Outbreak, Surveillance and Investigation Reports
Field Epidemiology Training Program, Bureau of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625901734-5, Fax: +6625918581, Email: osireditor@osirjournal.net, http://www.osirjournal.net

# Bayesian Statistics in Epidemiological Investigations

Ying Lu[1,*], Christopher J Gregory[1,2,3]

1 Division of Global Health Protection, Thailand Ministry of Public Health–US Centers for Disease Control and Prevention Collaboration, Thailand

2 Division of Global Health Protection, Centers for Disease Control and Prevention, United States

3 Current affiliation, Division of Vector-borne Diseases, Centers for Disease Control and Prevention, United States

* Corresponding author, email address: ryluzhang@yahoo.com

On a stormy night on 31 May 2009, Air France Flight 447 took off from Rio de Janeiro bound for Paris and disappeared over the South Atlantic without trace. On board were 228 passengers and crew. After more than a year of searching for the plane wreckage yielded no results, and on the verge of giving up, the French Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile (BEA), the French authority responsible for safety investigations in civil aviation, contracted Metron, Inc. to launch a search using Bayesian methodology. Before Metron began physically searching the ocean, information was gathered from all previous searches, regardless of success or failure, and previous accidents occurring in the area and known ocean current dynamic factors. The Metron statisticians quantified the uncertainties in terms of probabilities to be used in formulating the search plan. The probabilities of the location where the plane went down were updated after each physical search. Using this strategic search plan, the plane wreckage and remaining bodies were found in less than a week. While the emotional news of the family members finally being reunited with the recovered bodies of their loved ones made the headlines, backstage the Metron statisticians realized the importance of the Bayesian method in solving a previously intractable mystery.

In this article, we will explore how the same Bayesian method can be used in epidemiology investigations starting with the long established history of Bayes' theorem in section one followed by the underlying philosophy behind the theorem in section two. Applications to various epidemiological investigations will be reviewed in section three and followed by an overall summary in section four.

## Section 1. History of Bayes' theorem

Over the past 200 years the use of statistics has revolutionized science, and Bayesian statistics has been presented and evolved during that time. Bayesian statistics is based on Bayes' theorem, or Bayes' rule, invented by the English statistician, philosopher and Presbyterian minister, Thomas Bayes, in the 18th century. This theorem provides a way to draw inferences not only from current study data, but also from other knowledge such as previous studies or expert opinion. During his lifetime, Bayes did not publish his theorem; it was only posthumously that his friend Richard Price published it. The first recognized Bayesian analysis was conducted by a French scholar, Pierre-Simon Laplace, in a study of birth data in France that included a total of 241,945 girls and 251,527 boys born in Paris from 1745 to 1770[1]. After calculating the probability that more boys than girls would be born, he concluded that the birth of more boys than girls was "a general law for the human race"[2]. This finding inspired further research that looked at factors that could influence the sex ratio.

During World War II (WWII), an English mathematician Alan Turing secretly used Bayes' rule to decrypt Germany's Enigma code by continuously updating the probability of the alphabetic letters using prior decrypted messages. He successfully located the U-boat submarines which were tying up thousands of ships and troops needed to support British war efforts[3]. This discovery was estimated to shorten the war by at least two years[4]. Even before the WWII, due to the rapid industrialization of the United States, Edward C. Molina, a leading expert in Bayesian theory, was urgently requested to evaluate the Bell telephone communication system to automate its labor intensive structure. Analyzing information of dialed phone numbers with the economics of various combinations of switches, selectors and trunking lines, Molina's calculating of probabilities based on Bayes' rule increased the automation capacity and reduced costs. This work played a critical role in making the Bell system more competitive[3]. The post-war building boom highlighted the need to set up the insurance schemes for sick and injured workers. Isaac M. Rubinow, a physician and statistician, and Albert Wurts Whitney, a specialist in insurance mathematics, used Bayes' rule to set up the models for actuaries in the insurance industry[3]. With its long established theory and successful applications, Bayesian statistics effectively informed decision making and solved the problems that were previously impossible to solve by conventional analysis.

