Evaluating a Modular Decision Support Application For Colorectal Cancer Screening

Laura G. Militello[1]; Julie B. Diiulio[1]; Morgan R. Borders[1]; Christen E. Sushereba[1]; Jason J. Saleem[2]; Donald Haverkamp[3]; Thomas F. Imperiale[4, 5, 6]

[1]Applied Decision Science, Dayton, Ohio;
[2]Department of Industrial Engineering, University of Louisville, Louisville, KY, USA;
[3]Centers for Disease Control and Prevention, Albuquerque, NM, USA;
[4]Department of Medicine, Indiana University School of Medicine;
[5]Regenstrief Institute;
[6]Richard L. Roudebush, VA Medical Center's Center of Innovation

Keywords
Health information technology, decision support techniques, evaluation methodology, colorectal cancer, cancer screening

Summary
Background: There is a need for health information technology evaluation that goes beyond randomized controlled trials to include consideration of usability, cognition, feedback from representative users, and impact on efficiency, data quality, and clinical workflow. This article presents an evaluation illustrating one approach to this need using the Decision-Centered Design framework.
Objective: To evaluate, through a Decision-Centered Design framework, the ability of the Screening and Surveillance App to support primary care clinicians in tracking and managing colorectal cancer testing.
Methods: We leveraged two evaluation formats, online and in-person, to obtain feedback from a range primary care clinicians and obtain comparative data. Both the online and in-person evaluations used mock patient data to simulate challenging patient scenarios. Primary care clinicians responded to a series of colorectal cancer-related questions about each patient and made recommendations for screening. We collected data on performance, perceived workload, and usability. Key elements of Decision-Centered Design include evaluation in the context of realistic, challenging scenarios and measures designed to explore impact on cognitive performance.
Results: Comparison of means revealed increases in accuracy, efficiency, and usability and decreases in perceived mental effort and workload when using the Screening and Surveillance App.
Conclusion: The results speak to the benefits of using the Decision-Centered Design approach in the analysis, design, and evaluation of Health Information Technology. Furthermore, the Screening and Surveillance App shows promise for filling decision support gaps in current electronic health records.

Correspondence to:
Laura G. Militello, MA
Applied Decision Science
5335 Far Hills Avenue, Suite 217
Dayton, Ohio 45429
Email: l.militello@applieddecisionscience.com
(937) 602–7844

Appl Clin Inform 2017; 8: 162–179
https://doi.org/10.4338/ACI-2016-09-RA-0152
received: September 11, 2016
accepted: December 5, 2016
published: February 15, 2017
Citation: Militello LG, Diiulio JB, Borders MR, Sushereba CE, Saleem JJ, Haverkamp D, Imperiale TF. Evaluating a modular decision support application for colorectal cancer screening. Appl Clin Inform 2017; 8: 162–179
https://doi.org/10.4338/ACI-2016-09-RA-0152

1. Background and Significance

The need for improved evaluation of health informatics applications is well-recognized. The difficulty of conducting evaluations of tools to be used in the complex and highly varied contexts in which healthcare takes place has been widely discussed [1-5]. The 2004 European workshop on New Approaches to Systematic Evaluation of Health Information Systems (HIS-EVAL) highlighted many of the challenges and proposed recommendations [6]. A few years later the U.S National Institute of Standards and Technology issued guidelines for evaluation of health information technology.[7–8] Both call for multidisciplinary approaches that go beyond randomized controlled trials to include consideration of usability, cognition, feedback from representative users, and impact on efficiency, data quality, and clinical workflow. This article presents a case study, illustrating one approach to managing the constraints and complexity of conducting evaluations in healthcare.

1.1 Objective

The objective of this project was to evaluate the Screening and Surveillance App, a modular decision support application intended to help primary care clinicians track and manage colorectal cancer screening for their patients. The intent was to assess the feasibility of this approach and begin to quantify the anticipated benefits. A Decision-Centered Design framework was used to help guide the analysis, design, and evaluation of the App.

1.2 Colorectal Cancer Screening

Colorectal cancer (CRC) is the second leading cause of cancer related deaths in the U.S.[9] Although screening tests are highly effective at detecting early stage cancers, screening rates remain lower for CRC than for other cancers such as breast and cervical [10–11]. Primary care clinician (PCC) recommendations are highly correlated with whether or not patients obtain screening [12–14]; thus, strategies to support PCCs represent one path to increasing screening rates [14–16].

1.3 Tracking and Managing Colorectal Cancer Screening

Electronic health records (EHRs) represent a natural platform for helping PCCs discuss CRC with their patients and recommend testing at appropriate intervals. In fact, many EHRs include clinical reminders for CRC screening. However, clinical reminders are limited in their effectiveness because they rarely include information about a specific patient's testing history and risk factors [17]. Although this information may be stored in the EHR, it is difficult to find. Tracking and managing this information is even more difficult because colorectal cancer testing can occur at long intervals and in different locations. The challenge of finding data that are fragmented and stored in multiple places in the EHR has been identified as a major safety related risk area in the NISTIR 7804–1 [18].

1.4 The Screening and Surveillance App: A Modular Approach

One strategy for increasing screening rates among patients is to improve the accessibility of CRC related information for their PCCs. We designed and developed a modular decision support application, the Screening and Surveillance App [19], to help PCCs track and manage their patients' CRC screening regardless of whether the patient is in routine screening mode, or in surveillance mode based on prior findings.

Modular applications provide a promising means to integrate cognitive support into EHRs [20-21]. The Screening and Surveillance App queries the EHR for specific CRC-related data, and displays the data in a way that supports the needs of the clinicians. This type of modular application can reduce the need to search the EHR for key information. All relevant information (e.g., test results and relevant guidelines) can be displayed together. Clinicians can easily compare and integrate information without having to remember data from one screen to the next. Although the core functionality of the app is modular (i.e., the identification of relevant data, display strategies, and algorithms are the same regardless of which EHR is used), the interface between the app and the EHR

must be tailored to each EHR to accommodate different application programming interfaces (APIs). The Screening and Surveillance App is depicted in ▶Figure 1.

The need for a conversation between patient and clinician about CRC screening is often triggered by a clinical reminder or alert in the EHR; however, it is often difficult to find the data needed to make a patient-centered CRC testing recommendation and have a meaningful conversation with the patient [19]. The Screening and Surveillance App is designed to ensure that the primary care clinician has key information at-a-glance when it is most needed. Although some CRC-related data may not be captured in the EHR, the clinician can be confident that whatever is available is currently displayed, reducing time spent looking for data that may or may not be available. Recommendations generated by the app are driven by national guidelines [9, 22], and by gastroenterologist recommendations based on previous testing. Similarly, risk factors include those recognized in the national guidelines (i.e., first-degree relative diagnosed with CRC).

