

Provided for non-commercial research and educational use only.
Not for reproduction or distribution or commercial use.



This article was originally published in a journal published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues that you know, and providing a copy to your institution's administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

<http://www.elsevier.com/locate/permissionusematerial>



Modeling patient arrivals in community clinics[☆]

Christos Alexopoulos^{a,*}, David Goldsman^a, John Fontanesi^b, David Kopald^b,
James R. Wilson^c

^a*School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive, Atlanta, GA 30332-0205, USA*

^b*Department of Pediatrics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0927, USA*

^c*Edward P. Fitts Department of Industrial Engineering, North Carolina State University, Campus Box 7906, Raleigh, NC 27695-7906, USA*

Received 15 January 2004; accepted 20 July 2005

Available online 15 February 2006

Abstract

We develop improved methods for modeling and simulating the streams of patients arriving at a community clinic. In previous practice, random (unscheduled) patient arrivals were often assumed to follow an ordinary Poisson process (so the corresponding patient interarrival times were randomly sampled from an exponential distribution); and for scheduled arrivals, each patient's tardiness (i.e., the deviation from the scheduled appointment time) was often assumed to be randomly sampled from a normal distribution. A thorough analysis of patient arrival times, obtained from detailed workflow observations in nine community clinics, indicates these assumptions are not generally valid, and the tardiness data sets for this study are best modeled by unbounded Johnson distributions. We also propose a nonhomogeneous Poisson process to model the random patient arrivals; we review a nonparametric approach to estimating the associated mean-value function; and we describe an algorithm for generating random patient arrivals from the estimated model. The adequacy of this model of random patient arrivals can be assessed by standard goodness-of-fit tests. These findings are important since testable scheduling optimization strategies must be based upon accurate models for both random and scheduled patient arrivals. The impacts on modeling, as well as implications for practice management, are discussed.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Patient arrival process; Clinic scheduling; Johnson system of distributions

1. Introduction

In a world of tightening resources and rising competition, healthcare facilities face increasing pressure to have patients seen faster and to exit clinic services more quickly. Improved patient flow needs to be balanced with ensuring adequate time to complete needed

clinical functions. Identifying those factors that contribute to unnecessary nonclinical patient wait times would seem a wise undertaking [1].

The balancing act begins with the process of scheduling and controlling patient arrivals [2,3]. Patients often vary from their scheduled arrival times, leaving clinic personnel to either wait for late arriving patients or hurry through services [4,5]. Failure to adequately control patient arrival patterns has a cascading effect that begins with a mismatch between staffing ratios and service demand [6], producing excessive patient waiting time [7,8], variability in clinical encounter times [9,10], incomplete preventive service delivery [11], and

[☆] This paper was processed by Donna C. Llewellyn and Paul Griffin. It is based on a presentation at the INFORMS Annual Meeting in Atlanta, Georgia, October 2003.

* Corresponding author. Tel.: +1 404 894 2361.

E-mail address: christos@isye.gatech.edu (C. Alexopoulos).

disgruntled staff [12]. Patients react with dissatisfaction with the clinical encounter to the point where they may leave the clinic before the exam is complete [13].

Several authors have studied patient arrival patterns [14–18]. Such analyses ought to consider appropriate probability distributions for modeling the arrival processes. To the contrary, review of time-motion studies of patient arrival times published in MEDLINE, COCHRANE and INFORMS databases from 1985 to the present found that, when mentioned, investigators explicitly assumed that the interarrival times corresponding to the unscheduled visits are independent and identically distributed (IID) random variables from the exponential distribution (i.e., arising from a Poisson process). In addition, these studies often modeled patient tardiness for scheduled appointments as a normal random variable. Clearly there is a need to analyze scheduled patient arrival “tardiness” patterns.

For our purposes, there are two types of patient arrivals: *random (unscheduled)* and *scheduled*. A clinic supervisor’s operating strategy must be able to accommodate the exigencies of patient arrival behavior, both for unscheduled as well as scheduled patients. For example, how often do unscheduled patients show up? Further, how late (or early) are the scheduled patients likely to be? A critical first-step in performing queuing analysis is the modeling of the arrival process for unscheduled patients, and the tardiness times for scheduled patients.

The thesis of the current paper, however, is *that the time-homogeneous Poisson process is rarely an appropriate model for random arrivals, and that the normal distribution often does not seem appropriate for tardiness*. The following discussion underlines the reasons for this assessment and motivates our approach.

Why not exponential interarrival times? The assumption of IID exponentially distributed interarrival times (or, equivalently, arrivals occurring as a Poisson process) is a popular one in the queuing literature. Poisson processes arise naturally when the arrivals are “random” (“memoryless”), that is, when the knowledge that an arrival has not occurred by a certain time gives no additional information about the time of the next arrival. Since we know, at least approximately, when scheduled arrivals will appear, it is obvious that scheduled arrivals cannot possibly satisfy the memoryless assumption of the exponential distribution.

Can unscheduled arrivals arise from a time-homogeneous Poisson process? Probably not. In order for an arrival process to be Poisson, three mathematical assumptions must be satisfied. First, arrivals can only occur one-at-a-time. Second, arrival patterns must

not change as the day progresses. And third, arrivals in disjoint portions of the day must be statistically independent of each other. Unfortunately, unscheduled arrivals can violate each of these assumptions. First, arrivals do not have to occur one-at-a-time. For example, it is possible that factors external to the clinic or the patient, such as bus schedules, traffic-light timing, car-pooling, or availability of parking spaces, impact patients and act to “cluster” their arrivals. Second, patient arrival patterns certainly change over the course of a day due to commuting patterns, bus schedules, lunch hours, work schedules, and many other factors. Third, a traffic jam or other type of transit-related delay may force a large group of patient arrivals to be highly dependent over an extended period of time during the day. The bottom line is that arrivals, whether scheduled or not, often do not form a homogeneous Poisson process.

Why not normal tardiness for scheduled patients? The normal distribution’s “bell-shaped” probability density function is symmetric about its mean value and tails off quickly as we progress away from the mean. The mean is presumably zero in the case of tardiness values (where a positive tardiness value corresponds to a late arrival, and a negative value corresponds to an early arrival). The normal distribution is appropriate for many physical phenomena ranging from crop yield to standardized test performance; further, the normal distribution, in the form of the Central Limit Theorem, is usually an appropriate model for averages of observations. But is the normal distribution always appropriate for the case of tardiness observations? The answer here is a resounding “no,” as we will demonstrate in the subsequent discussion and in Section 4.

