Proceedings of the 1996 Winter Simulation Conference
ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain

ANALYZING THE TEMPORAL ASSOCIATION BETWEEN HEALTH DISORDERS AND MEDICAL TREATMENTS USING PROBABILITY MODELS AND MONTE CARLO SIMULATION

Sheldon H. Jacobson

Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061-0118

Douglas J. Morrice

MSIS Department
The University of Texas at Austin
Austin, Texas 78712-1175
End to End Simulation Department
Schlumberger Austin Research Center
Austin, Texas 78720

ABSTRACT

Statistical analysis is often used in medical studies to provide evidence for an association between a treatment and a patient condition. The work in this paper is motivated by a recent medical study which analyzes the *temporal association* between a treatment (i.e., an implant procedure) and the rare autoimmune disorders, polymyositis and dermatomyositis (PM/DM). To address this association mathematically, this paper develops a probability model based on the multinomial distribution that can be used to make statistical inferences about the timing of incidences of a treatment/condition pair. This paper details an empirical study of this model using Monte Carlo simulation. It also describes some analytical results developed in subsequent research efforts. Data from the medical study illustrate the application of this model and its results.

1 INTRODUCTION

Medical studies often provide statistical data as evidence for an association between a treatment (e.g., medical procedure, drug, therapy) and a subsequent patient condition (e.g., disorder, syndrome, disease). Frequently reported statistics such as the number or proportion of a treatment condition pair are inherently *static* and fail to capture an additional dimension often found in patient data: the *timing* of the incidences of a patient condition after the treatment has been administered. The latter is referred to as *temporal association* (Cukier et al. 1993). Approaches designed to detect temporal association are potentially more powerful than static approaches in establishing connections between treatments and patient conditions. Whereas a static approach might consider the number of incidences in a given population of patients statistically insignificant, a temporal approach applied to the same data might find the timing of the incidences to be extremely rare. For example, while the total number of patients who develop a particular condition any time after receiving the treatment may be very close to the expected number, these patients may do so much sooner than expected.

This paper develops a mathematical model to study temporal association in a problem encountered in a medical study conducted by Cukier et al. (1993). Their study is designed to determine the existence of an association between the uncommon autoimmune disorders PM/DM and an implant procedure used to correct wrinkles due to aging, acne scars, and other superficial skin defects.

Other literature exists on the modeling of event timings in medical data. This research considers problems such as estimating failure time distributions, constructing regression models for censored data, and comparing survival time distributions between multiple groups of patients in clinical trials. (see, for example, Kalbfleisch and Prentice 1980 and Fleming and Harrington 1991). None of this work, however, directly addresses the temporal association problem considered in this paper.

To examine the temporal association, Cukier et al. (1993) conducted a retrospective cohort study that included approximately 345,000 patients who had received the implants over an eight year period. Of the 345,000, 9 patients developed one of the autoimmune disorders. Table 1 contains the timings of the disease diagnoses after the implants were given. A case is recorded in month 1 if it is diagnosed in the interval $(0, 1]$ months, month 2 if it is diagnosed in the interval $(1, 2]$ months, and so on. Months not included in the table had no new cases diagnosed.

Cukier et al. (1993) also construct a conditional probability distribution that reflects the temporal likelihood of a patient developing at least one of the autoimmune disorders in each one month post-implant period over a 96 month time horizon (96 months are considered because the maximum post-

Table 1: Timing of Diagnoses for the 9 Patients After the Implants Were Given

Month	1	2	4	6	7	8	25
Number of Cases	1	1	3	1	1	1	1

exposure observation period for any of the patients was eight years). This distribution is shown in Figure 1. These data reflect the likelihood of incidences of PM/DM in a patient population that grows from zero at the beginning of the study period to 345,000 patients after eight years, assuming the incidences were purely coincidental after the implants were given. Finally, Cukier et al. (1993) assess the likelihood of the incidence timings for the 9 patients who developed the disorders against the temporal probability distribution using Monte Carlo simulation.

The objective of this paper is to formulate a mathematical model to describe the problem studied in Cukier et al. (1993), and to present the results of a Monte Carlo study used to support the results in Cukier et al. (1993). The paper also describes some analytical results that extend the Monte Carlo simulation study. The remainder of the paper is organized as follows. In Section 2, the mathematical model is formulated. The model is not restricted to the particular problem studied in Cukier et al. (1993). In fact, it is general enough to study the temporal association between any patient condition and treatment. Section 3 describes the details of the Monte Carlo simulation model and the results used in Cukier et al. (1993). Section 4 describes some of the analytical results that extend this research. Section 5, summarizes the results and provides some concluding remarks.