1.5 Decision-Centered Design

The Screening and Surveillance App was designed and developed using a Decision-Centered Design Framework [23]. Decision-Centered Design focuses on designing to support decision making, sensemaking, and other macrocognitive functions during challenging and complex situations. Other Cognitive Engineering frameworks focus primarily on routine cases [24]. Decision support is most needed during challenging and unanticipated situations. Too often, systems are designed for routine operations but become brittle and inadequate when used in unexpected ways to address situations not anticipated in the design. Decision-Centered Design uses challenging situations as an organizing structure to guide the analysis, design, and evaluation of decision support applications. Although CRC screening is a routine part of care, clinicians often address CRC in the context of more complex macrocognitive activities, including determining how to allocate time during the patient encounter for patients with multiple issues, how to (efficiently) educate patients who may have misconceptions about colorectal cancer screening, and determining whether the patient has had prior CRC testing and what the findings were.

2. Evaluation Of The Screening and Surveillance App

We conducted both an online and an in-person evaluation in order to assess performance, workload, and usability of the Screening and Surveillance App. Our rationale was that meeting the design objective of reducing the need to search the EHR for CRC-related information would positively influence these aspects of clinical work. Research questions included: 1) Are participants able to answer questions about CRC-related patient data accurately using the App? 2) Does the App require less workload/mental effort than other EHRs to find CRC-related patient information? 3) Do participants rate the App to be highly usable and useful? For the in-person evaluation, we asked these questions as comparisons to the EHR currently used at VA medical centers. We asked an additional question: Are participants able to find CRC related patient data more efficiently using the App than the CPRS facsimile?

2.1 Methods

The online evaluation was determined to be exempt by the University of Dayton Institutional Review Board (IRB) because the research team had no access to personally identifiable information of survey respondents. All coordination with participants, including payment, was managed by a market research firm. The in-person evaluation was approved by the IRB of Indiana University Purdue University at Indianapolis and by the Research and Development Committee at the Richard L. Roudebush VA Medical Center, Indianapolis, IN.

2.1.1 Participants

For the online evaluation, a market research firm recruited 24 PCCs to participate. Inclusion criteria consisted of PCCs who treat patients who are eligible for CRC screening (aged 50–75) and who have

had at least three years of experience using an EHR. Participants were offered $50 as an incentive for participation. Our intent was to include at least 20 participants as recommended by NISTIR 7804 Electronic Health Record Usability Protocol [7] to capture most of the variance; however, we oversampled modestly in anticipation of potentially unusable data due to technical difficulties. No technical difficulties arose, and we were able to use data from all 24 participants.

For the in-person evaluation, we recruited 10 PCCs at the Richard L. Roudebush VA Medical Center in Indianapolis to participate in the study. Participants were recruited via email invitations and staff meetings. All participants were primary care clinicians with at least three years of experience using the VA's EHR, CPRS. A $100 gift card was offered as an incentive for participation.

Prior experience suggested that recruiting PCCs to participant in a 60-minute, face-to-face session would be challenging. Our goal was to recruit at least 10 participants. We consider this a meaningful sample based on prior usability research suggesting that with 10 participants, 80 percent of problems are found [25].

2.1.2 Materials

Online participants completed the evaluation independently on their own computer. The evaluation was administered using two online survey tools. Participants in the in-person comparison evaluation used a laptop computer and external monitor provided by the study team. Morae usability software was used to create a recording of the session, track time and mouse clicks, and deliver the Health ITUES. Materials also included:

Demonstration Video

For the online evaluation, we used a narrated, five-minute video demonstrating features of the Screening and Surveillance App. For the in-person comparison study, we used the same video and added a segment to explain how to access the App from the CPRS facsimile.

CPRS facsimile

For the in-person comparison study, we developed a CPRS facsimile so we could compare the App to the current EHR without the risk of using personally identifiable health information. We created the CPRS facsimile using Axure, an interactive wireframe tool. The CPRS user interface includes a desktop metaphor with tabs organizing different types of information. Tabs include a Cover Sheet, Problems, Meds, Orders, Notes, Consults, Surgery, D/C Summ, Labs, and Reports (▶Figure 2). It is not possible to display data from multiple tabs at the same time. To accommodate this limitation, many users open a second instance of CPRS and display it on a separate monitor. The Screening and Surveillance App was designed to reduce the effort required to find and examine data stored in different locations in the EHR. We added a button to the facsimile to access the Screening and Surveillance App for the evaluation.

Patient scenarios

For the online study, two screenshots of the app, each containing different mock patient data related to CRC were used. Mock patients were designed to represent challenging scenarios in which the PCC would need to integrate data from multiple sources to build a picture of the patient's CRC-screening history, risk, and recommended next steps.

For the in-person comparison study, we created four patient scenarios consisting of two pairs. The mock patients in each pair were very similar in terms of CRC testing history and risk factors. One mock patient of each pair was displayed using the Screening and Surveillance App. The other was displayed using the facsimile of CPRS.

Task Questions

To measure performance, we used a series of ten questions related to CRC screening for each patient. See ▶Table 1 for a list of task questions. In the online evaluation, participants read and entered their answers using the survey software. During the in-person evaluation, participants read and answered the questions using Microsoft Excel Visual Basic.

Subjective Workload Dominance (SWORD) scale

The SWORD is a comparative workload dominance measure that allows participants to compare the workload associated with using different interfaces to complete tasks.[26–27] Participants compared two interfaces (the Screening and Surveillance App and the EHR they typically use) and two tasks (find the date of the next recommended test and find the information needed to assess the patient's CRC risk factors). Pairwise comparisons for every combination of task and interface were made on a 17-point graphic rating scale. An example of the adapted SWORD used in the online evaluation is depicted in ▶Figure 3. Due to the difficulty of obtaining consistent data using the SWORD, it was only used for the online evaluation.

Rating Scale Mental Effort (RSME)

To replace the SWORD for the in-person evaluation, we used the RSME. The RSME [28] was used to measure participants' perceived mental effort. Participants were asked to make a horizontal mark on the vertical line to indicate how much mental effort was required. The RSME was completed on paper (▶Figure 4).

Health Information Technology Usability Evaluation Scale (Health ITUES)

The Health ITUES is a customizable usability instrument consisting of four subscales with predictive validity [29-30]. We used three of the four subscales: Quality of Life, Perceived Usefulness, and Perceived Ease of Use. The fourth subscale, User Control, was not administered because it had little relevance to the App. The participants recorded their response to each question on a 5-point graphic rating scale ranging from 1 (strongly disagree) to 5 (strongly agree). See ▶Table 2 for a list of the Health ITUES questions. For the in-person evaluation, a question was added to the Health ITUES: "I can easily remember how to access the Screening and Surveillance App."

