Title

The Building Blocks of Inter-operability

Journal

Applied Clinical Informatics, 08(02)

ISSN

1869-0327

Authors

Culbertson, Adam
Goel, Satyender
Madden, Margaret B
et al.

Publication Date

2017-04-01

DOI

10.4338/aci-2016-11-ra-0196

The Building Blocks of Interoperability

A Multisite Analysis of Patient Demographic Attributes Available for Matching

Adam Culbertson[1]; Satyender Goel[2]; Margaret B. Madden[2]; Niloufar Safaeinili[2]; Kathryn L. Jackson[2]; Thomas Carton[3]; Russ Waitman[4]; Mei Liu[4]; Ashok Krishnamurthy[5]; Lauren Hall[6]; Nickie Cappella[7]; Shyam Visweswaran[7]; Michael J. Becich[7]; Reuben Applegate[8]; Elmer Bernstam[8]; Russell Rothman[9]; Michael Matheny[9]; Gloria Lipori[10]; Jiang Bian[10]; William Hogan[10]; Douglas Bell[11]; Andrew Martin[12]; Shaun Grannis[12]; Jeff Klann[13]; Rebecca Sutphen[14]; Amy B. O'Hara[15]; Abel Kho[2]

[1]HIMSS, Chicago, IL;
[2]Northwestern University, Chicago, IL;
[3]Lousiana Public Health Institute, New Orleans, LA;
[4]Kansas University Medical Center, Kansas City, KS;
[5]University of North Carolina, Chapel Hill, NC;
[6]BaylorScott & White Health, Dallas, TX;
[7]University of Pittsburgh School of Medicine, Pittsburgh, PA;
[8]University of Texas at Houston, Houston, TX;
[9]Vanderbilt University, Nashville, TN;
[10]University of Florida, Gainesville, FL;
[11]University of California, Los Angeles, CA;
[12]Regenstrief Institute, Indianapolis, IN;
[13]Harvard Medical School, Boston, MA;
[14]USF Morsani College of Medicine, Tampa, FL;
[15]Census Bureau, Suitland, MD

Keywords

Record linkage, master patient index, data completeness, data collection, data validation and verification, data processing

Summary

Background: Patient matching is a key barrier to achieving interoperability. Patient demographic elements must be consistently collected over time and region to be valuable elements for patient matching.

Objectives: We sought to determine what patient demographic attributes are collected at multiple institutions in the United States and see how their availability changes over time and across clinical sites.

Methods: We compiled a list of 36 demographic elements that stakeholders previously identified as essential patient demographic attributes that should be collected for the purpose of linking patient records. We studied a convenience sample of 9 health care systems from geographically distinct sites around the country. We identified changes in the availability of individual patient demographic attributes over time and across clinical sites.

Results: Several attributes were consistently available over the study period (2005–2014) including last name (99.96%), first name (99.95%), date of birth (98.82%), gender/sex (99.73%), postal code (94.71%), and full street address (94.65%). Other attributes changed significantly from 2005–2014: Social security number (SSN) availability declined from 83.3% to 50.44% ($p<0.0001$). Email address availability increased from 8.94% up to 54% availability ($p<0.0001$). Work phone number increased from 20.61% to 52.33% ($p<0.0001$).

Conclusions: Overall, first name, last name, date of birth, gender/sex and address were widely collected across institutional sites and over time. Availability of emerging attributes such as email and phone numbers are increasing while SSN use is declining. Understanding the relative availability of patient attributes can inform strategies for optimal matching in healthcare.

Correspondence to:
Adam Culbertson
4300 Wilson Blvd., Suite 250
Arlington, VA 22203
aculbertson@himss.org

Appl Clin Inform 2017; 8: 322–336
https://doi.org/10.4338/ACI-2016-11-RA-0196
received: November 16, 2016
accepted: January 21, 2017
published: April 5, 2017
Citation: Culbertson A, Goel S, Madden MB, Jackson KL, Carton T, Waitman R, Liu M, Krishnamurthy A, Hall L, Cappella N, Visweswaran S, Safaeinili N, Becich MJ, Applegate R, Bernstam E, Rothman R, Matheny M, Lipori G, Bian J, Hogan W, Bell D, Martin A, Grannis S, Klann J, Sutphen R, O'Hara AB, Kho A. The building blocks of interoperability: A multisite analysis of patient demographic attributes available for matching. Appl Clin Inform 2017; 8: 322–336
https://doi.org/10.4338/ACI-2016-11-RA-0196
Funding
Mike Becich received grants by CTSI: UL1TR001857–01, and PCORI PaTH: CDRN-1306–04912.

1. Introduction

The Health Information Technology for Economic and Clinical Health Act (HITECH) passed in 2009 provided over thirty-five billion dollars of incentive funding over five years [1]. The financial incentive program helped to rapidly increase the adoption of electronic health records [2]. This emerging technology promised to allow providers to document and share patient health information with the promise to increase quality and lower costs of patient care by facilitating care coordination. However, the lack of healthcare interoperability has led to bottlenecks in delivering on these promises. One key barrier to achieving health care interoperability is the lack of a consistent means to identify and match patient records across multiple care sites [3].

