Bayesians versus Frequentists: Comparing Methods in Medical Statistics

Katelyn Nelson

Department of Statistics, Brigham Young University

WRTG 316: Technical Communication

Professor Marsha Maxwell

December 16, 2022

Bayesians versus Frequentists: Comparing Methods in Medical Statistics

In the last decade the debate between using frequentist or Bayesian methods to analyze medical statistics has become more prominent. In this literature review, I will address the arguments made for both sides to provide some clarity about which method is preferrable. Research shows that Bayesian methods offer new possibilities and flexibility that traditional methods don't have, but frequentist methods are more standardized and have functioned well for a long time. Despite these differences, Bayesian and frequentist methods both have value because when results from the two methods are compared, they usually lead to the same conclusions. However, it is important to note that many experts consider Bayesian methods superior in efficiency of trials, flexibility, interpretability, and comprehensiveness. Making an informed decision about which method to use can improve the quality, speed, and interpretability of the analyses done with medical data. As Radzvilas et al. (2021) remarks, the sometimes-intense debate between frequentist and Bayesian methods and their applications in medical statistics is "understandable" because the decisions made will "affect statistical practice, and thereby the health, wealth, and happiness of nations" (p. 13690).

For decades, even centuries, statisticians have made important decisions regarding the health and safety of the public by relying on frequentist processes to analyze medical data. This branch is 'traditional' and well-known statistics that relies on relative frequencies from collected data. Over time frequentist processes have been refined and standardized to meet requirements established by Institutional Review Boards (IRBs) (Jiménez-Fonseca et al., 2021, p. 3). In comparison, Bayesian processes draw from known information and collected data and are not standardized because they have only risen in popularity in the last twenty-five years,

appealing to statisticians because of the greater flexibility they offer (Ashby, 2006, p. 3607-3608). Although both methods have the same goal – to use data to make informed decisions – they deliver results in different ways that could lead to different conclusions. It is rare to get contradictory results, but even so the statistical community has been trying to settle which method is better in medical statistics because they operate under fundamentally different assumptions and deliver results in forms that have different interpretations. The intuitive interpretation of Bayesian results is appealing, but processes would need to be improved for it to reach the same standard as frequentist methods.

In this literature review I will begin by briefly explaining frequentist methods, Bayesian methods, and their key differences. Then I will address the growth of Bayesian methods in the last few decades which led to the debate between methodologies. Finally, I will compare the advantages and disadvantages of each method.

An Overview of Statistical Methods

Although there are many branches of statistical thought, I am going to focus on two common ones, frequentist and Bayesian. I will give an overview of these two branches to provide a foundation for their comparison later in this report. To help you understand the overview of each method I will use two analogies to demonstrate applications of frequentist and Bayesian thought.

Helpful Analogy

During the time of Aristotle, Hippocrates, and other philosophers there were two main types of medical thought. The first way of thinking was rationalistic; physicians in this group examined individual patients as isolated events and responded to their condition based on

assumed causes of their ailment. I like to think of this as an objective, textbook diagnosis – catalog symptoms and find the matching problem or disease. Physicians in the second group, the empirical thinkers, administered treatments based on their experience as a physician in addition to the patients' symptoms. Unlike the rationalists, these empirical physicians relied on the wisdom they accumulated during their time spent with patients and their time as a physician. In this analogy, the rationalist physicians are considered frequentists and the empirical physicians are Bayesians (Tsagkaris et al., 2022, p. 100805).

Frequentist Methods

As mentioned above, a frequentist approach to analyzing data is objectively based on the available data. Frequentist tests include confidence interval estimation, hypothesis and significance testing, and maximum likelihood estimation. Results and test statistics include p-values, confidence intervals, and other summary statistics (Radzvilas et al., 2021, p. 13696). Later in this literature review I will compare these tests and results to those of Bayesian methods, so having a basic understanding of them is important.