2.1.3 Procedure

For both the online and in-person evaluations, participants watched a demonstration video describing how the Screening and Surveillance App worked, proceeded through mock patient scenarios, answered questions related to workload (online) or mental effort (in-person), then answered questions related to usability. Due to the different contexts for each evaluation, there were some differences in recruitment and experimental procedure.

For the online evaluation, the market research firm recruited participants by sending an email with a link to access the website for the study. Screening questions were administered to ensure participants met the inclusion criteria. Respondents who did not meet the inclusion criteria were directed to a separate webpage thanking them for their interest. Those meeting the inclusion criteria watched a demonstration video of the Screening and Surveillance App.

Next, participants worked through two patient scenarios and the task questions related to CRC screening. To control for order effects, the order of the two patient scenarios was counterbalanced: half of the participants received Scenario 1 first and half of the participants received Scenario 2 first.

Following the patient scenarios, participants completed the SWORD to rate the workload required to work through the scenarios using the Screening and Surveillance App in comparison to the EHR used in their practice. After completing the SWORD, participants responded to the Health ITUES to measure usability and usefulness. To conclude the study, participants answered demographic questions and were directed to the market research firm's website for payment.

In-person evaluation sessions were conducted individually at the Indianapolis VAMC and a satellite facility. Sessions lasted approximately 60 minutes. At the beginning of each session, participants were provided an overview and given a chance to ask questions. Participants were given an information sheet and consent was verbally obtained to record the session. Next, participants answered demographic questions and watched the demonstration video of the App. Before the evaluation began, the facilitator guided each participant through a training scenario with the App. The training scenario allowed participants to ask any App related questions and to become familiar with using Visual Basic to answer the task questions.

During the evaluation, participants worked through four patient scenarios and answered CRC-related questions, using the App for two scenarios and the CPRS facsimile for two scenarios. The order of App versus CPRS facsimile was counterbalanced so that half of the participants interacted

with the App first and half interacted with the CPRS facsimile first. Following each patient scenario, participants completed the RSME to measure mental effort.

After the participants completed all of the patient scenarios, they completed two final RSME ratings: 1) an overall retrospective rating of the mental effort required to use the Screening and Surveillance App to find the information needed to make CRC testing recommendations, and 2) an overall retrospective rating of the mental effort required to use the version of CPRS they routinely use in their clinic to accomplish the same task. Next, participants completed the Health ITUES. To conclude the study, the participants were thanked and offered a $100 gift card for participation.

2.1.4 Analysis

Responses were imported into SPSS for statistical analysis. Incomplete data sets were removed prior to analysis.

Accuracy

We assessed the accuracy of responses to the task questions. Answers to each question were coded as "0" for incorrect or "1" for correct. For the comparison evaluation, we used a paired-samples *t*-test to compare the mean number of correct responses (out of 10) between the App and the CPRS facsimile. In addition to accuracy, we were also able to collect data related to three more performance measures during the in-person evaluation: time, screens accessed, and mouse clicks.

Time

We calculated the time spent completing each patient scenario. We used a paired-samples *t*-test to compare the mean time spent using the App versus the CPRS facsimile.

Screens Accessed

We tracked the number of screens participants accessed when completing each patient scenario. We used a paired-samples *t*-test to compare the mean number of screens accessed using the App versus the CPRS facsimile.

Mouse Clicks

The number of mouse clicks performed during the patient scenarios was recorded using Morae. We used a paired-samples *t*-test to compare the mean number of mouse clicks when using the App versus the CPRS facsimile.

Workload

For the SWORD used in the online evaluation, we calculated mean workload dominance for each combination of interface and task across participants (higher means correspond to higher perceived workload). We used independent sample *t*-tests to compare the means. To determine whether participants are consistent in their scoring, Saaty [31] proposes what is called a Consistency Ratio. If the value of the Consistency Ratio is smaller or equal to 10%, the inconsistency is acceptable. We calculated the Consistency Ratio for all 25 participants. Only 8 out of 25 participants had acceptable Consistency Ratios. Two of the 8 selected "equal" for every comparison. While technically consistent, we suspect that these participants were trying to finish the survey as quickly as possible. As a result, we excluded these two participants from analysis resulting in 6 remaining participants.

Mental Effort

For the in-person evaluation, we measured mental effort with the RSME, where 0 represents absolutely no mental effort and 150 represents the highest possible mental effort. Participants completed the RSME after each patient scenario and at the end of the study. We used a paired-samples *t*-test to compare the mean mental effort reported.

Usability

We measured usability with the Health ITUES, where 1 is the least positive response and 5 is the most positive. We calculated mean responses and standard deviations for each question. Ratings above a 3 were considered positive and ratings below 3 were considered negative.

2.2 Results

The most commonly used EHR reported by respondents was Epic, followed by Allscripts and NextGen. ▶Table 3 presents a list of all 14 EHRs mentioned. ▶Table 4 summarizes the participant demographics.

A summary of the performance, workload, and usability results of the online evaluation can be found in ▶Table 5, and a summary of the results of the in-person evaluation can be found in ▶Table 6. Results are described in detail below.

2.2.1 Performance

Performance was measured by the accuracy of responses to the task questions. Across both patient scenarios in the online evaluation, participants scored a mean of 8.71 out of 10 questions correct (*SD* = 1.33). We found no significant difference in accuracy between the two patient scenarios: the means were 8.75 (*SD* = 1.65) for Scenario 1 and 8.75 (*SD* = 1.33) for Scenario 2, $t(23) = 0.00$, $p = 1.00$ (data not shown). Similarly, we found no significant order effects between first and second scenario: the means were 8.63 (*SD* = 1.58) for the first scenario and 8.79 (*SD* = 1.38) for the second scenario, $t(23) = -0.62$, $p = 0.54$ (data not shown).

For the in-person evaluation, we measured performance in four ways: accuracy, time, screens accessed, and mouse clicks.

Accuracy

Across all patient scenarios, participants performed significantly better using the App than the CPRS facsimile. The App mean was 9.15 (*SD* = 0.78) and the CPRS mean was 6.95 (*SD* = 1.19), $t(9) = 6.41$, $p < 0.001$.

Time

Participants performed significantly faster using the App than the CPRS facsimile. The App mean was 187.31 seconds (*SD* = 57.18) and the CPRS mean was 262.90 seconds (*SD* = 63.81), $t(9) = -4.42$, $p = 0.002$, 95% CI [−114.31, −36.86]. The relative ratio is 1.47%.

Screens Accessed

Participants accessed significantly fewer screens while using the App than the CPRS facsimile. The App mean was 3.45 (*SD* = 0.76) and the CPRS mean was 10.45 (*SD* = 3.49), $t(9) = -6.36$, $p < 0.001$.

