Journal of the American Medical Informatics Association, 24(e1), 2017, e111–e120
doi: 10.1093/jamia/ocw124
Advance Access Publication Date: 28 August 2016
Research and Applications

Research and Applications

Identifying collaborative care teams through electronic medical record utilization patterns

You Chen,[1] Nancy M Lorenzi,[1,2] Warren S Sandberg,[1,3] Kelly Wolgast,[2,4] and Bradley A Malin[1,5]

[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA, [2]School of Nursing, Vanderbilt University, [3]Department of Anesthesiology, Vanderbilt University, [4]Healthcare Leadership Program, School of Nursing, Vanderbilt University, and [5]Department of Electrical Engineering and Computer Science, Vanderbilt University

Corresponding Author: You Chen, PhD, 2525 West End Ave, Suite 1475, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37203, USA. E-mail: you.chen@vanderbilt.edu

Received 29 March 2016; Revised 15 July 2016; Accepted 20 July 2016

ABSTRACT

Objective: The goal of this investigation was to determine whether automated approaches can learn patient-oriented care teams via utilization of an electronic medical record (EMR) system.

Materials and Methods: To perform this investigation, we designed a data-mining framework that relies on a combination of latent topic modeling and network analysis to infer patterns of collaborative teams. We applied the framework to the EMR utilization records of over 10 000 employees and 17 000 inpatients at a large academic medical center during a 4-month window in 2010. Next, we conducted an extrinsic evaluation of the patterns to determine the plausibility of the inferred care teams via surveys with knowledgeable experts. Finally, we conducted an intrinsic evaluation to contextualize each team in terms of collaboration strength (via a cluster coefficient) and clinical credibility (via associations between teams and patient comorbidities).

Results: The framework discovered 34 collaborative care teams, 27 (79.4%) of which were confirmed as administratively plausible. Of those, 26 teams depicted strong collaborations, with a cluster coefficient > 0.5. There were 119 diagnostic conditions associated with 34 care teams. Additionally, to provide clarity on how the survey respondents arrived at their determinations, we worked with several oncologists to develop an illustrative example of how a certain team functions in cancer care.

Discussion: Inferred collaborative teams are plausible; translating such patterns into optimized collaborative care will require administrative review and integration with management practices.

Conclusions: EMR utilization records can be mined for collaborative care patterns in large complex medical centers.

Key words: collaborative networks, electronic medical records, health care organization modeling, data mining

BACKGROUND AND SIGNIFICANCE

A potential factor driving high costs in US health care is a fragmented care delivery and payment structure that can lead to overtreatment and overpayment.[1,2] In response, health care organizations (HCOs) are increasingly investing in strategies to encourage greater coordination and collaboration among providers.[3–9] However, current care collaboration structures generally are specialty-centric and place great reliance on discrete services.[10–15] A greater understanding of current models for collaboration is an essential prerequisite to developing richer collaboration models.

Traditionally, 2 general classes of approaches have informed the creation and management of collaborative care. The first is based on the proactive creation of care teams that are oriented to diagnose, treat, and manage specific diseases.[10–12] This strategy leverages the

knowledge of established experts and professional affiliations, but the resulting collaborations focus narrowly on certain disorders or medical problem sets (eg, gastrointestinal and heart diseases), with difficulty in realizing expert care for unrelated comorbidities.[12] An increasingly recognized alternative is based on data science strategies, whereby the presence and composition of collaborations are learned through the analysis of administrative data, such as insurance claims.[16–23] However, to date, this perspective has been limited in application, for several notable reasons. First, by relying on claims data, the collaborations correspond only to those identified by payments for the primary problems ailing a patient at the time of care. Moreover, claims data only captures the interactions of care providers and patients based on billed conditions and procedures, but fail to capture other types of interactions (eg, writing clinical notes, administering medications, and simply checking on patient status), which are events critical to patient care but are rarely viewed outside of a health care system. A second hurdle is that the learned collaborative care models are rarely validated by administrative experts or clinical leaders in charge of transitioning such discoveries into practice.[5]

There has been limited investigation into how to establish and manage collaborative care that is patient-oriented by capturing the interactions of electronic medical record (EMR) users and patients based on operational actions.[5,9] EMR systems provide opportunities to (1) investigate how care providers collaborate in the care of patients and (2) investigate the presence of unrecognized ad hoc collaborations. This is notable because EMR systems are longitudinal and can provide a detailed view of the interactions among care providers in patient management across the health care organization.[16,24–26] They also document the majority of actions that care providers take with respect to their patients.[27–29] This information can enable researchers to focus on all elements of patient care, rather than solely on reimbursable services (which miss interactions of EMR users and patients based on operational actions). Such a view can shift the focus from service-oriented care teams to integrated patient-oriented care teams.[17]

RESEARCH DESIGN AND METHODS

For this study, we developed a novel unsupervised data-mining pipeline to learn collaborative care teams. We recognized that the patterns derived via unsupervised learning require review before transitioning into the clinical environment, such that we performed multiple types of evaluations and contextualization of the inferred care teams. Specifically, we conducted 2 types of evaluations of the system. First, we conducted an extrinsic evaluation to determine the plausibility of inferred teams with clinical and administrative experts. Second, we conducted an intrinsic evaluation to determine the collaborative strength (via a clustering coefficient) and clinical credibility (via associations with patient comorbidities) of each team. In this section, we introduce the data studied, the details of the pipeline, and the specific methods by which the evaluations were conducted.

Dataset

This study is based on 4 months' worth of de-identified data from the StarPanel EMR system of Vanderbilt University Medical Center (VUMC).[16,30] The data were collected during 4 months in 2010 and contain information on 486 documented operational areas, 10 659 HCO employees, 17 947 inpatients, and 5176 unique International Statistical Classification of Diseases Version 9 (ICD-9) billing codes.

This investigation leverages ICD-9 codes to group patients, and then models the interactions of care providers at the patient group level. We acknowledge that such billing codes are insufficient and may not be a completely accurate representation of precise patient status.[31,32] To address such a limitation, we rely on the phenome-wide association study (PheWAS) vocabulary, which was introduced to reduce variability in the definitions of clinical concepts in secondary data use scenarios.[33] Building on the successful application of PheWAS in various association studies,[34] we translated each ICD-9 code into its PheWAS term, each of which corresponds to a group of ICD-9 codes.[33] After such transformation, the dataset consisted of 1413 PheWAS codes.

The dataset consisted of 831 721 unique operational actions between the employees and the EMRs, 74 192 assignments of PheWAS codes to patients, and 10 667 affiliations of VUMC employees to operational areas. The interactions of employees are modeled based on (1) the operational actions between employees and patients and (2) the groups of patients inferred based on PheWAS codes.

