Formative evaluation of the accuracy of a clinical decision support system for cervical cancer screening

Kavishwar Balwant Wagholikar,[1] Kathy L MacLaughlin,[2] Thomas M Kastner,[3] Petra M Casey,[3] Michael Henry,[4] Robert A Greenes,[5,6] Hongfang Liu,[1] Rajeev Chaudhry[7]

[1]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA
[2]Division of Family Medicine, Mayo Clinic, Rochester, Minnesota, USA
[3]Division of Obstetrics–Gynecology, Mayo Clinic, Rochester, Minnesota, USA
[4]Division of Anatomic Pathology, Mayo Clinic, Rochester, Minnesota, USA
[5]Department of Biomedical Informatics, Arizona State University, Phoenix, Arizona, USA
[6]Department of Health Science Research, Mayo Clinic, Scottsdale, Arizona, USA
[7]Division of Primary Care Internal Medicine, Center for Innovation, Mayo Clinic, Rochester, Minnesota, USA

Correspondence to
Dr Kavishwar Wagholikar, Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 First Street SW, Rochester, MN 55901, USA; waghsk@gmail.com

ABSTRACT

Objectives We previously developed and reported on a prototype clinical decision support system (CDSS) for cervical cancer screening. However, the system is complex as it is based on multiple guidelines and free-text processing. Therefore, the system is susceptible to failures. This report describes a formative evaluation of the system, which is a necessary step to ensure deployment readiness of the system.

Materials and methods Care providers who are potential end-users of the CDSS were invited to provide their recommendations for a random set of patients that represented diverse decision scenarios. The recommendations of the care providers and those generated by the CDSS were compared. Mismatched recommendations were reviewed by two independent experts.

Results A total of 25 users participated in this study and provided recommendations for 175 cases. The CDSS had an accuracy of 87% and 12 types of CDSS errors were identified, which were mainly due to deficiencies in the system's guideline rules. When the deficiencies were rectified, the CDSS generated optimal recommendations for all failure cases, except one with incomplete documentation.

Discussion and conclusions The crowd-sourcing approach for construction of the reference set, coupled with the expert review of mismatched recommendations, facilitated an effective evaluation and enhancement of the system, by identifying decision scenarios that were missed by the system's developers. The described methodology will be useful for other researchers who seek rapidly to evaluate and enhance the deployment readiness of complex decision support systems.

INTRODUCTION

Although cervical cancer can be largely prevented with screening, it still continues to be a major cause of female cancer-related deaths.[1] Several national organizations have released guidelines for cervical cancer screening and surveillance.[2–5] However, the guidelines are complex and are based on a multitude of factors. Consequently, they cannot be easily recalled by care providers and many patients do not receive the optimal screening.[6–9]

As a potential solution we have previously developed and reported a prototype clinical decision support system (CDSS), which automatically analyzes patient data in the electronic health record (EHR), and suggests the guideline-based recommendation to care providers.[10] However, the system is susceptible to failures due to its complexity as it is based on

multiple guidelines and free-text processing. Another shortcoming of the prototype was that only a single guideline expert was involved in its development. Therefore, further evaluation was necessary to ensure the readiness of the system for deployment in clinical practice. This paper reports the methodology used to evaluate and improve the CDSS with participation of multiple users and experts, before clinical deployment. In contrast to the widely published summative evaluations that determine the post-deployment effectiveness/impact, the aim of this work is to perform a formative evaluation before deployment, in order to ensure the system's post-deployment effectiveness.

BACKGROUND
Cervical cancer screening

Worldwide, cervical cancer was diagnosed in approximately 530 000 women and resulted in approximately 275 000 deaths in 2008.[11] Despite the confirmed effectiveness of routine screening, the American Cancer Society estimates 12 170 cases of cervical cancer and 4220 deaths in the USA in 2012.[1] A meta-analysis of 42 multinational studies reported that over half of the women diagnosed with cervical cancer had inadequate screening or no screening, and that lack of appropriate follow-up of abnormal tests contributed to 12% of diagnoses.[12]

Cervical cancer screening/surveillance involves an evaluation of cervical cells (cytology) through a liquid-based specimen or Papanicolaou (Pap) smear. Human papilloma virus (HPV) testing may be additionally performed to detect the presence of high-risk strains of HPV (the cause of cervical pre-cancer and cancer). Several national organizations including the American Cancer Society, US Preventive Services Task Force, American College of Obstetricians and Gynecologists and the American Society for Colposcopy and Cervical Pathology have released guidelines for cervical cancer screening and/or management of abnormal screening tests.[2–5] However, the guidelines are complex and are based on a multitude of factors including age, risk factors for cervical cancer and previous screening test results. Therefore, recalling and following the evidence-based guidelines is challenging for care providers, as a result of which many patients do not receive optimal screening.[6–9]

Apart from efforts to improve guideline adherence of the providers, several other interventions focused on patients have been investigated in the past two decades.[13] The interventions to improve

screening rates are adjuvant to strategies for reducing the risk factors for HPV infection.[14] They can be broadly categorized as educational,[15] reminders,[16] interactive voice response[17] or telephone call,[18] counseling[19] and economic incentives.[20] Reminders and educational interventions have been found to be most effective.[21–23] With the growing use of EHR in the USA, the use of decision support systems such as ours to implement reminders for providers and patients has a high potential for improving the screening and surveillance rates.[24] The following subsection provides an overview of the challenges for the utilization of such systems.