Mouse Clicks

Participants performed fewer mouse clicks using the App than the CPRS facsimile; however, the difference was not significant $t(7) = 2.32$, $p = 0.053$. The App mean was 15.06 (*SD* = 4.89) and the CPRS mean was 24.56 (*SD* = 10.98). Note: we only analyzed data for 8 of the participants due to a Morae system error.

2.2.2 Workload

For the online evaluation, six of the 24 respondents provided consistent data. Analysis of these 6 responses suggests that for both tasks, the Screening and Surveillance App required less workload than the other EHRs used by participants. For Task A (find the information needed to assess the patient's colorectal risk factors), participants rated the App as requiring significantly less workload than their own EHR. The App mean was 12.25 (*SD* = 8.64) and the EHR mean was 38.38 (*SD* = 13.46), $t(5) = 3.80$, $p = 0.013$, $d = 1.55$, power = 0.96. For Task B (find the date of the next recommended test) participants rated the App as requiring significantly less workload than their own EHR. The App mean was 13.98 (*SD* = 6.86) and the EHR mean was 33.85 (*SD* = 15.21), $t(5) = 2.67$, $p = 0.045$, $d = 1.12$, power = 0.77.

2.2.3 Mental Effort

For the in-person evaluation, participants rated the App as requiring significantly less mental effort than the CPRS facsimile while they worked through the patient scenarios. The App mean was 20.78 (*SD* = 8.28) and the CPRS mean was 51.88 (*SD* = 20.72), $t(9) = -4.42$, $p = 0.002$. Retrospective rat-

ings show an even larger gap, with a mean rating of 16.90 (*SD* = 8.22) for the App, and a mean rating of 55.65 (*SD* = 19.36) for the version of CPRS they routinely use, $t(9) = -7.77$, $p < 0.001$.

2.2.4 Usability

The mean score across all Health ITUES subscales was 4.26 out of 5 (*SD* = 0.61) for the online evaluation, and 4.67 out of 5 (*SD* = 0.37) for the in-person comparison. All three subscales also received highly positive mean ratings for both evaluations. The Quality of Work Life subscale's mean for the online evaluation was 4.24 (*SD* = 0.84) and 4.67 (*SD* = 1.44) for the in-person evaluation. The Perceived Usefulness subscale's mean was 4.27 (*SD* = 0.58) for the online evaluation and 4.59 (*SD* = 0.46) for the in-person evaluation. The Perceived Ease of Use subscale's mean was 4.24 (*SD* = 0.69) for the online evaluation and 4.83 (*SD* = 0.31) for the in-person evaluation.

2.3 Discussion

Findings from both evaluations suggest that the Screening and Surveillance App is an effective decision support tool. Participants were able to answer questions about CRC related patient data accurately using the App. The App required less workload than other EHRs to find CRC related patient information. Participants rated the App to be highly usable and useful. Moreover, in comparison to the CPRS facsimile, the participants in the in-person evaluation showed improvements in performance, perceived mental effort, and usability with the Screening and Surveillance App. Participants were able to answer questions about CRC related patient data more accurately using the App compared with the CPRS facsimile. Participants were able to find CRC related patient data more efficiently using the App than the CPRS facsimile (based on time and screens accessed). Participants completed patient scenarios 29% faster with the App than the CPRS facsimile. Participants perceived the App to require less mental effort to use than the CPRS facsimile. Participants ranked the App as requiring "almost no effort" (13, 9%) to "a little effort" (26, 17%). Participants ranked CPRS as requiring "some effort" (37, 25%) to "rather much effort" (57, 38%). While participants did not compare the usability of the App to CPRS, they did rate the App very highly indicating that participants perceived the App as being useful and usable.

With regard to the scenarios used for the online evaluation, we found no differences in performance between the two patient scenarios, suggesting that they were equal in difficulty as intended. In addition, we found no order effects, suggesting that the order of scenario presentation did not significantly influence performance.

With regard to measurement instruments, we found the Health ITUES to be an effective tool for obtaining feedback about usability. The use of the SWORD for rating workload in the online evaluation was not as effective. Data from only six respondents were usable, reducing the robustness of workload findings. We initially selected the SWORD because it was designed to compare workload across interfaces thus allowing us to collect comparative data. Furthermore, research suggests that retrospective ratings of workload are more accurate than concurrent ratings [32]. However, we underestimated the complexity involved in making consistent ratings. The developers of the SWORD recommend a tutorial/training session for participants. Although we offered brief online training consisting of a practice rating of two familiar tasks using personal electronic devices, it was not sufficient. A more in-depth training session was not feasible for this evaluation given time limitations. We concluded that a simpler, unidimensional workload assessment instrument, such as the Rating Scale Mental Effort (RSME) [28] would be more effective for future studies with similar limitations.

One of the limitations of the in-person evaluation involved the use of a simulated setting. Many clinicians said that during a patient encounter, they would often ask patients about difficult-to-find information such as family history (See [9, 22] for relevant aspects of family history). This conversation is important not just for collecting historical data, but also for eliciting new information such as a relative recently diagnosed with CRC. Thus, asking the participants to rely solely information in the EHR without patient input was an artificiality of the study. The App is not intended to replace this conversation with the patient. Rather, it may reduce the need for the patient to recall data already captured in the EHR, preserving time for discussion of new information.

Another limitation of the simulation is that it is an imperfect representation of clinical workflow. For example, we would not expect clinicians to search for answers to each of the CRC-related questions for each patient. The time savings reported suggest that the Screening and Surveillance App would allow clinicians to find relevant data quickly, but we do not have an accurate view of the time savings likely to be realized during a patient encounter. Similarly, to further understand the importance of the perceived difference in mental workload will require additional testing of the App in a clinical setting.

With regard to overall limitations, it is important to point out that participants were volunteers who were offered $50 or $100 for participation in the online and in-person evaluations, respectively. Asking for volunteers may have skewed our sample to include clinicians who are looking for improvements in health information technology. Although experimenters made every effort to use neutral language (i.e., using language such as: "Our goal in this evaluation is to understand how easy or difficult the application is to use…"), the use of an incentive may have primed participants to respond favorably to the software being evaluated. With regard to experimental control, the study design would have been cleaner if we had recruited participants who had no experience with CPRS. For this study, however, we were particularly interested in exploring how the Screening and Surveillance App might change work for experienced clinicians using the Screening and Surveillance App in conjunction with their existing EHR. Another limitation with regard to recruiting is that we did not collect data about the study populations for comparison with the total eligible population, or the number of PCCs invited to participate before we achieved the desired sample size, limiting our ability to determine how well those who chose to participate represent the population of PCCs who use EHRs.

