Innovations in Chronic Care Delivery Using Data-Driven Clinical Pathways

Yiye Zhang, MS; and Rema Padman, PhD

According to the World Health Organization, 60% of all deaths, worldwide, can be attributed to chronic diseases such as diabetes, heart disease, stroke, and cancer; they are also a major cause of poverty and lack of economic development.[1] As part of a multi-pronged effort to address this challenge, innovations in chronic care delivery are beginning to leverage advanced statistical and machine learning models and algorithms to obtain new insights into care quality, outcomes, and cost.[2-4] Machine learning is the science of constructing algorithms that learn from large volumes of data in order to facilitate decision making by generating potentially new insights; it has gained widespread implementation across many industries today.[5] Just a few examples of machine learning applications are speech recognition, self-driving cars, and personalized online experiences.[6-8]

Although innovations driven by machine learning have seen tremendous success,[9,10] subsequently resulting in improved service performance, productivity, and growth,[11-13] for a variety of reasons, the healthcare industry has been relatively slow to incorporate these techniques into decision-support applications and to adapt to resulting changes.[14-16] For instance, in making treatment decisions, many clinicians may prefer to use clinical practice guidelines (CPGs) over predictions generated by machine learning algorithms—algorithms which may seem like a "black box" with little relevance to actual clinical decision making.[17] However, many of the current clinical decision support capabilities, whether CPG-embedded electronic health record (EHR) interactivity or computerized provider order entry (CPOE) application, are designed by humans and target the "average patient."

As the Precision Medicine initiative states,[18] we are now in an era in which clinical interventions need to be personalized and predictive, and so should decision support recommendations. To meet this objective, it is no longer sufficient to rely on CPGs, often created based on consensus opinions or randomized clinical trials that have strict enrollment criteria. Rather,

ABSTRACT

Objectives: Chronic diseases are common, complex, and expensive health conditions that can benefit from innovations in healthcare service delivery enabled by information technology and advanced analytic methods. This paper proposes a data-driven approach, illustrated in the context of chronic kidney disease (CKD), to develop clinical pathways of care delivery from electronic health record (EHR) data.

Study Design: We analyzed structured and de-identified EHR data from 2009 to 2013 of 664 CKD patients with multiple chronic conditions.

Methods: Machine learning algorithms were used to learn data-driven and practice-based clinical pathways that cluster patients into subgroups and model the co-progression of their encounter types, diagnoses, medications, and biochemical measurements. Given a pattern of biochemical measurements, our algorithm identifies the most probable clinical pathways, and makes predictions regarding future states, with and without temporal information. CKD stages, their complications, and common medications are included in the clinical pathways.

Results: Using the EHR data of 664 patients who were initially in CKD stage 3 and hypertensive, we identified 7 patient subgroups—each distinguished primarily by the type of complications suffered by the patients. Our algorithm demonstrates fair accuracy (up to 44% and 75%, respectively) in learning the most probable clinical pathways and predicting future states associated with temporal patterns of biochemical measurements and patient subgroups.

Conclusions: Data-driven clinical pathway learning summarizes multidimensional and longitudinal information from EHRs into clusters of common sequences of patient visits that may assist in the efficient review of current practices and identifying potential innovations in the care delivery process.

Am J Manag Care. 2015;21(12):e661-e668

Take-Away Points

- The availability of high-volume, time-stamped, and individual-level health data is beginning to facilitate clinical interventions that are personalized and predictive.
- Healthcare service delivery can benefit greatly from advanced statistical and machine learning models and algorithms that can learn personalized insights from electronic health record (EHR) data.
- Data-driven clinical pathways that describe the co-progression of encounter types, diagnoses, medications, and individual biochemical measurements can be learned from EHR data, using statistical and machine learning methods to support the review of current practices and innovate healthcare delivery approaches.
- Our proposed methodology is generalizable to other clinical conditions and can accommodate varying numbers of clinical and other relevant factors.

with the tremendous amount of data being accumulated in EHRs from the enactment of the Health Information Technology for Economic and Clinical Health (HITECH) Act as part of the American Recovery and Reinvestment Act,[19] healthcare service delivery can also benefit greatly from advanced statistical and machine learning models and algorithms that can learn potentially useful insights from large amounts of highly detailed data collected daily, as part of routine care delivered in multiple, diverse settings.

Traditional topics in machine learning include classification and unsupervised learning.[5] Classification refers to the method of labeling unknown data to target variables through training a classification model using labeled data. Logistic regression and naïve Bayes are examples of classification algorithms.[5] For example, Lee et al used logistic regression to predict 7-day mortality from heart failure in emergent care using initial vital signs, clinical and presentation features, and laboratory tests.[20] Unsupervised learning refers to the identification of latent groups in the data. Unlike classification, which is also called "supervised learning," unsupervised learning does not have true labels, typically does not have true labels, and users need to predefine the number of latent groups. K-means and hierarchical clustering are 2 of the most common unsupervised learning algorithms.[5]

