

**USING KNOWLEDGE OF BIOSTATISTICS IN
THE DAILY PRACTICE OF A DOCTOR.
SOFTWARE OF STATISTICAL RESEARCH
AND THE ORDER OF PRESENTATION OF
SCIENTIFIC WORKS.**



RESEARCH AND THE POPULATION PERSPECTIVE

The patient-perspective remains preeminent for good and undeniable reasons. Examining, identifying, and treating the patient for his or her benefit alone is a tremendous responsibility for which we are disciplined and trained. We should not turn our faces from this challenge.

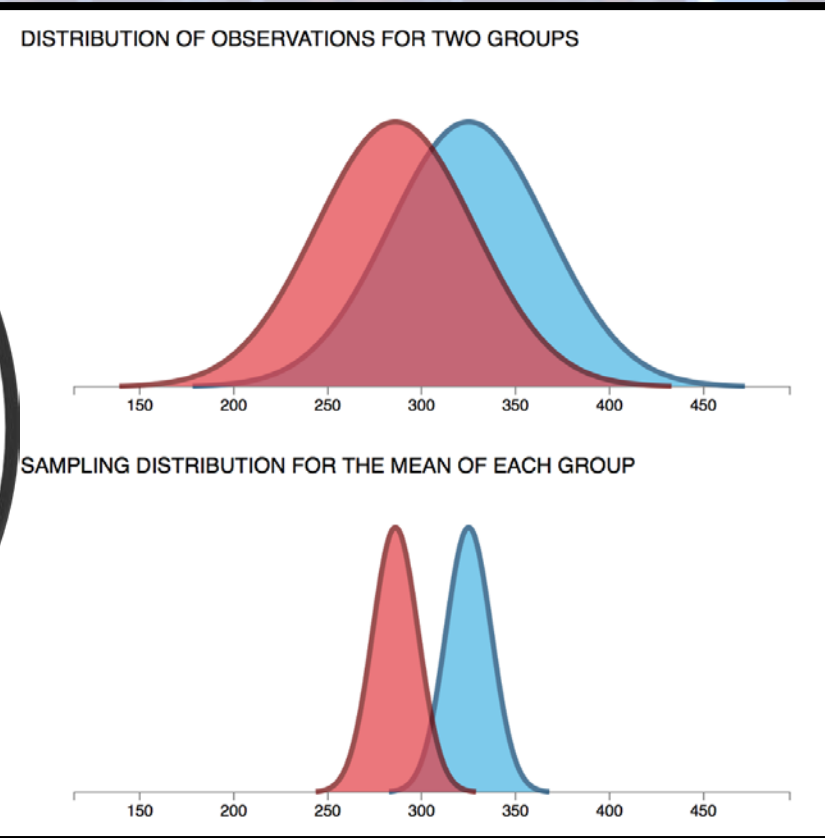
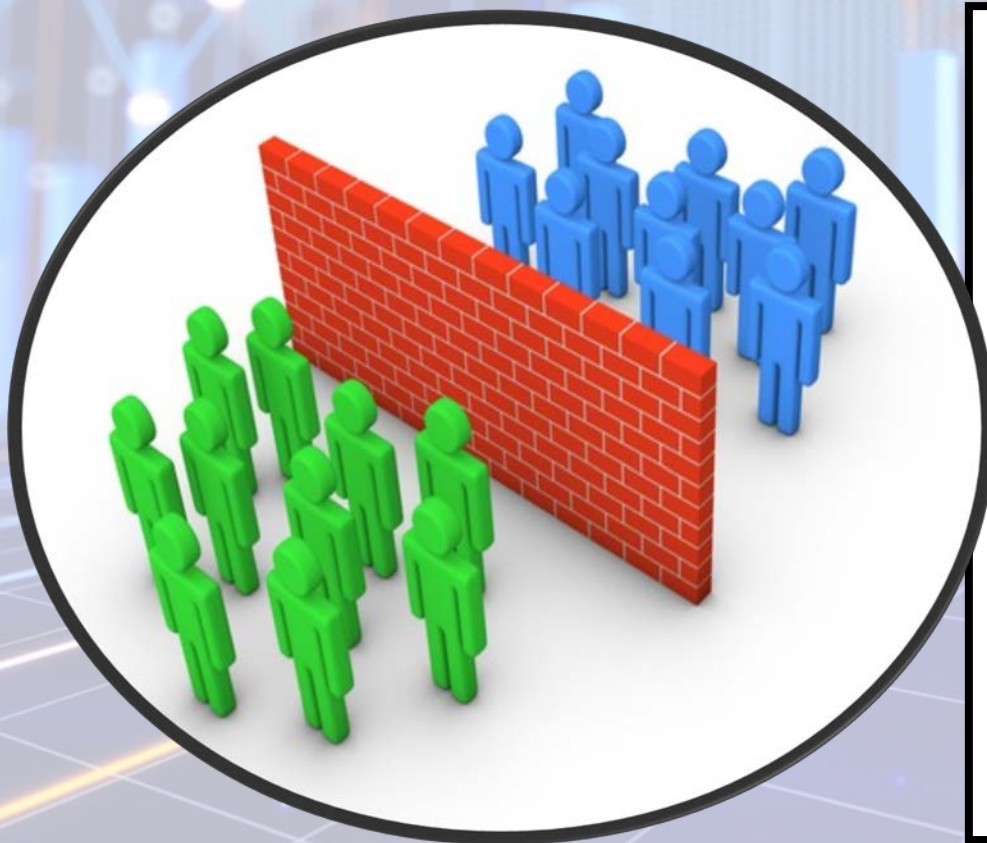


Yet the principles of research are different from those of medical practice. In fact, many physicians believe that in order to understand research tenets, they must embrace concepts that they have been specifically trained to shun.

One example of this disconnect between research and medical practice principles is the concept of the “average” patient. The average blood pressure produced by the active group in a clinical trial may be only 2-3 mm.Hg. lower than in the control group, a finding heralded as an important advance by the research community. However, physicians treating patients in the community are not impressed with the change, asking how a 2-3 mm Hg reduction could possibly matter in a patient where the daily stresses of work and family living generate changes in blood pressure that are more significant.