2 MODEL FORMULATION

The problem described in Section 1 can be restated to facilitate the construction of the underlying probability model. Suppose the units in the sample are balls (rather than patients who develop the disorder) and the monthly cells on the time scale are urns. An equivalent problem formulation is that of randomly distributing N balls into Q urns where the balls act independently. The probability of any ball falling into the qth urn is u_q, $q = 1, \ldots, Q$, where $\sum_{q=1}^{Q} u_q = 1$. Let N_q denote the number of balls placed in urn q ($\sum_{q=1}^{Q} N_q = N$). The underlying probability distribution for the $\{N_q, q = 1, \ldots, Q\}$ is the multinomial distribution, (see Johnson and Kotz 1977, page 108).

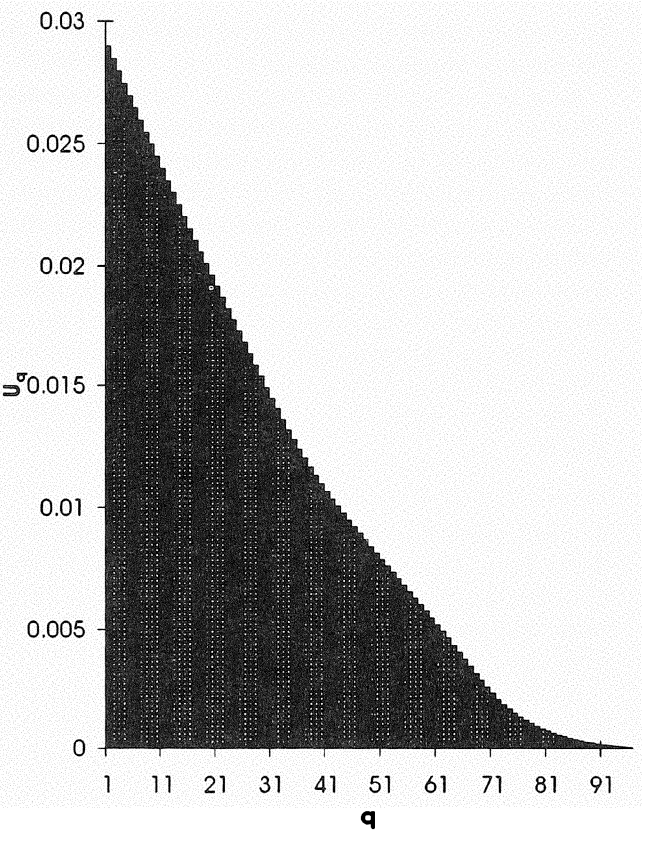

Figure 1: Temporal Probability Distribution of Patients Expected to Develop Coincidental Autoimmune Disorders

In the multinomial framework, there are various ways to characterize the temporal association between the disorders and the implant procedure. Cukier et al. (1993) identify the discrete random variable

$$R = \sum_{q=1}^{Q} qN_q \tag{1}$$

to be appropriate for their study and purpose, where R can assume the values $N, N+1, \ldots, QN$. Although other measures could be used to characterize temporal association, R is a natural quantity to consider for the medical study because it measures the mean latency period for a particular sample of patients. In other words, it can be interpreted as the cumulative number of patient-months until all patients in which the disorder occurred actually developed the disorder after receiving the implant treatment. For example, from the data in Table 1, the observed value of R is 61. In other words, it took a total 61 patient-months for all 9 patients to develop an autoimmune disorder after the implant treatment. Therefore, R is a way

to compare different patient group configurations on the basis of temporal ordering.

In order to study the temporal association, one must characterize the probability distribution of R and then assess the likelihood of observed values of R being generated from this distribution. For example, Cukier et al. (1993) assess the likelihood of the disorders occurring in the 9 patients as soon after treatment as they did. Stated in terms of R, they assess the likelihood of a randomly selected 9 patient group (from the multinomial distribution with probabilities given in Figure 1) having a total patient-month value of less than or equal to 61, which is just the tail probability

$$P\{R \leq 61\}. \tag{2}$$

A small value for this tail probability provides evidence for a strong temporal association between the disorders and the implant treatment since this indicates that disorders occur much sooner than expected from the patient population.

