Identifying Patients Who are Likely to Receive Most of Their Care from a Specific Health Care System: Demonstration via Secondary Analysis

Gang Luo¹, PhD; Peter Tarczy-Hornoch^{1, 2, 3}, MD; Adam B Wilcox¹, PhD; E Sally Lee⁴, PhD

¹Department of Biomedical Informatics and Medical Education, University of Washington, UW Medicine South Lake Union, 850 Republican Street, Building C, Box 358047, Seattle, WA 98195, USA

²Department of Pediatrics, Division of Neonatology, University of Washington, 1959 NE Pacific St, Seattle, WA 98195, USA ³Department of Computer Science and Engineering, University of Washington, 185 Stevens Way, Seattle, WA 98195, USA ⁴Population Health Analytics, UW Medicine Finance, University of Washington, UW Tower, 4333 Brooklyn Ave NE, Box 359427, Seattle, WA 98195, USA

luogang@uw.edu, pth@uw.edu, abwilcox@uw.edu, sallylee@uw.edu

Corresponding author:

Gang Luo, PhD Department of Biomedical Informatics and Medical Education, University of Washington, UW Medicine South Lake Union, 850 Republican Street, Building C, Box 358047, Seattle, WA 98195, USA Phone: 1-206-221-4596 Fax: 1-206-221-2671 Email: luogang@uw.edu

Abstract

Background: In the United States, health care is fragmented in numerous distinct health care systems including private, public, and federal organizations like private physician groups and academic medical centers. Many patients have their complete medical data scattered across several healthcare systems, with no particular system having complete data on any of them. Several major data analysis tasks like predictive modeling using historical data are considered impractical on incomplete data. **Objective**: Our goal is to find a way to enable these analysis tasks for a healthcare system with incomplete data on many of its patients.

Methods: This study presents, to the best of our knowledge, the first method to use a geographic constraint to identify a reasonably large subset of patients who tend to receive most of their care from a given healthcare system. A data analysis task needing relatively complete data can be conducted on this subset of patients. We demonstrated our method using data from University of Washington Medicine (UWM) and PreManage data covering use of all hospitals in Washington state. We compared ten candidate constraints to optimize the solution.

Results: For UWM, the best constraint is that the patient has a UWM primary care physician and lives within five miles of at least one UWM hospital. About 16.01% (55,707/348,054) of UWM patients satisfied this constraint. Around 69.38% (10,501/15,135) of their inpatient stays and emergency department visits occurred within UWM in the following six months, more than double the corresponding percentage for all UWM patients.

Conclusions: Our method can identify a reasonably large subset of patients who tend to receive most of their care from UWM. This enables several major analysis tasks on incomplete medical data that were previously deemed infeasible.

Keywords: data analysis; inpatients; emergency departments; health care system

1. Introduction

In the United States, health care is fragmented in numerous distinct health care systems including private, public, and federal organizations like private physician groups and academic medical centers. Frequently, a given healthcare system has incomplete medical data on many of its patients, as these patients' complete data are recorded across multiple healthcare systems [1, 2]. Finnell *et al.* [2] showed that during a three-year period in Indiana, 40.7% of emergency department visits came from patients who also had emergency department visits at other healthcare systems. Bourgeois *et al.* [1] showed that during a five-year period in Massachusetts, 56.5% of adult hospital encounters (inpatient stays and emergency department visits) came from patients who also had encounters at other hospitals. Incomplete data are particularly problematic in academic healthcare systems like University of Washington Medicine (UWM), where many patients are referred from other healthcare systems. As shown in the Results section, <1/3 of the hospital encounters for all UWM patients occur within UWM. Currently, several major data analysis tasks like predictive modeling using historical data are deemed impractical on incomplete data. This limits the applications based on these analysis tasks. For example, predictive modeling is widely used for identifying future high-cost patients's cost assume complete historical data [8-10], and are not used by a healthcare system with incomplete data on its patients. As a result, many future high-cost patients are not identified and enrolled in care management, contributing to undesirable outcomes.

