Robust Prediction of Treatment Times in Concurrent Patient Care

Rafael B. Fricks, *Member, IEEE,* Henry Tseng, Marjorie Veihl, Kishor S. Trivedi, *Life Fellow, IEEE,* and Roger C. Barr, *Fellow, IEEE*

Abstract— **Outpatient centers comprised of many concurrent clinics increasingly see higher patient volumes. In these centers, decisions to improve clinic flow must account for the high degree of interdependence when critical personnel or equipment is shared between clinics. Discrete event simulation models have provided clinical decision support, but rarely address high-volume clinics with shared resources. While highly complex models are now capable of representing clinics in detail, validation techniques often do not evaluate model predictive performance when presented with new data. Cross-validation provides a means to evaluate the robustness of model treatment time predictions when ongoing data collection in clinics is impractical. Ensuring robust predictions assures validity in the use of models to optimize clinic performance.**

We apply cross-validation in evaluating a model of glaucoma clinic service at Duke Eye Center. In-person observation is used to verify the accuracy of operations data collected through electronic health records (EHR). From the EHR data, we formulate a stochastic reward net model, employing phase-type distributions to represent treatment durations, and solved through discrete event simulation. The model is formulated in two configurations to represent (1) concurrent demand on clinic staff, or (2) independently functioning clinics. Evaluating these two alternatives in cross-validation studies, we find model prediction accuracy improves when interdependence is explicitly modeled in the examined setting.

I. Introduction

Multi-service centers account for an increasing share of outpatient visits in health care, however a minimal share of reported clinic flow models [1]. Centers are challenging to model, as they may have several clinics operating concurrently that share co-located imaging and testing. The interdependence between clinics that share critical resources makes planning in centers particularly difficult. Discrete event simulation (DES) provides tools for modeling clinics and evaluating new plans, however most applications focus on independent portions of a health care system [1]. In fact, reviews have shown relatively few models for interconnected clinics [1-3], such as in centers. Explicitly modeling dependence on shared staff or resources between clinics adds to model complexity. Alternatively, simplifying models by assuming independence may skew predictions of the length of patient visits and impact overall simulation model accuracy. Ultimately each model must compromise in choosing adequate detail to represent the studied clinic [1]. Rigorous validation protocols are necessary to ensure a model provides suitable accuracy not only in the data used to generate the model, but continues to predict subsequent data.

Many examples of validation in clinic flow models determine if suitable accuracy is achieved by comparing model predictions to the same data set used to 'train' a model via an appropriate statistical test (examples in [4-6]). Assessing model accuracy with respect to the same training data tends to underestimate model error in predicting new data [7]. In more complex models, this optimistic error estimate is usually attributed to a greater ability to represent nuances in a particular data set [7]. Though preferable for increasingly complex models, it is often impractical to measure clinic performance on an ongoing basis to have follow-up independent data to evaluate predictive ability. Several authors noted the difficulty of gathering clinic flow data such as visit durations for an initial modeling application. Cross-validation techniques provide methods to estimate how models predict future data by resampling existing data [7]. By subdividing an existing set of operations data into training and tests subsets, we can assess how the model predicts independent data sampled from comparable populations. These methods estimate the model's accuracy in predicting future clinic flow.

We employ cross-validation to estimate the effect of explicitly accounting for interdependence between high-volume glaucoma clinics at Duke Eye Center on model accuracy. Nearly 90 patients visit glaucoma clinics at DEC daily, with typically one specialty provider examining as many as 60 patients per day directly. Each patient visit can be divided into three distinct steps; (1) initial examination by a technician (workup), (2) any ordered testing or imaging such as visual field tests or retinal photography, (3) visit with the patient's physician. While patients are initially referred to these clinics for glaucoma indications, visits to a specialty clinic may include broader comprehensive tests such as visual acuity measurements or cataract evaluations. For clarity in this interim analysis we focus on the workup step, which is ubiquitous in many practices and is the first contact between patient and staff after check-in. We measure the durations of workups by extracting anonymized time stamp data from logged activities in electronic health records (EHR). The EHR durations are compared to a set of in-person observations of workups to verify consistency. We then use a phase-type distribution fitting algorithm [8] to parameterize a distribution for workup durations. These

*Research supported by the National Science Foundation, under the Graduate Research Fellowship Program, Grant No. DGF 1106401 J.

