Published in final edited form as:
J Biomed Inform. 2009 February ; 42(1): 82–89. doi:10.1016/j.jbi.2008.07.00.

Classification models for the prediction of clinicians' information needs

Guilherme Del Fiol, MD, MS[1,2] and **Peter J. Haug, MD**[1,2]

1 *Biomedical Informatics Department, University of Utah, Salt Lake City, UT*

2 *Intermountain Healthcare, Salt Lake City, UT*

Abstract

Objective—Clinicians face numerous information needs during patient care activities and most of these needs are not met. Infobuttons are information retrieval tools that help clinicians fulfill their information needs by providing links to on-line health information resources from within an electronic medical record (EMR) system. The aim of this study was to produce classification models based on medication infobutton usage data to predict the medication-related content topics (e.g., e.g., dose, adverse effects, drug interactions, patient education) that a clinician is most likely to choose while entering medication orders in a particular clinical context.

Design—We prepared a dataset with 3,078 infobutton sessions and 26 attributes describing characteristics of the user, the medication, and the patient. In these sessions, users selected one out of eight content topics. Automatic attribute selection methods were then applied to the dataset to eliminate redundant and useless attributes. The reduced dataset was used to produce nine classification models from a set of state-of-the-art machine learning algorithms. Finally, the performance of the models was measured and compared.

Measurements—Area under the ROC curve (AUC) and agreement (kappa) between the content topics predicted by the models and those chosen by clinicians in each infobutton session.

Results—The performance of the models ranged from 0.49 to 0.56 (kappa). The AUC of the best model ranged from 0.73 to 0.99. The best performance was achieved when predicting choice of the *adult dose*, *pediatric dose*, *patient education*, and *pregnancy category* content topics.

Conclusion—The results suggest that classification models based on infobutton usage data are a promising method for the prediction of content topics that a clinician would choose to answer patient care questions while using an EMR system.

Keywords

Information storage and retrieval; machine learning; clinical decision support systems; infobuttons; web usage mining; information needs

Corresponding Author: Guilherme Del Fiol, MD, MS, Intermountain Healthcare, 4646 Lake Park Blvd, Salt Lake City, UT, 84120, Phone: 801-442-6303/Fax: 801-442-6995, E-mail: E-mail: guilherme.delfiol@utah.edu.

I. INTRODUCTION

Numerous information needs arise in the course of patient care. It has been estimated that the frequency of information needs ranges from one to four questions per patient encounter [1], [2],[3]. A large percentage of these needs are not met, mostly because clinicians fail to find an answer to their need or because they opt not to pursue an answer [1],[4],[5]. A significant percentage of these needs are related to gaps in medical knowledge that could be filled by one of the numerous on-line health information resources that have become available since the advent of the World Wide Web [6],[7]. However, a number of barriers to the use of information resources at the point of care preclude a more frequent and effective use of such resources [4],[5]. These barriers include lack of time, doubt that an answer exists or can be easily found, and lack of access to resources that can directly answer the question.

It has been suggested that solutions to this problem should facilitate access to information that reflects the context in which information needs arise [8],[9]. “Infobuttons” are examples of such solutions [10]. Infobuttons attempt to predict the information needs that a clinician might have while using an electronic medical record (EMR) system and provide links to relevant content in an attempt to fulfill these needs (Figure 1) [11].

Infobuttons are typically implemented with a software component called an “Infobutton Manager” [12]. The core piece of the Infobutton Manager is a knowledge base that contains rules that map the various instances of context and the information needs that may arise in each of these instances to information resources [11]. In present Infobutton Manager implementations, these rules need to be manually designed and coded in the knowledge base. In previous studies, we have explored the feasibility of machine learning and Web usage mining techniques to enhance the prediction of information needs [13],[14]. In these studies, we found that infobutton usage data can be used to create classification models that accurately predict the information resources that a clinician is most likely to visit in a particular EMR context. In this present study, we conducted a similar investigation, but focus on prediction of the content topics that a clinician might find useful to fulfill her information needs.

A. Attribute selection techniques

Machine learning deals with the task of automatically inferring prediction models from data. Traditionally, the success of a learning method is dependent on its ability to identify a small subset of highly predictive attributes [15]. More recently, machine learning has been applied to domains characterized by remarkably high attribute dimensionality (with many of these attributes being irrelevant or redundant), relatively few training instances, and scarce availability of expert knowledge [16]. In these domains, the identification of a subset of optimal predictors almost invariably must be accomplished using automated methods.

Attribute selection is the process of identifying and removing as much of the irrelevant and redundant information as possible from a dataset. The reduction of dimensionality in a dataset presents a number of benefits, such as enabling algorithms to operate faster and more effectively, improving classification accuracy, improving data visualization, and enhancing understanding of the derived classification models [16].

Automated attribute selection methods can be classified into attribute ranking methods and attribute subset evaluation methods [15],[16]. The former methods assess the merit of individual variables for prediction independently of other attributes. These methods can be used as an initial screening to reduce dimensionality in large datasets or merely to produce a baseline. The latter methods assess the usefulness of subsets of attributes, accounting for redundancy and interactions among multiple attributes.

