LEARNING ATTRIBUTES OF DISEASE PROGRESSION FROM TRAJECTORIES OF SPARSE LAB VALUES

VIBHU AGARWAL

Biomedical Informatics Training Program, Stanford University Stanford, CA 94305, USA Email: vibhua@stanford.com

NIGAM H SHAH

Center for Biomedical Informatics Research, Shah Lab, Stanford University Stanford, CA 94305, USA Email: nigam@stanford.com

There is heterogeneity in the manifestation of diseases, therefore it is essential to understand the patterns of progression of a disease in a given population for disease management as well as for clinical research. Disease status is often summarized by repeated recordings of one or more physiological measures. As a result, historical values of these physiological measures for a population sample can be used to characterize disease progression patterns. We use a method for clustering sparse functional data for identifying sub-groups within a cohort of patients with chronic kidney disease (CKD), based on the trajectories of their Creatinine measurements. We demonstrate through a proof-of-principle study how the two sub-groups that display distinct patterns of disease progression may be compared on clinical attributes that correspond to the maximum difference in progression patterns. The key attributes that distinguish the two sub-groups appear to have support in published literature clinical practice related to CKD.

1. Introduction

It is common knowledge that diseases manifest differently in different people. Knowing the alternative progression patterns of a disease for a given population, as well as the clinical attributes associated with the patterns, is therefore of interest to patients, doctors as well as researchers¹. Knowing what to expect, empowers patients to make informed choices about their treatment options as well as plan a judicious acquisition of healthcare resources in the future. Furthermore, the ability to spot the unusual, and initiate a clinical evaluation in case the observed symptoms are anomalous with respect to known progression attributes, has the potential to improve the care delivery process. From the perspective of the care provider, knowing the attributes of the different paths of disease progression is essential for investigating risk factors associated with progression².

For a healthcare system preparing to care for an aging population, an understanding of disease paths as the basis for planning treatment can have a profound impact on the patient's wellness goals. For instance, it has been seen that classification of end-stage functional decline into four groups explains the observed patterns in a sample of older medicare decedents³. Insight into the most likely course of progression and the "signature" attributes, can prove invaluable to healthcare professionals. Finally, a knowledge of progression patterns is essential for discovering treatment options that alter disease progression. For instance, the stage duration as well as progression rates between normal aging and severe dementia, assessed via the Global Deterioration Scale in patients with Alzheimer's, show high heterogeneity⁴. A clinical evaluation of prospective therapies that seek to slow the cognitive decline in patients with Alzheimer's would need to be carried out in individuals with similar progression trends.

The general problem is of discovering patterns of clinical events associated with stages of progression and then classes of such sequential patterns. Generally, disease progression modelling efforts first learn a state transition model using comorbidity patterns and later infer the comorbidities that drive progression based on the observed symptoms⁵. However, for many diseases, the disease status can be reliably summarized by recording one or more physiological measures. Univariate measures such as Glycosylated Hemoglobin (Diabetes), Predicted Forced Vital Capacity (Scleroderma) and Estimated Glomerular Filtration Rate (Kidney Disease) are used routinely in medical practice. These measurements are typically recorded irregularly, and usually after long intervals, making the recorded trajectories sparse. For example, out of 18,342 patients with Type 2 Diabetes in our extract of patient data from the Stanford Clinical Data warehouse, only 8231 patients had two or more HbA1c measurements. The mean number of observations per patient was 7.49. An estimate of the disease progression path based on an observed trajectory of such measurements will have high variance. As a consequence, clusters derived from such path estimates are likely to be unstable.

We hypothesize that it is possible to learn clinically meaningful clusters of disease paths from sparse and irregular trajectories of lab values. In our earlier work² we have described a generative model for simultaneously modelling stages of progression in a cohort of Chronic Kidney Disease (CKD) patients, as well as discovering clusters of distinct progression sequence. In related work, there are prior efforts in creating finite dimensional representation of a trajectory captured by dense measurements, and cluster the trajectories using an appropriate similarity metric. Example

of successful path estimation with methods employing Gaussian Process regression often involve measurements in post-operative care or the intensive care unit, where physiological measurements are recorded regularly and relatively few observations are missing^{6–8}.