The failure to correctly match patients to their health care records has enormous financial and human cost. A Research and Development Corporation (RAND) study in 2005 estimated that an interoperable health care system could save as much as $371 billion annually [4]. A subsequent 2008 RAND study estimated that missing information, delayed treatment, and duplicate lab tests cost more than $8 billion annually [5]. In addition, each duplicate medical record can cost as much as $60 to resolve [6]. The magnitude of the problem is difficult to estimate but one study estimated potential duplicates using first and last name only as high as 40% [7]. Even more significant than financial cost is the potential harm to human health due to errors of patient matching. For example, a patient with an incorrectly matched record may receive the wrong medication, potentially causing an adverse drug event [8]. Additionally, a fragmented patient record resulting from inability to match health records may result in withholding lifesaving diabetes treatment [8]. The failure to adequately match patients can have significant results to patients, on the extreme end could lead to loss of life. For research purposes, duplicate patient records can lead to skewed results and inaccurate conclusions. The issue of duplicate patient medical records may lead to an increase in missed laboratory tests [9], an issue so significant that the Emergency Care Research Institute (ECRI) identified accurate patient identity as the second greatest safety concern in their annual list of the Top Ten Safety Concerns for Healthcare organizations in 2016. The inability to match records also has population health implications. For example, hospitals are required to track methicillin-resistant Staphylococcus Aureus (MRSA) cases and isolate infected patients. However, patients often travel to different healthcare institutions where their infection status is often unknown [11]. Failure to accurately identify patients can lead to increased rates of nosocomial MRSA infection [11].

Given the consequences resulting from the inability to accurately match patients, numerous groups have developed recommendations focusing on several key factors affecting the ability to match patients to their medical records. Two critical factors contribute to effective patient matching: (1) accurate and complete patient demographic attributes including; first name, last name, and date of birth used to ascertain match status, and (2) robust algorithms to classify record pairs as either a match or non-match. The quality of the data is an important factor that affects the ability to link patient health records. There are several dimensions of data quality that can be considered such as accuracy, completeness, consistency, and timeliness [12].

In this study we focused on the collection or completeness of different patient demographic attributes, and compared their overall and longitudinal availability among a convenience sample of geographically distinct healthcare institutions. The term availability refers to the presence of a non-null value entered into the field, excluding specific default values such as 999–9999 for each of the different patient demographic elements.

To be useful for patient matching, the demographic attributes have to be consistently collected by medical registration personnel and be highly available across time and region. Multiple stakeholder groups have published guidance regarding which patient demographic attributes should be collected to support patient matching. Healthcare Information and Management and Systems Society (HIMSS) recognized the importance of accurate patient matching and published the *Patient Identity Integrity White Paper* in 2009 [13], which recommended establishing a minimum set of data patient demographic attributes for matching. The Office of the National Coordinator for Health Information Technology (ONC) published *Privacy and Security Solutions for Interoperable Health Information Exchange* in 2009 [14], which made similar recommendations. ONC later published *Patient Identification and Matching Final Report in 2013* which further advocated studying the data attributes used in patient matching, development of a minimal set of patient demographic attributes,

and studying how new non-traditional attributes could improve match rates [15]. American Health Information Management Association (AHIMA) published *Patient Matching in Health Information Exchanges* in 2014 recommending a minimum set of patient data attributes for matching and offered a list of new attributes to be collected for matching [16]. In 2015, ONC released a paper *Connecting Health and Care for the Nation: A Shared Nationwide Interoperability Roadmap version 1.0,* which recommended identifying a core set of matching attributes to be included in health information exchanges [3]. More recently the Sequoia Project evaluated the completeness of common demographic data attributes and how each attribute affected match rates. They concluded that additional data attributes do not necessarily yield better matching results [17]. However the results were limited to a specific geographic region [17]. In summary, several patient matching stakeholders recommend standardizing existing patient demographic attributes, developing certification criteria for these standardization approaches, and studying additional non-traditional patient demographic attributes that may improve matching.

While previous recommendations identified attributes necessary for accurately matching medical records, we are unaware of any study characterizing the current availability of these patient demographic attributes on a national scale. Ideally each patient demographic attribute should be; unique, ubiquitous, unchanging, uncomplicated, uncontroversial, and easily accessible [18]. One of the most pertinent characteristics is the availability each patient demographic attribute and was the focus of this study. For an patient demographic attribute to be useful for patient matching algorithms they must be consistently collected across time and regions.

In this work we seek to evaluate the real-world availability of patient demographic attributes for the purpose of patient matching across the United States, and determine differences across time and region.

2. Research Methods

2.1 Study Design

The goal of our study was to better understand what patient demographic attributes different clinical sites collect that could be used for algorithms to determine rather two records correspond to the same patient matching. We compiled a list of 36 patient attributes pooled from previous guidance published over the past decade that make recommendations on what different patient demographic attributes are available [3, 13–16] (▶Table 1). A data dictionary defining these 36 attributes and pseudo-code for extracting attribute counts and percent availability from the site's EHR system or Electronic Data Warehouse were distributed to each of nine participating sites. Sites were instructed to return counts and percent availability from 2005 to 2014, overall, by each year and for the overall survey period combined. Only patients having at least one visit to the site since 2005 and who were between 18 and 89 years of age at time of extraction were included. For overall counts and percentages, patients were included only once in each attribute count (that is, if a patient was seen multiple times over the study period, his attribute information was included only once, as measured using the site's master patient index number at the time of last visit). For yearly counts and percentages, sites were requested to report the total number of patients seen in that year (denominator), and having the attribute present in that year (numerator); Sites were requested to report total number of patients seen in an individual year (denominator), and the presence the patient demographic attribute (numerator) for that year. When historical records of patient demographic attributes by year where not available, sites were requested to include patients in the numerator and denominator only for the year of their last visit. Metadata was sent from each site to Northwestern University for compilation and statistical analysis.

