Complementary methods of system usability evaluation: Surveys and observations during software design and development cycles

Jan Horsky, Kerry McColgan, Justine E. Pang, Andrea J. Melnikas, Jeffrey A. Linder, Jeffrey L. Schnipper, Blackford Middleton

Clinical Informatics Research and Development, Partners HealthCare, Boston, USA
Division of General Medicine and Primary Care, Brigham and Women’s Hospital, Boston, USA
Harvard Medical School, Boston, USA

Abstract

Poor usability of clinical information systems delays their adoption by clinicians and limits potential improvements to the efficiency and safety of care. Recurring usability evaluations are therefore, integral to the system design process. We compared four methods employed during the development of outpatient clinical documentation software: clinician email response, online survey, observations and interviews. Results suggest that no single method identifies all or most problems. Rather, each approach is optimal for evaluations at a different stage of design and characterizes different usability aspect. Email responses elicited from clinicians and surveys report mostly technical, biomedical, terminology and control problems and are most effective when a working prototype has been completed. Observations of clinical work and interviews inform conceptual and workflow-related problems and are best performed early in the cycle. Appropriate use of these methods consistently during development may significantly improve system usability and contribute to higher adoption rates among clinicians and to improved quality of care.

1. Introduction

There is a broad consensus among healthcare researchers, practitioners and administrators that although health information technology has the potential to reduce the risk of serious injury to patients in hospitals, significant differences remain among the multitude of electronic health record (EHR) systems with respect to their ability to achieve high safety, quality and effectiveness benchmarks [1–4]. In many instances, the intrinsic potential of EHRs for preventing and mitigating errors continues to be only partially realized and some implementations may, paradoxically, expose clinicians to new risks or add extra time to many routine interactions [5,6].

Research evidence and published reports on the successes, failures, best-practices, lessons learned and barriers overcome during implementation efforts have had only limited effect so far on accelerating the adoption of electronic information systems [7]. According to conservative estimates, at least 40% of systems either are abandoned or fail to meet business requirements, and fewer than 40% of large vendor systems meet their stated goals [8]. A recent national study reported that only four percent of physicians used a fully functional, advanced system and that 13% used systems with only basic functions [9].

Transition from paper records to electronic means of information management is an arduous process at large institutions and private practices alike. It introduces new standards and reshapes familiar practices often in ways unintended or unanticipated by the stakeholders. Clinicians object to forced changes in established workflows and familiar practices, long training times, and excessive time spent serving the computer rather than providing care [10,11].

Although the initial decline in efficiency generally improves with increased skills and sufficient time to adjust to new routines [12], systems themselves rarely evolve to better meet the demands and requirements of the clinical processes they need to support. A recent survey found an increase in the availability of EHRs over two years in one state, but the researchers also reported that routine use of ten core functions remained relatively low, with more than one out of five physicians not using each available function regularly [13]. An observational study of 88 primary care physicians identified key information management goals, strategies, and tasks in ambulatory practice and found that nearly half were not fully supported by available information technology [14].
Developing highly functional, versatile clinical information systems that can be efficiently and conveniently used without extensive training periods is predicated on incorporating rigorous and frequent usability evaluations into the design process. Iterative development methodology for graphical interfaces suggests evaluating and revising successive prototypes in a cyclical fashion until the product attains required characteristics. There are several common techniques that can be used to perform the evaluations that are either carried out entirely by usability experts or involve the input of intended users. Equally important is to see usability evaluation as situated within the context of challenges imposed by complex socio-technical systems [15] and within broader conceptual frameworks for design and evaluation such as those based on the theory of distributed cognition and work-centered research [16].