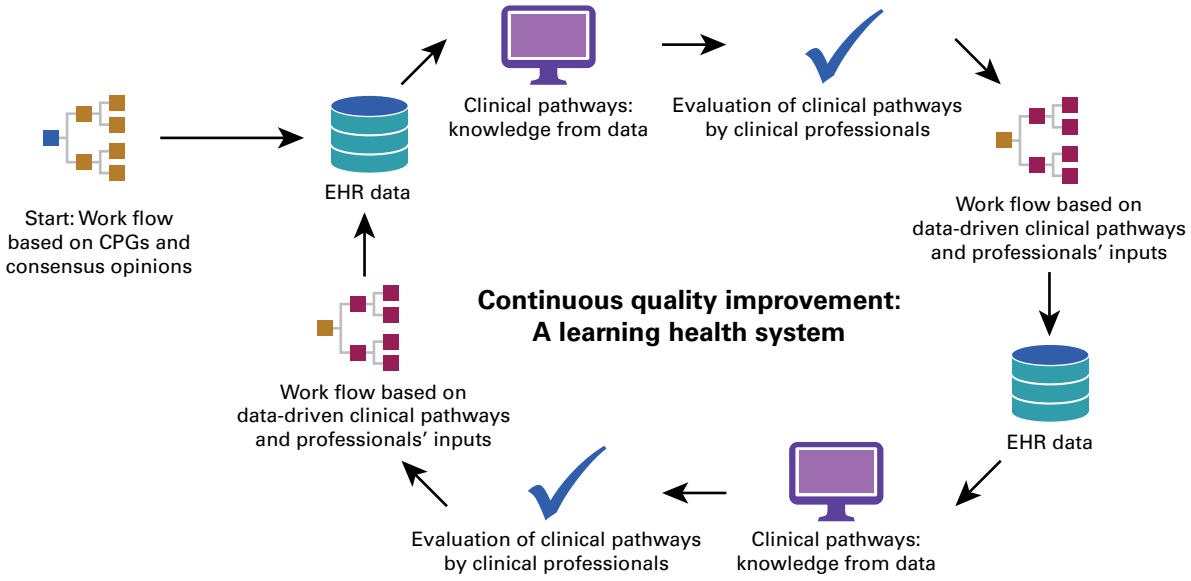
Zhang et al used a variant of the K-means clustering algorithm to design more efficient order sets from historical order data in a pediatric inpatient setting.[21] Order sets are groups of relevant orders traditionally clustered together by clinical experts and used within CPOE; this is an example of a manually designed healthcare information technology application that requires significant labor- and knowledge-intensive effort for maintenance and update. In the same study, Zhang et al demonstrated that order sets can be created using machine learning algorithms, with the resulting data-driven order sets requiring less physical and cognitive workload in usage because the methods were trained to find the optimal combinations of orders that matched, with order data generated from actual work flow. In addition to these classical approaches, many advanced machine learning algorithms have been developed and applied over the years to facilitate a more efficient, safer healthcare system.[22-25]

In this paper, we present a machine learning approach for learning the most probable, data-driven clinical pathways from the EHR data of patients with chronic kidney disease (CKD), and predicting the most probable upcoming interventions at any stage, given recent history. CKD is a chronic condition that currently affects more than 26 million US adults, with an additional 73 million at increased risk for the disease.[26] It is also associated with increased risk for cardiovascular disease and acute kidney injury (AKI), and the majority of the patients also suffer from comorbidities such as hypertension and diabetes.[26] Consequently, CKD management is complex and expensive, and a large proportion of the US Medicare budget every year is allocated for the treatment of CKD.[27] Specifically, the per person per year average cost of treating CKD was $23,128 in 2011—more than twice the average cost of treating non-CKD conditions in the Medicare population ($11,103).[27] With the cost increasing and quality of life decreasing as the disease progresses to end-stage renal disease (ESRD),[27] there is a growing imperative to pursue innovations in service delivery and management of CKD and other chronic conditions that may generate improved health outcomes, cost savings, and patient satisfaction.[4]

Additionally, generating the highest quality scientific evidence and associated practice recommendations for chronic conditions such as CKD is a continuing challenge for the healthcare field.[3] One of the most recent CPGs for CKD was published by the National Kidney Foundation's Kidney Disease Outcomes Quality Initiative in 2012, which is an update of its 2007 guideline. However, of its 7 key recommendations, only 2 recommendations received the highest grade from the Evidence Review Team of the guideline Work Group for strength of recommendation ("recommend" vs "suggest"), and the highest grade for quality of evidence ("high" vs "moderate," "low," "insufficient"), while other recommendations received lower grades for strength of recommendations and for the quality of evidence.[28]

In this paper, we propose that evidence from actual practices, particularly those that include large number of patients in local treatment settings over reasonable durations, may be used to assist guideline development. We present methods for knowledge extraction from data using machine learning algorithms, and demonstrate that such knowledge can be regarded as practice-based, data-driven clinical pathways.

Figure 1. Using Data-Driven Clinical Pathways in a "Learning" Care Delivery Environment



CPG indicates clinical practice guideline; EHR, electronic health record.

Clinical pathways translate CPG recommendations into an actionable plan such as flow charts, and are used by more than 80% of US hospitals for at least 1 intervention.[29] This research aims to develop clinical pathways not strictly based on CPGs, but practice-based evidence learned from data. An overall framework of our approach that supports a learning healthcare system is presented in Figure 1.

METHODS

Prior Work

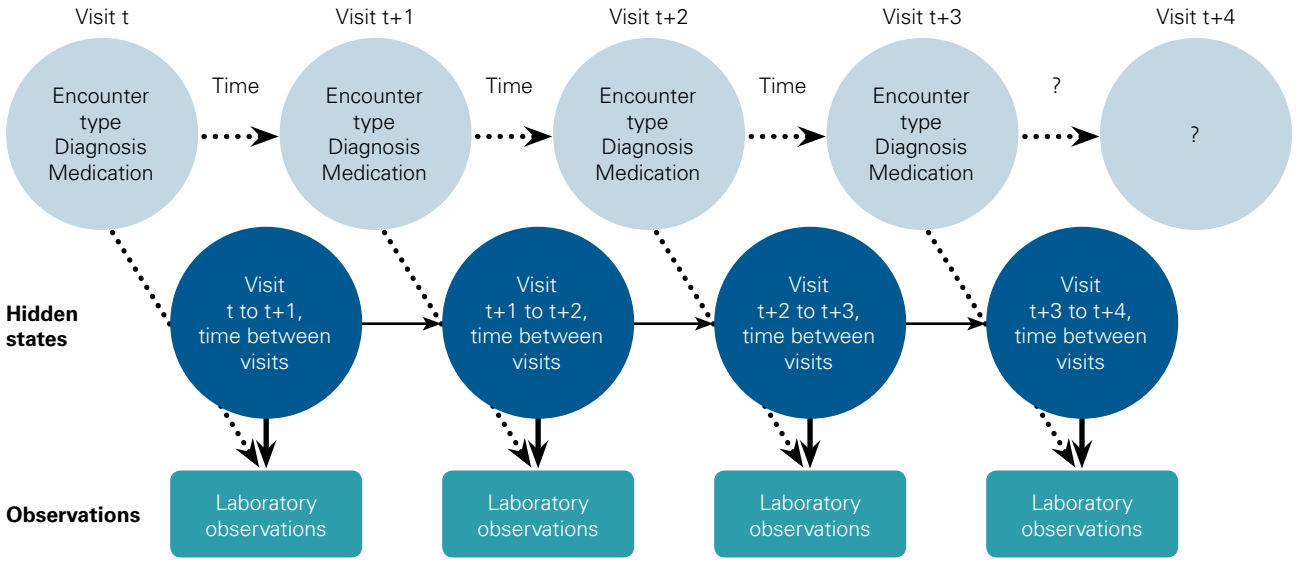
Data-driven clinical pathway learning has garnered research interest since the 1990s,[30-38] but there is limited research on machine learning approaches for the problem. Recently, Lakshmanan et al used a type of clustering algorithm, called DBScan, to cluster patients' history prior to pathway learning, and applied SPAM, an algorithm to find frequent patterns in pathways, to associate patterns with patient outcomes.[33] Huang et al used topic model, a recently developed probabilistic method, for learning latent topics from documents, to discover clinical pathway patterns from EHR event logs.[38] Zhang et al modeled clinical pathways as Markov chains that included the co-progression of multiple interventions and diagnoses, and visualized them to allow identification of variations in care and outcomes across latent patient subgroups.[39]

