Osteoporosis and Sarcopenia 6 (2020) 161–167

Contents lists available at ScienceDirect

Osteoporosis and Sarcopenia

journal homepage: http://www.elsevier.com/locate/afos

Review article

Common methodological issues and suggested solutions in bone research

Tuan V. Nguyen

Garvan Institute of Medical Research, St Vincent's Clinical School, UNSW Medicine, UNSW Sydney, School of Biomedical Engineering, University of Technology Sydney, 384 Victoria Street, Darlinghurst, NSW, 2010, Australia

ARTICLE INFO

Article history:
Received 26 July 2020
Received in revised form
12 November 2020
Accepted 19 November 2020
Available online 28 November 2020

Keywords:
Statistical methods
P-value
Confidence interval
Bayesian inference
Collider bias

ABSTRACT

Bone research is a dynamic area of scientific investigation that usually encompasses multidisciplines. Virtually all basic cellular research, clinical research and epidemiologic research rely on statistical concepts and methodology for inference. This paper discusses common issues and suggested solutions concerning the application of statistical thinking in bone research, particularly in clinical and epidemiological investigations. The issues are sample size estimation, biases and confounders, analysis of longitudinal data, categorization of continuous data, selection of significant variables, over-fitting, P-values, false positive finding, confidence interval, and Bayesian inference. It is hoped that by adopting the suggested measures the scientific quality of bone research can improve.

1. Introduction

Bone research commonly involves multifaceted studies. These studies may range from basic cellular experiments, clinical trials to epidemiological investigations. Most of these studies come down to 3 broad aims: assessing difference (ie, effect), association, and prediction. Do cells with one version of a gene express more of an enzyme than cells with another version? Does a new drug reduce the risk of fracture compared with placebo? Among hundreds of risk factors in a cohort study, which factors are associated with fracture? Can a new prediction model based on Caucasian populations be used for fracture risk assessment in Asian populations? The answer to these questions invariably involves statistical thinking.

Indeed, every stage of a research project – from study design, data collection, data analysis, to data reporting – involves statistical consideration. Statistical models and null hypothesis significance testing are powerful methods to discover laws and trends underlying observational data, and to help make accurate inference. Test of hypothesis can also help researchers to make decision of accepting or rejecting a null hypothesis, contributing to the scientific progress. Thus, reviewers and readers alike expect researchers to apply appropriate statistical models to obtain useful information from the data for creating new knowledge.

However, misuse of statistical methods has been common in biomedical research [1], and the problem is still persistent [2,3]. In the 1960s, a review of 149 studies from popular medical journals revealed that less than 30% of studies were methodologically 'acceptable' [4]. About 2 decades later, a review of 196 clinical trials on rheumatoid arthritis found that 76% of the conclusions or abstracts contained 'doubtful or invalid statements' [5]. In a recent systematic review of published studies in orthopedic journals, 17% of studies where conclusions were not consistent with results presented, and 39% of studies where a different analytical method should have been applied [6]. While the majority of statistical errors were minor, about 17% errors could compromise the study conclusion [6]. Apart from errors, there are deep concerns about the abuse of statistical methods that lead to misinterpretation of data and retraction of published studies. The bone research community has recently come to terms with a high profile retraction of papers by a bone researcher [7]. The misuse of statistical methods and misinterpretation of statistical analysis partly contribute to the problem of irreproducibility of research findings [8,9].

The recognition of the lack of reproducibility in biomedical research [10–12] has led to several discussions on how to improve the quality of bone research publications [13–15]. As an editor and

E-mail address: t.nguyen@garvan.org.au.
Peer review under responsibility of The Korean Society of Osteoporosis.

https://doi.org/10.1016/j.afos.2020.11.008

expert reviewer for several bone and medical journals over the past 25 years, I have identified major areas that need improvement, namely, reporting of study design, data analysis, and interpretation of P-values. In this article, I focus on the most common issues that appear repeatedly in the bone research literature, and then suggest possible solutions. My aim is to help bone research colleagues in providing relevant ideas and methods that are required to improve the reproducibility and accurate inference of their work.

1.1. Sample size

The founder of modern statistics, Karl Pearson, once said that "*the utility of all science consists alone in its method, not its material*" [16]. Although the same method can be used in different studies, it is the details of methodological activities that define the quality of the work. The description of details and activities of study design can be found in several guidelines such as CONSORT [17] for clinical trials, STROBE [18] for observational studies, and ARRIVE [19] for animal studies.

One important point of these guidelines is the description of sample size estimation. As a norm, studies with inadequate sample size have low sensitivity (eg, power) to uncover a true association. It is not widely appreciated that underpowered studies often produce statistically significant and exaggerated findings, but the findings have low probability of reproducibility [20].

Therefore, a clear explanation of sample size estimation and rationale, including primary outcome, expected effect size, type I and type II error, greatly help readers to assess the reliability of study findings [21]. Unfortunately, many bone science authors do not report how they arrived at the sample size. Moreover, most laboratory studies are based on a small number of animals, but there is no quantitative justification of the sample size [22]. As a result, it is very difficult to interpret a study's observed effect size in the absence of a hypothesized effect size that underlined the estimation of sample size.

