INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION, *21*(1), 73–89

Effective Display of Medical Laboratory Report Results on Small Screens: Evaluation of Linear and Hierarchical Displays

Heidi Lam
School of Engineering Science, Simon Fraser University, Canada

Arthur E. Kirkpatrick
School of Computing Science, Simon Fraser University, Canada

John Dill
School of Engineering Science, Simon Fraser University, Canada

M. Stella Atkins
School of Computing Science, Simon Fraser University, Canada

Two studies evaluated linear and hierarchy+elision small-screen display formats for clinical reasoning tasks. A controlled, quantitative study with 28 medically naive participants using a task abstracted from clinical use of laboratory results found that both display formats supported rapid and accurate decision making. Distribution of the search targets significantly affected speed, with decisions in linear format made 13% faster (4.7 sec) when all targets could be viewed on a single screen than when targets required scrolling between several screens and in hierarchical format 15% faster (5.1 sec) when all the targets were confined within one category. Performance was equivalent regardless of the relative order of the target results and data in the laboratory report. In a qualitative study, 7 physicians used the displays to perform a realistic diagnosis. Physicians were comfortable with both display formats, but preference varied with clinical experience. The 5 less experienced clinicians favored hierarchy+elision, whereas the 2 highly experienced clinicians tended to prefer the linear display.

Funding for this work was partially provided by the National Sciences and Engineering Research Council of Canada and Neoteric Technology Ltd. under an Industrial Postgraduate Scholarship. The work described in this article was performed as part of Heidi Lam's MASc thesis.

Heidi Lam is now at the Department of Computer Science, University of British Columbia, Canada.

Correspondence should be addressed to Heidi Lam, Department of Computer Science, University of British Columbia, 201–2366 Main Mall, Vancouver, BC V6T 1Z4 Canada. E-mail: hllam@cs.ubc.ca

1. DISPLAYING LABORATORY REPORT RESULTS ON SMALL SCREENS

Medical laboratory reports are essential tools for health care professionals in patient assessment, diagnosis, and long-term monitoring. For effective medical care, these reports must be up to date, complete, and accurate. Electronic displays on networked, handheld devices can provide physicians with more immediate, accurate laboratory reports than traditional paper files. Systems providing clinical information on mobile devices are already being fielded (e.g., PalmCIS; Chen et al., 2004). There are little to no empirical data on the effectiveness of displaying test results on small screens. In this article, we present empirical results comparing the display of laboratory reports on small screens using two popular display formats for lists: a simple linear list and a hierarchically structured list with selective display of subgroupings.

Displaying medical test results effectively is difficult on screens of any size. A single result is not enough to support decision making. It must be displayed in context with other results, supplemented by test-specific and patient-specific information. Test-specific information includes reference range, sampling date, and clinical information provided by the primary physician or laboratory, whereas patient-specific information includes age, gender, height, and weight. In short, a typical laboratory report contains a large amount of data, and if these are presented ineffectively, there is a danger that important information will be missed (Mayer, Wilkinson, Heikkinen, Ørntoft, & Magid, 1998). Even in prestigious medical institutions, available laboratory data can fail to elicit an appropriate clinical response in 30% of cases (Altshuler, 1994), and 9% to 31% of laboratory errors relate to inappropriate interpretation and utilization of laboratory results (Bonini, Plebani, Ceriotti, & Rubboli, 2002). Displays designed without accounting for the nature of laboratory data and the limitations of human perception may contribute to this problem (Mayer et al., 1998; Wright, Jansen, & Wyatt 1998).

This task is even more difficult when the display size is severely limited, as in the case of most handheld devices. In fact, various studies have shown that performance on small screens is less effective than on their desktop counterparts (Jones, Marsden, Mohd-Nasir, Boone, & Buchanan, 1999; Kamba, Elson, Harpold, Stamper, & Piyawadee, 1996), especially when the task is complex (Watters, Duffy, & Duffy, 2003).

