Journal of Biomedical Informatics 48 (2014) 84–93

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Decision support from local data: Creating adaptive order menus from past clinician behavior

Jeffrey G. Klann [a,b,e,*,1], Peter Szolovits [c], Stephen M. Downs [d,e], Gunther Schadow [e,2]

[a] Laboratory of Computer Science, Massachusetts General Hospital, One Constitution Center, Suite 200, Boston, MA 02129, United States
[b] Harvard Medical School, 25 Shattuck St, Boston, MA 02115, United States
[c] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Stata Center, 32 Vassar St, 32-254, Cambridge, MA 02139, United States
[d] Children's Health Services Research, Indiana University School of Medicine, 410 W. 10th St, Suite 1000, Indianapolis, IN 46202, United States
[e] The Regenstrief Institute for Health Care, 410 W. 10th St, Suite 2000, Indianapolis, IN 46202, United States

ARTICLE INFO

Article history:
Received 7 October 2013
Accepted 7 December 2013
Available online 16 December 2013

Keywords:
Clinical Decision Support
Data mining
Bayesian analysis

ABSTRACT

Objective: Reducing care variability through guidelines has significantly benefited patients. Nonetheless, guideline-based Clinical Decision Support (CDS) systems are not widely implemented or used, are frequently out-of-date, and cannot address complex care for which guidelines do not exist. Here, we develop and evaluate a complementary approach – using Bayesian Network (BN) learning to generate adaptive, context-specific treatment menus based on local order-entry data. These menus can be used as a draft for expert review, in order to minimize development time for local decision support content. This is in keeping with the vision outlined in the US Health Information Technology Strategic Plan, which describes a healthcare system that learns from itself.

Materials and methods: We used the Greedy Equivalence Search algorithm to learn four 50-node domain-specific BNs from 11,344 encounters: abdominal pain in the emergency department, inpatient pregnancy, hypertension in the Urgent Visit Clinic, and altered mental state in the intensive care unit. We developed a system to produce situation-specific, rank-ordered treatment menus from these networks. We evaluated this system with a hospital-simulation methodology and computed Area Under the Receiver–Operator Curve (AUC) and average menu position at time of selection. We also compared this system with a similar association-rule-mining approach.

Results: A short order menu on average contained the next order (weighted average length 3.91–5.83 items). Overall predictive ability was good: average AUC above 0.9 for 25% of order types and overall average AUC .714–.844 (depending on domain). However, AUC had high variance (.50–.99). Higher AUC correlated with tighter clusters and more connections in the graphs, indicating importance of appropriate contextual data. Comparison with an Association Rule Mining approach showed similar performance for only the most common orders with dramatic divergence as orders are less frequent.

Discussion and conclusion: This study demonstrates that local clinical knowledge can be extracted from treatment data for decision support. This approach is appealing because: it reflects local standards; it uses data already being captured; and it produces human-readable treatment-diagnosis networks that could be curated by a human expert to reduce workload in developing localized CDS content. The BN methodology captured transitive associations and co-varying relationships, which existing approaches do not. It also performs better as orders become less frequent and require more context. This system is a step forward in harnessing local, empirical data to enhance decision support.

Abbreviations: BN, Bayesian Network; ARM, Association Rule Mining; CPT, Conditional Probability Table; GES, Greedy Equivalence Search; ITS, Iterative Treatment Suggestion (the methodology defined in this manuscript); UVC, Urgent Visit Clinic.

* Corresponding author at: Laboratory of Computer Science, Massachusetts General Hospital, One Constitution Center, Suite 200, Boston, MA 02129, United States. Tel.: +1 617 643 5879; fax: +1 617 643 5280.
E-mail address: Jeff.Klann@mgh.harvard.edu (J.G. Klann).

[1] Dr. Klann is no longer affiliated with 'The Regenstrief Institute for Health Care'.
[2] Present address: Pragmatic Data LLC, 8839 Rexford Rd., Indianapolis, IN 46260, United States.

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.12.005

1. Introduction

A currently popular approach to improving the quality of health care is to make sure that similar cases are handled in similar ways, i.e., to reduce the variability of care [1]. Frequently this is accomplished through propagation of external protocols into practice, through mechanisms such as Clinical Decision Support (CDS) [2].

Unfortunately, computable CDS content is extremely expensive and time-consuming to create [3], maintain [4], and localize [5]. Consequently CDS has been much more slowly adopted than other components of Health Information Technology (HIT) [6]. Even when CDS available, the content is frequently inappropriate or incorrect [7]. Various projects are being undertaken to standardize computable CDS content in order to reduce the local implementer's work (e.g., [8]).

Still, standardized CDS does not address the following issues: the frequency of content change in medicine, physician attitudes toward guidelines, and terminology challenges. First, much content, both routine and complex, is not distilled into guidelines [9]. This might be quite common; in one study, the literature provided answers to primary care providers' routine clinical questions only 56% of the time [10]. Second, studies have shown that physicians value colleagues' advice at least as much as guidelines [11]. This might be because medicine is locally situated, and colleagues can provide a local frame of reference through which to decide if and how external guidelines relate to particular local cases [12]. Third, standardized content databases require translation of codes into standard terminologies, which is difficult and frequently causes failures in interoperability.

Electronic Medical Record (EMR) data is rapidly proliferating [13], in part due to the Meaningful Use incentive program [14]. These data offer the opportunity to harness local physician wisdom – how care is actually delivered – to augment and suggest protocols, vastly decreasing human effort in developing CDS content and making knowledge available in complex scenarios. It is possible to partially reconstruct physician decisions by aggregating the millions of treatment events in medical record systems. Such locally generated CDS content avoids the three issues discussed above. This fits into the Office of the National Coordinator for HIT's strategic plan, which centers on building a "learning healthcare system" that can perform dynamic analysis of existing healthcare data to glean various information, including best practices [15].

1.1. The wisdom of the crowd

Despite the incompleteness of guidelines and poor maintenance of expert-curated CDS, individual physician behavior is not reliable either. Studies show that care continues to be widely variable and that physicians' treatment does not align well with guidelines [16]. Therefore we suggest two important goals in the design of a CDS tool based on local wisdom.

First, the *average* behavior of many physicians is usually much better than any *individual* physician. Condorcet's jury theorem, upon which voting theory is grounded, proves that when each member in a group of independent decision makers is more than 50% likely to make the correct decision, averaging those decisions ultimately leads to the right answer [17]. If we believe that a physician is more likely than chance to make the correct decision, we can trust the averaged decision. The theorem does have two important caveats. First, it is only guaranteed to apply to binary choices (plus an unlimited number of irrelevant alternatives) [18]. Thankfully, many high-level medical decisions are of this type (e.g., "do I anticoagulate this patient or not?"). Second, crowd wisdom can become crowd madness when decision-makers are not truly independent but are influenced by some outside entity [19]. And of course, practitioners are influenced by colleagues, formularies, available equipment, local culture, etc. The Dartmouth Atlas project has found that the quality of care in a region is profoundly influenced by the 'ecology' of healthcare in that region, including resources and capacity, social norms, and the payment environment [20].