2.2 Sites and Settings

To achieve our research goal and gain a snapshot of the data attributes available for patient matching, we identified nine health care institutions representing a convenience sample of systems from geographically distinct sites around the country (▶Figure 1) from 7 of the 13 Patient Centered-Out-

comes Research Institutes (PCORI) [19] Clinical Data Research Networks (CDRNs), including REACHnet [20], GPC [21], Capricorn [22], MidSouth [23], OneFlordia [24], and PATH [25, 26], pSCANNER [27]. Data at each site was individually collected through queries of the institution's Electronic Data Warehouse (EDW). Final metadata results from each site were collected and compiled for analysis at Northwestern University.

2.3 Data Analysis

Metadata counts from the nine sites were aggregated and percent availability for each attribute was calculated across all sites. Overall counts were calculated by summing the overall numerators and denominators from each site; yearly counts from 2005–2014, were calculated by summing the yearly numerators and denominators from each site. The percent availability of a particular attribute was calculated as the number of patients with non-null (or non-missing) values divided by the total number of patients. Minimum, maximum, and mean percent availability were calculated. Variation and trends in attribute availability across time was assessed using Pearson's chi-squared tests, correlation coefficients, and Cochran-Armitage tests for trend. Additionally, to understand the contribution of each site to overall yearly trends, we calculated the proportion of total patients (overall and by year) for each site. All statistical analysis was performed using SAS v. 9.4.

3. Results

We received data from nine sites that were located in cities covering North, South, East, West, and Midwest regions of the United States.

3.1 Total Availability of Attributes

The aggregated results revealed which demographic attributes are most commonly collected. ▶Figure 2 highlights the overall average, minimum, and maximum percent availability for the 36 data attributes across all sites. Last name (99.96%), first name (99.95%), date of birth (99.82%), gender/sex (99.73%), state (95.95%), street address (94.99%), city (94.74%), postal code (94.71%), and full address (94.65%) all had overall availability near or above 95%. ▶Figure 2 shows the average, minimum and maximum availability for each patient demographic attribute. The attributes of country (full, abbreviated), race (OMB, free text), date of birth (birth year) were reported as one row in the figure.

3.2 Attributes that were Consistently Collected Across Location

▶Figure 3 compares attributes across sites. The values of first name, last name, date of birth, and gender/sex were all highly available across sites (range 98.97–100%). Geographic features that were not as consistently available included SSN (69.22-99.96%), full address (81.06–100%), and email address (0–87.00%).

3.3 Attributes that were Consistently Collected Across Time

▶Figure 4 displays the set of attributes having 95% or greater availability overall from 2005–2014. This group of patient demographic attributes was stable across time with a slight decrease from 2006–2007, where the percentage for address attributes dipped to 93%. This decrease is likely due to change in single site with a larger number of total patients, and was not statistically significant. The person attributes are the combination four attributes of last name, first name, date of birth and gender. The location attributes are the combination five attributes including city, state, street address, postal code, and full address.

3.4 Attributes that were Rapidly Changing

For other attributes, availability is not static over time. Social security number (SSN) and email address both exhibited notable changes (▶Figure 5). SSN showed a statistically significant decline in availability from 2005–2014 ($p<0.0001$). The overall percent availability for SSN dropped from 83.30% in 2005 to only 50.44% in 2014. Additionally, SSN availability across sites varied significantly; in our sample, one site did not report any collection of SSN, while other sites reported nearly complete capture (range 0–99.96%). On average, SSN was available for 69.22% of our study population over the entire course of data collection.

While SSN showed a steady decline in overall availability, the availability of email address increased over time ($p<0.0001$) (▶Figure 5). In 2005, only 8.94% of patients in our sample population had any email address recorded; by 2014, this percentage increased to 54.08%. Over the entire study period, email was available for 23.88%, and ranged from 0 to 87.00% availability by individual site.

Like email, phone number availability increased over time (▶Figure 6). Availability of both home and work phone number significantly increased over time ($p<0.0001$ for both). Home phone number availability increased from 62.27% to 92.31% from 2005 to 2014; availability for work phone number, while lower than that of home phone number, also increased (from 20.61% to 52.33%) over the study period. The availability of cell phone number also increased over time, however, this increase was not statistically significant ($p=0.0179$).

4. Discussion

This work evaluated availability of patient matching data attributes. We examined how the availability of these attributes changes over time and across regions in the United States. Previous patient matching recommendations identified a selection of attributes to be standardized and collected, but lacked information describing the availability of these patient demographic attributes. In addition, we observed that the collection of attributes changed over time and varied by region; understanding the trends in data collection practices are important for a nationwide patient matching strategy. Our results can inform institutional strategies regarding the choice of data attributes for patient matching, as understanding the trends of data collection practices are important for a nationwide patient matching approach.

We found several interesting results which could assist in formulating a nationwide patient matching strategy. We confirmed that some data attributes including first name, last name, and date of birth, gender/sex, and address (either full or partial) were highly available across time at each institution. However, future work is needed to determine the effectiveness of these attributes in supporting accurate patient matching in large demographic database of patient data. Work such as the Rand Study was done in 2008 includes estimates for false positives using SSA Death Master File. But the work now a bit dated and the work should be replicated using data that is representative of a health care system [5].

Beyond this initial set of highly available attributes, our results also suggest that email address and phone numbers (home, cell, work) may soon be more widely available. Our data showed large variation in the capture of each type of phone number in the medical record. Overall cell phone number availability was very low (6%) in 2014. However, this percentage may, in reality, be much higher, as studies have noted that 40% of adults live in 'cellphone-only households' in the United States [28]. This suggests that our results may show inaccurate capture of this variable in clinical systems, rather than simply a lack of availability. Additionally, home phone numbers may be associated with multiple people in a household, while cell phones tend to be associated with a single individual [29], and thus cell number may be preferred over home as an attribute for patient matching. The requirement of accurate entry of phone number and email address or protocols (including direct patient authentication) ensuring valid email addresses may make these attributes an increasingly valuable candidate for matching. Future work should examine the accuracy of patient matching when email address and phone numbers are included in the matching algorithm at varying levels of completeness. Email address should be considered as one of the core elements to be shared for linking data.