1.2. Biases and confounders

In uncontrolled and non-randomized studies, the association between exposure and outcome can be misled by biases and confounders. The list of biases and confounders are extensive [23], and these biases are almost always present in uncontrolled studies. Among the list of biases, *selection bias* is a major threat. Selection bias can arise in studies where participants were drawn from a sample that is very different from the general population, and as a result, it may distort the true association between exposure and outcome. The diagram below (Fig. 1) shows a hypothetical association between an exposure and an outcome in a population with a correlation coefficient being $r = -0.29$ ($P < 0.0001$; left panel); however, if a subset of the population was selected for analysis (right panel) then the association is no longer statistically significant ($r = -0.05$; $P = 0.72$). Thus, studies in subgroup of patients or non-representative samples have a high risk of reaching a wrong conclusion.

Confounding is a common threat to the validity of conclusions from observational studies. A confounder is defined as a variable that causes or influences both the exposure and outcome (Fig. 2, left panel). For instance, an association between low levels of physical activity and bone mineral density could be confounded by advancing age (i.e., a confounder). In osteoporosis research, confounding variables such as age, gender, comorbidities, and frailty could account for the observed association between bisphosphonates and mortality in observational studies [24].

Collider bias [25] is another threat to the validity of observational studies. A variable is considered a 'collider' if it is caused by both the exposure and the outcome. It should be noted that collider is different from confounder, which is defined as a variable that is the cause of both exposure and outcome (Fig. 2, right panel). For example, both fracture (outcome) and respiratory failure (exposure) can cause patients to be hospitalized, and in this case, hospitalization is the potential collider. The effect of collider bias is nicely illustrated by the spurious association between single nucleotide polymorphisms (SNP) and sex [26]. In this analysis, none of the 694 SNPs for height, as expected, was associated with sex (ie, the outcome) in a bivariate analysis; however, when height (ie, the collider) was added to the model, 222 SNPs were significantly associated to sex [26]. This example highlights that in association analysis, adjusting for factor that is causally related to the outcome can yield biologically meaningless but statistically significant association.

Regression-based adjustment is a powerful method to adjust for the effect of confounding variables, and help the inference to be more accurate. However, regression adjustment for a collider can yield a spurious association between exposure and outcome [26]. Some researchers have the tendency to adjust for all variables available with the intention to obtain the most unbiased association. For instance, some authors used weight, height, body mass index, and age in a regression model. Such an agnostic approach of adjustment may be counterproductive, because it runs the risk of over-adjustment and over-fitting, not to mention the problem of multicollinearity (ie, correlation among predictor variables). Not all associations require regression adjustment, and appropriate adjustment requires a careful consideration based on substantive knowledge. For instance, adjustment is not necessary for a covariate that does not induce the causal relationship between exposure and outcome [27].

1.3. Longitudinal data

In prospective cohort studies, individuals are repeatedly measured over time, enable the examination of individual evolution of outcome. The analysis of data from this type of study design is challenging, because (i) measurements within an individual are correlated, (ii) the duration between visits is different between individuals, and (iii) there are missing data. Some authors applied the analysis of variance to analyze such a longitudinal dataset, but this method cannot handle the difference in follow-up duration and missing data. If the within-subject correlation is not properly accounted for, it can lead to false positive findings and wrong confidence intervals [28]. Researchers are suggested to consider more modern methods such as generalized estimating equations [29] and the linear mixed effects model [30]. A major strength of these modern methods is that they can handle missing data while still accounting for variability within and between individuals.

Another common problem associated with longitudinal data analysis is the determination of rate of change for an individual. For studies that measure bone mineral density (BMD) before (denoted by x_0) and after (x_1) intervention, most researchers would calculate the percentage change as the difference between 2 measurements over the baseline measurement, ie, $(x_1 - x_0)/x_0 \times 100$, and then use the percentage change as a dependent variable for further analyses. Although this measure seems straightforward, it is not symmetric [31] and can result in misleading results [32]. A better and symmetric quantification of change should use the mean of 2 measurements as the denominator, ie, $(x_1 - x_0)/mean(x_0, x_1) \times 100$. For testing hypothesis concerning difference between treatments in before-after studies that involves a continuous outcome variable, the analysis of covariance is considered a standard method [33].