It is also worth noting that time and accuracy comparisons were made to a single EHR. Responses from the online evaluation participants who used a range of EHRs suggest that participants found that the Screening and Surveillance App required less workload to find CRC-related data than the EHRs they use every day. However, we do not have quantitative comparison data for other EHR interfaces.

3. Conclusion

This project demonstrates the successful application of the Decision-Centered Design Framework to the design of Health Information Technology. By designing scenarios representing challenging situations and including cognitive performance measures (e.g., key elements of Decision-Centered Design), we were able to examine the utility of the Screening and Surveillance App for filling decision support gaps in current EHRs. The positive performance on accuracy, efficiency, perceived workload/mental effort, and usability speak to the potential benefits of modular software applications and the larger design approach. It is possible that these benefits would be enhanced over time, as clinicians would become more efficient using the Screening and Surveillance App with practice.

This evaluation was innovative in two ways. First, we leveraged two evaluation formats, online and in-person. The online evaluation allowed us to obtain feedback from a broad range of PCCs; the in-person evaluation allowed us to obtain comparative data. These are common challenges involved in the evaluation of Health Information Technology. Rather than choosing a single evaluation format, we advocate leveraging multiple evaluation formats to compensate for the limitations of each. Using two evaluation formats allowed for a more complete picture of the potential impact of the Screening and Surveillance App. Second, we included additional measures (accuracy, efficiency, perceived workload, and usability) that extend the safety-focused measures recommended in the NISTIR 7804. This allowed us to more fully assess the ability of the Screening and Surveillance App to support clinicians in quickly finding and integrating the information they need to make timely, evidence-based, patient-centered screening recommendations.

4. Multiple Choice Question

Which assessment tool is ideal for measuring the workload of health information technology when time limitations are a concern?

- A. SWORD (Subjective Workload Dominance)
- B. RSME (Rating Scale Mental Effort)
- C. NASA TLX (Task Load Index)
- D. Health ITUES (Information Technology Usability Evaluation Scale)

Answer: (B. RSME) Including a measure for perceived workload is an important component of anticipating the impact of a new technology on operator performance and satisfaction. After implementing the SWORD during the online evaluation and the RSME during the in-person evaluation, we concluded that a simpler, unidimensional workload assessment instrument, such as the RSME would be more effective for future studies with similar time limitations.

Clinical Relevance Statement

Strategies for evaluating clinical decision support are evolving as the need to understand the potential impact of health information technology on decision making, workflow, and cognitive performance increase. The NISTIR 7804 safety measures provide a valuable foundation. To extend this foundation it is critical that future evaluations include challenging scenarios and measures tailored to assess cognitive performance in the context of specific tasks, in combination with more general measures of workload and safety.

Conflict Of Interest

The Screening and Surveillance App evaluated in this project was designed and developed under the same contract that funded this evaluation. Many of the authors of this manuscript contributed to both the design and the evaluation of the health education materials. Laura Militello is co-owner of Applied Decision Science where the work was undertaken. Donald Haverkamp is a representative of the sponsoring agency that funded this work.

Human Subject Research Approval

The procedures used in this project have been reviewed in compliance with ethical standards of the University of Dayton and the Indiana University Institutional Review Boards and with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects (http://www.wma.net/en/30publications/10policies/b3/17c.pdf).

Acknowledgments

We would like to thank all the busy primary care clinicians who volunteered to participate in this evaluation study. We would also like to thank Dr. Matthew Bair who helped with the design of scenarios and prototypes used in the evaluation.