Several attribute ranking and attribute subset evaluation methods have been proposed. In a recent benchmark study, Hall and Holmes compared many of these methods and identified the ones that performed best given a range of prediction problems and datasets [15]. The best performing methods according to this benchmark study were “Information Gain”, “Recursive Elimination of Features (Relief)” [17], “Correlation-Based Feature Selection (CFS)” [18], “Consistency-Based Subset Evaluation” [19], and “Wrapper Subset Evaluation”[20]. The first two are attribute ranking methods and the latter three are attribute subset evaluation methods. Table 1 summarizes the mechanism, advantages, and disadvantages of each method. The results of the benchmark study provide guidelines for the choice of attribute selection methods, but highlight that the method of choice for a particular learning problem depends on factors such as the characteristics of the dataset, computational processing time restrictions, and learning algorithm [15]. Therefore, attribute selection methods still need to be evaluated in light of the prediction problem at hand to determine an optimal choice.

B. Machine learning algorithms

Several types and variations of machine learning algorithms are available. Examples are rules, decision trees, nearest neighbor, Naïve Bayes, Bayesian networks, Multiple Regression, Neural Networks, and Support Vector Machines [21]. Each method has its advantages and disadvantages. For example, decision trees, Naïve Bayes, and rules tend to be faster than Bayesian networks and Support Vector Machines and perform reasonably well in most prediction problems. Yet, the latter two algorithms tend to outperform the former in situations where data are noisy or missing and attributes are not conditionally independent. Although benchmark studies have revealed some overall winners, the choice of a learning algorithm must be made in light of the characteristics of a given prediction problem, data source, and prediction performance metrics [22],[23].

A relatively recent type of learning method known as “ensemble learning” or “meta-learning” combines the output of multiple models, called “base learners”, to produce a final prediction [24],[25]. A meta-learning model frequently outperforms any of its individual members. In the present study, two meta-learning techniques are investigated: “Boosting” [26] and “Stacking” [27]. Boosting produces multiple base models of the same type in a sequence of learning iterations. In each iteration, training set cases that were misclassified by the model generated in the previous iteration are assigned a higher weight for training. Thus, Boosting allows subsequent models to focus on the examples that are more difficult to predict. Once a set of models is created, the final prediction for a given case is achieved by aggregating the predictions of the base models. Boosting works particularly well with base learning methods that tend to produce unstable models (i.e., models that easily become outdated with minor changes to the data distributions), such as decision trees and rules. Boosting generally performs well even when composed of a weak base learner. In the present study, we used an improved implementation of Boosting called “MultiBoost” [28].

Stacking also combines multiple models but, unlike Boosting, these models can be derived from a mixed set of learning algorithms. For example, a Stacking classifier may combine a Naïve Bayes, a decision tree, and a rule-based classifier. Each classifier makes its own prediction estimating probabilities for each class. The final prediction is then computed using a meta classifier, such as multiple linear regression, which uses the class probabilities of the base learners as attributes of the model. Stacking was initially proposed by Wolpert in 1992 [27]. Seewald implemented an enhanced version of Stacking, called StackingC, which improves the performance of Stacking on multi-class prediction problems [29]. Stacking generally performs better than the best single base learner contained in the ensemble [23], [24],[27].

II. METHODS

The study method consisted of five steps: identification of data sources, data cleaning, data preparation, automated attribute selection, and classification (training and performance evaluation). The three latter steps were done in Weka, an open source machine learning tool that contains Java-based implementations of the algorithms mentioned previously [21].

A. Study environment

This study was conducted at Intermountain Healthcare, a healthcare delivery network located in Utah and Southeastern Idaho. Clinicians at Intermountain have access to a web-based EMR called HELP2 Clinical Desktop [30]. A number of modules in the Clinical Desktop offer infobuttons, including laboratory results review, problem list, and medication order entry (Figure 1). Infobuttons are implemented in HELP2 using an Infobutton Manager [11],[31]. In 2007, an average of 885 users clicked on infobuttons at least once every month. These users conducted an average of approximately 4,000 infobutton sessions per month; 67% of these originated from the medication order entry module. Although this module is used primarily in the outpatient setting, clinicians in the inpatient setting have read-only access to infobuttons in the medication lists that are created in the outpatient environment. More detailed analyses of the usage and usefulness of infobuttons at Intermountain and other institutions are available elsewhere [31],[32],[33].

B. Data sources

The data sources used in this study were 1) the Infobutton Manager monitoring log, a database that keeps a detailed record of every infobutton session; 2) the Intermountain Enterprise Data Warehouse (EDW), a large repository of clinical and administrative data used for analytical purposes; and 3) the Intermountain terminology server [30]. From these sources, attributes considered to be potentially useful predictors were extracted and merged into one single dataset using SQL queries. The dataset was limited to medication infobutton sessions that occurred between May 15, 2007 and December 5, 2007. This dataset contains attributes that characterize the clinical user, the medication associated with the infobutton, the patient, and the topics that users selected. Since users can view multiple topics in one single session, the first topic that was accessed in a given session was considered the target class label for prediction. Table 2 contains a complete list of the 26 attributes that were included in the dataset.

C. Data cleaning and preparation

A series of ad-hoc steps were performed to reduce noise from the dataset, especially to remove sessions that may not have been conducted to fulfill real information needs. As a result, the following sessions were excluded from the dataset:

1. Sessions associated with test patients or conducted by information systems personnel.
2. Sessions where users clicked on more than four topics. We thought these sessions were more likely to be motivated by testing, demonstration, or training purposes than by patient care needs.
3. Sessions where the selected topic accounted for less than 0.5% of the sessions in the dataset and thus were considered not to have a sufficient number of cases for training (e.g., contraindications, medication class, dose adjustment, breast feeding).
4. Sessions that had a duration of less than six seconds. After evaluating the minimum time necessary to display and review infobutton results, it was felt that sessions less than six seconds represented uninformative interactions, probably user errors in invoking the infobutton service.