Organizational component learning modules

We translated utilization of an EMR into organizational components via a series of transformations. Figure 1 summarizes the process that translated utilization data into organizational components, while Figure 2 provides an example of the transformation to guide the reader. Initially, we characterized utilization of an EMR through 3 variables, which are represented as matrices: (1) $A_{\text{diagnosis}\times\text{patient}}$ charactering the assignment of diagnoses to patients (Figures 1a and 2a); (2) $B_{\text{patient}\times\text{user}}$ representing the management of patients by users (Figures 1c and 2c); and (3) $C_{\text{user}\times\text{operation}}$ representing the affiliations of users to their operational areas (Figures 1e and 2f). These data then proceed through 4 transformations. In the following, note that we use X' to represent the transposition of matrix X.

Transform 1 (Place individual patients into groups of patients with similar clinical concepts): It has been shown that interactions of HCO employees over a group of patients with similar clinical concepts are more meaningful for characterizing collaborative relations than interactions with respect to each patient.[35] For instance, certain hematologists may focus on the diagnosis of myeloid leukemia, while others may specialize in the diagnosis of lung malignancy. These care providers are potentially related to one another because they deal with cancer; however, if we fail to relate these diagnoses, we may not discover the relationship between the care providers.

Topic modeling has been shown to be an effective strategy to infer clinical concepts via EMR data.[24,36] By relying on topics rather than discrete diagnoses, patients with similar clinical concepts (topics) can be grouped together (Figure 1b). Thus, we use a topic inference strategy[30] (Diagnosis Topic Inference (DTI) Module in Figure 1) to derive clinical "concepts" from $A'_{\text{diagnosis}\times\text{patient}}$ (Figure 2a). Each concept (topic) is characterized as a vector representing probabilities of diagnosis codes assigned to it, as shown in $T_{\text{topic}\times\text{diagnosis}}$ (Figure 2b). Additionally, each patient is characterized by a vector of topics ($L_{\text{patient}\times\text{topic}}$ in Figure 2d) via DTI.

Transform 2 (Focus on interactions of users based on individual patients into interactions over groups of patients): This transformation (User Interaction Learning Module in Figure 1) proceeds as follows:

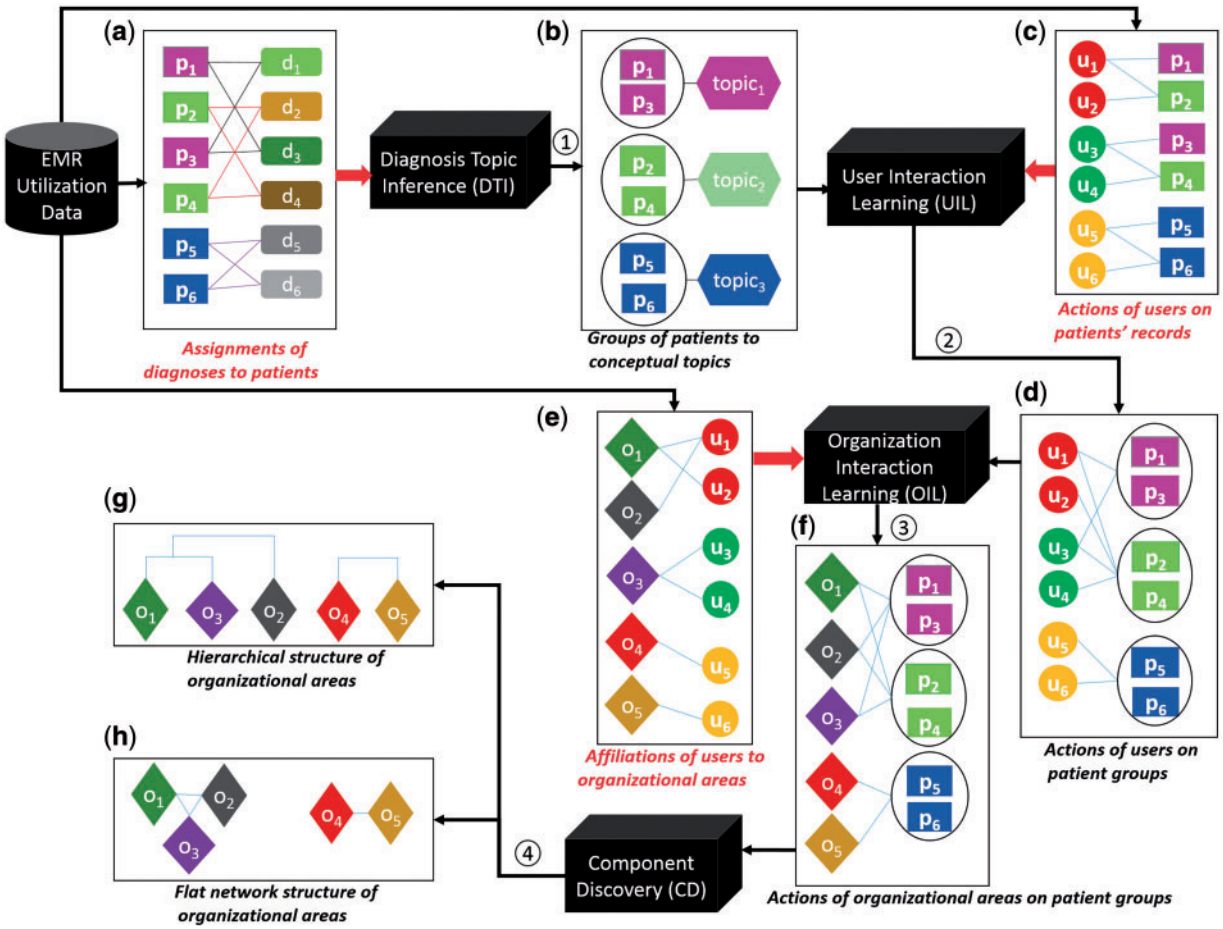

Figure 1. The process by which HCO components are learned through EHR system utilization. u_i = EMR user; p_i = EMR of a patient; d_i = diagnosis code assigned to an EMR; o_i = operational area affiliated with a user; topic_i = concept that represents a latent diagnostic pattern.

$$D_{\text{user}\times\text{topic}} = B'_{\text{patient}\times\text{user}} \times L_{\text{patient}\times\text{topic}} \quad (1)$$

Management of patients by users is shown in Figure 1c, with the formal $B'_{\text{patient}\times\text{user}}$ in Figure 2c. The derived interactions of users by patient groups are shown in Figure 1d, with the formal $D_{\text{user}\times\text{topic}}$ in Figure 2e.