In order to meaningfully cluster paths estimated from sparse measurement trajectories, it is possible to borrow support from other trajectories provided a large number of trajectories have been recorded for the full time grid. The Functional Clustering Model (FCM) proposed by James and Sugar⁹ models sparse trajectories as random effects, after fitting natural cubic splines to observations from each trajectory. In the work presented here, we cluster creatinine measurements from patients with Chronic Kidney Disease using the FCM. We then compare the distribution of clinical features between patients in different clusters, by defining a time window around the region of maximum discrimination between the clusters. Finally, we examine the features whose distribution is significantly different between clusters, and interpret the differences in the light of published literature on the management of Chronic Kidney Disease. Figure 1 illustrates our approach and overall workflow.

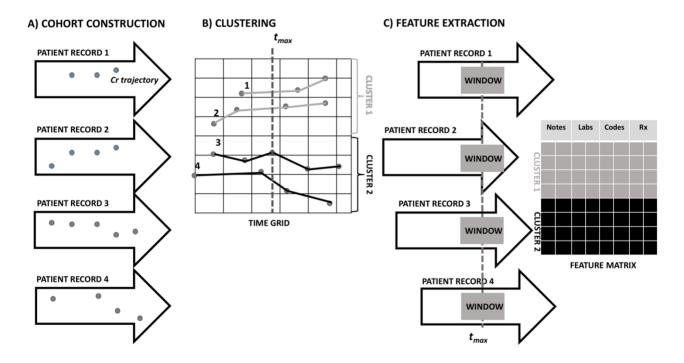


Figure 1A) Records for patients in the CKD cohort B) Clustering sparse trajectories of creatinine values. The time point at which the clustered trajectories are most discriminable is t_{max} C) Text concepts, lab results, ICD9 codes and prescriptions from a window centered at t_{max} in the patient records

2. Data

The patient dataset was extracted from the Stanford clinical data warehouse (SCDW), which integrates data from Stanford Children's Health (SCH) and Stanford Health Care (SHC). The extract comprises 2 million patients, with 49 million encounters, 35 million coded diagnoses and procedures, 204.8 million laboratory tests, 14 million medication orders as well as pathology, radiology, and transcription reports totaling over 27 million clinical notes. Our extract of the deidentified patient data from 01/1994 through 06/2013 from SCH and SHC is stored in a structured and indexed form within a MySQL relational database.

2.1. Cohort selection

In order to select patients with a diagnosis of Chronic Kidney Disease (CKD), we used the presence of the ICD-9 code 585.00 as our filtering criterion. We identified 959 CKD patients through this method, and kept 792 that had 3 or more creatinine measurements. We henceforth refer to this set of 792 patients as our cohort. Examining the sequence of creatinine measurements time over or "trajectories" from our cohort revealed significant sparsity in the creatinine measurements. Figure 2 shows the distribution of of the number of per patient measurements in our cohort, with a mean of 25.98 and median 13.

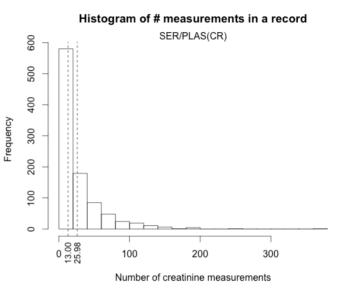


Figure 2 Distribution of number of measurements of creatinine per patient with the mean and median.