If tardiness is not normal, what is it? Law and Kelton [19] describe methods for fitting statistical distributions to data. First, the user attempts to identify candidate models based on summary statistics, graphical aids (e.g., histograms and box plots), or prior experience with the type of the data. If a parametric model seems plausible, the user proceeds with the estimation of the unknown parameters. This task can be accomplished by a variety of methods, including matching moments or quantiles, the method of maximum likelihood, and the method of least squares. Parameter estimation is followed by goodness-of-fit (GOF) tests, such as the omnibus chi-square test, the Kolmogorov–Smirnov (K–S) test, and the Anderson–Darling test. The majority of the aforementioned tasks can be accomplished by software packages, e.g., ExpertFit™ (Averill Law and Associates, 2005, Tucson, AZ). Packages for specific models are also available.

The remainder of this article proceeds as follows: Section 2 describes the experimental setup and the methodology used. Section 3 describes a practical method for modeling and generating nonhomogeneous Poisson arrival processes. Section 4 contains the results of our experiments. Section 5 provides concluding remarks and problems worthy of future study.

2. Experimental setting

The Partnership of Immunization Providers (PIP) is a collaborative public/private project created by the University of California, San Diego School of Medicine, Division of Community Pediatrics, in association with community clinics and small, private provider practices. PIP is funded by the Centers for Disease Control and Prevention. It is a healthcare delivery research enterprise with special emphasis on developing affordable and practical quality preventive health methodologies within those healthcare facilities serving the poorest and neediest of our nation. These partner clinics serve those living in areas known to have lower-than-average immunization rates, high unemployment, and ethnic diversity. PIP has applied multi-faceted strategies targeting provider practice, immunization-tracking capability, linkage to a computerized county tracking system, residency immunization curriculum, and quality improvement systems.

2.1. Data collection

A total of 225 patient observations were obtained during two different time periods (91 observations in the fall of 1999 and 134 in the late summer of 2000) as part of a larger workflow data collection effort. These observations were collected using a workflow data acquisition tool described in Fontanesi et al. [7]. This tool, the Observational Checklist of Patient Encounters (OCPE), includes data fields to record individual patient scheduled appointment times and actual arrival times.

2.2. Data analysis

To avoid anomalies that might result from operational changes or secular trends, both observational time periods were assessed separately. The difference between scheduled appointment times and actual patient arrival times (i.e., tardiness) was calculated; then empirical chi-square and K–S GOF tests were applied for a variety of distributions, including the normal and Johnson distributions [20]. The Johnson family of distributions will be discussed in detail in Section 3.

Within a specific data set, let X_i denote the i th patient's tardiness. In order for the forthcoming GOF

tests to work properly, we must establish independence among the sequential observations X_1, X_2, \dots . Independence can be assessed empirically (e.g., by the scatter plot of the pairs (X_i, X_{i+1})) or statistically (e.g., by von Neumann's test [21]).

To keep this article self-contained, we briefly review the relevant statistical tests—the von Neumann test for independence, and the chi-square and K–S tests for GOF.

von Neumann's test for independence. If the data X_1, X_2, \dots, X_n are IID, then the distribution of the statistic

$$C_n = \sqrt{\frac{n^2 - 1}{n - 2}} \left[1 - \frac{\sum_{i=2}^n (X_i - X_{i-1})^2}{2(n-1)S_n^2} \right] \quad (1)$$

is very close to the standard normal distribution, even for n as small as 8. We will denote the sample mean and sample variance of the data by \bar{X}_n and S_n^2 , respectively. Hence, one can reject the null hypothesis of independence at level α when $|C_n| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of the standard normal distribution. The value α is the user-specified Type I error, i.e., the probability of incorrectly rejecting the hypothesis of independence when this hypothesis is true. The p -value of this test is approximately $2[1 - \Phi(|C_n|)]$, where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal random variable. We remind the reader that the p -value of a test is the probability that a test statistic larger than the current one would be obtained if the hypothesized distribution were indeed correct. (Alternatively, the p -value is the smallest value of the Type I error for which the null hypothesis is rejected given the current test statistic.)

Chi-square GOF test. The chi-square is probably the most-popular GOF test, and it is simple to use, especially for data X_1, X_2, \dots, X_n arising from a continuous distribution having CDF $F(x)$. To conduct the most effective version of the test, we first divide the hypothesized distribution's support into k "equiprobable" intervals, that is, we identify values $a_0, a_1, a_2, \dots, a_k$ such that $F(a_i) = i/k$, or equivalently, $a_i = F^{-1}(i/k)$, for $i=0, 1, \dots, k$, where $F^{-1}(\cdot)$ is the inverse CDF. Notice that a_0 could be $-\infty$ and a_k could be $+\infty$. The respective intervals are $I_i = (a_{i-1}, a_i]$, $i = 1, \dots, k$, with the understanding that the first interval could be closed and the last interval could be open. We then compare the actual numbers of observations that fall in each interval, O_1, O_2, \dots, O_k , to the corresponding expected numbers, E_1, E_2, \dots, E_k , via the chi-square test statistic,

$$\chi_{\text{GOF}}^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

where $E_i = n[F(a_i) - F(a_{i-1})] = n/k$, for $i = 1, \dots, k$. Usually, one takes n sufficiently large or k small enough so that $E_i = n/k \geq 5$. The number of intervals can be chosen using Scott's [22] rule of thumb $k \approx \frac{5}{3} \sqrt[3]{n}$.

Clearly, a chi-square statistic that is "too large" indicates a poor fit between the observed data and the hypothesized distribution. More precisely, we reject the null hypothesis that $F(x)$ is the appropriate distribution if $\chi_{\text{GOF}}^2 > \chi_{1-\alpha, k-1-s}^2$, where $\chi_{1-\alpha, d}^2$ is the $(1 - \alpha)$ -quantile of the chi-square distribution with d degrees of freedom and s is the number of unknown parameters that are estimated by the method of maximum likelihood. If, for example, we hypothesize that the observations arise from a normal distribution with unknown mean μ and standard deviation σ , then we set $s = 2$, compute the maximum likelihood estimators (MLEs) $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma} = \sqrt{n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$, and evaluate χ_{GOF}^2 using $E_i = n\{\Phi[(a_i - \hat{\mu})/\hat{\sigma}] - \Phi[(a_{i-1} - \hat{\mu})/\hat{\sigma}]\}$, for $i = 1, \dots, k$. More discussion on parameter estimation follows the ensuing K–S GOF test synopsis.