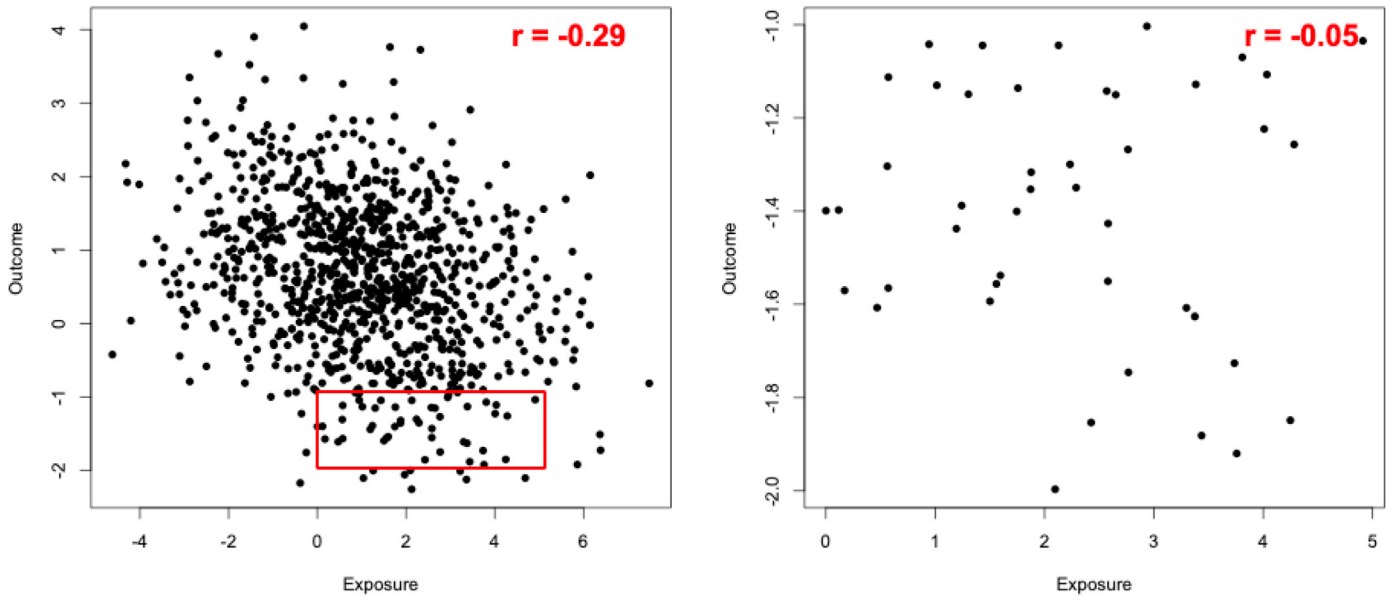

Fig. 1. Illustration of selection bias. There was a significant association between exposure and outcome in the population (left panel), but if a subset of individuals in red box were selected from the population, the association can be statistically non-significant (right panel).

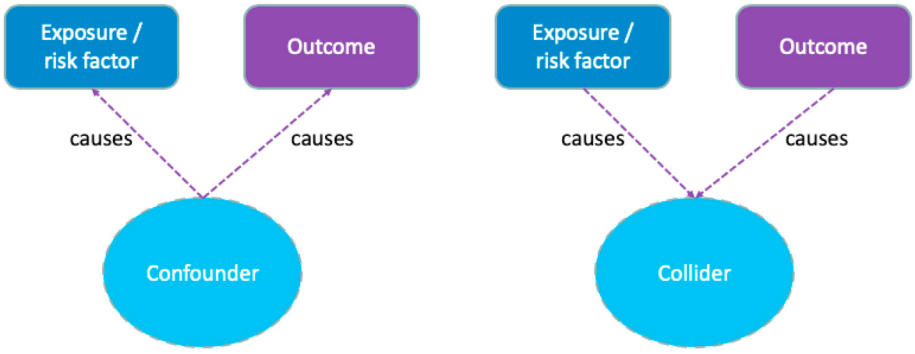

Fig. 2. Illustration of confounding variable and collider variable. A confounder is a variable that causes both exposure and outcome variables. A collider is a variable that is caused by both exposure and outcome variables. Regression model can be used to adjust for the effect of confounder, but it should not be used to adjust for the effect of collider.

1.4. Categorization of continuous variable

It is not uncommon to read bone research papers where the authors categorize continuous variables such as bone mineral density (BMD) into 2 distinct groups (eg, "osteporosis" and "non-osteoporosis"), or 3 groups (eg, osteoporosis, osteopenia, and normal), and then use the categorized variable as an exposure or an outcome for further analyses. While the World Health Organization's recommended BMD classification [34] is appropriate for *clinical/diagnostic* purposes, it is a bad practice for *scientific* purpose.

It has been repeatedly shown that such a categorization is unnecessary and can distort an association [35]. Apart from the risk of misclassification, the obvious problem with categorization of continuous variables is the loss of information. In the case of dichotomization, for example, all individuals above or below the cut-point is treated equaly, yet their prognosis could be vastly different. Therefore, the loss of information is increased (ie, more severe) when the number of categories is reduced. Categorization also reduces the efficiency of adjustment for confounders. In linear models, a categorized risk factor removes only 67% of the confounder compared to when the continuous type of the variable is used [36].

For scientific purposes, it is recommended that investigators do not categorize continuous variables in an analysis of association. Some continuous variables may exhibit a non-normal distribution, and in this case, it is instructive to consider more appropriate analyses such as spline regression or non-parametric smoother, and not to categorize continuous data.