In this paper, we combine clustering and temporal modeling to elicit common clinical pathways from the data. Specifically, given patient characteristics and a sequence of laboratory observations from multiple laboratory tests, we illustrate methods to learn the most probable sequence of clinical interventions that are associated with the laboratory observations, and to make predictions about patients' impending conditions as a result of the interventions. This approach allows us to link patients' biochemical responses with clinical interventions and with specific outcomes, thus providing a novel methodology for data-driven clinical pathway learning.

Clustering of Patients

To accommodate the heterogeneity in the patient population and improve model accuracy, we group patients according to similarity of their clinical history prior to pathway learning and prediction. We expect patients' pathways to branch out as their health conditions and corresponding treatments evolve in different ways. Therefore, prior to pathway learning and prediction, we use hierarchical clustering to cluster patients' pathways into subgroups according to longest common subsequence (LCS) distance measure.[40] LCS is the longest subsequence that 2 sequences have in common, while preserving the order of occurrence of the items in the sequences, but items are possibly separated. LCS has been widely applied in biomedical research as a similarity measure used in trajectory analysis and protein sequence analysis.[40] The distance measure, dLCS, is then computed as the difference between the sum of the lengths of 2 sequences and

Figure 2. Modeling the Treatment Process



"?" indicates component is to be predicted; t, the number of visits.

twice their LCS. (Details are in the eAppendix, available at www.ajmc.com.) Hence, dLCS is affected by the length of the identified subsequence, and the lengths of both sequences; for example, given the same length of LCS, dLCS is bigger for 2 long sequences than 2 short sequences. Therefore, clustering using dLCS allows us to group patients who not only share similarity in clinical interventions, but also have similar durations of treatment. The optimal number of clusters is determined using Silhouette, a measure commonly used in cluster analysis.[41] In this study, we consider clusters that have 10 or fewer patients as outliers, and plan to evaluate rare events and exceptions in future research.

Model

Figure 2 illustrates our modeling scheme for learning the clinical pathways. Given the time stamps associated with intervention data recorded in the EHR, we assume that each state in the data-driven clinical pathway is separated by at least 1 time unit (eg, day, week, month), and that each state may contain more than 1 type of intervention. For example, it is typical for a CKD patient to have a follow-up visit in the clinician's office, receive medication prescriptions, and have diagnostic codes assigned to the visit. Our data encoding anticipates such multidimensional and longitudinal features in the data. We assign a unique label for each unique combination of interventions occurring from a visit on the same day, such that patients' clinical interventions that span multiple categories, such as diagnosis, medication prescription, and encounter type, can be transformed into 1-dimensional pathways, as shown in the top row in Figure 2. Naturally, these interventions are related to one another over time in varying degrees. For instance, interventions that occurred within 6 months of each other may be more strongly correlated than those that occurred within 2 years of each other.

In the context of CKD management, we assume that interventions at visit t+2 are dependent on activities at visit t+1 and t, as shown in the middle row in Figure 2. For analytical tractability, and reflecting actual practice in the management of many health conditions, the time intervals between 2 consecutive visits are categorized as: 1) less than 3 months, 2) greater than 3 but fewer than 6 months, or 3) at least 6 months. These assumptions are practice- and condition-specific,[3] but can be readily modified for different settings. Patients' biochemical conditions, as reflected by their laboratory observations, are assumed to be influenced by the interventions, as shown in the bottom row in Figure 2. For the problem of clinical pathway learning described in this study, our goal is to learn the most probable sequence of clinical interventions given to patients with a particular trajectory of biochemical responses. Similarly, the prediction problem is to infer the most probable imminent interventions in the next state—most importantly, diagnostic codes—for these patients.

We model this treatment process as a hidden Markov model (HMM). HMM is a statistical model with a wide range of applications, such as in speech recognition and RNA sequence analysis.[42] It is defined by 5 elements: sequence of hid-

den states, sequence of observations, state transition probability distribution, observation probability distribution, and initial state distribution.[43] HMM is used to represent a process in which a sequence of observations is generated, and each observation is triggered by an underlying process that is hidden to us. For example, given a sequence of a patient's body temperatures, we may assume that the patient's health condition is affecting his or her body temperature. Therefore, the sequence of body temperatures form the observations in HMM, and health conditions represent HMM's sequence of hidden states.