Conversely, the availability of SSN has declined in recent years. This may be due to increased awareness of privacy related issues such as identify theft [30]. Social security number is a highly discriminating variable which should yield an accurate patient match in many situations using SSN alone. However, declining availability may adversely affect the ability to match patient records using this variable, and further research should identify the effect declining SSN has on match accuracy. The decline of SSN may lead to a rethink for the need to add additional highly specific data elements. Organizations that are highly reliant on SSN as a matching variable in the short term should consider the addition of emerging elements such as email. In the long term this finding may serve to reinvigorate the debate on what other highly specific data elements should be included. Effective nationwide patient matching strategies or at minimum shared best practices are urgently needed. It is hoped that this work can advance the discussion on the current state of patient data matching and will further encourage additional work in the field.

Study Limitations

There were several limitations of this work. First, we studied a small convenience sample of sites across the country that focused on academic medical institutions. The participating sites may not necessarily be representative of their regions. Additionally, some sites had a larger proportion of our overall patient population (▶Supplementary Table 1), and therefore, larger sites might have a stronger influence on the overall percent availability. However, the goal of the work was to gain a sense of the trends that are occurring nationally. Future work should extend the evaluation to a larger and more diverse number of sites more representative of the United States. Second, sites queried metadata from data warehouses; actual production systems may contain additional demographic attributes not identified in this work. Due to the limited scope of the work, we evaluated, arguably one of the simplest aspect of data quality: availability. Important dimensions of data quality such as the accuracy, consistency, accessibly, and believability were not examined [12]. This would require an examination of the data more closely than looking at meta-data counts. We did not evaluate if the accuracy of specific patient demographic attributes were correct; we only measured the presence of some non-null value for each field. For example the study did not evaluate whether John Doe was recorded as J. Doe or any other permutation of a given patient name that could occur. This work should be extended to drill down further into the dimensions of data quality such as accuracy and completeness. It should be noted that many of these aspects of data quality are difficult to measure as it requires an accepted reference that is considered the reference data set or gold standard. Future work would require a strong consideration of how to access HIPAA protected data and what Institutional Review Board (IRB) would need to be in place. This work stopped short of such in-depth analysis of meta-data and focused on a convenience sample over time and large geographic regions of the United States. We only looked at whether a given field contained a non-null value and how this changed over time and region. Future work should evaluate more complex data quality measures such as completeness, consistency, conformity, accuracy, integrity, timeliness or other frameworks for data quality [31]. However, this future work would require analysis at the patient level and would therefore require IRB review. Finally, while our study did span from 2005–2014, we were only able to examine a high level snapshot of the data at a specific period of time; repeating this analysis to add additional years would be useful to further identify trends in attribute availability.

5. Conclusion

Overall, analysis of data quality is required to measure the effectiveness of different algorithmic approaches to automatically match patient records using combinations of patient demographic attributes such as those identified in this work. Before advances can be made, it is essential to understand the availability, quality, and limitations of the currently available patient demographic data as well as understand how well current matching algorithms can perform given the quality of most highly available patient demographic attributes. This study identifies consistently collected attributes which could be used as an input for a patient matching algorithm. Future work may help

characterize how well existing matching algorithms perform on a reference data set in which matches and non-matches are known using the most highly available set of attributes identified in this work. Additionally, the study of highly available attributes can help inform a national discussion on what patient demographic elements are available for matching and which additional elements may be needed to achieve cross institutional interoperability. The ability to match patients to their health data is an essential component if the value of electronic medical records is to be realized.

Questions

What set of attributes were highly available in this particular study?
A) FN, SNN, LN, DOB
B) FN, LN, DOB, Address, Gender/Sex
C) SSN, Email, ADDRESS, Insurance Number
D) SSN, INSURANCE NUMBER

What set combination of attributes were rapidly decreasing and rapidly increasing over the survey period?
A) Email, Gender/Sex, DOB, SSN
B) FN, LN
C) MI, INSURANCE NUMBER
D) SSN, EMAIL

Clinical Relevance Statement

Patient matching is a critical barrier to achieving interoperability. The ability to matching patients is a function of the patient demographic elements available to match patients and the algorithms or methods used.

Conflicts of Interest

Abel Kho and Satyender Goel are stake holders in Health Data Link LLC, a PPRL software solution company.

Human Subjects Protections

This study did not collect actual patient data. The work only collected statistics on the meta-data about how often a demographic field contained a value other than null or default values. Therefore the work was exempt from requiring IRB approval since no actual patient data was used.

Acknowledgements

We would like to thank all the clinical sites that contributed meta-data counts to the study. Special thanks to Maggie Madden for leading the statistical analysis of the combined sites data. Jess Behrens for help in the development of figures.

