



Development of Biostatistics: From Past to Future

Biyoistatistiğin Gelişimi: Geçmişten Geleceğe

Sevilay KARAHAN

 0000-0002-8692-7266

Ahmet Ergun KARAAĞAOĞLU

 0000-0002-7024-7231

Department of Biostatistics, Hacettepe
University School of Medicine,
Ankara, Turkey

ABSTRACT

Biostatistics which is the application of statistics in the field of health and biology; provides powerful tools for creating questions, designing studies, developing measurements, and analyzing data and has an important place in determining the efficacy and safety of products such as drugs and vaccines. The impact of statistical sciences on medical and biological sciences has increased rapidly during the last few decades. Clinicians need to understand statistics well enough to follow up and evaluate empirical studies that provide an evidence base for clinical practice. Recent advances in biomedical research have created both new challenges and opportunities for statisticians and data scientists. Big data analytics, precision medicine, artificial intelligence, causal inference, and other new research resources inspire data scientists to develop modern statistical methods and innovative inference procedures. Therefore new philosophies such as causal models and prediction, new models such as graphical chain models and random effects models, faster computers and new clever algorithms for integration and maximization are needed. Without adequate investment in biostatistics, all medical research is at a significant risk of “drowning in data, but starving for knowledge”.

Keywords: Biostatistics; data science; big data.

ÖZ

İstatistiğin sağlık ve biyoloji alanındaki uygulaması olan biyoistatistik; soru oluşturma, çalışma tasarlama, ölçüm geliştirme ve verilerin analizi için güçlü araçlar sağlar ve ilaç, aşı gibi ürünlerin etkinlik ve güvenliğinin belirlenmesinde önemli bir yere sahiptir. İstatistik biliminin tıp ve biyolojik bilimlerdeki etkisi son yıllarda hızla artmıştır. Klinikisyenlerin, klinik uygulama için kanıtlar sağlayan deneysel çalışmalarını izlemek ve değerlendirmek için istatistiği yeterince iyi anlamaları gerekir. Biyomedikal araştırmalardaki son gelişmeler, istatistikçiler ve veri bilimcileri için hem yeni zorluklar hem de fırsatlar yaratmıştır. Büyük veri analitiği, hassas (kişiselleştirilmiş) tıp, yapay zeka, nedensel çıkarım ve diğer yeni araştırma kaynakları; veri bilimcilerine modern istatistiksel teknikler ve yenilikçi çıkarım yöntemleri geliştirme konusunda ilham vermektedir. Bu nedenle; nedensel modeller ve tahmin gibi yeni felsefelere, grafik zincir ve rastgele etki gibi yeni modellere, daha hızlı bilgisayarlara, entegrasyon ve maksimizasyon için yeni akıllı algoritmalara ihtiyaç vardır. Biyoistatistiğe yeterli yatırım yapılmazsa, tüm sağlık araştırmaları önemli bir “veride boğulma, ancak bilgi açlığından ölme” riski altındadır.

Anahtar kelimeler: Biyoistatistik; veri bilimi; büyük veri.

Corresponding Author

Sorumlu Yazar

Ahmet Ergun KARAAĞAOĞLU
ekaraaga@gmail.com

Received / Geliş Tarihi : 23.09.2021

Accepted / Kabul Tarihi : 02.11.2021

Available Online /

Çevrimiçi Yayın Tarihi : 18.11.2021

INTRODUCTION

How to translate a scientific question into a statistical one, and interpreting the result of a statistical analysis, are more important than knowing how to carry out the computations involved in the analysis. Scientific theory provides a model of how nature should behave; a scientific experiment generates data that show how nature actually behaves, and statistical inference is the bridge between model and data (1).