1.5. Selection of 'significant' variables

In many studies, the aim is to identify a set of predictor variables that are independently associated with a continuous outcome (in multiple linear regression) or a binary outcome (in multiple logistic regression). In the presence of hundreds or thousands of variables of interest, the number of possible sets of variables (or models) can be very large. For instance, a study with 30 variables can generate at least $2 \wedge 30 = 1{,}073{,}741{,}824$ possible models, and determining which models are associated with an outcome is quite a challenge.

Many researchers have traditionally used stepwise regression to select the 'best model'. While stepwise regression is a popular method for selecting a relevant set of variables, it has serious deficiencies [37]. It is not widely appreciated that stepwise regression does not necessarily come up with the best model if there are

redundant predictors. Consequently, variables that are truly associated with the outcome may not be identified by stepwise regression, because they do not reach statistical significance, while non-associated variables may be identified to be significant [38]. As a result, the model identified by stepwise regression is poorly reproducible.

For selection of relevant predictors, investigators are strongly suggested to consider more robust methods such as Bayesian model averaging [39,40] or LASSO [41] which has been shown to perform better than the stepwise regression. Still, the models identified by these methods are only suggestive in nature. Statistical algorithms do not have substantive knowledge about a clinical or biological problem that we researchers have. Therefore, the best models must be guided by substantive knowledge, not just by statistical method-driven model selection.

1.6. Over-fitting

Multivariable statistical model always runs the risk of being over-fitted, in the sense that the model is unnecessarily complex. When over-fitting happens, the model is not valid because it tries to explain the random part of the model rather than the association between variables. As a result, an over-fitting model may fit the data very well for a dataset at hand, but it fits poorly for a new and independent dataset.

Over-fitting often happens when the number of parameters in the model is greater than the number of events. There is a rule of thumb that each predictor in a multivariable model requires at least 10 events [42], but recent research has shown that this rule of thumb is simplistic. Theoretical studies show that the number of events in a multivariable prediction model is determined by (i) the incidence of disease, (ii) the number of risk factors, (iii) the proportion of variance explained, and (iv) shrinkage factor [43].

Modern methods such as LASSO [41] or ridge regression [44] can help reduce over-fitting. In particular, LASSO is a method that shrinks the model coefficients toward 0 by imposing a constraint on the sum of the parameter estimates. This imposition can help eliminate non-important predictors in the model, and hence reduce the over-fitting.

1.7. P-values

Much of scientific inference boils down to the interpretation of P-value. Since its inception in the 1920s, P-value has been ubiquitous in the scientific literature, such that it is sometimes considered a "passport for publication". Readers of biomedical research literature may have noticed that the interpretation of P-value in most papers was largely dichotomized into "significant" vs "non-significant", with $P = 0.05$ being the commonest threshold for declaring a discovery. In some not-so-extreme cases, researchers reach a conclusion of effect based on a finding with $P = 0.04$, but readily dismiss a result with $P = 0.06$ as a null effect. However, it is not widely appreciated that that P-values vary greatly between samples [45], such that a deletion or addition of a single observation can change the statistical significance of a finding. Therefore, the simple classification of finding into "significant" and "non-significant" based on the threshold of 0.05 is not encouraged. The conclusion of an effect should be based on full evidence, not limited to the levels of statistical significance alone.

P-value is a result of null hypothesis significant testing (NHST). However, few practicing scientists realize that NHST is the hybridization of 2 approaches: test of significance and test of hypothesis. This hybridization has generated a lot of confusion and misinterpretation of P-values. It is thus instructive to have a brief review of the thinking underlying the NHST approach.

In the paradigm of *significance testing*, a null hypothesis is proposed, then a test statistic (eg, t-test, chi-squared test) is computed from the observed data. An index, called P-value, representing the deviation between the test statistic and the null hypothesis is derived, with lower values being a signal of the degree of implausibility of the null hypothesis. The proponent of this significance testing approach, Sir Ronald Fisher, suggested that a finding with P-value of 0.05 or lower is considered statistically significant. In his own words: "*The value for which $P = 0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not*" [46] Fisher suggests that researchers should report exact P-values (eg, $P = 0.031$, not $P < 0.04$).

In the paradigm of *hypothesis testing*, a null hypothesis and an alternative hypothesis are proposed to assess 2 mutually exclusive theories about the population of interest. Two long-term rates of erroneous decisions are then defined prior to conducting data collection: (i) the probability of a false positive finding that will be made when the null hypothesis is true (also referred to as type I error or α); and (ii) the probability of a false negative finding that will be made when the null hypothesis is false (ie, type II error or β). Traditionally, researchers set $\alpha = 5\%$ and $\beta = 20\%$ in most studies. After the data have been collected and distilled into a test statistic, the test result is then compared with a theoretical cut-off value associated with type I error. If the test result is smaller than the cut-off value, then the null hypothesis is accepted; otherwise the null hypothesis is rejected. The hypothesis testing approach, developed by Jerzy Neyman and Egon Pearson in the 1930s, was designed so that *"in the long run of experience, we shall not be too often wrong"* [47]

NHST is the marriage between Fisher's significance testing and Neyman-Pearson's hypothesis testing approaches [48]. In NHST, P-value is compared with type I error rate α to reject (when $P \leq \alpha$) or accept (when $P > \alpha$) the null hypothesis. As can be seen, this is actually a mis-marriage of 2 different approaches, because the P-value from significance testing is a local measure of evidence for a specific study, but the type I error and type II error from hypothesis testing are global measures from independent studies taken as a totality.

