Offline Testing of the ATHENA Hypertension Decision Support System Knowledge Base to Improve the Accuracy of Recommendations

S.B. Martins, MD, MSc[1], S. Lai, MD[2], S. Tu, MS[3], R. Shankar, MS[3], S.N. Hastings, MD[4], B.B. Hoffman, MD[5], N. DiPilla, M,B.B.S[1] and M.K. Goldstein, MD[1,3]

[1]GRECC,VA Palo Alto Health Care System, Palo Alto, CA; [2]Santa Clara Valley Medical Center, San Jose, CA; [3]Department of Medicine, Stanford University School of Medicine, Stanford, CA; [4]Durham VA Medical Center, Durham, NC; and [5]VA Boston Health Care System, West Roxbury, MA

Abstract

ATHENA-HTN is a clinical decision support system (CDSS) that delivers guideline-based patient-specific recommendations about hypertension management at the time of clinical decision-making. The ATHENA-HTN knowledge is stored in a knowledge-base (KB). Changes in best-practice recommendations require updates to the KB. We describe a method of offline testing to evaluate the accuracy of recommendations generated from the KB. A physician reviewed 100 test cases and made drug recommendations based on guidelines and the "Rules" (descriptions of encoded knowledge). These drug recommendations were compared to those generated by ATHENA-HTN. Nineteen drug-recommendation discrepancies were identified: ATHENA-HTN was more complete in generating recommendations (15); ambiguities in the Rules misled the physician (3); and content in the Rules was not encoded (1). Three new boundaries were identified. Three updates were made to the KB based on the results. The offline testing method was successful in identifying areas for KB improvement and led to improved accuracy of guideline-based recommendations.

Introduction

Evidence-based clinical-practice guidelines (CPGs) provide busy clinicians with recommendations based on expert evaluation of the medical literature. Clinical decision-support systems (CDSS) encoding CPGs, integrated with patient data from an electronic medical record, can facilitate the translation of research into practice by providing clinicians with guideline-based patient-specific recommendations[1]. Developers of new or updated CDSS's must ensure that the system generates correct recommendations[1].

We developed a knowledge-based CDSS, formerly ATHENA DSS, now known as ATHENA-Hypertension (HTN). ATHENA-HTN, built with the EON architecture developed at Stanford Medical Informatics, includes a knowledge base (KB) that models medical knowledge about managing primary hypertension and a guideline interpreter that serves as execution engine[2]. The KB is modeled in Protégé software[3]. ATHENA-HTN is designed to process individual patients' clinical data from an electronic source against hypertension knowledge in the KB to generate patient-specific recommendations that are displayed at point of care (clinical visit) for clinical management.

Evaluation of the accuracy of recommendations generated by a CDSS includes several steps at different stages of development and deployment of the system (Figure 1). In this paper, we focus on testing the KB and the execution engine to ensure that the recommendations generated are faithful to the clinical practice guidelines we were encoding.

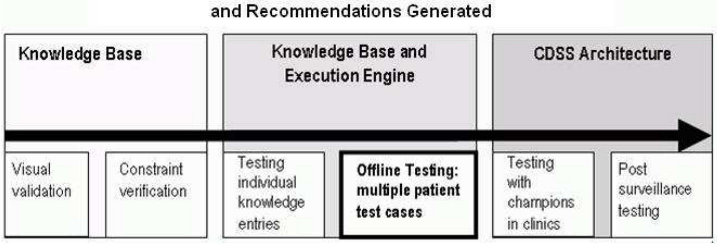
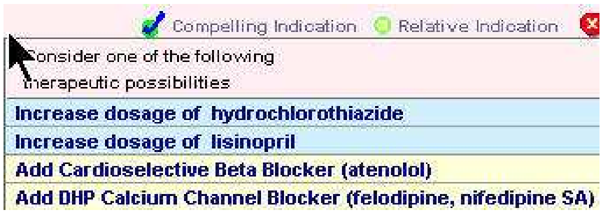

Phases in Testing CDSS Accuracy of the Knowledge Encoded and Recommendations Generated
Knowledge Base
Knowledge Base and Execution Engine
CDSS Architecture
Visual validation
Constraint verification
Testing individual knowledge entries
Offline Testing: multiple patient test cases
Testing with champions in clinics
Post surveillance testing