This leads to our second design requirement. Even when averaging decisions, it is impossible to guarantee that results are not influenced by these caveats. Therefore we do not seek to *replace* manual content development with automatically generated CDS content. Instead, our goal is to *complement* content development with knowledge distilled from EMR data. To this end, it was important to choose a data mining approach which produces output that a human expert could understand and update before inserting it into a clinical system.

1.2. Mining EMR data

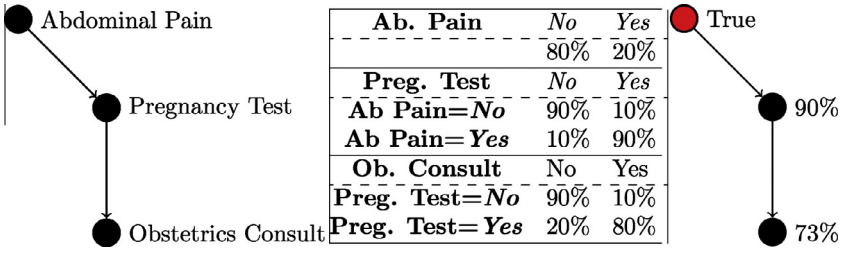
A handful of studies have explored methods to abstract treatment decisions captured in EMR data into knowledge bases [21–25] or to find knowledge on-demand [26]. The majority of work in abstracting EMR data have used variations of Amazon.com's pairwise Association Rule Mining (ARM) algorithm [27], which has shown good results when capturing global linkages where little variability exists (e.g., drugs used for HIV treatment) [28]. However, researchers have struggled with both transitive associations and the long, static lists of associations that do not take context into account. In one case, the results of such an approach required a great deal of manual editing before incorporation into a decision support system [29]. Other studies have used this approach only as a rudimentary starting point for content developers. For example, the condition-treatment linkages in the National Drug File Reference Terminology (NDF-RT) were 'jumpstarted' by this approach [24].

Bayesian Networks (BNs) are an appealing alternative for mining wisdom from EMR data. BNs are a powerful multivariate, probabilistic reasoning paradigm that naturally model interactions among associations. BNs have a two-phase lifecycle. First, they are constructed, either by hand – which has been widespread in medical informatics research (see e.g., [30]) – or more recently from databases of observational data [31]. Such 'structure learning algorithms', as they are called, take into account transitive associations and co-varying relationships that pairwise rule mining cannot. Therefore, BN structure learning might be able to make sense out of the tangled correlations in clinical data that have hampered other approaches. The second phase of the BN lifecycle is its use – rather than being static networks or rules, BNs enable rapid, iterative exploration of decisions as context evolves.

In a previous study, we piloted a BN approach to produce static order menus for complications of inpatient pregnancy [32]. Our results were very promising, but our scenarios were fixed, they only explored one small domain of medicine, and they relied on the opinion of a single nurse practitioner to evaluate our results. In this study, we more fully flesh out our previous work to use BNs to learn the typical successions of orders made by clinicians for a variety of types of cases. Next, we build a recommendation system that responds adaptively to suggest the most common next orders based on what has been ordered and diagnosed previously. Third, we evaluate this system on hospitalization order-entry data in a multitude of scenarios across four domains. Finally, we undertake a brief comparison of this dynamic approach to a static ARM-like approach.

1.3. Objective

Our goal was to develop a methodology to produce adaptive, patient-tailored, situation-specific treatment advice from order-entry



Ab. Pain	*No*	*Yes*
	80%	20%
Preg. Test	*No*	*Yes*
Ab Pain=*No*	90%	10%
Ab Pain=*Yes*	10%	90%
Ob. Consult	No	Yes
Preg. Test=*No*	90%	10%
Preg. Test=*Yes*	20%	80%

Fig. 1. An example Bayesian Network (left), the Conditional Probability Tables associated with it (middle), and the posterior probabilities given the evidence of 'Abdominal Pain' (right).

data, which can be used as a draft for expert review, in order to minimize development time for local decision support content. We used Bayesian Networks because of their adaptive nature and their ability to account for transitive associations and co-varying relationships. Also, they are human-readable and could therefore be curated by a human expert. We built and evaluated a recommendation system that dynamically suggests the most common next order based on what has been ordered previously. We also compared it to a static ARM-like approach.

2. Material and methods

2.1. Bayesian Networks and induction from data

A BN is a directed graph of vertices (nodes) and edges connecting those vertices. Embedded in each node is a Conditional Probability Table (CPT), which specifies the probability of each node state given the state of each parent. In this work, we induce BNs that represent the probabilistic relationships among orders and diagnoses. Then, as specific orders are placed and diagnoses made in a specific case, we instantiate the variables corresponding to those actions in the network (known as *evidence*), which revises the probabilities for other orders in the BN to the posterior probability that they would be placed conditioned on the previous actions. This allows us to rank remaining orders by their probability of occurring. In our interface, we present these ranked order menus to the user as orders are placed, in descending order of probability. We do not present diagnoses on the order menus, because the goal is to suggest treatments, leaving diagnosis to clinicians. An example of a simple BN, the underlying probabilistic relationships, and the revised posterior probabilities given evidence is shown in Fig. 1. The methodology, Iterative Treatment Suggestion (ITS), is summarized in Table 1. We implemented this methodology in Java using the SMILE toolkit [33], a freely available toolkit for network inference. A prototype of this interface can be seen in Fig. 2.

Table 1
A formal description of the ITS methodology for suggesting orders via a Bayesian Network. This parallels the graphical example in Fig. 2.

Algorithm: Iterative Treatment Suggestion (ITS)

Where:
G is a Bayesian Network Model
O is a set of possible orders, initially including all orders in G
D is a set of possible diagnoses, including all diagnoses in G
E is a set of evidence, initially containing all D set to *false*

Do:
1. Update beliefs (compute the posterior probability of all $O \notin E$)
2. Create a list of all $O \notin E$ in descending order of posterior probability, optionally stopping at a predefined threshold
3. Display the list and D to the user and wait for the user to choose an order or diagnosis from the list
4. Move the order from O to E, or set the diagnosis to *true* in E