Clinical decision support

CDSS[25 26] have been developed for a variety of decision problems including preventive services,[27 28] therapeutic management,[29] prevention of adverse events,[30] diagnosis,[31 32] risk estimation,[33] and chronic disease management.[34] CDSS have been found to improve health service delivery across diverse settings, but there is sparse evidence for their impact on clinical outcomes.[35] The potential positive impact of CDSS on the quality of care is not always realized, because the systems are not always utilized or are not implemented effectively.[26] Some of the possible reasons for ineffective implementations are alert fatigue,[36] lack of accuracy,[37] lack of integration with workflow,[38] and prolonged response time.[39]

Formative evaluations to ensure the acceptable levels of the above performance parameters may play a crucial role for effective implementation.[40] In contrast to the widely published summative evaluations that determine the impact/effectiveness of the system, the aim of formative evaluations is to address the factors that will determine the effectiveness, during the development phase itself.[41] Formative evaluations have been emphasized as critical components of EHR implementation[42] and health information technology projects in general.[43] Formative evaluations to rectify failure points of a CDSS before deployment may enhance the effectiveness of deployment in the clinical setting.

Our CDSS is particularly prone to multiple points of failure, because it is based on a complex model synthesized from multiple guidelines, it requires highly accurate natural language processing (NLP), which can be a challenging task, and it utilizes data from a multitude of information sources[10] (see figures 1 and 2). Moreover, the CDSS is aimed to be comprehensive—to generate screening and surveillance recommendations for all female primary care patients in the institution, which is a major advancement over current systems.[44–46]

Therefore, a rigorous validation is required for our system to ensure user acceptability and clinical impact. This paper reports the methodology used to evaluate and improve the CDSS with participation of multiple users and experts, before clinical deployment. The objective is to ensure that the recommendations of the CDSS are of sufficient accuracy to be acceptable and useful to the providers. Testing for usability and work-flow integration are excluded from the scope of the current study.

METHODS

The recommendations of potential end-users for a random sample of patients were recorded and compared to the recommendations generated by the CDSS. Mismatched recommendations were resolved by independent experts, and an error analysis was performed to improve the CDSS. The study was conducted using a web-based application. The detailed methodology is as follows.

Overview of CDSS architecture

As shown in figure 1, the CDSS has three modules: data module, guideline engine, and NLP module. The latter two modules contain respective rulebases, viz a guideline rulebase for representing the screening and management guidelines and a NLP rulebase for interpreting cervical cytology (Pap) reports. When the CDSS is initiated for a particular patient, the guideline engine parses the guideline rules (figure 2) and queries the data module for the required patient parameters. The data module in turn interfaces with the EHR to retrieve the patient information and when the data involves free-text information, for example, a cytology report, the data module calls the NLP module to extract the relevant variables. Based on its constituent rules the guideline engine continues to seek patient parameters, until it has sufficient data to compute the recommendation. The architecture of the CDSS is elaborated elsewhere.[10]

Expert review of guideline model

Before initiating this study, the guideline model (rulebase implemented in the system) was reviewed and approved by several experts who did not participate in the development of the CDSS prototype. Figure 2 shows the flowchart representation of the system's guideline model.

Construction of test set

We randomly selected 6033 patients who had visited Mayo Clinic Rochester in March 2012 and had consented to make their medical records available for research. The CDSS was run

Figure 1 Architecture of the system. CDSS, clinical decision support system; EHR, electronic health record.
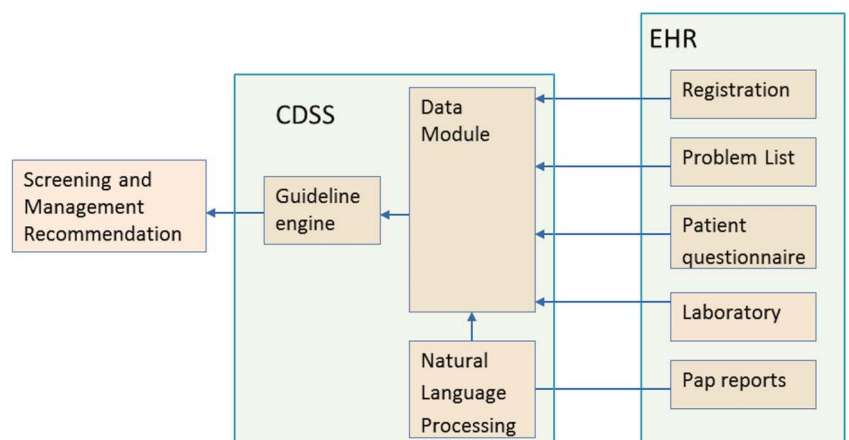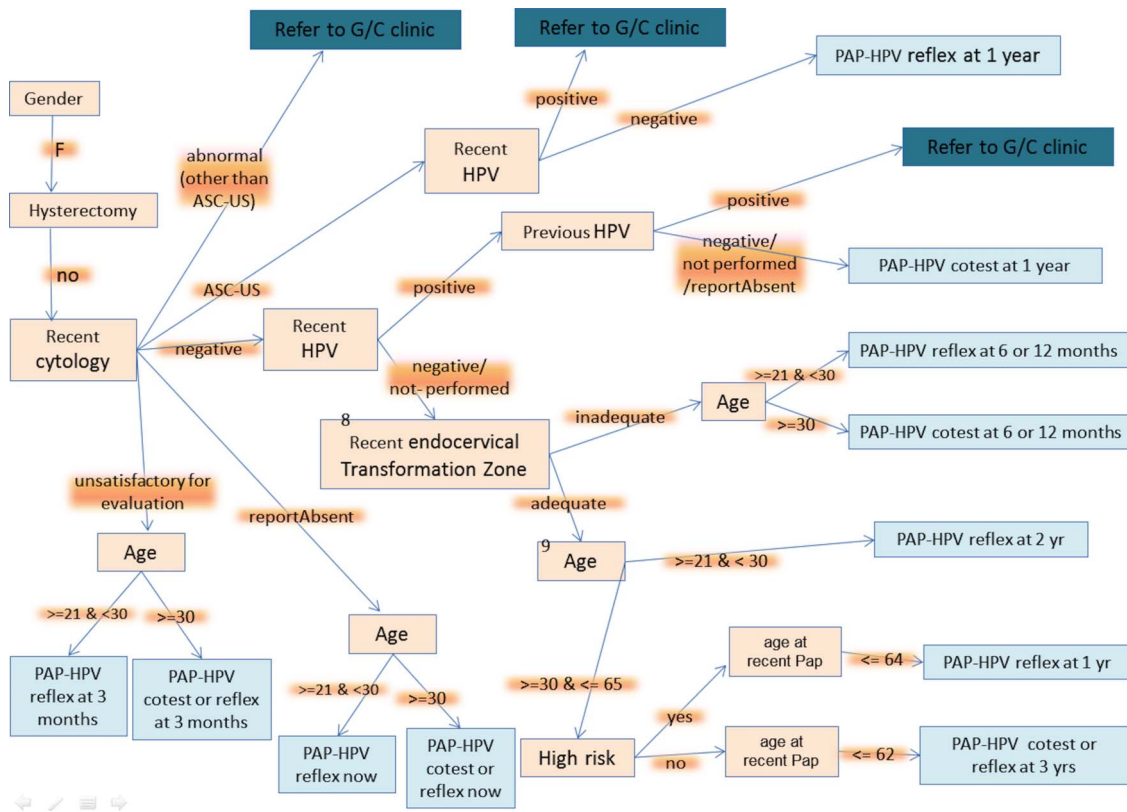
Figure 2 Guideline flowchart for the proof of concept system. It represents the guideline rulebase implemented in the clinical decision support system. ASC-US, atypical squamous cells of undertermined significance; G/C, gynecology clinic; HPV, human papilloma virus; PAP, Papanicolaou.