Based on the properties of the multinomial distribution, certain aspects of the distribution of R can be easily quantified. The mean and the variance of R are

$$E[R] = \sum_{q=1}^{Q} q u_q N \tag{3}$$

and

$$Var[R] = \sum_{q=1}^{Q} q^2 u_q (1-u_q) N - 2 \sum_{q} \sum_{<r} q r u_q u_r N. \tag{4}$$

Using the values for the $\{u_q\}$ shown in Figure 1, along with $Q = 96$, and $N = 9$, yields $E[R] \approx 229.056$ and $Var[R] \approx 3226.70$. An analytic upper bound for the tail probability in (2) can be obtained from the one-sided Chebyshev inequality (Ross 1988, page 352):

$$P\{R < 62\} \leq \frac{3226.70}{3226.70 + (229.056 - 62)^2} \approx 0.104. \tag{5}$$

Expression (5) illustrates that the tail probability , $P\{R \leq 61\}$ is bounded above by 0.104. This means close to 90% of all possible 9 patient groups would have a patient-month total of greater than or equal to 62. While this indicates that the observed R for the 9 patient group in Table 1 is somewhat rare, it is far from definitive since statistical significance in most studies is declared on a probability of 0.05 or less. In most instances, the Chebyshev inequality is quite conservative and does not provide a tight upper bound. In fact, this will be demonstrated in Section 4 to be the case.

Section 3 provides the details of the Monte Carlo simulation model that was used to generate the results reported in Cukier et al. (1993). We demonstrate that the model is able to produce a very precise estimate of the tail probability in (2).

3 A MONTE CARLO SIMULATION APPROACH

The relationship between the multinomial distribution and R results in a straightforward Monte Carlo simulation model for this problem. The simulation input was generated by the linear congruential pseudo random number generator RAND from Law and Kelton (1991), pages 449-450. According to the authors, this generator is portable and well-tested (see also Marse and Roberts 1983). In our Monte Carlo experiment, nine pseudo uniform random variates were transformed using inversion (Law and Kelton 1991, page 469) to produce a sample of nine random variates from the multinomial distribution. Then, R was computed from the multinomial sample. An estimate of the probability distribution for R was generated by performing 1,000,000 replications of this experiment. The following algorithm summarizes the logic used in a simulation program written in FORTRAN:

Algorithm 1 *Initialize the Number of Samples, N, Q, and the $\{u_q\}$*

Compute the cumulative probability distribution, C_q, for u_q, $q = 1, 2, \ldots, Q$

For $I = 1$, Number of Samples

Set $F_R = 0$, for $R = N, N+1, \ldots, QN$ (initialize the distribution of observed frequencies for R to zero)

Set $R = 0$

Set $N_q = 0$ for $q = 1, 2, \ldots, Q$ to zero

For $J = 1, N$

Sample U (Uniform(0,1) random variate)

Set $K = 1$

While $U > C_K$ do

$K = K + 1$

EndWhile

$N_K = N_K + 1$

EndFor J

For $L = 1, Q$

$R = R + N_L L$

EndFor L

$F_R = F_R + 1$

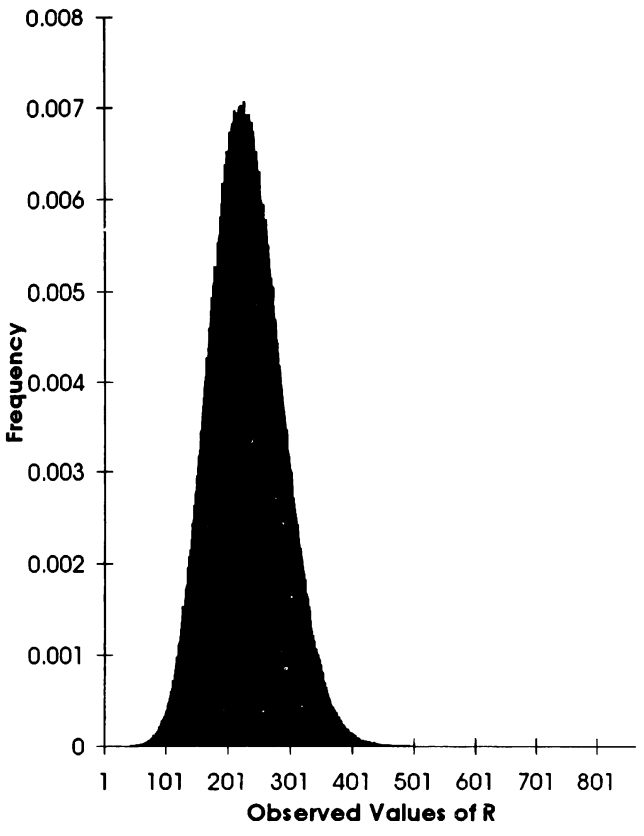

Figure 2: Observed Frequency for R from a Monte Carlo Simulation Run of 1,000,000 Samples