Figure1: Testing accuracy of CDSS recommendations

Newly developed or updated software will contain errors[4]. Myers defines testing as "the process of executing a program with the intent of finding errors"[4]. Successful testing involves detecting errors and addressing them. The choice of method for software testing depends on the development phase and the objectives. Regression testing is a standard method used in software engineering to evaluate changes and updates for errors. This method recognizes that updates and changes are particularly error prone and may introduce errors (regression) into a previously well-functioning system. A set of test cases with known correct output can be run in the updated system to determine whether the updates have regressed the software. As a first step toward generating the set of test cases with known correct output, a domain expert reviews sets of sample data and indicates the correct system output. The domain expert's outputs are then compared to the software system's outputs, and discrepancies are investigated and resolved[5].

In designing a CDSS, developers make decisions about the knowledge content and about the consistency, depth and coverage of the knowledge encoded in the system[5]. Each of these areas is pertinent to testing. For example, testers must consider the boundaries of the KB and ensure that the system "fails gracefully" at the boundaries[6].

The knowledge encoded in the version of the ATHENA-HTN KB addressed here was based on the Sixth Report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure (JNC6)[7] and the VA Diagnosis and Management of Hypertension in the Primary Care Setting[8]. Sample recommendations are shown in Figure 2. As new evidence emerges about best clinical practices from clinical trials or guidelines, it is important to update the KB of a CDSS. Once updated, the KB requires testing to ensure continued accuracy of recommendations: not only to ensure that the newly added material is correct, but also to ensure that unanticipated errors were not introduced.



Consider one of the following therapeutic possibilities

Increase dosage of hydrochlorothiazide

Increase dosage of lisinopril

Add Cardioselective Beta Blocker (atenolol)

Add DHP Calcium Channel Blocker (felodipine, nifedipine SA)

Figure 2. Partial view of drug recommendations in ATHENA-HTN

Some evaluations of clinical decision support systems have focused on comparing the generated recommendations with actual practice, with an implicit assumption that the recommendations generated by the system are consistent with the guidelines but without presenting testing data to document the accuracy[9,10].. In this paper, we describe our method of offline testing , applied after an update of the KB and prior to clinical deployment of the updated system.

Methods

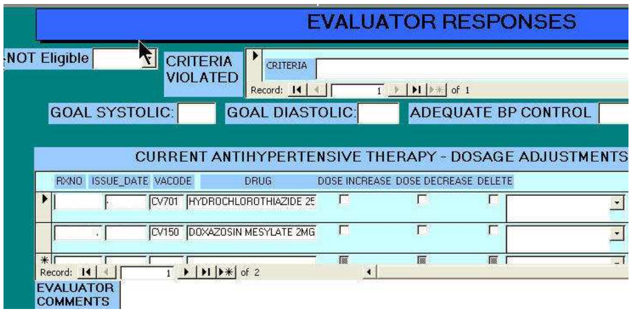
A physician reviewed 100 cases abstracted from real patient data and made drug recommendations for hypertension management. These drug recommendations were compared with those from ATHENA-HTN for the same cases, looking for errors and for program behavior at the boundaries, to identify areas for correction or improvement. The comparison of physician recommendations with ATHENA-HTN recommendation allowed a cross-check on both.

Case Selection – We used the set of 100 test cases from the original offline testing of the CDSS that was completed prior to initial deployment. These anonymized cases were randomly selected from a data set of 1000 consecutive patients with a hypertension diagnosis in the VA Palo Alto electronic medical records. The cases were stratified into five clinical categories: diabetes mellitus; coronary artery disease; heart failure; conjunction of diabetes mellitus with heart failure; and hypertension with none of these co-morbidities. The clinical categories were based on a regional VA performance guideline in use at that time. Diagnoses were based on ICD 9 codes.

Rules Document – In designing the original ATHENA-DSS and in updating it with new clinical information, we maintained a document (Rules) containing a narrative form of the information we had attempted to encode in the KB. Tierney et al have described a challenge in automating guidelines, due to the presence of ambiguity and the lack of explicit definitions in the published guidelines[11]. To minimize ambiguity in specifying the knowledge we were encoding, we wrote the Rules with more explicit definitions than the published guidelines.

