MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-935

Surveying a New Multi-Institution Clinical Data Research Network

Marc Rosenman[a,b], Margaret Madden[a], Elissa Oh[a], Satyender Goel[a,c], Abel Kho[a,c]

[a] *Center for Health Information Partnerships, Northwestern University, Chicago, Illinois, USA*
[b] *Department of Pediatrics, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA*
[c] *Department of Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA*

Abstract

Cultivated by the Patient-Centered Outcomes Research Network (PCORnet), thirteen regional clinical data research networks (CDRNs) are taking shape across the U.S. The PCORnet common data model was carefully planned, and the data marts assembled by the more than 80 data-contributing institutions (nodes) are undergoing, in 2016-2017, a series of data characterization cycles. PCORnet will adjudge each node's—and thereby, in a significant way, each CDRN's—readiness or unreadiness for multi-institution research. Certifying each node's quality and fidelity is of course essential. But in understanding network readiness there is an additional, vital dimension—one that has received too little attention. It is the development of knowledge about the nature of a CDRN's data, in its federated sense. With visualizations, how might one grasp the meta-data of a CDRN? We outline an approach that builds upon the HealthLNK Data Repository, a forerunner to the Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN) CDRN.

Keywords:

Patient Outcome Assessment; Metadata; Electronic Health Records.

Introduction

Cultivated by the Patient-Centered Outcomes Research Network (PCORnet), thirteen regional clinical data research networks (CDRNs) are taking shape across the U.S [1-3].

PCORnet is part of the Patient Centered Outcomes Research Institute, which was authorized by the Patient Protection and Affordable Care Act of 2010. The PCORnet common data model was carefully planned, and the data marts assembled by the more than 80 data-contributing institutions (nodes) are undergoing, in 2016-2017, a series of data characterization cycles. PCORnet will adjudge each node's—and thereby, in a significant way, each CDRN's—readiness or unreadiness for multi-institution research. Certifying each node's quality and fidelity is of course essential. But in understanding network readiness there is an additional, vital dimension—one that has received too little attention. It is the development of knowledge about the nature of a CDRN's data, in its federated sense. With visualizations, how might one grasp the meta-data of a CDRN? We outline an approach that builds upon the HealthLNK Data Repository [4-6], a forerunner to the Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN) CDRN [7,8].

The HealthLNK Data Repository is a de-identified assembly of electronic health records (EHR) of adults 18-89 years old, from seven Chicago health care institutions: five large academic medical centers, one large county health care system, and a network of community health centers. HealthLNK also created a software application to merge and de-duplicate the patient identifiers across the institutions [4]. Data from the seven institutions are thereby woven together at the individual level via de-identified hashing. While the repository is maintained in a centralized system, housed in an enterprise data warehouse behind a secure firewall at Northwestern University, HealthLNK is a shared resource, to provide insight on the health of the Chicago community, and to identify opportunities to improve care. As a new repository it has proved efficient in providing data for studies of patterns of care among patients with these diverse conditions: diabetic ketoacidosis [6], diabetic retinopathy [9], gastrointestinal endoscopic pro- cedures [10], systemic lupus erythematosus [12], and non- emergent conditions in the emergency department [12].

HealthLNK was designed by many of the same Chicago institutions that subsequently developed CAPriCORN, which is one of the thirteen PCORnet CDRNs. CAPriCORN includes the seven HealthLNK institutions plus four more. The hashing-and-matching software that HealthLNK developed for merging, de-duplicating, and de-identifying patient identifiers across the institutions has been adopted by CAPriCORN and some other CDRNs in PCORnet. For all of these reasons HealthLNK provides an opportunity to describe a CDRN at a federated level. How does one go about exploring a federated repository's strengths and weaknesses, and building its meta-data? The purpose of this activity is not to test a hypothesis but rather is to build pre-research knowledge. It may help investigators shape hypotheses and methods. It also may guide decisions about future sponsored opportunities to pursue.

Here we provide an overview of HealthLNK (as reflective of a nascent CDRN) from various angles:

- Contributions and interdigitations by data type:
  - Demographics
  - Diagnoses
  - Procedures
  - Laboratory results
  - Vital signs
  - Medications
- Overlap between the institutions' populations
- Emergency, inpatient, and outpatient encounters
- Subpopulations based on particular health conditions, or number of institutions visited
- Mortality
- Geographic scope

Methods

Retrospective data from 2006 to 2012 were available in the HealthLNK data repository. We used SQL queries to intersect the six principal data tables: demographics, diagnoses, procedures, laboratory results, vital signs, and medications.