EndFor I

Calculate the cumulative sum of F_R *for* $R = N, N+1, \ldots, QN$

Output F_R *and the cumulative sums for* $R = N, N+1, \ldots, QN$.

Figure 2 is a distribution of the observed frequencies of values for R from the Monte Carlo simulation run. This plot (suitably standardized) serves as an unbiased estimate of the probability mass function (Law and Kelton 1991, page 362). It appears to be symmetric about the approximate mean value of 229.056 with very little probability in the tails beyond the value of $R \leq 61$.

The $P\{R \leq 61\}$ was estimated by,

$$\hat{P} = \frac{(\#\text{ of samples for which the observed R} \leq 61)}{1,000,000}. \tag{6}$$

The data illustrates that the probability of observing $R \leq 61$ is extremely low. Out of the 1,000,000 replications, 92 fall at or below the value of 61. Therefore, a point estimate for $P\{R \leq 61\}$ is $\hat{P} = 9.2 \times 10^{-5}$. Since the sample size is large and (6) is a proportion, large sample theory can be used to produce an approximate confidence interval for $P\{R \leq 61\}$. Using the procedure found in Bickel and Doksum (1977), page 160, an approximate $(1-\alpha)\times 100\%$, $(0 < \alpha < 1)$ confidence interval for $P\{R \leq 61\}$ is

$$\frac{S + \frac{k_\alpha^2}{2} \pm k_\alpha \sqrt{\frac{S(\eta - S)}{\eta} + \frac{k_\alpha^2}{4}}}{\eta + k_\alpha^2}$$

where S is the number of Monte Carlo simulation replicates observed in the tail, k_α is the $1 - \frac{\alpha}{2}$ percentile from the normal distribution, and η represents the total number of simulation replicates. In this example, $S = 92$, $k_\alpha = 1.96$ (for a 95% confidence level), and $\eta = 1,000,000$. Hence, the observed confidence interval is $(7.5024\times 10^{-5}, 1.1282\times 10^{-4})$. Bickel and Doksum (1977) state that the use of this confidence interval is satisfactory as long as the smaller of $(1,000,000)(P\{R \leq 61\})$ or $(1,000,000)(1 - P\{R \leq 61\})$ is at least 5. For the example in this paper, the value is estimated to be 92.

4 ANALYTICAL APPROACHES

The probability distribution for R can be determined analytically by addressing the following combinatorial problem: find all possible values of $\{N_q, q = 1, \ldots, Q\}$ such that

$$\sum_{q=1}^{Q} N_q = N \tag{7}$$

and

$$\sum_{q=1}^{Q} qN_q = R. \tag{8}$$

Once the possible combinations of $\{N_q\}$ are known, the probability that R equals some prespecified value can be calculated analytically using the probabilities from the multinomial distribution. Since we are interested in a cumulative tail probability, this procedure would have to be repeated for each fixed value of R. Total enumeration for all the $\{N_q, q = 1, \ldots, Q\}$ that satisfy the above two constraints for different values of R is quite difficult because the number of possible combinations of the $\{N_q\}$ grows exponentially in the size of Q, N, and R (see Jacobson and Morrice 1996). Hence a total enumeration approach is impractical for realistic sized problems such as the one considered in Cukier et al. (1993).

Fortunately, this problem can be solved without using total enumeration. In order to define this procedure, some additional notation is needed. Let n_q

be the total number of balls in the first q urns, $q = 1, 2, \ldots, Q$. By definition, $n_q = \sum_{i=1}^{q} N_i$. In addition, let

$$R_q = \sum_{i=1}^{q} iN_i$$

and

$$p_{q,n,r} = P\{R_q = r, n_q = n\}$$

for $q = 1, 2, \ldots, Q$, $n = 1, 2, \ldots, N$, $r = n, n+1, \ldots, qn$. Finally, let $V_{q,n,r}$ be the total number of patient samples (or vectors) of size n that are possible for the first q urns with $R_q = r$ and $n_q = n$ balls.