3. Methods

3.1. Functional Clustering Model

In order to cluster sparse observations in patient histories, we make use of the Functional Clustering Model (FCM)⁹ which solves the problems of high variance due to sparsity as well as that of unequal variances due to irregular time instants. While a full description of the functional clustering approach and model fitting procedure is described by James and Sugar, the intuition behind the procedure is to project each curve onto a finite dimensional space using a natural cubic spline basis and cluster the resulting coefficients. However, instead of treating the basis coefficients as parameters James and Sugar model these are random effects that are ascertained by the clustering algorithm, via a global optimization over all curves. Doing so allows us to borrow strength across curves, allowing the method to work with sparsely or irregularly sampled curves, provided that the total number of observations is large enough. The model is fit to the data using an Expectation Maximization like procedure that iteratively updates the FCM parameters. As shown in Figure 3, each track is represented as a vector of measurements Y_i such that

$$Y_i = g_i + \epsilon_i \tag{1}$$

where g_i represents the true (unobserved) measurements at the same instants and ϵ_i is the vector of random observation errors $\epsilon_i \sim N(0, \sigma^2 I)$. The observation errors are assumed to be uncorrelated with each other and with g_i . We assume membership in one among G clusters. The true observations are modelled by using natural cubic splines as basis functions s(t) so that

$$g_i(t) = s(t)^T \eta_i \tag{2}$$

where η_i is the vector of spline coefficients. The FCM uses a random effects model for the η_i , assuming a normal distribution of values around a cluster mean as follows

$$\eta_i = \mu_{z_i} + \gamma_i \tag{3}$$

where μ_{z_i} represents the mean of the *i*th cluster of tracks and $\gamma \sim N(0, \Gamma)$. An additional parameterization step allows for a low dimensional representation of η_i given by

$$\eta_i = \lambda_0 + \Lambda \alpha_{z_i} + \gamma_i \tag{4a}$$

where
$$\mu_k = \lambda_0 + \Lambda \alpha_k$$
 (4b)

Thereafter the FCM can be represented as below

$$Y_i = Si^T (\lambda_0 + \Lambda \alpha_{z_i} + \gamma_i) + \epsilon_i$$
(5)

where *Si* is the matrix of basis expansions of all time points for track *i*, α_{z_i} is a *h* vector representing the mean of the *i*th cluster of trajectories in a low dimensional space. The low dimensional representation is achieved via $\mu_k = \lambda_0 + \Lambda \alpha_k$ where both μ_k , λ_0 are vectors in \mathbb{R}^p , Λ is a *p* x *h* matrix and $h \leq \min(p, G-1)$. To ensure a unique solution for μ_k , λ_0 and Λ , we require

$$\sum \alpha_i = 0 \tag{6a}$$

$$\Lambda^T S^T (\sigma^2 I + S \Gamma S^T)^{-1} S \Lambda = I \tag{6b}$$

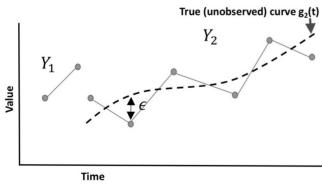


Figure 3 Modeling functional data

The form of (6b) ensures that $Cov(\alpha_i) = I$ for all *i*, when the observations for each track are measured at the same time points. The above formulation allows us to project every trajectory into *h* dimensional space to obtain the corresponding $\hat{\alpha}_i$, such that the proximity between $\hat{\alpha}_i$ and α_k represents the likelihood that the track *i* belongs to cluster *k*. Since $\hat{\alpha} = \Lambda^T S^T (\sigma^2 I + S\Gamma S^T)^{-1} (Y - S\lambda_0)$, the vector $\Lambda^T S^T (\sigma^2 I + S\Gamma S^T)^{-1}$ may be thought of as weights that determine the proximity to α_k , with the highest weight having the most influence on cluster assignment. This allows us to use $\Lambda^T S^T (\sigma^2 I + S\Gamma S^T)^{-1}$ as a discriminant function, for determining the time points that provide maximum cluster discrimination.