Kolmogorov–Smirnov (K–S) GOF test. Again, consider data X_1, X_2, \dots, X_n arising from a continuous distribution, which we hypothesize to be the continuous CDF $F(x)$. The K–S test works with the empirical (sample) CDF of the data,

$$F_n(x) = \frac{\# \text{ of } X_i \leq x}{n}, \quad -\infty < x < \infty. \quad (3)$$

Assuming that $F(x)$ is indeed the true distribution, it is well known, via the Glivenko–Cantelli lemma [23], that, as the sample size n becomes large, the empirical CDF $F_n(x)$ converges uniformly to $F(x)$ for all x . Now let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ denote the order statistics based on the sample X_1, X_2, \dots, X_n . Then $F_n(X_{(i)}) = i/n$ for $i = 1, \dots, n$, and so it stands to reason that we can expect the "uniform" result that $F(X_{(i)}) \approx i/n$ if the hypothesized distribution is reasonable.

The K–S test quantifies both the maximum deviation of the empirical CDF above or below the predicted uniform line and permits assessment on both the upper and lower bounds of the distribution fit [24]. In particular, the K–S test rejects the hypothesized distribution when the test statistic

$$\begin{aligned} D_n &= \sup_x |F(x) - F_n(x)| \\ &= \max \left\{ \max_{1 \leq i \leq n} \left| \frac{i}{n} - F(X_{(i)}) \right|, \right. \\ &\quad \left. \max_{1 \leq i \leq n} \left| F(X_{(i)}) - \frac{i-1}{n} \right| \right\} \end{aligned} \quad (4)$$

is larger than a tabulated quantile based on the sample size n and the Type I error α .

2.3. Parameter estimation

For most candidate distributions, it turns out that one needs to estimate the unknown parameters, usually via the method of maximum likelihood, moment matching, quantile matching, or the method of least squares. For instance, the exponential distribution requires the estimation of a single rate parameter, the normal requires estimation of the mean and variance, and the three-parameter Weibull requires estimates for its location, shape, and scale parameters. In our work, we studied the fit of distributions from the Johnson system. Our choice of the Johnson system is due to its generality and flexibility since it contains bounded and unbounded distributions, and asymmetric and symmetric distributions (including the normal distribution). Routines for generating realizations from its families are incorporated in standard computer simulation languages such as ArenaTM (Rockwell Automation, 2005, Milwaukee, WI) and AutoModTM (Brooks Automation, 2005, Chelmsford, MA).

The Johnson system [20] contains four parametric families whose CDFs can be expressed in the form

$$F(x) = \Phi\{\gamma + \delta f[(x - \xi)/\lambda]\}, \quad (5)$$

where $\gamma \in (-\infty, \infty)$ and $\delta > 0$ are shape parameters, $\lambda > 0$ is a scale parameter, $\xi \in (-\infty, \infty)$ is a location parameter, and $f(x)$ is one of the following functions:

$$f(x) = \begin{cases} \ln(x) & \text{for the lognormal } (S_L) \text{ distribution,} \\ \sinh^{-1}(x) & \text{for the unbounded } (S_U) \text{ distribution,} \\ \ln[x/(1-x)] & \text{for the bounded } (S_B) \text{ distribution,} \\ x & \text{for the normal } (S_N) \text{ distribution.} \end{cases} \quad (6)$$

One can select a unique distribution from this system based on the first four moments. In particular, there is a unique distribution for each feasible combination of the skewness β_1 and the kurtosis β_2 (see Fig. 1 in Swain et al. [25]). Notice that S_N is the normal distribution with mean $\mu = \xi - (\gamma\lambda/\delta)$ and standard deviation $\sigma = \lambda/\delta$. Further, it can be shown that the S_U distribution has mean $\mu = \xi - \lambda \exp[1/(2\delta^2)] \sinh(\gamma/\delta)$.

The computation of MLEs for the four parameters within the families S_U and S_B is a difficult computational problem; further, the MLEs are often sensitive to departures from the assumed distribution. This

necessitates the estimation of the parameters by other methods, e.g., moment matching, quantile matching, and least-squares fitting.

The least-squares methodology works as follows. Recall that $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ are the order statistics based on a sample X_1, X_2, \dots, X_n from $F(x)$. Then for $i = 1, \dots, n$, the transformed realization $U_{(i)} = F(X_{(i)})$ has the distribution of the i th uniform order statistic; that is, the i th smallest observation in a random sample of size n from the uniform distribution on the interval $(0, 1)$. In particular,

$$U_{(i)} = \Phi\{\gamma + \delta f[(X_{(i)} - \xi)/\lambda]\} \quad \text{and} \\ \theta_i = E(U_{(i)}) = \frac{i}{n+1}, \quad i = 1, \dots, n. \quad (7)$$

If one writes $U_{(i)} = \theta_i + \varepsilon_i$, then the “errors” ε_i are translated uniform order statistics with mean zero and covariances

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = \frac{\theta_i(1 - \theta_j)}{n + 2}, \quad 1 \leq i \leq j \leq n. \quad (8)$$

Define the vectors

$$\mathbf{U} = [U_{(1)}, \dots, U_{(n)}]', \quad \boldsymbol{\theta} = [\theta_1, \dots, \theta_n]' \quad \text{and} \\ \boldsymbol{\varepsilon} = \mathbf{U} - \boldsymbol{\theta}. \quad (9)$$

The fitting of the X_i is based on the minimization of a quadratic form $\boldsymbol{\varepsilon}'\mathbf{W}\boldsymbol{\varepsilon}$ in the n -dimensional Euclidean space, where \mathbf{W} is the matrix associated with the quadratic form (see below for choices of \mathbf{W}). The least-squares optimization problem is

$$\text{minimize}_{(\gamma, \delta, \lambda, \xi)} g(\gamma, \delta, \lambda, \xi; \mathbf{W}) \equiv \boldsymbol{\varepsilon}'\mathbf{W}\boldsymbol{\varepsilon}$$

subject to

$$\delta > 0 \\ \lambda \begin{cases} 0 & \text{for } S_U \\ > X_{(n)} - \xi & \text{for } S_B \\ = 1 & \text{for } S_N \text{ and } S_L \end{cases} \\ \xi \begin{cases} < X_{(1)} & \text{for } S_L \text{ and } S_B \\ = 0 & \text{for } S_N. \end{cases} \quad (10)$$

The assignment $\mathbf{W} = \mathbf{I}$ yields ordinary least-squares (OLS) estimators, whereas $\mathbf{W} \neq \mathbf{I}$ results in weighted least-squares estimators. The most frequently used weight matrices are $\mathbf{W} = \mathbf{V}^{-1}$, where $\mathbf{V} = [\text{Cov}(\varepsilon_i, \varepsilon_j)]$, and $\mathbf{W} = \mathbf{D} = \text{diag}\{1/\text{Var}(\varepsilon_1), \dots, 1/\text{Var}(\varepsilon_n)\}$.