Transform 3 (Focus on interactions of users into interactions of their affiliated operational areas): The relations of operational areas are more stable and consistent in comparison to the interaction relationships of users.[16,30] Based on this observation, we transformed the interactions of users by patient groups (Figure 1d) into interactions of operational areas by patient groups (Figure 1f) through Equation 2 (Organization Interaction Learning Module in Figure 1).

$$E_{\text{operation}\times\text{topic}} = C'_{\text{user}\times\text{operation}} \times D_{\text{user}\times\text{topic}} \quad (2)$$

The affiliations of users to their operational areas are shown in Figure 1e, with the formal $C'_{\text{user}\times\text{operation}}$ in Figure 2f. The formal area by patient group (characterized by topics) $E_{\text{operation}\times\text{topic}}$ is shown in Figure 2g.

Transform 4 (Focus on interactions of operational areas into organizational components): To infer an organizational component (which we represent as a network of related operational areas), we needed to measure the collaborations among operational areas. To do so, we invoked a cosine similarity measure (we selected cosine measure because evidence suggests it is effective for comparing such diagnosis topics[24]) to learn the strength of the relation for a pair of operational areas as:

$$R_{\text{operation}\times\text{operation}}(i,j) = \frac{E(i)\cdot E(j)}{|E(i)|\cdot|E(j)|} \quad (3)$$

where $E(i)$, $E(j)$ are row vectors of E. The formal representation of collaborations among operational areas is shown in Figure 2h. Organizational components are learned according to a bottom-up nearest neighbor clustering algorithm over the collaborations of operational areas.[37] We relied upon this approach because the method can discover a set of nearest neighbors for each operational area and subsequently group these neighbors into a hierarchical structure. The transformation (Component Discovery Module in Figure 1) hierarchically clusters the nearest neighboring operational areas into components. The relations between components and operational areas are represented by a binary matrix $H_{\text{component}\times\text{operation}}$. In this matrix, a cell value of 1 indicates that an operational area belongs to a component, while 0 indicates otherwise.

To understand the process of hierarchical clustering, we list an example to show how the hierarchical network as shown in Figure 1g is constructed. The cosine similarity between operational areas O_1 and $O_3$is 1 (Figure 2h), which is the largest similarity (or smallest distance) in this example. As such, O_1 and O_3 are the first

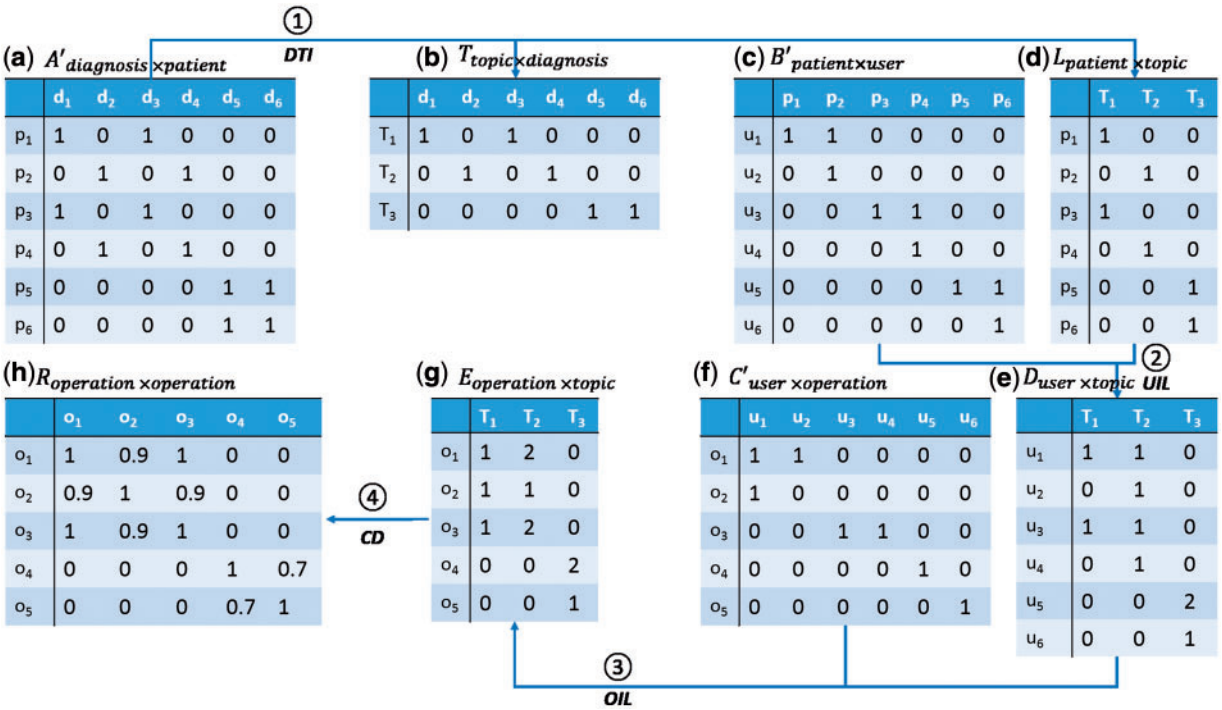

Figure 2. An example of the inference of collaborative networks from EMR utilization data.

organizational areas to be connected. We further find the similarity between O_2 and O_2 and O_3 is 0.9, so O_2 is the next to be connected. Finally O_4 and O_5 are connected with a similarity of 0.7.

Our clustering algorithm terminates the clustering process at a distance threshold where the merging of large clusters becomes frequent. This avoids clustering all the operational areas into several big clusters. The justification for determination of the number of components in this work can be found online in Supplement S1.

Hypothesis tests for component plausibility

For an extrinsic evaluation, we investigated whether clinical and administrative experts could distinguish an inferred organizational component from a randomly generated component (in terms of capability for collaborative patient management). To do so, we designed a survey that consisted of ⟨inferred, random⟩ pairs of operational areas in the HCO, which we then asked the clinical and administrative experts to review for plausibility.

Survey questions. We recruited experts to answer questions of the following form: "*To what extent do you believe Vanderbilt University Medical Center (VUMC) employees in the displayed group of operational areas collaborate to manage patients?*" For each question, we provided 5 candidate answers (*Not at all likely*, *Slightly likely*, *Moderately likely*, *Very likely*, and *Completely likely*). To perform hypothesis testing (see below), we converted these answers into integer values (Likert score) in the range 1–5 (eg, *Not at all likely* is mapped to 1). Details of the survey design and questions are in Table S36 of Supplement S3.

Pretest. The survey was pretested in the Research Electronic Data Capture (REDCap) management system[38] with 1 physician, 1 nurse, and 1 hospital administrator who were not affiliated with the research team. It was found that the survey could be completed in <20 minutes, which was deemed to be an acceptable amount of time by the respondents. Feedback from the participants further indicated that we should alphabetize the operational areas of each component for more convenient viewing and reference during the survey.

Survey administration. Next, we invited 26 participants who were knowledgeable professionals with a diverse array of expertise (eg, physicians, nurses, and administrators). Each potential survey respondent was emailed an introduction to the goals of the survey and a link to the online REDCap survey.