We provide here only an overview of the FCM and point the reader to a full description of the model, constraints and the fitting algorithm⁹. Since intuitively one expects to see a small number of distinct trajectory patterns, we clustered the creatinine measurements using small values of G. We obtained two nearly indistinguishable clusters with G=3, which suggests that G=2 may be an appropriate number of clusters for the creatinine trajectories in our cohort. After fitting the FCM with G=2 and making cluster assignments for every trajectory, we plotted the discriminant function for the two clusters in order to distinguish the time at which the two groups of trajectories most differ from each other.

3.2. Feature Engineering and Analysis

For each patient record in our cohort, we construct a one-year time window centered around the maximum discriminating time point as identified by the discriminant function described earlier. Within the time window, we represent the structured and unstructured data within a patient record as features from four categories – terms (or concepts), prescriptions, laboratory test results and diagnosis codes. Prescriptions, laboratory test results and diagnosis codes were taken from the structured record whereas terms were extracted from free text. We normalize terms into concepts in the same manner as in our earlier studies involving text mining on clinical notes—essentially using UMLS term-to-concept maps with suppression rules to weed out ambiguous mappings as described by Jung et al¹⁰. Such mapping reduces the total number of features as well as reduces the number of correlated features since synonyms get mapped to the same concept.

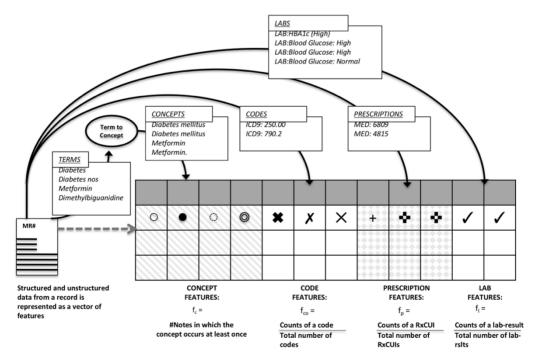


Figure 4 Features based on normalized counts of text concepts, ICD9 codes, prescriptions and lab results from the 1 year window around the maximum discriminating time point in the patient record

For concepts, we used the number of distinct notes in which the concept occurs at least once (note frequency) as the feature representation. For prescriptions and diagnostic codes, we used the normalized counts of the active ingredient for each medication (RxNorm concept unique identifier) and the normalized counts of each International Classification of Diseases, revision 9 (ICD9) code as the respective features. For laboratory test results, we utilized the categorical result status for each ordered test (high/ normal/low or normal/abnormal) as recorded in the Electronic Health Record (EHR) and calculated a feature based on the normalized counts for each test-result instance in the record. Our feature construction method is illustrated in Figure 4 which depicts our feature matrix along with the four categories (concepts, diagnosis codes, prescriptions and laboratory results) of data elements within the patient record from which the features are sourced. Finally, we performed an enrichment analysis on the feature matrix for the two clusters using Fisher's exact test, setting a false discovery rate of 5% to adjust for multiple testing. All analysis was performed in R using the Aphrodite¹¹ and the fclust¹² APIs.

4. Results

Figure 5 shows a randomly selected subset of 50 trajectories from each of the two clusters, indicating the overall progression pattern in each cluster. The thick lines corresponding to each cluster mean show progression patterns that the respective cluster represents. In case of cluster 1, the mean trajectory indicates that creatinine levels begin to rise around the age of 65 years, peak at around 72 years and then decrease. The trajectories in cluster 2 suggest an overall better control on creatinine levels.