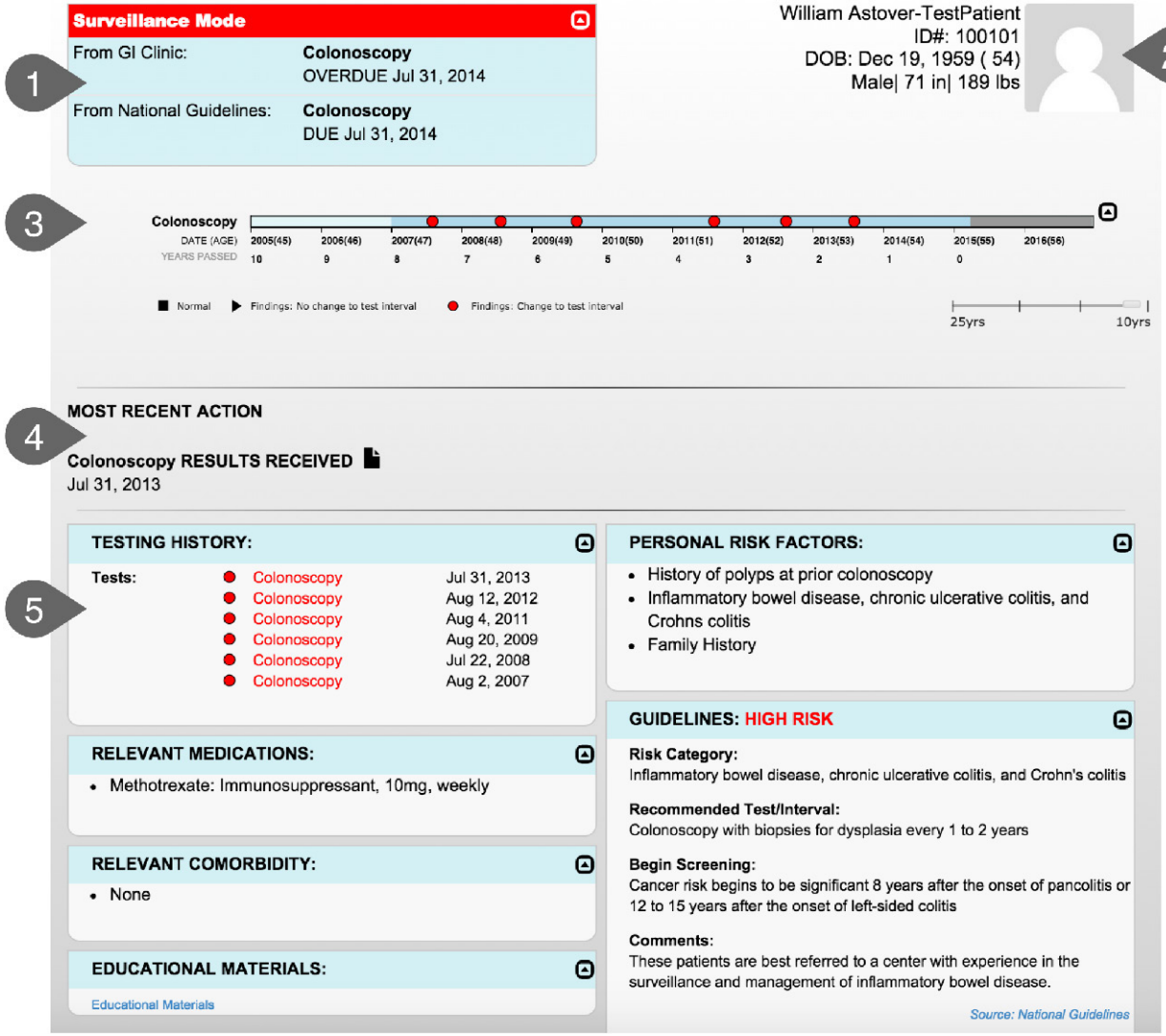

Fig. 1 1. At-A-Glance Information: Provides key information including whether the patient is in screening or surveillance mode, recommended testing options, and due date of next test. 2. Demographics: Provides basic information about the patient such as age and date of birth. A photo is included to reduce wrong patient errors. 3. Interactive Timeline: Provides a way to quickly visualize screening history including test results. 4. Most Recent Action: Quickly locate where the patient is in the testing process (i.e., Test Ordered, Test Scheduled, Test Completed, Results Received, or Test Declined). 5. Patient-Centered Dashboard: Provides additional information and links to relevant guidelines for complex cases. The dashboard includes a tabular format of the timeline, personal risk factors, relevant guideline recommendations, relevant medications and comorbidities, and links to educational materials and the guidelines.

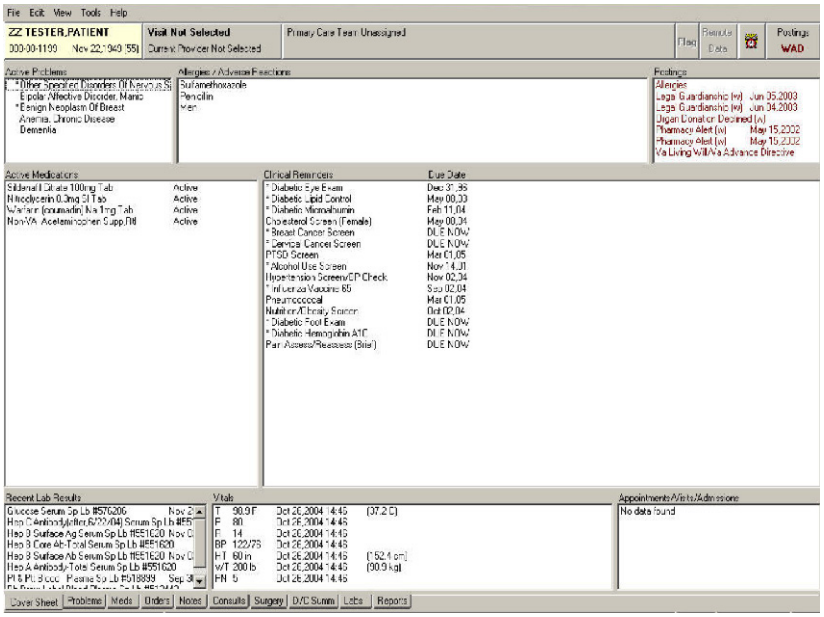

Fig. 2 CPRS user interface uses a desktop metaphor with tabs organizing different types of information

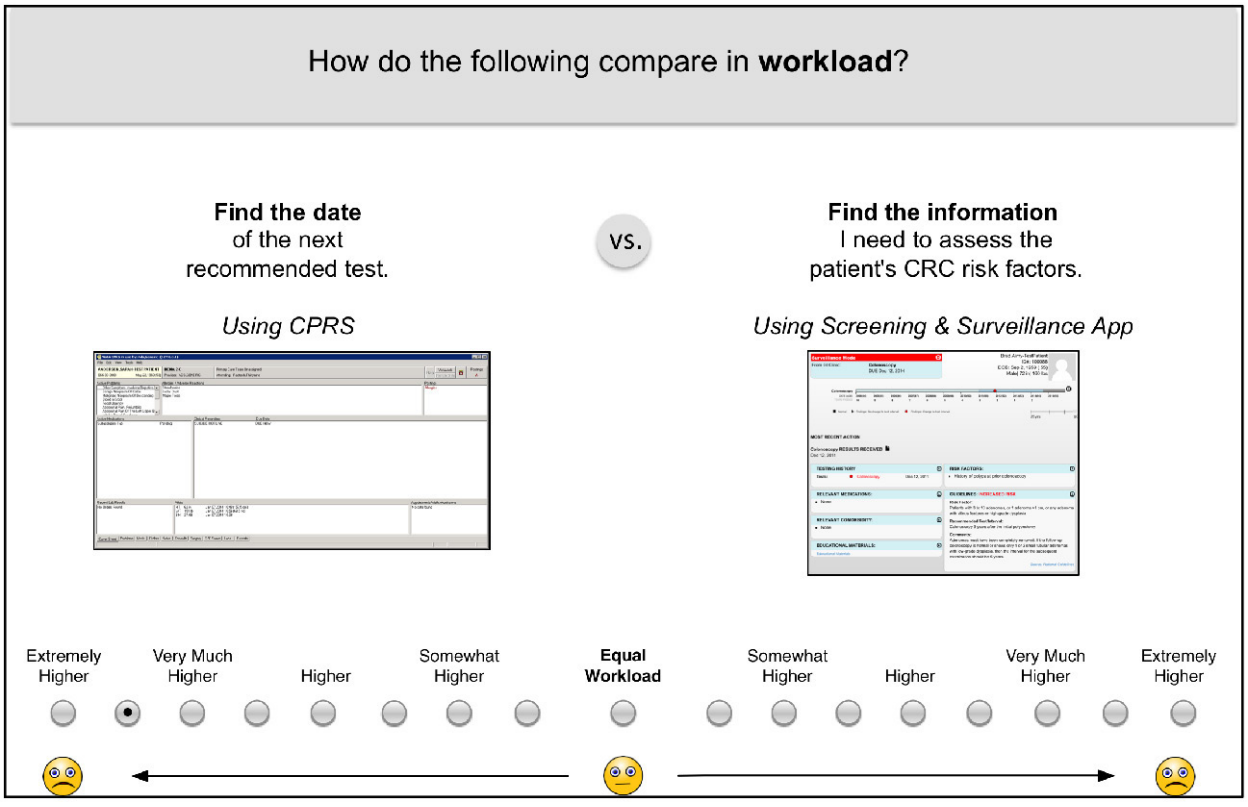

Fig. 3 Adapted Subjective Workload Dominance (SWORD) scale to measure perceived workload.