For this Venn diagram we report the number of patients in the six-way and five-way intersections, and the single data type zones. We also determined in which institution(s) each patient had at least one diagnosis record. These diagnosis data are the substrate for a six-set Venn diagram, and some smaller diagrams focused on inpatient or emergency department encounters in a set of four institutions. In the analyses of diagnosis records we excluded one institution because it had data from only two years, and because a six-way Venn diagram is easier than a seven-way version to interpret visually. The template for our six-way Venn diagram is the one authored by Jeremy Carroll, PhD (then of Hewlett-Packard) [13].

Because CDRNs may have particular value in examining those patients who sought health care in more than one institution, we described additional dimensions of some scenarios for multi (or single) institution use, per patient. We examined institutional cross-over for patients with cancer, and for those who were the victim of a stabbing or gunshot wound. We also examined, among patients who had one or more diagnosis records in two institutions, how many diagnosis records the patient had from each institution. We also calculated the mortality rate by number of institutions visited.

The geographic analysis of Cook County (Chicago and nearby suburbs) and DuPage County (additional western suburbs), Illinois reflects the density of the HealthLNK population: number of unique patients in HealthLNK in each zip code divided by the total zip code population (U.S. Census). The map was drawn using ArcGis software (Esri, Redlands, CA). We included all seven institutions in the geographic analysis.

The geographic analysis of Cook County (Chicago and nearby suburbs) and DuPage County (additional western suburbs), Illinois reflects the density of the HealthLNK population: number of unique patients in HealthLNK in each zip code divided by the total zip code population (U.S. Census). The map was drawn using ArcGis software (Esri, Redlands, CA). We included all seven institutions in the geographic analysis. Encounter dates in HealthLNK are provided by the data-contributing institutions as MM/YYYY. HealthLNK's rules forbid comparison of the institutions by name.

Results

There are 3,697,707 unique patients in the 2006-2012 instance of the HealthLNK database (seven institutions). The numbers of unique patients by data type are shown in Table 1.

Table 1 – Patient Counts by HealthLNK Table

Data Table	Unique Patients with Data Type	% Database Population Total
Demographics	3,085,215	83%
Diagnosis	2,602,509	70%
Procedures	1,882,573	51%
Laboratory results	1,587,832	43%
Vital signs	1,955,212	53%
Medications	1,444,084	39%

Based on records in the Diagnosis Table (using the six institutions with more than two years of data), the distribution of number of distinct calendar months (with an encounter) per patient is shown in Figure 1. Across the 7-year retrospective period, about half of the patients had encounters in no more than two distinct calendar months. Patients by combination of data types are depicted in Figure 2. About one-sixth of the patients have all six data types, and slightly more than one-third have five of six data types. About one-fourth have only one data type (mostly demographics or vital signs); these single-type scenarios suggested

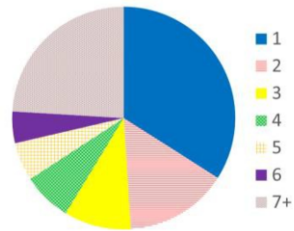


Figure 1 – Number of unique calendar months in which each patient had encounters (based on the Diagnosis Table)

that some data were omitted by one or two institutions in the extract/transfer/load procedures; their code will be revised in the next data refresh. Figure 3 shows the number of institutions visited per patient.

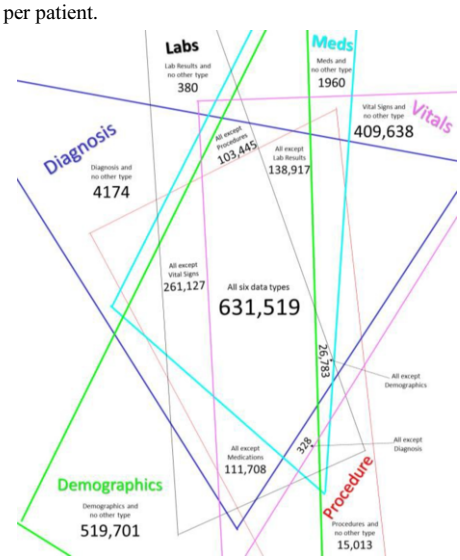


Figure 2 – Venn diagram (with truncated edges), six data types: the one, five, and six-way intersections are shown