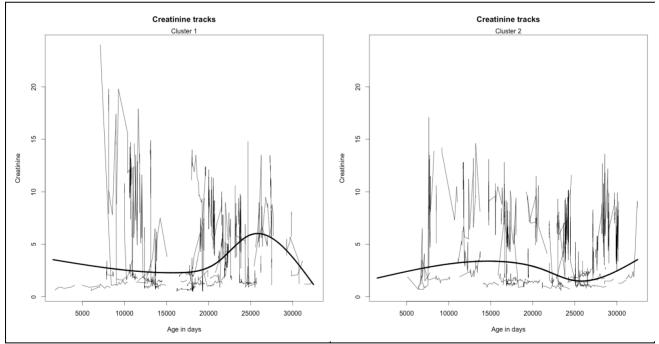


Figure 5 50 randomly drawn trajectories from each of the two clusters. The time offsets represent the age (in days) at which the measurement was taken. The heavy lines are the respective mean trajectories

The maximum value of the discriminant function occurs at a time offset $t_{dmax} = 26,383$ days (72 years), corresponding to the peak in the mean Creatinine trajectory in cluster 1. Drawing concepts, ICD9 codes, prescriptions and lab results from the EHR records of patients in each of

the two clusters for the 1-year time window centered on t_{dmax} yields 1046 features. The rate of progression to end stage renal disease depends on a number of risk factors, such as the presence of CKD-associated illnesses, nutrition issues and adherence to medication¹³ and we expect features from 1 year window to represent events associated with clinically significant decline in renal function in advanced-stage patients¹⁴.

Rank	Feature ID (p)	Names	Rank	Feature ID (p)	Names
1	lab.3023314.0	Hematocrit	16	obs.4274025	Disease
	(3.27e-05)	[Volume Fraction] of Blood by		(7.86e-04)	
		Automated count:			
2	lab.3000963.0	Hemoglobin:	17	obs.4322976	Procedure
	(4.06e-05)			(8.01e-04)	
3	lab.3000905.0	Leukocytes	18	obs.4229881	Weight loss
	(5.92e-05)	[#/volume] in Blood by		(8.46e-04)	
		Automated count:			
4	obs.4187395	Reflux	19	obs.46233416	Assessment
	(7.19e-05)			(9.53e-04)	
5	obs.4187458	Review of systems	20	obs.4143467	Chief complaint
	(1.46e-04)			(1.09e-03)	
6	obs.4243768	Auscultation	21	lab.3009261.0	Glucose [Presence]
	(1.78e-04)			(1.22e-03)	in Urine by Test
_	1 4110((2	D 1 / 1	22	1 4200224	strip:
7	obs.4118663	Related	22	obs.4209224	Cyst
8	(2.06e-04) obs.31967	Nausea	23	(1.51e-03) obs.441408	Vomitina
0	(3.23e-04)	Nausea	23	(1.67e-03)	Vomiting
9	lab.3013682.0	Urea nitrogen	24	obs.4147571	Follow-up
	(3.25e-04)	serum/plasma:	24	(1.82e-03)	ronow-up
10	obs.442985	Male	25	obs.4267147	Platelet count
10	(3.85e-04)	1,1,1,1,1	20	(1.85e-03)	
11	lab.3014051.0	Protein [Presence]	26	lab.3022621.0	pH of Urine by
	(4.03e-04)	in Urine by Test strip:		(1.85e-03)	Test strip:
12	obs.254761	Cough	27	obs.77670	Chest pain
12	(4.21e-04)	Cough	27	(1.86e-03)	enest pain
13	obs.4077953	Therapy	28	lab.3035350.0	Ketones urine
	(4.88e-04)	- ·· r J		(1.86e-03)	dipstick:
14	obs.4099313	Urinalysis	29	lab.3004501.45876384	Glucose lab:High
	(5.19e-04)	-		(1.96e-03)	Č –
15	obs.4329041	Pain	30	obs.4303558	Touch
	(7.37e-04)			(1.98e-03)	

Table 1 Top ranked significant features

A Fisher's exact test of enrichment for each feature with respect to the two clusters, using a false discovery rate of 5% to adjust for multiple testing, identified 133 enriched features, of which 30 top ranked features along with their respective p-values are presented in Table 1.