Statistics can be defined as the science of extracting information from data in the presence of variability and uncertainty and is a key tool that relates the diversity of each patient's observations to more abstract concepts such as clinical features, natural histories, clinical response and risks (2, 3). Studies with appropriate biostatistical support, from design to analysis and reporting, are the best experimental studies, therefore it plays a critical role in health research (3). Biostatistics which is the application of statistics in the field of health and biology; provides powerful tools for creating questions, designing studies, developing measurements, and analyzing data and has an important place in determining the efficacy and safety of products such as drugs and vaccines (4). The names such as biostatistics, biometrics, biometry and even bioinformatics, direct the same general enterprise of the use and development of statistical theory and methods to address design, analysis and interpretation of information in the biological sciences. Whatever it is called, it has always been and continues to be of great importance in the conduct of scientific investigations in agriculture, life sciences in general, ecology, forestry, medicine and public health. By combining statistical reasoning with knowledge of the real scientific problems, statisticians can and have made high-profile contributions both to the science and to the policy. With the explosion of data collection in areas like genomics, medical imaging, environmental sciences, with the increasing reliance on sophisticated mathematical modeling to explain biological phenomena; and with the increasing public focus on displaying "statistics" at every turn and disseminating results of studies, gained statistical sciences increasing importance (5). A biostatistician's unique contribution to a research team is the ability to measure uncertainty and generate robust inferences from data. Due to the increasing complexity and amount of health-related data, the need for biostatistics expertise in research teams is expanding and evolving (6).

One often faces a dilemma when asked to define *biostatistics*. As indicated by Khurshid et al. (7) the problem begins with the word itself. Its roots are Greek where the component *bios* involves biology; the study of living things and the component *statistics* involves the amassing, tracking, analysis, and application of data. A large variety of terms in scientific literature are interchangeably used: biostatistics, biometry, biometrics, biological statistics, medical statistics, clinical statistics, biostatistical science, sometimes even biomedical statistics, medical biostatistics, environmental statistics, pharmaceutical statistics, biopharmaceutical statistics, and public health statistics. The terminology is inconsistent and is confusing at best. Looking at these terms without an understanding, it seems that they all deal with different topics. Despite these differences, the terms are used to mean the same thing.

Chiang (8) in his discussion on an article by M. Zelen, who indicated that it was difficult to classify biostatistics as a discipline and proposed the term *biostatistical science*, stated that he did not share his (Zelen's) difficulty in classifying biostatistics as a discipline. He defined biostatistics as a discipline that is concerned with the development and application of statistical theory and methods for the study of phenomena arising in the life sciences and whether biostatistics is or is not a discipline depends on the amount and quality of knowledge that has been developed and accumulated in the field. Although biostatistics could not be considered as a discipline till 1940's, by the drastic change experienced after then, researches shifted from descriptive statistics to the development of theoretical basis for the field. As a result, biostatistics today contains a respectable body of knowledge both in quality and quantity and is built on a solid theoretical foundation. As a conclusion Chiang (8) stated that more people in mathematical statistics will focus on biostatistical problems, and biostatistics will increase in its contribution and importance in the advancement of science (8).

Biostatistics is the application and development of statistical techniques for biological sciences and biostatistics is defined as "statistical method(s) in medicine and the health sciences" (9).

The impact of statistical sciences on medical and biological sciences has increased rapidly during the last few decades. Physicians practice on the basis of clinical knowledge, which is framed after a series of tests, treatments and statistical analyses. A physician may not have a sound knowledge of statistical principles or techniques, but the information he uses in the clinical decision-making process is undoubtedly always based on statistical evidence. However, conclusions drawn from the statistical evidence may be inaccurate or misleading and therefore, without a sound understanding of statistics, a physician may not be able to reach the most appropriate decision.

Statistics is a discipline that has changed science, medicine, and public policy. Biostatistical principles are necessary in all branches of biology and medicine, and have become a mandatory part of medical research. The Human Genome Project, for example, while relying on advanced biological techniques, also depends heavily on statistical techniques for extracting the right data out of a large pool of gene sequences. The evolution of (bio)statistics, is parallel to development in other scientific fields, particularly medicine (7).

The statistical analysis should be part of the scientific method used to test the hypothesis and should be planned before the study is started. Thus, the study of probabilities and the way they have to be interpreted in medical practice tell you something about scientific method. And scientific method is the only thing keeping you all from being quacks (10).

HISTORICAL DEVELOPMENT OF STATISTICS AND BIOSTATISTICS

In earlier times when scientists started to understand and explore the nature and their environment the very first thing they used to do is to measure and record some