The broad objective of this study was to compare data gathered by four usability evaluation methods and discuss their respective utility at different stages of the software development process. We hypothesized that no single method would be equally effective in characterizing every aspect of the interface and human interaction. Rather, an approach that employs a set of complementary methods would increase their cumulative explanatory value by applying them selectively for specific purposes. Our narrower goal was to formulate recommendations for designers and evaluators of health information systems on the effective use of common usability inspection methods during the design and development cycle. This report expands a brief discussion of methods used in the design, pilot testing, and evaluation of the Smart Form in a previous publication [17].

2. Background

The reasons why one system may be preferred over another by clinicians and perform closer to expectations are often complex, vary with local conditions and almost always include financing, leadership, prior experience and training. Among the core predictors of quick adoption and successful implementation are the design quality of the graphical user interface and functionality, along with socio-technical factors [7]. Usability has a strong, often direct relationship with clinical productivity, error rate, user fatigue and user satisfaction that are critical for adoption. The system must be fast and easy to use, and the user interface must behave consistently in all situations [18]. At the same time, the system must support well all relevant clinical tasks so that a clinician working with the computer can achieve higher quality of care. The Healthcare Information and Management Systems Society (HIMSS) considers poor usability characteristics of current information technology as one of the major factors, and “possibly the most important factor” hindering its widespread adoption [19].

Historically, developers and designers have failed to tap the experiential expertise of practicing clinicians [20]. The lack of a systematic consideration of how clinical and computing tasks are performed in the situational context of different clinical environments often results in designs that are off the intended mark and fail to deliver improvements in safety and efficiency. For example, in an experiment that examined the interactive behavior of clinicians entering a visit note, researchers compared the sequence and flow of items on an electronic note form that was implied by the designed structure to actual mouse movements and entry sequences recorded by a tracking software and found substantial difference between the observed behavior and prior assumptions by the designers [21].

Existing usability studies mainly employ research designs such as expert inspection, simulated experiments, and self-reported user satisfaction surveys. Unfortunately, a large body of research indicates that self-reports can be a highly unreliable source of data, often context-dependent, and even minor changes in question wording, format or order can profoundly affect the obtained results [22].

While analyses that rely predominantly on a single method may produce incomplete or unreliable results, there is considerable evidence of the effectiveness of comprehensive approaches that combine two or more methods, as important redesign ideas rarely emerge as sudden insights but may evolve throughout the work process [23,24]. For example, during the development of a decision support system, designers employed field observations, structured interviews, and document analyses to collect and analyze users’ workflow patterns, decision support goals, and preferences regarding interactions with the system, performed think-aloud analyses and used the technology acceptance model to direct evaluation of users’ perceptions of the prototype [25]. A careful workflow analysis could lead to the identification of potential breakdown points, such as vulnerabilities in hand-offs, and communication tasks deemed critical could be required to have a traceable electronic receipt acknowledgment [26]. The advantage of informing the design from its conception with close insights into local needs and actual practices the software will support is reflected in the fact that “home-grown” systems show a higher relative risk reduction than commercial systems [1].

Iterative development of user interfaces involves the steady refinement of the design based on user testing and other evaluation methods [27]. The complexity and variability of clinical work requires correspondingly complex information systems that are virtually impossible to design without usability problems in a single attempt. Experts need to create a situation in which clinicians can instill their knowledge and concern into the design process from the very beginning [28]. Changing or redesigning a software system as complex as an EHR after it has been developed (or implemented) is enormously difficult, error-prone, and expensive [29,30]. Iterative evaluations early in the process allow larger conceptual revisions and refinements to be done without excessive effort and resources [31].

The software developed, tested and deployed in a pilot program in this study, the Coronary Artery Disease (CAD) and Diabetes Mellitus (DM) Smart Form (Fig. 1), was a prototype of an application intended to assist clinicians with documenting and managing the care of patients with chronic diseases [17]. Integrated within an outpatient electronic record, it allowed direct access to laboratory and other coded data for expedient entry into new visit notes. The Smart Form also aggregated reviewing of prior notes and laboratory results to create disease-relevant context for the planning of care, and provided actionable decision support and best-practices recommendations. The anticipated benefit to clinicians includes savings in time required to look up, collect, interpret and record clinical data into a note, and an increase in the quality and completeness of documentation that may contribute to improved patient care.