to compute the screening and surveillance recommendations for these patients. Based on the recommendations the patients were mapped to the branches in the guideline flowchart for cervical cancer screening/management (figure 2). This flowchart was developed before the 2012 updates in the national guidelines.[2–5] Each pathway in the flowchart corresponds to a distinct combination of patient variables, and it represents a unique decision scenario. As some decision scenarios occur more frequently during practice than others, a randomly selected test set can be biased towards the frequent decision scenarios. Therefore, to ensure that the evaluation was not biased to the frequent scenarios, we performed stratified random sampling, restricting the selection to a maximum of 14 cases per decision scenario. The total number of cases in the test set was 196.

User participation

We invited 89 potential users of the CDSS to participate in this study. The recruitment was done by sending mass emails as well as by specifically contacting potential users. The participants were of diverse background and training. They included staff consultants, residents and nurse practitioners from the institution's departments of family medicine, internal medicine and obstetrics and gynecology. We created a web-based application to collect the recommendations of the healthcare providers for the test set (figure 3). The web application was deployed on the institution's internal network.

Collection of provider recommendations

The web system was available from 12 April 2012 to 4 May 2012. When a participant logged into the system, a 1-min training video was presented. Subsequent to the video presentation, the web system randomly selected (without repetition) a case

number from the test set and presented it to the participants. The participants assessed the information for the presented case by chart review using the EHR system, and recorded the most appropriate guideline-based recommendation for the case, by selecting the appropriate options in the web system's interface (figure 3). In addition to the template recommendation options, a free-text box was provided, to allow the participants to input recommendations that were not covered in the template options. Each participant completed seven different cases. The web system also recorded the time taken by the providers to input their recommendations.

Analysis

The care providers' recommendations were compared with those of the CDSS (figures 4 and 5). When there was a mismatch in the recommendations, the case was reviewed by one
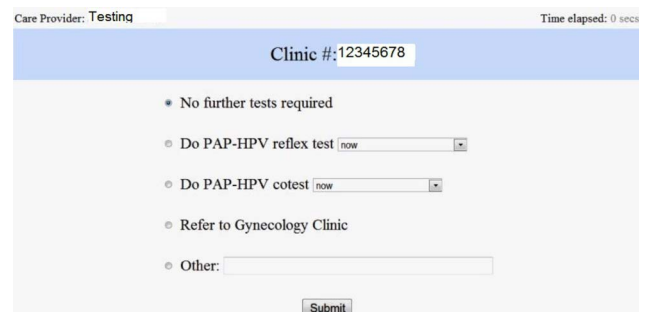

Figure 3 Interface of the web-system used by care providers to participate in the study. HPV, human papilloma virus; PAP, Papanicolaou.
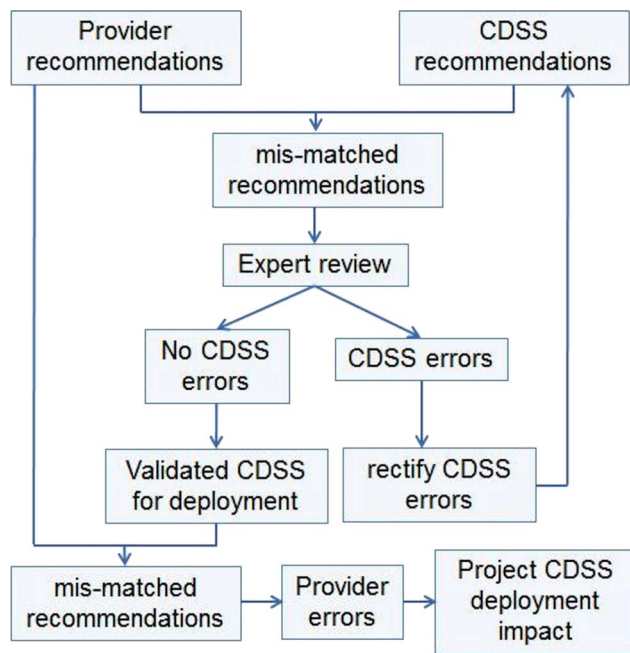
Figure 4 Study design. CDSS, clinical decision support system.

of two experts who did not participate in the development of the prototype, to decide if the CDSS or the provider recommendation was more accurate/optimal. If the CDSS was found to be less optimal, an error analysis was performed to identify the fault in the CDSS. The CDSS was then improved to correct the identified errors.