Effective information display on any screen depends on many factors, including visual properties (e.g., font, formatting, and color), interaction technique, navigation and overview tools (e.g., bookmarks, search engines, and summarization), and device characteristics (e.g., brightness and contrast). Although all these factors can contribute to the eventual success or failure of the display, we focus in this work on layout of lists of laboratory results. We believe that the layout is an important consideration in displaying laboratory reports, with other factors and associated tools providing enhancements.

Although lists are common data structures, it is still currently unclear how to best display them on small screens. Traditionally, lists are displayed linearly or, if there is an overarching hierarchical organization, hierarchically. We present empirical evaluations of the display of laboratory reports in linear and hierarchical for-

mats. To compensate for the display size limitation, we explore visualization techniques that harness human perceptual capabilities. We take an integrated approach, considering both the diagnostic process that physicians apply to these reports and such social factors as familiarity of format and the medical profession's deliberately conservative approach to innovation.

2. DISPLAYING LABORATORY REPORT RESULTS ON SMALL SCREENS

The obvious choices for displaying lists of laboratory results would be displaying the data either as a linear scrolling list or with hierarchy+elision, grouping the items as a tree with nodes that can be expanded or collapsed. These two options are shown in Figure 1.

Theoretical analyses of effectiveness of list displays have argued for the use of a structured, hierarchical format (Furnas, 1997; Tullis, 1997). Grouping list items improves navigability of the original list by providing shortcut paths and introducing new viewing links. As a result, grouping can reduce the diameter of the navigation space and improve traversability (Furnas, 1997). For small screens, this grouping can result in substantially reduced scrolling to reach the target.

Whereas hierarchical grouping improves navigability, elision adds further benefits. When all the nodes are collapsed, category headings provide an overview of the data structure, and elision allows selective viewing of the groups. On small screens, users can view relevant information while maintaining the context on a single screen. These benefits have costs. Elision hides information: Users are now required to be familiar with the grouping scheme to predict the correct group for their target information. Furthermore, action is required before the information can be read.

Hierarchy+elision also requires more screen space. An extra row is required for category headings, reducing the number of items displayable without vertical scrolling. Hierarchical indentation of items reduces the total number of columns

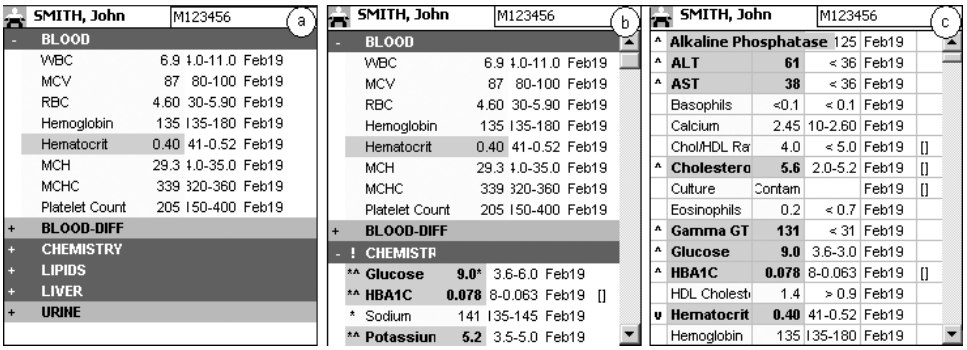

FIGURE 1 Formats for list display: (a) Hierarchy+elision with the single group, Blood, expanded; (b) two groups, Blood and Chemistry, expanded; and (c) simple linear list display.

displayable without horizontal scrolling. Increasing the hierarchy depth consumes further space.

