The Role of Mathematical Modeling in Medical Research: "Research Without Patients?"

Richard B. Chambers, MSPH

Outcomes Assessment Department, Alton Ochsner Medical Foundation, New Orleans, LA

Computer controlled mathematical models of medical outcomes are commonly found in the current medical literature. What is less common is an understanding of the methods used to construct such models, leaving the consumers of medical research to accept the interpretations as presented. A basic knowledge of the concepts used to generate models will provide the clinician with the insight needed to critically evaluate medical literature based on mathematical models.

Chambers RB. The role of mathematical modeling in medical research: "Research without patients?" The Ochsner Journal 2000; 2: 218-223.

The development of computerized mathematical models used to simulate medical outcomes is a growing area of specialization (1-6). A current MEDLINE search of articles using mathematical models yielded 43,764 articles dating from 1966. The majority (97%) of the manuscripts including mathematical models were published only since 1990. Since 1999, 9219 articles were published. That is 21% of the medical manuscripts using mathematical modeling over the last 35 years published in only this last year.

Clinicians and administrators are accepting the conclusions drawn from modeling, often without realizing the data are simulated. I have often been asked to comment on a journal article only to realize well into the critique that my clinical colleague did not know that the tables, charts, and figures were referring to computer generated cases. The surprise was best phrased by the question, "Do you mean that we can do research without patients?" The answer is, "Yes and no."

The "yes" part of the answer hinges on the soundness of the methodologies employed. Regression methods, the most commonly seen in modeling, use some variation of the classical linear model, y=mx+b, according to a transformation or derivation that plots a math function closely describing the data (1,7,8). This is not new to biometrics, but the historical use is to compare two groups by the parameters of their lines from measured observations. Using the derived regression to predict outcomes in individuals from the same population has always been an accepted application of regression. Making the jump to using mathematical modeling to generate simulated patient populations, and even model their outcomes for therapies of the future, is a more difficult stretch.

The "no" part of the answer is rooted in the skepticism to believe something that was not only not seen and measured by the reader, but was also not seen or measured by the math modelers. Stopping here, however, can deprive the reader of the benefits of mathematical modeling. Some problems simply cannot be solved with a single math function or formula (8). One solution is to repeat trial and error tests, possibly over many lifetimes. Another is to simulate the process in a computer model. The keys to the validity of modeling are the known dependent probabilities, associated variances, and coefficients determining the relative significance of each factor to the model (1,7,8). This means that a model must be based on sound research, with actual data that are widely accepted as valid by the medical science community.

Mathematical modeling is presented by various names like predictive modeling, simulation, or decision analysis. By far the most common methodology is the Markov Chain Monte Carlo simulation. The two parts of this method each have their own Mesh headings on MEDLINE, and together they have evolved into the acronym MCMC (pronounced "mac-mac"). Understanding the process of a MCMC simulation can go far in making one a better consumer of mathematical modeling because it contains the elements basic to modeling by any other name (8).

Table 1. The probability of drawing a red ball for draws 1-5, 10, 100, and 1000 for the first 10 individual simulations of the Markov Chain example. The values used to describe the possible outcomes in the text are in boldface.

Markov chain	Probability of choosing a red ball to draw number: 1	2	3	4	5	10	100	1000
1	0.00	0.50	0.00	**0.50**	1.00	0.50	0.50	0.50
2	0.00	0.50	0.00	0.50	0.00	0.50	0.50	0.50
3	0.00	0.00	0.50	**1.00**	0.50	0.00	0.00	1.00
4	0.00	0.50	0.00	0.50	0.00	0.50	0.50	0.50
5	0.00	0.00	0.50	**0.00**	0.50	1.00	0.00	0.00
6	0.00	0.00	0.50	1.00	0.50	0.00	0.00	1.00
7	0.00	0.50	0.50	1.00	0.50	1.00	0.00	0.00
8	0.00	0.50	0.50	1.00	0.50	1.00	0.00	1.00
9	0.00	0.50	0.00	0.00	0.50	1.00	1.00	1.00
10	0.00	0.50	0.00	0.50	0.00	1.00	0.00	1.00

Markov Chain

A Markov Chain, first used in the 1940s to model nuclear reactions, is simply a series of conditional probabilities in a fixed, dependent order (1). Used by physicists for this limited application, this technique went unknown to the statistical community until the 1970s when it was generalized to any application for which one could not derive a single probability function (1). The first practical applications appeared in the 1980s in the fields of neuroscience (1) and economics (7).