After these steps were concluded, the dataset contained 3,078 cases. Class labels were distributed as follows: *adult dose* (58.6%), *patient education* (18.6%), *adverse effects* (11.2%), *pediatric dose* (4.8%), *pregnancy category* (2.2%), *drug interactions* (1.6%), *precautions* (1.6%), and *how supplied* (1.5%).

To reduce attribute dimensionality and missing data in preparation for the attribute selection and classification steps, three data transformation processes were applied:

1. Attributes with high dimensionality were transformed to reduce the number of possible values. For example, the medications associated with infobutton sessions are represented in the infobutton monitoring log at the clinical drug level (i.e., ingredient, strength, and presentation). The clinical drugs were transformed into higher drug classification levels according to a hierarchical knowledge base that is available in the terminology server[1]: *Parent level 1* (high-level drug class, such as antibiotic and anti-inflammatory), *Parent level 2* (specific drug class, such as aminoglycoside, third generation cephalosporin, and beta-blocker), *Parent level 3* (main drug ingredient, such as furosemide and warfarin), and a combination of these three levels (*merged parent level*). The latter was obtained by identifying the most specific level that had a minimum number of five cases for training. The goal was to achieve a balance between amount of information and data sparseness.
2. Attributes that conveyed the same information from different sources were merged into a third attribute. For example, user *discipline* and *specialty* were obtained from two data sources in the EDW: the HELP2 audit data mart and the human resources data mart. The former is more complete, though not always accurate, while the latter is very accurate, but only contains information about employed physicians and registered nurses. The human resources data mart was used as a master and the HELP2 audit data mart as a source for missing values.
3. Numeric attributes, such as *years of clinical practice*, were discretized with the Fayyad and Irani algorithm [34]. The *patient age* attribute was also discretized according to age categories in the MeSH (Medical Subject Headings) code system[2].

After the data cleaning and preparation steps were completed, the final dataset was exported to a comma delimited file and then converted to Weka's file format.

D. Attribute selection

In this step, we followed a similar process to the one proposed by Hall and Holmes [15]. First, attribute ranking algorithms (Information gain and Relief) were executed to obtain a baseline and exclude attributes that were clearly useless. Second, three attribute subset evaluation algorithms (CFS, Consistency, and Wrapper) were executed to obtain optimal attribute subsets. Next, the attribute subsets were used as inputs for training classification models based on nine learning algorithms. Finally, the performance of the attribute sets was compared with a baseline that contained all attributes in the original data source. The performance in this step was measured by 10-fold cross-validation using two thirds of the original dataset. When no statistical difference was found between two or more algorithms, the method that produced the smallest attribute set was selected as the optimal choice for a given learning method.

E. Classification

Nine learning methods were used in this study: Naïve Bayes, rules (PART algorithm [35]), decision tree (C4.5 algorithm [36]), boosted Naïve Bayes, boosted rules, boosted decision tree,

[1] National Drug Data File™, First Data Bank, Inc., San Bruno, CA,http://www.firstdatabank.com
[2]Medical Subject Headings. National Library of Medicine, Bethesda, MD, http://www.nlm.nih.gov/mesh/

Bayesian network, Support Vector Machine (SVM) (Platt's sequential minimal optimization algorithm [37]), and Stacking (StackingC algorithm [29]). The MultiBoost algorithm developed by Webb was used for Boosting [28]. These methods were chosen for being state-of-the-art examples that encompass a variety of machine learning techniques [21]. Boosting was not coupled with Bayesian network and SVM because these two techniques tend to produce stable models that are not benefited by Boosting. For the StackingC model, we used a variety of base learning methods that demonstrated good performance in ad-hoc experiments conducted prior to this study: Naïve Bayes, boosted rules, boosted decision tree, SVM, and Bayesian network.

Classification models were trained with two thirds of the original dataset. The remaining one third of the original data set was set aside for testing. Ten test sets were then obtained from this original test set by randomly sampling cases with replacement (bootstrap) until each new test set was 80% of the size of the original test set.

F. Measurements

Performance in the attribute selection and classification experiments was measured in terms of agreement (kappa) between the output of each classifier and the topics that the users had actually selected. Classification performances by topic were measured in terms of the area under the ROC curve (AUC). Statistical differences among multiple algorithms were verified using the Friedman's test. If a significant difference was found, multiple pair wise comparisons were made. The Nemenyi post-hoc test was used for adjustment of multiple comparisons as recommended by Demšar [38].

III. RESULTS

A. Attribute selection

Attribute ranking indicated that the five strongest individual predictors were *avg reads*, *orders entered*, *avg writes*, *patient age*, and *parent level 3*. The *interaction count* attribute was ranked last by both attribute ranking methods. Table 3 lists the individual attribute scores according to the information gain and Relief attribute ranking algorithms. None of the attribute subset evaluation methods significantly improved the performance of the classifiers over the baseline. However, CFS was considered the optimal method in most cases because it produced the smallest attribute subset (six attributes; best in five learning methods), followed by Consistency (10 attributes; best in three learning methods), and Wrapper (average of 13 attributes; best in one learning method). Table 4 lists the attributes that each of the three attribute subset evaluation methods identified. Table 5 lists the best attribute subset evaluation methods by learning algorithm.