important features. Therefore the history of statistics dates back to ancient times. The first appearance of statistics in the world began with the census of soldiers and keeping records of events such as marriage, birth and baptism. As it can be understood from here, the first statistical studies were mostly on descriptive statistics. The origins of probability and statistics are usually found 1650 - 1700 period by dealing with the mathematical explanation of games of chance and the study of mortality data. During the 1700's, in the Age of Enlightenment, things gradually began to change. The scientific method, based on empirical evidence about diseases and the impacts of different interventions, began to be applied. However, the study of medical treatment remained almost entirely qualitative. At the end of the eighteenth century, Gilbert Blane described the research process as hinging almost exclusively on clinical reasoning, with no hint of quantitative analysis (11). This was also the age of Scientific Revolution and many well-known scientists like Galileo and Newton, although not influencing its development, were the early users of probability and statistics (1). Not too long after, however, statistical ideas would start gaining importance. By the early nineteenth century, the concept of probability, essentially in its current form, was well known. Pierre-Simon de Laplace was the leading scientific theorist of the era, and a strong believer in the potential of statistical analysis in various fields. In medical research, he advocated comparing rates of success between alternative therapeutic interventions. With the use of quantitative data, scientists such as John Graunt, William Petty, Edmund Halley, Jacob Bernoulli started to work in the field of mathematical statistics. Graunt modeled population growth, and Halley applied statistical models to the insurance industry. Deparsier calculated the average life expectancy by creating life tables. Bernoulli's law of large numbers is an important milestone. Theorems developed by important scientists such as Laplace, Legendre, Gauss and Poisson are still used today (12).

The person who laid the foundations of biometrics is Quetelet, who is considered the father of modern statistics. Quetelet was the first to combine the methods of anthropology and social statistics with results from probability theory and mathematical statistics. He produced works such as "On the development of man and his abilities or the experience of social physics", "On the social system and the laws that govern it" and "Anthropology". Russian statisticians such as Chebyshev, Markov and Kolmogorov also produced very important works on statistics (12).

Two of those who played an important role in the application of statistics to biology are the British statisticians Francis Galton and Karl Pearson. Darwin's cousin, Galton, published work on anthropology and genetics. Similarly, Pearson successfully studied the heredity and variability problems of organisms. He is the developer of methods such as correlation coefficient and chi-square statistics, which are still widely used today (13). He is also the founder of *Biometrika*, one of the most important journals in the field, which has been published since 1901. Gosset, a student of Pearson and publishing his work under the pseudonym "student"; worked on small sample theory (12).

Another person who made great contributions to biometrics is Ronald Yelmer Fisher. As both a biologist-experimenter and mathematician-statistician, Fisher brought not only new methods but also new ideas to biometrics. He revolutionized the theory and practice of statistics, especially as it applies to agricultural experimentation. Fisher is revered not only as a great statistician but also as one of the great geneticists of the twentieth century. Two of Fisher's fundamental contributions to experimental design were the ideas of randomization and blocking (1).

Fisher's revolutionary innovations in experimental design and analysis proved enormously successful in several fields, especially in agricultural research and industrial engineering. Fisher's methods of significance testing, building upon W.S. Gosset's initial breakthrough of the t-test in 1908, finally solved the problem of how to analyze experiments with modest sample sizes. His designs based on random assignment of treatments to experimental units provided not only a firm basis for calculating p-values for significance tests, but also a means of eliminating bias resulting from uncontrolled "causes" influencing the outcome. One very important but controversial direction of this development was initiated by Jerzy Neyman and E.S. Pearson. These pioneers regarded statistics primarily as a way to guide decisions, an approach that Fisher found inappropriate for scientific research. However, the Neyman-Pearson decision-theoretic methods, including such concepts as confidence intervals, Type I and Type II error rates and statistical power, became widely accepted. These tools proved particularly useful for research related to industrial product development, production and testing. The first randomized clinical trial, directed by Hill in 1946 under the auspices of the British Medical Council, demonstrated clearly the benefits conferred by streptomycin for the treatment of tuberculosis. Hill's pioneering efforts began shifting medical opinion toward a greater appreciation of the value to society of definitive findings that could be obtained via randomized clinical trials (11).

When it comes to today, Marvin Zelen noted the emergence of a field called the "science of biostatistics", referring to the application of statistics, probability, computation, and mathematics to a subject area. Also, Zelen said, "The future of Biostatistics Science will be closely related to computation" (13).

Till 1980's, "pure science" was considered "better science." Important break-throughs in the beginning of 1990's, primarily in the field of molecular biology, have made it increasingly clear that the future is bright for interdisciplinary and multidisciplinary efforts and that academic institutions should prepare themselves for this trend (5).