The assignment $\mathbf{W} = \mathbf{V}^{-1}$ produces the minimum variance linear unbiased estimators for the parameters

$(\gamma, \delta, \lambda, \xi)$. The matrix \mathbf{V}^{-1} has the tri-diagonal form

$$[\mathbf{V}^{-1}]_{ij} = (n + 1)(n + 2) \times \begin{cases} 2 & \text{if } i = j, \\ (-1) & \text{if } |i - j| = 1, \\ 0 & \text{if } |i - j| > 1, \end{cases}$$

and the respective objective function for the optimization problem (10) is

$$g(\gamma, \delta, \lambda, \xi; \mathbf{V}^{-1}) = 2(n + 1)(n + 2) \\ \times \left[\sum_{i=1}^n \varepsilon_i^2 - \sum_{i=2}^n \varepsilon_i \varepsilon_{i-1} \right]. \quad (11)$$

Unfortunately, this method is often problematic; see Section 2.2 and Appendix A of Kuhl and Wilson [26] for details.

The matrix $\mathbf{W} = \mathbf{D}$ yields the diagonally weighted least-squares (DWLS) parameter estimators that are highly-recommended by Swain et al. [25]. After a bit of algebra, we see that

$$g(\gamma, \delta, \lambda, \xi; \mathbf{D}) = (n + 2) \sum_{i=1}^n \frac{\varepsilon_i^2}{\theta_i(1 - \theta_i)}. \quad (12)$$

The optimization problems with objective functions (11) or (12) can be solved by the FITTR1 software package described in Swain et al. [25]. This portable package can be downloaded from the site <http://www.ie.ncsu.edu/jwilson> and can perform all aspects of distribution fitting, including statistical GOF tests (chi-square and K–S) and point-by-point comparison between the empirical CDF $F_n(x)$ and the fitted CDF $\hat{F}(x) = \Phi\{\hat{\gamma} + \hat{\delta} f[(x - \hat{\xi})/\hat{\lambda}]\}$ with the estimated parameters $\hat{\gamma}$, $\hat{\delta}$, $\hat{\lambda}$, and $\hat{\xi}$. The OLS and DWLS estimates are computed based on the optimization algorithm of Marquardt [27]. In the absence of initial values, the algorithms use estimates based on moment matching. In addition to the above two methods, alternative estimators can be obtained by the method of moments, matching sample quantiles (Section 5 of Venkatraman and Wilson [28]), or by minimizing the L_1 or L_∞ norm of the difference between the empirical CDF and the fitted Johnson CDF.

Remark. A couple of facts relevant to GOF tests are in order (for additional details, see Chapter 6 of Law and Kelton [19]): (a) The chi-square test is (asymptotically) valid only when all parameters of the hypothesized CDF are known or when any unknown parameter is replaced by its respective MLE (in this case, we subtract one degree of freedom for each such parameter). Also, the chi-square test is not quite as robust as the Kolmogorov–Smirnov test, as it is highly dependent on

the number/size of chosen intervals. (b) The K–S test is valid for any sample size n provided that all parameters of $F(x)$ are known or that $F(x)$ is amongst the exponential, normal/lognormal, and Weibull distributions and all unknown parameters have been estimated by the method of maximum likelihood.

Although the normal distribution is only one constituent of the FITTR1 package, we conducted the relative GOF tests using the normal-distribution-based MLEs of the mean and standard deviation, i.e., \bar{X}_n and $\sqrt{(n-1)/n}S_n$.

On the other hand, the GOF tests for the S_U and S_B Johnson distributions were conducted with least squares estimates in place of the unknown parameters. This practice is common (although rarely mentioned explicitly). For example, moment-matching estimates are used by Arena's Input Analyzer with GOF tests. Furthermore, any concerns related to the validity of a GOF test should be relaxed when the p -value is either low (there is significant evidence against the null hypothesis) or high.

3. A nonparametric approach to modeling and simulating unscheduled patient arrivals

Suppose that we are given a time interval $(0, S]$ over which we observe several independent replications (realizations) of a stream of unscheduled patient arrivals, and that this stream of arrivals constitutes a nonhomogeneous Poisson process (NHPP) with a time-dependent arrival rate $\lambda(t)$ for $t \in (0, S]$. For example, the observation interval $(0, S]$ might represent the time period on each weekday during which unscheduled patients may walk into a clinic—say, between 9:00 A.M. and 5:00 P.M.

Suppose that k realizations of the arrival stream over this observation interval have been recorded so that we have n_i patient arrivals in the i th realization for $i = 1, 2, \dots, k$; and thus we have a total of $n = \sum_{i=1}^k n_i$ patient arrivals accumulated over all realizations of the arrival stream. Moreover, let $\{t_{(i)} : i = 1, \dots, n\}$ denote the overall set of arrival times for all unscheduled patients expressed as an offset from the beginning of the observation interval $(0, S]$ and then sorted in increasing order. Thus, for example, if we observed $n = 250$ patient arrivals over $k = 5$ days, each with an observation interval of length $S = 480$ min, then $t_{(1)} = 2.5$ min means that over all 5 days, the earliest patient arrival occurred 2.5 min after the clinic opened its doors to unscheduled arrivals on one of those days; and similarly, $t_{(2)} = 4.73$ min means that the second-earliest patient

arrival occurred 4.73 min after the clinic opened its doors to unscheduled arrivals on one of those days.

Given that $\lambda(t)$ represents the rate of arrival of unscheduled patients for each t in the observation interval $(0, S]$, we see that the mean-value function $\mu(t)$ representing the expected number of arrivals during the interval $(0, t]$ is given by $\mu(t) = \int_0^t \lambda(s) ds$. We take $t_{(0)} \equiv 0$ and $t_{(n+1)} \equiv S$ so that a piecewise linear estimator of $\mu(t)$ for $t \in (0, S]$ is

$$\hat{\mu}(t) = \frac{in}{(n+1)k} + \left\{ \frac{n[t - t_{(i)}]}{(n+1)k[t_{(i+1)} - t_{(i)}]} \right\} \quad \text{if} \\ t_{(i)} < t \leq t_{(i+1)} \quad \text{and} \quad i = 0, 1, \dots, n; \quad (13)$$

(see Leemis [29]). Eq. (13) provides a basis for modeling and simulating unscheduled patient-arrival streams when the arrival rate exhibits a strong dependence, for example, on the time of day.