In the planning stage of the development, two experts, including a physician, conducted focus groups with approximately 25 physicians who described their usual workflows, methods for acute and chronic disease management, attitudes towards decision support, and their wants and needs, and summarized emerging themes [17].

3. Methods

We have conducted four different studies of usability and human–computer interaction that were intended to collect two types of data: comments elicited directly from clinicians working with the Smart Form, and findings derived from formal evaluations by
usability experts. We rigorously maintained distinctions between direct, free-style comments made by clinicians and objective findings by usability experts. Comments were always direct expressions of clinicians that originated either spontaneously or in response to a question, written or verbal. Findings, on the other hand, were expert opinions and recommendations based on field notes, interviews, focus groups and on direct observation of clinicians interacting with the Smart Form.

The reason why we chose to count and compare comments and findings instead of actual problems is the uncertainty in determining whether any two or more user reports describe identical problems, as comments may sometimes be vague, too general or without the proper context to match them to unique problems. Since we could not differentiate all problems in a consistent manner, we decided to report the comments and findings themselves as approximations to actual problems.

In the first study, clinicians sent their comments by email during a 3-month pilot period in which they used the module for the documentation of actual visits. Another set of comments, in the second study, were entered in an online survey at the end of the pilot. We also extracted direct quotes of clinicians from transcripts of interviews and think-aloud protocols that were completed as parts of usability evaluation in the remaining two studies. The findings, in contrast, were formulated entirely by usability experts as the result of a series of evaluation studies (third and fourth) and published in technical reports.

Each comment and a finding were assigned to a usability heuristic category independently by two researchers. The classification scheme was specific to the healthcare domain and its development is described in detail in a section below. The number of comments and findings in each category was compared to assess the descriptive power of each data collection method for specific usability characteristic. For example, we would contrast the different proportion of comments from each source that contributed to the total number of observations in each category.

The four data collection methods are described in detail below. Think-aloud studies were conducted by a usability expert at our institution and walkthroughs and evaluations by independent professional evaluators on contract basis.

3.1. Email via an embedded link

The Smart Form was integrated within the outpatient clinical records system and used by 18 clinicians for 3-months (March to May, 2006) in the course of their regular clinical work to write visit notes for patients with coronary artery disease and diabetes. They had the option of opening a free-text window on their desktops at any time by clicking on a link embedded in the application and typing in their comments. The messages were collected in a database and logged with a timestamp and the sender’s name.

3.2. Online survey

Fifteen participants received an email with a link to an online survey in May 2006. Questions about satisfaction, frequency of use and problems had multiple-choice responses and were accompanied by two open-ended questions, “What changes could be made to the Smart Form that would make you more likely to use it?” and “What improvements can be made to the Smart Form before you would recommend it to other clinicians?” Completion was voluntary and rewarded with a $20 gift certificate.

3.3. Think-aloud study and observations

We recruited six primary care physicians and specialists (four women) to participate in usability and interaction studies. Evaluations were conducted in the clinicians’ offices at six different clinics and lasted 30–45 min. Subjects were asked to complete a series of interactive tasks described in a previously developed clinical scenario. A researcher played the role of a patient during each session to provide a realistic representation of an office visit. Medical history, current medications and the presence of diabetes and CAD were included in a narrative paragraph that was accompanied by
supporting electronic documentation of prior visits, lab results, vitals and demographic information in a simulated patient record.

Subjects were instructed to verbalize their thoughts (to think-aloud) as they were completing the tasks and interacting with the Smart Form. Video and audio recordings of each session were made with Morae usability evaluation software installed on portable computers. The verbal content was transcribed for analysis to be used together with the resulting screen captures. In a debriefing period after completion, subjects were asked follow-up questions to elaborate or elucidate their actions and reasoning. The results of this study were compiled in a technical report.