The classical example used to teach Markov Chain theory is the random draw of one of two balls from a bag with replacement. We begin with unpainted balls. When an unpainted ball is drawn, a coin is flipped to decide to paint it red or black. The ball is painted and put back in the bag. When a red ball is drawn, it is painted black; when a black ball is drawn, it is painted red. Because the same individual ball can be drawn sequentially, it is not possible to derive a probability function to predict the probability of drawing a red ball from the bag at any given trial. Since there are only two balls, there are three possible probabilities of drawing a red ball at any one draw. One could be drawing from two black, two red, or one of each color at any given draw. The possibility at each draw depends upon the entire sequence of events, from the first draw to the draw in question. Every time the experiment is repeated, the n^{th} trial can present a different possibility. There is the additional possibility of sequentially drawing the same individual ball in runs of varying length during the experiment. Convergence is achieved when the model looses perceivable dependence on the starting point (9). The time prior to convergence is referred to as a "burn in" period. Runs of sequential draws of the same individual ball have more effect on the chain of events during the burn in period.

A computer simulation was written to illustrate this example (Table 1). In the first Markov Chain, the probability of drawing a red ball on the draw 4 is 0.50 because after draw 3 there was one red and one black ball in the bag. In the third Markov Chain, the probability of drawing a red ball on the draw 4 is 1.00, and in the fifth Markov-chain the same request has a probability of 0.00.

The problem of calculating event probabilities in this classical example is used because only modeling can solve it; a global math function or formula is not possible, although processes that can be solved by a global math function can also be modeled. Usually, to solve a math problem, one would prefer to solve the single function, but to test a process or the effect of sequential occurrences at the extremes of the known variances, it is often useful to develop a model. This is where regression methods are employed. While regression functions from least-squares methods are used, it is more common to see Bayesian methods. Bayesian-derived coefficients are encountered when outcomes are expressed as probabilities, such as logistic or probit regression (8).

Monte Carlo

Monte Carlo simulation came into useful application in the same era as Markov-Chain processes (1, 7). Monte Carlo simulation is a series of random draws, simulating an event within the known parameters of the probability distribution of the event (1, 7). The name originated with early developers of the method who used a roulette wheel to generate random numbers. The wheel generated a gambling atmosphere as well, inspiring remembrance of the famous Monaco city, Monte Carlo (10).

To illustrate the analytical synergy of these methods, and why the combination is more common than the individual parts alone, let's expand the random ball draw example into a Monte Carlo simulation. The same software that ran the individual experiments for the previous example was modified to loop 10,000 times and write the resulting probabilities of drawing a red ball on draws 1-5, 10, 100, and 1000. Since the three possible probabilities are 0.0, 0.5, and 1.0, we know that the average probability will migrate toward 0.5, and that the error term will reach equilibrium. We do not know, however, at which draws in the model the mean will reach 0.5 or the error term will stabilize.

Table 2. Simple descriptive statistics of the probability of drawing a red ball for draws 1-5, 10, 100, 1000 for the Monte Carlo example of n=10,000.

Draw Number	Mean	Standard Deviation	Median
1	0.00	0.00	0.00
2	0.25	0.25	0.00
3	0.37	0.33	0.50
4	0.44	0.35	0.50
5	0.47	0.35	0.50
10	0.50	0.36	0.50
100	0.50	0.36	0.50
1000	0.50	0.36	0.50

As shown in Table 2, the mean probability reached 0.5 by the tenth draw, and the standard deviation (used for the error term) stabilized at 0.36. Further analysis will be insightful because there are two extremes by which these parameters can be reached. First, most values can be very near the mean, minimizing the error term. Second, half of the values can equal the minimum value and half the maximum value. This gives an error term of maximum magnitude, or a "worst case scenario" error term. Comparing the frequencies of the possible probabilities between various depths into the chain can reveal when the model stabilizes. For this evaluation, two nonparametric analyses were run. A sign test, to pin down when the possible outcomes stabilized symmetrically around 0.50, and the Kolmogorov-Smirnov Test, a non-parametric analysis of variance (11). Frequency table analysis using either the Pearson or Cochran X^2 tests is possible, and would yield confirming results, but the presentation is clearer with the chosen tests.