This study presents, to the best of our knowledge, the first method to use a geographic constraint to identify a reasonably large subset of patients who tend to receive most of their care from a specific healthcare system. This is to enable these data analysis tasks on incomplete medical data. Although the healthcare system has incomplete data on many of its patients, it has more complete data on this subset of patients. For a data analysis task requiring relatively complete medical data, we can conduct the task for this subset of patients, with the understanding that the analysis results apply to only this subset of patients rather than all patients of the healthcare system. This could be an improvement compared with the current practice of not conducting the task at all, in cases when conducting the task on all patients is impractical. Our previous work [11] sketched the method's main goal, but did not complete the method, do a computer coding implementation, or evaluate the method's performance. This study aims to fill these gaps and demonstrate our method using data at UWM.

2. Methods

Patient population

The patient cohort included all adult patients (age \geq 18) who had encounters at UWM facilities (hospitals and clinics) with information stored in UWM's enterprise data warehouse during the one-year period of April 1, 2016 - March 31, 2017. In this paper, an encounter can be of any type, unless it is explicitly specified as a hospital encounter or an outpatient visit. UWM is the largest academic healthcare system in Washington state and has both hospitals and clinics for adults.

Data set

We used administrative data in UWM's enterprise data warehouse during the two-year period of April 1, 2015 - March 31, 2017. The data set included encounter and primary care physician (PCP) information of our patient cohort. We also used PreManage data that UWM has on all of its patients during the six-month period of April 1, 2017 - September 30, 2017. PreManage is Collective Medical Technologies Inc.'s commercial product offering encounter and diagnosis data on hospital encounters (inpatient stays and emergency department visits) at many U.S. hospitals [12]. PreManage data cover all hospitals in Washington state. Starting April 1, 2017, UWM has been receiving relatively complete PreManage data on its patients. In this paper, we chose April 1, 2017 as the index date separating the prior and subsequent periods for the analysis task.

Our constraint-based patient identification method

Our goal is to use a constraint to identify a reasonably large subset of patients who tend to receive most of their care from UWM. We considered three UWM hospitals whose administrative and clinical data are stored in UWM's enterprise data warehouse: Harborview Medical Center, University of Washington Medical Center, and Northwest Hospital. All three hospitals are in Seattle, Washington. We considered the following candidate constraints that all include the component of living within r miles of at least one of the three UWM hospitals, with r being a parameter whose optimal value was to be determined in the study:

- (1) **Distance only**: The patient lives within *r* miles of at least one of the three UWM hospitals. Intuitively, with everything else being equal, the closer a patient lives to UWM hospitals, the larger portion of his/her care tends to be received from UWM. Also, the smaller the *r*, the smaller the number of UWM patients satisfying the constraint.
- (2) **PCP** (primary care physician): The patient has a UWM PCP, and lives within *r* miles of at least one of the three UWM hospitals. UWM PCPs tend to make referrals within UWM. Hence, intuitively, compared with others, patients with a UWM PCP may receive a larger portion of their care from UWM.
- (3) ≥2 encounters in the past year: The patient had ≥2 encounters at UWM facilities in the past year (April 1, 2016 March 31, 2017), and lives within r miles of at least one of the three UWM hospitals. Intuitively, patients with more previous encounters at UWM facilities may receive a larger portion of their care from UWM in the future.
- (4) ≥ 2 encounters in the past two years: The patient had ≥ 2 encounters at UWM facilities in the past two years (April 1, 2015 March 31, 2017), and lives within *r* miles of at least one of the three UWM hospitals.
- (5) \geq 1 hospital encounter in the past year: The patient had \geq 1 hospital encounter in one of the three UWM hospitals in the past year, and lives within *r* miles of at least one of the three UWM hospitals.
- (6) \geq 2 hospital encounters in the past year: The patient had \geq 2 hospital encounters across the three UWM hospitals in the past year, and lives within *r* miles of at least one of the three UWM hospitals.
- (7) \geq 2 hospital encounters in the past two years: The patient had \geq 2 hospital encounters across the three UWM hospitals in the past two years, and lives within *r* miles of at least one of the three UWM hospitals.
- (8) ≥ 1 outpatient visit in the past year: The patient had ≥ 1 outpatient visit to UWM in the past year, and lives within *r* miles of at least one of the three UWM hospitals.
- (9) ≥ 2 outpatient visits in the past year: The patient had ≥ 2 outpatient visits to UWM in the past year, and lives within r miles of at least one of the three UWM hospitals.
- (10) \geq **2** outpatient visits in the past two years: The patient had \geq 2 outpatient visits to UWM in the past two years, and lives within *r* miles of at least one of the three UWM hospitals.