B. Classifier performance

The performance of the classifiers showed an overall moderate level of agreement, with average kappa scores ranging between 0.47 (rules) and 0.56 (Stacking). Table 6 lists the performance of the nine learning methods overall (kappa) and at predicting individual topics (AUC). Table 7 shows pair wise comparisons between the nine learning methods in terms of kappa. Stacking, SVM, and Bayesian network were the best methods overall. Although there was no statistical difference among the Stacking, Bayesian network, and SVM classifiers, Stacking outperformed the other two competitors in all 10 bootstrapped test sets. With the exception of Naïve Bayes, the boosted algorithms were slightly superior to their non-boosted counterparts, but the difference was not statistically significant.

The learning methods showed varied performance levels regarding the prediction of each individual class. Overall, the AUC for *pediatric dose*, *patient education*, *pregnancy*

category, and *adult dose* was high. Conversely, the AUC for *drug interactions*, *adverse effects*, *how supplied*, and *precautions* was not as good, but still acceptable. Stacking ranked among the highest methods at predicting every one of the topics.

IV. DISCUSSION

This study supplements previous work related to the construction of classification models based on usage data to predict clinicians' information needs [13],[14]. The proposed method can be used to develop classification models that can be integrated into existing Infobutton Manager implementations, potentially improving the effectiveness of infobuttons. For example, users could be taken automatically to the content topic that a classification model predicts to be the most relevant in a particular context. Information needs prediction models will also enable different approaches to the delivery of context-sensitive information in EMR systems. For example, succinct information on candidate topics (i.e., the ones that the models predict to be the most relevant in a given context) can be more easily accessible via a keyboard shortcut or dynamically displayed in a sidebar as medication orders are entered, reviewed, or refilled. In all these alternatives, the goal is to present the minimal amount of information to support quick decisions, reducing unnecessary navigation steps and exposure to irrelevant information [39].

A. Attribute selection

The attribute ranking algorithms were, in general, consistent with the conclusions of the attribute subset evaluation algorithms. For example, the five strongest attributes according to the Relief and Information Gain algorithms (i.e., *avg reads*, *orders entered*, *avg writes*, *patient age*, and *parent level 3*) were also among the subsets identified by the CFS, Consistency, and Wrapper attribute subset evaluation algorithms. This confirms that attribute ranking methods may provide a useful baseline ranking to guide the next steps in the attribute selection process, for example allowing the elimination of useless attributes.

Unlike the benchmark study conducted by Hall and Holmes [15], in our study none of the attribute subset evaluation methods significantly improved the performance of the models over the baseline (i.e., all attributes). A potential explanation for the lack of detectable differences is that the attributes used in our study were hand selected based on domain knowledge, so that the initial attribute set was already close to optimal. Conversely, Hall and Holmes used datasets available in the University of California Irvine repository, which is a standard for machine learning benchmark studies [15]. Despite the lack of significant performance improvement, attribute subset evaluation methods, notably CFS, eliminated redundant and useless attributes, producing more compact models.

CFS was the best attribute subset evaluation algorithm overall, contradicting the Hall and Holmes study where Wrapper was the best method. CFS is particularly good at removing redundant attributes, and our dataset included several attributes that tried to capture the same type of information using different data sources or semantic levels (e.g., *discipline*, *specialty*, multiple *drug parent levels*). Therefore, CFS may have been able to identify the strongest among each set of redundant attributes. Our study confirmed other known advantages of CFS: it generally produces more compact models, it is much faster to execute, and its results are independent of the target learning method [15]. Nevertheless, Consistency was the best method for rules, boosted rules and boosted decision tree, and Wrapper was best for decision tree. Therefore, a comparison of attribute subset evaluation methods is important in future experiments or applications that deal with different learning algorithms and infobutton usage data from other sources.

The CFS method identified a combination of strong predictors that characterize the user (i.e., *avg reads*, *avg writes*, *orders entered*), the patient (i.e., *age*), and the medication associated

with the infobutton (i.e., *parent level 3*). This confirms the belief that context influences information needs and shows that context can be portrayed in multiple dimensions.

Although the results do not imply causal relationships among attributes and the topics that clinicians decide to view, potential explanations for the associations found can be proposed. The *avg reads*, *avg writes*, and *orders entered* attributes are indicators of the volume and nature of EMR use. Intermountain clinicians are more likely to use the HELP2 Clinical Desktop in the outpatient than in the inpatient setting. Therefore, it is expected that outpatient clinicians have higher values for *avg reads* and *avg writes*. The nature of the care process, patient characteristics, and EMR use in the outpatient setting differs from the inpatient; as a result, it is likely that different information needs will arise in these two settings as well. In addition, clinicians who enter medication orders have higher values for the *orders entered* variable than those who only read data from the EMR. The medication order entry process probably leads to different information needs than read-only consultation about medications that have been ordered in the past. In summary, the volume of EMR write and read activity might serve as a surrogate for the types of activities and roles that a clinician routinely performs. This surrogate seems to be more accurate and specific than other attributes that describe users, such as specialty and discipline.

Patient age is likely to be a strong determinant of the type of dose information that is requested. In fact, further inspection of the dataset revealed that the majority of the infobutton sessions associated with pediatric patients were related to *pediatric dose*, while other topics were seldom viewed.

Finally, it is reasonable to assume that the characteristics of a medication influence the nature of information needs. For example, medications that are frequently prescribed may lead to different questions than those that are rarely used. Similarly, medications that are associated with various adverse effects may be more likely to raise questions about adverse effects. *Parent level 3* was the most specific drug classification level that was available in the dataset. Other less specific levels were not as useful in predicting the types of information needs associated with the medication.