The problem is to determine the distribution for $R_Q (\equiv R)$, i.e., $p_{Q,N,r}$. The following theorem provides the basis for a recursive approach to calculating $p_{q,n,r}$.

Theorem 1 *The probability $p_{q,n,r}$ can be computed using the recursion*

$$p_{q,n,r} = \sum_{m=0}^{h(q,n,r)} p_{q-1,n-m,r-qm} \binom{n}{m} u_q^m \qquad (9)$$

for $q = 2, 3, \ldots, Q$, $n = 1, 2, \ldots, N$, and $r = n, n+1, \ldots, qn$, where $h(q, n, r) = min\{n, \lceil r/q \rceil\}$. The initial boundary conditions for this recursion are

1. $p_{q,n,r} = 0$ *for* $q = 1, 2, \ldots, Q, n = 1, 2, \ldots, N$ *and* $r = 1, 2, \ldots, N-1$.
2. $p_{1,n,n} = u_1^n$ *for* $n = 1, 2, \ldots, N$.
3. $p_{q,0,0} = 1$ *for* $q = 1, 2, \ldots, Q$.
4. $p_{q,0,r} = 0$ *for* $r = 1, 2, \ldots, qN$ *and* $q = 1, 2 \ldots, Q$.

Proof: See Jacobson and Morrice (1996).

The recursion in (9) can be successively applied, starting with the values $q = 2$ and $n = 1$, until the desired values for $q(= Q)$, $n(= N)$, and r are reached. A recursive algorithm is described in Jacobson and Morrice (1996).

A corollary to Theorem 1 can be used to determine the number of vectors for each possible value of R.

Corollary 1 *The number of vectors, $V_{q,n,r}$, can be computed using the recursion*

$$V_{q,n,r} = \sum_{m=0}^{h(q,n,r)} V_{q-1,n-m,r-qm} \binom{n}{m} \qquad (10)$$

for $q = 2, 3, \ldots, Q$, $n = 1, 2, \ldots, N$, and $r = n, n+1, \ldots, qn$, where $h(q, n, r) = min\{n, \lceil r/q \rceil\}$. The initial boundary conditions for this recursion are

1. $V_{q,n,r} = 0$ *for* $q = 1, 2, \ldots, Q, n = 1, 2, \ldots, N$ *and* $r = 1, 2, \ldots, N-1$.
2. $V_{1,n,n} = 1$ *for* $n = 1, 2, \ldots, N$.
3. $V_{q,0,0} = 1$ *for* $q = 1, 2, \ldots, Q$.
4. $V_{q,0,r} = 0$ *for* $r = 1, 2, \ldots, qN$ *and* $q = 1, 2, \ldots, Q$.

Proof: See Jacobson and Morrice (1996).

To solve the recursion in (9) for all values of r ($r = 1, 2, \ldots, QN$), in the worst case, QN recursions must be constructed for each value of r, where each recursion has at most $N+1$ parts in the summation. Therefore, the computational effort needed to determine the entire distribution for R is polynomial in Q and N, where the largest polynomial term is of order Q^2N^3. Solving the recursion in (10) for the number of possible vectors requires the the same amount of computational effort as solving the recursion in (9). In addition,

Lemma 1

$$\sum_{r=1}^{QN} V_{Q,N,r} = Q^N.$$

Proof: See Jacobson and Morrice (1996).

The recursive procedure derived from Theorem 1 was implemented in FORTRAN. This program can be used to characterize the entire distribution for R in the Cukier et al. (1993) study. The probability distribution is given in Figure 3. For $R \leq 61$, the $P\{R \leq 61\} = 1.0250 \times 10^{-4}$. This value validates the Monte Carlo simulation results reported by Cukier et al. (1993) and described in the last section, since the confidence interval produced by the simulation covers 1.0250×10^{-4}.

Convolving the distribution in Figure 1 with itself N times provides an alternative way to formulate the probability distribution of R. This follows since R is also the sum of sample of size N from the distribution in Figure 1. The following theorem provides a recursive procedure for calculating the probability distribution of R using this convolution formulation.