Analysis. To assess whether the experts found the learned organizational components to be plausible, we conducted a series of hypothesis tests, each of which can be summarized as: "*For a given pair of ⟨inferred, random⟩ components, experts can distinguish the inferred from the random component.*" We applied a linear regression model as shown in Equation 5 to determine the Likert score for a pair of inferred and random components.

$$\text{Likert Score} = \alpha + \theta \times \beta \tag{4}$$

where $\theta 1(\text{inferred}), 0(\text{random})\}$ represent the inferred and random components, respectively. Under this model, the Likert score for the random component is α ($\theta = 0$) and for the inferred component is $\alpha + \beta$ ($\theta = 1$). As such, the value of β corresponds to the difference of Likert scores for inferred and random components.

Hypothesis test. We used the Likert scores as observations to infer β via linear regression models. We then used an analysis of variance[39] to test the significance of $\beta \neq 0$ against a null hypothesis $\beta = 0$. We tested the hypothesis at the 2-sided $\alpha = 0.05$ significance level. For each test, we used a power of 0.9 to calculate number of respondents needed to confirm the test.

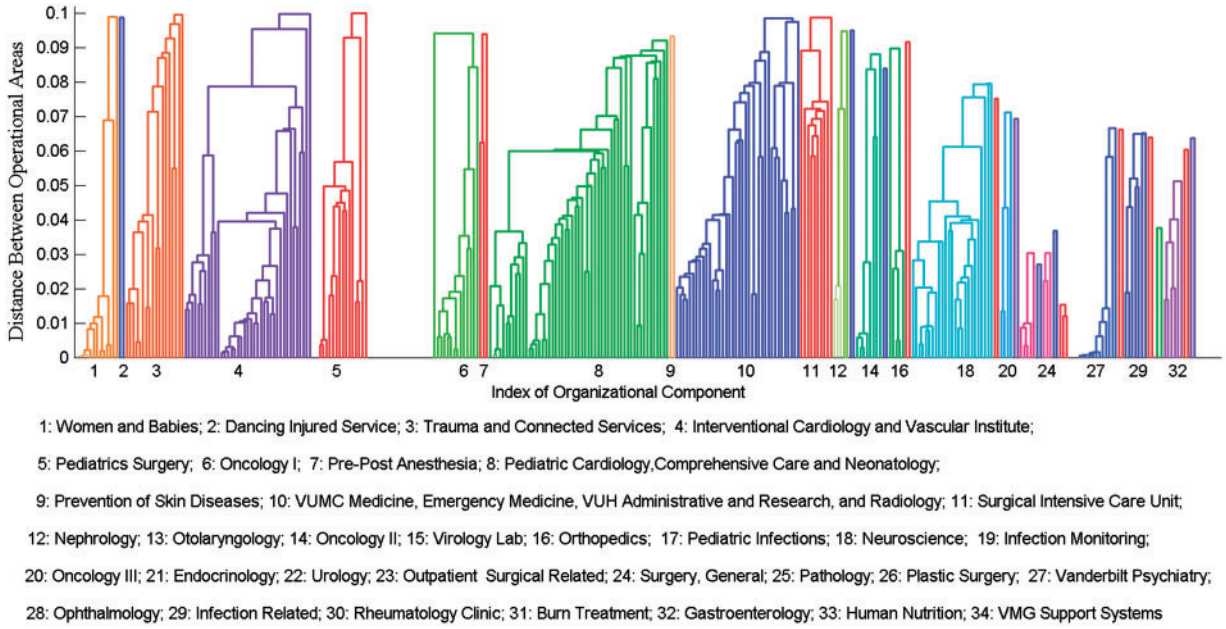

Figure 3. The organizational components learned from 4 months of inpatient EMR utilization. Note that the smaller the distance between 2 operational areas is, the stronger the collaboration between the affiliated employees. The empty gaps between components are due to cutting the dendrogram above a value of 0.1. They correspond to the inducing of independent operational areas. The composition of each component, in terms of its operational areas in the HCO, can be found online in Supplement S1.

Strength of collaboration within a component

It should be recognized that an inferred organizational component only represents a hierarchical structure of operational areas; it does not prioritize the strength of collaboration to yield functional collaborative care. We anticipate that the more tight-knit the collaboration within a component is, the greater the opportunity for establishment of collaborative care. Thus, we set out to conduct an intrinsic evaluation to investigate the strength of the collaboration between all operational areas in an organizational component.

To do so, we composed a network of the operational areas within a component (as shown in Figure 1h), where the relations of operational areas in $R_{\text{operation}\times\text{operation}}$ are represented by corresponding edges. To systematically investigate the strength of the collaboration within a network, we measured its corresponding cluster coefficient.[40] This measure (which ranges from 0 to 1) is positively correlated with the strength of the collaborations in a component.

Associations between components and diagnoses

To provide clinically relevant cues for HCOs to know if the learned organizational components are for the right patient groups, we conducted an intrinsic evaluation of clinical credibility via association mining between the organizational components and diagnostic conditions, in the form of PheWAS codes. The associations between diagnoses and components are measured as:

$$F_{\text{component}\times\text{disgnosis}} = H_{\text{component}\times\text{operation}} \times E_{\text{operation}\times\text{topic}} \times T_{\text{topic}\times\text{diagnosis}} \quad (5)$$

Specifically, a PheWAS code was associated with an organizational component if its probability to that component (Equation 6) was >0.3. This threshold is based on the observation that, for a majority of the learned components, a predicted probability around 0.3 leads to a clear separation between the codes.

RESULTS

The results are organized around (1) the discovered components, whose face validities were confirmed by the clinical and administrative experts, (2) the cluster coefficients for the components, and (3) the associations between components and patient comorbidities. We close this section with an illustrative example of an organizational component associated with oncology management.

Organizational components

As shown in Figure 3, the pipeline discovered 34 organizational components for the indicated VUMC inpatient setting. In aggregate, the components covered 317 of 486 (65%) of the HCO operational areas whose affiliated employees accessed EMRs during the study period, which suggests that the health care system is highly collaborative. It is not surprising that certain operational areas remain less integrated, because not all HCO areas are expected to function in a collaborative manner due to highly specialized, rare services. Since this investigation focuses on networks of interactions, we removed the independently functioning operational areas from further consideration.

For convenience, we refer to the ith component as C_i. It can be seen in Figure 3 that the VUMC inpatient setting decomposes into a set of functional collaborating networks (eg, C_1, Women and Babies; C_3,Trauma and Connected Services; and C_{29}, Infections). The HCO areas for each component can be found online in Supplement S2.