Until the user closes the session

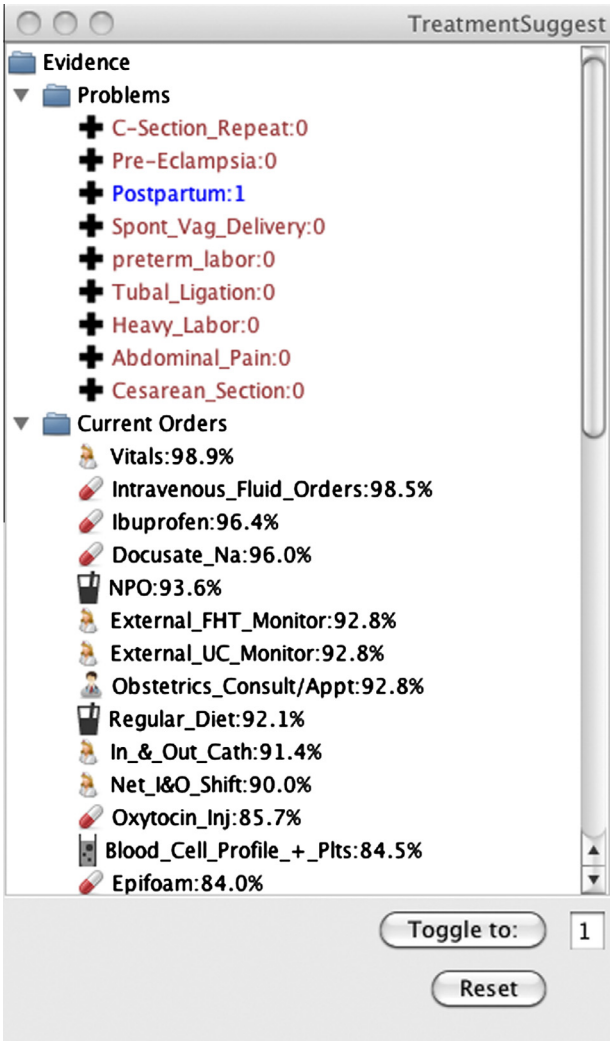


Fig. 2. A prototype implementation of Iterative Treatment Suggestions (ITS). The panel shows the current evidence (labeled 0 or 1) and the possible orders in descending probability order. As orders and diagnoses are placed (the toggle button), the evidence is revised and the posterior probability of possible orders given the network is recalculated.

2.2. Inducing Bayesian Networks from data

A common approach to induce a Bayesian Network from data (called structure learning) is a greedy search-and-score methodology. From a set of disconnected nodes, edges are added, removed, and reversed until a network is found that best explains a training dataset according to a scoring function. Here we used the BDeu scoring function [34]. A greedy search is used because a complete

Table 2
The co-occurring diagnoses and complaints in each domain-specific network, listed by their prevalence in the test sets. 0% indicates the co-occurrence was only present in the training set. Diagnoses were used as evidence as they appeared in the test cases, and were not part of the predictive evaluation.

Pregnancy, Inpatient	(%)	Back pain, ED	(%)	Hypertension, UVC	(%)	Medical, ICU	(%)
Postpartum	89	Vehicle Accident	4	Med Refill	27	Hypotension	<1
Cesarean Section	4	Neck Pain	3	Diabetes Mellitus	16	AIDS	0
Spont Vag Delivery	2	Abdominal Pain	3	Back Pain	6	Drug Abuse	0
Tubal Ligation	1	Chest pain	2	Abscess	6	Diabetes Mellitus	0
Pre-Eclampsia	1	UTI	2	Coronary Artery Disease	4	Encephalopathy	0
Preterm Labor	1	Headache	1	Toothache	4	Anemia	0
Abdominal Pain	1	Knee Pain	1	Cellulitis	4	Hypoglycemia	0
C-Section Repeat	<1	Hypertension	1	Headache	3	Hypokalemia	0
Failed Induction	<1	Med Refill	1	COPD	3	Sepsis	0
Failed induction	0	Shoulder Pain	<1	Hyperlipidemia	<1		

Table 3
For each domain, the weighted average AUC (Area Under the Receiver–Operator Curve) and position in menu at time of order, where 1 is the top suggestion). Weighting is by frequency of order.

Domain	Weighted average	
	AUC	Position
Inpatient pregnancy	.844	3.91
Medical intensive care unit	.781	5.72
Back pain in the emergency department	.765	5.83
Hypertension in the Urgent Visit Clinic	.741	4.88

exploration of all possible graphs is combinatorial, and so is therefore not possible on networks of more than a few nodes [35].

The most powerful greedy search is arguably the Greedy Equivalence Search (GES) [36]. Rather than searching Bayesian Networks, it searches what are known as 'equivalence classes' of Bayesian Networks. These are groups of Bayesian Networks that all are probabilistically equivalent. If an optimal Bayesian Network exists for the given dataset, GES will always find it. Therefore, we used a GES implementation in the freely available Tetrad toolkit [37].

2.3. Hospital simulation methodology

To evaluate ITS in the myriad of evolving clinical situations, we chose to compare how well the suggestion menus predict the actual next action taken in a hospitalization. Therefore we wrote a program to simulate hospitalizations on our test set using the ITS methodology. As in ITS (Table 1), our program places each order in the hospitalization in succession, adding it to the 'evidence' in the network, and recalculating the posterior probabilities for variables in the network. It also adds diagnoses as evidence at the appropriate time step in the hospitalization. After each order in the hospitalization, our program records the posterior probabilities in the menu (step 2), in order to calculate performance in predicting the next order. To determine order succession within each hospitalization, we used the time and session information in our order-entry data. Where two orders had the same recorded time, we used both possible orderings and kept the higher-scoring

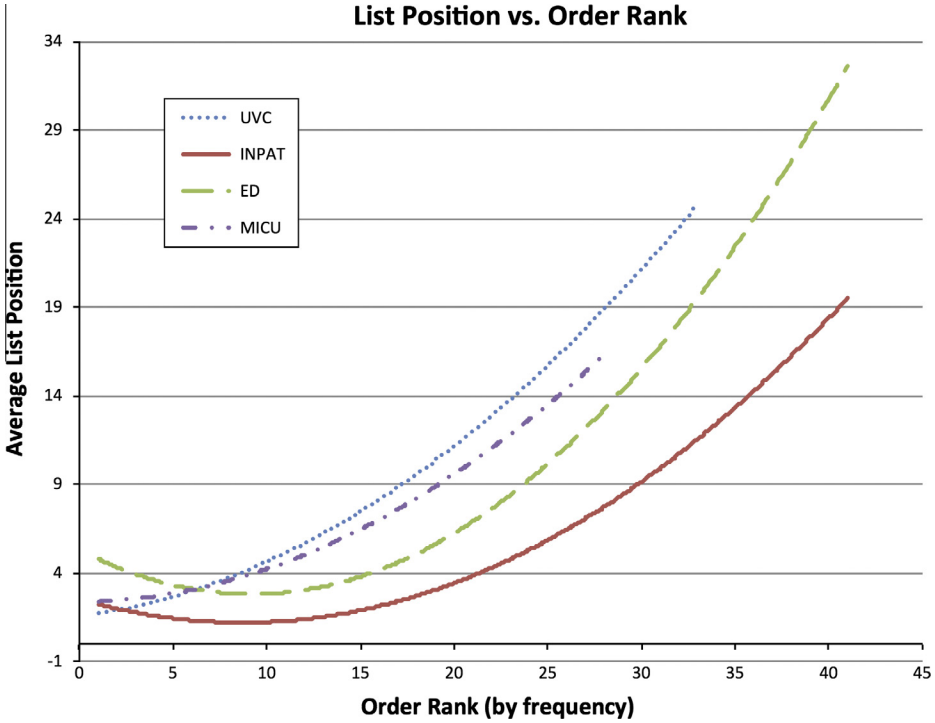


Fig. 3. The average position in the list at the time of order vs. the frequency rank of the order in the test sets.

combination. In the event an order was placed more than once, subsequent placements were ignored (because our system allows orders to be entered as evidence only once).