This mis-marriage has generated to a lot of misconceptions of P-values [49,50]. Most researchers interpret P-value as the probability of null hypothesis (eg, no effect, no association), and consequently 1 minus P-value is implicitly viewed as the probability that the alternative hypothesis (eg, presence of effect, association) is true; however, such an unconditional interpretation is wrong. Actually, P-value is the probability of obtaining results as extreme as the observed results *when* the null hypothesis is true – it is a conditional probability. Thus, if an effect size with $P = 0.06$, it means that when the null hypothesis is true, a value of the effect size as or more extreme than what was observed occurs in 6% of all samples; it does *not* mean that the null hypothesis is true in 6% of all samples. In other words, the effect size observed, or smaller, occurs in $1 - P = 94\%$ of all samples under the assumption that the null hypothesis of no effect is true.

Because the P-value threshold of 0.05 is traditionally considered 'statistically significant', and statistical significance is associated with a greater chance of publication, some researchers have involved in questionable research practices such as "P-hacking" [51]. P-hacking is a practice of data manipulation in conscious or subconscious way that produces a desired P-value. These include multiple subgroup analyses of an outcome, categorization of continuous data, data transformation, and selection of statistical tests. By manipulating data in such ways, an absolutely negative data can produce a statistically significant result in 61% of the time [51].

1.8. Multiple testing, large sample size, and false discovery rate

In recent years, national registries have provided researchers with opportunities to test hundreds or thousands of hypotheses, with many more tests being unreported. As a norm, the more one searches, the more one discovers unexpected and false findings. It can be shown that the probability of false positive findings is an exponential function of the number of hypothesis tests. For instance, at the alpha level of 5%, a study testing for association between 50 risk factors and an outcome, there is a 92% probability that the study will find at least one 'significant' association, even if there is no association between any of the risk factors and the outcome. In genomic research, the P-value threshold of 5×10^{-8} has become a standard for common-variant genome wide association studies, but there is no such threshold for registry-based research. Researchers using registry based data are suggested to take measures (such as Bonferroni's procedure or Tukey's test) to adjust P values from multiple testing so that the nominal P-value is less than 0.05, and to report the false discovery rate [52].

Studies with very large sample size pose serious challenges in the inference of association. For a given effect size, P-value is a reflection of sample size, in the sense that studies with very large sample size almost always reject the null hypothesis. In the 1950s, Lindley showed that a statistically significant finding from a study with very large sample size may represent strong evidence for the null effect, and this is later known as "Lindley's Paradox" [53]. For example, an observed proportion of 49.9% is consistent with the null hypothesis of 50.0% (P = 0.95) when the sample size is 1000 individuals; however, when the sample size is 1,000,000, P = 0.045 which is against the null hypothesis at the a level of 0.05. In other words, studies with very large sample size are very likely to find small P-values, but their evidence against the null hypothesis is very weak.

The implication is that the level of 5% may not be applicable to large sample size studies. Researchers need to adjust the observed P-value in large sample size studies. Good proposed a simple adjustment called *Q* or *standardized P*-value [54]: $Q = P\sqrt{n/100}$, where P is the actual P-value, *n* is the sample size. Thus, when $n = 100$, the standardized P-value *Q* is the same as the observed P-value. Good suggested that $Q > 1$ can be interpreted as support for the null hypothesis. Thus for $n = 1{,}000{,}000$ and P = 0.045, $Q = 4.5$, which is an evidence for the null hypothesis. Another solution is to set an 'optimal' α level based on a hypothesized effect size and cost of errors [55].

Many researchers mistaken the P-value as a false discovery rate. According to this view, a finding with P = 0.05 is equivalent to a false discovery rate of 5%. However, such an interpretation is also wrong. It can actually be shown that in the agnostic scenario a finding of P = 0.05 is equivalent to a false discovery rate of at least 30% [56]. It can also be shown that a P-value of 0.001 corresponds to a false discovery rate of 1.8% [57]. Thus, there is a call that the routine P-value should be lowered to 0.005 [58] or 0.001 [9] to minimize false discovery rate. The implication of these consideration is that researchers should not regard any result with P > 0.005 as an evidence of discovery.

1.9. Confidence interval

Researchers are almost always interested in knowing the size of an effect or magnitude of association which is not conveyed by P-value. Confidence interval provides likely values of effect size within an interval (usually taken as 95%) that are compatible with a study's observed data. Thus, confidence interval is a very useful complementary information pertaining to the practical significance of findings. For instance, a study testing the effect of supplementation of vitamins C and E during pregnancy concluded that the supplementation "*does not reduce the risk of death or other serious outcomes in their infants*" [59]. However, actual data showed that the relative risk of death or serious outcome (relative risk 0.79; 95% confidence interval, 0.61 to 1.02) clearly favored the supplementations group, even though P = 0.20.