As shown in Table 3, frequencies of possible outcomes between draws 2-5 and every other draw are statistically dissimilar. Draws 10, 100, and 10,000 are statistically similar. That is, the model has reached convergence at the tenth draw in the chain, as proven by the stability of the model through the 1000th draw in the simulation. Once convergence is achieved, the model is assumed to be stable and can be used for the intended purpose (8,9).

Putting It All Together

A simple medical model will put all of the concepts together. This is a simplified example to illustrate the concepts; modeling a specific medical process or outcome requires the inclusion of many cofactors that make interpretation and presentation more difficult. First, a theoretical rare disease, D, is always fatal if untreated within a short time after onset (this allows us to skip time-dependent covariates). The first known treatment, A, has a success rate of 0.40 from a study of 200 subjects. Treatment B was recently tested against A in a trial of the same size that confirmed the success rate of A and established the success rate of B at 0.50. Figure 1 shows the outcomes of the two groups and the significant p-value of 0.04. Is this enough information to decide to make treatment B the treatment of choice? Remember, disease D is a rare event and these studies took years and big budgets to coordinate nationwide data collection. No additional outcome studies will be around any time soon. Furthermore, from only 200 subjects the success rates of 0.40 and 0.50 have 95% confidence intervals (0.33-0.47 and 0.43-0.57, respectively). The 95% confidence interval is the range in which we expect to find the success rates for 95 of the next 100 studies of the same size (11). The observed overlap may lead one to suspect that null studies are certainly possible as well as studies reaching the opposite conclusion.

The one clinical trial can be modeled into many trials and give estimates of the range of outcomes (which we suspect when we observe the confidence intervals) and the likelihood of reaching the same conclusion in successive clinical trials. The simulation can assign the outcome four different ways by two different divisions. The first division is the object of the simulation, which can be either individuals or populations. A model based on individuals is a more lifelike simulation and will yield more variation. Population-based models might be better suited for public health or community medicine applications. From our example, treatment group A can have a success rate of 0.40, but any patient in treatment group A cannot have a success rate of 0.40. Each patient's success rate must be either 0 or 1. The assignment of the 0 or 1 is the next division of strategy. If the confidence interval or other measure of variance is not known, each unit can have a random number between 0 and 1 generated and tested against the known rate (0.40 for group A in our example). If the random number for a unit is below 0.40, the unit is a success. If the random number is above 0.40, the unit is a failure. Since the error term is not known, this method is said to be a parameter estimation method and is called Gibbs sampling (8,12). (The name comes from the first use of this strategy in pixel imaging where the Gibbs probability distribution is used [8].) In assigning a binary outcome, this method yields a larger error term, but in applications where the error term is already at an extreme of small or large, this is not a concern. It is the only choice when the error term is unknown or a parameter such as sample size or a highly disputed denominator (as occurs in national databases or national surveillance) is encountered. In our example, the error terms are known, so, rather than compare the generated random number to 0.40, it can be compared to a rate randomly selected from the range of the confidence interval. In this most

Table 3. Statistical comparison of frequencies of possible outcomes between draws. P-values greater than 0.1 lead to the conclusion that the frequencies of possible outcomes are not different between these draws.

Comparing Draw	To Draw	Sign Test p-values	Kolmogorov-Smirnov Test p-values	Wilcoxon Signed Ranks Test p-values
2	3	<0.0001	<0.0001	<0.0001
2	4	<0.0001	<0.0001	<0.0001
2	5	<0.0001	<0.0001	<0.0001
2	10	<0.0001	<0.0001	<0.0001
2	100	<0.0001	<0.0001	<0.0001
2	1000	<0.0001	<0.0001	<0.0001
3	4	<0.0001	<0.0001	<0.0001
3	5	<0.0001	<0.0001	<0.0001
3	10	<0.0001	<0.0001	<0.0001
3	100	<0.0001	<0.0001	<0.0001
3	1000	<0.0001	<0.0001	<0.0001
4	5	<0.0001	<0.0001	<0.0001
4	10	<0.0001	<0.0001	<0.0001
4	100	<0.0001	<0.0001	<0.0001
4	1000	<0.0001	<0.0001	<0.0001
5	10	<0.0001	<0.0001	<0.0001
5	100	<0.0001	<0.0001	<0.0001
5	1000	<0.0001	<0.0001	<0.0001
10	100	0.4230	0.9991	0.3150
10	1000	0.4165	0.9999	0.3081
100	1000	0.9999	0.9999	0.9999

realistic simulation, we are comparing a randomly generated number between 0 and 1 to a randomly generated number between the lower and upper bounds of the 95% confidence interval to decide whether the case is a success or a failure. In the case of population-based simulation, the population's success rate is drawn from the range of the confidence interval.