Another confusing message for the physicians laboring to understand research issues is the concept of variability. Treating physicians clearly understand that patients are very different from one another, unique in an infinite number of ways. We take this into account, adjusting the patient's visit schedule in accordance with the patient's stability or our concerns. We never think of modifying variability. Yet, researchers become experts at controlling, adjusting, or even removing variability.

THE DIFFERENCE BETWEEN THE TWO GROUPS OF WORKERS IS PERSPECTIVE. PHYSICIAN VIEWS DEVELOP FROM THE PATIENT-BASED PERSPECTIVE, WHILE RESEARCHERS ARE COMMONLY POPULATION-BASED. EACH PERSPECTIVE USES DIFFERENT TOOLS BECAUSE ITS GOAL'S ARE DIFFERENT.



THE NEED FOR INTEGRATION



- The two examples cited above reflect the difficulties that complicate the physician's attempt to integrate relevant research findings into their body of knowledge. The different points of view between the population and patient perspective induce a tension in physicians because their goals are different. Yet it is up to us to integrate these disparate points of view, accepting the best of each while simultaneously understanding their limitations.
- We can begin this incorporation process by repudiating the idea that individual patient preeminence is negotiable. In practice, the patient must come first — subjugating our primary concern for the patient to any other perspective leads to unethical, avoidable damage to patients and their families. However, embracing the patient perspective blinds us to an objective view of the effect of our interventions.

We as practicing physicians often cannot see the results of our own inter-ventions in an objective light. For example, if patients who respond well to our treatments return to see us, while the poor responders drift away, we are lulled into a false sense of confidence in the positive utility of the intervention.

On the other hand, if patients who are invigorated by our treatment do not return, while only the dissatisfied come back to demonstrate the inadequacies of our therapeutic approach, we may prematurely abandon an important and promising therapy.

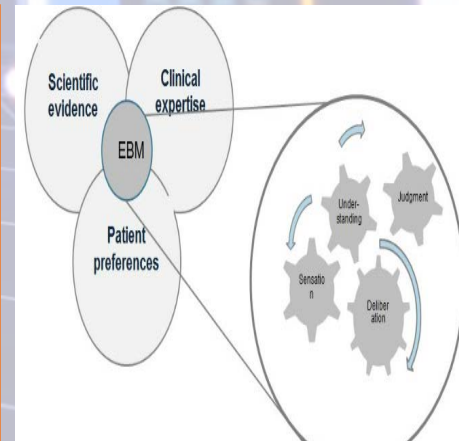
The fact is that some patients improve (or deteriorate) regardless of what we do. By concentrating on what we observe, to the exclusion of what should be considered but is unobservable, we lose sight of an unbiased view of the intervention.



TRAP

An objective view of therapy begins with a balanced mindset, requiring that we respond to our patients' trust in us by modulating our own belief in the therapies that we advocate. We physicians as a group have become too easily enamored of promising theories, e.g., the arrhythmia suppression hypothesis or the immunotherapy/interferon hypothesis. These theories, well researched at the basic science level, are replete with detail, holding out tremendous promise for our patients. This combination of innovation and hope can be hypnotic. Furthermore, initial Findings in individual patients can lead to a domino effect of individual anecdotes, discussion groups, articles, and treatment recommendations. This set of theories can easily take on a life of their own.

Herein lies the trap. Well-developed and motivated theories are commonly wrong and unhelpful in medicine. Basing patient practice on an attractive but false theory has had and will continue to have devastating effects on patients, their families, and our communities. Accepting the conclusion before formally testing it is a wonderful principle in religion, but it is antipodal to science.



In order to keep our guard up, we must also remind ourselves, somewhat painfully, that professional and compassionate doctors who placed their patients first as we do, who were true to their oaths then as we are now, were often the source of destructive, barbaric treatments. For hundreds of years, compassionate physicians allowed infected wounds to fester because an ancient text declared authoritatively and erroneously that purulence was beneficial, and that cleanliness during surgery was unnecessary. More recently, we have learned that hormone replacement therapy for women as commonly administered for years is harmful.

The harmful actions of physicians based on our faith in false beliefs are the sad, natural consequences of accepting the patient-preeminence philosophy as an unbiased one. We must therefore recognize that, while the patient perspective is appropriate, we require additional tools to observe the effects of our therapies objectively. The population-perspective provides these tools. Thus, the population perspective does not represent a threatening viewpoint, but an additive one, and its consideration does not lead to philosophy replacement but perspective enlargement.

CLINICAL VERSUS RESEARCH SKILLS



However, the fine clinical motivation for carrying out research can obstruct the conduct of that research. The impetus to use research to address an unmet medical need is closely aligned with an essential creed of modern medicine, to relieve pain and suffering. However, this can easily translate to the sense that great haste must be expended in the research effort. Left unchecked, this ideology can generate the belief that speed is everything, and that everything—discipline, standards, dialogue, methodology, quality, and good judgment — should be sacrificed for it. All of these principles add value to the scientific effort but unfortunately are shunned for the sake of rapid progress.

MODEL FRAMEWORKS FOR MEDICAL DECISION MAKING



It is a poorly publicized fact that, in addition to the basic science courses and clinical rotations that they must do during their training, physicians also take courses in biostatistics and medical decision making. In these courses, prospective physicians learn some math and statistics that will help them as they sort through different symptoms, findings, and test results to arrive at diagnoses and treatment plans for their patients. Many physicians, already bombarded with endless medical facts and knowledge, shrug these courses off. Nevertheless, whether they learned it from these courses or from their own experiences, much of the reasoning that physicians use in their daily practice resembles the math behind some common machine learning algorithms. Let's explore that assertion a bit more in this section as we look at some popular frameworks for medical decision making and compare them to machine learning methods.