Researchers have developed a number of systems based on the hierarchy+elision technique to display lists on small screens. The Power Browser (Buyukkokten, Molina, Paepcke, & Winograd, 2000) and the WebTwig browser (Jones, Buchanan, & Mohd-Nasir, 1999) provide hierarchy+elision outlines of Web sites, where users can view more or fewer levels of the tree structure by expanding or collapsing the tree display. LibTwig (Jones, Jones, & Deo, 2004) provides search engine results in a hierarchy+elision list grouped by key phrases extracted from the documents. For text, Buyukkokten, Molina, and Paepcke (2001) proposed the Accordion Summarization technique, which displays text paragraphs on a Web page by stages: only the first line, the first three lines, then the whole paragraph. These studies supported simple information search (e.g., searching for a phone number on a personal Web page) by a rather generic user group.

Despite the longevity of the discussion over display formats and the number of published theoretical analyses and demonstration systems just noted, we have not found any empirical evaluations of these display formats. Indeed, the seemingly straightforward question, "Which format is better?" is ill formed in the task where the user is searching for multiple values. For this task, performance of each format is sensitive to the distribution of the values. If the values are distributed so that they appear close together—say, on one screen—the task will be completed more quickly than if their distribution requires scrolling the display. The optimal distribution will be different for each display format.

2.1. Specific Issues in Supporting Clinical Reasoning

The context of clinical reasoning is substantially different from the contexts for previous studies. The task is more complicated, requiring synthesis of multiple results instead of a simple visual search. Furthermore, members of the medical community have strong expectations about the proper format for display and use of data, expectations that may not be met by techniques designed for commercial Web sites or personal calendar applications.

As a display for medical laboratory data, the hierarchical format has several advantages. Physicians are familiar with existing paper systems, which present results hierarchically, and laboratory data seem well suited to a hierarchical display. Given that physicians use laboratory tests to investigate a specific clinical problem at hand, a limited number (30–50) of tests are usually requested, falling within one to five categories. These characteristics make the data well suited for hierarchical display. In addition, because most disease pathology tends to follow specific organ or physiological systems, grouping results together by systems tends to cluster relevant results further.

The elision technique extends these benefits. First, a completely collapsed display provides an overview for overall patient status assessment and a guide for navigation. The collapsed categories thus provide context to the physicians while viewing specific test results in opened categories. Second, experiments have shown that medical decision making based on laboratory reports is faster and less error

prone if all required data can be viewed on one page (Nygren, Wyatt, & Wright, 1988). Grouping of tests based on organ or physiological systems makes it more likely that relevant results will fit on a single screen. Third, elision allows physicians to selectively view test groups, which can further concentrate relevant result. For example in Figure 1a, the physician has selectively viewed the Blood group, and in Figure 1b, the physician has selectively viewed the Blood and the Chemistry groups, while collapsing the Blood-Diff group between them.

The apparent suitability of hierarchy+elision for clinical results must be qualified by the amount of the physician's clinical experience. Studies in medical cognition (Patel, Arocha, & Kaufman, 1994; Patel, Groen, & Patel, 1997) have found that physicians' clinical reasoning depends on the level of expertise and domain knowledge. Domain experts use laboratory tests to refine and evaluate their hypotheses and to plan therapy. Because each test is ordered for a reason, these experts need only see the test result, which they interpret within the framework of their prior hypotheses. In this case, results should be displayed in a manner that maximizes speed of access, such as an alphabetic list. This process avoids the layer of complexity imposed by test grouping, as some tests can be potentially grouped under more than one organ system and some systemic tests may be difficult to categorize under a single system. However, the reasoning of clinicians inexperienced in the domain of the clinical case may not be as structured and coherent, and they often use laboratory tests to generate more hypotheses rather than to evaluate prior hypotheses. For these physicians, more structured test reports might facilitate more structured clinical reasoning. Such differences in knowledge management are not specific to the medical domain. In the context of recalling C programming knowledge, Ye and Salvendy (1994) found that experts apparently organized their knowledge as larger but fewer chunks than those of novices, enabling the experts to hold more information in their working memory.