The sequence of hidden states in an HMM has a first-order Markov property, which states that the current state only depends on the previous state.[44] Therefore, we regard the middle row in Figure 2 as the sequence of hidden states and the bottom row as the sequence of observations. Parameters of the HMM, such as transition probabilities of hidden states in the Markov chain, are learned from the data using the expectation-maximization (EM) algorithm.[43] Given HMM parameters, we can perform both the clinical pathway learning and prediction tasks through HMM decoding, which calculates the sequence of hidden states with the highest probability given the sequence of observations and the probability distribution of the model. Details of the model and algorithm are described further in the eAppendix and prior studies.[39]

Table 1. Clinical Pathway Components

Category	Component (number of unique patients)
Encounter type	Office (664), hospital (99), education (28)
Diagnosis	Main diagnoses: CKD stage 1-5 (7, 107, 664, 87, and 4, respectively), ESRD (18), hypertension (664), acute kidney injury (48)
	Complications: hyperparathyroidism (311), anemia (296), proteinuria (149), hyperkalemia (118), acidosis (63), hyperphosphatemia (23), glomerulonephritis (24), urinary obstruction (20), volume depletion (7), rhabdomyolysis (1)
Medication (drug class)	Angiotensin-converting enzyme inhibitors (94), angiotensin II receptor blockers (75), diuretics (133), statins (85)
Laboratory test	Albumin (664), calcium (664), creatinine (664)

CKD indicates chronic kidney disease; ESRD, end-stage renal disease.

RESULTS

Descriptive Statistics

We demonstrate the methodology using a real-world data set of 664 patients, with visits from 2009 to 2013 extracted from the EHR, who suffered from CKD and associated complications. The gender ratio is nearly equal. Over 67% of the patients are aged at least 70 years, and nearly 95% are Caucasian. Components considered as part of clinical pathways and the number of unique patients who had each component in their EHR are listed in Table 1. These components were selected for their relevance in CKD management, per consultation with clinicians, but can be extended to include additional details. All 664 patients had initial diagnoses of CKD stage 3 and hypertension, but not diabetes, and none of the patients had anemia or hyperparathyroidism initially. These patients either progressed to advanced CKD stages and ESRD, or improved to CKD stages 1 and 2. Most of them subsequently developed some of the complications listed in Table 1.

Clustering of Patients

The number of clusters, k, was determined to be 7 using the highest silhouette value (0.189) from hierarchical clustering. Table 2 describes the characteristics of each group in detail, indicating that hierarchical clustering using dLCS was able to divide patients into subgroups that differ on treatment frequency, duration, and outcome at the end of the study period. For example, 95% of the patients in subgroup 5 showed improvement in their conditions at the end of the study period, while none worsened, after being in the clinic for an average of 26.9 months. Subgroup 3 is the largest subgroup, and it also has the smallest average dLCS, suggesting that patients are more similar to one another compared with other subgroups. Subgroup 2, which needs to be investigated further, had a mixture, with 14% of patients who improved and 20% who worsened. The final column in Table 2 lists complications of CKD that the majority of patients suffer from in each group.

Clinical Pathway Learning and Prediction

Table 3 summarizes the accuracies associated with predicting the imminent interventions and diagnoses, such as prescription of diuretics and episodes of AKI, and learning the most probable pathways for sample subgroups 3, 4, and 5. We chose these 3 subgroups because of their larger subgroup sizes, and interesting final outcomes at the end of the study. We tested the accuracies using the most common sequence of laboratory observations (LOs) from 3 consecutive visits, and the number of patients who experienced such patterns is listed under the column, "Number of patients who had LOs." Training and testing were performed through a variant of the leave-one-out cross-validation method.[45] Learning and prediction were done with respect to the most common sequence of LO in each subgroup. It is interesting to note that the common biochemical patterns in subgroups 3 and 4

Table 2. Patient Subgroup Characteristics

Group	Patients, n	Visits, m (minimum, maximum)	Average dLCS	Average Treatment Duration (months)	Improved Patients, n (%)	Worsened Patients, n (%)	Common Complications
1	77	2, 14	7.0	31.1	1 (1%)	8 (10%)	Hyperparathyroidism, proteinuria
2	149	2, 15	6.3	19.1	21 (14%)	30 (20%)	Hyperparathyroidism
3	219	2, 17	4.2	29.8	2 (1%)	9 (4%)	None
4	82	2, 15	5.4	24.7	2 (2%)	10 (12%)	Anemia, hyperparathyroidism
5	41	2, 12	5.0	26.9	39 (95%)	0 (0%)	None
6	51	2, 16	5.7	27.4	3 (6%)	5 (10%)	Proteinuria
7	45	2, 18	5.9	28.2	0 (0%)	5 (11%)	Anemia

dLCS indicates longest common subsequence distance measure; m, minutes.

are the same, but the model identified different clinical pathways for these 2 groups, which require further examination. "Pathway with time"/"Pathway without time" measure accuracy of learning the entire pathway, including/not including the actual time duration between 2 visits, respectively. Similarly, "Future visit with time"/"Future visit without time" measure the prediction accuracy for patient's future interventions, with/without time durations between visits. Each state variable contains information on the presence or absence of 3 encounter types, 19 diagnoses, and 4 drug classes, in addition to 3 different durations between visits. Therefore, the probability of accurate learning and prediction, on a random try, is extremely low compared with the results from our algorithm.

We also examined the false negative and false positive rates in the prediction of an imminent condition such as AKI. We define a false negative to be a case where patients' CKD stages are worse than predicted, or patients developed AKI, which our methods failed to predict. A false positive is defined as patients' CKD stages being better than predicted, or prediction of AKI when no AKI developed in reality. We include AKI in this analysis because it is a serious adverse outcome: it often requires hospitalization and can be fatal.[46] We were able to obtain false-positive and false-negative rates that are as low as 0%, although this result needs to be validated using a much larger sample. Nevertheless, the learning and prediction algorithms show promise in identifying common pathways of treatments, but these need to be analyzed further to better delineate effective interventions in the various subgroups.