Significance testing, one of the most common frequentist tools, is a way to compare data to an existing hypothesis and then draw conclusions on the validity of the claim. A significance test results in a test statistic called a p-value. If data has a low p-value, that means there is a very small probability of observing that data or similar data if the null hypothesis were true, which gives us reason to be skeptical of the null hypothesis. It is important to note that this test does not tell us the probability of the data being an accurate representation of the population, rather it gives insight about the validity of the test used to analyze the data and obtain that p-value (Wagenmakers et al., 2008 and Hittner & Fasina, 2021). Often used with p-

values, confidence intervals provide a range of plausible values for the estimated parameter based on the sample data (Radzvilas et al., 2021, p. 13703). To draw conclusions from a p-value or a confidence interval, you need to choose a significance level and a confidence level. A significance level (α) is the probability of rejecting the null hypothesis when it is true, which is chosen depending on how 'willing' you are to make this mistake. Often, we default to $\alpha = 0.05$ (Kostis & Dobrzynski, 2020, p. 109). The confidence level will be $(1 - \alpha)$. Again, we often use $(1 - 0.05) = 0.95$. If a p-value is below α, then the results are said to be significant and there is reason to believe that the null hypothesis is false. A confidence interval is interpreted with $100(1 - \alpha)\%$ confidence, indicating the confidence that if the procedure were repeated many times, it will indicate the same results $100(1 - \alpha)\%$ of the time. Again, it is important to note that a p-value is not the probability that the null hypothesis is true or false (Ferreira et al., 2020, p. 206). Similarly, a confidence interval is not a statement about the estimated value, but it "[describes] the procedure used to estimate" that value (Radzvilas et al., 2021, p. 13703).

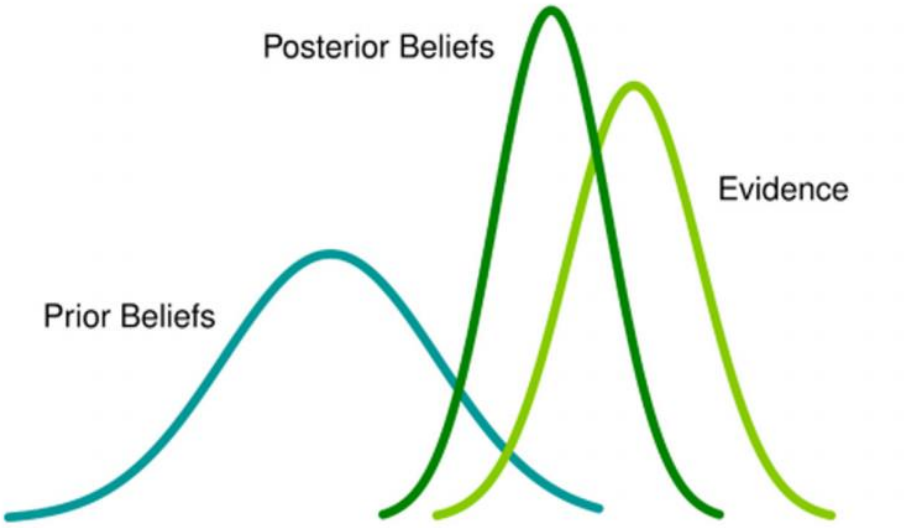
Bayesian Methods

Unlike a frequentist approach, Flores et al. (2022) points out that a "Bayesian approach involves (A) incorporating prior knowledge about variables and parameters, (B) learning from new, observed data and (C) using both forms of information to reach a balanced conclusion." Each of these pieces – the prior knowledge, data, and their balance – are represented by probability distributions, which show the possible and probable values a parameter can take (Ferreira et al., 2020, p. 202 and Wagenmakers et al., 2008, p. 189).