Component plausibility

The survey was completed by 23 of the 26 invited experts (88.5%). Demographics of the participants who completed the survey are in Table S35 of Supplement S3. Table 1 reports the average difference between the Likert scores for each learned organizational component and its randomized counterpart. When the difference $\beta > 0$, it indicates that the learned organizational component scored higher

Table 1. Survey results from HCO clinical and administrative experts ($n = 23$) regarding the plausibility of organizational components based on EMR utilization **No.**

	Organizational Component	Likert Score Difference	Respondents Required for Statistical Significance	*P*-value
Confirmed to be statistically significant on a t-test at the 2-sided $\alpha = 0.05$ significance level				
C_1	Women and Babies	1.174	11	7.9×10^{-5}
C_2	Dancing Injured Service	1.000	11	6.2×10^{-5}
C_3	Trauma and Connected Service	1.000	18	1.7×10^{-3}
C_4	Interventional Cardiology and Vascular Institute	1.348	11	6.8×10^{-5}
C_5	Pediatric Surgery	1.652	6	2.6×10^{-8}
C_6	Oncology I	1.652	5	5.1×10^{-9}
C_7	Pre-Post Anesthesia	2.870	3	1.0×10^{-13}
C_{12}	Nephrology	1.348	9	1.3×10^{-5}
C_{13}	Otolaryngology	0.695	22	9.5×10^{-3}
C_{14}	Oncology II	1.652	7	9.5×10^{-7}
C_{16}	Orthopedics	1.826	6	3.8×10^{-8}
C_{17}	Pediatric Infections	1.783	7	1.0×10^{-6}
C_{18}	Neuroscience	1.565	8	5.3×10^{-6}
C_{19}	Infection Monitoring	1.913	8	3.4×10^{-6}
C_{20}	Oncology III	1.565	10	4.0×10^{-5}
C_{22}	Urology	1.217	14	4.2×10^{-4}
C_{23}	Outpatient Surgical Related	1.174	23	6.5×10^{-3}
C_{24}	Surgery, General	1.261	11	8.4×10^{-5}
C_{25}	Pathology	1.565	9	1.7×10^{-5}
C_{26}	Plastic Surgery	2.609	4	3.7×10^{-12}
C_{27}	Vanderbilt Psychiatry	1.696	6	1.3×10^{-7}
C_{28}	Ophthalmology	2.913	3	1.6×10^{-13}
C_{29}	Infection Related	1.261	13	4.4×10^{-4}
C_{30}	Rheumatology Clinic	1.261	19	8.2×10^{-5}
C_{31}	Burn Treatment	2.565	3	1.1×10^{-14}
C_{32}	Gastroenterology	1.174	17	1.5×10^{-3}
C_{34}	Vanderbilt Medical Group Support Systems	0.913	23	7.4×10^{-3}
Not confirmed to be statistically significant				
C_8	Pediatric Cardiology, Comprehensive Care, and Neonatology	0.478	68	1.1×10^{-1}
C_9	Prevention of Skin Diseases	0.260	320	4.4×10^{-1}
C_{10}	VUMC Medicine and Emergency Medicine, Vanderbilt University Hospital Administration and Research, and Radiology	0.391	112	2.1×10^{-1}
C_{11}	Surgical Intensive Care Unit	0.608	40	3.7×10^{-2}
C_{15}	Virology Lab	0.304	158	2.8×10^{-1}
C_{21}	Endocrinology	0.478	49	5.6×10^{-2}
C_{33}	Human Nutrition	0.782	51	6.2×10^{-2}

Each row shows the distance between the Likert score of the inferred organizational component and its randomized counterpart. Note that a positive distance indicates the inferred component received a higher Likert score.

(and is more plausible) than the random entry. The clinical and administrative experts always scored the organizational component as more plausible. Moreover, the Likert scores for 27 of the 34 inferred components (79%) were statistically significantly higher than the randomized component (2-sided $\alpha = 0.05$ confidence level).

Cluster coefficients

Figure 4 depicts the cluster coefficient for each component. Let us consider C_8, Pediatric Cardiology, Pediatrics Comprehensive Care, and Neonatology, which exhibits the smallest cluster coefficient ($\sim$0.12, in the lower-right section of Figure 4). Figure 4 shows values representative of the collaboration strength for each component. Figure 5 illustrates the interactions among the operational areas within each organizational component. As depicted, the network for C_8 (in the lower-left section of the figure) is large (with 55 connected operational areas) and has a relatively low overall density with 3 subnetworks. To build collaborative care for this network, administrative leaders would need to consider the interactions of the operational areas within the entire component, as well as the interactions among the subnetworks. By contrast, C_1, Women and Babies (upper-right section of figure), exhibits a single dense set of interactions among its members and thus has a large cluster coefficient (> 0.8 in Figure 4). This indicates that administrative leaders would only need to consider the interactions of operational areas within this single unified component.

Organizational component and comorbidity associations

It was found that 119 PheWAS codes were associated with the 34 organizational components (Figure S23 of Supplement S4). Each component was associated with approximately 5–10 PheWAS codes. To illustrate such relationships, let us consider the PheWAS codes associated with C_1, Women and Babies. It is evident that this component is responsible for complications associated with childbirth. Specifically, the associated PheWAS codes include abnormality in fetal heart rate or rhythm, abnormality of pelvic soft tissues and organs complicating pregnancy, late pregnancy and failed

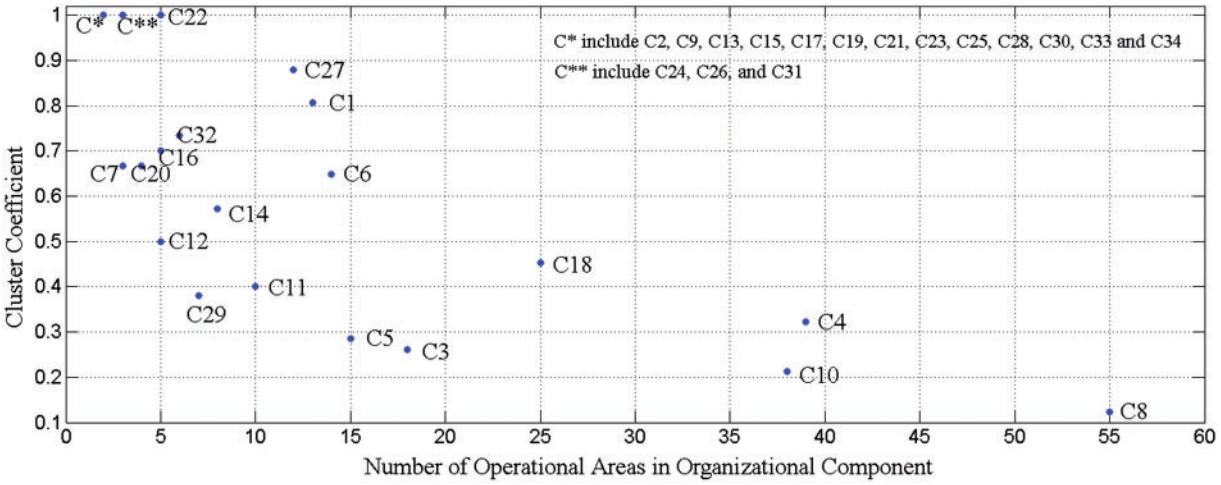

Figure 4. The strength of collaboration among the operational areas of a component as a function of its size. C* and C** correspond to the sets of components with 2 and 3 operational areas, respectively.