To perform GOF testing on the fitted mean-value function $\hat{\mu}(t)$ for $t \in (0, S]$, we recommend the following cross-validation technique. Suppose that in addition to the k realizations of the target arrival process that were used to compute the estimated mean-value function $\hat{\mu}(t)$, we observe one additional realization $\{A'_i : i = 1, 2, \dots, n'\}$ that is independent of the previously observed realizations, with the i th patient arriving at time A'_i for $i = 1, \dots, n'$. If the target arrival stream is in fact an NHPP with mean-value function $\mu(t)$ for $t \in (0, S]$, then the transformed arrival times $\{B'_i = \mu(A'_i) : i = 1, 2, \dots, n'\}$ constitute a homogeneous Poisson process with an arrival rate of 1; and the corresponding transformed interarrival times $\{X'_i = B'_i - B'_{i-1} : i = 1, 2, \dots, n'\}$ (with $B'_0 \equiv 0$) constitute a random sample from an exponential distribution with a mean of 1. It follows that an appropriate test for the adequacy of the fitted mean-value function $\hat{\mu}(t)$ as an approximation to the true mean-value function $\mu(t)$ is to apply the Kolmogorov–Smirnov test to the data set $\{X''_i = \hat{\mu}(A'_i) - \hat{\mu}(A'_{i-1}) : i = 1, 2, \dots, n'\}$ (with $A'_0 \equiv 0$) using the hypothesized distribution function $F(x) = 1 - e^{-x}$ for all $x \geq 0$. For a comprehensive discussion of other techniques for assessing the goodness of fit of estimated arrival processes, see [26,30,31].

If the mean-value function $\hat{\mu}(t)$ passes the GOF test outlined above, then we can use the following simulation algorithm of Leemis [29] to generate a new stream of arrival times $\{A_i : i = 1, 2, \dots\}$ over the time interval $(0, S]$ with the same general pattern of dependence on time as in (13)—that is, with the arrival rate at each time t in the interval $(0, S]$.

```

[1] Set  $i \leftarrow 1$  and  $N \leftarrow 0$ .
[2] Generate  $U_i \sim \text{Uniform}(0, 1)$ .
[3] Set  $B_i \leftarrow -\ln(1 - U_i)$ .
[4] While  $B_i < n/k$  do
    Begin
        Set  $m \leftarrow \lfloor \frac{(n+1)kB_i}{n} \rfloor$ ; Set  $A_i \leftarrow t_{(m)} + \{t_{(m+1)} - t_{(m)}\} \left\{ \frac{(n+1)kB_i}{n} - m \right\}$ ;
        Set  $N \leftarrow N + 1$ ; Set  $i \leftarrow i + 1$ ; Generate  $U_i \sim \text{Uniform}(0, 1)$ ,
        Set  $B_i \leftarrow B_{i-1} - \ln(1 - U_i)$ .
    End

```

Note that in the simulation algorithm given above, $\lfloor z \rfloor$ denotes the greatest integer (or floor) function so that, for example, $\lfloor 3.12 \rfloor = 3$. Moreover, the total number of arrivals generated by this algorithm on one simulated realization of the arrival stream is given by the random variable N ; and provided that $N > 0$, the i th patient will arrive at time A_i for $i = 1, \dots, N$.

The main advantage of this approach to modeling and simulation of time-dependent arrival processes is that it does not require the assumption of any particular functional form for the way in which the arrival rate $\lambda(t)$ depends on the time t since the beginning of the observation interval $(0, S]$. Moreover as $k \rightarrow \infty$, so that the number of realizations of the target arrival process becomes large, with probability 1 the estimated mean-value function $\hat{\mu}(t)$ of Eq. (13) converges to the true mean-value function $\mu(t)$ for all $t \in (0, S]$. This means that the simulation algorithm given above (which is based on “inversion” of $\hat{\mu}(t)$ so that $A_i = \hat{\mu}^{-1}(B_i)$ for $i = 1, \dots, N$) is also asymptotically valid as $k \rightarrow \infty$. For more information on this approach to modeling and simulation of time-dependent arrival processes, see Leemis [32].

Remark. The United Network for Organ Sharing (UNOS) carried out a remarkable large-scale application of a simplified variant of this approach to modeling and simulating patient-arrival streams in the development and use of the UNOS Liver Allocation Model (ULAM) for analysis of the cadaveric liver-allocation system in the United States (see [33]). ULAM incorporated models of (a) the streams of liver-transplant patients arriving at 115 transplant centers, and (b) the streams of donated organs arriving at 61 organ procurement organizations in the United States—and virtually all these arrival streams exhibited strong dependencies on the time of day, the day of the week, and the season of the year as well as pronounced geographic effects.

Remark. As we indicated in the introduction, extraneous conditions might force the creation of “batched” arrivals. This issue can be addressed by means of *compound* Poisson processes (homogeneous or nonhomogeneous): the arrivals of groups are generated by the underlying Poisson process, and the sizes of the groups are IID random variables from some discrete distribution that is independent from the Poisson process.

4. Results

We examined two sets of data collected during two different days. The first set contained $n = 91$ observations, and the second set contained 134 observations. The times are in minutes. The nonparametric Kruskal–Wallis test [19, Chapter 6] with Type I error $\alpha \leq 0.25$ rejected the hypothesis that the data sets are homogeneous; hence, they cannot be merged into a single set. Among the continuous distributions in Chapter 6 of Law and Kelton, only the Johnson S_U (unbounded) and S_N (normal) appeared to be reasonable candidates based on graphical techniques (histograms, box plots, quantile–quantile plots, and probability–probability plots) and statistical GOF tests.