2.2. Existing PDA Applications for Clinical Use

Both the linear and the hierarchy+elision techniques have been used to display medical results on small screens. In some applications, numerical results of standard hematology tests are displayed as linear lists of six or seven items (e.g., Mobile MedData Charts, Medical Communication Systems, Inc., Woburn, MA; Patient Tracker, Shatalmic, Layton, UT). PatientKeeper (PatientKeeper, Inc., Newton, MA) displays standard laboratory test panels (collections of related tests) in a modified tabular format, with each panel occupying one row of the display. Because only numerical results are displayed and the layout of test panels varies with geographic region, the correspondence between test name and numerical result is not obvious. These displays also assume the physicians would be familiar with the units and reference ranges for the test, but those conventions also differ among laboratories and geographic regions. In contrast to using linear lists, PalmCIS displays test panels using a form of hierarchy+elision (Chen et al., 2004). In the initial screen, only the panel names and their corresponding numerical results are visible, whereas all other information (e.g., test name, date, clinical information) is available on pull-down menus. Details of individual panels can be accessed as separate linear lists, with each panel on a separate page.

The work to date on the clinical use of PDAs has considered broader issues of providing medical reference materials and electronic patient records. The few systems that have presented laboratory results have not specifically focused on display techniques for those results. We are not aware of published usability evaluations of display techniques or their degree of support for medical reasoning. Due to the various trade-offs in the hierarchy+elision and linear formats, we performed two studies to guide design choices. In our quantitative study, we focused on the effects of search target distribution on task performance, whereas in the qualitative study we looked at the effects of domain expertise on the choice of display formats.

3. EVALUATION OF THE HIERARCHY+ELISION AND THE LINEAR FORMATS

We performed a quantitative and a qualitative evaluation of these display formats for small-screen display of medical results.

3.1. The Quantitative Study

This study measured the time and accuracy of the two display formats for a task similar to clinical decision making. Given the difficulty of obtaining physicians as participants for such a study, we instead had participants with no medical knowledge perform a task abstracted from clinical reasoning. As previously discussed, physicians generally develop a schema while interpreting a report for clinical decision making. The study task should therefore contain several elements of the clinical task: (a) Provide a list of search targets to the participants corresponding to the laboratory results the physicians look up; (b) require evaluations using rule-based logic statements to mimic the physician's task of determining the status of a particular test result as elevated, depressed, or normal; and (c) require drawing a conclusion from all the evaluations.

For both displays, we hypothesized that the task time and error rate would be lower when the required data were contained in a single screen or a single category rather than when the data were scattered across two or more screens or categories. We hypothesized that this effect would be more pronounced for the hierarchy+elision display than for the linear display, because a mismatch would require closing and opening nodes, an act that may be more disruptive to short-term memory than scrolling. Furthermore, we hypothesized that if the presentation order of the task targets matched the ordering of the data or data category in the displays, the task time and error rate would be lower.

3.2. The Qualitative Study

The controlled quantitative study was designed to give reliable estimates of response time and accuracy, but this structure prevented it from accounting for two important issues. First, because our participants were naive to both the abstract experimental task and the clinical reasoning task of actual interest, their results were

likely to predict only the preferences of physicians that are not experienced in the clinical case domain and would not apply to domain experts. Second, because our participants were not members of the medical community, the controlled study could not account for the social factors that influence acceptance of the two display formats, or even the acceptance of handheld displays at all.

The qualitative study was designed to determine the impacts of these two factors on the use of the two display formats. We interviewed physicians and observed them reasoning about a case study, using laboratory results displayed on a handheld device.

4. STUDY 1—QUANTITATIVE STUDY

4.1. Design

Participants. Mass e-mail messages were used to recruit 32 participants. Of these, 4 were excluded from further analysis during data screening (see Results section). The remaining 28 were predominantly men (20 men, 8 women) and from three university departments (22 from computing science, 4 from interactive arts, and 2 from engineering science). Their mean age was 27 years (SD = 6; min = 22, max = 48). All were right-handed. Participants were paid $10 CDN for the session.