All of the top 30 features are found to be enriched in cluster 1 and an examination of the features suggests concordance with the known attributes of disease severity. For example the top 2 features (lab orders for hematocrit and hemoglobin) suggest the presence of Anemia and Thrombocytopenia respectively, which are two common comorbidities in advanced CKD that are thought to occur as a result of reduced erythropoietin secretion¹⁵. An observation related to platelet counts (feature rank 25) corroborates this view. Similarly, automated blood count is routinely measured in CKD patients, particularly in those patients requiring management of Anemia. Recently studies show that spikes in granulocyte and monocyte count in CKD patients are associated with progression to end stage renal disease¹⁶. A finding of Proteinuria on dipstick urinalysis is amongst the early signs of kidney disease, however high levels of protein in the urine is an indicator of nephritic syndrome and is associated with edema, increased cholesterol levels and other comorbidities that increase the risk of CKD progression¹⁷. Lab orders for pH and ketone measurement in urine (feature rank 26 and 28 respectively) suggest diabetic ketoacidosis which is an uncommon but life threatening complication in chronic kidney disease. Poor glucose regulation (feature rank 29) further supports the view that several of the distinguishing attributes have etiological linkages with diabetic complications.

5. Discussion

Longitudinal patient data from a large sample of patients offers an opportunity to characterize the variability in how phenotypes progress over time. However, discovering patterns of progression for chronic diseases is challenging because of the irregularity and sparsity in the observations. Trajectory estimates derived from irregularly sampled and sparse observations have high variance which leads to unstable cluster definitions. The irregular sampling also poses an additional challenge – the variance of the estimated curve coefficients are different for each trajectory. The FCM described in the methods section addresses the problem by treating the curve coefficients as random effects and by projecting each curve into a subspace, such that the covariance normalized distance from the cluster center in this subspace, represents the probability of cluster membership. Using the FCM to cluster creatinine trajectories of CKD patients results in two clusters with distinct mean trajectories. Features based on counts of clinical attributes, from a windowed segment of the patients' EHR records at the point of maximal cluster separation, show a significantly different distribution in the two clusters; and many are supported by medical literature on CKD. However, several of our significant features refer to general CKD attributes and do not appear to have an obvious connection with disease severity or progression. Further, reducing the false discovery rate to 1% did not give any statistically significant features.

We acknowledge limitations of our approach. We made use of the ICD9 code for CKD for defining our cohort. Such a method could have a positive predictive value (PPV) as low as 53%¹⁸ if relying solely on the ICD9 code. In which case, the the selected patients may not have the clinical indicators for the disease, while creatinine measurements could still be available since creatinine may be ordered routinely as part of a basic metabolic panel. We mitigate this issue by requiring at least three creatinine measurements for members of the analysis cohort. Further still, creatinine values are known to be altered through processes that are independent of renal function and standard practice requires that the estimated Glomerular Filtration Rate be used for assessing

kidney health¹⁹. The FCM also implicitly assumes that the unobserved time points are missing at random. Given that our data comes from a referral facility, it is likely that there exists a "disease severity bias" in the missing observations.

The alternative to using ICD9 codes for cohort identification is to use a robust algorithm that has been validated to achieve a high PPV^{20} . Patient records extracted from claims data may possibly provide better longitudinal coverage compared to EHR data from a tertiary care facility. Addressing our study's limitations through the aforementioned remedial measures appears feasible and we anticipate doing so in follow up work.

6. Conclusion

Being able to account for individual variability in the progression of diseases is of value to the practitioner, the patient as well as the researcher. For chronic diseases, learning the clinical attributes of the disease progression paths is possible by using a method for clustering irregularly sampled, sparse trajectories of disease markers, by defining a time window in which the clusters are most discriminable, and identifying discriminating features based on that time window. Our results from clustering creatinine trajectories of CKD patients demonstrate the feasibility of the approach.

7. Acknowledgements

We would like to thank our colleagues Sarah Poole and Jassi Pannu for their contributions to this study. This work was funded by NLM R01 LM011369.