R. B. Fricks is with the Department of Biomedical Engineering, Duke University, Durham, NC 27705 USA (e-mail: rafael.fricks@duke.edu).

H. Tseng, is with the Department of Ophthalmology, Duke University Health System, and Duke University School of Medicine, Durham, NC 27705 USA. (e-mail: henry.tseng@duke.edu).

M. Veihl, is with the Department of Gastroenterology, previously Ophthalmology, Duke University Health System, Durham, NC 27705 USA. (e-mail: marjorie.veihl@duke.edu).

K. S. Trivedi, is with the Department of Electrical and Computer Engineering, Duke University Health System, Durham, NC 27705 USA. (e-mail: ktrivedi@duke.edu).

R. C. Barr is with the Department of Biomedical Engineering, Duke University, Durham, NC 27705 USA (e-mail: roger.barr@duke.edu).

parameterized distributions are finally combined in a stochastic reward net (SRN) model which is solved by discrete event simulation, using a modification of a previous study in this setting [9]. The SRN model in this current study can be reconfigured to represent the resource demands of adjacent clinics explicitly. Finally, we quantify the tradeoff in accuracy by computing the mean root squared error (MRSE) between model and data sets when adjacent, concurrently operating clinics are represented or omitted.

II. Methods

In this section we outline a multi-tiered approach for producing robust models of health care service. In-person observation and input from physicians, administrators, and staff informed the three-step generalization of visits to glaucoma clinics. From observation, the workup step is emphasized due to its clear delineation for measurement purposes, as well as its criticality in setting the pace of patient flow. We now detail how the workup step is measured, modeled, parameterized, and validated.

A. Data Collection

Workup durations are measured through event logging by staff using existing point-of-care stations. In point-of-care event logging, clinic staff use additional inputs added to their typical workflow in the electronic health record interface to log the start and end times of defined treatment steps, such as workup. Staff providing treatment are solely responsible for logging events. Such a system has been in place at Duke Eye Center since late 2015, and yielded approximately 22,000 raw observations across all specialties in the first three-month period available, January through March of 2016. We focus on 3082 glaucoma clinic visits during this period, which encompasses the previous study [9]. From this period, we extract 12 complete schedules for patient arrivals to one high-volume glaucoma clinic. This clinic sees consistent traffic, averaging 56 patients per day, while sharing technicians with an average of 33 additional glaucoma patients from other practitioners.

Compared to external observers, event logging has allowed for continuous measurement with relatively little overhead. Logging reduces ambiguity in start and end times as staff recognize when treatment has been performed. There are disadvantages in coverage and accuracy. In our experience, this form of entry is particularly subject to staff cooperation. Logging activities tend to be omitted during peak periods as patient care always takes priority, and paradoxically these periods are critical to quality improvement decisions. Records with inaccurate start, finish, or other omissions are difficult to identify without additional context. Other forms of context, such as patient complexity, may be eliminated in anonymization steps before data analysis. Models must estimate treatment duration as well as account for system dynamics such as staff occupancy during peak periods, with potentially imperfect data.

Given the concern for potential inaccuracies in reported workup durations, we independently verify these timings by comparing to in-person observation. During a nine-month period spanning 2017, 97 observations were gathered from shadowing eye center clinics using a data collection form. The mean root squared error between observation to model, and EHR data to model, were compared to independently verify that EHR samples are representative of the population.

B. Model Formulation and Parameterization

We wish to model the time until workups are complete, in a manner that still provides insight into system dependencies and can be used for planning resource allocations. Consider the duration of a patient visit from check-in to workup completion as a random variable Y with probability density $f_y(t)$. Since the workup is the first step in a patient visit, Y is the sum of how long the workup procedure itself took, and any wait time until the workup began (once a technician begins a patient exam it is effectively not interrupted or restarted). Denoting the delay time X_d and the workup time X_w also as random variables, we assume that the delay duration is independent of the workup duration, and can write an expression for the probability density function of completion time Y as a convolution (1-2)

$$Y = X_d + X_w \tag{1}$$

$$f_Y(t) = f_{X_d}(t) * f_{X_w}(t) \tag{2}$$

EHR logging provides several direct measurements of the workup time, X_w, as well as the completion time Y. The X_w data are used to parameterize a phase-type distribution from the class described in [8], using the implementation in BuTools 2 software package [10]. Phase-type distributions have been effective in fitting casual measurements for similar applications [11, 12]. Y data is used in validation of the composite model, which estimates the convolution result (2).