3.4. Walkthroughs, expert evaluations and interviews

A team of professional health informatics consultants carried out independently usability assessment and walkthroughs and conducted interviews with six primary care physicians and specialists (two women) whose experience with the application ranged from novice to expert. The results of the evaluation were presented in a technical report.

3.5. The development of heuristic usability assessment scheme

Four sets of usability heuristics with a substantial theoretical overlap have been generally accepted and are widely used in professional evaluations: Nielsen’s 10 usability heuristics [33] (derived from the results of a factor analysis of about 250 problems), Shneiderman’s Eight Golden Rules of Interface Design [34], Tognazzini’s First Principles of Interaction Design [35], and a set of principles based on Edward Tufte’s visual display work [36]. These approaches were recently integrated into a single Multiple Heuristics Evaluation Table by identifying overlaps and combining conceptually related items [37].

These general heuristics sets have been used to evaluate healthcare-related applications [38–41] and consumer-health websites [42]. A set of aggregated Nielsen’s and Shneiderman’s heuristics was proposed by Zhang and colleagues [43] for HIT and applied to the evaluation of an infusion pump [44] and a clinical web application [45]. However, the categories and guidelines do not specifically address biomedical or clinical concepts. Our goal was to formulate additional categories to increase their cumulative explanatory power.

To this end we analyzed all 155 statements about usability problems collected during the study to identify emergent themes following the grounded theory principles [46]. Two researchers then independently assigned the statements into heuristic categories, either general or modified according to newly identified themes. Several iterative coding sessions and discussions ensued, and as a result of extensive comparison and refinement, 12 heuristic categories were formulated (Table 3).

3.6. Participants

All data were collected from 45 clinicians within Partners Healthcare practice network who participated in either part of the study (with a small overlap). Most were primary care physicians (73%), about half were female (53%), and the mean age of the group was 48 years.

4. Results

Analyses were performed separately on comments by clinicians and on findings by usability experts. Results are presented in the following sections and contrasted.

4.1. Comments by clinicians

Results for comments are summarized in Table 1. There were 155 comments from 36 clinicians obtained either in the form of written communication (email and survey) or transcribed from direct verbal quotes (interview and evaluation). We received 85 emails from nine clinicians (reflecting a 50% response rate), and 20 free-text comments were entered in the online survey by 15 clinicians (54% response). Six clinicians who participated in usability evaluations made 26 comments and another six clinicians made 24 distinct comments during interviews.

Over a half of all responses (55%) were emails, and about equal numbers were obtained from the survey, evaluations and interviews (15%, 13% and 17%, respectively). The most common form of a response that constituted about a third of collected data (N = 54) was an email classified as either a Biomedical, Control or Fault category. Comments from the other three sources were most likely to be classified in the following categories: Customization and Control for survey (N = 9, 45%), Transparency and Workflow for evaluations (N = 14, 54%), and Cognition and Workflow for interviews (N = 13, 54%). Overall, the Control, Cognition and Biomedical categories described about a half of all data (52%), and about a third (35%) was classified in the Customization, Workflow and Technical categories. There were no Consistency or Context comments.

Although email was the most prevalent form of communication in the set, its proportion was different within each heuristic category (Fig. 1). For example, it added up to 80% or more in three categories (Terminology, Fault and Biomedical) and to a majority (61%) in the Control category, but only one was classified as related to Workflow. Written response was more likely to be used for the reporting of technical, biomedical and interaction problems (e.g., Fault, Biomedical, Terminology, Control), while verbal comments often related to Workflow or Transparency difficulties. For example, almost 90% of comments made during evaluations were clustered in just four categories and similar distribution was found in data from interviews.