In each candidate constraint, distance is no longer a factor when $r=+\infty$.

Data analysis

Using the distVincentyEllipsoid function in R's geosphere package version 1.5-5 [13], we computed the ellipsoid great circle distance between a patient's home and a UWM hospital based on the longitude and latitude coordinates of the patient's 5-digit home address zip code and the hospital's address. This distance serves as a rough proxy of the travel distance between the patient's home and hospital, is easy to compute, and is sufficient for our patient identification purpose, as shown in the Results section. For other researchers wanting to adopt our constraint-based patient identification method for their studies, using zip codes instead of exact patient home addresses can facilitate data acquisition because a limited data set is easier to obtain than an identified one.

We compared performance across the ten candidate constraints for identifying patients likely to receive most of their care from UWM. We used administrative data in UWM's enterprise data warehouse to check whether a patient satisfied a specific constraint. For each candidate constraint, we computed the percentage of UWM patients satisfying it. For all patients satisfying the constraint, we used PreManage data to compute the percentage of their hospital encounters that occurred within UWM in the following six months (April 1, 2017 - September 30, 2017). Since hospital encounters are usually much more expensive than other encounters, this percentage reflects the portion of these patients' care received from UWM. In computing this percentage, every patient satisfying the constraint was included, regardless of whether the patient had ≥ 1 hospital encounter in the following six months. When selecting the final constraint to be used, we struck a balance between the following two criteria:

- (1) **Criterion 1**: The percentage of UWM patients satisfying the constraint should be as large as possible. As multiple data analysis tasks will be conducted on these patients, this will maximize the usefulness of the applications based on these tasks.
- (2) Criterion 2: For the patients satisfying the constraint, the percentage of their hospital encounters that occurred within UWM should be as large as possible. This is to maximize the degree of completeness of the medical data that UWM has on these patients. For the data analysis tasks that will be conducted on these patients, this degree impacts the biases in the analysis results.

As mentioned in the Discussion section, the selected constraint has a special property, increasing our confidence that the patients identified by the constraint also tend to incur most of their outpatient visits within UWM.

To show how the constraint-based method works for individual UWM hospitals, for all patients satisfying the selected constraint and each of the three UWM hospitals, we used PreManage data to compute the percentage of these patients' hospital encounters that occurred at the UWM hospital in the following six months.

Ethics approval

The institutional review board of UWM reviewed and approved this study, and waived the need for informed consent for all patients.

3. Results

Table 1 shows the demographic characteristics of our patient cohort.

UWM's enterprise data warehouse during	April 1, 2016 - March 31, 2017.
Demographic characteristics	Number of patients ($n = 348,054$
Age	
18 to <30	64,311 (18.48%)
30 to <65	209,033 (60.06%)
≥65	74,710 (21.47%)
Gender	
Male	154,511 (44.39%)
Female	193,506 (55.60%)
Unknown or not reported	37 (0.01%)
Race	
American Indian or Alaska native	5,107 (1.47%)
Asian	33,595 (9.65%)
Black or African American	25,224 (7.25%)
Multiple races	2,772 (0.80%)
Native Hawaiian or other Pacific islander	2,500 (0.72%)
White	232,354 (66.76%)
Unknown or not reported	46,502 (13.36%)
Ethnicity	
Hispanic	21,436 (6.16%)
Non-Hispanic	270,043 (77.59%)
Unknown or not reported	56,575 (16.25%)
Insurance	
Private	167,640 (48.16%)
Public	154,008 (44.25%)
Self-paid or charity	26,406 (7.59%)