DISCUSSION

This paper provides a brief overview of machine learning approaches to assist medical decision making, and introduces a methodology, as well as an application that illustrates the development of data-driven clinical pathways through mining of EHR data. This approach may facilitate timely extraction of potential new evidence that could become the basis for new clinical trials, and may also serve as "shared baselines" to be used within a local practice for work flow and population health management.[47] Patient-focused applications derived from our research, particularly those that visualize the clinical pathway and provide related patient-oriented recommendations and educational resources, may enhance patients' understanding of their diseases and treatments, thus facilitating shared decision making.

An important ongoing study is to develop prediction models for other significant outcomes of interest in the management of CKD and its complications. Also, we need to evaluate these data-driven clinical pathways, especially their divergence and rare events, and their predictions with input from clinical professionals. As a growing number of healthcare organizations pilot new care delivery and payment models, such as the accountable care organizations,[48] exploring disease trajectories that incorporate the interactions of clinical interventions and their associated outcomes may also provide useful insights on the cost effectiveness of treatments, which organizations can leverage for implementing innovative care delivery practices.

A crucial prerequisite for success in the application of advanced machine learning methods to healthcare delivery is data quality. It is not uncommon for computational scientists to spend significant effort in cleaning EHR data before analysis. In addition, even after months of processing, there are often still missing data and errors, some arising from the mismatch between actual work flows and process assumptions, subjecting the analytical results to bias. Such inefficiency can be minimized by careful obser-

Table 3. Performance of the Clinical Pathway Learning and Prediction Models

Sub-Group	Most Probable Pathway Patterns	Number of Patients Who Had LOs	Prediction Accuracy With and Without Time Between Visits					
			Pathway With Time	Pathway Without Time	Future Visit With Time	Future Visit Without Time	False-Negative Rate	False-Positive Rate
3	V284[a]-V284-V284, (all at least 6 months apart)	43	35%	44%	63%	67%	0%	0%
4	arb873[b]-arb988[c] (at least 6 months apart) –V2026[d] (3-6 months apart)	14	7%	7%	36%	50%	0%	0%
5	V609[e]-V609-V609 (all 3-6 months apart)	12	17%	42%	25%	75%	0%	25%

LO indicates laboratory observation. V284, arb873, arb988, V2026 and V609 are encoding of unique visits.
The following are listed in order of encounter type; diagnosis; medication:
[a]V284: office visit; chronic kidney disease (CKD) stage 3, hypertension; no medication prescription.
[b]arb873: office visit; CKD stage 4, hypertension, anemia, acidosis, hyperphosphatemia, volume depletion, glomerulonephritis, urinary obstruction; angiotensin receptor blocker (ARB).
[c]arb988: office visit; CKD stage 3, hypertension, anemia, acidosis, hyperphosphatemia, volume depletion, glomerulonephritis, urinary obstruction; ARB.
[d]V2026: office visit; CKD stage 3, hypertension, anemia, hyperparathyroidism, acidosis, hyperphosphatemia, glomerulonephritis, urinary obstruction; no medication prescription.
[e]V609: office visit; CKD stage 2, hypertension; no medication prescription.

vation and understanding of the care delivery context, and planning of the data storage with a range of options available depending on the data size.[49] At the same time, methods have been developed, such as imputation and approximate inference algorithms, that can accommodate missing data. For example, in this paper, we used the EM algorithm to infer the parameters of HMM. Furthermore, diversity is innate to most healthcare data, and we found it to be one of the biggest challenges in accurately inferring clinical pathways, requiring large amounts of data and robust methods for analysis and inference. In this paper, we examined encounter type, diagnosis, medication prescriptions, and biochemical measurements, but our data representation is flexible with regard to the number of clinical factors of interest. Therefore, when sufficient curated data becomes available, factors such as medical expenses and behavioral information can also be incorporated to enrich the learned pathways and personalized predictions of health and cost outcomes.

CONCLUSIONS

This paper presents additional promising evidence of the potential of machine learning applications for clinical decision making. We develop and demonstrate a methodology to facilitate more targeted management of patients with complex chronic conditions using data-driven clinical pathways. Clinical pathways are learned from a healthcare organization's EHR data by summarizing multidimensional clinical history as chronologically organized sequences, capturing information on the co-progression of encounter types, diagnoses, medications, and biochemical measurements. Further, we link clinical pathways to a few outcomes within subgroups of patients with reasonable accuracy using hierarchical clustering and HMM. Applying our methodology to relevant EHR data on 664 patients with CKD stage 3 and hypertension, we identify clinical pathways that may be compared with current CPG recommendations in future studies, and contribute to the development of shared-baseline within hospitals. These methods and broad findings from EHR data are generalizable and can be adapted to other clinical conditions to support efficient review of treatments and outcomes and to aid clinical professionals and patients in making more informed treatment and management decisions.

Acknowledgments

The authors are very grateful to the forward-thinking physicians and staff of the community nephrology practice, Teredesai, McCann & Associates, PC, in Western Pennsylvania, who generously provided detailed, de-identified data from their 20-year electronic health record for this study. We particularly thank Pradip Teredesai, MD, FACP; Qizhi Xie, MD, PhD; Nirav Patel, MD; and staff members Linda Smith and Audra Barletta, who gave us important clinical and technical information about the data and the key characteristics of CKD, AKI, and their treatments. This study was designated as Exempt by the Institutional Review Board at Carnegie Mellon University.

Author Affiliations: The H. John Heinz III College, Carnegie Mellon University (YZ, RP), Pittsburgh, PA.

Source of Funding: This study is part of a doctoral thesis at Carnegie Mellon University and has no funding source.

Author Disclosures: Dr Padman and Ms Zhang report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

Authorship Information: Concept and design (YZ, RP); acquisition of data (YZ, RP); analysis and interpretation of data (YZ, RP); drafting of the manuscript (YZ); critical revision of the manuscript for important intel-

lectual content (YZ, RP); statistical analysis (YZ); administrative, technical, or logistic support (RP); and supervision (RP).

Address correspondence to: Yiye Zhang, MS, The H. John Heinz III College, Carnegie Mellon University, 4800 Forbes Ave, Pittsburgh, PA 15213. E-mail: yiyez@andrew.cmu.edu.

