman KE. Clinical significance testing in rehabilitation research: what, why, and how? Phys Ther Rev 2007;**12**:287–96.

39 Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease specific quality of life questionnaire. J Clin Epidemiol 1994;**47**:81–7.