Projection of CDSS impact on clinical practice

The CDSS was modified and re-evaluated on the test set, in order to ensure that the errors identified in the above analysis

Figure 5 Summary of test set construction and CDSS evaluation results, showing number of cases in each step of the study. CDSS, clinical decision support system; EHR, electronic health record.

were rectified. Finally, we compared the recommendations of the corrected CDSS with those of the providers to identify provider errors. These cases were analyzed to identify the decision scenarios that were difficult for the providers, in order to project the potential of the CDSS to assist with the decisions. The average time taken by the providers to make the recommendations was computed, after excluding outliers.

RESULTS

Figure 5 summarizes the results of the CDSS evaluation. Of the 89 providers who were invited to participate in the study, 28 agreed to participate, and finally 25 completed the exercise of annotating the test cases with their recommendation. A total of 175 cases was annotated by the participants. The CDSS was found to generate an error flag for six cases because it could not obtain the pathology reports due to bugs in the interface to the EHR system. In the remaining 169 cases, the recommendations by the healthcare providers did not match the recommendation made by the CDSS for 75 cases.

The mismatch cases were presented to one of two experts (who co-authored this paper). The experts reviewed the recommendations and decided on the final optimal recommendation for the patient. The experts were blinded to the identity of the healthcare provider who made the recommendation for the individual cases. The CDSS was found to be suboptimal compared to the provider in 22 cases. Therefore, the accuracy computed to 147/169=87.0% (figure 5 and table 1).

CDSS error analysis

Analysis of the 22 CDSS failure cases, led to identification of 12 errors/failure points in the CDSS (table 2 and figure 6). The errors were classified as modeling errors and programming errors. Modeling errors are due to deficiencies in the system's guideline rulebase/model, for example, missing a decision scenario, or incorrect logic. Programming errors include errors/bugs
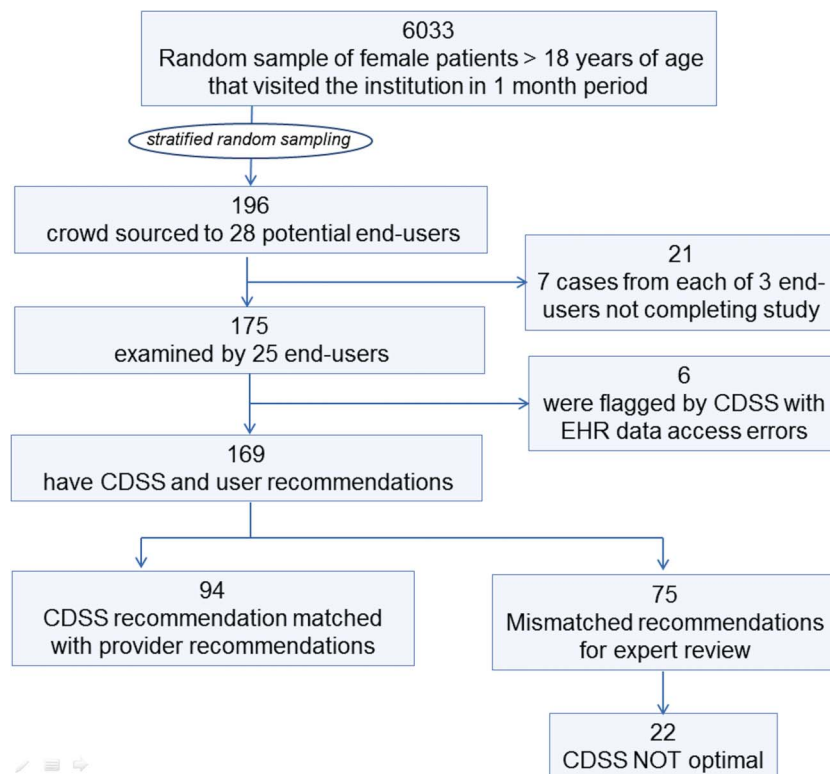
Table 1 Distribution of CDSS errors over different decision scenarios

Test cases	Patient variables							Grouped decision scenarios	% Incorrect by CDSS
	Hysterectomy	Age (years)	Recent HPV	Previous HPV	High risk	ETZ adequate	Cytology		
12	Yes							Hysterectomy	17
3		≥66			No		No report	Cytology report absent	13
3		≥30 & <66					No report		
8		≥21 & <30					No report		
10		<21					No report		
13			Pos				ASCUS	ASCUS	4
11			Neg				ASCUS		
2			NP				ASCUS		
1			Pos	pos			Neg	Cyto negative and HPV pos	0
7			Nos	pos/NP/absent			Neg		
2					Yes	No	Neg	Cyto and HPV neg and ETZ inadequate	25
2		<21			No	No	Neg		
2		>65			No	No	Neg		
1		≥30	Neg/NP		No	No	Neg		
6		≥30	Neg/NP		No	No	Neg		
11		≥21 & <30	Neg/NP		No	No	Neg		
15		≥30	Neg/NP		No	Yes	Neg	Cyto and HPV negative	14
18		≥30 & <66	Neg/NP		No	Yes	Neg		
10		≥21 & <30	Neg/NP			Yes	Neg		
10		≥30	Neg/NP		Yes	Yes	Neg	Normal cytology high risk	0
5		≥66			No		Unsatis.	Unsatis.	33
1		≥66			Yes		Unsatis.		
5		≥30 & <66					Unsatis.		
1		≥21 & <30					Unsatis.		
10							Abnormal (other than ASCUS)	Abnormal (other than ASCUS)	0
169									13

The combination of patient variables corresponds to decision scenarios that are grouped for readability and interpretation in the last two columns.
ASCUS, atypical squamous cells of undertermined significance; CDSS, clinical decision support system; Cyto, cytology; ETZ, endocervical transformation zone; HPV, human papillomavirus; NP, not performed; Unsatis, unsatisfactory for evaluation.

in the developed software, for example, incorrect rounding for age cut-off. The CDSS was robust in extracting the patient information from the EHR, except for history of hysterectomy. A summary of the errors is as follows (figure 6):

▶ The upper age limit for screening recommendation was not set, because the approach was to err on the side of caution and let the provider overrule the system's recommendation for stopping screening (errors 1, 2, 4 and 9). This has now been rectified by considering the high-risk status of the patients.