B. Classifier performance

The prediction performance of the learning methods evaluated in this study was overall very good. In general, topics that had more cases available for training were associated with better performance. According to the AUC metric, the best performance was achieved with the *adult dose*, *pediatric dose*, *patient education*, and *pregnancy category* topics. Moderate levels were obtained with *adverse effects* and *drug interactions*. *Precautions* and *how supplied*, which accounted for the minority of the cases in the training set, had the worst performance among the possible topics. Further research is necessary, perhaps using a larger dataset, to improve the prediction performance of these least frequent topics. As more usage data become available, models could also enable the prediction of topics that were not included in the study dataset, such as *drug class*, *breast feeding category*, and *contraindications*. Nevertheless, models that predict the topics that are most frequently viewed are likely to be sufficient for integration with an operational environment.

Stacking, SVM, and Bayesian network were the best methods overall in terms of agreement with the users' actual needs. Nevertheless, Stacking provided more uniform results, always ranking among the top three methods (with regard to AUC) for each topic. This is consistent with previous studies where Stacking overcame the individual performance of its base classifiers [23],[27]. Boosted algorithms did not significantly improve the performance of their non-boosted counterparts, perhaps because the base algorithms used in this study (i.e., Naïve

Bayes, C4.5, and PART) are strong learners themselves, unlike the ones used in previous comparisons [23],[26].

C. Limitations

Due to specific characteristics of the environment, the EMR system, and the data available at our institution, it is possible that the attribute sets and classification models developed in this study will not generalize to other institutions. However, this study provides a guide to other institutions regarding the subset of attributes and learning methods that can be evaluated and used in their settings.

D. Future studies

Classification models might improve the ability of infobuttons to present the most relevant information to clinicians, potentially resulting in time savings, less cognitive effort, and higher success at meeting information needs. Nevertheless, further research is warranted to improve the prediction performance of content topics that are less commonly viewed. Alternatives to be pursued include capturing larger training datasets, investigation of attributes that were not used in this study (e.g., care setting, a more detailed description of the task being performed), and the assessment of other more recent meta-learning techniques that have shown potential to outperform Stacking [40].

This study focused on medication-related infobutton sessions. Further research is necessary to develop classification models to predict content topics that are needed in different EMR contexts, such as problem lists, laboratory results, and clinical notes.

As a next step, we are now incorporating classification models in the Infobutton Manager component at Intermountain Healthcare. This process is raising new research questions, such as usability issues and long term updating and evaluation of the model.

V. CONCLUSION

This study supports the hypothesis that prediction models based on infobutton usage data are a promising solution for predicting the information needs that a clinician might have in a particular context while using an EMR system. The information needs are strongly affected by the characteristics of users, patients, and medications. The models here evaluated had a good overall performance, especially at predicting the *adult dose*, *pediatric dose*, *patient education*, and *pregnancy category* topics. Further research is warranted to improve the performance of other topics that are less commonly viewed. Stacking was the best method overall, but other methods, such as SVM, Bayesian network, and Naïve Bayes also performed well. Finally, the proposed method consists of a sound alternative for the enhancement of Infobutton Managers, helping clinicians fulfill their information needs at the point of care.

Acknowledgements

The authors would like to acknowledge Joyce A. Mitchell, Scott P. Narus, James J. Cimino, and Chuck Norlin, for carefully reviewing the manuscript and for providing valuable comments and suggestions on the study design and results.

This project was supported in part by National Library of Medicine Training Grant 1T15LM007124–10 and National Library of Medicine grant R01-LM07593.

References

1. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med 1985;103:596–9. [PubMed: 4037559]

2. Gorman P. Information needs in primary care: a survey of rural and nonrural primary care physicians. Medinfo 2001;10:338–42.

3. Osheroff SA, Forsythe DE, Buchanan BG, Bankowitz RA, Blumenfeld BH, Miller RA. Physicians' information needs: analysis of questions posed during clinical teaching. Ann Intern Med 1991;114:576–81. [PubMed: 2001091]

4. Norlin C, Sharp AL, Firth SD. Unanswered questions prompted during pediatric primary care visits. Ambul Pediatr 2007;7(5):396–400. [PubMed: 17870649]

5. Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. J Am Med Inform Assoc 2005;12:217–24. [PubMed: 15561792]

6. Magrabi F, Coiera EW, Westbrook JI, Gosling AS, Vickland V. General practitioners' use of online evidence during consultations. Int J Med Inform 2005;74:1–12. [PubMed: 15626631]

7. Westbrook JI, Coiera EW, Gosling AS. Do online information retrieval systems help experienced clinicians answer clinical questions? J Am Med Inform Assoc 2005;12:315–21. [PubMed: 15684126]

8. Forsythe DE, Buchanan BG, Osheroff JA, Miller RA. Expanding the concept of medical information: an observational study of physicians' information needs. Comput Biomed Res 1992;25(2):181–200. [PubMed: 1582194]

9. Lomax EC, Lowe HJ. Information Needs Research in the Era of the Digital Medical Library. Proc AMIA Annu Fall Symp 1998:658–62.

10. Cimino JJ, Elhanan G, Zeng Q. Supporting Infobuttons with terminological knowledge. Proc AMIA Annu Fall Symp 1997:528–32. [PubMed: 9357682]

11. Cimino, JJ.; Del Fiol, G. Infobuttons and Point of Care Access to Knowledge. In: Greenes, RA., editor. Clinical decision support: the road ahead. Academic Press; 2006.

12. Cimino JJ, Li J, Bakken S, Patel VL. Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users. Proc AMIA Annu Fall Symp 2002:170–4.

13. Del Fiol G, Haug PJ. Use of classification models based on usage data for the selection of infobutton resources. Proc AMIA Annu Fall Symp 2007:171–175.