Theorem 2 *The probability $p_{q,n,r}$ can be computed using the convolution formula*

$$p_{q,n+1,r} = \sum_{i=\max(1,r-q)}^{r-1} p_{q,n,i} u_{r-i}$$

for $q = 1, 2, \ldots, Q$, $n = 1, 2, \ldots, N-1$, $r = n+1, n+2, \ldots, q(n+1)$. The initial boundary conditions for this formula are

$$p_{q,1,r} = \begin{cases} u_r & \text{if } r \leq q \\ 0 & \text{if } r > q \end{cases}$$

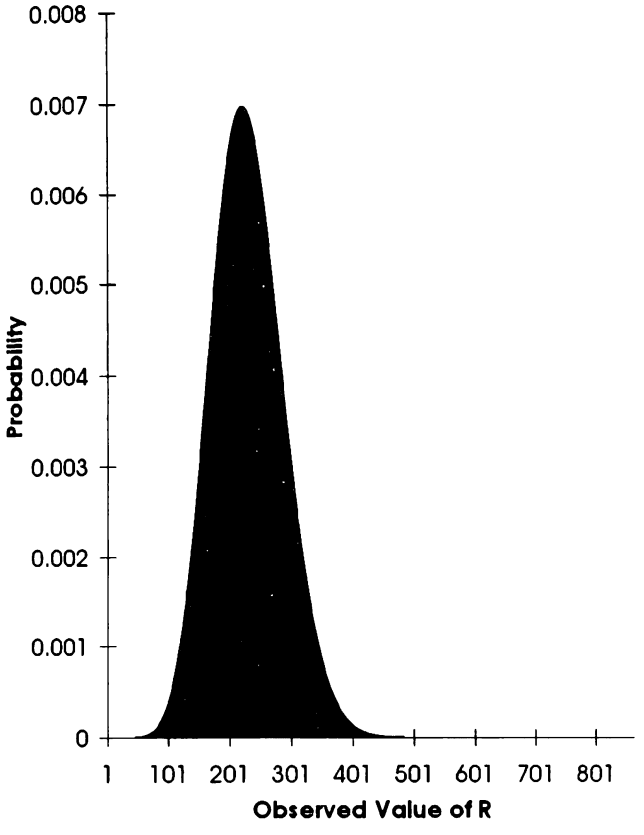

Figure 3: Probability Distribution for R

for $r = 1, 2, \ldots, QN$, $q = 1, 2, \ldots, Q$.

Proof: See Jacobson and Morrice (1996).

When the convolution formula in Theorem 2 is set up as a recursive algorithm, this algorithm requires polynomial computational effort in Q and N to generate the entire distribution for R. To obtain this distribution for Q alone, the largest polynomial term is of order Q^2N^2. However, if this distribution is required for all $q = 1, 2 \ldots, Q$, then the largest polynomial term is of order Q^3N^2.

The convolution formulation has the advantage over the formulation in Theorem 1 in computational speed if the distribution for R is required for the single value Q. However, if the distribution of R is required for all values of $q = 1, 2 \ldots, Q$, then the Theorem 1 formulation has the advantage over the convolution formulation in computational speed. This follows from the order of the largest polynomial terms for the two approaches.

Theorems 1 and 2 eliminate the need to perform a Monte Carlo simulation to calculate the probability distribution for R. In addition to being able to produce exact results, the algorithms associated with Theorems 1 and 2 are very fast relative to the simulation described in the last section. In particular, the simulation takes about 21.6 CPU seconds on the DEC AlphaServer 2100 4/275 to produce the data for Figure 2. On the same machine, the algorithm from Theorem 1 produces the data for Figure 3 in about 0.33 seconds of CPU time and the convolution approach produces the same results in about 0.14 seconds of CPU time. Perhaps the only advantage of the simulation over the results from Theorems 1 and 2 is that the simulation produces actual patient samples whereas the algorithms from the theorems do not produce such samples explicitly. However, sample counts are provided by the results in Corollary 1 and Lemma 1.

5 CONCLUSIONS

This paper develops an analytical model for describing the temporal association between disorders and treatments in medical studies. The model was inspired by a medical study conducted by Cukier et al. (1993). A complete description of the Monte Carlo simulation approach used by Cukier et al. (1993) has been provided. In addition, a recursive procedure and a convolution procedure that provide analytical solution to their problem have been described. The solution algorithms using these formulations are very efficient and can be used to solve realistically sized problems.