Physician Choice – A general internist (SL) who had no previous involvement in the development of ATHENA-HTN or the KB update reviewed test cases. The physician extensively reviewed the Rules so that he was familiar with the knowledge encoded in the KB. The Rules and the original guidelines were available to the physician for further reference while reviewing test cases.

Information Presentation – Patient information was presented to the physician in an MS Access© evaluation form detailing patient data available to the CDSS (Figure 3): diagnosis, medications, allergy/ adverse drug reactions, most recent blood pressures and selected laboratory values (sodium, potassium, creatinine, lipid profile and urinary protein).



EVALUATOR RESPONSES

NOT Eligible

CRITERIA VIOLATED

CRITERIA

Record: 1 of 1

GOAL SYSTOLIC:

GOAL DIASTOLIC:

ADEQUATE BP CONTROL

CURRENT ANTIHYPERTENSIVE THERAPY - DOSAGE ADJUSTMENTS

RXNO	ISSUE_DATE	VACODE	DRUG	DOSE INCREASE	DOSE DECREASE	DELETE
		CV701	HYDROCHLOROTHIAZIDE 25	□	□	□
		CV150	DOXAZOSIN MESYLATE 2MG	□	□	□

Record: 1 of 2

EVALUATOR COMMENTS

Figure 3. Partial view of physician evaluation form

Physician Review of Cases – The physician reviewed the 100 test cases. He identified which patients met ATHENA-HTN eligibility criteria (for example, primary rather than secondary hypertension) and made recommendations about anti-hypertensive drug therapy based on the Rules.

Comparison of Drug Recommendations – We compared the drug recommendations made by the physician with those generated by ATHENA-HTN, using a custom-designed comparison program followed by a manual re-check. A team of physicians on the development team (SM, MG) and the evaluator physician (SL) reviewed and characterized each discrepancy between drug recommendations (by the physician and by ATHENA-HTN) by consensus.

Results

Eligibility for ATHENA-HTN – The physician and ATHENA-HTN agreed on the exclusion of six test cases. The physician correctly excluded one further test case that met exclusion criteria (hyperplasia of the renal artery leading to secondary hypertension), a diagnostic code that was missing from the KB. ATHENA-HTN correctly excluded one case (malignant hypertension) that met criteria for exclusion but was not specified in the Rules. The remaining 92 cases were used for assessment of drug recommendations.

Drug Recommendations – For the 92 eligible test cases, ATHENA-HTN made 181 drug recommendations with an average of two recommendations per test case (median 2, range 0-5). The physician made 184 drug recommendations with an average of two recommendations per test case (median 2, range 0-5).

There were 27 drug recommendation discrepancies between physician and ATHENA-HTN. Eight discrepancies originated from unclear presentation of drug information to the physician in the evaluation form: the pharmacy text for instructions on how to take the medication (commonly known as SIG) included non-standard acronyms understood by the pharmacy but misunderstood by the physician. The remaining 19 true discrepancies between ATHENA-HTN and the physician were characterized as follows:

ATHENA-HTN was more comprehensive than the physician in adding, substituting or increasing drug therapy, where the criteria were clear in the Rules (15 cases). For example, in two cases the physician missed increasing the dose of anti-hypertensive therapy and in three cases ATHENA-HTN recommended adding a non-dihydropyridine calcium channel blocker per the guidelines while the physician did not.

A recommendation was in the Rules, but not encoded in the KB (one case). ATHENA-HTN did not suggest the addition of a thiazide diuretic in a patient with inadequately controlled blood pressure. The Rules clearly stated that a thiazide is a relative indication in all situations.

Ambiguity existed in the Rules (three cases). Despite our efforts to specify the Rules in detail, some ambiguity persisted. For example, in the KB, drug dose ranges (low, medium, high) were specified for each drug, with the "high" range selected to represent the top of the dose response curve (maximal effect) such that greater efficacy would not be expected from further increase of the drug dose in most patients. The Rules were not explicit about the dose ranges or their functionality, so the physician suggested an increase in the dose of a drug that was already in the "high" range but not the maximum FDA-approved dose of the drug.