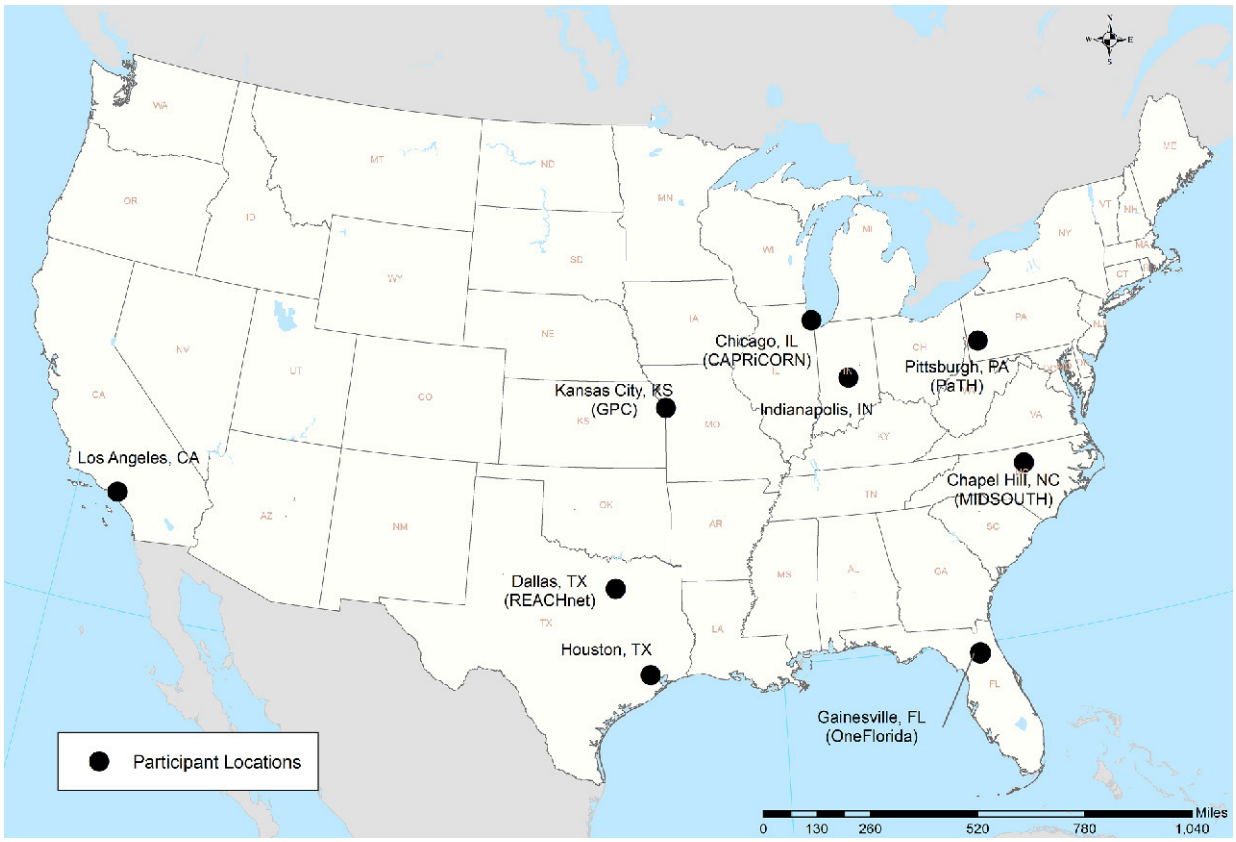

Fig. 1 Geographical distribution of participating sites

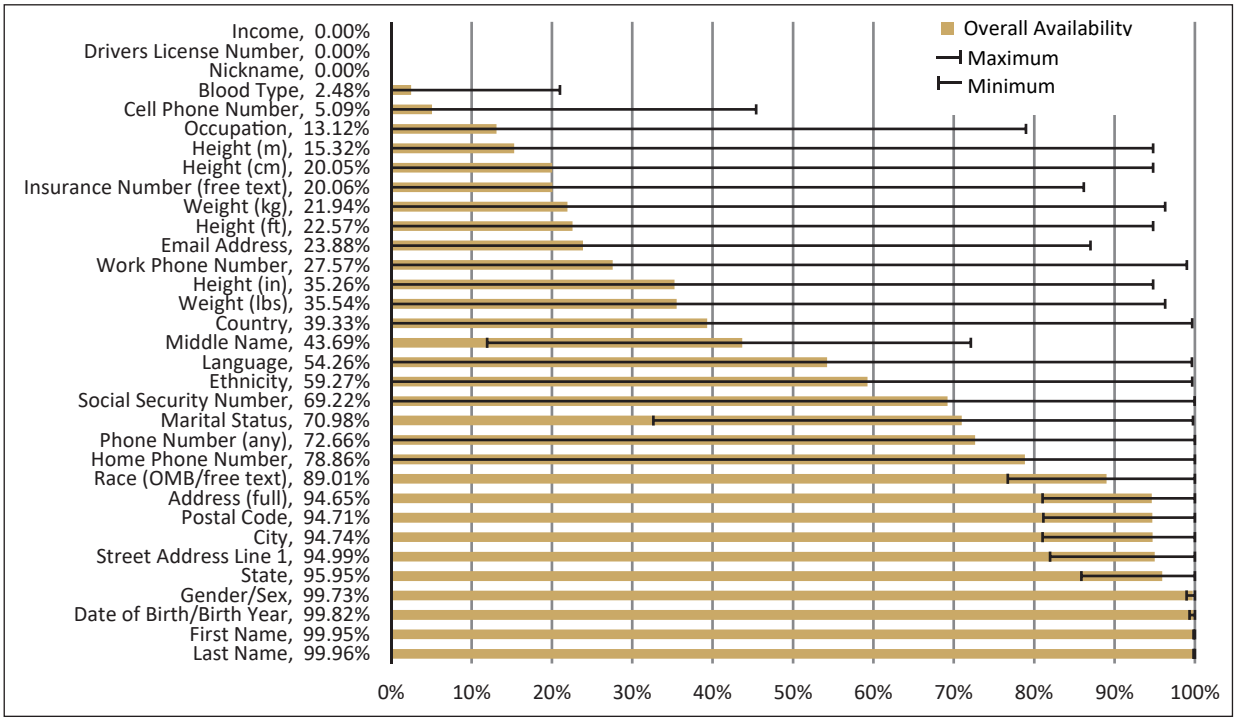

Fig. 2 Percent availability of patient demographic attributes across all regions from 2005–2014