In estimating the delay distribution, which is dependent on the number of idle technicians, patients, and residual workup time currently in the system, we account for the delays by limiting the number of available technicians, noting the patients-to-technician ratio (PTR). The model hypothesizes that delays in service arise due to unavailable staff, which is evaluated in results. Resource blocking is modeled by the SRN diagram in Figure 1. This SRN uses the phase-type distributions to model workup durations in place of standard timed transitions. Additionally, black-filled timed transitions indicate entry points for the emulated arrivals, which are extracted directly from the EHR records. These two accommodations are departures from strict SRN formalisms, but provide a useful shorthand for the implemented simulation while maintaining a clear system logic.

The SRN is solved using our implementation of discrete event simulation in MATLAB 2017a (Mathworks, Inc.). Each schedule is simulated 10000 times, for both model

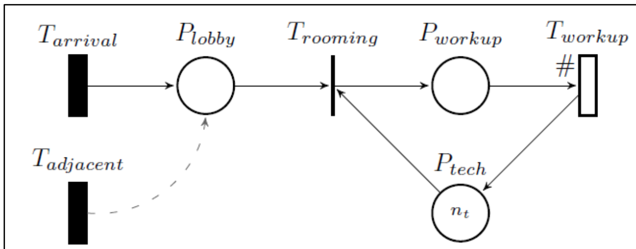

Figure 1. Glaucoma clinic entry SRN diagram. Emulated schedules generate patients in the lobby, which then proceed through workup as techs become available. Adjacent arrivals can be toggled to examine the impact, indicated by the dashed arc from the adjacent transition

variations. Each simulation run executes an input schedule by generating random number deviants, retaining each predicted instance of $x_d,\ x_w,\ y$ from check-in to workup completion.

C. Validation

The proposed model is evaluated using k-folds cross validation [7]. This variant on cross validation is performed by randomizing the order of the original data set at the patient visit level, and then dividing the set into k equal portions referred to as "folds." One fold is selected as the validation set per test, and all other folds are used in training the model. This procedure is repeated k times using each fold as validation set once, providing k opportunities to test a model against independent validation data that was not used in fitting the model, evaluating model performance beyond the training data without undue emphasis on the validation set.

Given a training set $\widehat{Z_1}$ with m_1 observations occurring at time t_{m_1}, and similarly a validation set $\widehat{Z_2}$, denote a hypothesized model for the cumulative distribution function (CDF) of Y trained on $\widehat{Z_1}$ as $H_{\widehat{Z_1}}(t)$. We can compute the training and validation errors of a proposed model as the mean root squared error from model to each observation in the empirical CDF of a set S (denoted as $\hat{F}_S(t)$) using (3-4)

$$Err_{trn} = \frac{1}{m_1}\sum_{1}^{m_1}\sqrt{\left(\hat{F}_{\widehat{Z_1}}(t_{m_1}) - H_{\widehat{Z_1}}(t_{m_1})\right)^2} \qquad (3)$$

$$Err_{val} = \frac{1}{m_2}\sum_{1}^{m_2}\sqrt{\left(\hat{F}_{\widehat{Z_2}}(t_{m_2}) - H_{\widehat{Z_1}}(t_{m_2})\right)^2} \qquad (4)$$

This error can be used to select the model that minimizes the average validation error across all folds. Alternatively, validation criteria can be defined as a model that reduces the average error below a threshold. We report the average error and plot the distribution functions.

III. Results

In-person observations are consistent with EHR event log measurements as well as models estimated from EHR data, with MRSE from Observation to Model of 5.39%; comparing EHR to Model is notably 1.84% (Figure 2). The model from EHR predicts observation timings with minimal error, and so subsequent parameterization uses the larger volume event logging data. All measurements are regarded as precise to the nearest minute in subsequent analysis.

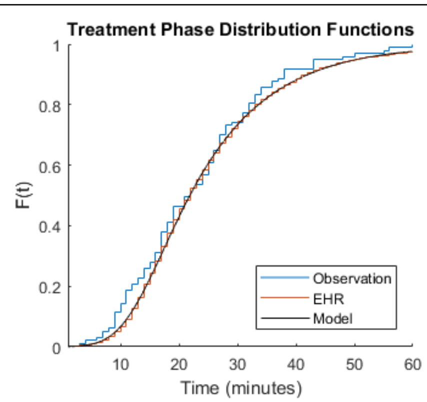

Figure 2. Comparison of in-person observations to event logging collected through electronic health record (EHR) system. These data are redundant measures of the duration of a workup, once a technician begins seeing a patient. Empirical data are evaluated against a phase-type distribution model of workup duration, estimated from EHR data.