Physician Comments – Physician entered comments when reviewing a test case. Ten comments led to identification of issues about program boundaries:

Identifying previously unidentified ATHENA-HTN boundaries (three cases). ATHENA-HTN and the physician agreed on drug recommendations for these cases. For example, the physician commented that the patient was taking sotalol and that despite it being in an anti-hypertensive drug class (beta adrenergic receptor antagonist), it was being used as an anti-arrhythmic agent. Managing an anti-arrhythmic was beyond the scope of the program, so recommendations for a drug that is both an anti-hypertensive and an anti-arrhythmic represents a boundary issue for the program.

Other ATHENA-HTN's boundary issues (seven cases). These comments represented a mix of newly identified limitations and deliberate (and thereby previously known) limitations of the system due to design decisions about what was outside the scope of the system. We had elected not to encode recommendations to decrease drug therapy and the physician suggested decreasing therapy (three cases). We had also decided during system design that, while we would provide alerts about high or low potassium values, we would not attempt to make drug recommendations to add, increase or decrease potassium supplements (one case). The MD identified a patient taking a drug that might increase blood pressure. The program design did not include a review of drugs outside the anti-hypertensive drug class for their potential effect on blood pressure. It was beyond the scope of the program to recognize this. (We plan to introduce this feature in a future version of the KB.) The physician also recognized two cases in which the patient had prescriptions for two different drugs from the same anti-hypertensive class. We had designed the KB so that it would not add a second drug from a drug class already in use, but we had not designed it to detect that a patient had prescriptions for two drugs in the same class.

Updates were made to the KB based on testing results (three cases). Discrepancies and physician comments led to three KB updates. Exclusion criteria were updated to include renal hyperplasia, in addition to renal artery stenosis, as it may be associated with secondary hypertension. Hydrochlorothiazide was designated as having a relative indication for all patients without absolute contraindications. Sotalol was re-categorized as an anti-arrhythmic agent rather than an anti-hypertensive to avoid recommending substitutions or dose increases.

Discussion

Our approach to pre-deployment testing is grounded in the recognition that all software contains errors and that a successful test is one that detects errors so they can be addressed. Updating software (in our case the KB) by modifying the content has the potential to introduce unintended errors. The regression method for testing the updated KB in ATHENA-HTN was successful in identifying areas for improvement. Offline testing led to three updates of the KB to ensure accuracy of generated recommendations.

Application of our method of pre-deployment testing led to several important observations. Effective testing requires a gold standard of correct information. At first glance, it might seem that a physician review of patient data would immediately yield answers that would serve as that gold standard. However, for a large number of cases — each with extensive data — fatigue, distraction and brief inattention can lead to incomplete recommendations by an physician. We note that none of the physician recommendations were incorrect or would have been clinically inappropriate. Rather, the physician in a few cases did not include all the possible recommendations that would have been allowed per the guidelines. Inclusion of all possible responses, which might seem compulsive or obsessive in a human being, is exactly one of the things that computer programs do well[10]. Our procedure of having both the physician and the software provide recommendations, then comparing them and resolving discrepancies in a consensus discussion with additional physicians, allowed us to define a gold standard set of answers for the test cases, which we can use in future testing.

It became evident during our testing procedure that it is very difficult to represent the CDSS knowledge base in narrative form. We identified areas in which the Rules had remaining ambiguities open to misinterpretation by the physician. The purpose of the Rules is to share the content of the KB with clinicians in the very precise formats needed for encoding the knowledge. Specifying the guideline knowledge to be encoded allows for the detection of ambiguities and gaps in the original guidelines so that these can be resolved by clinical experts rather than programmers. However, the Rules add another layer that requires maintenance and reconciliation with the KB. Our testing procedure allowed us to identify areas in which the Rules document was not an accurate representation of what was in the KB.

Several features of our pre-deployment testing method should be noted. We used cases drawn from actual patient data files rather than simulated cases constructed to match the concepts programmed into the CDSS. We wanted opportunities to challenge the system with the breadth of actual patient data.