 Table 1. Demographic characteristics of adult patients who had encounters at UWM facilities with information stored in UWM's enterprise data warehouse during April 1, 2016 - March 31, 2017

Figures 1 and 2 show the percentage of UWM patients satisfying each of the ten candidate constraints. The percentage increases with *r*, initially quickly when *r* is small and then more slowly as *r* becomes larger. Recall that *r* is the maximum allowed distance in miles between the patient's home and the closest UWM hospital. As UWM mainly serves the Seattle metropolitan area, 88.92% (=309,483/348,054) of UWM patients live within 60 miles of at least one of the three UWM hospitals. About 44.76% (=138,530/309,483) of these patients live within 5 miles.



Figure 1. The percentage of University of Washington Medicine (UWM) patients satisfying each of the 10 candidate constraints. PCP: primary care physician.



Figure 2. The percentage of University of Washington Medicine (UWM) patients satisfying each of the 10 candidate constraints, when *r* is ≤ 10 . PCP: primary care physician.

For each of the ten candidate constraints and all of the patients satisfying it, Figures 3 and 4 show the percentage of their hospital encounters that occurred within UWM in the following six months. With a few exceptions when r is small, as r increases, the percentage decreases, initially quickly when r is small and then more slowly as r becomes larger. This is consistent with our intuition that with everything else being equal, patients living further from UWM hospitals are less likely to use them. Regardless of how small r is, this percentage never approaches 100%, partly because UWM patients could also use several



non-UWM hospitals that are within one mile of certain UWM hospitals. Since we want this percentage to be as large as possible, we should choose *r* to be ≤ 5 or ≤ 10 , depending on the constraint.

Figure 3. The percentage of hospital encounters that occurred within the University of Washington Medicine (UWM) in the following 6 months for each of the 10 candidate constraints and all patients satisfying it. PCP: primary care physician.



Figure 4. The percentage of hospital encounters that occurred within the University of Washington Medicine (UWM) in the following 6 months for each of the 10 candidate constraints and all patients satisfying it, when r is ≤ 10 . PCP: primary care physician.

In selecting the final constraint to be used, we struck a balance between Criteria 1 and 2 listed at the end of the Methods section. The PCP constraint significantly outperforms six other constraints on Criterion 1. When *r* is ≤ 6 , the PCP constraint outperforms all of the other constraints under Criterion 2. Also, when $r=+\infty$ and distance is no longer a factor, no constraint outperforms the PCP constraint with $r\leq 6$ under Criterion 2. Figure 5 shows the percentage of UWM patients satisfying the PCP constraint, as well as the percentage of these patients' hospital encounters that occurred within UWM in the following six months. When the PCP constraint was used with r=5, 16.01% (=55,707/348,054) of UWM patients satisfied the constraint. For these patients, 69.38% (=10,501/15,135) of their hospital encounters occurred within UWM in the following six months. In comparison, for all UWM patients, 31.80% (=39,171/123,162) of their hospital encounters occurred within UWM in the following six months.



Figure 5. The percentage of University of Washington Medicine (UWM) patients satisfying the PCP constraint and the percentage of these patients' hospital encounters that occurred within UWM in the following 6 months. PCP: primary care physician.

For each of the three UWM hospitals and all patients satisfying the PCP constraint, Figure 6 shows the percentage of their hospital encounters that occurred at the UWM hospital in the following six months. The percentage varies across the three UWM hospitals. As *r* increases, the percentage decreases at similar rates across the three UWM hospitals.