Using the recorded posterior probabilities and the actual next order placed, we were able to compute the Area Under the Receiver–Operator Curve (AUC). This measures *discriminability*, equivalent to the probability that when an order is placed, it will be ranked higher than at previous times. We used the approach in Hanley and McNeil [38] to calculate the AUC directly without first calculating the full ROC curve. The formula is as follows:

$$\mathrm{AUC}(\vec{T},\vec{F}) = \frac{\sum_t^{\vec{T}} \sum_f^{\vec{F}} \begin{cases} 1 & t > f \\ 0.5 & t = f \\ 0 & t < f \end{cases}}{\|\vec{T}\| * \|\vec{F}\|} \quad (1)$$

Here $\vec{T}$ is a list of posterior probabilities for true instances of a particular order, and $\vec{F}$ is the corresponding list for false instances.

We also computed the average position an order appears in the menu at the time it is selected. This measures *accuracy* by reporting the average list length required for 100% precision. The value is between one and the total number of orders in the network, where one is the top of the menu (and is therefore the best outcome).

2.4. Comparison with Association Rule Mining

To compare our approach to pairwise Association Rule Mining (ARM), we developed a variant of the ITS hospital simulation methodology. It performs the same analysis of average menu position but it uses a static menu of orders, which are arranged in descending frequency of co-occurrence with the main diagnosis in each domain (e.g., pregnancy in inpatient pregnancy). To facilitate direct comparison, the orders selected by GES were used to generate the menu in each domain.

2.5. Evaluation

2.5.1. Data source

For evaluation, we chose four modalities of medicine: inpatient medicine, the emergency department (ED), the Urgent Visit Clinic (UVC), and the intensive care unit (ICU). Each modality reflects a different aspect of medicine. Inpatient care focuses more on treatment than diagnosis in a longer-term stay, the ED involves a shorter stay involving both diagnosis and treatment, the UVC involves a very brief 'stay' focused on diagnosis, and the ICU involves tightly-correlated actions for very specific care.

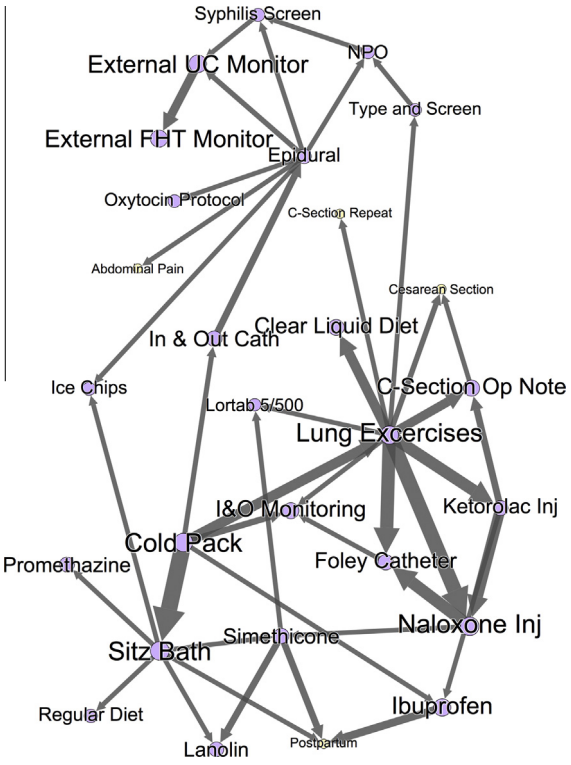


Fig. 4. A portion of the inpatient pregnancy networks. This figure shows the Markov Blankets of C-Section Operative Note, Ext. UC Monitor, and Sitz Bath, three nodes with high AUC in Table 4. These three Markov Blankets comprise the majority of the total graph, and the graph forms one single connected component – indicating strong relationships between all nodes in this network. Orders are purple; problem/complaints are yellow. Node/label size is proportional to AUC, and edge weight is an approximation of the strength of relationship. Notice the highly-correlated clusters, e.g. Sitz bath and other postpartum treatments (cold pack, ice chips, lanolin, etc.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Order name, AUC, and average menu position (#) of the ten best and worst order predictions in each domain. 'Best' and 'worst' are chosen by AUC (higher is better). Menu position, showing the average location in the suggestion menu just before selection, is also reported (lower is better).

Pregnancy, Inpatient			Back pain, ED			Hypertension, UVC			Altered mental state, MICU		
Name	AUC	#	Name	AUC	#	Name	AUC	#	Name	AUC	#
Sitz Bath	1.00	1.0	Abdomen CT	1.00	1.0	aPTT	1.00	1	Vancomycin Level	0.94	7.6
Cold Pack	1.00	1.1	Pelvis CT	1.00	1.2	Cardiac Markers	0.99	1.4	Ventilator Adjustment	0.94	2.7
Naloxone Inj	1.00	1.2	Peripheral Smear	0.96	11.6	ESR Test	0.97	6.7	Phosphorus Test	0.93	2.1
Lung Exercise	0.99	1.1	Cardiac Markers	0.96	2.2	Protime	0.95	1.8	Magnesium Level	0.91	2.9
Morphine (PCA)	0.99	2.0	Blood Cell Profile	0.95	1.7	Blood Culture	0.93	13.2	Basic Metabolic Panel	0.90	4.3
Ext. UC Monitor	0.99	1.0	Lipase	0.94	3.4	Drug Abuse Urine Test	0.92	3.7	Cardiac Markers	0.84	11.2
Ibuprofen	0.98	1.1	Vaginal Infection Test	0.93	5.7	BNP Test	0.91	6.8	Esomeprazole	0.84	6.9
Ext. FHT Monitor	0.97	1.1	Chest CT	0.92	9.4	Blood Cell Profile	0.88	2.3	Glucose	0.82	7.5
Docusate Na	0.96	1.2	Spine Cervical CT	0.92	4.7	Urine Culture	0.83	11.3	IV Fluids	0.82	1.7
I&O Monitoring	0.94	1.2	Comp. Metabolic	0.92	3.3	Dental Consult	0.82	6.4	Vancomycin	0.81	4.3
NPO	0.73	1.5	Phys. Therapy Consult	0.64	12.4	Hgb A1c	0.72	22	Zosyn	0.72	11.5
IV Lock	0.73	9.8	Lumbar Spine CT	0.63	29.4	Medicine Consult	0.69	2.6	NPO	0.71	12.7
Syphilis Screen	0.73	9.5	Knee Xray	0.62	23.4	Med Follow-up Consult	0.67	5.2	SCD	0.71	7.1
Ice Chips	0.72	15.8	Wrist Xray	0.61	38.1	Dermatology Consult	0.66	13.5	EKG	0.70	10.5
IV Fluids	0.71	1.1	Sports Med. Consult	0.60	31.7	Lateral Chest Xray	0.62	6.3	Restraints	0.68	3.5
Drugs Urine Test	0.71	27.8	EPIC Referal	0.59	29.0	Physl Therapy Consult	0.57	21.4	Frontal Chest Xray	0.68	3.7
Oxytocin Protocol	0.68	23.8	Neurosurgery Consult	0.59	13.8	Head CT	0.56	28.0	Albuterol	0.63	17.7
Type and Screen	0.65	13.2	Medicine Consult	0.58	5.8	TSH	0.50	22.0	Furosemide	0.55	15.3
Lortab 5/500	0.60	2.9	Med Follow-up Consult	0.57	6.5	T4-Free Level	0.50	28.0	Prealbumin	0.51	12.5
Morphine	0.50	22.7	Lumbar Spine MRI	0.53	17.5	Knee Xray	0.50	23.8	Arterial blood gas	0.50	5.7