## Section 2. Bayesian Philosophy

The philosophy behind Bayes' theorem is that knowledge of an interest is best provided not only from data of a single study, but also from incorporating other relevant information or prior knowledge of the interest. Many people apply Bayes' rule subconsciously, but in Bayesian statistical analysis, all the common sense ways of double checking can be quantified into a probability measure to estimate, or predict, uncertain situations to assist in decision making. This methodology imitates the approach that clinicians use routinely in diagnosing patients or the common sense that people use in daily life. For example, when an employer needs to hire qualified staff, the employer usually not only interviews the candidates, but also checks past work references of the candidates before making a decision. In this example, the qualification of a candidate is the main interest. The employer gets the evidence about the candidate's qualification not only through a face-to-face interview (the data from the study), but also through reference checking (other relevant information). Thus, the employer has comprehensive knowledge about the qualifications of a candidate based on two sources in order to select the best candidate.

The examples above and in section 1 include three components: uncertain prior knowledge of reality, the data generated based on current investigation or study, and the posterior probability of the reality. The Bayesian approach incorporates prior knowledge about the reality in the form of a prior distribution, which is then updated by information in the data, in the form of a likelihood function. Quantifying the prior distribution with the likelihood function generates a posterior distribution of the reality, which contains updated knowledge taking into account the information added by the data. This principle is Bayes' theorem[1].

## Section 3. Bayesian in Epidemiological Investigations

Over time, Bayes' theorem has been applied in many areas, including epidemiology and medical research. After WWII, the claim that smoking potentially caused lung cancer was fiercely debated. Epidemiological studies had been conducted because a spike in lung cancer incidence was observed after the wars when smoking was very prevalent. Among the research, the famous "Doll and Hill" case-control study that was published in 1954 showed a strong association between smoking and lung cancer[5]. In fact, both the authors were motivated to quit smoking because of this finding. The potential causal relationship between lung cancer and smoking attracted more attention from epidemiologists and the public health sector. A famous epidemiological investigation using Bayesian analysis was conducted in the 1950s by Jerome Cornfield, a biostatistician from the U.S. National Institutes of Health (NIH). He used lung cancer incidence data from the NIH as prior information and combined it with Doll and Hill's study data to calculate the probability of developing lung cancer caused by smoking[6]. Cornfield's analyses contributed in definitively establishing the causal relationship between smoking and lung cancer and this is considered as the most influential Bayesian analysis in the 1950s. One consequence has been that many populations, such as men in America, have seen a sharp decrease of lung cancer mortality since the 1990s because of decreased smoking prevalence starting in the 1960s[7].

Another impactful Bayesian epidemiological analysis was the re-assessment of mammogram screening as a tool in preventing breast cancer in the United States. Before 2009, it was recommended that all women aged 40 years or older have an annual mammogram to diagnosis early stage breast cancer. However, the majority of women who tested positive had a subsequent negative result by ultrasound indicating that they were free from breast cancer. The main interest in the reassessment was "what is the probability of breast cancer given a positive mammogram result?" The researchers used Bayes' rule to consider the prevalence of breast cancer as the prior probability in the population and the test result (data of the study) to update the probability of breast cancer given a positive test. For example, if there was a hypothetical population of 10,000 with breast cancer prevalence 0.4%, there would be 40 true breast cancer cases (10,000*0.4%). As sensitivity of the mammogram was 80%, there would be 40*80% = 32 true positive cases among breast cancer cases. Because of 90% specificity of the test, 996 would also test positive but be free of breast cancer. So the probability of having breast cancer given a positive mammogram would be only ~3% (32/(32+996)) (Table 1). This means that 97 out of 100 women with positive mammograms would have a false positive test, causing unnecessary worry and recommendations for further testing leading to a waste of money. With this evidence, in 2009, the United States Preventive Services Task Force changed the breast cancer screening recommendation against routine screening starting at age 40[8].

**Table 1. Mammogram screening result in a hypothetical population with prevalence of breast cancer 0.4%, sensitivity as 80%* and specificity as 90%† of mammogram**

| Mammogram | Breast cancer | Not breast cancer | Total |
|---|---|---|---|
| Positive | 32 | 996 | 1,028 |
| Negative | 8 | 8,964 | 8,972 |
| Total | 40 | 9,960 | 10,000 |