4.2. Findings by usability evaluators

The results are summarized in Table 2. There were 47 findings extracted from expert reports. Over two thirds were classified into just three categories: Cognition, Customization and Workflow. In contrast, none were in the Fault, Speed or Terminology categories and only one was classified as Biomedical. Technical and biomedical concepts were generally not represented in the evaluations.

4.3. Comments and findings comparison

We contrasted all 47 findings with a subset of 105 comments that included only email and survey. Findings were derived from

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Comments by heuristic category and source.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristic category</td>
<td>Email N (%</td>
</tr>
<tr>
<td>Biomedical</td>
<td>21 (81)</td>
</tr>
<tr>
<td>Cognition</td>
<td>12 (46)</td>
</tr>
<tr>
<td>Control</td>
<td>17 (61)</td>
</tr>
<tr>
<td>Customization</td>
<td>7 (29)</td>
</tr>
<tr>
<td>Fault</td>
<td>16 (94)</td>
</tr>
<tr>
<td>Speed</td>
<td>3 (43)</td>
</tr>
<tr>
<td>Terminology</td>
<td>4 (100)</td>
</tr>
<tr>
<td>Transparency</td>
<td>4 (36)</td>
</tr>
<tr>
<td>Workflow</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Totals</td>
<td>85 (55)</td>
</tr>
</tbody>
</table>
Table 2
Findings by Heuristic Category and Source.

<table>
<thead>
<tr>
<th>Heuristic category</th>
<th>Evaluation N</th>
<th>Interview N</th>
<th>Total findings N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biomedical</td>
<td>1</td>
<td>0</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Cognition</td>
<td>10</td>
<td>6</td>
<td>16 (34)</td>
</tr>
<tr>
<td>Control</td>
<td>2</td>
<td>4</td>
<td>6 (13)</td>
</tr>
<tr>
<td>Customization</td>
<td>2</td>
<td>7</td>
<td>9 (19)</td>
</tr>
<tr>
<td>Consistency</td>
<td>0</td>
<td>1</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Context</td>
<td>1</td>
<td>1</td>
<td>2 (4)</td>
</tr>
<tr>
<td>Transparency</td>
<td>5</td>
<td>0</td>
<td>5 (11)</td>
</tr>
<tr>
<td>Workflow</td>
<td>7</td>
<td>0</td>
<td>7 (15)</td>
</tr>
<tr>
<td>Totals</td>
<td>28</td>
<td>19</td>
<td>47 (100)</td>
</tr>
</tbody>
</table>

Table 3
Description of Heuristic Evaluation Categories.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistency</td>
<td>Hierarchy, grouping, dependencies and levels of significance are visually conveyed by systematically applied appearance characteristics, perceptual cues, spatial layout, text formatting and pre-defined color sets. Behavior of controls is predictable. Language in commands, labels and warnings is standardized.</td>
</tr>
<tr>
<td>Transparency</td>
<td>The current state is apparent and possible future states are predictable. Action effects, their closure and failure are indicated.</td>
</tr>
<tr>
<td>Control</td>
<td>The interruption, resumption and non-linear or parallel task completion is possible. Direct access to data across levels of hierarchy, backtracking, recovery from unwanted states and reversal of actions are possible.</td>
</tr>
<tr>
<td>Cognition</td>
<td>Content avoids extraneous information and excessive density. Representational formats allow perceptual judgment and unambiguous interpretation. Cognitive effort is reduced by minimalistic design, formatting and use of color, allowing fast visual searches. Recognition is preferred over recall. Conceptual model corresponds to work context and environment.</td>
</tr>
<tr>
<td>Context</td>
<td>Terms, labels, symbols and icons are meaningful and unambiguous in different system states. Alerts and reminders perceptually distinguish between general (disease, procedure, guidelines) and patient-specific content.</td>
</tr>
<tr>
<td>Terminology</td>
<td>Medical language is meaningful to users in all contexts of work, compatible with local variations and established terms.</td>
</tr>
<tr>
<td>Biomedical</td>
<td>Biomedical knowledge used in rules and decision support is current and accurate, reflecting guidelines and standards. It is evident how suggestions are derived from data and what decision logic is followed.</td>
</tr>
<tr>
<td>Safety</td>
<td>Complex combinations of medication doses, frequencies, units and time durations are disambiguated by appropriate representational formats and language, entries are audited for allowed value limits. Omissions are mitigated by goal and task completion summary views. Errors are prevented from accumulating and propagating through the system.</td>
</tr>
<tr>
<td>Customization</td>
<td>Preferred data views, organization, sorting, filtering, defaults, basic screen layout and behavior are persistent over use sessions and can be defined individually or according to role.</td>
</tr>
<tr>
<td>Fault</td>
<td>Software failures and functional errors are minimal, do not compromize safety and prevent the loss of data.</td>
</tr>
<tr>
<td>Speed</td>
<td>Minimal latency of screen loads and high perceived speed of task completion.</td>
</tr>
<tr>
<td>Workflow</td>
<td>Navigation, data entry and retrieval does not impede clinical task completion and the flow of events in the environment.</td>
</tr>
</tbody>
</table>