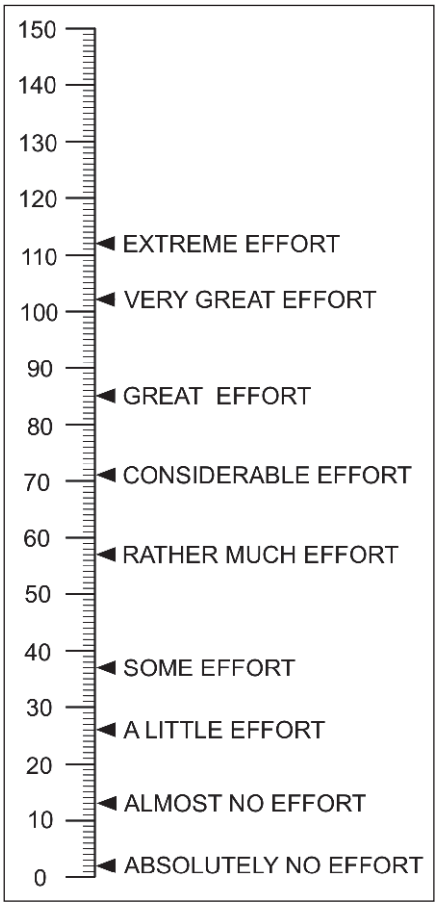

Fig. 4 Rating Scale of Mental Effort (RSME)

Table 1 Task Questions

1. Please describe your colorectal cancer testing recommendation for this patient.
2. Which risk factor makes this patient eligible for colorectal cancer testing now?
3. Is this patient in screening or surveillance mode?
4. Does the patient have a family history relevant to colorectal cancer?
5. Is the patient taking any medications that could be relevant to colorectal cancer testing (if yes, please specify)?
6. What was the date of the patient's most recent colorectal cancer test?
7. Were there any findings from the patient's most recent colorectal cancer test that could change the testing interval (even if it didn't change the interval for this patient)?
8. Which colorectal cancer test modalities has this patient used?
9. What was the most recent action taken for the patient relevant to colorectal cancer testing?
10. Does the date of the GI clinic's recommended next test match the date of the national guideline's recommended next test?

Table 2 Health ITUES Questions

Quality of Work Life
1. I think the SSA would have a positive impact on CRC testing rates.
2. I think the SSA would be an important part of tracking and managing patients' CRC testing.
3. I think the SSA would increase my confidence in CRC testing recommendations.
Perceived Usefulness
4. Using the SSA makes it easier to find information relevant to CRC testing for my patients.
5. Using the SSA enables me to find information about my patients' CRC testing more quickly.
6. Using the SSA makes it more likely that I will be able to make patient-centered recommendations about CRC testing.
7. The SSA is useful for making CRC testing recommendations.
8. I am satisfied with the SSA for providing the necessary information to make CRC testing recommendations.
9. I can make CRC testing recommendations in a timely manner because of the SSA.
10. Using the SSA would increase productive discussions about CRC testing with my patients.
11. would be able to find relevant patient info for making CRC testing recommendations whenever I use the SSA.
Perceived Ease of Use
12. I am comfortable with my ability to use the SSA.
13. Learning to operate the SSA was easy for me.
14. I would be easy for me to become skillful at using the SSA.
15. I find the SSA easy to use.

EHR	Frequency
Epic	8
AllScripts	5
Next Gen	5
Practice Fusion	4
E-Clinical Works	3
Centricity	3
Cerner	3
Athena	1
Davlone	1
Emd	1
Greenway Prime Suite	1
Spring Charts	1
USAR Health Records	1
Visionary	1

Table 3 Electronic Health Records used by online evaluation respondents

Demographic		Frequency (%)			
		Online		In Person	
		N = 24		N = 10	
Years as a Primary Care Clinician	0–5	2	(8.3%)	2	(20%)
	6–10	6	(25.0%)	3	(30%)
	11–15	5	(20.8%)	2	(20)%
	16–20	4	(16.7%)	2	(20%)
	21+	7	(29.2%)	1	
Percent Time Clinical (versus admin, training, etc.)	0–25%	1	(4.2%)	2	(20%)
	26–50%	0		2	(20%)
	51–75%	0		0	
	76–100%	23	(95.8%)	6	(60%)
Percent of Patients Aged 50–75	0–25%	0			
	26–50%	11	(45.8%)		
	51–75%	12	(50.0%)		
	76–100%	1	(4.2%)		
Years of EHR experience	3–5	10	(41.7%)	5	(50%)
	6–8	8	(33.3%)	2	(20%)
	9–11	4	(16.7%)	1	(10%)
	12+	2	(8.3%)	2	(20%)

Table 4 Participant demographics

Table 5 Results of Online Evaluation (N = 24)

Measure	App Mean (SD)	EHR Mean (SD)	*t*(df)	*p*
Performance				
Accuracy	8.71 (1.33)	-	-	-
Workload				
SWORD Task A	12.25 (8.64)	38.38 (13.46)	3.80 (5)	0.013
SWORD Task B	13.98 (6.86)	33.85 (15.21)	2.67 (5)	0.045
Usability				
Health ITUES Avg.	4.26 (0.61)	-	-	-
Quality of Work Life	4.24 (0.84)	-	-	-
Perceived Usefulness	4.27 (0.58)	-	-	-
Perceived Ease of Use	4.24 (0.69)	-	-	-

Table 6 Results of In-Person Comparison (N = 10)

Measure	App Mean (SD)	App Min./Max. (Range)	CPRS Mean (SD)	CPRS Min./ Max. (Range)	*t*(df)	*p*
Performance						
Accuracy	9.15 (0.78)	8.00–10.00 (2)	6.95 (1.19)	4.50–8.50 (4)	6.14 (9)	< 0.001
Time	187.31 (57.18)	115.00–281.65 (166.65)	262.90 (63.81)	154.00–345.50 (191.50)	-4.42 (9)	= 0.002
Screens Accessed	3.45 (0.76)	3.00–5.00 (2.00)	10.45 (3.49)	6.00–17.50 (11.50)	-6.36 (9)	< 0.001
Mouse Clicks	15.06 (4.89)	6.50–23.50 (17.00)	24.56 (10.98)	12.50–50.00 (37.50)	2.32 (7)	= 0.053
Mental Effort						
RSME Patient Scenarios	20.78 (8.28)	7.05–31.50 (24.00)	51.88 (20.72)	25.00–85.00 (60.00)	-4.42 (9)	= 0.002
RSME Retrospective	16.90 (8.22)		55.65 (19.36)		-7.77 (9)	< 0.001
Usability						
Health ITUES Avg.	4.67 (0.37)		-		-	-
Quality of Work Life	4.67 (0.44)		-		-	-
Perceived Usefulness	4.59 (0.46)		-		-	-
Perceived Ease of Use	4.83 (0.31)		-		-	-