A simple way to think about the process of a Bayesian analysis is using your prior knowledge, gathering data, and combining the two to provide a complete view of the data called a posterior distribution. Previous knowledge that informs our inference is called a prior distribution. A prior distribution models how we think the data looks based on what we already know with no information from the data or trial (Teira, 2011, p. 275). Likewise, we use a likelihood distribution to model the data. After gathering data, we can use Bayes theorem to weigh the prior and likelihood to get a posterior distribution that is informed by our existing knowledge and the collected data as seen in Figure 1 (Teira, 2011, p. 276 and Wagenmakers et al., 2008, p. 190). Then we use the posterior distribution to inform us about the population of interest. Some common summaries of a posterior distribution that describe the parameter of interest are the mean, variance, and a credible interval. For example, the mean of the posterior distribution would give a 'point estimate' of the parameter (Teira, 2011, p. 276).

Figure 1

Weighing a Prior and Likelihood to Get a Posterior Distribution



Note. From "Limitations of randomized clinical trials," by J. B. Kostis & J. M. Dobrzynski, 2020, *The American Journal of Cardiology, 129*, 111 (https://doi.org/10.1016/j.amjcard.2020.05.011)

One unique aspect of inference from a posterior distribution is that it can be updated as we update the posterior. Flores et al. (2022) reasons, “For example, if we already have an estimate of a parameter's mean from related previous studies, this value could be used as the mean of a prior distribution…” (see also Ferreira-González & Ferreira-González, 2022, p. 167). Not only can we use estimates of parameters values to choose prior distributions, but we can also use previous posterior distributions. After all, a popular adage is “todays posterior is tomorrows prior” (Wagenmakers et al., 2008). This means that as more information becomes available, we can improve the posterior distribution and get more accurate results.

Key Differences

Although frequentist and Bayesian statistics can both be used to inform us about a population parameter, they use methods that deliver results in different formats. A Bayesian approach uses probabilities assigned to the hypotheses before doing any testing, while frequentists rely solely on the data gathered in the testing stage. Furthermore, Bayesians can provide inference on a specific hypothesis while frequentists focus more on the performance of a specific test (Radzvilas et al., 2021, p. 13691), which makes Bayesian methods particularly good for estimation while frequentist methods are good for dichotomous testing (Kostis & Dobrzynski, 2020, p. 111). These differences and their advantages and disadvantages applied to medical statistics will be discussed in more detail later.

The Rise of Bayesian Methods

Frequentist methods have dominated the history of statistics, but overtime Bayesian methods have become more common, especially in the medical community with research and clinical trials. According to Ashby (2006):

> Twenty-five years ago, Bayesian statistics barely got mentioned in the same breath as medical statistics; now the two are completely intertwined. No conference on Bayesian statistics in complete without medical applications, no conference on medical statistics is complete without some Bayesian approaches. (p. 3607)

To understand the conflict between methods, I will give background about how Bayesian statistics has evolved in the last few decades.

Influence of Technology

One of the reasons that there was little use of Bayesian analysis twenty-five years ago is because it is very complicated computationally, but computers can easily handle the difficult computations. So as computers became more accessible in the 1980s, Bayesian statistics exploded in popularity and complexity (Teira, 2011, p. 256). Common tasks that require statistical software include obtaining posterior distributions, using simulation algorithms to create approximations, and processing large amounts of data, which are all basic parts of Bayesian analysis. These methods are becoming more appealing because they quickly and easily create complicated models with continually improving computational technology (Wagenmakers et al., 2008, p. 181). Some posterior distributions that were impossible to calculate by hand have been simulated using software (Ferreira et al., 2020, p. 206).

Impact of the COVID-19 Pandemic

Some recent spikes of advancements in Bayesian technology in the medical field were a response to the COVID-19 pandemic. With patients in critical conditions, Bayesian approaches could respond to the need for faster trials and informed care because most frequentist significance tests were organized too quickly to gather sufficient participants to learn anything

substantial from the data (Teira, 2011, p. 293 and Jiménez-Fonseca et al., 2021, p. 2). The development of Randomized Embedded Multifactorial Adaptive Platform (REMAP) Trials, which are analyzed using Bayesian statistics, require fewer participants to collect informative data. Because of this REMAP trials are replacing randomized clinical trials (RCTs), which are analyzed with frequentist statistics. REMAP trials are also being used more because they can study several treatment domains at once whereas RCTs can only study one. This makes a clinical trial more adaptable and informative, which provides patients with faster treatment (Jiménez-Fonseca et al., 2021).