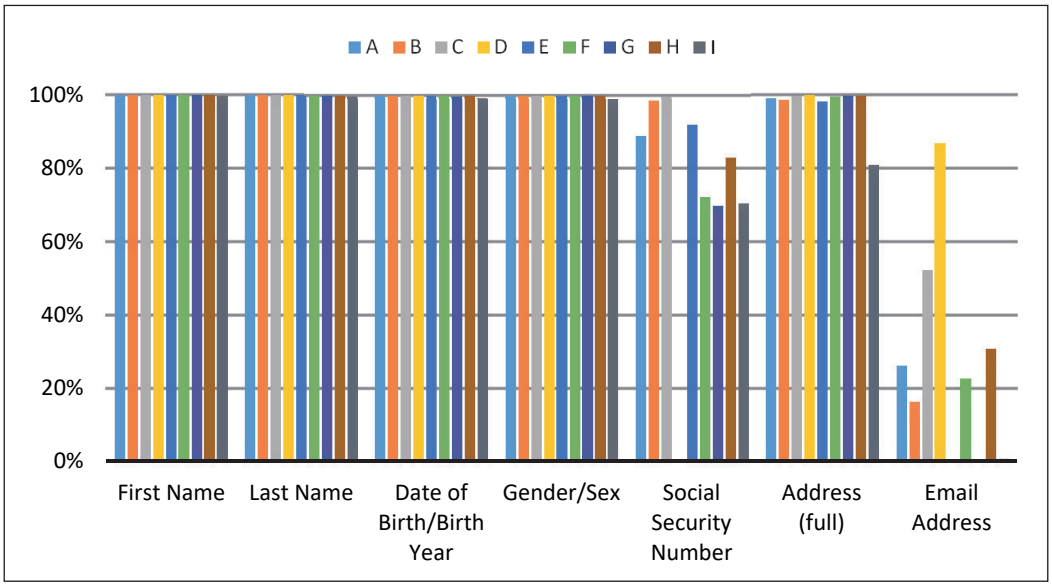

Fig. 3 Comparison of patient demographic attributes across sites (A-I)