References

1. Kaplan B. Evaluating informatics applications-clinical decision support systems literature review. Int J Med Inform 2001; 64: 15–37.
2. Kaplan B. Evaluating informatics applications-some alternative approaches: theory, social interactionism, and call for methodological pluralism. Int J Med Inform 2001; 64: 39–56.
3. Littlejohns P, Wyatt JC, Garvican L. Evaluating computerised health information systems: hard lessons still to be learnt. BMJ 2001; 326: 860–863.
4. Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykänen P, Rigby M. STARE-HI-Statement on reporting of evaluation studies in Health Informatics. Int J Med Inform 2009; 78: 1–9.
5. Hanauer D, Zheng K. Measuring the impact of health information technology. Applied clinical informatics 2012; 3(3): 334–336.
6. Ammenwerth E, Brender J, Nykänen P, Prokosch HU, Rigby M, Talmon J. Visions and strategies to improve evaluation of health information systems: Reflections and lessons based on the HIS-EVAL workshop in Innsbruck. Int J Med Inform 2004; 73: 479–491.
7. Lowry SZ, Quinn MT, Ramaiah M, Schumacher RM, Patterson ES, North R, Zhang, J, Gibbons MC, Abbot P. Technical evaluation, testing, and validation of the usability of electronic health records. National Institute of Standards and Technology Interagency/Internal Report (NISTIR) 7804. 2012.
8. Lowry SZ, Ramaiah M, Patterson ES, Brick D, Gurses AP, Ozok A, Simmons D, Gibbons MC. Integrating Electronic Health Records into Clinical Workflow: An Application of Human Factors Modeling Methods to Ambulatory Care. National Institute of Standards and Technology Interagency/Internal Report (NISTIR) 7988. 2014.
9. Levin B, Lieberman DA, McFarland B, Smith RA, Brooks, D, Andrews KS, Dash C, Giardiello FM, Glick S, Levin TR, Pickhardt P. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. CA Cancer J Clin 2008; 58: 130–160.
10. Frazier AL, Colditz GA, Fuchs CS, Kuntz KM. Cost-effectiveness of screening for colorectal cancer in the general population. JAMA 2000; 284(15): 1954-1961.
11. Zauber AG, Winawer SJ, O'Brien MJ, Landsdorp-Vogelaar I, van Ballegooijen M, Hankey BF, Shi W, Bond JH, Schapiro M, Panish JF, Stewart ET. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N Engl J Med 2012; 366(8): 687-696.
12. Subramanian S, Klosterman M, Amonkar MM, Hunt TL. Adherence with colorectal cancer screening guidelines: a review. Prev Med 2004; 38: 536–550.
13. Brawarsky P, Brooks DR, Mucci LA, Wood PA. Effect of physician recommendation and patient adherence on rates of colorectal cancer testing. Cancer Detect Prev 2004; 28: 260–268.
14. Janz NK, Wren PA, Schottenfeld D, Guire KE. Colorectal cancer screening attitudes and behavior: a population-based study. Prev Med 2003: 627–634.
15. Hudson SV, Ohman-Strickland P, Cunningham R, Ferrante JM, Hahn K, Crabtree BF. The effects of teamwork and system support on colorectal cancer screening in primary care practices. Cancer Detect Prev 2007; 31: 417–423.
16. Sarfaty M, Wender R. How to increase colorectal cancer screening rates in practice. CA Cancer J Clin 2007; 57: 354–366.
17. Saleem JJ, Militello LG, Arbuckle N, Flanagan M, Haggstrom DA, Linder JA, Doebbeling BN. Provider perceptions of colorectal cancer screening decision support at three benchmark institutions. In: AMIA Annual Symposium Proceedings 2009; 558-562.
18. Lowry SZ, Ramaiah M, Patterson ES, Prettyman SS, Simmons D, Brick D, Latkany P, Gibbons MC. Technical Evaluation, Testing, and Validation of the Usability of Electronic Health Records: Empirically based use cases for validating safety – enhanced usability and guidelines for standardization. National Institute of Standards and Technology Interagency/Internal Report (NISTIR) 7804–1. 2015.
19. Militello LG, Saleem JJ, Borders MR, Sushereba CE, Haverkamp D, Wolf SP, Doebbeling BN. Designing colorectal cancer screening decision support: a cognitive engineering enterprise. Journal of Cognitive Engineering and Decision Making 2016; 10(1): 74–90.
20. Kawamoto, K. Integration of knowledge resources into applications to enable clinical decision support: architectural considerations. In: Greenes R, ed. Clinical Decision Support: The road ahead. Burlington, MA: Elsevier 2007; 503–539.
21. Sim I, Gorman P, Greenes R, Haynes RB, Kaplan B, Lehnmann H, Tang PC. Clinical decision support systems for the practice of evidence-based medicine. J Am Med Inform Assoc 2001; 8(6): 527–534.
22. U.S. Preventive Services Task Force. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Annals of Internal Medicine 2008: 1406–1422.

23. Militello LG, Dominguez CO, Lintern G, Klein G. The Role of Cognitive Systems Engineering in the Systems Design Process. Systems Engineering 2009; 13(3): 261–273.
24. Kirwan B, Ainsworth LK. A Guide to Task Analysis. London: Tailor & Francis 1992.
25. Faulkner, Laura. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. Behavior Research Methods, Instruments, & Computers 2003; 35(3): 379–383.
26. Vidulich MA, Tsang PS. Absolute magnitude estimation and relative judgment approaches to subjective workload assessments. In: Proceedings of the Human Factors Society 31st Annual Meeting 1987; 1057–1061.
27. Vidulich MA. The use of judgment matrices in subjective workload assessment: The Subjective Workload Dominance (SWORD) technique. In: Proceedings of the Human Factors Society 33rd Annual Meeting 1989; 1406–1410.
28. Zijlstra F, Van Doorn L. The construction of a subjective effort scale. Technical Report, Delft University of Technology 1985; 40.
29. Yen PY, Wantland D, Bakken S. Development of a customizable health IT usability evaluation scale. In: AMIA Annual Symposium Proceedings 2010; 917.
30. Yen PY, Sousa KH, Bakken S. Examining construct and predictive validity of the Health-IT Usability Evaluation Scale: confirmatory factor analysis and structural equation modeling results. J Am Med Inform Assoc 2014; 21(2): 241–248.
31. Saaty TL. How to make a decision: The analytic hierarchy process. Eur J Oper Res 1990; 48(1): 9–26.
32. Vidulich MA, Ward GF, Schueren J. Using the Subjective Workload Dominance (SWORD) Technique for Projective Workload Assessment. Hum Factors 1991; 33(6): 677–691.