We first analyzed data set 1. The independence of the observations was assessed and verified via a scatter plot and the von Neumann test (the test statistic is $C_{19} = 0.713$ with a p -value of 0.476). The sample mean and sample standard deviation of the observations are $\bar{X}_{91} = -3.49$ and $S_{91} = 19.51$, respectively. Rows 3–4 of Table 1 list the test statistics and the associated p -values for the equiprobable chi-square and K–S GOF tests. Both tests indicate the poor fit of the normal distribution. Next, we used the FITTR1 package to fit a Johnson distribution other than the normal one. This package indicated that the only proper distribution is the S_U unbounded Johnson distribution, and then proceeded to compute the

Table 1
Chi-square and Kolmogorov–Smirnov tests for normal and Johnson S_U distributions

Data set 1 ($n = 91, \bar{X}_{91} = -3.49, S_{91} = 19.51$)	Statistic	p -value
Normal distribution:		
Equiprobable chi-square test (10 intervals)	29.11	0.0062
K–S test	0.148	≤ 0.025
Unbounded Johnson (S_U) distribution: ($\hat{\gamma} = 0.224, \hat{\delta} = 0.750, \hat{\lambda} = 7.814, \hat{\xi} = 1.188$)		
Equiprobable chi-square test (10 intervals)	11.75	0.23
K–S test	0.081	0.59
Data set 2 ($n = 134, \bar{X}_{134} = 8.88, S_{134} = 18.67$)		
Normal distribution:		
Equiprobable chi-square test (15 intervals)	23.61	0.051
K–S test	0.083	≤ 0.025
Unbounded Johnson (S_U) distribution: ($\hat{\gamma} = -0.596, \hat{\delta} = 1.630, \hat{\lambda} = 24.270, \hat{\xi} = -1.757$)		
Equiprobable chi-square test (15intervals)	16.45	0.29
K–S test	0.046	0.93

following DWLS estimates: $\hat{\gamma} = 0.224, \hat{\delta} = 0.750, \hat{\lambda} = 7.814$, and $\hat{\xi} = 1.188$. Rows 7–8 of Table 1 contain the outcomes of chi-square and K–S tests for the unbounded Johnson distribution. Fig. 1 superimposes the empirical CDF $F_{91}(\cdot)$, the fitted Johnson CDF, and the fitted normal CDF. The good fit of the Johnson model is evident from the large p -values and the tight fit between the empirical and fitted densities. Fig. 2 overlays the plots of the fitted unbounded Johnson and normal densities. Notice that the fitted Johnson probability density function (PDF) has “fat” tails and is a bit asymmetric, neither characteristic of a normal distribution.

We next analyzed the second data set with $n = 134$ observations. The sample mean and sample standard deviation of the observations are $\bar{X}_{134} = 8.88$ and $S_{134} = 18.67$, respectively. Again, the normal distribution gave a poor fit; this is evident from the low p -values of the equiprobable chi-square and K–S tests listed in rows 11–12 of Table 1 as well as from Figs. 3 and 4.

The FITTR1 package indicated that within the Johnson system only the S_U distribution yields a good fit and computed the following DWLS estimates: $\hat{\gamma} = -0.596, \hat{\delta} = 1.630, \hat{\lambda} = 24.270$, and $\hat{\xi} = -1.757$. Fig. 3 depicts the empirical CDF $F_{134}(\cdot)$, the fitted unbounded Johnson CDF, and the fitted normal CDF whereas Fig. 4 displays the respective fitted Johnson and normal PDFs. Rows 15–16 of Table 1 list the outcomes of equiprobable chi-square and K–S GOF tests. The good fit of the Johnson model is again apparent from the large p -values (especially those associated with the K–S test).

Remark. We also submitted the second data set to ExpertFit’s automated fitting procedure. Amongst 18 continuous distributions with unbounded support, the Johnson S_U distribution was the overwhelming winner, with the three-parameter log-logistic distribution a close second (the remaining 16 distributions performed miserably). ExpertFit used the method of quantile matching to derive the following estimates of S_U ’s parameters: $\hat{\gamma} = -0.576, \hat{\delta} = 1.548, \hat{\lambda} = 21.741$, and $\hat{\xi} = -0.775$. Both S_U and log-logistic distributions passed the K–S test, with the S_U distribution yielding a smaller test statistic; but the log-logistic distribution failed the equiprobable chi-square test based on 15 intervals at levels $0.10 \leq \alpha \leq 0.20$.

5. Conclusions

A number of interesting findings arose from our study. First, unscheduled arrivals may not follow an ordinary Poisson process (i.e., the interarrival times are not IID exponential); in fact, unscheduled arrivals may not even follow a more-general NHPP (i.e., a Poisson process whose arrival rate changes over time). Second, it appears from our multiple GOF measurements that tardiness times for scheduled patients should not be assumed to have a normal distribution; in our study, the tardiness times were not even particularly symmetric. For these reasons, we must warn modelers of clinic processes to exercise caution when choosing the distributions of the random variables (such as interarrival and service times) that drive simulation programs.

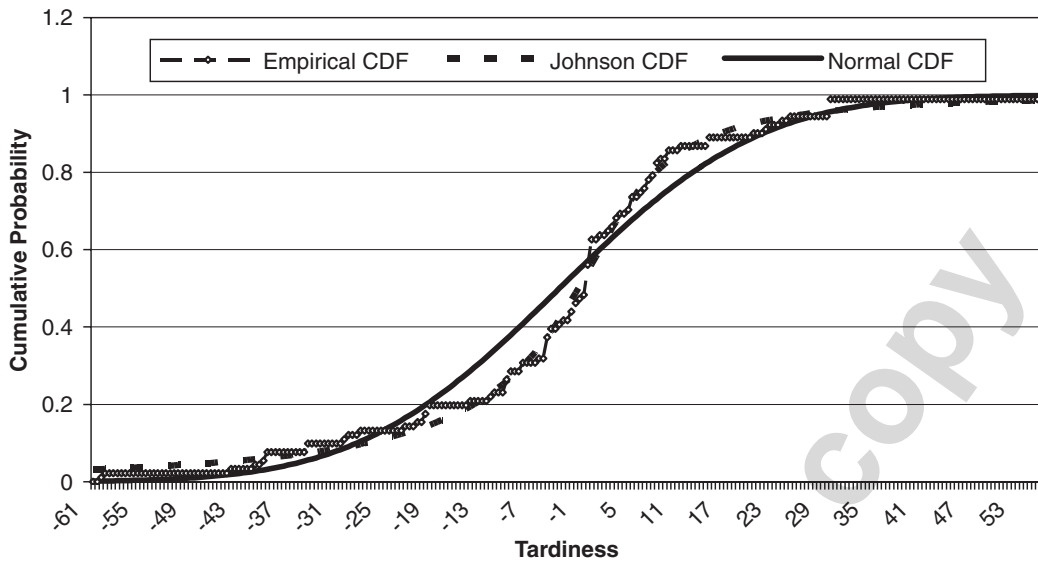


Fig. 1. Empirical and fitted CDFs for data set 1.