REFERENCES

1. Chronic diseases and health promotion. World Health Organization website. http://www.who.int/chp/en/. Published 2015. Accessed November 6, 2015.

2. Abra G, Patel M, Moore D, et al. Trend-bearing Chronic Kidney Disease Care Model. Stanford University website. http://cerc.stanford.edu/fellowships/docs/CERC4modelsummary2.11.2013_3PMpdf.pdf. Published 2013. Accessed November 6, 2015.

3. Zhang Y, Padman R, Wasserman L, Patel N, Teredesai P, Xie Q. On clinical pathway discovery from electronic health record data. *IEEE Intelligent Systems.* 2015;30(1):70-75.

4. Saria S. A $3 trillion challenge to computational scientists: transforming healthcare delivery. *IEEE Intelligent Systems.* 2014;29(4):82-87.

5. Bishop CM. *Pattern Recognition and Machine Learning.* New York, NY: Springer-Verlag; 2006.

6. Rabiner L, Juang B-H. *Fundamentals of Speech Recognition.* Upper Saddle River, NJ: Prentice Hall; 1993.

7. Stavens D, Thrun S. A self-supervised terrain roughness estimator for off-road autonomous driving. Presented in: proceedings of Conference on Uncertainty in AI (UAI); July 13-16 2006; Cambridge, MA. https://dsl-pitt.org/uai/papers/06/p469-stavens.pdf. Accessed November 6, 2015.

8. Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing.* 2003;7(1):76-80.

9. Kusiak A. Innovation: a data-driven appraoch. *Internatl J Production Econom.* 2009;122(1):440-448.

10. Miles I. Innovation in services. In: Fagerberg J, Mowery DC, Nelson RR, eds. *The Oxford Handbook of Innovation.* New York, NY: Oxford University Press; 2005.

11. Miles I. Service Innovation. In: Maglio PP, Kieliszewski CA, Spohrer JC, eds. *Handbook of Service Science.* New York, NY: Springer; 2010:511-533.

12. Hall BH, Lotti F, Mairesse J. Innovation and productivity in SMEs: empirical evidence for Italy. *Small Bus Econ.* 2009;33(1):13-33.

13. Tether B, Howells J. Changing understanding of innovation in services. *Innvoation Services.* 2007;9:21-60.

14. Goldzweig CL, Towfigh A, Maglione M, Shekelle PG. Costs and benefits of health information technology: new trends from the literature. *Health Aff (Millwood).* 2009;28(2):w282-w293.

15. Shekelle PG, Morton SC, Keeler EB. Costs and benefits of health information technology. *Evid Rep Technol Assess (Full Rep).* 2006(132):1-71.

16. Kayyali B, Knott D, Kuiken SV. The big-data revolution in US health care: accelerating value and innovation. McKinsey & Company website. http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care. Published April 2013. Accessed November 6, 2015.

17. Moxey A, Robertson J, Newby D, Hains I, Williamson M, Pearson SA. Computerized clinical decision support for prescribing: provision does not guarantee uptake. *J Am Med Inform Assoc.* 2010;17(1):25-33.

18. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med.* 2015;372(9):793-795.

19. HITECH Act enforcement interim final rule. HHS website. http://www.hhs.gov/ocr/privacy/hipaa/administrative/enforcementrule/hitechenforcementifr.html. Accessed November 6, 2015.

20. Lee DS, Stitt A, Austin PC, et al. Prediction of heart failure mortality in emergent care: a cohort study. *Ann Intern Med.* 2012;156(11):767-775.

21. Zhang Y, Padman R, Levin JE. Paving the COWpath: data-driven design of pediatric order sets. *J Am Med Inform Assoc.* 2014;21(e2):e304-e311.

22. Saria S, Rajani AK, Gould J, Koller D, Penn AA. Integration of early physiological responses predicts later illness severity in preterm infants. *Sci Transl Med.* 2010;2(48):48ra65.

23. Chen JH, Podchiyska T, Altman RB. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records [published online July 21, 2015]. *J Am Med Inform Assoc.* 2015. pii:ocv091.

24. Halpern Y, Choi Y, Horng S, Sontag D. Using Anchors to Estimate Clinical State without Labeled Data. Paper presented at: AMIA Annual Symposium Proceedings; November 2014; Washington, DC.

25. Neill DB, Cooper GF. A multivariate Bayesian scan statistic for early event detection and characterization. *Mach Learn.* 2010;79(3):261-282.

26. Coresh J, Selvin E, Stevens LA, et al. Prevalence of chronic kidney disease in the United States. *JAMA.* 2007;298(17):2038-2047.

27. United States renal data system: 2013 atlas of CKD & ESRD. United States Renal Data System website. http://www.usrds.org/atlas.aspx. Published 2013. Accessed November 6, 2015.

28. National Kidney Foundation. KDOQI clinical practice guideline for diabetes and CKD: 2012 update. *Am J Kidney Dis.* 2012;60(5):850-886.

29. Rotter T, Kinsman L, James E, et al. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. *Cochrane Database System Rev.* 2010(3):CD006632.

30. Lin F, Chou S, Pan S, Chen Y. Mining time dependency patterns in clinical pathways. *Int J Med Inform.* 2001;62(1):11-25.

31. Huang CW, Syed-Abdul S, Jian WS, et al. A novel tool for visualizing chronic kidney disease associated polymorbidity: a 13-year cohort study in Taiwan. *J Am Med Inform Assoc.* 2015;22(2):290-298.

32. Zhang Y, Padman R, Wasserman L. On learning and visualizing practice-based clinical pathways for chronic kidney disease. Presented at: American Medical Informatics Association 2014 Annual Symposium; November 2014; Washington, DC.

33. Lakshmanan GT, Rozsnyai S, Wang F. Investigating clinical care pathways correlated with outcomes. In: *Business Process Management.* Berlin, Heidelberg, Germany; Springer; 2013: 323-338.

34. Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. *Artif Intell Med.* 2012;56(1):35-50.

35. van der Aalst WMP, van Dongen BF, Herbst J, Maruster L, Schimm G, Weijters AJMM. Workflow mining: a survey of issues and approaches. *Data Knowl Eng.* 2003;47(2):237-267.