Some researchers tend to mistakenly interpret confidence interval as a test of significance. In this view, a 95% confidence interval does not include the null hypothesis value is interpreted as statistically significant. On the other hand, a 95% confidence interval includes the null hypothesis value is considered statistically non-significant. However, confidence interval is a result of estimation, and it should *not* be interpreted within the framework of significance testing. Accordingly, a confidence interval from 0.61 to 1.02 should be interpreted that the data are compatible with a 49% reduction of risk or a 2% increase in risk. Thus, confidence interval should be named as "*Compatibility Intervals*" [60].

While reporting confidence intervals has been almost a norm in clinical research papers, it is still not widely adopted in animal research. Investigators in basic as well as translational research are suggested to report confidence interval for key measures in their papers.

1.10. Bayesian inference

A 95% confidence interval (CI) from *a* to *b* is sometimes interpreted as there is a probability of 95% that the true value lies between *a* and *b;* however, this interpretation is strictly incorrect. The actually interpretation of confidence interval requires a mental exercise: if the study were repeated infinite number of times with different samples, and a 95% CI is obtained for each time, then 95% of the intervals would contain the true value. That interpretation is based on the *frequentist school of inference*. Admittedly, it is not easy to comprehend the true meaning of CI.

The statement that 'there is a probability of 95% that the true value lies between *a* and *b*' can only be derived from a Bayesian analysis. A Bayesian analysis uses the Bayes' theorem to synthesize the prior information of an effect and the existing data to produce the posterior probability of an effect [61]. The posterior probability can directly provide the kind of answer that researchers want to have: *given the observed data, what is the probability that there is an effect/association*. Just as patients would like to know what is the probability of having a disease after seeing a test result, researchers want to know what is the probability of an effect after seeing result of a test statistic. P-value cannot answer that question; Bayesian analysis can.

Bayesian analysis allows the reporting of direct probability statements about any magnitude of difference that is of clinical interest [62,63]. For instance, a meta-analysis of 8 randomized controlled trials showed that supplements of calcium and vitamin D (CaD) reduced the risk of fracture in both community dwelling and institutionalized individuals [64]. Using a Bayesian analysis [65], we showed that the there was a 95% chance that the risk ratio of fracture associated CaD supplements ranges between 0.68 and 1.02. Moreover, there is a 44% probability that CaD supplements reduce fracture risk by at least 15% [65]. Sometimes, P-value based results are not necessarily consistent with a Bayesian analysis. For instance, based on the frequentist inference, the effect of alendronate on hip fractures may be interpreted as statistically non-significant at the alpha level of 5%; however, result of a Bayesian analysis indicated that there is a 90% probability that alendronate reduced fracture risk by at least 20% [66]. Although the Bayesian school of inference has been suggested as a paradigm of inference in the 21st century [67], its application in the medical research is

Table 1
List of common issues and suggested solutions.

Issue	Suggested solution
Lack of sample size justification	Provide a statement of sample size estimation, including hypothesized effect size, type I and type II error.
Confounders and biases	Regression adjustment, but be aware of over adjustment and unnecessary adjustment.
Data-dependent categorization of continuous data	Avoid categorization of continuous data. Use spline regression or non-parametric smoother.
Dichotomization of *P*-values into "significance" and "non-significance" based on the threshold of P = 0.05	Avoid dichotomization of P-value. Report actual P-values. Consider P < 0.001 or P < 0.005 as a threshold for discovery declaration.
Selection of 'significant' variables	Avoid stepwise regression. Consider LASSO and Bayesian Model Averaging methods.
Over-fitting: number of predictors is greater than the number of events	Consider LASSO and ridge regression analysis
Analysis of variance for longitudinal data	Consider linear mixed effects model as an alternative to repeated measures analysis of variance.
Multiple tests of hypothesis	Consider adjustment for multiple tests of hypothesis. Consider false discovery reporting [77].
P-value in very large sample size study	Consider Good's adjustment [54].
Interpretation of different P-values as different effect sizes	Avoid. Report confidence intervals.
Quantification of uncertainty of effect size	Consider Bayesian analysis.

still modest. The low level of uptake of Bayesian methods in medical research is partly due to the difficulty in choosing prior distributions that capture a reasonable amount of background knowledge. Many researchers used expert opinions for determining prior distribution, but this can create many biased problems. Nevertheless, in most cases, prior distributions can be generated from previously published data or from probability distributions that reflect a range of background knowledge about an association: non-informative, sceptical to optimistic [68]. Bone researchers are encouraged to consider Bayesian analysis and interpretation more often in their studies.

2. Conclusions

Statistical errors can arise in every phase of a study, from experimental design, data analysis to interpretation (Table 1). Data are products of experiment, and data quality is a consequence of experimental design. Good experimental design, whether it is animal study or clinical trial, is essential for generating high quality data. For a well-designed study with high quality data, simple statistical methods suffice in most cases, and the chance of statistical errors is low. Data can be adjusted, but study design cannot be reversed. Therefore, it is very important that issues concerning study design (eg, sample size, control, matching, blocking, randomization, measurements) should be considered at the beginning of a research project to minimize subsequent errors.