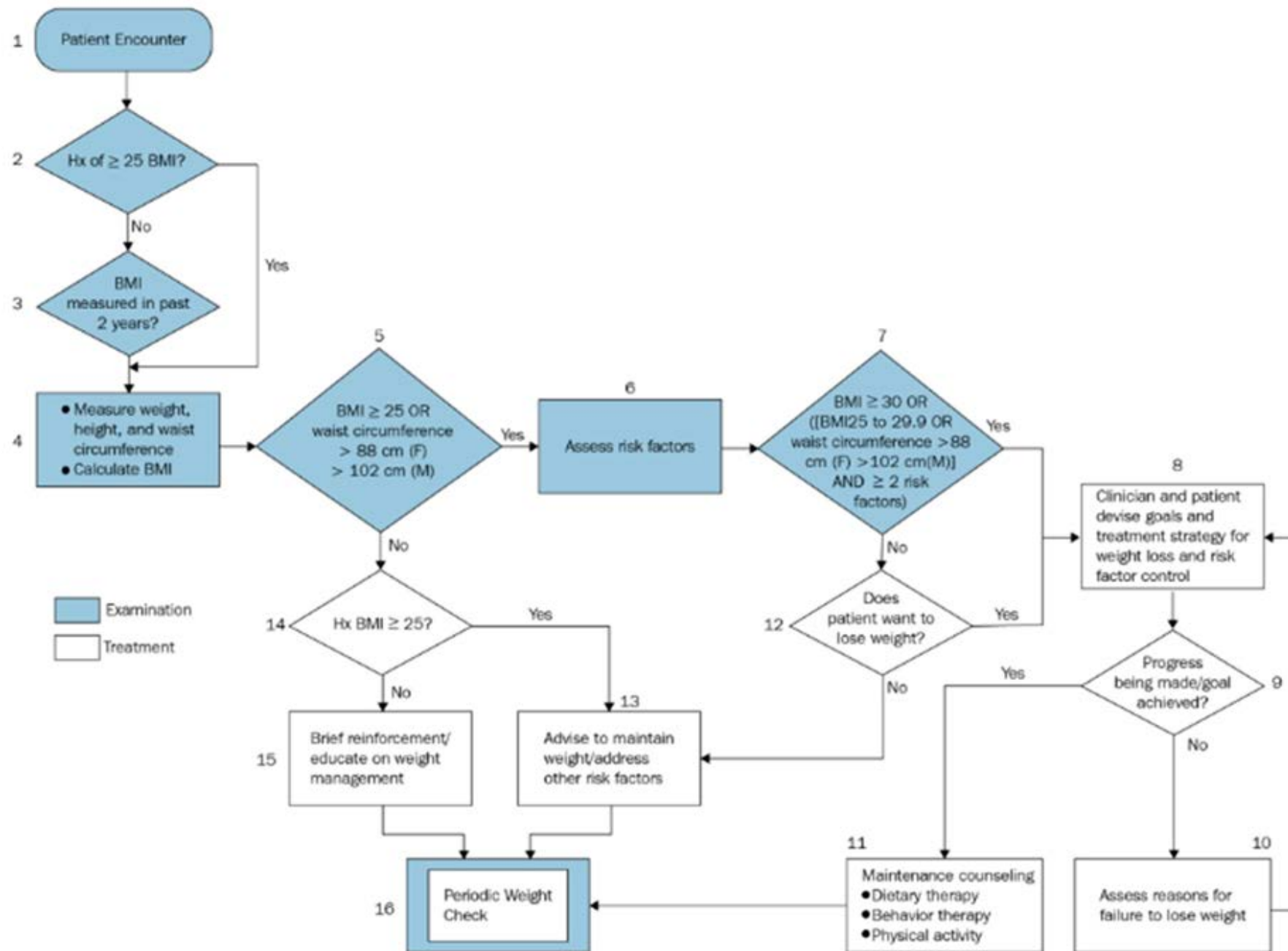
TREE-LIKE REASONING

We are all familiar with tree-like reasoning; it involves branching into various possible actions as different decision points are met. Here we look at tree-like reasoning more closely and examine its machine learning counterparts: the decision tree and the random forest.

CATEGORICAL REASONING WITH ALGORITHMS AND TREES

In one medical decision making paradigm, the clinical problem can be approached as a tree or an algorithm. Here, an algorithm does not refer to a "machine learning algorithm" in the computer science sense; it can be thought of as a structured, ordered set of rules to reach a decision. In this type of reasoning, the root of the tree represents the initiation of the patient encounter. As the physician learns more information while asking questions, they come to various branch or decision points where the physician can proceed in more than one route. These routes represent different clinical tests or alternate lines of questioning. The physician will repeatedly make decisions and pick the next branch, reaching a terminal node at which there are no more branches. The terminal node represents a definitive diagnosis or a treatment plan.