Figure 6. For each of the 3 University of Washington Medicine (UWM) hospitals and all patients satisfying the PCP constraint, the percentage of their hospital encounters that occurred at the UWM hospital in the following 6 months. PCP: primary care physician.

4. Discussion

Principal findings

By striking a balance between Criteria 1 and 2, we chose the PCP constraint with r=5 as the final one to be used. Using our constraint-based method to identify the right subset of patients, we more than doubled the percentage of patient hospital encounters that occurred within UWM in the following six months: from 31.80% (=39,171/123,162) to 69.38% (=10,501/15,135). Moreover, as each identified patient has a UWM PCP, we are confident that the identified patients incurred most of their outpatient visits within UWM in the following six months, even if we do not have data to verify this.

Potential use of our results

Our results show that for patients living within five miles of at least one of the three UWM hospitals, UWM provides most of their care and has reasonably complete medical data on them. For a data analysis task requiring relatively complete data, such as predictive modeling using historical data, we can conduct the task on this subset of patients and obtain useful results, even if conducting the task on all UWM patients is impractical. For example, we can build a predictive model to identify future high-cost patients among this subset [3]. Enrolling such patients in care management can help prevent high costs and improve outcomes [5-7].

Our results show that patients living further from the three UWM hospitals tend to receive a smaller portion of their care from UWM. This suggests UWM to consider using different preventive interventions for patients living at differing distances from the UWM hospitals, e.g., for care management to achieve better results. For patients who will receive only a small portion of their care from UWM, it is difficult for UWM to use expensive preventive interventions in a cost-effective manner.

This study used PreManage data covering adult patients in all age groups. It cannot be done using Medicare claims data that mainly cover patients aged ≥ 65 and patients with certain disabilities and diseases. Similar to many other healthcare systems, UWM does not have complete claims data covering all of its patients' healthcare use both within and outside of UWM. We could use claims data to do a similar study for another healthcare system, if it has complete claims data covering all of its patients' healthcare use both within and outside of its patients' healthcare use both within and outside of that system.

This study used PreManage data to validate the PCP constraint's effectiveness for UWM. However, the PCP constraint does not depend on PreManage data's availability and can be used by another healthcare system even if it cannot access PreManage data. In this case, one way to estimate the PCP constraint's effectiveness is to survey some of its patients about their healthcare use both within and outside of the system.

This paper focuses on identifying patients likely to receive most of their care from a specific healthcare system. If multiple healthcare systems exchange data, we could use a similar method to identify patients likely to receive most of their care from these healthcare systems combined. This could enable several data analysis tasks across these healthcare systems.

Limitations

This study has several limitations that can serve as interesting areas for future work:

- So far, UWM has accumulated PreManage data only over a limited period. After UWM accumulates more PreManage data, we should redo our analysis, check the percentage of patient hospital encounters that occur within UWM in the next 2-3 years, and see whether any of our conclusions will change.
- (2) This study demonstrates our constraint-based patient identification method at a single healthcare system, UWM, which provides both inpatient and outpatient care mainly for an urban area. To understand how our method generalizes, we should repeat our analysis on several other healthcare systems, some mainly serving urban areas and others offering many services in rural areas, and see whether the optimal constraint will change. For a healthcare system offering many services in rural areas, we would expect the optimal value of r to be >5, as patients are more scattered in rural areas than in urban areas.
- (3) For a healthcare system with incomplete medical data on many of its patients, we can use our method to identify a subset of patients, on whom the healthcare system has more complete data, and estimate the data's incompleteness level on this subset of patients. For a data analysis task, using incomplete data to do the analysis on this subset of patients would produce biased results, which could still be better than no result if the degree of bias is acceptable. Yet, the exact relationship between data incompleteness level and degree of bias in the analysis results is unknown. In particular, we have no idea of the threshold for the data's incompleteness level, beyond which the analysis' conclusion could become invalid. To address this issue, we can take a reasonably complete data set from another healthcare system like Kaiser Permanente, remove different portions of the data set, and check the resulting impact on the analysis results. This will help us understand whether our method is good enough for enabling the data analysis task in the current healthcare system.