From the original set of 3082 visit data, 2624 valid estimates of Y and 2360 valid estimates of X_w are retained after excluding record imperfections such as missing time stamps. A phase-type distribution with 15 phases is used to model workup durations, parameterized from training data in each fold. One parameterization's results are plotted alongside the empirical data (Fig. 2). Five folds were sampled from the EHR data, all folds showed comparable performance.

Table 1 displays results from the simulation, where three variations of the model (Fig. 1) are considered. The concurrent model has the adjacent arrivals enabled, which represent patients in clinics outside of the primary high-volume clinic studied. These results are contrasted to an independent clinic model (Ind. 4T/Ind. 3T) which disables adjacent arrivals and models dedicating technicians solely to one clinic. The 3T/4T designations refer to the number of technicians assigned, to provide a range of similar patient-to-technician ratio (PTR) that provide upper and lower bounds to the optimal concurrent model PTR. Finally, error is reported as a percent, as the MRSE as computed can be interpreted as the average vertical distance between empirical data distribution and model. This distance, measured in percent, represents the average percentile deviation from a model estimate to a typical data set.

IV. Discussion

Choosing which details to represent plays a pivotal role in producing robust predictive models. New modeling techniques [12], as well as wide access to powerful tools for simulation [1] and data fitting [8] make complex models readily available. Compared to the previous study [9] modeling this clinic, the model evaluated here has a drastically simpler SRN structure, but provides remarkable accuracy. Quantitative evaluation is critical in determining when adding model complexity has raised model accuracy.

Accuracy assessments must account for inherent flaws in data collection and potential variations or outliers in the measurand. One interpretation of overfitting phenomena is that accuracy has been improved with respect to a sample by representing measurement noise. Adding more complexity that improves accuracy with respect to one sample may

TABLE I. SIMULATION SYSTEM PARAMETERS AND RESULTS

Metric	Model		
	Concurrent	***Ind. 4T***	***Ind. 3T***
Average[a] Total Patients	89.08	56.08	56.08
Technicians Assigned	5	4	3
Patient-Tech Ratio	17.82	14.02	18.69
Average[b] Training Error	2.38%	15.58%	12.50%
Average[b] Validation Error	2.81%	15.48%	12.60%

a. Average of total patients in all twelve arrival schedules

b. Average of error computed in each fold cross-validation test diminish accuracy when evaluated versus a similar, independent sample. Independent evaluations in cross-validation provide an indication of when overfitting may occur, and are necessary to produce robust models that remain predictive outside of the initial sample.

The consistent reduction in error seen by explicitly modeling concurrency indicates a worthwhile addition of complexity to improve accuracy. The concurrent model minimizes error compared to the independent alternatives, by approximately 10-12%. There is little discernible change in training versus validation errors, further indicating a robust model that continues to perform as anticipated when new data is presented. Interpreted in this application, the model will accurately predict patient flow in subsequent clinic days. Plotting the predicted distributions along the measured data from an arbitrary fold (Fig. 3), the concurrent model results are clearly distinguished from the alternatives. While PTR values are similar, concurrency better captures the transient availability of technicians, where moment-to-moment a probabilistic number of technicians is available to any given clinic. Internally reviewing simulation traces further emphasizes this point. We conclude that representing explicit dependence improves accuracy by an appreciable margin and confirms the need for interconnected models.

Lastly, we find that independent in-person observations of workup durations do not substantially deviate from EHR data. The methods sharply differ in data acquisition rate, where event logging data collection is approximately 200 times faster than in-person observation, without the need for additional personnel. Such substantially larger data sets potentially capture more patient variety than observation alone, which can be used to further refine models. Given the cost of continued in-person observation, we conclude from these results that EHR data are sufficient to produce models that accurately reflect clinic flow.

V. Conclusions and Future Work

Decisions at outpatient centers impact hundreds of patients daily. As clinic flow data become available, more centers will form models to test plans. Rigorous evaluation ensures more robust predictive models. Cross-validation techniques can be used to evaluate when simulative models capture sufficient detail. We demonstrate one application in modeling patient intake at a high-volume glaucoma clinic.