The results of our MCMC simulation are summarized in Table 4. Of 100 simulations of the known clinical trial, 91 observed that treatment B had a better success rate than treatment A. Of the simulations with observed outcomes confirming the clinical trial, only 50 (55%) had significant p-values (mean = 0.21, standard deviation = 0.29). Comparing the studies demonstrating the superiority of treatment B with studies refuting that finding for statistical significance with a Fisher's exact test gives a p-value of 0.0013 (Figure 2). The interpretation is that a study refuting the superiority of treatment B is less likely to be statistically significant. The model gives us reason to not be so confident about the

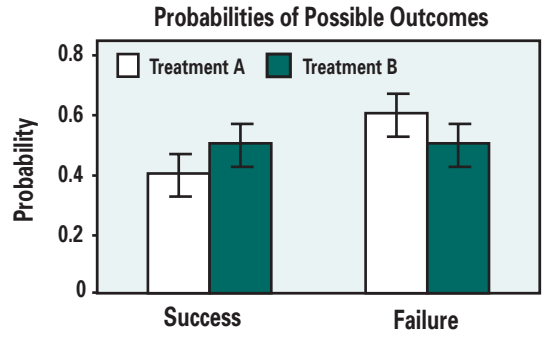

Figure 1. A comparison of hypothetical treatments, of sample size 200, for a hypothetical disease for the medical model example. The chi-square test (X^2) =4.04 with 1 degree of freedom yields a p-value of 0.04. The risk ratio of treatment failure in treatment A relative to B is 1.2 with a 95% confidence interval of 1.03 to 1.39. The statistical inference of this single study would be that treatment B was superior to treatment A.

Table 4. Simple descriptive statistics of the medical model example comparing hypothetical treatments in a simulation of n=100 clinical trials.

	Significant	Non-Significant	Total
Confirming B>A	50	41	91
Refuting B>A	0	9	9
	Mean	Standard Deviation	Median
p-value	0.21	0.29	0.06

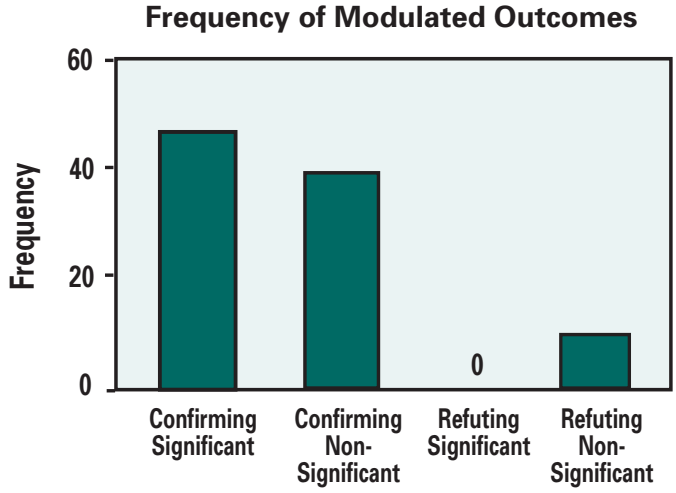

Figure 2. Status of statistical significance between simulations with confirming results to those refuting the superiority of treatment B. The p-value of the Fisher's exact test of this comparison is 0.0013. The statistical inference is that statistical significance is dependent on (more likely) confirming the superiority of B.

finding in the clinical trial. While I might personally prefer to be on treatment B if I had disease D, I can predict that if the clinical trial were repeated 100 times nine research centers out of the 100 would not want to switch from treatment A at all.

Discussion

Modeling is increasingly common in the medical literature and is more often becoming the basis for managed care policy (3-6). It is most commonly encountered for predicting outcomes, as in the simple example used in this paper (13-15). Predicting the way an intervention or a technology might drive demand on a medical system is another variant of this use (16). This is helpful when limitations like a rare event prohibit repeating actual studies or expanding research on actual patients.