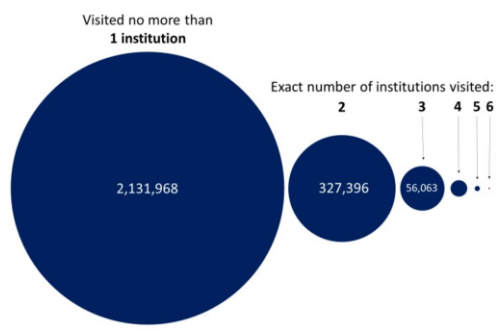


Figure 3 – Number of institutions visited per patient (based on the Diagnosis Table). Circles scaled by number of patients

Characteristics of patients by number of institutions visited are shown in Table 2. In the Diagnosis Table, we studied International Classification of Diagnosis, 9th Revision (ICD-9) codes, by number of institutions per patient. Among the small population with an encounter at all six institutions, more than 50% had an ICD-9 code for depression, and more than 40% for lack of housing. The frequencies of ICD-9 codes for drug abuse, chronic alcoholism, and suicidal ideation were also above 40%.

Table 2 – Patient Characteristics by N of Institutions Visited

Characteristic	Number of Institutions Visited				
	1	2	3	4	5
%Women	54	61	65	65	57
Race					
% Black	26	46	57	71	77
% White	49	41	30	22	20
% Other/declined	25	13	13	7	3
Insurance Status					
% Medicare	18	20	24	28	32
% Medicaid	7	13	15	20	25
% Commercial	57	39	29	21	13
% Self-Pay	11	19	21	21	22
% No Charge	0	0	1	1	1
% Other	8	9	10	9	7
Morality (deaths/10,000)	1.7	5.4	4.8	7.7	12.6

** The denominator for this column is <100 patients.*

When we examined the encounter types by "E," "I," and "O" (emergency department, inpatient, and outpatient, respectively), we noticed that one institution, which we know has a busy emergency department, had sent no type "E" but a very large number of type "I" records. Because of this apparent data transfer error, we excluded that institution from the "EIO" analyses. We also excluded the network of community health centers (all outpatient) and the institution with two years of data. Among the other four institutions, patient characteristics by number of institutions visited is show in Table 3.

Table 3 – Patient Characteristics by N of Institutions Visited (Inpatient and Emergency Department)

Characteristic	Number of Institutions Visited				
Inpatient Institutions	**1**	**2**	**3**	**4**	**5**
% Women	49	62	59	46	26
Race					
% Black	50	26	40	51	53
% White	17	61	55	45	48
% Other/declined	33	13	5	4	0
Age (Median)	45	51	56	53	44
ER Institutions	**1**	**2**	**3**	**4**	**5**
% Women	61	57	64	59	41
Race					
% Black	50	43	70	76	67
% White	18	45	25	21	32
% Other/declined	32	12	5	3	1
Age (Median)	40	40	38	40	42.5

Characteristic	Number of Institutions				
Inpatient Institutions	**0**	**1**	**2**	**3**	**4**
% Women	49	62	59	46	26
Race					
% Black	50	26	40	51	53
% White	17	61	55	45	48
% Other/declined	33	13	5	4	0
Age (Median)	45	51	56	53	44
ER Institutions	**0**	**1**	**2**	**3**	**4**
% Women	61	57	64	59	41
Race					
% Black	50	43	70	76	67
% White	18	45	25	21	32
% Other/declined	32	12	5	3	1
Age (Median)	40	40	38	40	42.5

We also examined use of multiple institutions, within three subpopulations (again using the six institutions analyzed for Figure 3). A cohort with any malignancy diagnosis (which may include some patients who were evaluated and were found not to have cancer) was defined based on the presence of any ICD-9 diagnosis code in the range 140 to 209.99. A cohort with melanoma was defined based on at least one ICD-9 diagnosis code containing 172. A cohort with melanoma was defined based on at least one ICD-9 diagnosis code containing 172. A cohort with knife or gun- shot injury was defined by this set of ICD-9 E-codes: E965, E965.0, E965.1, E965.3, E965.4, E965.5, E965.8, E965.9, E966, E970, or E974. This analysis (Figure 4) considered vis- its of any type, for any diagnosis, among these three cohorts.

Those with any malignancy diagnosis are counted on the primary y-axis; those with melanoma or knife or gunshot injury on the secondary y-axis. Use of more than one institution was least among those with melanoma. Use of more than one institution was more common among those with knife or gun- shot injury.