FUTURE OF BIOSTATISTICS

Recent advances in biomedical research have created both new challenges and opportunities for statisticians and data scientists. Big data analytics, precision medicine, artificial intelligence, causal inference, and other new research resources inspire data scientists to develop modern statistical methods and innovative inference procedures (14). Nowadays scientists are confronted with larger data sets and more complex data. Therefore new philosophies

such as causal models and prediction, new models such as graphical chain models and random effects models, faster computers and new clever algorithms for integration and maximization are needed (15).

The recently popularized concept of data science is a combination of mathematics-statistics, field knowledge and computer knowledge (Figure 1). Data is the most precious mine of the information age. Data scientists are needed to process this mine and turn data into information. Data science requires fusion of statistical thinking and information technology (16).

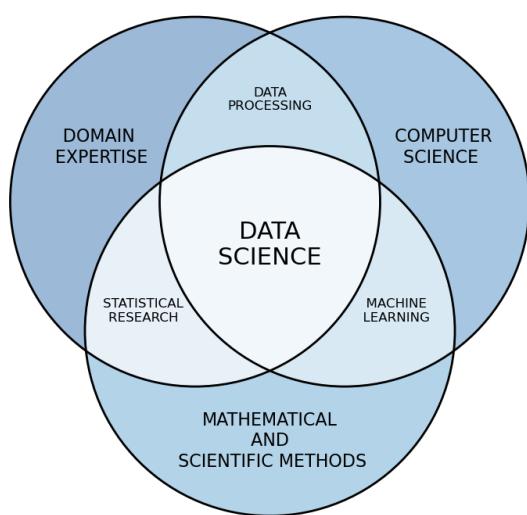


Figure 1. Components of data science (16)

Millions of dollars are spent on medical and public health research around the world each year. The correct use of this expenditure depends on the correct analysis and interpretation of the data. Investments in this research are at risk if insufficient attention is paid to biostatistics. As the data collected grows, the risk of not understanding the structure of this data and not using the right statistical approaches in the analysis will also increase (3).

Clinicians need to understand statistics well enough to follow up and evaluate empirical studies that provide an evidence base for clinical practice. They should realize the statistical aspects of the clinical literature, evaluate the strengths and weaknesses of the presented analyzes, and strengthen their active participation in research. A perfectly chosen and applied analysis would be misleading best if it is done from data collected using incorrect data or data collected using an incorrect measurement technique, or at the wrong time. To quote the aphorism often uttered in introductory statistics classes, "garbage in, garbage out" (2). An in-depth education is needed to establish the connection between health problems and statistical methods. Biostatistics is a basic discipline required by the clinician, master's or doctoral student / graduate who conducts research in a very wide area. This situation sometimes causes biostatistics to be seen as a simple technical tool rather than a basic discipline (3). Although many experienced researchers value collaboration with

biostatisticians, biostatistics is sometimes considered an ancillary service rather than an academic discipline. To maximize the contribution of biostatisticians to research, biostatistics must be integrated with the purpose of research (6). Shallow and misuse of statistics can easily lead to unscientific applications. Biostatistics requires not only knowledge of statistical methods but also the use of technical skills, including computation (3). On the other hand, sometimes a large gap occurs between biostatistician's advanced statistical model and clinicians' perspective. Both sides should strive to build a bridge that will bridge this gap (15).

In the twenty-first century, there has been a significant increase in data collection at low costs. Large and frequent data are collected in the fields of molecular biology, health science, engineering, geology, climatology, economics, finance and humanities. For example, in biomedical research, MRI, fMRI, microarray, and proteomics data are often collected for each subject, involving hundreds of subjects; in molecular biology, large sequencing data is rapidly becoming available; thousands of high-resolution images are collected in natural resource exploration and agriculture; millions of transactions are recorded every day in business and finance. Huge volumes of data are recorded on social media platforms. The frontiers of science, engineering, and the humanities differ in problems in studies, but still share a common theme: big or complex data is being collected and new information needs to be discovered. Big data and new scientific research have a strong influence on statistical thinking, methodological development and theoretical work. It also challenges traditional statistical theory, methods, and computation. Many new insights and phenomena need to be discovered and new statistical tools need to be developed (17, 18).

The application of data-intensive biomedical assays and technologies, such as DNA sequencing, proteomics, imaging protocols, and wireless health monitoring devices have revealed a great deal of inter-individual variation with respect to mechanisms and factors that influence disease. This has led to the belief that interventions must be tailored (i.e., 'personalized') to the features each patient possesses. In the context of personalized medicine (precision medicine), if it is not known a priori what intervention might 'match' a patient's profile, then it becomes an empirical question as to which intervention might be most appropriate for that patient (19).