14. Del Fiol G, Haug PJ. Infobuttons and classification models: a method for the automatic selection of on-line information resources to fulfill clinicians' information needs. J Biomed Inform. 2007 Dec 8; [Epub ahead of print]

15. Hall MA, Holmes G. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Transactions on Knowledge and Data Engineering 2003;15:1–16.

16. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 2003;3:1157–82.

17. Kira K, Rendell L. A practical approach to feature selection. Proceedings of the Ninth International Conference on Machine Learning 1992:249–256.

18. Hall MA. Correlation-based feature selection for discrete and numeric class machine learning. Proceedings of the 17th International Conference on Machine Learning 2000:359–366.

19. Liu H, Setiono R. A probabilistic approach to feature selection: A filter solution. Proceedings of the 13th International Conference on Machine Learning 1996:319–327.

20. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997;97:273–324.

21. Witten, IH.; Frank, E. Data Mining: Practical machine learning tools and techniques. Vol. 2. Morgan Kaufmann; San Francisco: 2005.

22. King R, Feng C, Shutherland A. Statlog: comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence 1995;9(3):259–287.

23. Caruana R, Niculescu-Mizil A. An Empirical Comparison of Supervised Learning Algorithms. Proceedings of the 23rd International Conference on Machine Learning 2006;148:161–168.

24. Chan PK, Stolfo SJ. On the accuracy of meta-learning for scalable data mining. Journal of Intelligent Information Systems 1997;8(1):5–28.

25. Vilalta R, Drissi Y. A Perspective view and survey of meta-learning. Artificial Intelligence Review 2002;18(2):77–95.

26. Schapire RE. The strength of weak learnability. Machine Learning 1990;5(2):197–227.

27. Wolpert DH. Stacked generalization. Neural Networks 1992;5:241–260.
28. Webb GI. MultiBoosting: A technique for combining boosting and wagging. Machine Learning 2000;40(2):159–196.
29. Seewald A. How to make Stacking better and faster while also taking care of an unknown weakness. Proceedings of the 19th International Conference on Machine Learning 2002:554–561.
30. Clayton PD, Narus SP, Huff SM, Pryor TA, Haug PJ, Larkin T, et al. Building a comprehensive clinical information system from components. The approach at Intermountain Health Care. Methods Inf Med 2003;42:1–7. [PubMed: 12695790]
31. Del Fiol G, Rocha R, Clayton PD. Infobuttons at Intermountain Healthcare: Utilization and Infrastructure. Proc AMIA Annu Fall Symp 2006:180–4.
32. Maviglia SM, Yoon CS, Bates DW, Kuperman G. KnowledgeLink: impact of context-sensitive information retrieval on clinicians' information needs. J Am Med Inform Assoc 2006 Jan-Feb;13(1): 67–73. [PubMed: 16221942]
33. Cimino JJ. Use, usability, usefulness, and impact of an infobutton manager. Proc AMIA Annu Fall Symp 2006:151–5.
34. Fayyad UM, Irani KB. Multi-interval discretization of continuous-valued attributes. Proceedings of the International Joint Conference on Uncertainty in AI 1993:1022–1027.
35. Frank E, Witten IH. Generating accurate rule sets without global optimization. 15th International Conference on Machine Learning 1998:144–151.
36. Quinlan, R. C4.5: Programs for machine learning. Morgan Kaufmann Publishers; San Mateo, CA: 1993.
37. Platt, J. Fast training of support vector machines using sequential minimal optimization. In: Schöelkopf, B.; Burges, C.; Smola, A., editors. Advances in kernel methods - support vector learning. MIT Press; 1998.
38. Demšar J. Statistical comparison of classifiers over multiple data sets. Journal of Machine Learning Research 2006;7:1–30.
39. Weir CR, Nebeker JJ, Hicken BL, Campo R, Drews F, Lebar B. A cognitive task analysis of information management strategies in a computerized provider order entry environment. J Am Med Inform Assoc 2007;14(1):65–75. [PubMed: 17068345]
40. Caruana R, Niculescu-Mizil A, Crew G, Ksikes A. Ensemble selection from libraries of models. Proceedings of the 21st international conference on machine learning 2004;69:18–25.