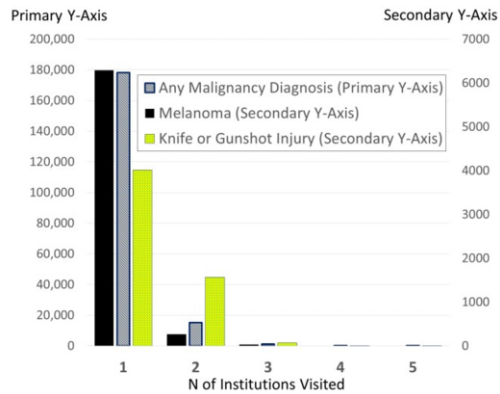
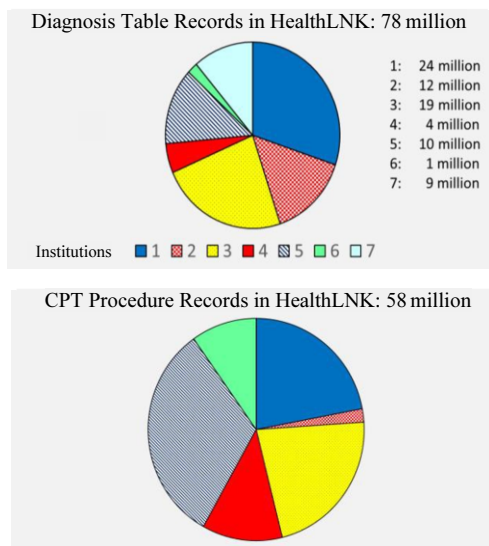


Figure 4 – N of institutions visited, among three cohorts

Apart from the issue noted above wherein one institution sent its type "E" encounters with a type "I" label, the Diagnosis Table and Procedure Table records sent to HealthLNK by the participating institutions were strong (Figures 5a and 5b).



Figures 5 – Distribution of Diagnosis (top) and Procedure (bottom) Table records, respectively, by institution

By contrast, not all of the institutions sent a full set of laboratory results. It can be helpful when visualizing a new CDRN to compare the distribution of institutions across different laboratory results. Figure 6 provides examples.

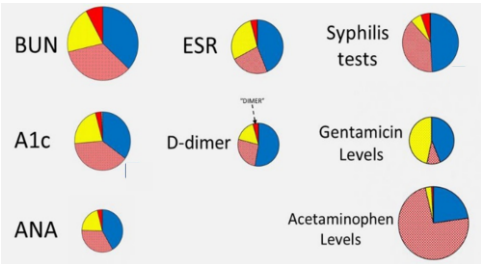


Figure 6 – Eight laboratory results by institution (BUN: Blood Urea Nitrogen; ESR: Erythrocyte Sedimentation Rate; A1c: Hemoglobin A1c; ANA: Anti-nuclear antibody)

In this way we discerned more than the fact that one institution provided only diabetes-related results. The distinctive institutional distributions of gentamicin levels and of acetaminophen levels reflects, it turns out, differences in actual care patterns (one of the hospitals in HealthLNK truly predominates in the number of patients evaluated with suspicion of acetaminophen toxicity) as well as technical (informatics) idiosyncrasies on the part of the data-supplying institutions.

The geographic distribution of the population (including all 7 institutions) in HealthLNK is illustrated in Figure 7. When we examined the patients with visits to two institutions, we found that in many cases, there existed only one (or a few) diagnosis record(s) in one of the two institution that the patient visited (Figure 8).

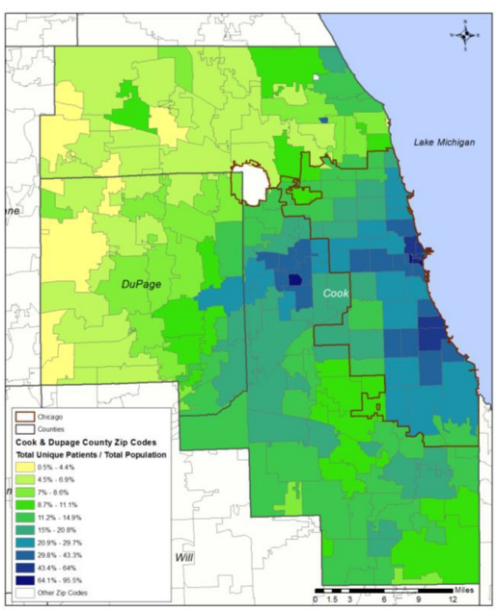


Figure 7 – Zip codes in darker blue have a larger population in HealthLNK (as a proportion of U.S. Census population)

More than 40% of these patients had only one or two diagnosis records in the second institution. About 33% of these patients had more than five diagnosis records in the second institution.