Future work extends this model to subsequent treatment phases, as well as verifying that the model generalizes to other practices as additional data sets are incorporated in the cross-validation scheme. Highly predictive clinic models provide tantalizing prospects in optimization. More complex models require additional research into decision support and efficient optimization.

Acknowledgment

The authors would like to thank the physicians, staff, and administration at Duke Eye Center. We would also like to thank the participation of undergraduate volunteers from Duke University in shadowing the clinics and providing observational data.

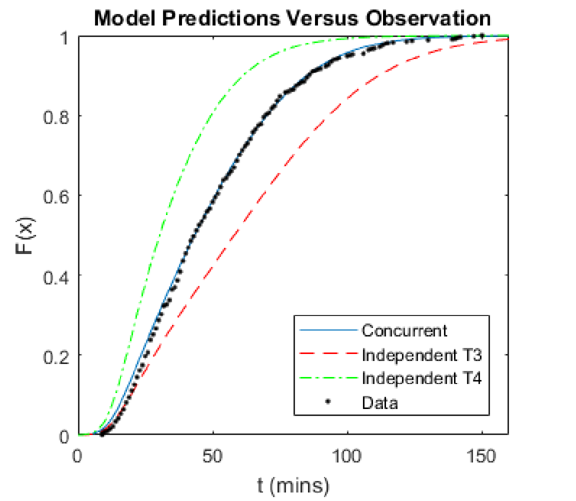

Figure 3. Concurrent clinics model versus independent clinic approximation with 3 and 4 technicians. Validation set data consistently aligns with the concurrent model predictions in all folds.

References

[1] M. M. Gunal and M. Pidd, "Discrete event simulation for performance modelling in health care: a review of the literature," *J of Sim,* vol. 4, pp. 42-51, 03//print 2010.

[2] J. B. Jun, S. H. Jacobson, and J. R. Swisher, "Application of Discrete-Event Simulation in Health Care Clinics: A Survey," *The Journal of the Operational Research Society,* vol. 50, pp. 109-123, 1999.

[3] D. Fone, S. Hollinghurst, M. Temple, A. Round, N. Lester, A. Weightman*, et al.*, "Systematic review of the use and value of computer simulation modelling in population health and health care delivery," *J Public Health Med,* vol. 25, pp. 325-35, Dec 2003.

[4] E. Alfonso, X. Xie, V. Augusto, and O. Garraud, "Modeling and simulation of blood collection systems," *Health Care Manag Sci,* vol. 15, pp. 63-78, Mar 2012.

[5] M. R. Hribar, S. Read-Brown, I. H. Goldstein, L. G. Reznick, L. Lombardi, M. Parikh*, et al.*, "Secondary use of electronic health record data for clinical workflow analysis," *Journal of the American Medical Informatics Association,* vol. 25, pp. 40-46, 2018.

[6] H. Takagi, Y. Kanai, and K. Misue, "Queueing network model for obstetric patient flow in a hospital," *Health Care Management Science,* vol. 20, pp. 433-451, September 01 2017.

[7] T. Hastie, *The elements of statistical learning : data mining, inference, and prediction*. New York: Springer, 2009.

[8] A. Thummler, P. Buchholz, and M. Telek, "A novel approach for fitting probability distributions to real trace data with the EM algorithm," in *2005 International Conference on Dependable Systems and Networks (DSN'05)*, 2005, pp. 712-721.

[9] R. B. Fricks, H. H. Tseng, M. Pajic, and K. S. Trivedi, "Transient performance & availability modeling in high volume outpatient clinics," in *2017 Annual Reliability and Maintainability Symposium (RAMS)*, 2017, pp. 1-6.

[10] G. Horvath and M. Telek, "BuTools 2: a Rich Toolbox for Markovian Performance Evaluation," presented at the proceedings of the 10th EAI International Conference on Performance Evaluation Methodologies and Tools on 10th EAI International Conference on Performance Evaluation Methodologies and Tools, Taormina, Italy, 2017.

[11] M. Fackrell, "Modelling healthcare systems with phase-type distributions," *Health Care Management Science,* vol. 12, p. 11, 2008.

[12] K. S. Trivedi and A. Bobbio, *Reliability and availability engineering : modeling, analysis, and applications*. New York, NY, USA: Cambridge University Press, 2017.