We extracted data for four domain-specific BNs from the four selected modalities as follows:

1. *Choosing chief diagnosis:* We focused our domains on the most frequent diagnosis/complaint for the four modalities: visits involving pregnancy in inpatient medicine, back pain in the ED, hypertension in the UVC, and 'altered mental state' in the Medical ICU (MICU).
2. *Data extraction:* We extracted and de-identified 3 years of inpatient order-entry data from the local county hospital in Indianapolis (2007–2009) and chose visits that corresponded with each domain. This involved 9228 ED back pain, 1821 UVC hypertension, 4843 inpatient pregnancy, and 1546 'altered mental state' MICU visits.
3. *Variable selection:* For each domain, we selected 50 variables: the 40 most frequent orders and the 10 most frequent co-occurring diagnoses and complaints. Orders were of low granularity, which ensured sufficient data for predictive power; for example, medication orders only included the type of medicine (e.g., vancomycin), not the route, dose, or frequency. The diagnoses and complaints used in our networks can be seen in Table 2. Note that sometimes less than ten are shown because fewer than ten diagnoses/complaints co-occurred with the diagnosis.
4. *Train/test split:* We split each data set into a training (2/3 of admissions) and test set (1/3).

2.5.2. Computational approach

Using these four data sets, we applied and evaluated the BN and ARM methods as follows:

1. *Network induction:* Via GES (Section 2.2), we induced four Bayesian Networks using each of the four training sets. Because GES will discard nodes that do not have predictive power, sometimes the resulting networks contained fewer than 50 nodes. This was most notable in the ICU network, where only 25 orders were retained.
2. *Hospitalization simulation:* We ran our ITS hospital-simulation program (Section 2.3) on each the each of the four networks using their corresponding test set, which collected statistics on AUC and average position in the menu at time of selection.

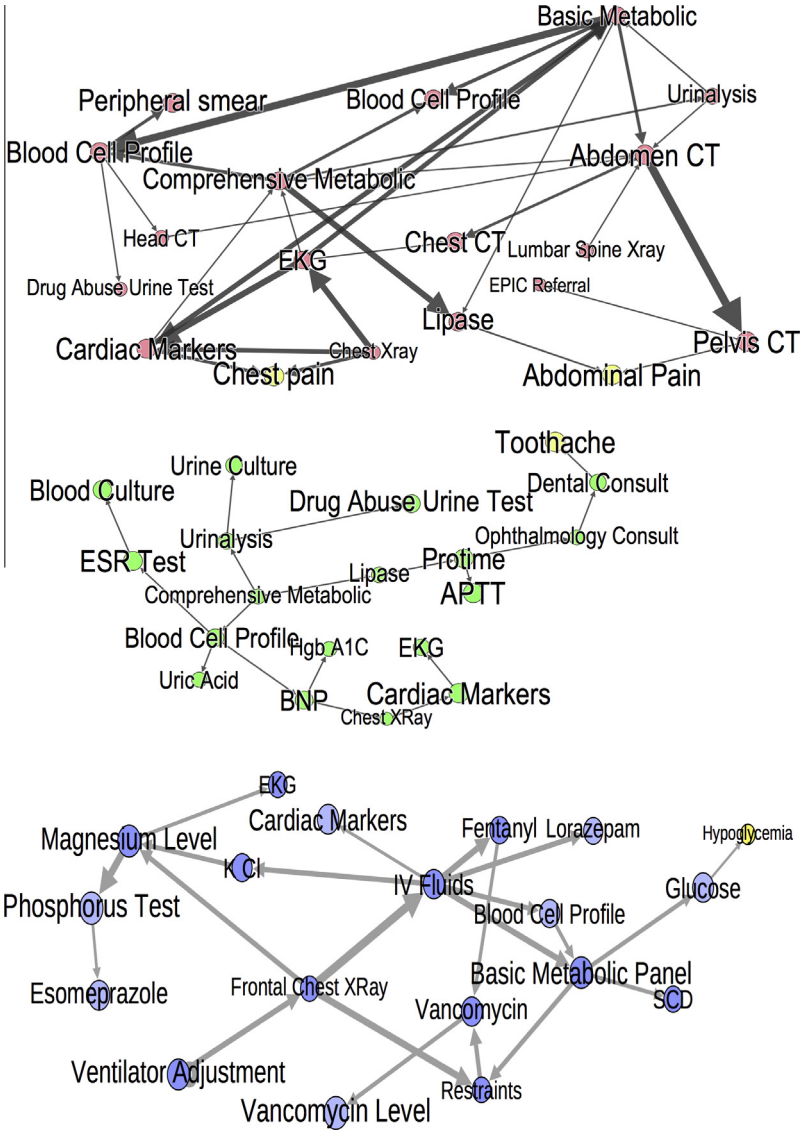


Fig. 5. High AUC nodes from Table 4 with their parents and children in all domains but inpatient. MICU is blue (bottom), UVC is green (middle), and ED is red (top). Problems/complaints are yellow. Node/label size is proportional to AUC, and edge weight is an approximation of the strength of the relationship. Here, notice the logical clusters and intuitively correct relationships. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. *Visualization:* We wrote a program to export the networks into Gephi format. Gephi is an open-source network visualization tool [39]. We wrote a Gephi script to select the Markov Blankets for a set of nodes. A *Markov Blanket* of a node is its parents, children, and siblings, and is frequently used as a heuristic for the set of most relevant variables in prediction [40]. This allowed us to visually examine nodes in a graph and their most important neighbors.
4. *Comparison to Association Rule Mining:* We ran our ARM-based hospital-simulation (Section 2.4), which collected statistics on average position in a static menu at time of selection.

3. Results and discussion

A standard desktop computer induced each network (step 1) in less than 30 min and ran the ITS hospital-simulation program (step 2) in an average of 5 min. Table 3 shows summary statistics: average AUC and average menu position, weighted by the frequency of each order. Fig. 3 shows trendlines of the average position vs. order rank by frequency. For each domain, the 10 orders in which the system performed best and worst (by AUC) are shown in Table 4.

Figs. 4–6 show portions of the graph structure (step 3). Fig. 4 shows the Markov Blankets around some nodes in the pregnancy network with high AUC. Fig. 5 shows nodes with high AUC and their parents and children in the other three networks. Fig. 6 does the same with nodes of low AUC. Note that arrow directions should not be interpreted as showing causality, only a statistical association.

Finally, Table 5 and Fig. 7 compare the BN approach (step 2) to an ARM approach (step 4). Table 5 shows the weighted and unweighted average difference in list length between ARM and BN. Fig. 7 shows average menu position vs. order rank by frequency using ARM. It is directly comparable to Fig. 3 for the BN approach.