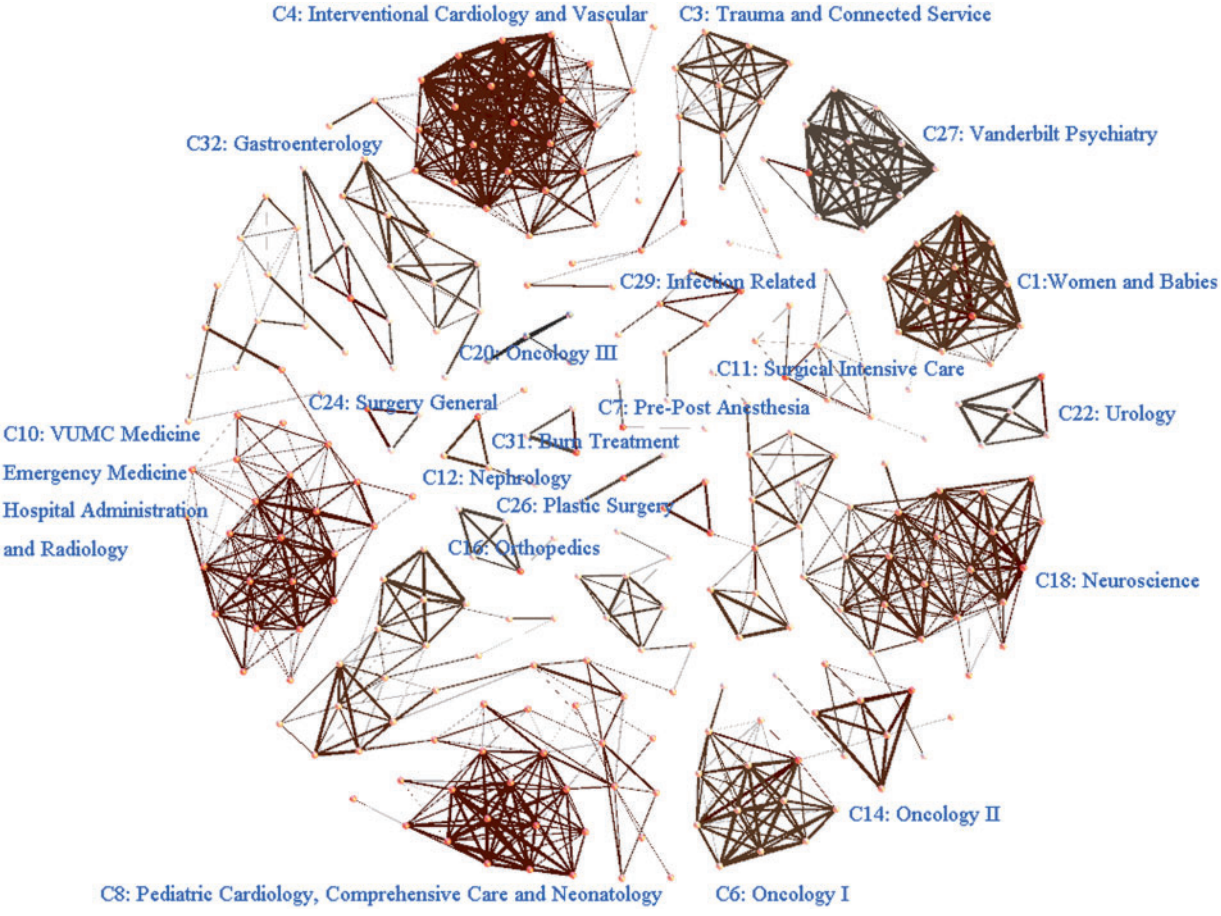

Figure 5. A network view of the organizational components inferred from EMR utilization records. A node corresponds to an operational area, and an edge is the interaction relation between 2 operational areas.

induction, morbid obesity, obstetrical/birth trauma, other conditions of the mother complicating pregnancy, and problems associated with amniotic cavity and membranes. Supplement S2 online provides the associations for each component.

An illustrative example of evaluation of organizational components

To gain further understanding of the plausibility evaluation of the inferred organizational components, we worked with several oncologists to interpret the relationship between the operational areas in

C_6, Oncology I. This component is notable because the operational areas exhibit a large clustering coefficient (~0.65), and thus it was anticipated to be highly collaborative.From a clinical perspective, this component was associated with patients diagnosed with various hematologic cancers, such as acute myelogenous leukemia, myelodysplastic syndrome, and multiple myeloma (PheWAS codes as shown in Table 2).

As depicted in Figure 6, the operational areas Hematology/Stem Cell Clinic (index 7) and Myelosuppression (index 8) were the first to be linked by our pipeline. This is likely because the care providers in the bone marrow transplantation (BMT) unit located in the outpatient and inpatient settings access the same patient charts. This is intuitive from an HCO management perspective, because there is a close collaboration between the outpatient transplant unit and the inpatient marrow suppression unit. Patients often move from one to the other over the course of the 100-day acute period of a BMT.

Table 2. Top 20 PheWAS codes associated with organizational component C_6, Oncology I.

Code	Description	Predicted probability
585.1	Acute renal failure	1.000
197	Chemotherapy	0.769
204.21	Myeloid leukemia, acute	0.688
288.11	Neutropenia	0.534
509.1	Respiratory failure	0.506
284	Aplastic anemia	0.496
480	Pneumonia	0.489
401.1	Essential hypertension	0.478
198.2	Secondary malignancy of lung	0.463
198	Secondary malignant neoplasm	0.443
202.2	Non-Hodgkins lymphoma	0.440
081	Infection/inflammation of internal prosthetic device, implant, or graft	0.430
198.6	Secondary malignancy of bone	0.409
287.3	Thrombocytopenia	0.390
284.1	Pancytopenia	0.358
783	Fever of unknown origin	0.357
198.4	Secondary malignant neoplasm of liver	0.355
198.1	Secondary malignancy of lymph nodes	0.348
204.4	Multiple myeloma	0.343
198.5	Secondary malignancy of brain/spine	0.319

The predicted probability is normalized into a range of [0,1] by using min–max normalization.