Displays. The testing software was run on a laptop, for better flow control and data recording than would have been possible on a handheld device. The laptop had a resolution of 1280 × 768, in which the data display area was 220 × 330 pixels (4 × 6 cm), approximately the display screen size of a common "Pocket PC," with a maximum of 15 rows of readable text information. All input was performed with an optical mouse. Examples of the two presentation techniques are shown in Figure 2.

There were six categories (Table 1), each with 10 items. From this pool of 60 items, 30 were available on a given trial for display using either the hierarchy+eli-

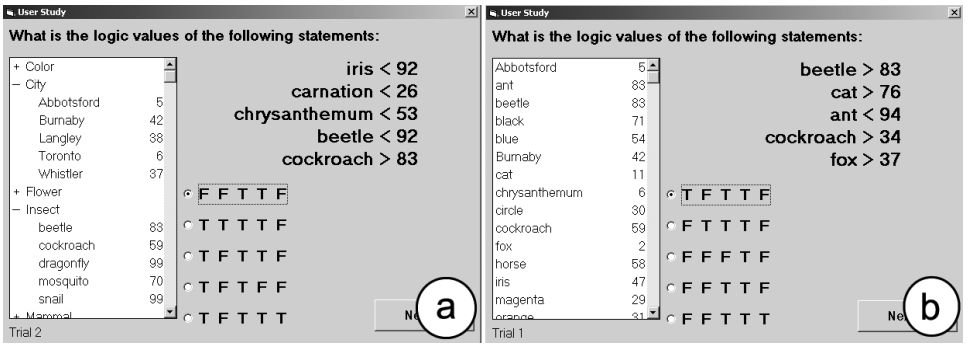

FIGURE 2 Interfaces for the Study 1 task: (a) hierarchy+elision display, and (b) linear display.

Table 1: Categories and Members for Study 1 Task

Category	*Member Items*
Color	black, blue, green, magenta, orange, pink, purple, red, white, yellow
City	Abbotsford, Burnaby, Halifax, Kelowna, Langley, Ottawa, Surrey, Toronto, Vancouver, Whistler
Flower	buttercup, carnation, chrysanthemum, daffodil, iris, orchid, primrose, rose, sunflower, tulip
Insect	ant, beetle, cockroach, dragonfly, grasshopper, fly, mosquito, snail, spider, whitefly
Mammal	cat, coyote, dog, fox, horse, ox, pig, sheep, tiger, zebra
Shape	circle, diamond, hexagon, pentagon, rectangle, rhombus, octagon, parallelogram, square, triangle

sion or the linear display. In the hierarchy+elision display, items were grouped by category, with display of 5 items per category, selected from the 10 candidates. Maintaining 5 items per category kept the branching factor within the recommended range for menus (Kiger, 1984; Miller, 1981). A survey of test panels provided by most laboratories revealed that most test panels consist of 1 to 10 items. Data points were arranged alphabetically within each category. Categories could be expanded or collapsed by clicking the + or – sign on the group title. For the linear display, data names were sorted alphabetically. Both displays used vertical scroll bars.

Task. Our abstract clinical reasoning task required determining the truth values of five target statements (upper-right corner of the screen, Figure 2) based on the data presented (Figure 2 left, using one of the two display formats) and matching these values to one of five options (Figure 2, lower right). We believe this task to be a valid abstraction of typical medical reasoning, a belief that was confirmed by interviewing an emergency physician. A pilot study (Lam, 2004) demonstrated that the strategy of evaluating the target statements sequentially from top to bottom produced the best times and accuracies, so participants were instructed to use this strategy.

Experimental design. The display formats were evaluated using a within-subject, randomized complete block design. Two factors were varied: target distribution and target order. For linear displays, target distribution was defined as either small (all targets found within a single page) or large (targets were scattered among two or more pages). For hierarchy+elision displays, target distribution varied from one to five of the listed categories instead of using the *small* and *large* criterion, as it would be difficult to predict if the participant fit all the targets in a single screen due to the optional closing of categories.