* National Cancer Institute at National Institute of Health, USA

† New England Journal of Medicine

Bayesian methodology is increasingly embraced by investigators in the United States Centers for Disease Control and Prevention (US CDC) in various studies. Using Bayesian methodology[9] in a pneumonia prevalence study by the Thailand Ministry of Public Health and US CDC, incidence of chest radiograph confirmed pneumonia in rural Thailand in children under five years old was 38% higher than estimated from conventional analysis. In a multi-site Pneumonia Etiology Research for Child Health (PERCH) study, Bayesian methods were similarly utilized to estimate the etiologies of childhood pneumonia[10,11]. In this study, multiple diagnostic tests were used to detect the specific pathogen that potentially could cause pneumonia, but none of the tests were 'gold standard' reference test with 100% sensitivity and specificity. In conventional analysis, one test must be considered as the perfect gold standard. However, Bayesian methodology can take into account the uncertain accuracy of all the tests (<100% sensitivity and specificity) and better estimate the probability of pneumonia etiology. The same Bayesian methods used in PERCH can be applied in other studies[10]. For example, the etiology of neonatal infections in South Asia case-control study aims to determine the etiology of serious neonatal infections, including sepsis and meningitis, using Bayes' theorem[12].

## Section 4. Summary

Given that the principles of Bayes' theorem were established many years ago and many successful applications were demonstrated in various fields, Bayesian analysis is considered a very important statistical method. However, it wasn't widely appreciated even among statisticians until the late 20th century, when computers with high speed computation capacity became available. Using this new capacity, the statisticians from the British Medical Research Council and Imperial College developed openBUGS for Linux and WinBUGS for Windows (https://www.mrc-bsu.cam.ac.uk/software/bugs/), free software used to facilitate previously difficult to calculate Bayesian formulas. With more powerful personal computers and the availability of free software and more statisticians with Bayesian expertise, the scientific community can embrace analysis using Bayesian methods more easily than ever before. Nowadays, in more and more institutions, Bayesian methodology is taught along with conventional statistical methodology so that more researchers from all disciplines, including public health, are able to apply it. Bayesian methods can solve many problems that weren't solvable by conventional methods due to its ability to effectively deal with uncertainties by considering multiple sources of information. In this booming information era, Bayesian methodology will almost certainly play an increasingly important role in statistical analyses for epidemiological investigations.

## Acknowledgements

We thank Dorothy L. Southern for providing help in scientific writing.

## Suggested Citation

Lu Y, Gregory CJ. Bayesian statistics in epidemiological investigations. OSIR. 2018 Jun;11(2):24-7.

## References

1. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd ed. New York: CRC Press; 2014.
2. Hald A. A detailed discussion of Laplace's birth studies. 1998.
3. Mcgrayne SB, The theory that would not die. London: Yale university press; 2012.
4. Copeland J. Alan Turing: The codebreaker who saved 'millions of lives'. 2012 [cited 2018 Feb 5]. <http://www.bbc.com/news/technology-18419691>.

5. Doll R. Hill AB, The mortality of doctors in relation to their smoking habits; a preliminary report. Br Med J. 1954 Jun 26;1(4877):1451-5.

6. Cornfield J. The estimation of the probability of developing a disease in the presence of competing risks. Am J Public Health Nations Health. 1957;47(5):601-7.

7. Islami F, Torre LA, Jemal A. Global trends of lung cancer mortality and smoking prevalence. Transl Lung Cancer Res. 2015 Aug;4(4):327-38.

8. Siu AL, U.S. Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2016 Feb 16;164(4):279-96. Epub 2016 Jan 12.

9. Lu Y, Baggett HC, Rhodes J, Thamthitiwat S, Joseph L, Gregory CJ. Bayesian latent class estimation of the incidence of chest radiograph-confirmed pneumonia in rural Thailand. Epidemiol Infect. 2016 Oct;144(13):2858-65. Epub 2016 Mar 2.

10. Deloria Knoll M, Fu W, Shi Q, Prosperi C, Wu Z, Hammitt LL, et al. Bayesian estimation of pneumonia etiology: epidemiologic considerations and applications to the pneumonia etiology research for child health study. Clin Infect Dis. 2017 Jun 15;64(suppl_3):S213-S227.

11. Wu Z, Deloria-Knoll M, Hammitt LL, Zeger SL. Partially latent class models for case-control studies of childhood pneumonia aetiology. JR Stat Soc C. 2016;65:97-114.

12. Saha SK, Islam MS, Qureshi SM, Hossain B, Islam M, Zaidi AK, et al. Laboratory methods for fetermining rtiology of neonatal infection at population-based sites in South Asia: the ANISA study. Pediatr Infect Dis J. 2016;35(5 Suppl 1):S16-22.