3.1. Analysis of BN approach

The evaluation of our treatment suggestion system on four domain-specific BNs against test cases drawn from the same environments showed fairly strong overall performance. In particular, our treatment suggestion menus correctly suggest common orders in a short list: 3.91–5.83 items (Table 3). A length of five accurately suggests more than the top 20 inpatient pregnancy orders and emergency department back pain orders (Fig. 3). Also, the system's average AUC is high (74–84%, also in Table 3), meaning that common orders are ranked higher at the time they are ordered than prior to ordering.

There was high variance in performance on individual orders (AUC 0.5–0.99), both across and within domains (Tables 3 and 4). Within a domain, some orders are suggested almost exactly when they should be, such as a cold pack in pregnancy visits and a pelvis CT in the ED. Other orders appear at the bottom of long menus and are not predicted much better than chance, such as a neurology consult in the ED. Performance varied across domains as well. Inpatient pregnancy had a weighted average AUC .884 and menu position 3.91 (Table 3), and even the least frequent orders required a menu length of only half the total orders (Fig. 3). In the other domains, average AUC and menu length were notably worse and the least frequent orders required a menu length containing at least 75% of possible orders.

Figs. 4–6 shed light on this phenomenon. For high AUC nodes (Figs. 4 and 5), the network diagrams are tight clusters with connections that make intuitive sense. For example, postpartum is directly connected to adjuncts like simethicone, toothache is connected to a dental consult, and related tests like magnesium and phosphorus levels are linked. This clustering and intuitiveness indicates that the correct amount of context was provided for these nodes. The pregnancy network formed one giant cluster, which likely explains its high overall performance. The low-performing nodes in the other networks were either part of smaller subnetworks, or, in the case of the MICU, relied on infrequent diagnoses that were not in the test set (Fig. 6). Relationships among low-performing nodes were frequently almost linear and had non-intuitive connections, indicating transitive associations due to missing context. For example, restaints is directly connected to vancomycin (see Fig. 6, MICU) – both might be appropriate when a patient has an infection causing delirium, but they are not predictive of each other. Also a general medicine consult does not directly predict a diagnosis of diabetes (see Fig. 6, UVC), nor does a lumbar spine X-ray directly suggest a knee X-ray. The context needed

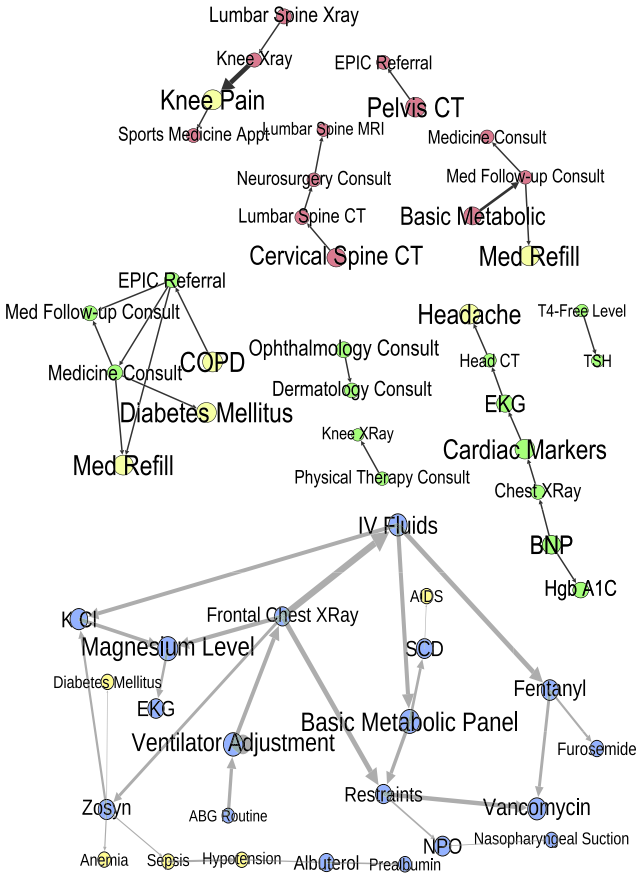


Fig. 6. Low AUC nodes from Table 4 with their parents and children in all domains but inpatient. Notice the linear chains, multiple subnetworks, connection to infrequent diagnoses, and transitive relationships. This indicates appropriate context is lacking for these nodes. (Top to bottom: ED, UVC, MICU.)

Table 5
For each domain, the weighted average position in menu at time of order, where 1 is the top suggestion, for the BN and ARM approaches. Weighting is by frequency of order. Also shows the weighted and unweighted difference in average list length (ARM-BN).

Domain	Weighted average position			Unweighted
	BN	ARM	Difference	Difference
Inpatient pregnancy	3.91	5.67	+1.76	+2.73
Medical intensive care unit	5.72	5.95	+0.23	+1.14
Back pain in the emergency department	5.83	9.87	+4.04	+7.64
Hypertension in the Urgent Visit Clinic	4.88	6.06	+1.18	+4.04

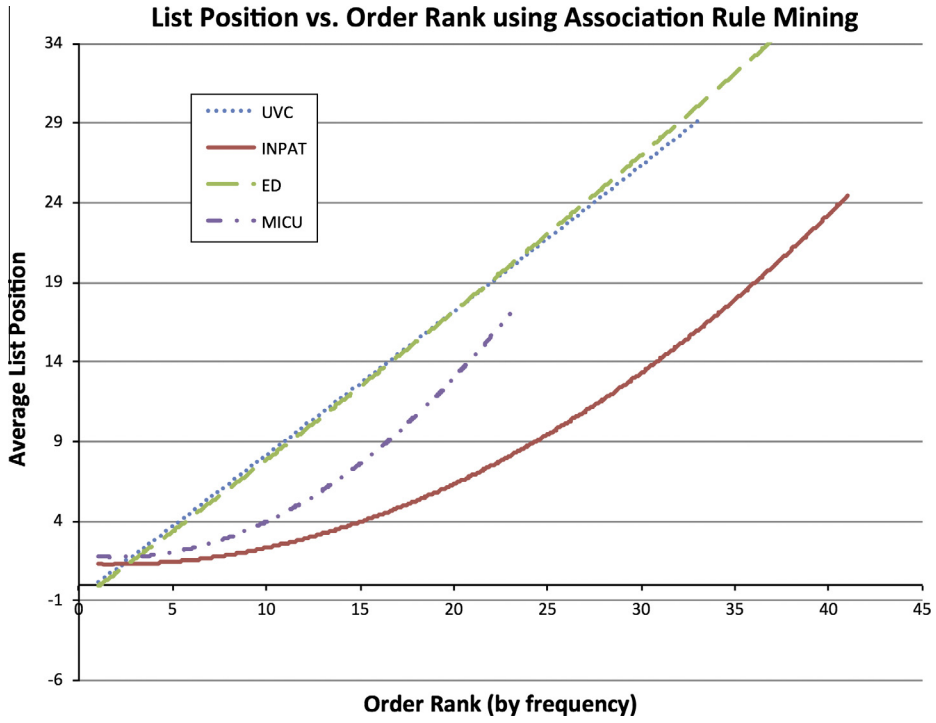


Fig. 7. Using an Association Rule Mining approach, the average position in the list at the time of order vs. the frequency rank of the order in the test sets.

likely includes: additional well-chosen orders and diagnoses, external information about patient health status, test results, and family history. This points to the need for additional data sources and more principled feature selection.