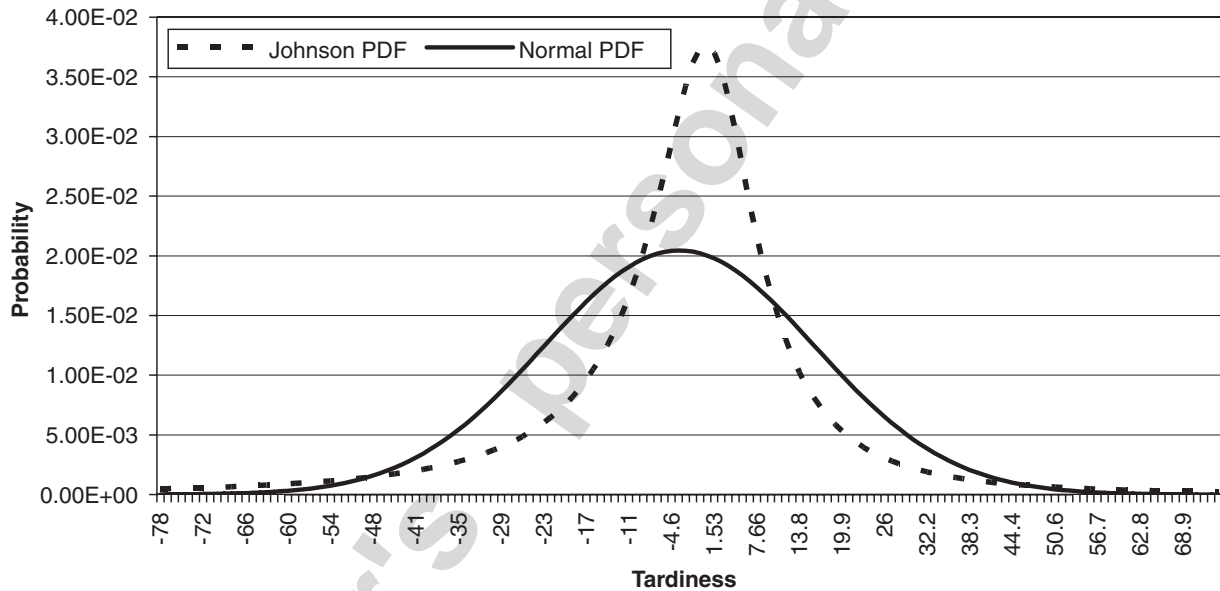


Fig. 2. Fitted Johnson and normal densities for data set 1.

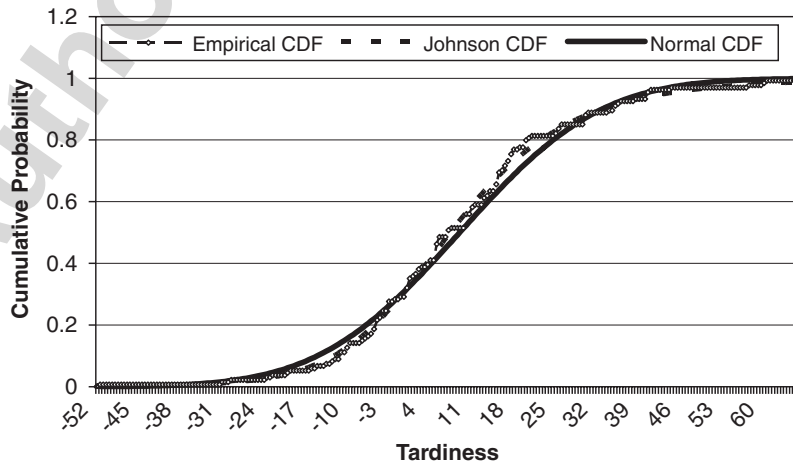


Fig. 3. Empirical and fitted CDFs for data set 2.

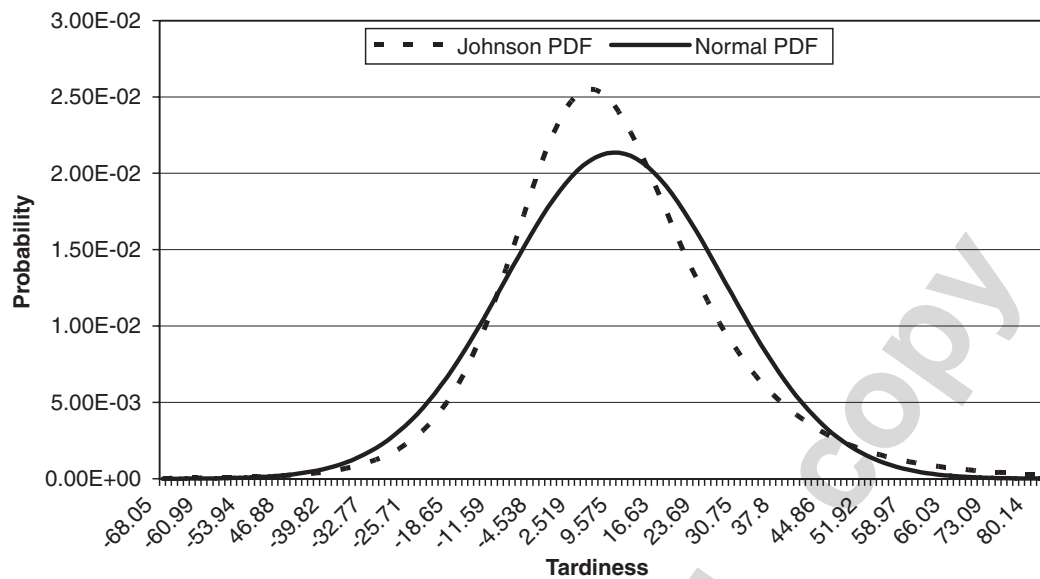


Fig. 4. Fitted Johnson and normal densities for data set 2.

What do we still need to do? An important task that remains is to examine the consequences of misidentifying a particular distribution. For example, the penalty for using a slightly incorrect tardiness distribution may not be particularly burdensome, as long as the first two moments (expected value and variance) are “close enough” to the actual values. On the other hand, clinics that fail to properly characterize patient tardiness when creating scheduling strategies are likely to experience impedance in matching demand to resources [34,35]. It is therefore of interest to determine which random inputs are “most vital” to the simulation and have the potential to cause the greatest problems if modeled incorrectly; this problem can be studied via a simulation-based sensitivity analysis. We are also interested in obtaining sound models for the other distributions that the simulation requires, e.g., service-time distributions, probabilities of certain tasks being performed correctly, etc. This task simply requires more and better data, and is ongoing through the OCPE observational data collection facility. We wish to point out that the Johnson system is not a panacea for modeling patient tardiness, it is just a flexible alternative to common distributions (e.g., beta, triangular, etc.). In cases where tardiness distributions exhibit multimodal behavior and distinct causes for those modes cannot be identified, it may be necessary to fit sufficiently flexible distribution families, and for this purpose the Bézier family of probability distributions may be an attractive alternative [36].