Discussion

In this survey of a forerunner to a PCORnet CDRN, we illustrated several dimensions in which it may be helpful to build understanding of the meta-data. In analyses after review of Table 1, we found that not all institutions sent laboratory data. One institution sent only diabetes-related laboratory results, in preparation for a specific research project. We then further learned about nuances by laboratory test. We also conducted indepth visualizations of the quality of various data elements in several domains (data not shown), such as variation between (and within) institutions in microbiology culture laboratory results.

Not unexpectedly, Figure 1 shows that most patients in a multi-institution, outpatient and inpatient EHR database have relatively few health care encounters over the years. Venn diagrams like that in Figure 2 are useful in planning potential studies of quality-of-care metrics. The zone with all (five) other data types, except for medication data, will be important to consider when calculating metrics in which the numerator is based on the use of a particular class of medication. Not unexpectedly, Figure 3 shows that the vast majority of patients visited no more than one institution over the years. Under 3% visited more than two institutions.

Table 2 shows that those who visited a higher number of institutions visited were more likely to men, African-American, and patients with publicly-financed or self-pay health care coverage. Mortality was strongly associated with number of institutions visited between 1 and 5. There were no deaths in the small group of patients who visited 6 institutions. The latter group had a very high rate of psychiatric diagnoses, sub- stance abuse, and homelessness. Table 3 shows that among patients with at least one hospitalization during the study period, those who were hospitalized in multiple institutions over time were more likely to be men and to be African-American. The direction of these associations was the same for the number of emergency department institutions visited over the years.

Figure 4 suggests that the degree of "crossover" between institutions will have to be examined project by project, as patterns may well vary among different study cohorts. The contrasts between Figures 5a/5b and 6 suggest that additional exploration of (federated) meta-data for laboratory tests is needed.

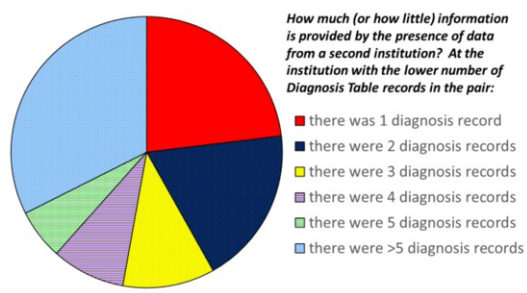


Figure 8 – Diagnosis records per patient in the one of the two institutions, among patients who visited two institutions

The geographic distribution in Figure 7 is consistent, in general, with the locations of the HealthLNK institutions. Within the city limits, HealthLNK's population penetrance is lowest on the northwest and southwest sides of Chicago. Those neighborhoods are served, to a substantial degree, by hospitals that are not part of HealthLNK. In this context, it is worth considering what percentage HealthLNK is, of Chicago's overall health care plant. Based on publically available data, we calculate that the HealthLNK hospitals account for approximately 40% of the acute care adult inpatient beds in the city, and approximately 30% of the labor and delivery beds. Figure 8 suggests additional health services research

work we may pursue in examining patients' use of more than one institution in a CDRN.

CDRNs and other multi-institution networks vary in geographic scope, types of constituent institutions, degree of centralization of data (and, hence, meta-data), etc. Those in a city (or any locale) might ask what proportion of total hospital beds in the area are within the network. The techniques we describe in examining overlaps of data types and populations may assist other networks in understanding care patterns and the network's relationship with its milieu. One might ask questions like: For newborns in a CDRN, to what extent does the CDRN have data on well-child visits in the ensuing years? What are the implications for child health research projects?

Networks can also build meta-data knowledge about particular data elements. In Table 2 above, we use the six categories of health insurance stored in HealthLNK. But what is the consistency and fidelity with which the institutions mapped their granular health insurance values to the six categories? We have begun comparing these mapping details and will build our meta-data knowledge to the point where we plan to offer new guidance to all networks in categorizing insurance.

The principal limitation of this project is that it is pre-research visualization rather than hypothesis driven work. Nevertheless, we think that this type of work is necessary—and needs more attention paid to it—in order to support the more specific research projects that will follow.

Conclusion

The HealthLNK project has substantially informed the work now underway in the CAPriCORN CDRN. In turn, the work in both CAPriCORN and PCORnet will help standardize and will enhance future versions of the HealthLNK repository.

Acknowledgements

This study was not sponsored.