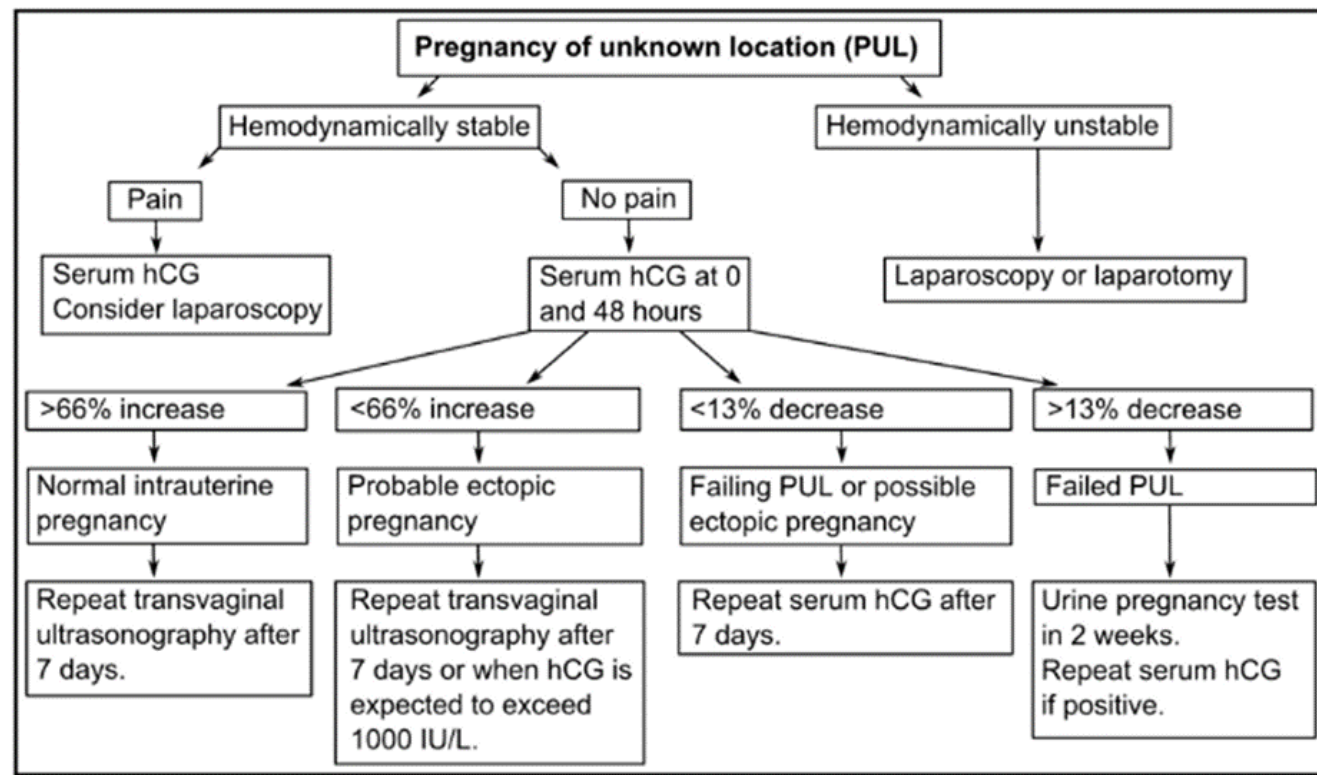




For example, suppose we have a female patient with several clinical variables that are measured:

BMI = 27, waist circumference = 90 cm, and the number of cardiac risk factors = 3.

Starting at node #1, we skip from Node #2 directly to Node #4, since the BMI > 25. At Node #5, again the answer is "Yes." At Node #7, again the answer is "Yes," taking us to the management plan outlined in Node #8



A second example of an algorithm that combines both diagnosis and treatment is shown as follows. In this algorithm for the diagnosis/treatment of pregnancy of an unknown location, a hemodynamically stable patient with no pain (a patient with stable heart and blood vessel function) is routed to have serum hCG drawn at **0 and 48 hours** after presenting to the physician. Depending on the results, several possible diagnoses are given, along with corresponding management plans.

Algorithms have a number of advantages. For one, they model human diagnostic reasoning as sequences of hierarchical decisions or determinations. Also, their goal is to eliminate uncertainty by forcing the caretaker to provide a binary answer at each decision point. Algorithms have been shown to improve standardization of care in medical practice and are in widespread use for many medical conditions today not only in outpatient/inpatient practice but also prior to hospital arrival by emergency medical technicians (EMTs).

However, algorithms are often overly simplistic and don't consider the fact that medical symptoms, findings, or test results may not indicate 100% certainty. They are insufficient when multiple pieces of evidence must be weighed for arriving at a decision.

PROBABILISTIC REASONING AND BAYES THEOREM

A second, more mathematical way of approaching the patient involves initializing the baseline probability of a disease for a patient and updating the probability of the disease with every new clinical finding discovered about the patient. The probability is updated using Bayes theorem.

Using Bayes theorem for calculating clinical probabilities

Briefly, Bayes theorem allows for the calculation of the post-test probability of a disease, given a pretest probability of disease, a test result, and the 2x2 contingency table of the test. In this context, a "test" result does not have to be a lab test; it can be the presence or absence of any clinical finding as ascertained during the history and physical examination. For example, the presence of chest pain, whether the chest pain is substernal, the result of an exercise stress test, and the troponin result all qualify as clinical findings upon which post-test probabilities can be calculated. Although Bayes theorem can be extended to include continuously valued results, it is most convenient to binarize the test result before calculating the probabilities.

To illustrate the use of Bayes theorem, let's pretend you are a primary care physician and that a 55-year-old patient approaches you and says, "I'm having chest pain." When you hear the words "chest pain," the first life-threatening condition you are concerned about is a myocardial infarction. You can ask the question, "What is the likelihood that this patient is having a myocardial infarction?" In this case, the presence or absence of chest pain is the test (which is positive in this patient), and the presence or absence of myocardial infarction is what we're trying to calculate.

Calculating the baseline MI probability

To calculate the probability that the chest-pain patient is having a myocardial infarction (MI), we must know three things:

The pretest probability

The 2x2 contingency table of the clinical finding for the disease in question (MI, in this case)

The result of this test (in this case, the patient is positive for chest pain)

Calculating the post-test probability of MI given the presence of chest pain

Now that we have LR+, we multiply it by the pretest probability to get the post-test probability:

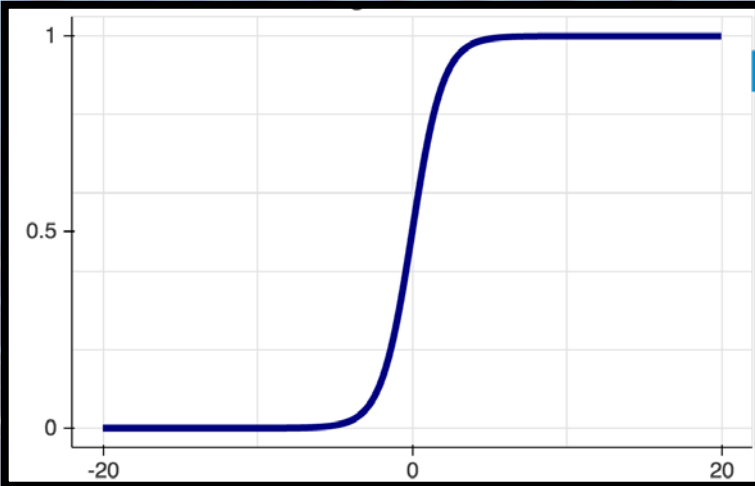
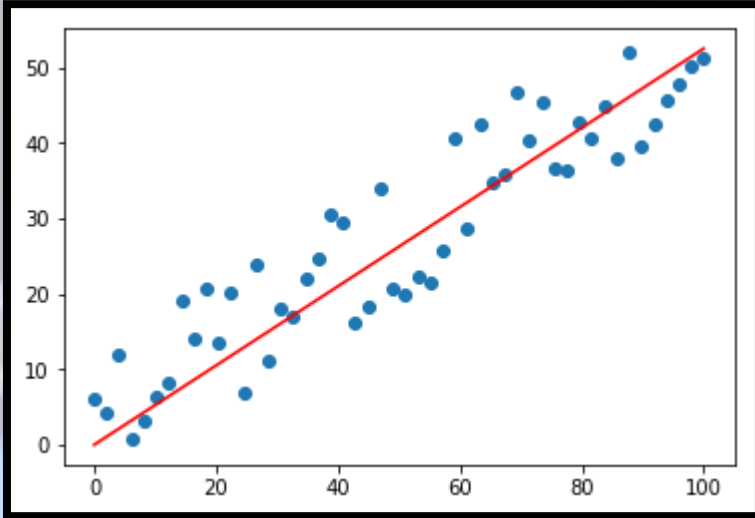
$$\text{Post-Test Probability} = 0.05 \times 2.85 = 14.3\%.$$