CONCLUSION

With the onset of the big data era, more and more people are trying to draw conclusions from this data. However, the spread of software developed by those who do not have sufficient knowledge about statistical methods poses an increasing risk. Big data requires both an advanced understanding of basic statistical concepts and methods and a mastery of computational tools such as dimension reduction and machine learning. More data does not always mean better data, and more analysis does not necessarily mean better science. The quality and reproducibility of research findings depend on the design of the data collection process and the sources of limitations and biases of the research. Without adequate investment in biostatistics, all medical research is at a significant risk of "drowning in data, but starving for knowledge" (3).

Ethics Committee Approval: Since our study was a review, ethics committee approval was not required.

Conflict of Interest: None declared by the authors.

Financial Disclosure: None declared by the authors.

Acknowledgements: None declared by the authors.

Author Contributions: Idea/Concept: SK, AEK; Design: SK, AEK; Data Collection/Processing: SK, AEK; Analysis/Interpretation: SK, AEK; Literature Review: SK, AEK; Drafting/Writing: SK, AEK; Critical Review: SK, AEK.

REFERENCES

- Diggle PJ, Chetwynd AG. Statistics and scientific method: An introduction for students and researchers. 1st ed. New York: Oxford University Press; 2011.
- Barkan H. Statistics in clinical research: Important considerations. *Ann Card Anaesth.* 2015;18(1):74-82.
- Lee KJ, Moreno-Betancur M, Kasza J, Marschner IC, Barnett AG, Carlin JB. Biostatistics: a fundamental discipline at the core of modern health data science. *Med J Aust.* 2019;211(10):444-6.e1.
- O'Neill RT. FDA's critical path initiative: a perspective on contributions of biostatistics. *Biom J.* 2006;48(4):559-64.
- Molenberghs G. Biometry, biometrics, biostatistics, bioinformatics, ..., bio-X. *Biometrics.* 2005;61(1):1-9.
- Welty LJ, Carter RE, Finkelstein DM, Harrell FE Jr, Lindsell CJ, Macaluso M, et al; Biostatistics, Epidemiology, and Research Design Key Function Committee of the Clinical and Translational Science Award Consortium. Strategies for developing biostatistics resources in an academic health center. *Acad Med.* 2013;88(4):454-60.
- Khurshid A, Ageel MI, Anwer S. Evolution of (bio)statistics in medical research: Fifty eight years of "Numbering Off". *Int J Crit Stat.* 2013;4(1):5-17.
- Chiang CL. What is biostatistics? *Biometrics.* 1985;41(3):771-5.
- Kotz S, Read CB, David L. Biostatistics. In: Kotz S, Balakrishnan N, Read CB, Vidakovic B, editors. 2nd ed. New Jersey: John Wiley & Sons; 2005. p. 551-63.
- Morgan PP. What are my chances of understanding biostatistics? *CMAJ.* 1986;134(10):1105-6.
- Weisberg HI. Statistics and clinical trials: Past, present and future. In: *JSM Proceedings, Section on Statistical Education.* Alexandria, VA: American Statistical Association; 2011. p. 1547-61.
- Biduchak A, Grytsiuk M, Chornenka Z, Domanchuk T. History of formation and development of biostatistics. *Current Issues of Social Studies and History of Medicine.* 2020;2(26):84-90.
- Öğüş E. To be together medicine and biostatistics in history: Review. *Türkiye Klinikleri J Biostat.* 2017;9(1):74-83.
- Zhao Y, Abebe A, Qi L, Zhang M, Zhang X. New advances in biostatistics. *J Probab Stat.* 2019;Special Issue:1352310.
- van Houwelingen HC. The future of biostatistics: expecting the unexpected. *Stat Med.* 1997;16(24):2773-84.
- Akdeniz F. New trends and developments in statistics. *Soc Sci Res J.* 2015;4(4):1-11.
- Fan J, Lin X, Liu JS. New developments in biostatistics and bioinformatics. New Jersey: Higher Education Press; 2009.
- Sathian B, Sreedharan J. Importance of biostatistics to improve the quality of medical journals. *WebmedCentral BIOSTATISTICS.* 2012;3(5):WMC003332.
- Schork NJ. Randomized clinical trials and personalized medicine: A commentary on deaton and cartwright. *Soc Sci Med.* 2018;210:71-73.