Another interesting discovery is that AUC is not always strongly correlated with menu position. Two examples can be seen in Table 4. A peripheral blood smear in the emergency department has high AUC but an average menu position of 11.6, and an order for Lortab (a narcotic painkiller) in inpatient pregnancy appears near the top of the suggestion menus but has AUC of only 0.60. In the first case, we suspect that although the blood smear's probability increases just prior to it actually being ordered, it is never high enough to outweigh other orders. In the second case, we believe the order stays at the top of the menu until it is picked because it has a high prior probability. We therefore conclude that choosing order-specific probability thresholds might be appropriate.

3.2. Comparison to ARM

Our results confirm previous results regarding ARM approaches: while an ARM approach can readily detect the most common associations, the strength of less common associations depend on context (e.g., previous orders and diagnoses) that ARM cannot capture.

In Table 5, there is a relatively small difference in weighted average menu length between the two approaches (Table 5), especially in smaller domains like the ICU (difference +0.23 items). This indicates similar performance for the most common orders. However, the unweighted difference is larger (+1.14 to +7.64 items), suggesting that the BN approach is having more impact on less common orders.

Comparing Fig. 7 (ARM) to Fig. 3 (BN) confirms this. Fig. 3 displays a slow increase in menu length as more orders are included, but Fig. 7 shows a much steeper rise. With the BN approach, a length of five accurately suggests an average of 16 orders (Fig. 3). The same menu length with the ARM approach accurately suggests only 9 orders on average (Fig. 7). Performance degrades rapidly as menu length increases. This confirms the BN approach's overall superior performance.

3.3. Limitations and future directions

This research is predicated on the assumption that average patterns in the data represent reasonably good care for future patients. As detailed in the Section 1.1, in many decision-making problems, average patterns do in fact represent 'crowd wisdom' [41], but 'crowd madness' – the domination of bad decisions in a group – can occur as well. Automatically discriminating wisdom from madness is important future work. Presently the 'wisdom' discovered should be reviewed by experts and aligned with guidelines before deployment.

The other principal limitation is that our models currently rely only on a small set of orders and diagnoses. We do not include other important factors such as test outcomes and physiologic changes. Also, we evaluated the networks using time-stamped data but the algorithm we used to learn networks does not utilize time information. Additionally, among orders and diagnoses, we choose the most frequent. All of this biases our system to short-term decisions that can be made with minimal context. We believe accuracy will be improved significantly with context-aware feature selection and temporal extensions to BN structure learning.

Our system and evaluation do not currently accommodate multiple orders of the same item within a hospitalization. Upon examination of our training sets, only orders in the ICU occurred multiple times on average per hospitalization. However, in the ICU, 16 orders (e.g. ventilator protocol changes, IV fluids, and common tests) do occur with multiplicity, and for this we need to develop a more complex methodology. We are exploring use of a 'temporal window' around the actual occurrence of the order in which we consider it a true instance.

The BN approach requires networks to remain relatively small, or data requirements and computational complexity become intractable [42]. We do not believe this makes them unattractive to 'big data' problems, but it will require an approach to intelligently create sets of largely independent domain-specific networks. We also plan to explore structure-learning algorithms that scale to larger data sets.

Our comparison to ARM was a side-by-side comparison that might have unfairly benefitted ARM. For one, only items chosen by GES were used in the menu – and some of the dropped associations might have been incorrect transitive associations. Also, including less common orders might show even more difference between BN and ARM. Further comparison is important future work.

Finally, our evaluation measures – AUC and menu position – only capture two aspects of the approach's predictive performance – discriminability and precision. There are many other classification evaluation measures (see for example [43]). For this methodology, it would also be valuable to measure the menu's utility as a decision-making aid. This could be done computationally using a decision-theoretic approach like decision curve analysis [44], or by soliciting feedback regarding sample menus from potential users.

4. Conclusion

The proliferation of medical data in EMRs offers an opportunity to abstract these data for use in Clinical Decision Support. Both the challenges associated with creating localized decision support and the incompleteness of guideline recommendations make this an important task. Existing approaches using pairwise Association Rule Mining produce long static lists that accurately capture only common, direct associations.

In this work, we have developed and implemented a system using Bayesian Network learning to discover the typical successions of orders made by clinicians from local order-entry data, which we have used as an adaptive recommendation system to suggest the most common next orders based on what has been ordered and diagnosed previously. We used a hospitalization-simulation evaluation methodology to determine how well our system reproduces reasonable behavior in four medical domains.

Our system performed fairly well on average in all domains but had variance that suggested future improvements. It performed best in inpatient pregnancy (weighted average AUC .844, weighted average menu position 3.91) and worst in the Urgent Visit Clinic (weighted average AUC .741, weighted average menu position 4.88). Our system had near-perfect performance on some orders (e.g., cold pack in inpatient pregnancy) but very poor performance on others (e.g., arterial blood gas monitoring in the medical intensive care unit). Higher performance appears to correlate with the presence of more factors needed to predict the order.

Comparing our system to an ARM-based equivalent, we found that only the most common orders are accurately suggested by both systems, and that a menu length of five suggested only about half as many orders accurately in ARM vs. BN. This confirms that despite the future work needed in our system, it does outperform existing approaches.

This study is a step forward in clinical knowledge-abstraction systems. Such a system could eventually be part of the envisioned "learning health system," in which a variety of clinical users – including researchers, administrators, and physicians – could dynamically analyze vast amounts of data for improved decision-making. This could be used for e.g., workload reduction in developing localized CDS, or as a method to quickly analyze local practice patterns.

Contributorship statement

Dr. Klann designed and implemented the study and wrote the manuscript.

The other authors served as advisors, helping to conceptually devise portions of the study, revise the methodology and implementation strategy, and provide feedback on the study design. The authors each offered particular expertise: Dr. Szolovits in machine learning approaches on clinical data; Dr. Downs in decision modeling and Bayesian Networks, and Dr. Schadow in clinical data mining and data analysis.

All authors also edited, contributed to, and approved the manuscript.

Acknowledgments

Thanks to Jeff Warvel for providing both data and expertise regarding the county-hospital order-entry system; and to Siu Hui for her insights into statistics and evaluation approaches. This work was performed at the Regenstrief Institute, Indianapolis, IN and at the Massachusetts General Hospital Laboratory for Computer Science, Boston, MA. This work was supported in part by Grant 5T15 LM007117-14 from the National Library of Medicine.