The operational areas Radiology Oncology Housestaff (index 4), Radiation Oncology Housestaff (index 5), and Radiation Oncology (index 6) are the next 3 areas linked The integration of these 3 operational areas serves as an indirect confirmation of the power of the data in EMR utilization logs. More specifically, it is clear that operational areas 4 and 5 refer to the same clinical concept. However, the problem of multiple aliases for the same concept is common in legacy systems, and particularly in the case of VUMC, where employees are permitted to specify their affiliations. These areas are likely members of the organizational component, because patients receiving bone marrow transplant are often treated on a daily basis for 2–4 weeks (depending on the type of transplant) by radiation oncologists.

The operational areas Bone Marrow Processing Lab (index 2) and Bone Marrow Registry (index 3) are the next 2 areas linked. These operational areas are related because of the data requirements for Center for International Blood and Marrow Transplant Research registry reporting. BMT is a highly regulated procedure with substantial data collected to populate national registries. Data elements such as donor characteristics, cell dose, processing viability, and antigen assays are all extracted from the lab and recorded.

Finally, once patients are discharged, they often return to the clinic for routine checkups. This is an activity that is associated with the operational areas Hematology/Stem Cell Clinic (index 7), Outpatient Clinical Pharmacy (index 10), Cancer Call Center (index 11), and Hematology/Oncology (index 12).

DISCUSSION

This study provides evidence that the utilization records in EMR systems can be translated into knowledge that is relevant for the

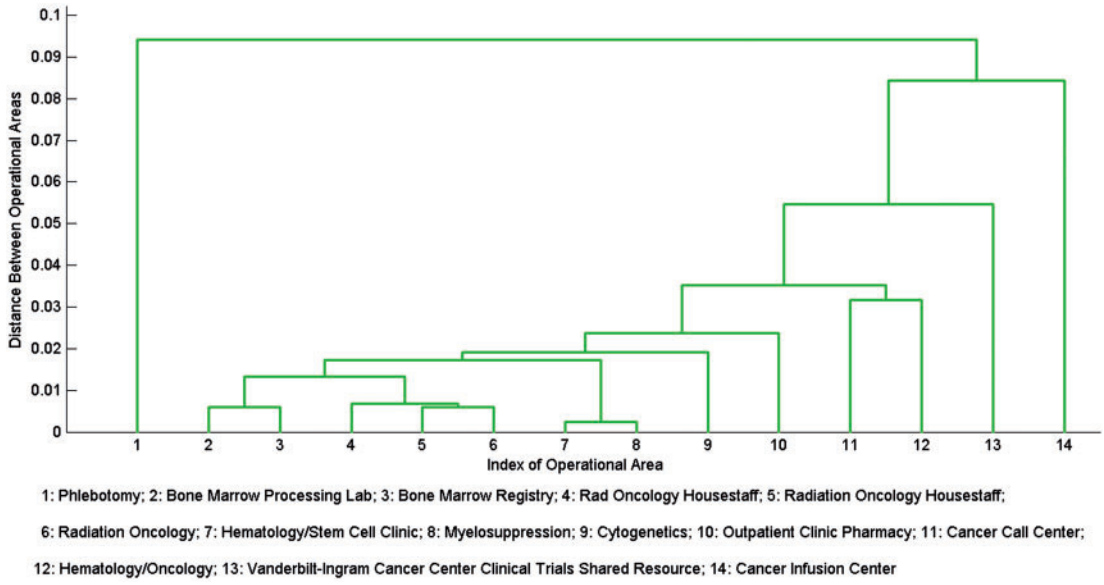

Figure 6. The hierarchical structure of the 14 operational areas that comprise C_6, Oncology I.

definition of collaborative networks. Moreover, this knowledge can be contextualized with collaboration strength and clinical concepts that associate with such networks in a meaningful manner. While this investigation indicates that data-driven methods can provide insight into HCO management, there are several limitations that should be recognized, which can serve as guidance for future investigations.

One of the more notable limitations is that clinical and administrative experts did not find certain organizational components to be statistically different from random elements in terms of plausibility. We believe there were several notable reasons for a lack of confirmation. First, the number of observations (ie, 23 pairs of inferred organizational components and randomized groups of operational areas) was underpowered in certain situations. As shown in Table 1, a greater number of experts would be needed to determine if the difference was statistically significant. Yet achieving such a number may be challenging, because the power analysis indicates that over 500 experts are necessary for C_9, Prevention of Skin Diseases. Also, certain employee actions may not be documented in the EMR system. Moreover, the rate at which interactions are undocumented could be higher for unconfirmed organizational components. This is a particularly plausible scenario for C_{15}, Virology Lab.

Second, this work leverages previous studies[24,30,35] finding that collaborative networks inferred at the patient group level are more stable and meaningful than those at the individual patient level, but does not quantify the differences between these 2 types of networks. Further studies need to be done to capture the differences in terms of the structure and operational relations of these 2 types of networks.

Third, this study only focused on the collaboration of HCO employees within an organizational component, and not their coordinated behavior. We note that it is challenging to infer coordinated relations through EMR systems due to asynchronous and replicated documentation of employees' actions. Our study aimed to reduce the influence of such problems through inference-based data mining (eg, grouping of patients based on common PheWAS codes). However, as data-driven HCO modeling progresses, it will be necessary to model coordination in a systematic and automated manner.

Fourth, this investigation only focused on a 4-month period in 2010. These data are sufficient to make our claim that EMR utilization data align with the expectations of the employees of a medical center. However, to make specific teaming recommendations, the volume of data and time period should be enlarged to evaluate the relevance and stability of care teams over time.

Finally, the data studied were selected from a single HCO. It is necessary to confirm that analogous new (and traditional) organizational components can be identified through utilization of EMRs at other institutions, or that failing to identify matching components can be explained in terms of observable organizational differences between HCOs. This is a nontrivial challenge, because it will require modeling components on a wide range of diseases across the entire EMR system, as well as recruiting knowledgeable clinical and administrative experts. Nonetheless, such replication is critical to ensure reproducibility and applicability in practice.

CONCLUSIONS

This paper introduces a novel data-driven framework based on utilization of an HCO's EMR system to discover collaborative organizational components. This was done by applying the framework to 4 months' worth of EMR utilization logs at VUMC. We validated the plausibility of the majority of the components with 23 clinical and administrative experts and further showed collaboration strength and correlated patient conditions of each component. We believe that such a data-driven method can enable HCOs to establish, refine, and manage collaborative care across large complex health care systems.