References

- 1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13(6):395-405. doi:10.1038/nrg3208.
- 2. Yang J, McAuley J, Leskovec J, LePendu P, Shah N. Finding progression stages in timeevolving event sequences. *Proc 23rd Int Conf World wide web*. 2014:783-794. doi:10.1145/2566486.2568044.
- 3. Lunney JR, Lynn J, Hogan C. Profiles of Older Medicare Decedents. *J Am Geriatr Soc.* 2002;50(6):1108-1112. doi:10.1046/j.1532-5415.2002.50268.x.
- 4. Komarova NL, Thalhauser CJ. High degree of heterogeneity in Alzheimer's disease progression patterns. *PLoS Comput Biol*. 2011;7(11):e1002251. doi:10.1371/journal.pcbi.1002251.
- 5. Wang X, Wang F. Unsupervised Learning of Disease Progression Models. 2014. doi:10.1145/2623330.2623754.
- 6. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS One.* 2013;8(6):e66341. doi:10.1371/journal.pone.0066341.
- 7. Pimentel M, Clifton D, Clifton L, Tarassenko L. Modelling Patient Time-Series Data from Electronic Health Records using Gaussian Processes. *Adv neural Inf Process Syst Work Mach Learn Clin Data Anal.* 2013:1-4.
- 8. Pimentel, MAF, Clifton DA TL. Gaussian process clustering for the functional characterisation of vital-sign trajectories. In: *Machine Learning for Signal Processing*

(*MLSP*), 2013 IEEE International Workshop On.; 2013:1-6. doi:10.1109/MLSP.2013.6661947.

- 9. James GM, Sugar C a. Clustering for Sparsely Sampled Functional Data. *J Am Stat Assoc.* 2003;98(462):397-408. doi:10.1198/016214503000189.
- Jung K, LePendu P, Iyer S, Bauer-Mehren A, Percha B, Shah NH. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. *J Am Med Inform Assoc*. 2015;22(1):121-131. doi:10.1136/amiajnl-2014-002902.
- 11. OHDSI. Aphrodite. https://github.com/OHDSI/Aphrodite. Accessed October 3, 2016.
- 12. Gareth J. Fclust. http://www-bcf.usc.edu/~gareth/research/fclustdoc.pdf. Accessed October 3, 2016.
- 13. Thomas R, Kanso A, Sedor JR. Chronic kidney disease and its complications. *Prim Care*. 2008;35(2):329-344, vii. doi:10.1016/j.pop.2008.01.008.
- 14. Zhang A-H, Tam P, LeBlanc D, et al. Natural history of CKD stage 4 and 5 patients following referral to renal management clinic. *Int Urol Nephrol*. 2009;41(4):977-982. doi:10.1007/s11255-009-9604-3.
- 15. Akimoto T, Ito C, Kotoda A, et al. Challenges of caring for an advanced chronic kidney disease patient with severe thrombocytopenia. *Clin Med Insights Case Rep.* 2013;6:171-175. doi:10.4137/CCRep.S13238.
- Agarwal R, Light RP. Patterns and prognostic value of total and differential leukocyte count in chronic kidney disease. *Clin J Am Soc Nephrol*. 2011;6(6):1393-1399. doi:10.2215/CJN.10521110.
- 17. Mehdi U, Toto RD. Anemia, diabetes, and chronic kidney disease. *Diabetes Care*. 2009;32(7):1320-1326. doi:10.2337/dc08-0779.
- Cipparone CW, Withiam-Leitch M, Kimminau KS, Fox CH, Singh R, Kahn L. Inaccuracy of ICD-9 Codes for Chronic Kidney Disease: A Study from Two Practice-based Research Networks (PBRNs). *J Am Board Fam Med.* 28(5):678-682. doi:10.3122/jabfm.2015.05.140136.
- 19. Samra M, Abcar AC. False estimates of elevated creatinine. *Perm J.* 2012;16(2):51-52.
- 20. Nadkarni GN, Gottesman O, Linneman JG, et al. Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu Symp Proc.* 2014;2014:907-916.