▶ Some of the error cases were due to the system stopping screening after the patient's 65th birthday. In these cases the age limit was applied after rounding the age (error 2). Therefore, to define the age explicitly and avoid rounding, the guideline model has been changed to the condition of <66 instead of ≤65 as defined earlier.

▶ History of hysterectomy was missed when it was reported in the problem list. This was a programming error that was resolved (error 6). In one case, hysterectomy was not mentioned in the problem list but occurred in the clinical notes, which are not searched by the system. This case was resolved after concepts that implied hysterectomy, for example, 'vaginal wall prolapse after hysterectomy' were included for determining history of hysterectomy, as this concept was present in the patient's problem list (error 10).

▶ The scenario of atypical squamous cells of undertermined significance (ASCUS) cytology with HPV not performed was not anticipated. This has now been included in the corrected model (error 7). A report of inadequate endocervical transformation zone is now ignored for high-risk patients, because it does not impact their management. This is because they are already having annual screening (error 12).

After the errors were rectified in the CDSS, it was found to generate optimal recommendations for all but one failure case. The one case that could not be resolved was due to the inability of the CDSS to identify history of hysterectomy in a patient, when both the problem list and patient annual questionnaire database had no documentation about the patient's hysterectomy. The experts inferred that the patient had undergone hysterectomy from the clinical notes. The CDSS failed because it was not designed to perform NLP on clinical notes to extract this information.

Provider errors analysis
After the recommendations of corrected CDSS were compared to those recorded by the providers, the providers were found to provide suboptimal recommendations in 56 of the 169 cases (33.1%), which is 34 (20.1%) more cases with suboptimal

Table 2 Listing and classification of CDSS errors (corresponds to figure 6)

Grouped decision scenarios	Error description	Error number	Type of error
Hysterectomy	Missed history of hysterectomy in problem list	6	Programming
	Missed a case of hysterectomy when not mentioned in problem list, but found in clinical note. This information is now obtained from patient provided data sources	10	Programming
Report absent	If cytology report is absent and age is <21 years, recommendation should be perform Pap-HPV reflex at age 21 years, instead of saying no recommendation	11	Modeling
	When the cytology report is not found, there needs to be an upper age limit for recommending screening 'now' for low-risk patients. For high-risk patients screening should be recommended even when age >65 years	4	Modeling
ASCUS	Missing decision scenario: when cervical cytology is ASCUS and HPV is not performed recommendation should be 'Cytology at 6 and 12 months'	7	Modeling
Cyto and HPV neg ETZ inadequate	When the cytology and HPV are negative but ETZ is inadequate examine age instead of age at recent report to recommend next screening	5	Programming
	When cytology and HPV are negative and in adequate ETZ, recommend Pap-HPV reflex at 6 or 12 months if last test was co-test, or recommend Pap-HPV co-test at 6 or 12 months if last test was reflex	8	Modeling
	When the cytology and HPV are negative but ETZ is inadequate, there is a need for upper age limit	9	Modeling
	Excluded consideration of inadequate ETZ for high-risk patients	12	Modeling
Normal cytology	For determining high risk exclude CIN1	3	Modeling
	For recommending screening for high risk patients, the upper age limit cut-off needs to be removed, as they would continue annual screen even if >65 years old. For low-risk patients with normal cytology, the upper age limit cut-off needs to be corrected	2	Modeling
Unsatisfactory for evaluation	For low-risk patients, there needs to be an upper age limit for recommending repeat test after 3 months	1	Modeling

ASCUS, atypical squamous cells of undertermined significance; CDSS, clinical decision support system; CIN1, cervical intraepithelial neoplasia 1; HPV, human papilloma virus; PaP, Papanicolaou; ETZ, endocervical transformation zone.

recommendations compared to the CDSS. Several of these patients had abnormal screening reports such as abnormal (other than ASCUS) cytology, ASCUS cytology, positive HPV or inadequate endocervical transformation zone. Some of the provider errors were due to incorrect determination of the risk status of the patient, due to boundary conditions such as age cut-offs. The mean time taken by the providers to make the recommendation was 1 min 39 s.

Figure 6 Modified guideline flowchart. The number in red circles corresponds to the errors described in the text and table 2. The yellow rectangles circumscribe the elements that were appended or modified to make the corrections. ASC-US, atypical squamous cells of undertermined significance; ETZ, endocervical transformation zone; G/C, gynecology clinic; HPV, human papilloma virus; PAP, Papanicolaou.

DISCUSSION

The study facilitated a comprehensive evaluation of the CDSS on a large and diverse set of patients that covered nearly all possible decision scenarios. The CDSS was evaluated to have a fair accuracy, and by performing the error analysis of failure cases the CDSS was considerably improved.