Although the focus of this article is on bone research, the errors identified here are also discussed in other areas of research [69–71]. Most of these errors come down to the practice of null hypothesis significance testing and P-value, which is the subject of intense debate among methodologists and practicing scientists [72]. It is recognized that the P-value overstates the evidence for an association, and that its arbitrary threshold of 0.05 is a major source of falsely interpreted true positive results. About 25% of all findings with P < 0.05, if viewed in a scientifically agnostic light, can be regarded as either meaningless [73] or as nothing more than chance findings [74]. There have been calls to ban P-value in scientific inference [75,76]. However, it is likely that the P-value is here to stay. Although P-value does not convey the truth, it is a useful measure that helps distinguish between noise and signal in the world of uncertainty. What is needed is the interpretation of P-value should be contextualized within a study and biological plausibility. It is hoped that this review helps improve statistical literacy along all phases of research.

Conflicts of interest

The author declares no competing interests.

Acknowledgments

Professor Nguyen is supported by a Leadership Fellowship of the Australian National Health and Medical Research Council (Grant Number APP1195305). **ORCID** Tuan V. Nguyen: 0000-0002-3246-6281.

References

[1] Altman DG. The scandal of poor medical research. BMJ 1994;308:283–4. 6924.
[2] Diong J, Butler AA, Gandevia SC, Heroux ME. Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. PLoS One 2018;13:e0202121.
[3] Brown AW, Kaiser KA, Allison DB. Issues with data and analyses: errors, underlying themes, and potential solutions. Proc Natl Acad Sci U S A 2018;115:2563–70.
[4] Schor S, Karten I. Statistical evaluation of medical journal manuscripts. J Am Med Assoc 1966;195:1123–8.
[5] Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. Contr Clin Trials 1989;10:31–56.
[6] Parsons NR, Price CL, Hiskens R, Achten J, Costa ML. An evaluation of the quality of statistical design and analysis of published medical research: results from a systematic survey of general orthopaedic journals. BMC Med Res Methodol 2012;12:60.
[7] Bolland MJ, Avenell A, Gamble GD, Grey A. Systematic review and statistical analysis of the integrity of 33 randomized controlled trials. Neurology 2016;87:2391–402.
[8] Colquhoun D. The reproducibility of research and the misinterpretation of p-values. R Soc Open Sci 2017;4:171085.
[9] Johnson VE. Revised standards for statistical evidence. Proc Natl Acad Sci U S A 2013;110:19313–7.
[10] Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
[11] Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. J Am Med Assoc 2005;294:218–28.
[12] Baker M. 1,500 scientists lift the lid on reproducibility. Nature 2016;533:452–4. 7604.
[13] Nguyen TV, Rivadeneira F, Civitelli R. New guidelines for data reporting and statistical analysis: helping authors with transparency and rigor in research. J Bone Miner Res 2019;34:1981–4.
[14] Jilka RL. The road to reproducibility in animal research. J Bone Miner Res 2016;31:1317–9.
[15] Manolagas SC, Kronenberg HM. Reproducibility of results in preclinical studies: a perspective from the bone field. J Bone Miner Res 2014;29:2131–40.
[16] Pearson K. The grammar of science. Cosimo Classics 2007.
[17] Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. J Pharmacol