Strengths and Weaknesses

New Bayesian procedures in medical statistics that challenge current frequentist practices are what led to the debate between methods. Because the methods are fundamentally different it is difficult to compare them based on procedures, so I will compare them using categories that indicate where one method might be preferable over the other.

Efficiency

There have been complaints regarding a 'drug lag' since the late 1960s when the United States Food and Drug Administration (FDA) implemented regulations that are still in place today (Teira, 2011, p. 293), but faster trials using Bayesian analysis have been able to improve this lag. Current FDA regulations in medical techniques and pharmaceuticals require three phases of trials before they can be approved for use (What Are the Different Types of Clinical Research?, 2018; see also Teira, 2011, p. 292), which means that patients have been denied important drugs because they have not finished all three phases of testing. A Bayesian trial, such as a REMAP trial, is faster than the frequentist alternative, an RCT, and speed up the FDA's clinical

trial phases, which is beneficial for both the patients who want faster access to life-saving drugs and the industry that is producing them (Teira, 2011, p. 293).

According to Jiménez-Fonseca et al. (2021), RCTs pose "several dilemmas" that make them slower than Bayesian trials "including being slow, relatively inefficient, complex, and yielding broad conclusions […] that are scarcely generalizable." REMAP trials became popular during the COVID-19 pandemic because a Bayesian trial was able to be organized more quickly than RCTs or other frequentist trials. This is primarily because they require fewer participants to draw conclusions about the effects of a vaccine (Teira, 2011, p. 256), making them better suited to respond to the urgent demand. A patient needing an important treatment is less concerned about the dichotomy of RCTs and p-values, than they are about making fast and reasonable decisions that can be drawn from Bayesian inference.

Flexibility

"There is no such thing as a single Bayesian approach" which makes it more flexible in the data it can analyze and its applications to a wide variety of clinical trial designs (Teira, 2011, p. 278). One example of Bayesian flexibility is the situations it can be applied to; Bayesian hypothesis testing and models use marginal probabilities, which do not change in nested and non-nested models, whereas frequentist models rely on a nested design (Wagenmakers et al., 2008, p. 195).

It is also important to understand that frequentist calculations are highly dependent on the sample and experiment design. Wagenmakers (2008) explains a situation where a researcher asks a participant 17 questions, where 13 are answered correctly, and uses two different models to test the significant of these results. Using a binomial model, the p-value was

.049, but using a negative binomial model resulted in a p-value of .021 (p. 184). Both models are appropriate ways to approach the situation, but they gave different results based on the intention of the researcher. This indicates that if your research question or your approach to an analysis changes, you could get different results even if your data remains the same.

In contrast, this kind of flexibility is expected in a Bayesian analysis, which allows multiple tests to be performed without correction and benefits from updates in the samples and data used. In frequentist hypothesis testing, if you test multiple different comparisons or different values of ALPHA, then your calculations need to account for the fact that the more you test, the more likely you are to find significant results. Using Bayesian posterior distributions, you can test different things without worrying about corrections in the calculations (Ferreira et al., 2020, p. 205). Furthermore, changes in sample space or data are discouraged in a frequentist analysis after it has been completed, but they improve the inference offered by a Bayesian posterior. “Incremental learning” (Jiménez-Fonseca et al., 2021, p. 3) is particularly suited for Bayesian statistics which can take advantage of the constant evolution of data in the medical field, especially during an outbreak like COVID-19 that requires a learn-as-you-go approach. Using today’s posterior as tomorrow’s prior (Wagenmakers et al., 2008, p. 193) to update our inference gives Bayesian models an aspect of flexibility that frequentists will never have.