A novel application for mathematical modeling is the determination of sample size requirement (17). Estimates of the population parameters can direct a simulation that increases one patient at a time until a statistically significant difference is detected between the experimental groups. A series of such simulations can give investigators a range and midpoint of a sample size that should satisfy the test of their hypothesis. This is most useful in experimental designs with categorical, nonparametric, or otherwise non-normally distributed data. In some of these circumstances there are no functions to determine sample size, and where math functions are established the simulation method usually predicts a much more reasonable sample size requirement.

Another innovative use of MCMC is estimation of missing data points (18). Most strategies to replace missing values use a point of central tendency like the mean or median. Such strategies usually have cutoff criteria for the minimum allowable proportion of missing fields to allow "filling in." Usually more

than 50% of the data for the case and the variable (throughout all cases) must be present. Such homogenous value replacement effectively reduces the variance. MCMC estimated values preserve the actual variance.

In all of the applications, math models in medicine are a reality to be faced. The impact of this, or any other, research methodology will only be to the betterment of patient care if the medical community at large undertakes a basic understanding, guiding the application of findings based on reasonable confidence after critical review. Modeling can answer otherwise unanswerable questions and greatly expand our knowledge base from actual study data. It can do all of this efficiently and risk-free, without real patients. However, all logical and mathematical components of a model must be based on valid research of actual patients.

References

1. Draper D. Bayesian Hierarchical Modeling. Department of Mathematical Sciences, Univ. of Bath, UK 2000.
2. Dawson-Saunders B, Trapp RG. Basic & Clinical Biostatistics, 2nd ed. Norwalk, CT: Appleton & Lang, 1994; 233-267.
3. Chalfin DB. Decision analysis in critical care medicine. Critical Care Clinics 1999; 15;647-661.
4. Kucey DS. Decision analysis for the surgeon. World J Surg 1999; 23:1227-1231.
5. Tom E, Schulman KA. Mathematical models in decision analysis. Infec Control Hosp Epidemiol 1997; 18:65-73.
6. Hagen MD. Decision analysis: A review. Family Med 1992; 24:349-354.
7. Mooney CZ. Monte Carlo Simulation. Series: Quantitative Applications in the Social Sciences. University of Iowa Department of Political Science, series #07-116. Newbury Park, CA: Sage Publications,1997.
8. Gamerman D. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. New York: Chapman & Hall Texts in Statistical Science, 1999.
9. Kass RE, Carlin BP, Gelman A, et al. Markov Chain Monte Carlo in practice: a roundtable discussion. American Statistician 1998, 52; 93-100.
10. History of Monte Carlo Method, Sabri Pllana. www.geocities.com/CollegePark/Quad/2435/
11. Conover WJ. Practical Nonparametric Statistics 2nd Edition. New York: Wiley 1980.
12. Albert PS, Waclawiw MA. A two-stage Markov chain for heterogeneous transitional data: a quasi-likelihood approach. Stat Med 1998; 17:1481-1493.
13. Mezzetti M, Robertson C. A hierarchical Bayesian approach to age-specific back-calculation of cancer incidence rates. Stat Med 1999; 18:919-933.
14. Knorr-Held L, Besag J. Modeling risk from a disease in time and space. Stat Med 1998; 17:2045-2060.
15. Ng ETM, Cook RJ. Modeling two-state disease processes with random effects. Lifetime Data Anal 1997; 3:315-335.
16. Dakins ME, Toll JE, Small MJ, et al. Risk-based environmental remediation: Bayesian Monte Carlo analysis and the expected value of sample information. Risk Anal 1996; 16:67-79.
17. Greenbaum IF, Fulton JK, White ED, et al. Minimum sample sizes for identifying chromosomal fragile sites from individuals: Monte Carlo estimation. Hum Genet 1997; 101:109-112.
18. Tu XM, Kowalski J, Jia G. Bayesian analysis of prevalence with covariates using simulation-based techniques: applications to HIV screening. Stat Med 1999; 18:3059-3073.

Richard Chambers is a Biostatistician in the Outcomes Assessment Department of the Alton Ochsner Medical Foundation Division of Research and the Statistical Consultant for The Ochsner Journal.