reports of evaluation and interviews that already contained reinterpreted verbal comments of the subjects. We therefore excluded comments made during evaluations from the comparison. Comments and findings showed divergent trends in characterizing usability aspects of the Smart Form (Fig. 3). Comments were more likely to describe discrete, clearly manifested and highly specific problems and events, such as software failures or concerns about medical logic or language (e.g., Control, Biomedical, Fault, Terminology). Findings derived from usability evaluation, on the other hand, tended to explain conceptual problems related to overall design and the suitability of the electronic tool to clinical work (e.g., Consistency, Context, Workflow). Both methods contributed about equally to the description of problems with human interaction (e.g., Cognition, Customization).

4.4. Implementation of design changes to a revised prototype

Individual comments and findings most often referred to single, discrete problems. Some problems were reported by several clinicians or were identified by multiple methods. The 155 analyzed comments and findings reported 120 unique problems (77% ratio), and 12 problems were simultaneously described by more than one method (10% ratio). We have iteratively implemented design changes into the prototype on the basis of 56 reported problems (47%). Most of the problems that led to subsequent changes (34) were reported by email.

5. Discussion

Our data analysis has identified the relative strengths and weaknesses of the four evaluation approaches, their distinct utility and appropriateness for characterizing different usability concepts, and their cumulative explanatory power as a set of complementary methods used at specific points of the development lifecycle. The large number of comments that clinicians provided were a rich source of reports on software failures, slow performance and potential conflicts and inconsistencies in biomedical content, while usability experts generally gave comprehensive assessments of problems related to human interaction and workflow, including characterizations of problems with interface design and layout that negatively affect cognitive and perceptual saliency of displayed information. The core principles, attributes and expected results for each method are summarized in Table 4 and discussed in depth in the following sections.

5.1. Email

An email link embedded in the application is available to everyone and at all times, allowing almost instantaneous reporting of problems as they occur. Informaticians and computer technology specialists can learn from these comments how the software performs in authentic work conditions and how well it supports clinicians in complex scenarios that commonly arise from the combination of personal workflows and preferences, unexpected events, and unusual, idiosyncratic, unplanned or non-standard interaction patterns. The wide range of conditions that affect performance and contribute to errors and failures would not be possible to anticipate and simulate in the laboratory. Performance measures in actual settings also give evidence of the technical and conceptual strengths of the design. Insights from these reports give designers a unique opportunity to make the application more robust and tolerant of atypical interaction, more effective in managing and preventing errors, and more appropriate for the clinical task it supports.