36. Egho E, Jay N, Raïssi C, Nuemi G, Quantin C, Napoli A. An approach for mining care trajectories for chronic diseases. In: *Artificial Intelligence in Medicine.* Berlin, Heidelberg, Germany; Springer; 2013: 258-267.

37. Yang W, Su Q. Process mining for clinical pathway: literature review and future directions. Paper presented at: 11th International Conference on Service Systems and Service Management; June 2014; Bejing, China.

38. Huang Z, Dong W, Ji L, Gan C, Lu X, Duan H. Discovery of clinical pathway patterns from event logs using probabilistic topic models. *J Biomed Inform.* 2014;47:39-57.

39. Zhang Y, Padman R. Patel N. Paving the COWPath: learning and visualizing clinical pathways from electronic health record data [published online September 28, 2015]. *J Biomed Inform.* 2015. pii: S1532-0464(15)00202-6.

40. Elzinga CH. Sequence analysis: metric representations of categorical time series. *Socio-logical Methods and Research.* 2006.

41. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Computational and Applied Mathematics.* 1987;20:53-65.

42. Eddy SR. What is a hidden Markov model? *Nat Biotechnol.* 2004;22(10):1315-1316.

43. Rabiner LR, Juang BH. An introduction to hidden Markov models. *IEEE ASSP Magazine.* 1986;3(1):4-*16.*

44. Karlin S. *A First Course in Stochastic Processes.* San Diego, CA: Academic Press; 2014.

45. Geisser S. P*redictive Inference (Monographs on Statistics & Applied Probability [book 55]).* New York, NY: Chapman and Hall/CRC; 1993.

46. Chertow GM, Burdick E, Honour M, Bonventre JV, Bates DW. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *J Am Soc Nephrol.* 2005;16(11):3365-3370.

47. James BC, Savitz LA. How Intermountain trimmed health care costs through robust quality improvement efforts. *Health Aff (Millwood).* 2011;30(6):1185-1191.

48. Shortell SM, Wu FM, Lewis VA, Colla CH, Fisher ES. A taxonomy of accountable care organizations for policy and practice. *Health Serv Res.* 2014;49(6):1883-1899.

49. Elmasri R, Navathe SB. *Fundamentals of Database Systems.* 6th edition. New York, NY: Pearson; 2010. ■

www.ajmc.com Full text and PDF

eAppendix

This appendix provides some additional details associated with representing time-stamped, high volume, and granular electronic health record (EHR) data, our modeling approach using hidden Markov model (HMM), and grouping patients using hierarchical clustering.

A1. Representation

An extract of the EHR describing a patient's visits is shown in the eAppendix **Table**. We model each visit, *V*, as a list of clinical activities drawn from 4 major components, as follows:

$$V: \{visit\ purpose;\ diagnosis;\ medication;\ procedure\}.$$

Laboratory activity is not part of *V*, as explained in section 2.3. Each unique *V* is given an encoding label, such as *V1*: {Office; CKD stage 3, hypertension; diuretics; Doppler}, and *V2*: {Education; CKD stage 4, hypertension, diabetes; not applicable (N/A); N/A}. Therefore, a patient's (*p*'s) sequence of visits, Q^p, can be represented using the relevant combinations of *encoding labels* alone, in the order in which the visits occur, such as V1-V2-V1-V3. The numbering in the encoding labels simply distinguishes one unique visit from another, and does not represent temporal factors. For example, a sequence may be V2-V2-V1, depending on the actual visit contents of the patient. For clarity, some V-labels, such as *arb873*, start with abbreviated name of the drug class, such as angiotensin II receptor blockers, prescribed from the visit.

Table. EHR Extract for a Patient's Visit

Date	Category	Entry	Laboratory Results
2/1/12	Visit Purpose	Office	Calcium = 8.4-9.5 mg/dL Creatinine = 1.2-1.6 mg/dL Hemoglobin = 13.5+ g/dL
	Procedure	N/A	
	Medication	Angiotensin-converting-enzyme (ACE) inhibitors, diuretics, statins	
	Diagnosis	Acute kidney injury, chronic kidney disease (CKD) stage 3, hypertension	
5/1/12	Visit Purpose	Hospital	Calcium = 8.4-9.5 mg/dL Creatinine = 1.6-2.2 mg/dL Hemoglobin = 13.5+ g/dL
	Procedure	Renal ultrasound	
	Medication	ACE inhibitors, diuretics, statins	
	Diagnosis	AKI, CKD stage 3, hypertension	
6/1/12	Visit Purpose	Office	Calcium = 8.4-9.5 mg/dL Creatinine = 1.6-2.2 mg/dL Hemoglobin = below 13.5 g/dL
	Procedure	N/A	
	Medication	ACE inhibitors, statins	
	Diagnosis	CKD stage 3, hypertension, anemia	

A2. Model

The problem of clinical pathway learning can be stated as follows: Given p patients, each having $m_p = 1,...M$ visits, and an associated set of biochemical measurements, (a) learn the most probable course of treatment for patients with a certain pattern of biochemical changes, and (b) predict the most probable future state of a patient's diagnosis and/or treatments.

Each patient's visits can be chronologically ordered into a sequence using the visit contents and their representation outlined in *A1*. So, collectively for p patients, there are p sequences. In this paper, we model such sequences of visits as a Markov chain.[1] A Markov chain is a sequence of random variables with memory-less property. In a first-order Markov chain, each state in the Markov chain depends only on its previous state. Hence, we make an assumption that treatment decisions are made mostly based on information from 2 previous visits, which corresponds with accepted practices in chronic disease management at the study site. Instead of modeling a second-order Markov chain, we create state variables in which each variable is a transition between 2 visits, such as V1-V2, and V2-V3. We also encode temporal information in our state variables, such that transitions that occurred *less than 3 months*, *3 to 6 months*, and *at least 6 months* apart can be differentiated. The state variable space for the Markov chain can be described as VV = (VV1, VV2, …., VVs), where VV1:{V1-V2 in less than 3 months} and VV2: {V1-V2 in 3 to 6 months}, and so on, with s such possible transitions. The numbering in the encoding labels, such as "VV1" and "VV2", is only to distinguish each VV from one another in the set of *labels* and does not represent time. To illustrate the idea, an actual sequence of VV may be: VV1-VV3-VV1, which translates into: {V1-V2 in less than 3 months}-{V2-V1 in at least 6 months}-{V1-V2 in less than 3 months}. A patient p's sequence Q^p can be represented both in terms of V-labels and of VV-labels.