Sometimes people argue that there are disadvantages to how flexible a Bayesian model is and that the subjectivity of choosing prior distributions can damage the validity of Bayesian results. While this is true, it would be naive if we failed to recognize that there are subjective choices involved in every statistical analysis (Flores et al., 2022, p.1167). If we take the example

of using different models to obtain a p-values, given by Wagenmakers et al. (2008), that was a subjective choice as well, but we are more accustomed to that kind of subjectivity.

Ethics and Safety

Ethics is an important part of all statistics, but it should be the primary consideration of clinical trials because trials are conducted on individuals whose health and safety are at risk. There are arguments for both competing methods in different ethical situations, so it is important to understand the balance (Palmer, 1993, p. 220).

An ethical advantage of frequentist methods is that it uses a "first do not harm" principle, claiming that it is safer to avoid using a treatment until there is sufficient evidence to use it rather than risking it because of a preconceived expectation of success (Tsagkaris et al., 2022, p. 100805). Gathering such evidence using frequentist test requires randomization, which is considered controversial in clinical trials because some patients are not offered the chance to receive a possibly life-changing treatment. The benefit of this randomization is that it is impartial from a regulatory point of view (Teira, 2011, p. 257). Moreover, frequentist methods have a longer history in medical statistics and have more standardized processes and regulations (Teira, 2011, p. 274). Unlike frequentist methods, Bayesian methods are not standardized or as regulated because of their newness. However, Bayesian methods can perform trials with a smaller number of participants, exposing fewer people to the risks of experimental drugs (Teira 2011, p. 256).

The biggest conflict between the ethics of these methods is deciding which to value more: the individual or the collective. Individual ethics is "concerned with the well-being of each patient" in the trial, while collective ethics "emphasizes the common good for society"

(Palmer, 1993, p. 220). Palmer (1993) argues that in the early stages of new treatments an individual ethic should be prioritized by using Bayesian methods that don't require random treatment assignment so that doctors and patients can discuss what might be more beneficial for them. Consequently, during the later stages of clinical trials when the drugs have already undergone preliminary testing, a collective ethic using frequentist methods should be prioritized to begin to explain the general effect. This suggests that as the phases of a clinical trials progresses, the method used could shift from Bayesian to frequentist to match the shift in ethical goals. As in most cases, the ethical considerations and methods will depend largely on the unique trial because neither method is ethically superior in all situations.

Interpretability

To draw conclusions that drive decision-making once an analysis is complete, the results need to be interpreted in context of the specific trial. Frequentist results provide clear conclusions with dichotomous (yes or no) outcomes, but the metrics included in Bayesian results provide more direct and intuitive interpretations that give a better understanding of the data.

As mentioned in the overview of the methods, frequentist statistics describe the validity of tests and procedures that were used to obtain the results (Radzvilas et al., 2021, p. 13703). For example, when using significant testing a "p-value does not provide direct evidence for or against the alternative hypothesis" (Hittner & Fasina, 2021, p. 73). Frequentist confidence intervals have the same problem. Because they are based on p-values, frequentist interpretations lead to dichotomous conclusions, which means that "either the null hypothesis is supported or it is not supported" (Hittner & Fasina, 2021, p. 73). In context of a clinical trial,

this dichotomy of outcomes could mean deciding that either a drug is harmful, or it is not. Bayesian methods, however, directly address the alternative hypothesis. Using a posterior distribution, an estimate or range can be found for a parameter (Ferreira et al., 2020, p. 204), which might be used to quantify the possible harm a drug might cause. Then, based on that estimation, decisions can be made about whether these results are acceptable and safe.