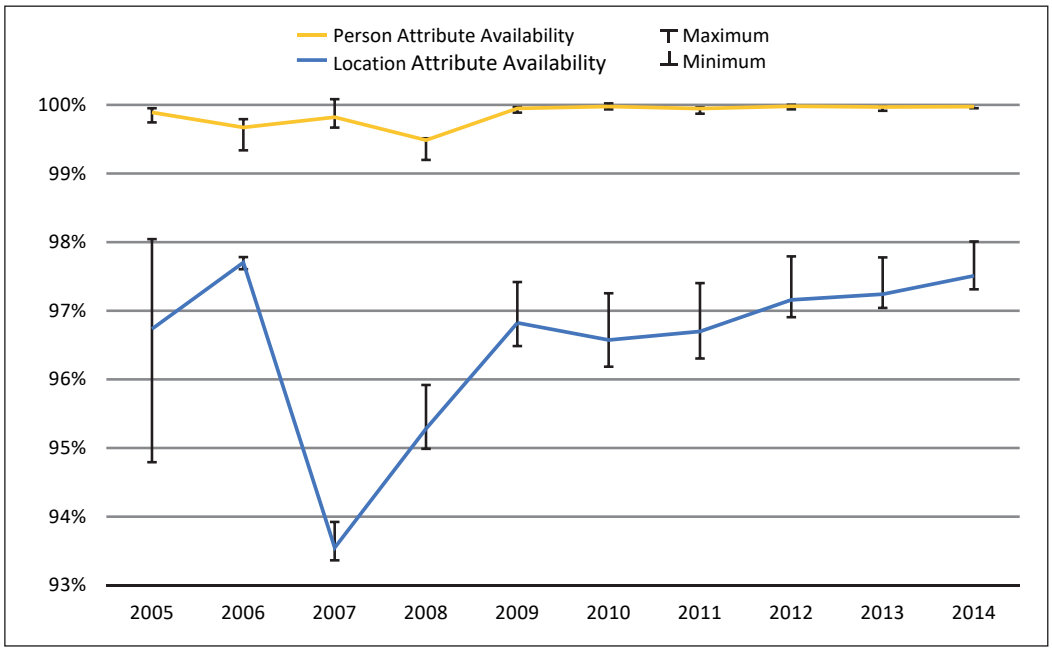

Fig. 4 Attributes with high availability over time

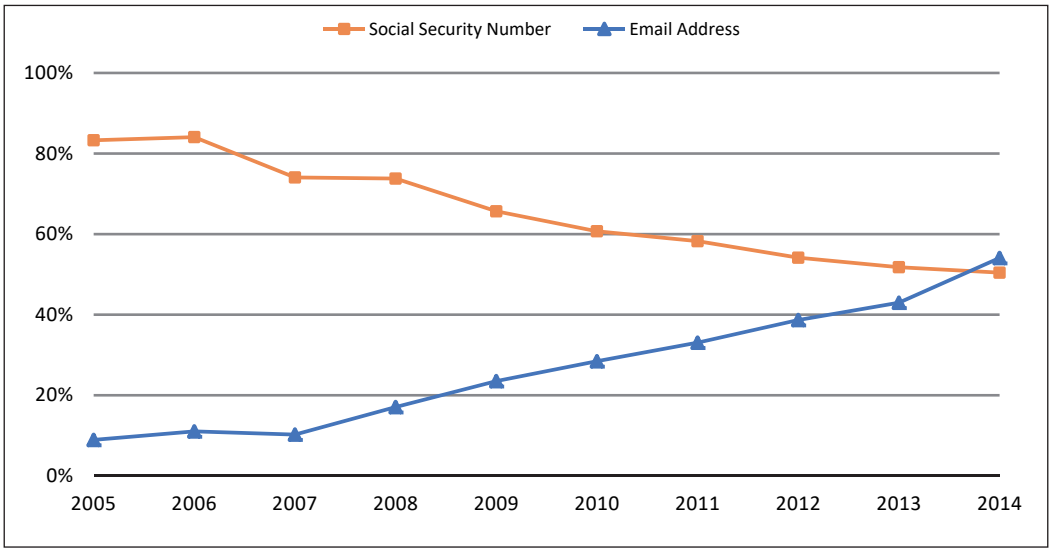

Fig. 5 Availability of SSN email over time

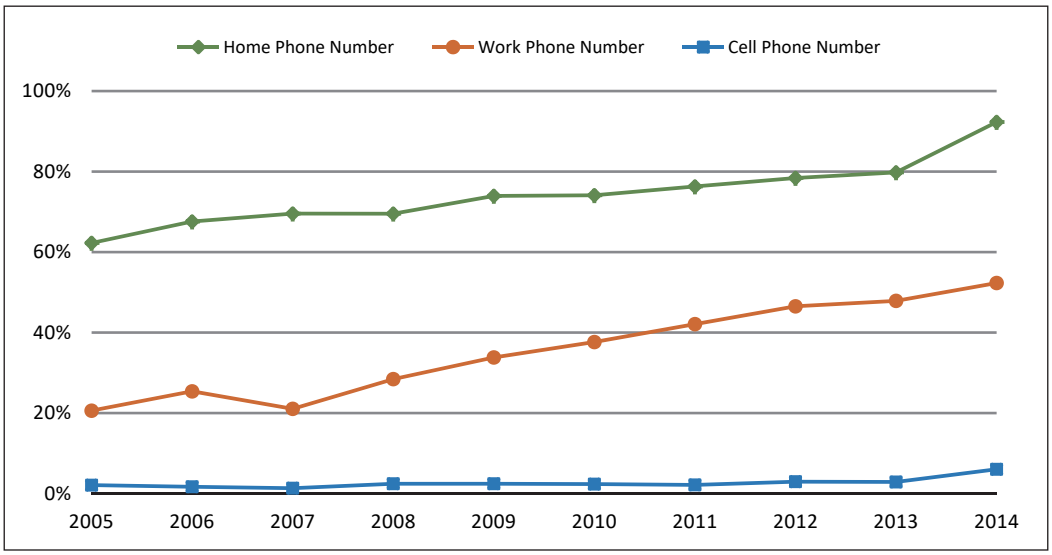

Fig. 6 Emerging attributes of phone availability plotted across time

Table 1 Patient level demographic attributes overall and by year for individual institutions.

First Name	City	Email Address	Income
Middle Name	State	Nickname	Marital Status
Last Name	Postal Code	Insurance Number (free text)	Height (cm)
Date of Birth	Country Abbreviation	Driver's License Number	Height (m)
Birth Year	Country Full Name	Race (OMB)	Height (in)
Gender	Phone Number (any)	Race (free text)	Height (ft)
Social Security Number	Home Phone Number	Ethnicity	Weight (lbs)
Address (full)	Cell Phone Number	Language	Weight (kg)
Street Address Line 1	Work Phone Number	Occupation	Blood Type