It was important to have an evaluator who was not otherwise involved in the ATHENA-HTN project. The testing was designed to provide the closest possible approximation of real-life applications of the CDSS in order to maximize error detection. Consequently, there is clear value in having the physician-evaluator be representative of the end-user population. Testing by multiple physicians could increase the error detection, but feasibility would become a constraint due to limited availability of busy physicians for this purpose.

When considering the success of the offline testing method for evaluating KBs, it is also important to consider the interaction between the testing system and the physician. In our study there was some difficulty with the presentation of drug doses in the interface used by the MD, which led to discrepancies in drug recommendations. Clear presentation of patient data in a format similar to that used by ATHENA-HTN, rather than in a standard pharmacy representation, would reduce the risk of misinterpretation of patient data.

The offline testing method demonstrated, unsurprisingly, that no matter how complex a CDSS for one particular condition may be, it will never contain enough rules to cover all clinical situations. This is where the interaction between a physician and the CDSS is most important. It is also where the design team’s judgment is particularly important, especially in setting boundaries for the Rules so that the system will be useful yet not unmanageably large or complex. Identifying and acknowledging both the boundaries and the expected behavior of the system are important to diffuse unrealistic expectations in end-users.

Conclusion

The offline testing method of the KB and execution engine was successful in identifying areas for improvement in the KB/execution engine software. Accordingly, updates to the KB were subsequently made. Boundaries of the KB content were better defined, and improvements to the offline testing method were identified. ATHENA-HTN was deployed after the knowledge-base update generated recommendations with improved accuracy.

Acknowledgements

This work was supported in part by VA HSR&D CPG-97-006, CPI-275 and NIH LM05708. The views expressed are those of the authors and not necessarily those of the Dept. Veterans Affairs or other affiliated organizations.

References

1. Goldstein MK, Coleman RW, Tu SW, Shankar RD, O'Connor MJ, Musen MA, et al. Translating research into practice: organizational issues in implementing automated decision support for hypertension in three medical centers. JAMIA 2004 Sep-Oct;11(5):368-76.
2. Goldstein MK, Hoffman BB, Coleman RW, Musen MA, Tu SW, Advani A, et al. Implementing clinical practice guidelines while taking account of changing evidence: ATHENA DSS, an easily modifiable decision-support system for managing hypertension in primary care. Proc AMIA Symp. 2000:300-4.
3. Gennari J, Musen M, Fergerson R, Grosso W, Crubezy M, Eriksson H, et al. The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. Int J Hum Comput Stud. 2003;58(1):89-123.
4. Myers G, Sandler C, Badgett T, Thomas T. The Art of Software Testing. 2nd Ed. Hoboken, NJ: John Wiley & Sons; 2004.
5. Friedman C, Wyatt J. Evaluation Methods in Medical Informatics. NY: Springer-Verlag; 1997.
6. Miller RA. Evaluating evaluations of medical diagnostic systems. JAMIA. 1996 ;3(6):429-31.
7. The Sixth Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Arch Intern Med. 1997;157(21):2413-46.
8. VA/DoD Evidence-Based Clinical Practice Guideline Working Group VHA, Department of Veterans Affairs , and Health Affairs, Department of Defense. Diagnosis and Management of Hypertension in the Primary Care Setting. Washington, DC: Office of Quality and Performance publication; 1999.
9. Leibovici L, Gitelman V, Yehezkelli Y, Poznanski O, Milo G, Paul M, et al. Improving empirical antibiotic treatment: prospective, nonintervention testing of a decision support system. J Intern Med. 1997 Nov;242(5):395-400.
10. Persson M, Mjorndal T, Carlberg B, Bohlin J, Lindholm LH. Evaluation of a computer-based decision support system for treatment of hypertension with drugs: retrospective, nonintervention testing of cost and guideline adherence. J Intern Med. 2000 Jan;247(1):87-93.
11. Tierney W, Overhage J, Takesue B, Harris L, Murray M, Vargo D, et al. Computerizing guidelines to improve care and patient outcomes: the example of heart failure. J Am Med Inform Assoc. 1995 Sep-Oct;2(5):316-22.