References

[1] D.A. Corley, H.S. Feigelson, T.A. Lieu, E.A. McGlynn, Building data infrastructure to evaluate and improve quality: PCORnet, *J Oncol Pract* **11** (2015):204-206.

[2] L.H. Curtis, J. Brown, R. Platt, Four health data networks illustrate the potential for a shared national multipurpose big-data network, *Health Aff (Millwood)* **33** (2014): 1178- 1186.

[3] C. Parke, J. Cook, T. Carton, S. Rao, The Louisiana clinical data research network: leveraging regional and national re- sources to improve clinical research efficiency, *Ochsner J* **14** (2014):718-723.

[4] A.N. Kho, J.P. Cashy, K.L. Jackson, A.R. Pah, S. Goel, J. Boehnke, J.E. Humphries, S.D. Kominers, B.N. Hota, SA.Sims, B.A. Malin, D.D. French, T.L. Walunas, D.O. Meltzer, E.O. Kaleba, R.C. Jones, W.L. Galanter, Implementation linkage tool in Chicago. *J Am Med Inform Assoc* **22** (2015):1072-1080.

[5] W.L. Galanter, A. Applebaum, V. Boddipalli, A. Kho, M. Lin, D. Meltzer, A. Roberts, B. Trick, S.M. Walton, B.L. Lambert, Migration of patients between five urban teaching hospitals in Chicago, *J Med Syst* **37** (2013), 1-8.

[6] J.A. Mays, K.L. Jackson, T.A. Derby, J.J. Behrens, S. Goel, M.E. Molitch, A.N. Kho, A. Wallia, An evaluation of recurrent diabetic ketoacidosis, fragmentation of care, and mortality across Chicago, Illinois, *Diabetes Care* **39** (2016):1671-1676.

[7] A. Solomonides, S. Goel, D. Hynes, J.C. Silverstein, B. Hota, W. Trick, F. Angulo, R. Price, E. Sadhu, S. Zelisko, J. Fischer, B. Furner, A. Hamilton, J. Phua, W. Brown, S.F. Hohmann, D. Meltzer, E. Tarlov, F.M. Weaver, H. Zhang, T. Concannon, A, Kho, Patient-centered outcomes research in practice: the CAPriCORN infrastructure, *Stud Health Technol Inform* (2015):584-588.

[8] A.N. Kho, D.M. Hynes, S. Goel, A.E. Solomonides, R. Price, B. Hota, S.A. Sims, N. Bahroos, F. Angulo, W.E. Trick, E. Ta lov, F.D. Rachman, A. Hamilton, E.O. Kaleba, S. Badlani, S.L. Volchenboum, J.C. Silverstein, J.N. Tobin, M.A. Schwartz, D. Levine, J.B. Wong, R.H. Kennedy, J.A. Krishnan, D.O. Meltzer, J.M. Collins, T. Mazany, CAPriCORN Team, CAPriCORN: Chicago area patient-centered outcomes re- search network, *J Am Med Inform Assoc* **21** (2014):607-611.

[9] D.D. French, J.J. Behrens, K.L. Jackson, A.N. Kho, T.L. Walunas, C.T. Evans, M. Mbagwu, C.E. Margo, P.J. Bryar, Payment reform needed to address health disparities of undiagnosed , *Ophthalmol Ther* (2016):123-131.

[10] K.L. Jackson, S. Goel, A.N. Kho, R.N. Keswani, Distance from hospital impacts adverse event detection after outpatient endoscopy, *Gastrointest Endosc* **85** (2016):380-386.

[11] T.L. Walunas, K.L. Jackson, A.H. Chung, K.A. Mancera-Cuevas, D.L. Erickson, R. Ramsey-Goldman, A. Kho, Disease outcomes and care fragmentation among patients with systemic lupus erythematosus. *Arthritis Care Res (Hoboken)* (2016) [Epub ahead of print]

[12] J. Fishman, S. McLafferty, W. Galanter, Does spatial access to primary care affect emergency department utilization for nonemergent conditions? *Health Serv Res* (2016). [Epub ahead of print]

[13] Diagrams Made from Triangles. [The Electronic Journal of Combinatorics website, edited June 2005] Accessed 12/23/2016. Available from: http://www.combinator- ics.org/files/Surveys/ds5/VennTriangleEJC.html.

Address for correspondence

Marc Rosenman, MD

625 N. Michigan Avenue, 15th Floor Chicago, IL, 60611, USA

Email: marc.rosenman@northwestern.edu
Phone: 312-503-3738