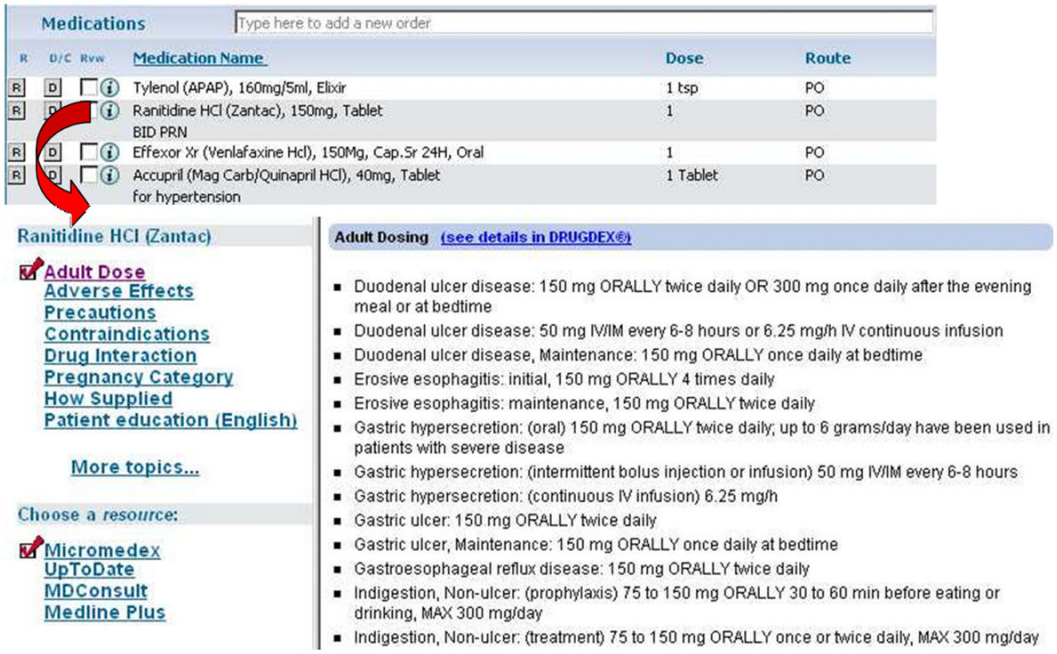

Figure 1.

A screen shot of the medication order entry module in the HELP2 Clinical Desktop system (above). Infobuttons are available next to each of the medications in the patient's medications list. When an infobutton is clicked, an infobutton navigation panel and a content page are displayed (below). The navigation panel offers a list of relevant content topics (e.g, adult dose, adverse effects, patient education) and resources that users can choose from (lower left).

Table 1

Automated attribute selection methods used in the machine learning experiments.

Algorithm	Description	Advantages	Disadvantages
Information gain	Each attribute in a dataset is assigned a score based on the additional information that an attribute provides regarding a class in terms of entropy reduction.	Simple and fast. Good for prediction problems where the high dimensionality limits the application of more sophisticated methods.	Does not account for redundancy and interactions among attributes.
Relief [17]	Randomly samples an instance from the data and then locates its nearest neighbor from the same and opposite class. The values of the attributes of the nearest neighbors are compared to the sampled instance and used to update relevance scores for each attribute.	Same as information gain.	Same as information gain.
Correlation- based feature selection (CFS) [18]	Merit of a given attribute is calculated taking into account the correlation of the attribute with the target class as well as the correlation of the attribute with other attributes in the dataset. Attributes with stronger correlation with the target class and weaker correlation with other attributes are ranked higher.	Fast and independent of the target learning method. Accounts for redundancy.	Does not account for potential interactions between attributes.
Consistency- based [19]	Identifies attribute sets whose values divide the data into subsets containing a strong single class majority.	Independent of the target learning method. Accounts for redundancy and interactions among attributes.	Slower than **correlation-based feature selection.**
Wrappers[20]	Uses a target learning algorithm to estimate the worth of attribute subsets. A search algorithm is used to test as many combinations of attributes as possible and find an optimal solution.	Accounts for redundancy and interactions among attributes. Generally give better results than other techniques because candidate solutions are evaluated using the target learning algorithm.	Specific to the learning algorithm that is used to evaluate the worth of the subsets (has to be rerun for each learning algorithm). Slower than the other methods, precluding its application to datasets with high dimensionality and slow learning algorithms.

Table 2

Attributes used in the machine learning experiments.

Attribute name	Description	Source
User		
Avg reads	Average number of data read events requested monthly from the HELP2 clinical data repository.	EDW (HELP2 audit data mart)
Avg writes	Average number of data write events requested monthly from HELP2 clinical data repository.	EDW (HELP2 audit data mart)
Orders entered	Average number of medication orders entered monthly in the HELP2 clinical data repository.	EDW (HELP2 audit data mart)
HELP2 discipline	User's discipline (e.g., physician, registered nurse) according to the HELP2 audit data mart.	EDW (HELP2 audit data mart)
HELP2 specialty	User's specialty (e.g., pediatrics, internal medicine) according to the HELP2 audit data mart.	EDW (HELP2 audit data mart)
HR discipline	User's discipline (e.g., physician, registered nurse) according to the human resources data mart.	EDW (human resources data mart)
HR specialty	User's specialty (e.g., pediatrics, internal medicine) according to the human resources data mart.	EDW (human resources data mart)
Merged discipline	A combination of *HELP2 discipline* and *HR discipline*. *HR discipline*'s missing values completed with *HELP2 discipline*'s data.	EDW (HELP2 audit and human resources data marts)
Merged specialty	A combination of *HELP2 specialty* and *HR specialty*. *HR specialty*'s missing values completed with *HELP2 specialty*'s data.	EDW (HELP2 audit and human resources data marts)
Years of practice	Number of years of clinical practice. A measurement of a clinician's experience.	
Search concept		
Parent level 1	High level medication class (e.g., antibiotics, hypotensives, psychoactive drugs).	Terminology server (First Data Bank)
Parent level 2	Specific medication class (e.g., third generation cephalosporins, beta-blockers, selective serotonin reuptake inhibitors).	Terminology server (First Data Bank)
Parent level 3	Main drug ingredient (e.g., ceftriaxone, propranolol, sertraline).	Terminology server (First Data Bank)
Merged parent level	A combination of the different parent levels seeking a trade off between amount of information (*parent level 3* contains more information) and data sparseness(*parent level 1* is the least sparse). (see Section II. C.: Data Cleaning and Preparation for details)	Terminology server (First Data Bank)
DEA class code	Degree of potential abuse and federal control of a drug.	First Data Bank
Drug class code	Availability of a drug to the consumer(over the counter vs. prescription required).	First Data Bank
AHFS class	Primary drug therapeutic class according to the American Hospital Formulary Service, which is maintained by the American Society of Health System Pharmacists (ASHP).	First Data Bank
Maintenance drug	A flag indicating whether a drug is used chronically or not.	First Data Bank
Interaction count	Number of severe drug interaction rules that a given drug participates on. Provides a measurement of the likelihood that a given drug will interact with others.	First Data Bank
Patient		
Age	Numeric patient age.	Clinical data repository data mart
Age group	Age according to Medical Subject Headings (MeSH).	Clinical data repository data mart
Gender	Patient's gender.	Clinical data repository data mart
Medications	Number of active medications in the patient's medications list.	Clinical data repository data mart
Problems	Number of active problems in the patient's problems list.	Clinical data repository data mart
Other		