FUNDING

This research was supported, in part, by the National Institutes of Health under grants R00LM011933 and R01LM010685.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

YC performed the data collection and analysis, methods design, survey design and survey data collection, hypothesis design, experiment design, evaluation and interpretation of experiments, and writing of the manuscript. NL performed survey design and collection of the survey data, interpretation of the organizational components, and writing of the manuscript. WS performed evaluations of inferred components, interpretation of the organizational components, and writing of the manuscript. KW performed interpretation of the organizational components and writing of the manuscript. BM performed data collection and analysis, survey and hypothesis design, evaluation and interpretation of experiments, and writing of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We would like to thank Dr Mark Frisse for providing constructive directions. We further thank Dr Matthew Rioth and Dr Travis Osterman for detailed interpretation of the inferred oncology organizational component. The datasets used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU, which is supported by institutional funding and by Vanderbilt Clinical and Translational Science Awards grant ULTR000445 from the National Center for Advancing Translational Sciences and the National Institutes of Health.

REFERENCES

1. Peterson MW. Emerging developments in postsecondary organization theory and research: fragmentation or integration. *Educ Res*. 1985;14(3):5–12.
2. Stange KC. The problem of fragmentation and the need for integrative solutions. *Ann Fam Med*. 2009;7(2):100–103.
3. Richardson WC, Briere R, eds. Crossing the quality chasm: a new health system for the 21st century. Committee on Quality of Health Care in America, Institute of Medicine. Washington, DC: National Academy Press; 2001.
4. Fisher ES. Building a medical neighborhood for the medical home. *N Engl J Med*. 2008;359:1202–1205.
5. Asarnow JR, Rozenman M, Wiblin J, Zeltzer L. Integrated medical-behavioral care compared with usual primary care for child and

adolescent behavioral health: a meta-analysis. *JAMA Pediatr.* 2015;169(10):929–937.
6. Cebul RD, Rebitzer JB, Taylor LJ, Votruba ME. Organizational fragmentation and care quality in the U.S. healthcare system. *J Eco Pers.* 2008; 22:93–113.
7. Grne O, Garcia M. Integrated care: a position paper of the WHO European office for integrated health care services. *Int JIntegr Care.* 2001;1(e21):1–10.
8. Reid RJ, Coleman K, Johnson EA, *et al*. The Group Health Medical Home at year two: cost savings, higher patient satisfaction, and less burnout for providers. *Health Aff.* 2010;29:835–843.
9. Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. Defining comorbidity: implications for understanding health and health services. *Ann Fam Med.* 2009;7:357–363.
10. Rittenhouse DR, Shortell SM. The patient-centered medical home: will it stand the test of health reform? *JAMA.* 2009;301(19):2038–2040.
11. Bergeson SC, Dean JD. A systems approach to patient-centered care. *JAMA.* 2006;296(23):2848–2851.
12. Katon WJ, Lin EHB, Korff MV, *et al*. Collaborative care for patients with depression and chronic illnesses. *N Engl J Med.* 2010;363:2611–2620.
13. Grumbach K, Bodenheimer T. Can health care teams improve primary care practice? *JAMA.* 2004;291:1246–1251.
14. Wagner EH. The role of patient care teams in chronic disease management. *BMJ.* 2000;320:569–572.
15. Berwick D. What patient-centered should mean: confessions of an extremist. *Health Aff.* 2009;28(4):w555–w565.
16. Chen Y, Nyemba S, Malin B. Auditing medical records accesses via healthcare interaction networks. *AMIA Annu Symp.* 2012;2012:93–102.
17. Carley K, Lee J. Dynamic organizations: organizational adaptation in a changing environment. *Adv Stra Manage.* 1998;15:269–297.
18. Carley K. Computational organization science: a new frontier. *PNAS.* 2002;99(3):7257–7262
19. Merrill J, Bakken S, Rockoff M, Gebbie K, Carley K. Description of a method to support public health information management: organizational network analysis. *J Biomed Inform.* 2007;40:422–428.
20. Uddin S, Khan A, Piraveenan M. Administrative claim data to learn about effective healthcare collaboration and coordination through social network. *Proc Int Conf Sys Sci.* 2015;3105–3114.
21. Cunningham FC, Ranmuthugala G, Plumb J, Georgiou A, Westbrook J, Braithwaite J. Health professional networks as a vector for improving healthcare quality and safety: a systematic review. *BMJ Qual Saf.* 2012;21:239–249.
22. Uddin S, Hossain L, Hamra J, Alam A. A study of physician collaborations through social network and exponential random graph. *BMC Health Serv Res.* 2013;13(234):1–14.
23. Merill JA, Sheehan BM, Carley KM, Stetson PD. Transition networks in a cohort of patients with congestive heart failure. *Appl Clin Inform.* 2015;6:548–564.
24. Chen Y, Ghosh J, Bejan CA, *et al*. Building bridges across electronic health record systems through inferred phenotypic topics. *J Biomed Inform.* 2015;55:82–93.
25. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med.* 2010;2(57):57cm29.
26. Kho AN, Pacheco JA, Peissig PL, *et al*. Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. *Sci Transl Med.* 2011;3(79):79re1.
27. Wu S, Chaudhry B, Wang J, *et al*. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Ann Int Med.* 2006;144(10):742–752.
28. Stead WW, Lin HS. National Research Council Committee on Engaging the Computer Science Research Community in Health Care Informatics, Computational Technology for Effective Health Care: immediate steps and strategic directions. *NAP.* 2009.
29. Unertl KM, Johnson KB, Lorenzi NM. Health information exchange technology on the front lines of healthcare: workflow factors and patterns of use. *J Am Med Inform Assoc.* 2012;19:392–400.
30. Chen Y, Lorenzi N, Nyemba S, Schildcrout JS, Malin B. We work with them? Health workers' interpretation of organizational relations mined from electronic health records. *Int J Med Inform.* 2014;83:495–506.
31. Peissig P, Costa VS, Caldwell MD, *et al*. Relational machine learning for electronic health record driven phenotyping. *J Biomed Inform.* 2014;52:260–270.
32. Tanpowpong P, Broder-Fingert S, Obuch JC, Rahni DO, Katz AJ, Leffler DA. Multicenter study on the value of ICD-9-CM codes for case identification of celiac disease. *Ann Epidemiol.* 2013;203:136–142.
33. Denny JC, Ritchie MD, Basford MA, *et al*. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics.* 2010;26:1205–1210.
34. Denny JC, Bastarache L, Ritchie MD, *et al*. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol.* 2013;31:1102–1111.
35. Chen Y, Nyemba S, Malin B. Detecting anomalous insiders in collaborative information systems. *IEEE Trans Dependable Secure Comput.* 2012;9: 332–344.
36. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. *J Biomed Inform.* 2015;58:156–165
37. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mac Learn Res.* 2003;3:993–1022.
38. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap): a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009;42:377–381.
39. David C, Hoaglinab E. The hat matrix in regression and ANOVA. *The Am Stat.* 1978;32(1):17–2240.
40. Watts DJ, Strogatz S. Collective dynamics of small-world networks. *Nature.* 1998;393:440–442.