The formative evaluation based on the reference set annotated by the care providers led to the identification of several failure points in the system. Several logical steps necessary to apply the national guidelines were missed when the guideline model was inspected by the experts before the study. The use of representative cases and their decision annotations by the care providers in this study helped draw attention to the particular scenarios in which the logical steps were missed. The task of modeling the free-text guidelines as rules is challenging due to ambiguity of the natural language used in the guidelines, and due to the difficulty in envisioning decision scenarios that can occur in clinical practice.[47 48] Our results indicate that guideline models based on abstraction from textual guidelines need to be tested with consistency checks on real-life cases. This finding is consistent with earlier research that demonstrates the critical importance of carefully analyzing the reasons for practising clinician disagreements with decision support, in order to improve CDSS design and effectiveness.[49]

The analysis identified situations/factors when the CDSS was prone to make errors, for example, hysterectomy cases. It also identified guideline areas in which the care providers need decision support. The providers were found to have difficulties in decision making for cases with abnormal findings, as reported by other studies.[12 50] Lack of follow-up referral after a positive screening test has also previously been documented in the context of colorectal cancer screening.[51 52] As the patients with abnormal screening reports are especially at risk of developing cancer, the screening/surveillance recommendations made by the providers can have far-reaching consequences for the patients. The CDSS was notably found to perform consistently well for such patients, and its deployment can be expected to improve the quality of the screening services considerably. Moreover, the CDSS can lead to provider time savings of 1 min 39 s per patient consultation, as determined in this study.

An alternative approach to evaluate the CDSS before deployment in clinical practice is to conduct a pilot study with a subset of potential end-users, who will verify the system's recommendation and provide feedback for improving the system. There are several disadvantages to this approach: the evaluation will be biased towards frequently occurring decision scenarios unless a special effort is made to identify the less frequent but high impact scenarios in the evaluation; and there will be a risk of missing validation for rare but important decision scenarios. Our approach of identifying distinct decision scenarios for the evaluation by using the prototype CDSS helped avoid bias towards the frequent decision scenarios, and allowed for an efficient utilization of the efforts of the participating providers and experts.

Similarly, our approach to blind the users to the CDSS recommendation has an advantage over seeking user feedback after deployment, because in the post-deployment setting, the user's judgment can be influenced by knowledge of the output of the CDSS.[49] Consequently, in the latter approach some of the failure points may be missed. Moreover, it may not be possible to project the clinical impact of the system, due to the modification of user behavior. With the current approach the decision scenarios that were difficult for the users were identified, and the usefulness of the system after deployment could be projected. Another advantage is that the end-users are not directly exposed to the CDSS before the formative evaluation; therefore, there is no loss of user confidence.[53]

A difficulty in performing CDSS evaluation is that it is often not feasible to involve a large number of users in system evaluation. The crowd-sourcing approach used in this study allowed a large number of users to participate, which in turn facilitated the construction of a large reference dataset of real-life decision scenarios. Consequently, the CDSS could be evaluated comprehensively for a wide variety of scenarios.

Literature on CDSS mainly consists of summative evaluations measuring impact on service and clinical outcomes.[54 55] Studies on performance aspects of the CDSS are rare, which suggests a lack of effort to ensure effective implementation. Our results demonstrate that such studies may be increasingly needed as complex CDSS that have an increased risk of failures are developed. Furthermore, research into developing efficient and practically feasible methods for pre-deployment evaluation of CDSS is called for. We believe that the approach described will be useful for developing complex systems that support wider and more complex domains of care.[28 56] The formative evaluation to ensure that the decision model itself is accurate will facilitate subsequent enquiries after deployment for quantifying guideline adherence of the providers, and for measuring clinical impact.

Crowd-sourcing can be useful for the development and validation of decision support applications. McCoy et al[57] have earlier used crowd-sourcing for building a knowledge base of problem–medication pairs. In their institution it was mandatory for clinicians to link prescriptions to patient problems, and McCoy et al[57] leveraged the resulting database as a resource to construct their knowledge base. On the other hand, our approach was to seek volunteer effort from the care providers for creating a gold standard for validating the CDSS.

Our analysis identified that the incompleteness of problem list and patient-provided information for hysterectomy is a challenge to accurate working of the CDSS for the subset of patients with hysterectomy. We plan to extend the NLP module of the CDSS to identify history of hysterectomy from clinical notes, if more such patients are encountered in the future.[56] Overall, the CDSS has a high level of accuracy, and has the potential to improve providers' recommendations especially in the high utility areas of the guidelines, and can thereby significantly advance the quality of screening. However, the corrected CDSS was not tested with new cases, which would be of benefit to determine whether further discrepancies in recommendations need to be addressed. We expect that the majority of the errors have been identified in the current analysis, and we plan to perform additional evaluations with a different set of cases to ensure system accuracy before deployment.[58]

We restricted the scope of the evaluation to accuracy and did not test the usability and integration with workflow, which are also major factors that determine utilization and clinical impact of the CDSS. These will be tested separately with pilot studies. Nonetheless, we expect that elimination (or at least minimization) of the issue of delivering the correct recommendations will facilitate the subsequent pilots.

Limitations

The use of an unfamiliar interface may have induced participants' mistakes, although we had provided a training video and designed a simple interface to record the participants'

recommendations. On the other hand, the participants were focused on the task of making screening decisions, and their performance can be expected to be better than target users who will have other tasks during the patient visit. As a result of these factors, further research is necessary to determine the usefulness of our approach to quantify provider errors. Nevertheless, our results indicate that the methodology is useful to identify qualitatively the areas for decision making that are difficult for the providers.

Updated cervical cancer screening guidelines were published at the end of our evaluation period.[59 60] It is possible that some of the participating providers were aware of the forthcoming change in the guideline and provided recommendations in accordance with the anticipated guideline.

We limited the expert review to cases in which there was a mismatch in recommendations of the CDSS and the providers, because the proportion of errors is expected to be high in this subset of cases. Consequently, there is a chance of missing erroneous decisions, when the recommendations of both the provider and CDSS are not optimal. However, such cases are expected to be small in number and are likely to have a representation in the mismatch group. The strategy of focusing on the mismatch group facilitates a judicious use of the expert reviewers' efforts.

Double blinding of reviewers was not done. It may be useful to blind the expert reviewers as to whether the source of the recommendations was the care provider or CDSS.