Pharmacother 2010;1:100–7.
[18] von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61:344–9.
[19] Kilkenny C, Browne W, Cuthill IC, Emerson M, Altman DG. Animal research: reporting in vivo experiments: the ARRIVE guidelines. J Gene Med 2010;12:561–3.
[20] Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14:365–76.
[21] McAlinden C, Khadka J, Pesudovs K. Precision (repeatability and reproducibility) studies and sample-size calculation. J Cataract Refract Surg 2015;41:2598–604.
[22] Sena E, van der Worp HB, Howells D, Macleod M. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci 2007;30:433–9.
[23] Sackett DL. Bias in analytic research. J Chron Dis 1979;32:51–63.
[24] Bergman J, Nordstrom A, Hommel A, Kivipelto M, Nordstrom P. Bisphosphonates and mortality: confounding in observational studies? Osteoporos Int 2019;30:1973–82.
[25] Pearce N, Richiardi L. Commentary: three worlds collide: berkson's bias, selection bias and collider bias. Int J Epidemiol 2014;43:521–4.
[26] Day FR, Loh PR, Scott RA, Ong KK, Perry JR. A robust example of collider bias in a genetic association study. Am J Hum Genet 2016;98:392–3.
[27] Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 2009;20:488–95.
[28] Gibbons RD, Hedeker D, DuToit S. Advances in analysis of longitudinal data. Annu Rev Clin Psychol 2010;6:79–107.
[29] Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42:121–30.
[30] Detry MA, Ma Y. Analyzing repeated measurements using mixed models. J Am Med Assoc 2016;315:407–8.
[31] Berry DA, Ayers GD. Symmetrized percent change for treatment comparisons. Am Statistician 2006;60:27–31.
[32] Tu YK. Testing the relation between percentage change and baseline value. Sci Rep 2016;6:23247.
[33] Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol 2001;1:6.
[34] WHO. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. Technical report series 843. Geneva: WHO; 1994.
[35] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127–41.
[36] Becher H, Grau A, Steindorf K, Buggle F, Hacke W. Previous infection and other risk factors for acute cerebrovascular ischaemia: attributable risks and the characterisation of high risk groups. J Epidemiol Biostat 2000;5:277–83.
[37] Harrell FEJ. Regression modeling strategies. New York, NY: Springer; 2001.
[38] Smith G. Step away from stepwise. Journal of Big Data 2018;5:32.
[39] Genell A, Nemes S, Steineck G, Dickman PW. Model selection in medical research: a simulation study comparing Bayesian model averaging and stepwise regression. BMC Med Res Methodol 2010;10:108.
[40] Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging dor linear regression models. J Am Stat Assoc 1997;92:179–91.
[41] Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997;16:385–95.
[42] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.
[43] Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276–96.
[44] Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017;38:1805–14.
[45] Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods 2015;12:179–85.
[46] Fisher RA. Statistical methods for research workers. London: Oliver and Boyd; 1950.
[47] Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Phil Trans Roy Soc Lond 1933;231:289–337.
[48] Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. Ann Intern Med 1999;130:995–1004.
[49] Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol 2008;45:135–40.
[50] Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31:337–50.
[51] Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 2011;22:1359–66.
[52] Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet 2007;81:208–27.
[53] Lindley D. A statistical paradox. Biometrika 1957;44:187–92.
[54] Good IJ. Standardized tail-area probabilities. J Stat Comput Simulat 1982;16:65–6.
[55] Mudge JF, Baker LF, Edge CB, Houlahan JE. Setting an optimal alpha that minimizes errors in null hypothesis significance tests. PLoS One 2012;7:e32734.
[56] Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci 2014;1:140216.
[57] Sellke T, Bayarri MJ, Berger JO. 1 Calibration of p values for testing precise null hypotheses. Am Statistician 2001;55:62–71.
[58] Ioannidis JPA. The proposal to lower P value thresholds to .005. J Am Med Assoc 2018;319:1429–30.
[59] Rumbold AR, Crowther CA, Haslam RR, Dekker GA, Robinson JS, Group AS. Vitamins C and E and the risks of preeclampsia and perinatal complications. N Engl J Med 2006;354:1796–806.
[60] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567:305–7.
[61] Diamond GA, Kaul S. Prior convictions: Bayesian approaches to the analysis and interpretation of clinical megatrials. J Am Coll Cardiol 2004;43:1929–39.
[62] Nguyen TV. Pharmacogenetics of anti-resorptive therapy efficacy: a Bayesian interpretation. Osteoporos Int 2005;16:857–60.
[63] Nguyen TV, Pocock N, Eisman JA. Interpretation of bone mineral density measurement and its change. J Clin Densitom 2000;3:107–19.
[64] Weaver CM, Alexander DD, Boushey CJ, Dawson-Hughes B, Lappe JM, LeBoff MS, et al. Calcium plus vitamin D supplementation and risk of fractures: an updated meta-analysis from the National Osteoporosis Foundation. Osteoporos Int 2016;27:367–76.
[65] Frost SA, Nguyen TV. Uncertain effects of calcium and vitamin D supplementation on fracture risk reduction. Osteoporos Int 2016;27:2647–8.
[66] Nguyen ND, Eisman JA, Nguyen TV. Anti-hip fracture efficacy of bisphosphonates: a bayesian analysis of clinical trials. J Bone Miner Res 2006;21:340–9.
[67] Ruberg SJ, Harrell FEJ, Gamalo-Siebers M, LaVange L, Lee JJ, Price K, et al. Inference and decision making for 21st-century drug development and approval. Am Statistician 2018;73:319–27.
[68] Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Feldman BM. Methods to elicit beliefs for Bayesian priors: a systematic review. J Clin Epidemiol 2010;63:355–69.
[69] George BJ, Beasley TM, Brown AW, Dawson J, Dimova R, Divers J, et al. Common scientific and statistical errors in obesity research. Obesity 2016;24:781–90.
[70] Katz JN, Losina E. Uses and misuses of the P value in reporting results of orthopaedic research studies. J Bone Joint Surg Am 2017;99:1507–8.
[71] Borg DN, Lohse KR, Sainani KL. Ten common statistical errors from all phases of research, and their fixes. Pharm Manag PM R 2020;12:610–4.
[72] Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol 2010;25:225–30.
[73] Matthews R. Why should clinicians care about Bayesian methods? J Stat Inf Plan 2001;94:43–58.
[74] Berger JOST. Testing a point null hypothesis: the irreconciliability of p values and evidence (with discussion). J Am Stat Assoc 1987;82:112–22.
[75] Nelder J. From statistics to statistical science. Statistician 1999;48:257–69.
[76] Trafimow D, Marks M. Editorial. Basic Appl Soc Psychol 2015;37:1–2.
[77] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 1995;57:289–300.