Another consideration is that Bayesian results can be interpreted a lot more intuitively than frequentist results, which are often misinterpreted. People naturally want to interpret frequentist results about the hypotheses in an intuitive way, but it is difficult because they address the procedures rather than the hypotheses. Teira (2011) explains:

> P−values and confidence intervals are often misinterpreted in the medical literature as if they provided direct probabilities for particular events in clinical trials. If this is not just a misunderstanding, but rather the expression of the sort of probability assignment the medical profession is interested in, this is an argument for the Bayesian approach, in which these probabilities can be correctly calculated. (p. 291).

Bayesian credible intervals and posterior estimates are much more intuitively interpreted than frequentist confidence intervals and p-values. Ferreira et al. (2020) gives an example of frequentist versus Bayesian interpretation. In a clinical context, a frequentist will ask “What is the probability of having a temperature > 39.5°[C] with a diagnosis of influenza?”, but a Bayesian will ask “What is the probability of having the flu, knowing that the temperature is > 39.5°[C]?” which makes more sense in how we diagnose patients (p. 206). Some Bayesian models can become less intuitive as they become more complicated (Ferreira-González & Ferreira-González, 2022, p. 168).

While Bayesian methods do take more time to prepare, the preparation allows for a more “holistic understanding” (Flores et al., 2022, p. 1188) of the data that makes for richer inference about the parameter of interest (Hittner & Fasina, 2021, p. 75). Especially crucial in medical statistics, interpretability of results and understanding data can mean the difference between making a mistake or being informed enough to make the correct decision.

Performance

The two methods we are considering differ in many ways, but their performance when used in decision-making is generally the same. In the three examples written about by Flores et al. (2022), Bayesian and frequentist methods provided similar point estimates and results that would lead researchers to make similar conclusions regardless of method and the added complexity in the Bayesian calculations. Furthermore, in a simulation-based study Radzvilas et al. (2021) remarks that as sample sizes increased, Bayesian posterior probabilities and frequentist confidence intervals converged to be very close to the sample frequency and both methods encourage good decisions.

Although most of the time the decisions will be the same in either method, there are a few cases where one method performs better than the other. A Bayesian approach does not give as accurate results when the prior distribution is biased (Radzvilas et al., 2021, p. 13690). A frequentist prior performs poorly with high confidence levels and low α levels (Radzvilas et al., 2021, p. 13690, 13740). Additionally, in comparing COVID-19 test positivity (TP) rates, both methods performed well, except when comparing two countries whose TP rates were very close in magnitude. In that case the Bayesian results detected the smaller difference much better than the frequentist methods (Hittner & Fasina, 2021, pp. 75, 72). In some instances

when sample sizes were very large, frequentist tests had a lot of statistical power and were more likely to show significance even if the difference in TP rates was small (Hittner & Fasina, 2021, p. 74).

Another weakness in frequentist methods that is worth noting is the inconclusive nature of p-values. If a p-value is not significant, for example $\mathrm{p} = 0.057$, regardless of how close it is to the arbitrary α threshold, no knowledge is gained and there is nothing conclusive to aid in decisions because that p-value doesn't provide enough evidence (Jiménez-Fonseca et al., 2021, p. 2). When making decisions, the dichotomous nature of frequentist hypothesis testing is at a big disadvantage compared to Bayesian estimation methods. Furthermore, in 2016 the American Statistical Association (ASA) discredited reliance on p-values when they released a statement saying that "scientific conclusions and business of policy decisions should not be based only on whether a p value passes a specific threshold" (Kostis & Dobrzynski, 2020, p. 110).

Conclusion

As seen in the developments during the COVID-19 pandemic, the applications of Bayesian methods in medical statistics are changing what is possible in research and clinical trials. Frequentist methods are still a valuable and credible tool, but if Bayesian methods continue to develop, their advantages in the speed, interpretation, flexibility, and estimation of effects will make them the dominate method in medical statistics. As Wagenmakers et al. (2008) says, "inside every Non-Bayesian, there is a Bayesian struggling to get out" (p. 198) because we naturally want to interpret our data as a Bayesian would. However, since both methods should be given serious consideration (Radzvilas et al., 2021, p. 13690) the most

important thing is to choose which method to use based on the specific trial and application of your research. Kostis et al. (2020) encourages using Bayesian analyses if you value estimation over testing, which seems to be where medical statistics is heading.