CONCLUSION

Our case study demonstrates that the approach to crowd-source the construction of the reference recommendations set, coupled with the expert review of mismatched decisions, can facilitate an effective evaluation of the accuracy of a CDSS. It is especially useful to identify decision scenarios that may be missed by the system's developers. The methodology will be useful for researchers who seek rapidly to evaluate and enhance the deployment readiness of next generation decision support systems that are based on complex guidelines.

REFERENCES

1 Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. *CA Cancer J Clin* 2012;62:10–29.

2 Saslow D, Runowicz CD, Solomon D, *et al*. American Cancer Society guideline for the early detection of cervical neoplasia and cancer. *CA Cancer J Clin* 2002;52:342–62.

3 ACOG Committee on Practice Bulletins—Gynecology. ACOG Practice Bulletin no. 109: cervical cytology screening. *Obstet Gynecol* 2009;114:1409–20.

4 Screening for Cervical Cancer. 2003 [cited; http://www.uspreventiveservicestaskforce.org/uspstf/uspscerv.htm (accessed 10 Oct 2012).

5 Wright TC Jr, Massad LS, Dunton CJ, *et al*. 2006 Consensus guidelines for the management of women with abnormal cervical cancer screening tests. *Am J Obstet Gynecol* 2007;197:346–55.

6 Yabroff KR, Saraiya M, Meissner HI, *et al*. Specialty differences in primary care physician reports of Papanicolaou test screening practices: a national survey, 2006 to 2007. *Ann Intern Med* 2009;151:602–11.

7 Saraiya M, Berkowitz Z, Yabroff KR, *et al*. Cervical cancer screening with both human papillomavirus and Papanicolaou testing vs Papanicolaou testing alone: what screening intervals are physicians recommending? *Arch Intern Med* 2010;170:977–85.

8 Lee JW, Berkowitz Z, Saraiya M. Low-risk human papillomavirus testing and other nonrecommended human papillomavirus testing practices among U.S. health care providers. *Obstet Gynecol* 2011;118:4–13.

9 Roland KB, Soman A, Benard VB, *et al*. Human papillomavirus and Papanicolaou tests screening interval recommendations in the United States. *Am J Obstet Gynecol* 2011;205:447; e1–8.

10 Wagholikar KB, Maclaughlin KL, Henry MR, *et al*. Clinical decision support with automated text processing for cervical cancer screening. *J Am Med Inform Assoc* 2012;19:833–9.

11 Arbyn M, Castellsague X, de Sanjose S, *et al*. Worldwide burden of cervical cancer in 2008. *Ann Oncol* 2011;22:2675–86.

12 Spence AR, Goggin P, Franco EL. Process of care failures in invasive cervical cancer: systematic review and meta-analysis. *Prev Med* 2007;45:93–106.

13 Yabroff KR, Zapka J, Klabunde CN, *et al*. Systems strategies to support cancer screening in U.S. primary care practice. *Cancer Epidemiol Biomarkers Prev* 2011;20:2471–9.

14 Shepherd JP, Frampton GK, Harris P. Interventions for encouraging sexual behaviours intended to prevent cervical cancer. *Cochrane Database Syst Rev* 2011;(4):CD001035.

15 Young JM, Ward JE. Randomised trial of intensive academic detailing to promote opportunistic recruitment of women to cervical screening by general practitioners. *Aust NZ J Public Health* 2003;27:273–81.

16 Tseng DS, Cox E, Plane MB, *et al*. Efficacy of patient letter reminders on cervical cancer screening: a meta-analysis. *J Gen Intern Med* 2001;16:563–8.

17 Corkrey R, Parkinson L, Bates L, *et al*. Pilot of a novel cervical screening intervention: interactive voice response. *Aust NZ J Public Health* 2005;29:261–4.

18 Miller SM, Siejak KK, Schroeder CM, *et al*. Enhancing adherence following abnormal Pap smears among low-income minority women: a preventive telephone counseling strategy. *J Natl Cancer Inst* 1997;89:703–8.

19 Downs LS Jr, Scarinci I, Einstein MH, *et al*. Overcoming the barriers to HPV vaccination in high-risk populations in the US. *Gynecol Oncol* 2010;117:486–90.

20 Marcus AC, Kaplan CP, Crane LA, *et al*. Reducing loss-to-follow-up among women with abnormal Pap smears. Results from a randomized trial testing an intensive follow-up protocol and economic incentives. *Med Care* 1998;36:397–410.

21 Everett T, Bryant A, Griffin MF, *et al*. Interventions targeted at women to encourage the uptake of cervical screening. *Cochrane Database Syst Rev* 2011;(5):CD002834.

22 Zapka J, Taplin SH, Price RA, *et al*. Factors in quality care—the case of follow-up to abnormal cancer screening tests—problems in the steps and interfaces of care. *J Natl Cancer Inst Monogr* 2010;2010:58–71.

23 Wright A, Poon EG, Wald J, *et al*. Randomized controlled trial of health maintenance reminders provided directly to patients through an electronic PHR. *J Gen Intern Med* 2012;27:85–92.

24 Kim JJ. Opportunities to improve cervical cancer screening in the United States. *Milbank Q* 2012;90:38–41.

25 Greenes RA, ed. *Clinical decision support: the road ahead 1ed*. Massachusetts: Academic Press, 2006.

26 Berner ES. *Clinical decision support systems: state of the art*. Rockville (MD): Agency for Healthcare Research and Quality; 2009, Report No: AHRQ Publication No.09-0069-EF.

27 Sequist TD, Zaslavsky AM, Marshall R, *et al*. Patient and physician reminders to promote colorectal cancer screening: a randomized controlled trial. *Arch Intern Med* 2009;169:364–71.

28 Wagholikar KB, Sohn S, Wu S, *et al*. Clinical decision support for colonoscopy surveillance using natural language processing. IEEE Healthcare Informatics, Imaging, and Systems Biology Conference. San Diego, CA: University of California, 2012.