The large number and variety of email reports and their often fragmentary content make them often hard to interpret. For example, it is difficult for clinicians to recall accurately the relevant and descriptive details of errors that were made or problems that were encountered during complex interactions with multi-step or interleaving tasks, and to convey a meaningful description of the event. However, informaticians may need details about the system state, work context or preceding actions that are often lacking in spontaneous and short messages to evaluate how a problem originated.
and its potential consequences. The usually large volume of emails accumulated over time also contains repetitive, idiosyncratic and inaccurate reports that may be of little value and need to be excluded. A self-selection bias among respondents (e.g., novice users may be underrepresented) may accentuate marginal problems or conceal more serious ones. Difficulties of more conceptual character may be only rarely reported through comment messages, as was evident from the analysis of our data (e.g., the distribution of comments in heuristic categories, Fig. 2).

Among the most significant advantages of embedded email response links are their inexpensive implementation, network-wide availability, real-time response and continuous, active data collection. These characteristics make email an excellent data collection method during pilot testing of release candidate versions and after the release of full versions. There is a high probability of quickly discovering technical problems, an opportunity to review medical logic for decision support tools that may not have been tested in complex scenarios (e.g., a patient with multiple comorbidities and drug prescriptions), and a likelihood of finding inconsistencies in terminology or ambiguities in language and expressions. For example, of the 56 changes and corrections we implemented in the prototype, 36 (64%) such problems were reported and identified in emails.

This method requires the software to be in the stage of a fully functional prototype or in its final release form. It may therefore be too laborious or expensive to make significant conceptual changes in design at that point. However, our data suggest (e.g., the proportions of comments to specific concepts in Fig. 2) that most of email-reported problems concern specific biomedical content, terminology and technical glitches that may be relatively easy.
to correct without large-scale changes in the code and screen layout.

5.2. Survey

Survey is another form of direct clinician response that we used in this study and it shares several characteristics with email communication, such as a potentially wide reach, economy of administration, a tendency for self-selection bias, relatively low response rate and the brevity of its form. Unlike email links, surveys are structured and contain a pre-determined set of questions to elicit responses and opinions on narrow topics of interest. They do not allow reporting problems in real time, however, and require respondents to recall and interpret past events at the time the survey is completed. This may be difficult, as our data suggest that free-text answers to open-ended questions did not contain references to specific and detailed biomedical and technical problems, the most frequent categories represented in emails (see Table 1).

Rather, clinicians tended to describe more broadly defined difficulties with screen control, navigation and customization. The content in surveys, as in other direct forms of communication, is often subjective, reflecting personal opinion, and therefore, of lower descriptive value and accuracy than data gathered in professional evaluations [22]. A substantial period of time needs to be allowed for potential survey respondents to work with a fully working prototype or the completed application before they can form meaningful opinions and gain a measure of proficiency. Surveys can be administered periodically for comparisons over time and can be timed to coincide with important events such as pilot release, shortly after full release.

5.3. Usability evaluations and interviews

The most telling indicators of conceptual flaws in the design come from the observation of human interaction errors [47]. They can provide insights into discrepancies between expected and actual behavior and identify inappropriate and ambiguous representational formats of information on the screen that impairs its accurate interpretation [48]. Errors are rarely reported directly in emails or in surveys, as the responders are not often aware of their own mistakes. For example, observation experts in our study reported that a clinician during a simulated task “could not tell whether the patient was taking Aspirin, assumed that urinalysis could only be ordered on paper and did not notice the save button,” an insight that would not be gained by introspection and recall.

Usability inspection methods in which experts alone evaluate the interface, such as the cognitive walkthrough and heuristic evaluation, provide predominantly normative assessments. In other words, they report how well the interface supports the completion of a standardized task that can be reasonably expected to be performed routinely, and measure the extent to which the design adheres to general usability standards. These methods produce reference models of interaction that can be compared to evidence from field observations.