The Markov chain considered here is time-homogeneous, meaning that the transition probability is independent of the state. Hence we make the assumption that the treatment regime is time-invariant. A time-homogenous Markov chain allows the calculation of the state transition probability distribution, which is the probability of a state transitioning to itself or other states in the chain.

A3. Modeling Treatment Process as HMM

The treatment process can be modeled as an HMM defined by 5 elements: sequence of hidden states, sequence of observations, state transition probability distribution, observation probability distribution, and initial state distribution.[2] Figure 2 from the main text depicts this treatment process. The top row 1 describes the actual sequence of visits for each patient, represented by V-labels (V_i), and the middle row is the sequence of hidden states, represented by VV-labels (VV_j), which we model as a Markov chain as described in *A2*. The bottom row is a

sequence of biochemical measurements (o_k), associated with each visit. In an HMM, the sequence of observations is visible, and each observation is dependent only on its corresponding hidden state in the sequence of hidden states. Observation probability distribution is the probability that observation o_k is emitted from hidden state VV_j. Initial state distribution is the distribution of each state in the first time unit. The solid arrows in Figure 2 show the direct relationships modeled in the HMM, and the dotted lines show the implicit but clinically relevant relationships in the model. HMM takes into account sampling bias, transition probability of hidden states, and probability of each hidden state emitting observations, such that learning and predictions can be more applicable to future data. In this paper, we accept the HMM assumption that observations at time *t* are only dependent on the hidden state at time *t*, but we realize that observations may in fact have an association of the hidden state at time *t+1*. Such structure can be tested using alternate models in future studies. Since the structure of the model is for now assumed to be known from the data that contains missing values, the estimation of parameters θ^* is performed using the EM algorithm,[3] assuming that each sequence Q^p of patient *p* is independent.

A4. Clinical Pathway Learning and Prediction

Clinicians may treat differently patients who come with similar laboratory observations due to a variety of reasons including comorbidities, medications, and practice variations. The treatments, in turn, will directly affect patients' biochemical data. Hence, we are interested in learning the most probable sequence of hidden states as clinical pathways that reflect such structural relationships. Given a sequence of observations $(o_1, \ldots, o_k)$, the most probable sequence of states is found using the Viterbi algorithm, a dynamic programming algorithm for decoding the most probable sequence of hidden states in HMM.[4] In addition, as shown in Figure 2, the set-up of the Markov chain state variables VV_j allows not only identification of the most probable sequence of hidden states, but also prediction of future visit in the actual visit sequence with V_i. Therefore, decoding a state in the Markov chain automatically reveals the content of the 2 visits associated with this state, one known and one unknown, and the time difference between these 2 visits.

Each set of biochemical measurements contains measurements from multiple laboratory tests. They are discretized into appropriate ranges, and combined as one variable, to be a single

observation in the HMM. If patients had multiple measurements of the same laboratory test, such as creatinine = 2.0 mg/dL and creatinine = 2.6 mg/dL in 2 separate readings, we take the average of the 2, 2.3 mg/dL, and categorize the average value into its valid range. In order to select appropriate sequence of measurements, we use Sequential Pattern Discovery using Equivalence classes (SPADE), a type of frequent sequence mining technique.[5] This technique finds patterns of biochemical measurements with high support. Support is the percentage of patients who have a given pattern in their sequences.

To evaluate the model, we apply a modified version of the leave-one-out cross-validation. In each iteration, we train the model using all patients who have had the pattern of interest, except for one. Then we evaluate the pathways and predictions against the test patient. This process is repeated until we have test results for all patients who have had the pattern of interest, and we report the average prediction accuracy, false negative rate, and false positive rate.

A5. Clustering Metric

We use hierarchical clustering to cluster patients into subgroups. Distance matrix used in clustering contains longest common subsequence (LCS) distance between each pair of patient sequences represented with V-labels.[6] Each patient has one and only one sequence. LCS is the longest subsequence that each pair of patient sequences, (Q^i, Q^j), where $1 \leq i, j \leq P$, $i \neq j$, and P is the number of patients, have in common, while preserving the order of occurrence of each item in the sequence, but possibly separated.

$$LCS(Q^i, Q^j) = \max\{|u|: u \in S(Q^i, Q^j)\},$$

where |u| is the length of the common subsequence for the pair of sequences (Q^i, Q^j), S(Q^i, Q^j) is the nonempty set of common subsequences of sequences $Q^i$and Q^j. LCS distance measure, dLCS, is defined as:

$$dLCS(Q^i, Q^j) = |Q^i| + |Q^j| - 2LCS(Q^i, Q^j).$$

eAPPENDIX REFERENCES

1. Norris JR. *Markov Chains (Cambridge Series in Statistical and Probabilistic Mathematics [book 2])*. Cambridge, UK: Cambridge University Press; 1999.

2. Rabiner LR, Juang BH. An introduction to hidden Markov models. *IEEE ASSP Magazine*. 1986;3(1):4-16.

3. Leroux BG. Maximum-likelihood estimation for hidden Markov models. *Stoch Proc Appl*. 1992;40:127-143.

4. Forney GD Jr. (1973). The viterbi algorithm. Proceedings of the IEEE, 61(3), 268-278.

5. Zaki MJ. SPADE: an efficient algorithm for mining frequent sequences. *Machine Learning*. 2001;42(1):31-60.

6. Elzinga CH. Sequence analysis: metric representations of categorical time series. *Sociological Methods and Research*. 2006.