5. Conclusions

To the best of our knowledge, for a healthcare system with incomplete medical data on many of its patients, we provided the first method to use a geographic constraint to identify a reasonably large subset of patients who tend to receive most of their care from the system. Our results show that our method performs reasonably well at UWM. Our method opens the door for conducting several major analysis tasks on incomplete medical data, which were previously deemed impractical to undertake.

Authors' contributions

GL was mainly responsible for the paper. He conceptualized and designed the study, performed literature review and data analysis, and wrote the paper. SL, PT, and AW offered feedback on various medical issues, contributed to conceptualizing the presentation, and revised the paper. SL also contributed to conceptualizing the study. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

Acknowledgments

We thank Xinran Liu, Richard Crouch, Lisa Swan, Kristal Mauritz-Miller, Christie Benevich, Philip J. Brewster, Ross Lordon, Lucy Wang, and Mika N. Sinanan for helpful discussions. GL was partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number R01HL142503. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

List of abbreviations

PCP: primary care physician UWM: University of Washington Medicine

References

- 1. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Arch Intern Med 2010;170(22):1989-95. PMID:21149756
- 2. Finnell JT, Overhage JM, Grannis S. All health care is not local: an evaluation of the distribution of emergency department care delivered in Indiana. AMIA Annu Symp Proc. 2011;2011:409-16. PMID:22195094
- Curry N, Billings J, Darin B, Dixon J, Williams M, Wennberg D. Predictive Risk Project Literature Review. London: King's Fund. http://www.kingsfund.org.uk/sites/files/kf/field/field_document/predictive-risk-literature-reviewjune2005.pdf. Archived at http://www.webcitation.org/6aga2XBFC
- 4. Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. Health Aff (Millwood) 2004;Suppl Web Exclusives:W4-427-36. PMID:15451964
- Vogeli C, Shields AE, Lee TA, Gibson TB, Marder WD, Weiss KB, Blumenthal D. Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. J Gen Intern Med 2007;22 Suppl 3:391-5. PMID:18026807
- 6. Caloyeras JP, Liu H, Exum E, Broderick M, Mattke S. Managing manifest diseases, but not health risks, saved PepsiCo money over seven years. Health Aff (Millwood) 2014;33(1):124-31. PMID:24395944
- Nelson L. Congressional Budget Office. Lessons from Medicare's demonstration projects on disease management and care coordination. https://www.cbo.gov/sites/default/files/112th-congress-2011-2012/workingpaper/WP2012-01 Nelson Medicare DMCC Demonstrations 1.pdf. Archived at http://www.webcitation.org/6agZcFxD1
- 8. Duncan I. Healthcare Risk Adjustment and Predictive Modeling. Winsted, CT: ACTEX Publications Inc.; 2011. ISBN:1566987695
- 9. Ash A, McCall N. Risk assessment of military populations to predict health care cost and utilization. http://www.rti.org/pubs/tricare_riskassessment_final_report_combined.pdf, 2005. Archived at http://www.webcitation.org/6aga7wYZC
- 10. Iezzoni LI. Risk adjustment for measuring health care outcomes, 4th ed. Chicago, IL: Health Administration Press; 2013. ISBN:1567934374
- 11. Luo G, Stone BL, Johnson MD, Tarczy-Hornoch P, Wilcox AB, Mooney SD, Sheng X, Haug PJ, Nkoy FL. Automating construction of machine learning models with clinical big data: proposal rationale and methods. JMIR Res Protoc 2017;6(8):e175. PMID:28851678
- 12. PreManage of Collective Medical Technologies Inc. http://collectivemedicaltech.com/what-we-do-2/premanage/. Archived at http://www.webcitation.org/6oOi8en1e
- 13. R's geosphere package homepage. https://cran.r-project.org/web/packages/geosphere/index.html. Archived at http://www.webcitation.org/72VG1whxE