Attribute name	Description	Source
Task	Action that the user performs in HELP2 when decided to click on an infobutton. Possible values are “medication order entry” and “medications list.”	Infobutton Manager log
Topic	First content topic that the user selected to view in a given infobutton session. The dataset contained eight possible topics that users could have selected in an infobutton session: adult dose, pediatric dose, drug interactions, adverse effects, patient education, pregnancy category, how supplied, and precautions.	Infobutton Manager log

Table 3

Attribute ranking scores according to the information gain and Relief individual attribute ranking algorithms. The lower the score, the stronger the attribute.

	Information gain	Relief
Avg reads	3	1.3
Orders entered	4.1	1.9
Avg writes	5.5	2.8
Age	10	4
Parent level 3	1	15.7
Merged parent level	5.4	10.5
Parent level 2	7	9.4
HR specialty	12	5
Age group	11	6.4
AHFS class	8	13.1
Parent level 1	13	10.7
Years of practice	17	7
HR discipline	14.9	11.3
HELP2 discipline	16	14.6
Merged specialty	14.1	17
Merged discipline	18.1	18.9
HELP2 specialty	19.4	18.5
Medications	19.5	21.4
Problems	22.7	19.6
Maintenance drug	21.9	21.6
Gender	21.4	23
DEA class code	24	24
Drug class code	25	26
Task	26	25

Table 4

Attributes selected by each of the attribute subset evaluation algorithms. Attributes that are not listed were not included in any of the optimal attribute subsets.

Attribute	CFS	Consistency	Wrapper (decision tree)
Avg reads	X	X	X
Avg writes	X	X	X
Orders entered	X		
HELP2 discipline			X
HELP2 specialty		X	
HR specialty			X
Parent level 3	X	X	
DEA class code			X
Drug class code			X
AHFS class		X	
Maintenance drug		X	X
Age	X		X
Age group	X	X	X
Gender		X	X
Medications		X	X
Problems		X	
Task			X

Table 5

Optimal attribute selection method per learning algorithm.

	Attribute selection method	Number of attributes
Naïve Bayes	CFS	6
Rules	CFS	6
Decision tree	Consistency	10
Boosted Naïve	Consistency	10
Boosted rules	Wrapper	7
Boosted tree	Consistency	10
Bayesian network	CFS	6
SVM	CFS	6
Stacking	CFS	6

Table 6

Overall performance of learning algorithms (kappa) and by topic (AUC).

	Kappa	AD	PD	PE	PC	AE	DI	HS	PR
NB	0.50	0.83	0.99	0.95	0.94	0.79	0.80	0.64	0.73
Rules	0.47	0.8	0.89	0.86	0.73	0.70	0.74	0.48	0.61
Tree	0.50	0.8	0.97	0.90	0.88	0.72	0.78	0.61	0.68
NB+	0.51	0.81	0.99	0.91	0.85	0.71	0.69	0.60	0.70
Rules+	0.49	0.81	0.97	0.91	0.68	0.70	0.68	0.74	0.57
Tree+	0.50	0.81	0.98	0.91	0.75	0.70	0.65	0.59	0.60
BN	0.52	0.86	0.99	0.96	0.91	0.78	0.83	0.58	0.75
SVM	0.54	0.76	0.99	0.92	0.92	0.79	0.71	0.72	0.71
Stacking	0.56	0.85	0.99	0.95	0.93	0.78	0.79	0.73	0.75

Legend: NB = Naïve Bayes; NB+ = Boosted Naïve Bayes; Rules+ = Boosted rules; Tree+ = Boosted decision tree; BN = Bayesian network; SVM = Support Vector Machine; AD = adult dose; PD = pediatric dose; PE = patient education; PC = pregnancy category; AE = adverse effects; DI = drug interactions; HS = how supplied; PR = precautions.

Table 7

Pair wise comparison between learning algorithms according to kappa. The numbers in the cells measure the difference between each pair wise comparison according to the Nemenyi test (significance is achieved when the absolute value of the score is at least 3.8). Positive numbers denote that the classifier in the row was better than the one in the column.

	NB	Rules	Tree	NB+	Rules+	Tree+	BN	SVM	Stacking
NB	-	2.1	0.1	−1.7	0.3	−0.8	−3.1	−4.1*	−5.4*
Rules	-	-	−2	−3.8*	−1.8	−2.9	−5.2*	−6.2*	−7.5*
Tree	-	-	-	−1.8	0.2	−0.9	−3.2	−4.2*	−5.5*
NB+	-	-	-	-	2	0.9	−1.4	−2.4	−3.7
Rules+	-	-	-	-	-	−1.1	−3.4	−4.4*	−5.7*
Tree+	-	-	-	-	-	-	−2.3	−3.3	−4.6*
BN	-	-	-	-	-	-	-	−1	−2.3
SVM	-	-	-	-	-	-	-	-	−1.3
Stacking	-	-	-	-	-	-	-	-	-

*Statistically significant.