CRITERION TABLES

The use of criterion tables is partially motivated by an additional shortcoming of Bayes theorem: its sequential nature of considering each finding one at a time. Sometimes, it is more convenient to consider many factors simultaneously while considering diseases. What if we imagined the diagnosis of a certain disease as an additive sum of select factors? That is, in the MI example, the patient receives a point for having positive chest pain, a point for having a history of a positive stress test, and so on. We could establish a threshold for a point total that gives a positive diagnosis of MI. Because some factors are more important than others, we could use a weighted sum, in which each factor is multiplied by an importance factor before adding. For example, the presence of chest pain may be worth three points, and a history of a positive stress test may be worth five points. This is how criterion tables work.

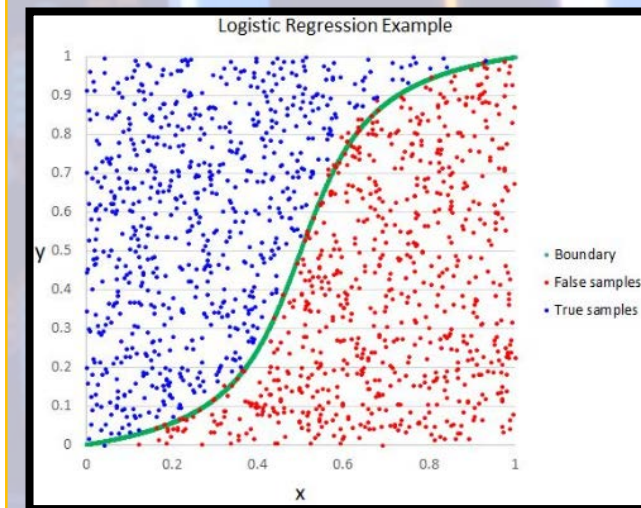
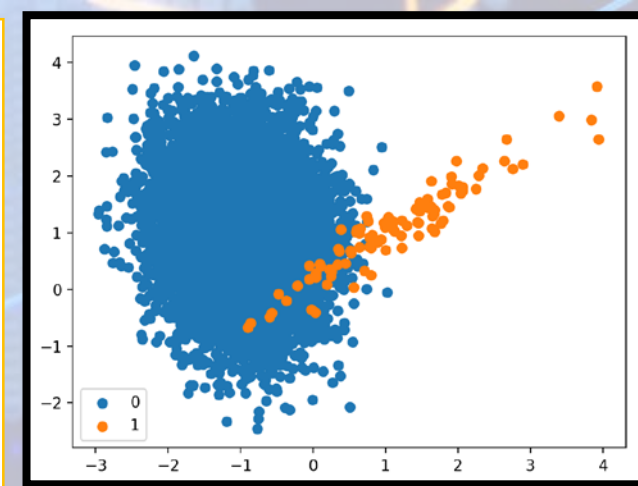
CLINICAL FINDING	SCORE
CLINICAL SYMPTOMS OF DEEP VEIN THROMBOSIS LEG SWELLING, PAIN WITH PALPATION	3.0
ALTERNATIVE DIAGNOSIS IS LESS LIKELY THAN PULMONARY EMBOLISM	3.0
HEART RATE 100 BEATS PER MINUTE	1.5
IMMOBILIZATION LOR 3 DAYS OR SUREERV IN THE PREVIOUS 4 WEEKS	1.5
PREVIOUS DIAGNOSIS OF DEEP VEIN THROMBOSIS PULMONARV EMBOLISM	1.5
HEMOPTYSIS	1.0
PATIENT HAS CANCER	1.0
RISK STRATIFICATION	
LOW RISK FOR PE	2.0
MEDIUM RISK FOR PE	2.0 – 6.0
HIGH RISK FOR PE	6.0

CORRESPONDING MACHINE LEARNING ALGORITHMS - LINEAR AND LOGISTIC REGRESSION



- Notice that a criterion table tends to use nice, whole numbers that are easy to add. Obviously, this is so the criteria are convenient for physicians to use while seeing patients. What would happen if we could somehow determine the optimal point values for each factor, as well as the optimal threshold? Remarkably, the machine learning method called logistic regression does just this.
- Logistic regression is a popular statistical machine learning algorithm that is commonly used for binary classification tasks. It is a type of model known as a generalized linear model.
- Logistic regression is like linear regression, except that it applies a transformation to the output variable that limits its range to be between 0 and 1. Therefore, it is well-suited to model probabilities of a positive response in classification tasks, since probabilities must also be between 0 and 1.

- Logistic regression has many practical advantages. First of all, it is an intuitively simple model that is easy to understand and explain. Understanding its mechanics does not require much advanced mathematics beyond high school statistics, and can easily be explained to both technical and nontechnical stakeholders on a project.
- Second, logistic regression is not computationally intensive, in terms of time or memory. The coefficients are simply a collection of numbers that is as long as the list of predictors, and its determination only involves several matrix multiplications (see the preceding second equation for an example). One caveat to this is that the matrices may be quite large when dealing with very large datasets (for example, billions of data points), but this is true of most machine learning models.
- Third, logistic regression does not require much preprocessing (for example, centering or scaling) of the variables (although transformations that move predictors toward a normal distribution can increase performance). As long as the variables are in a numeric format, that is enough to get started with logistic regression.
- Finally, logistic regression, especially when coupled with regularization techniques such as lasso regularization, can have reasonably strong performance in making predictions.