Finally, how can we use this work in a practical way? One of PIP’s primary goals is to provide clinics

with tools to allocate staff and appointment schedules in such a way to give the maximum number of patients the highest quality service. Thus, ongoing work includes optimal scheduling policies for staff and patients, improvement of medical practices to insure better levels of treatment, and the development of clinic policies in cases of unusual levels of patient demand. Of course, one of the most obvious ways to study these problems is through the use of computer simulation techniques, provided that adequate and reliable data are obtained.

References

- [1] Meza JP. Patient waiting times in a physician’s office. *American Journal of Managed Care* 1998;4:703–12.
- [2] Christl HL. Some methods of operations research applied to patient scheduling problems. *Medical Progress Through Technology* 1973;2:19–27.
- [3] Perros P, Frier BM. An audit of waiting times in the diabetic outpatient clinic: Role of patients’ punctuality and level of medical staffing. *Diabetic Medicine* 1996;13:669–73.
- [4] Kalai E, Kamien MI, Rubinovitch M. Optimal service speeds in a competitive environment. *Management Science* 1992;38:1154–63.
- [5] Kapustiak J, Ling H. Evaluation of patient waiting times at an academic ophthalmology clinic. *Journal of Medical Practice Management* 2000;15:228–33.
- [6] Asa AY, Carter MW, Nagle LM. A decision support system to meet the fluctuating needs of a hospital nursing unit. *Medinfo* 1995;8(2):1418.
- [7] Fontanesi JM, DeGuire M, Chiang J, Holcomb K, Sawyer MH. Application of workflow analysis tools in outpatient primary care settings. *Joint Commission Journal on Quality Improvement* 2000;26:654–60.

- [8] Waghorn A, McKee M. Surgical outpatient clinics: Are we allowing enough time? *International Journal for Quality in Health Care* 1999;11:215–9.
- [9] Dexter F. Design of appointment systems for preanesthesia evaluation clinics to minimize patient waiting times: A review of computer simulation and patient survey studies. *Anesthesia & Analgesia* 1999;89:925–31.
- [10] Saunders CE, Makens PK, Leblanc LJ. Modeling emergency department operations using advanced computer simulation systems. *Annals of Emergency Medicine* 1989;18:134–40.
- [11] Fontanesi J, De Guire M, Holcomb K, Kopald D, Sawyer MH. Can the doctor still see me: What happens when patients arrive late? *Journal of Medical Practice Management* 2003;18(5):239–43.
- [12] Reeves C. How many staff members do you need? *Family Practice Management* 2002;9(8):45–9.
- [13] Kyriacou DN, Ricketts V, Dyne PL, McCollough MD, Talan DA. A 5-year time study analysis of emergency department patient care efficiency. *Annals of Emergency Medicine* 1999;34:326–35.
- [14] Callahan NM, Redmon WK. Effects of problem-based scheduling on patient waiting and staff utilization of time in a pediatric clinic. *Journal of Applied Behavioral Analysis* 1987;20:193–9.
- [15] Clague JE, Reed PG, Barlow J, Rada R, Clarke M, Edwards RH. Improving outpatient clinic efficiency using computer simulation. *International Journal for Health Care Quality Assurance* 1997;10:197–201.
- [16] Dada M, Babad Y. In: 1991 Joint meeting of The Operations Research Society of America/The Institute of Management Science (ORSA/TIMS). Anaheim, CA: Institute for Operations Research and Management Science (INFORMS); 1991.
- [17] Hashimoto F, Bell S. Improving outpatient clinic staffing and scheduling with computer simulation. *Journal of General Internal Medicine* 1996;11:182–4.
- [18] Jennings M. Audit of a new appointments system in a hospital outpatient clinic (published erratum appears in *BMJ* 1991 Feb 23; 302(6774):455) (see comments). *British Medical Journal (Clin Res Ed)* 1991;302:148–9.
- [19] Law AM, Kelton WD. *Simulation modeling and analysis*, 3rd edition. Boston: McGraw-Hill; 2000.
- [20] Johnson NL. Systems of frequency curves generated by methods of translation. *Biometrika* 1949;36:149–76.
- [21] von Neumann J. Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 1941;12:367–95.
- [22] Scott DW. On optimal and data-based histograms. *Biometrika* 1979;66:505–610.
- [23] Karr AF. *Probability*. New York: Springer; 1993.
- [24] Stephens M. Use of the Kolmogorov–Smirnov, Cramér–von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society* 1970;Series B:115–20.
- [25] Swain J, Venkatraman S, Wilson JR. Least-squares estimation of distribution functions in Johnson’s translation system. *Journal of Statistical Computation and Simulation* 1988;29:271–97.
- [26] Kuhl ME, Wilson JR. Least squares estimation of nonhomogeneous Poisson processes. *Journal of Statistical Computation and Simulation* 2000;67:75–108.
- [27] Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the SIAM* 1963;11:431–41.
- [28] Venkatraman S, Wilson JR. Modeling univariate populations with Johnson’s translation system—description of the FITTRI software. Research Memorandum 87–21. School of Industrial Engineering, Purdue University, West Lafayette, Indiana; 1987.
- [29] Leemis LM. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process. *Management Science* 1991;37:866–900.
- [30] Lee S, Wilson JR, Crawford MM. Modeling and simulation of a nonhomogeneous Poisson process having cyclic behavior. *Communications in Statistics—Simulation and Computation* 1991;20(2&3):777–809.
- [31] Kuhl ME, Wilson JR, Johnson MA. Estimating and simulating Poisson processes having trends or multiple periodicities. *IIE Transactions* 1997;29:201–11.
- [32] Leemis LM. Nonparametric estimation and variate generation for a nonhomogeneous Poisson process from event count data. *IIE Transactions* 2004;36:1155–60.
- [33] Harper AM, Taranto SE, Edwards EB, Daily OP. An update on a successful simulation project: The UNOS liver allocation model. In: Joines JA, Barton RR, Kang K, Fishwick PA, editors. *Proceedings of the 2000 winter simulation conference*. Piscataway, NJ: Institute of Electrical and Electronics Engineers; 2000. p. 1955–62.
- [34] Podgorelec V, Kokol P. Genetic algorithm based system for patient scheduling in highly constrained situations. *Journal of Medical Systems* 1997;21:417–27.
- [35] Reilly T, Marathe V, Fries B. A delay-scheduling model for patients using a walk-in clinic. *Journal of Medical Systems* 1978;2:303–13.
- [36] Wagner MAF, Wilson JR. Using univariate Bézier distributions to model simulation input processes. *IIE Transactions* 1996;28:699–711.