The model formulation presented in this paper is general and not restricted to the specific problem given in the cited medical study. Studies similar to Cukier et al. (1993) are common in the medical literature. Hochberg (1993) cites several examples and specifically discusses Cukier et al. (1993) and Bridges et al. (1993) as similar types of studies. The latter focuses on a possible connection between silicon breast implants and rheumatic disease.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Richard A. Beauchamp from the Texas Department of Health, Austin, Texas and Dr. Merwin W. Hemphill from the Texas Air Control Board, Austin, Texas for their help and advice on this research. The first author acknowledges support from the NSF (DMI-9409266, DMI-9423929) and the AFOSR (F49620-95-1-0124). The second author acknowledges the financial support of the CBA / GSB Faculty Research Committee of the College of Business and the End to End Simulation Department at Schlumberger Austin Research.

REFERENCES

Bickel, P. J. and K. A. Doksum. 1977. *Mathematical statistics: basic ideas and selected topics.* Oakland, California: Holden-Day.

Bridges, A. J., C. Conley, G. Wang, D. E. Burns, and F. B. Vasey. 1993. A clinical and immunological evaluation of women with silicon breast implants and symptoms of rheumatic disease. *Annals of Internal Medicine* 118:929–937.

Cukier, J., R. A. Beauchamp, J. S. Spindler, S. Spindler, C. Lorenzo, and D. E. Trentham. 1993. Association between bovine collagen dermal implants and a dermatomyositis or a polymyositis-like syndrome. *Annals of Internal Medicine* 118:920–928.

Fleming, T. R. and D. P. Harrington. 1991. *Counting processes and survival analysis.* New York: John Wiley and Sons, Inc..

Hochberg, M. C. 1993. Cosmetic surgical procedures and connective tissue disease: the cleopatra syndrome revisited. *Annals of Internal Medicine* 118:981–983.

Jacobson, S. H. and D. J. Morrice. 1996. A mathematical model for assessing the temporal association between health disorders and medical treatments. Technical Report, The University of Texas at Austin, Austin, Texas.

Johnson, N. L. and S. Kotz. 1977. *Urn models and their applications: an approach to modern discrete probability theory.* New York: John Wiley and Sons, Inc..

Kalbfleisch J. D. and R. L. Prentice. 1980. *The statistical analysis of failure time data.* New York: John Wiley and Sons, Inc..

Law A. M. and W. D. Kelton. 1991. *Simulation modeling and analysis.* 2nd ed. New York: McGraw-Hill, Inc..

Marse K. and S. D. Roberts. 1983. Implementing a portable FORTRAN uniform(0,1) generator. *Simulation* 21:135–139.

Ross, S. 1988. *A first course in probability.* 2nd ed. New York: Macmillan Publishing Company.

AUTHOR BIOGRAPHIES

SHELDON H. JACOBSON is an Associate Professor in the Department of Industrial and Systems Engineering at Virginia Polytechnic Institute and State University (Virginia Tech). Before joining Virginia Tech, he served for five years on the faculty in the Department of Operations Research at Case Western Reserve University. He has a B.Sc. and M.Sc. in Mathematics from McGill University, and a Ph.D. in Operations Research from Cornell University. He has served as the Advanced Tutorial Track Coordinator at both the 1994 and the 1995 Winter Simulation Conferences. He also served as the Doctoral Colloquium Coordinator at both the 1993 and the 1994 Winter Simulation Conferences. He served as the Treasurer for the InfORMS College on Simulation (1994-1996). His research interests include simulation optimization and sensitivity analysis, frequency domain approaches to analyzing simulation outputs, and issues related to the complexity of analyzing structural properties of discrete event simulation models.

DOUGLAS J. MORRICE is an Associate Professor in the Department of Management Science and Information Systems at The University of Texas at Austin. He is currently on sabbatical leave in the End to End Simulation Department at Schlumberger Austin Research Center in Austin, Texas. Dr. Morrice received his undergraduate degree in Operations Research at Carleton University in Ottawa, Canada. He holds an M.S. and a Ph.D. in Operations Research and Industrial Engineering from Cornell University. His research interests include discrete event and qualitative simulation modeling and the statistical design and analysis of large scale simulation experiments. Dr. Morrice is a member of the The Institute for Operations Research and Management Science (InfORMS) and the Council of Logistics Management. He served as the Secretary for the InfORMS College on Simulation (1994-1996) and is Co-Editor of the *Proceedings of the 1996 Winter Simulation Conference.*