COMPLEX CLINICAL REASONING

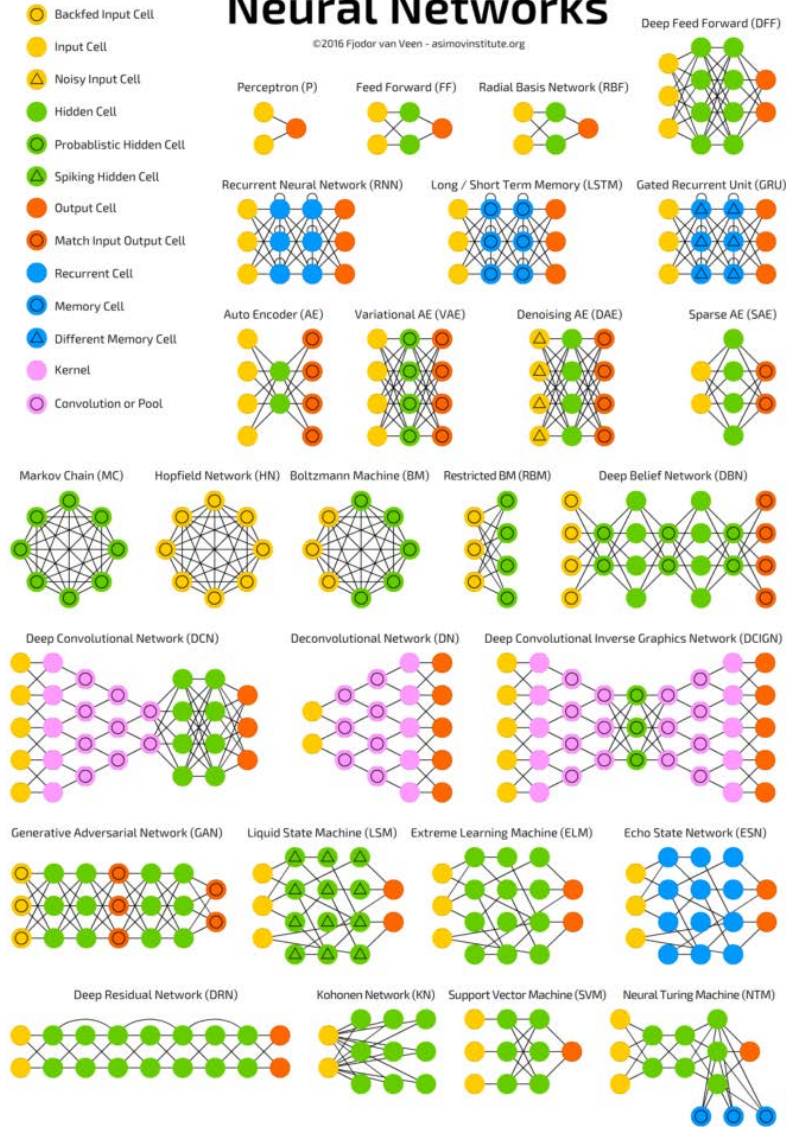
- Imagine that an elderly patient complaining of chest pain sees a highly experienced physician. Slowly, the clinician asks the appropriate questions and gets a representation of the patient as determined by the features of that patient's signs and symptoms. The patient says they have a history of high blood pressure but no other cardiac risk factors. The chest pain varies in intensity with the heartbeat (also known as pleuritic chest pain). The patient also reports they just came back to the United States from Europe. They also complain of swelling in the calf muscle. Slowly, the physician combines these lower level pieces of information (the absence of cardiac risk factors, the pleuritic chest pain, the prolonged period of immobility, a positive Homan's sign) and integrates it with memories of previous patients and the physician's own extensive knowledge to build a higher level view of this patient and realizes that the patient is having a pulmonary embolism. The physician orders a V/Q scan and proceeds to save the patient's life.
- Such stories happen every day across the globe in medical clinics, hospitals, and emergency departments. Physicians use information from the patient history, exam, and test results to compose higher level understandings of their patients. How do they do it? The answer may lie in neural networks and deep learning.



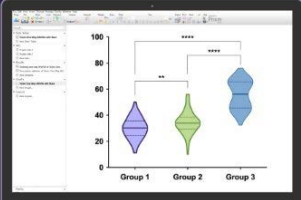
CORRESPONDING MACHINE LEARNING ALGORITHM - NEURAL NETWORKS AND DEEP LEARNING

A mostly complete chart of Neural Networks

©2016 Fjodor van Veen - asimovinstitute.org



A neural network is modeled after the nervous system of mammals, in which predictor variables are connected to sequential layers of artificial "neurons" that aggregate and sum weighted inputs before sending their nonlinearly transformed outputs to the next layer. In this fashion, the data may pass through several layers before ultimately producing an outcome variable that indicates the likelihood of the target value is positive. The weights are usually trained by using the backpropagation technique, in which the negative difference between the correct output and predicted output is added to the weights at each iteration.



GraphPad Prism

- The graphing and statistical analysis tool by GraphPad is one of the most popular tools in this review. Since its original development for biologists in both academic and industry they have stood on their strength like doing understandable statistics, and retracing every analysis. The software helps the user perform basic statistical methods needed in laboratory researchers and clinicians like t tests, nonparametric comparisons, one- and two-way ANOVA, analysis of contingency tables, and survival analysis.
- The best part of the software is its interpreted result analysis page that is prepared at the end of the analysis. The language used is very simple and straightforward without the use of technical jargons of statistics that a researcher looking at biological interpretations is not much concerned about. The descriptives and assumptions made during the analysis are explained during detail post-analysis. The events that have led to the findings of the analysis can be clearly back-traced with their order retained. The data that has been lost or unused in the analysis is well located in the same table used for data analysis represented by asterisked blue italics. The data can be annotated and then can be shared through LabArchives, a cloud-based laboratory notebook.
- The software allows automation without programming by retaining the graph(s) or entire data analysis by just cloning the resident file or graph. This also adds to its pertaining behaviour of automatic re-analysis of data in cases where any of the data points are altered, all on runtime without any need to redo the analysis performed or graph drawn. The live-update and graph cloning save a lot of precious time by avoiding repetition of the routine every time the same analysis is done.
- The Prism is not a full-fledged statistical program but its application in biology can be claimed to be near being complete. Given its ease of use, and the knowledge of statistics amongst biologists there is perfect complementation by this software to the community it serves, thereby claiming the majority of share in both data analysis and statistical data analysis. The graphing capability is of equal potential as the statistical analysis and just remains the reason for its dominance.
- The software also has the ability to import raw files in basic formats like csv, txt, more popular xls and more of a modern-day standard XML file, apart from its own pzf format. Clairfeuille T et al. analyzed electrophysiological data through Prism 7, in addition to Excel and Origin Pro, to investigate the fast inactivation of voltage-gated sodium channels. GraphPad Prism was used to perform statistical analysis to study the effect of sleep deprivation on tau and amyloid beta accumulation, the cognitive deficits of Alzheimer's disease, the BMP-dependent cardiac contraction, the role of the mGluR5-Erk pathway in tuberous sclerosis complex, the importance of TCR signaling to natural killer T cell development], the role of Lsd1 during blood cell maturation, the role of miR-146a in mouse hematopoietic stem cells, investigate mitochondrial division in yeast, the contributions of mast cells to dengue virus-induced vascular leakage, the synaptic transmission and cognitive function impairment caused by reduction of the cholesterol sensor SCAP in the brains of mice, and the modulation effect of USF1 in molecular and behavioral circadian rhythms in mammals, among others.