References

Ashby, D. (2006). Bayesian statistics in medicine: A 25 year review. *Statistics in Medicine, 25*(21), 3589-3631. https://doi.org/10.1002/sim.2672

Ferreira, D., Barthoulot, M., Pottecher, J., Torp, K. D., Diemunsch, P. & Meyer, N. (2020). Theory and practical use of Bayesian methods in interpreting clinical trail data: A narrative review. *British Journal of Anaesthesia, 125*(2), 201-207. https://doi.org/10.1016/j.bja.2020.04.092

Ferreira-González, I. & Ferreira-González, C. (2022). Bayesian vs frequentist statistics: Afraid of losing the reference? *REC: Interventional Cardiology, 4*(3), 167-168. https://doi.org/10.24875/RECICE.M22000292

Flores, R. D., Sanders, C. A., Duan, S. X., Bishop-Chrzanowski, B. M., Oyler, D. L., Shim, H., Clocksin, H. E., Miller, A. P. & Merkle, E. C. (2022). Before/after Bayes: A comparison of frequentist and Bayesian mixed-methods-effects models in applied psychological research. *British Journal of Psychology, 113*(4), 1164-1194. https://doi.org/10.1111/bjop.12585

Hittner, J. B. & Fasina, F. O. (2021). Statistical methods for comparing test positivity rates between countries: Which method should be used and why? *Methods, 195*, 72-76. https://doi-org.byu.idm.oclc.org/10.1016/j.ymeth.2021.03.010

Jiménez-Fonseca, P., Salazar, R., Valentí, V., Carmona-Bayonas, A. & Agnelli, G. (2021). Learning in times of stress: Lessons from COVID-19 that will last throughout this century. *European Journal of Internal Medicine, 96*, 1-4. https://doi.org/10.1016/j.ejim.2021.11.002

Kostis, J. B. & Dobrzynski, J. M. (2020). Limitations of randomized clinical trials. *The American Journal of Cardiology, 129*, 109-115. https://doi.org/10.1016/j.amjcard.2020.05.011

Palmer, C. R. (1993). Ethics and Statistical Methodology in Clinical Trials. *Journal of Medical Ethics, 19*(4), 219–222. http://www.jstor.org/stable/27717332

Teira, D. (2011). Frequentist versus Bayesian clinical trials. *Handbook of the Philosophy of Science, 16*, 255-297. https://doi.org/10.1016/B978-0-444-51787-6.50010-6

Tsagkaris, C, Papazoglou, A.S., Moysidis, D.V., Papadakis, M.& Trompoukis, C. (2022). Bayesian versus frequentist clinical research now and then: Lessons from the Greco-Roman medical scholarship. *Ethics, Medicine, and Public Health, 23*, 100805. https://doi.org/10.1016/j.jemep.2022.100805

Radzvilas, M., Peden, W. & De Pretis, F. (2021). A battle in the statistics wars: A simulation-based comparison of Bayesian, frequentist and Williamsonian methodologies. *Synthese, 199*(5), 13689–13748. https://doi.org/10.1007/s11229-021-03395-y

Wagenmakers, EJ., Lee, M., Lodewyckx, T. & Iverson, G.J. (2008). Bayesian Versus Frequentist Inference. In *Bayesian Evaluation of Informative Hypotheses* (pp. 181-207). Springer. https://doi.org/10.1007/978-0-387-09612-4_9

What Are the Different Types of Clinical Research? (2018, November 3). U.S. Food & Drug Administration; *FDA*. https://www.fda.gov/patients/clinical-trials-what-patients-need-know/what-are-different-types-clinical-research