References

[1] Corrigan JM, Donaldson MS, Kohn LT, Maguire SK, Pike KC. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: Institute of Medicine; 2001.
[2] Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med 2003;163:1409–16. http://dx.doi.org/10.1001/archinte.163.12.1409.
[3] Waitman LR. Pragmatics of implementing guidelines on the front lines. J Am Med Inform Assoc 2004;11:436–8. http://dx.doi.org/10.1197/jamia.M1621.
[4] Geissbuhler A, Miller RA. Distributing knowledge maintenance for clinical decision-support systems: the 'knowledge library' model. Proc AMIA Symp 1999:770. 10566464.
[5] Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005;293:1223–38. http://dx.doi.org/10.1001/jama.293.10.1223.
[6] Zhou L, Soran CS, Jenter CA, Volk LA, Orav EJ, Bates DW, et al. The relationship between electronic health record use and quality of care over time. J Am Med Inform Assoc 2009;16:457–64. http://dx.doi.org/10.1197/jamia.M3128.
[7] Van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006;13:138–47. http://dx.doi.org/10.1197/jamia.M1809.
[8] Standards & Interoperability (S&I) Framework. Health eDecisions Homepage. <http://wiki.siframework.org/Health+eDecisions+Homepage> [accessed 29.05.13].
[9] Sittig DF, Wright A, Osheroff J, Middleton B, Teich J, Ash J, et al. Grand challenges in clinical decision support. J Biomed Inf 2008;41:387–92. http://dx.doi.org/10.1016/j.jbi.2007.09.003.
[10] Gorman PN, Ash J, Wykoff L. Can primary care physicians' questions be answered using the medical journal literature? Bull Med Libr Assoc 1994;82:140–6. 7772099.
[11] Haug JD. Physicians' preferences for information sources: a meta-analytic study. Bull Med Libr Assoc 1997;85:223–32. 9285121.
[12] Perley CM. Physician use of the curbside consultation to address information needs: report on a collective case study. J Med Libr Assoc 2006;94:137–44. PMCID: PMC1435836.
[13] Ford EW, Menachemi N, Phillips MT. Predicting the adoption of electronic health records by physicians: when will health care be paperless? J Am Med Inform Assoc 2006;13:106–12. 16221936.
[14] Blumenthal D, Tavenner M. The 'meaningful use' regulation for electronic health records. N Engl J Med 2010;363:501–4. http://dx.doi.org/10.1056/NEJMp1006114.
[15] Office of the National Coordinator for Health IT. Federal health information technology strategic plan 2011–2015; 2011. <http://www.healthit.gov/sites/default/files/utility/final-federal-health-it-strategic-plan-0911.pdf>.
[16] McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, et al. The quality of health care delivered to adults in the United States. N Engl J Med 2003;348:2635–45. http://dx.doi.org/10.1056/NEJMsa022615.
[17] Condorcet M. Essay sur l'application de l'analyse de la probabilité des decisions: Redues et pluralité des voix. l'Imprimerie Royale; 1785.
[18] Arrow KJ. A difficulty in the concept of social welfare. J Political Econ 1950;58:328–46.

[19] Austen-Smith D, Banks JS. Information aggregation, rationality, and the Condorcet Jury theorem. Am Political Sci Rev 1996;90:34–45. http://dx.doi.org/10.2307/2082796.
[20] Fisher E, Goodman D, Skinner J, Bronner Kristen. Health care spending, quality, and outcomes, The Dartmouth Institute for Healthcare Policy and Clinical Practice 2009.
[21] Hasan S, Duncan GT, Neill DB, Padman R. Towards a collaborative filtering approach to medication reconciliation. AMIA Annu Symp Proc 2008:288–92. PMID:18998834.
[22] Wright A, Chen E, Maloney FL. Using medication data and association rule mining for automated patient problem list enhancement. AMIA Annu Symp Proc 2009:707.
[23] Klann J, Schadow G, McCoy JM. A recommendation algorithm for automating corollary order generation. Proc AMIA Symp 2009:333–7. 20351875.
[24] Carter JS, Brown SH, Erlbaum MS, Gregg W, Elkin PL, Speroff T, et al. Initializing the VA medication reference terminology using UMLS metathesaurus co-occurrences. Proc AMIA Annu Symp 2002:116–20. PMID: 12463798.
[25] McCoy AB, Wright A, Laxmisan A, Ottosen MJ, McCoy JA, Butten D, et al. Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications. J Am Med Inform Assoc 2012;19:713–8. http://dx.doi.org/10.1136/amiajnl-2012-000852.
[26] Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med 2011;365:1758–9. http://dx.doi.org/10.1056/NEJMp1108726.
[27] Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Comput 2003:76–80. http://dx.doi.org/10.1109/MIC.2003.1167344.
[28] Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform 2010;43:891–901. http://dx.doi.org/10.1016/j.jbi.2010.09.009.
[29] Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, Ramelson HZ, et al. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. J Am Med Inform Assoc 2011;18:859–67. http://dx.doi.org/10.1136/amiajnl-2011-000121.
[30] Heckerman DE, Nathwani BN. Toward normative expert systems: Part II. Probability-based representations for efficient knowledge acquisition and inference. Methods Inf Med 1992;31:106–16. PMID 1635462.
[31] Heckerman D. A tutorial on learning with Bayesian Networks. Innovations in Bayesian Networks, <http://dx.doi.org/10.1007/978-3-540-85066-3_3>.
[32] Klann J, Schadow G, Downs S. A method to compute treatment suggestions from local order entry data. Proc AMIA Symp 2010:387–91. PMID: 21347006.
[33] Druzdzel MJ. SMILE: structural modeling, inference, and learning engine and GeNIe: a development environment for graphical decision-theoretic models. In: Proceedings of the 16th national conference on artificial intelligence and the 11th innovative applications of artificial intelligence conference; 1999. p. 902–3 [ACM ID: 315504]. <http://portal.acm.org/citation.cfm?id=315149.315504> [accessed 16.03.11].
[34] Buntine W. Theory refinement on Bayesian Networks. In: Proceedings of the seventh conference (1991) on uncertainty in artificial intelligence; 1991. p. 52–60 [ACM ID: 114105]. <http://portal.acm.org/citation.cfm?id=114098.114105> [accessed 18.07.11].
[35] Eaton D, Murphy K. Exact Bayesian structure learning from uncertain interventions. AI Stat 2007:107–14.
[36] Chickering DM. Optimal structure identification with greedy search. J Mach Learn Res 2003;3:507–54. <http://jmlr.org/papers/volume3/chickering02b/chickering02b.pdf> [accessed 20.12.13].
[37] Ramsey J. Tetrad project homepage; 2011. <http://www.phil.cmu.edu/projects/tetrad/tetrad4.html> [accessed 06.03.10].
[38] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36. PMID:7063747.
[39] Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks; 2009. <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.
[40] Tsamardinos I, Aliferis CF. Towards principled feature selection: relevancy, filters and wrappers. In: Proceedings of the ninth international workshop on artificial intelligence and statistics; 2003.
[41] Surowiecki J. The wisdom of crowds. Random House, Inc.; 2005.
[42] Chickering DM, Heckerman D, Meek C. Large-sample learning of Bayesian Networks is NP-hard. J Mach Learn Res 2004;5:1287–330.
[43] Medlock S, Ravelli ACJ, Tamminga P, Mol BWM, Abu-Hanna A. Prediction of mortality in very premature infants: a systematic review of prediction models. PLoS One 2011;6:e23441. http://dx.doi.org/10.1371/journal.pone.0023441.
[44] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74. http://dx.doi.org/10.1177/0272989X0629536, PMID: 17099194PMCID: PMC2577036.