References

1. Where Is HITECH's $35 Billion Dollar Investment Going? 2016. Available from: http://healthaffairs.org/blog/2015/03/04/where-is-hitechs-35-billion-dollar-investment-going/.
2. Hsiao C-J, Jha AK, King J, Patel V, Furukawa MF, Mostashari F. Office-based physicians are responding to incentives and assistance by adopting and using electronic health records. Health Affairs 2013: 10.1377/hlthaff. 2013.0323.
3. The Office of the National Coordinator for Health Information Technology. Connecting Health and Care for the Nation: A Shared Nationwide Interoperability Roadmap; October 6th 2015. Available from https://www.healthit.gov/sites/default/files/hie-interoperability/Interoperibility-Road-Map-Supplemental.pdf.
4. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, Taylor R. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health affairs 2005; 24(5): 1103-1117.
5. Hillestad R, Bigelow J, Chaudhry B, Dreyer P, Greenberg MD, Meili RC, Ridgely MS, Rothenbery J, Taylor R. Identity crisis: An examination of the costs and benefits of a unique patient identifier for the US health care system: RAND Corporation; 2008.
6. Thornton SN, Hood SK, editors. Reducing duplicate patient creation using a probabilistic matching algorithm in an open-access community data sharing environment. AMIA Annual Symposium Proceedings; 2005: American Medical Informatics Association.
7. McCoy AB, Wright A, Kahn MG, Shapiro JS, Bernstam EV, Sittig DF. Matching identifiers in electronic health records: implications for duplicate records and patient safety. BMJ quality & safety 2013; 22(3): 219-224.
8. McDonald CJ. Computerization can create safety hazards: a bar-coding near miss. Annals of Internal Medicine 2006; 144(7): 510-516.
9. Joffe E, Bearden CF, Byrne MJ, Bernstam EV. AMIA Annual Symposium Proceedings 2012, November 3: 1269–1275; Chicago, IL.
10. ECRI. Top Ten Patient Safety Concerns for 2016 [cited 2016 August 8th]. Available from: https://www.ecri.org/Pages/Top-10-Patient-Safety-Concerns.aspx.
11. Kho AN, Lemmon L, Commiskey M, Wilson SJ, McDonald CJ. Use of a regional health information exchange to detect crossover of patients with MRSA between urban hospitals. Journal of the American Medical Informatics Association 2008; 15(2): 212-216.
12. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection: Springer Science & Business Media; 2012.
13. HIMSS. Patient Identity Integrity Toolkit. Chicago, IL: HIMSS; December 2009. Available from: http://www.himss.org/library/healthcare-privacy-security/patient-identity.
14. Dimitropoulos, Linda L. Privacy and security solutions for interoperable health information exchange[Internet]. Chicago; RTI International; 2007 [cited 2016 Nov 2016]. Available from: http://media.khi.org/news/documents/2009/08/28/HISPC_Privacy_and_Security_Solutions.pdf
15. Morris G, Farnum G, Afzal S, Robinson C, Greene J, Coughlin C. Patient identification and matching final report [Internet]. Bailtimore: Audicous Inquiry; 2014 [cited 2016 Sep 12]. Available from: https://www.healthit.gov/sites/default/files/patient_identification_matching_final_report.pdf.
16. Lusk KG, Neysa Noreen RH, Godwin Okafor RH, Kimberly Peterson MH, Erik Pupo MB. Patient Matching in Health Information Exchanges. Perspectives in Health Information Management AHIMA [Internet]. 2014 [cited 2016 Nov 11]]; Available from: http://perspectives.ahima.org/wp-content/uploads/2014/12/PatientMatchinginHIEs.pdf
17. The Sequoia Project. A Framework for Cross-Organizational Patient Identity Matching. The Sequoia Project; 2015 Nov. 10th. Available from http://sequoiaproject.org/wp-content/uploads/2015/11/The-Sequoia-Project-Framework-for-Patient-Identity-Management.pdf.
18. Timothy D. McFarlane BED, Shaun J. Grannis. Client Registries: Identifying and Linking Patients. 1st Edition. Elsevier. Chapter 11, p. 163–182.
19. Pcornetwork. About PCORnet – PCORnet. 2016. Available from: http://www.pcornet.org/.
20. ReachNET 2016. Available from: http://www.reachnet.org/.
21. What is the Greater Plains Collaborative? Greater Plains Collaborative (GPC) 2016. Available from: http://www.gpcnetwork.org/.
22. CAPriCORN. The Chicago Area Patient Outcomes Research Network 2016. Available from: http://capricorncdrn.org/.
23. COLLABORATE. Welcome to the Mid-South Clinical Data Research Network (CDRN) 2016. Available from: https://midsouthcdrn.mc.vanderbilt.edu/collaborate/.

24. OneFlorida 2016. Available from: http://onefloridaconsortium.org/.
25. PaTH Network. 2016. Available from http://pathnetwork.org/.
26. Amin W, Tsui FR, Borromeo C, Chuang CH, Espino JU, Ford D, Hwang W, Kapoor W, Lehmann H, Martich GH, Morton S, Paranjape A, Shirey W, Sorensen A, Becich MJ, Hess R, Path Network Team. PaTH: towards a learning health system in the Mid-Atlantic region. Journal of the American Medical Informatics Association 2014; 21(4): 633-636.
27. Ohno-Machado L, Agha Z, Bell DS, Dahm L, Day ME, Doctor JN, Gabriel D, Kahlon MK, Kim KK, Hogarth M, Matheny ME, Meeker D, Nebeker JR, pScanner Team. pSCANNER: patient-centered Scalable National Network for Effectiveness Research. Journal of the American Medical Informatics Association 2014; 21(4): 621-626.
28. Rainie L, Zickuhr K. Always on Connectivity. Washington, DC: Pew Research; 2015 Aug. 26th. Available from http://www.pewinternet.org/2015/08/26/americans-views-on-mobile-etiquette/.
29. Busse B, Fuchs M. Prevalence of Cell Phone Sharing. Survey Methods: Insights from the Field. 2013. 2013 March 3 [Cited 2016 Nov 12th]. Available from Retrieved from: http://surveyinsights.org/?p=1019.
30. Madden M. More online Americans say they've experienced a personal data breach. Washington, DC: Pew Research Center; 2014 April 14th. Available from: http://www.pewresearch.org/fact-tank/2014/04/14/more-online-americans-say-theyve-experienced-a-personal-data-breach/.
31. Smartbridge. Data Done Right: 6 Dimensions of Data Quality (Part 1). Houston, TX: Smartbridge; 2013 Aug 9th. Available from: http://smartbridge.com/data-done-right-6-dimensions-of-data-quality-part-1/.

Supplemental Table 1 Table list the percent contribution for each site by year and overall.

Site Name	All Years	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
A	6%	3%	2%	3%	3%	2%	2%	2%	2%	2%	6%
B	12%	19%	31%	23%	25%	22%	20%	19%	17%	18%	19%
C	3%	0%	0%	0%	0%	0%	0%	0%	3%	3%	7%
D	14%	6%	7%	8%	14%	22%	27%	31%	35%	37%	38%
E	7%	7%	6%	5%	5%	4%	4%	4%	4%	3%	4%
F	13%	11%	8%	6%	7%	7%	6%	6%	6%	6%	7%
G	8%	27%	22%	17%	20%	17%	16%	15%	13%	12%	0%
H	11%	8%	7%	5%	6%	5%	5%	5%	5%	5%	8%
I	26%	18%	16%	34%	20%	21%	19%	19%	15%	13%	12%
median	11%	8%	7%	6%	7%	7%	6%	6%	6%	6%	7%
min	3%	0%	0%	0%	0%	0%	0%	0%	2%	2%	0%
max	26%	27%	31%	34%	25%	22%	27%	31%	35%	37%	38%