MS Excel

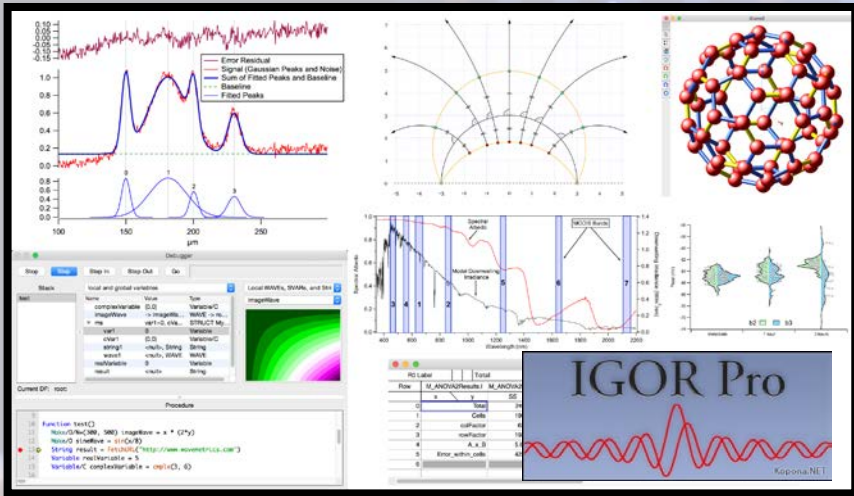
- Without any need of introduction, Microsoft Corp Excel is used widely in statistical analysis per the dataset taken for this review. The program has a wider reach and knowledge of use is quite widespread that the amount of unknown is very less about the way-of-use and thus the ease-of-use reaches the highest among the reviewed software. The program has functions that perform simple and complex mathematical and statistical functions one at a time. The syntax of writing any function is well accompanied by the user-friendly highlighted help alongside the function to be used. It is intuitive as it was the first graphical user interfaced data analysis software that was introduced in the world of computers in the 1990s. The interpretations are not descriptive and so the user needs to obtain the inferences depending on his level of knowledge of statistics. The data representation is not all-encompassing as like any other graphing software with a simple example of the inability to plot a two-sided y-axis graph.
- Although the Excel program is likely not be cited explicitly in articles (thus its actual number of citations is likely to be much higher); it is cited, as examples, among studies to investigate the functional property of the CK2 kinase in *Drosophila*, the molecular mechanism of organismal death in nematodes, the effect of JAK2 mutations on hematopoietic stem cells, mitochondrial division in yeast, the contributions of mast cells to dengue virus-induced vascular leakage, the role of cMyBP-C in the process of cardiac muscle contraction, the effect of colitis on tumorigenesis, the presence of Lgr5+ stem cells in mouse intestinal adenomas, the regulation of fine motor coordination by AMPAR in BG cells, and the regulation of meiotic non-crossover recombination by FANCM ortholog.



SAS

SAS is the largest purveyor of advanced analytics, and its statistical software is used in a diverse array of scientific and engineering enterprises and organizations. SAS software was used to perform statistical analyses to study the effect of CO₂ enhancement on organic carbon decomposition, the influence of genetic polymorphisms on complex traits and fitness in plants, the regulation of genetic traits by IGF pathway, the regulation of oogenic processes by MARF1, the molecular mechanism of novelty seeking in honey bees, the effect of nonrandom pollinator movement on reinforcing selection, the promotion of beneficial heart growth by fatty acids from Burmese python, the regulation of wing pattern evolution in butterfly by optix gene, and the coevolution with pathogen in *C. elegans*.

Wavemetrics Igor Pro

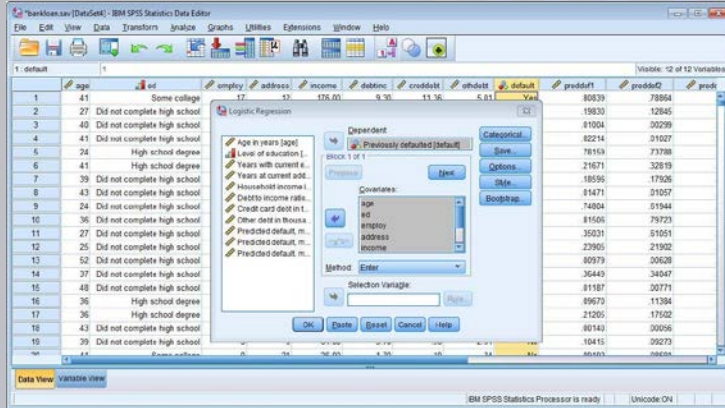


Wavemetrics Igor, a scientific and technical data analysis program, has multiple capabilities, such as data processing, statistical analysis, image processing and analysis, graphing, and 3D and volume visualization. The program has been cited in studies including electrophysiology, analytical ultracentrifugation, high-speed atomic force microscopic observations, imaging, calcium imaging, and X-ray emission spectroscopy.