Ethnographic and observational methods such as think-aloud studies, on the other hand, derive data from analyzing unscripted and natural interactions with the software by non-experts with various levels of computer and task-domain skills. They are therefore inherently descriptive and analytic and allow researchers to make inferences about the clarity and suitability of the design to the task from observed competencies and errors. Usability experts can integrate findings about interaction errors with interface evaluations, cognitive walkthroughs and heuristic evaluations into a comprehensive analysis and formulate optimal strategies for making modifications to the interface. Normative and descriptive methods together constitute a comprehensive evaluation of design in progress that can be repeated iteratively early in the process to refine data representation and interaction concepts in each successive version.

Findings from experts in this study have been clearly focused on conceptual and interaction-related aspects of the Smart Form (Table 2). The structured format of think-aloud studies follows pre-defined clinical scenarios that generally contain validated biomedical data and unambiguous terminology that do not represent potential problems to be reported in evaluations. Comments from clinicians working with the software in real settings, however, are more descriptive of specific factual, technical and biomedical errors that observational studies frequently do not capture. The
relative proportions of expert findings and clinicians' comments in each heuristic category and their respective tendency to describe different aspects of the software are clearly evident in Fig. 3.

Experts can also capture more easily positive aspects of the design and confirm successful trends. For example, an evaluator reported that “the subject seemed comfortable navigating around and understands how to update medications in the system.” Email responses are often initiated at the time of a failure or when an error is encountered, but rarely when the system is working well. In effect, successful performance is characterized by uneventful and well-progressing work which is apparent to observers but not often reported back to designers by clinicians.

Interviews with clinicians are usually done in conjunction with observations to elucidate aspects of collected data that require proper context for interpretation, and also as “debriefings” at the end of after think-aloud studies. The results of expert evaluations commonly incorporate insights and findings from interviews into comprehensive reports.

Expert evaluations are indispensable during the initial design stages when even significant corrections and reconceptualizations are still possible without incurring steep penalties in time and development effort.

6. Conclusion

This study has been conducted to characterize and compare four usability evaluation methods that were employed by the research team during the design and pilot testing of new clinical documentation software. We have also formulated a classification scheme of heuristic usability concepts that incorporates established principles and extends them for evaluations specific to the clinical software domain.

Our results suggest that no single method describes better than others all or most usability problems, but rather that each is optimally suited for evaluations at different points of the design and deployment process, and that they all characterize different aspects of the interface and human interaction. The studies and assessments we have performed were embedded in the design process and spanned the entire development cycle.

Heuristic evaluations and ethnographic observations of actual clinical work by usability experts inform and guide conceptual and workflow-related changes and need to be performed iteratively early in the design cycle so that they can be incorporated without excessive effort and time. Responses elicited directly from clinicians and other users through email links and surveys report mostly technical, biomedical, terminology and control problems that may occur in a wide variety of workflows and idiosyncratic use patterns.

The evaluations were conducted on the relatively small scale of a pilot study. However, the smaller size may be typical of many software development efforts at large academic and healthcare centers. The findings and lessons learned in this study may, therefore, be of interest to information system designers, developers and researchers and development centers affiliated with hospitals and directly related to their experiences with the design and improvement of clinical information systems. We have outlined a methodological approach that is applicable to most development processes of software intended for healthcare information systems.

We plan to formally validate and possibly revise the set of heuristics we formulated and apply it to the evaluation of an information system in its entirety that will also include judgments about safety that were not performed in this pilot study.

Health information technology is still in its nascent state today. Order entry systems, for example, still represented only a second generation technology in 2006 and had many limitations that precluded their meaningful integration into the process of care [49]. Applications not appropriately matched to clinical tasks tend to be chronically underused and may be eventually abandoned [21].

Acknowledgments

The Smart Form research was supported by Grant 5R01HS015169-03 from the Agency For Healthcare Research And Quality. We wish to thank Alan Rose, Ruslana Tsurikova, Lynn Volk and Svetlana Turovsky for their contribution and expertise in data collection and initial interpretation, and to all clinicians who participated in the four studies as subjects.

References