29 Strom BL, Schinnar R, Bilker W, *et al*. Randomized clinical trial of a customized electronic alert requiring an affirmative response compared to a control group receiving a commercial passive CPOE alert: NSAID–warfarin co-prescribing as a test case. *J Am Med Inform Assoc* 2010;17:411–15.

30 Etchells E, Adhikari NK, Wu R, *et al*. Real-time automated paging and decision support for critical laboratory abnormalities. *BMJ Qual Saf* 2011;20:924–30.

31 Elkin PL, Liebow M, Bauer BA, *et al*. The introduction of a diagnostic decision support system (DXplain) into the workflow of a teaching hospital service can

decrease the cost of service for diagnostically challenging Diagnostic Related Groups (DRGs). *Int J Med Inform* 2010;79:772–7.

32 Aronsky D, Fiszman M, Chapman WW, *et al*. Combining decision support methodologies to diagnose pneumonia. *Proceedings of the AMIASymposium* 2001:12–16.

33 Holt TA, Thorogood M, Griffiths F, *et al*. Automated electronic reminders to facilitate primary cardiovascular disease prevention: randomised controlled trial. *Br J Gen Pract* 2010;60:e137–43.

34 Gilutz H, Novack L, Shvartzman P, *et al*. Computerized community cholesterol control (4C): meeting the challenge of secondary prevention. *Isr Med Assoc J* 2009;11:23–9.

35 Bright TJ, Wong A, Dhurjati R, *et al*. Effect of clinical decision-support systems: a systematic review. *Ann Intern Med* 2012;157:29–43.

36 Kesselheim AS, Cresswell K, Phansalkar S, *et al*. Clinical decision support systems could be modified to reduce 'alert fatigue' while still minimizing the risk of litigation. *Health Aff (Millwood)* 2011;30:2310–17.

37 Wagholikar KB, Sundararajan V, Deshpande AW. Modeling paradigms for medical diagnostic decision support: a survey and future directions. *J Med Syst* 2012;36:3029–49.

38 Perna G. Clinical alerts that cried wolf. As clinical alerts pose physician workflow problems, healthcare IT leaders look for answers. *Healthc Inform* 2012;29:18, 20.

39 Lee F, Teich JM, Spurr CD, *et al*. Implementation of physician order entry: user satisfaction and self-reported usage patterns. *J Am Med Inform Assoc* 1996;3:42–55.

40 Friedman CP. 'Smallball' evaluation: a prescription for studying community-based information interventions. *J Med Libr Assoc* 2005;93(4 Suppl.):S43–8.

41 Manias E, Bullock S, Bennett R. Formative evaluation of a computer-assisted learning program in pharmacology for nursing students. *Comput Nurs* 2000;18:265–71.

42 McGowan JJ, Cusack CM, Poon EG. Formative evaluation: a critical component in EHR implementation. *J Am Med Inform Assoc* 2008;15:297–301.

43 Lorenzi NM, Novak LL, Weiss JB, *et al*. Crossing the implementation chasm: a proposal for bold action. *J Am Med Inform Assoc* 2008;15:290–6.

44 Banks E, Chudnoff S, Freda MC, *et al*. An interactive computer program for teaching residents pap smear classification, screening and management guidelines: a pilot study. *J Reprod Med* 2007;52:995–1000.

45 Sittig DF, Wright A, Osheroff JA, *et al*. Grand challenges in clinical decision support. *J Biomed Inform* 2008;41:387–92.

46 Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009;42:760–72.

47 Tu SW, Musen MA, Shankar R, *et al*. Modeling guidelines for integration into clinical workflow. *Stud Health Technol Inform* 2004;107:174–8.

48 Peleg M, Tu S, Bury J, *et al*. Comparing computer-interpretable guideline models: a case-study approach. *J Am Med Inform Assoc* 2003;10:52–68.

49 Hoeksema LJ, Bazzy-Asaad A, Lomotan EA, *et al*. Accuracy of a computerized clinical decision-support system for asthma assessment and management. *J Am Med Inform Assoc* 2011;18:243–50.

50 Berkowitz Z, Saraiya M, Benard V, *et al*. Common abnormal results of pap and human papillomavirus cotesting: what physicians are recommending for management. *Obstet Gynecol* 2010;116:1332–40.

51 Turner B, Myers RE, Hyslop T, *et al*. Physician and patient factors associated with ordering a colon evaluation after a positive fecal occult blood test. *J Gen Intern Med* 2003;18:357–63.

52 Carlson CM, Kirby KA, Casadei MA, *et al*. Lack of follow-up after fecal occult blood testing in older adults: inappropriate screening or failure to follow up? *Arch Intern Med* 2011;171:249–56.

53 Chin T. Doctors pull plug on paperless system. American Medical News 2003 [cited; http://www.ama-assn.org/amednews/2003/02/17/bil20217.htm. (accessed 10 Oct 2012).

54 Bright TJ, Wong A, Dhurjati R, *et al*. Effect of clinical decision-support systems: a systematic review. *Ann Intern Med* 2012;157:29–43.

55 Jaspers MW, Smeulers M, Vermeulen H, *et al*. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *J Am Med Inform Assoc* 2011;18:327–34.

56 Wagholikar K, Sohn S, Wu S, *et al*. Workflow-based data reconciliation for clinical decision support: case of colorectal cancer screening and surveillance. AMIA Summits Translation Science Proceedings. San Francisco, CA, 2013.

57 McCoy AB, Wright A, Laxmisan A, *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;19:713–18.

58 Chaudhry R, Wagholikar K, Decker L, *et al*. The Innovations in the delivery of primary care services using a software solution: the Mayo Clinic's Generic Disease Management System. *Int J Person Centered Med* 2012;2:361–7.

59 Saslow D, Solomon D, Lawson HW, *et al*. American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. *J Low Genit Tract Dis* 2012;16:175–204.

60 Moyer VA. Screening for cervical cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med* 2012;156:880–91; W312.