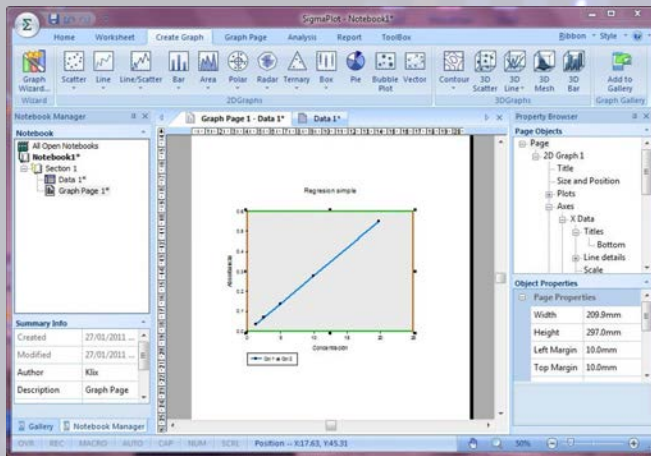


OriginLab publishes Origin, a scientific graphing and data analysis software. It has been cited in studies including fluorescence intensity analysis, FRET-FLIM measurement, electrophysiological analysis, biochemical assays, among other data analyses.

IBM SPSS

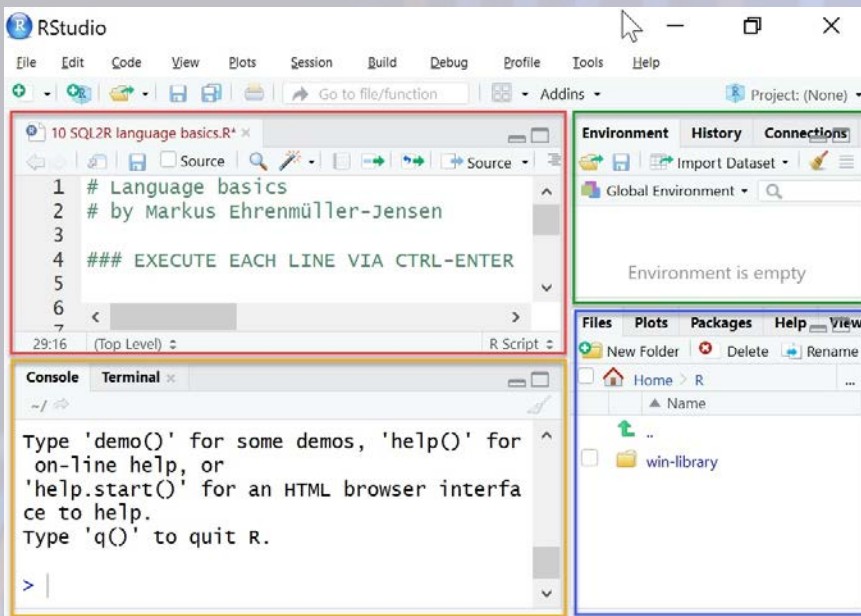


- The most comprehensive of the statistical tools available among those listed in the review, SPSS is a cross-disciplinary tool (biology, statistics, social sciences, etc.) with equal depth. The pre-requisite knowledge of statistics is needed to understand the software as in most cases the input needs to be defined appropriately. The methodology to be used for analysis is at times understood by the software but in most cases to be defined specifically and not that direct for a novice in statistics.
- This business analytics software can handle large scale data with ease, unlike any other software. The software can help a researcher at various level of analytical process like planning, data collecting, data access, data management and preparation, analysis, reporting and deployment.
- A very large set of possibilities can be enlisted in each step, for example, in data preparation, in case of a very large dataset, what can be a nightmare with other softwares can be a thoroughfare as it happens to sort a dataset based on unique properties which cannot be easily enlisted by visual glance or any other software. Herein SPSS can not just sort the dataset based on unique features but obtain characteristics like count on the number of samples with the feature and easily tabulate the information. The use and the immense potential of SPSS can be felt in particular if the dataset is very large. The amount of computational time it takes is very short even for a huge scale of data. SPSS has many modules apart from those that assist in specific set of calculations like bootstrapping to test stability and reliability of models, data categorisation for complex data and high-dimensional data, performing exact tests with both categorical and non-parametric data problems on small and large datasets, forecasting to predict trends and develop forecasts quick and easy, neural-network based methods that can related complex relationships in the data and regression methods to predict categorical outcomes and apply non-linear regression procedures.
- The extensive capabilities in the broad spectrum of statistics is the greatest utilitarian benefit of IBM SPSS given its capacity in handling very large datasets. The expected medium level statistical knowledge of the user to unravel its true potential as a statistical software is its bottleneck.
- SPSS software has been cited in multiple studies for statistical analysis.



SigmaPlot

- The plotting software SigmaPlot from Systat Software Inc. is a full fledged graphical representation software. It can import data ready for representation in a graphical representation in numerous formats from plain ASCII files, txt, csv, excel, MS Access, SigmaScan, ODBC complaint databases, run SQL queries on tables, Tablecurve 2D, 3D and import the data in all formats as well as graphical formats like BMP, JPEG, GIF, PNG, HTML, TIFF, PDF, PSD, EPS, etc.
- SigmaPlot can save templates so as to reduce the process of repeating format and other customizations made to a graph if such a type of graph has to be created with some other dataset. The built-in macro-language interface allows visual basic compatible programming, macro-recorder to save and play-back. One can export the graphs directly to powerpoint ready for presentation and also insert into a word file under progress to accommodate the analyzed graph.
- The program allows all sorts of symbols relevant for scientific representation to be incorporated in the graphs and the Report Editor is user-friendly for incorporation of reports into documents like Word, Excel via cut, copy and paste format. The report editor can be directly transferred to HTML page and shared online without any knowledge of web programming. SigmaPlot has plugins compatible with Excel thereby complementing and enhancing the Excel graphing tools. Tools like curve-fitting and other mathematical functions from the built-in library enhance the quality of data representation via SigmaPlot.
- In short, SigmaPlot is an excellent plotting, graphing tool as its name suggests but the value addition is very little using its built-in functions.



R programming language

R, a free statistics and visualization programming language, is gaining popularity among researchers in recent years. It has over 6000 packages, contributed by volunteers, covering a wide range of disciplines from molecular biology, phylogeny, to stock markets. Bioconductor is a software used for analyzing of high-throughput genomic data, with 934 packages as of Dec 2014, such as edgeR, BioMartRt, and Minfi. Among the articles surveyed, R version 2.15.2 was used to perform statistical analysis to study how the hippocampus played its dual role in spatial navigation and episodic memory and version 2.9.2 was used to study the mechanism by which CaMV perceives aphid vectors. Chopra S et al generated volcano plots of prostaglandin lipidomics analysis in R studio with the bioconductor limma package. Litke JL et al used R 3.4.1 and RStudio 1.0.157 for statistical analysis and data plotting.