Target order had two levels: matched and unmatched. The definition of *matched* depended on the display format. For the linear display, the target order was considered matched if the targets were listed in alphabetic order, as the linear display was

also alphabetically sorted. For the hierarchy+elision display, the target order was considered matched if the target category was ordered, although the targets themselves could be out of alphabetic order.

Trials were blocked by display type. For the linear display, both factors were fully crossed in a single 2 × 2 design. For the hierarchy+elision interface, the two factors were only partially crossed to reduce the total number of trials required. Each experimental session consisted of 32 trials, 16 linear and 16 hierarchy+elision. For the 16 linear trials, there were four conditions, each with four trials:

1. Small distribution with ordered target
2. Small distribution with unordered target
3. Large distribution with ordered target
4. Large distribution with unordered target

For the 16 hierarchy+elision trials, there were five conditions:

1. Data spanning a single category (Category 1, by definition ordered targets)
2. Data spanning two categories (Category 2, ordered and unordered targets)

3.–5. Data spanning three/four/five categories (Category 3/4/5, unordered targets)

To limit the total number of conditions in our experiment, we tested only target ordering with targets distributed among two adjacent categories, as we suspected effects of target ordering would likely be masked by target distribution when data span more than two categories.

Study protocol. Participants were apprised of the minimal risks of participation and told they could stop the experiment at any time. They were next instructed in the task for the two display techniques. Participants were advised to take breaks between trials if they wished. Participants did 2 practice trials with a linear display and, due to its higher complexity, 3 practice trials with the hierarchy+elision display. They then performed 16 trials with each display format, with the order of presentation counterbalanced across participants. Finally, participants answered nine questions concerning subjective impressions of the displays. Session times ranged from 30 to 45 min.

4.2. Results

Some participants had great difficulty with the task, performing much less accurately than the rest. They appeared to be unable to perform this kind of diagnostic task. Because we wished to use the results of the quantitative study to estimate performance by medical professionals with demonstrated skill at diagnosis, we set a threshold of a minimum accuracy of 75%. Furthermore, a design error in the interface software would cause users to occasionally terminate a trial prematurely. To

guard against this case, we eliminated any participant who had at least one trial that was performed in less than 10 sec. Of the original 32 participants, these criteria eliminated 2 for accuracy and 2 for time. All results are reported for the remaining 28 participants.

Response time data were analyzed using analysis of variance with participants as a blocked factor. When indicated, post hoc analysis was done using Tukey's honestly significant difference. Because accuracy data were binary, they were analyzed using the nonparametric Kruskal–Wallis test.

Mean task completion time across all experiments was 34.6 sec (SD = 14.2). Mean accuracy across all individual trials and participants was 90% (SD = 30%).

Target distribution and target order in linear format. Task times for the two target distributions are shown in Figure 3. Target distribution was significant, $F(1, 417) = 21.0$, $p < .0001$, with target data displayed on one screen 16% faster than target data on more than one screen, confirming our hypothesis for time. There was no significant effect of target order on response time, $F(1, 417) = 0.26$, $p = .61$, but a significant Distribution × Order interaction effect, $F(1, 417) = 7.13$, $p = .008$. Our hypothesis for accuracy was not confirmed, however, as there was no significant difference in accuracy among the four conditions, $\chi^2(3, N = 444) = 1.72$, $p = .63$.

In the open-ended qualitative questions (Table 2), the most positive aspect of the linear display was that there was no need to understand the meaning of the data name, or even read the whole name, to locate the data point (16 reports). The most negative aspects were (a) excessive scrolling (23 reports), (b) visual clutter on the small screen (4 reports), and (c) confusion of adjacent data names when they were similar (4 reports).

Target distribution and target order in hierarchical format. Figure 4 shows the task times for target distributions for the hierarchical format. As with the linear for-

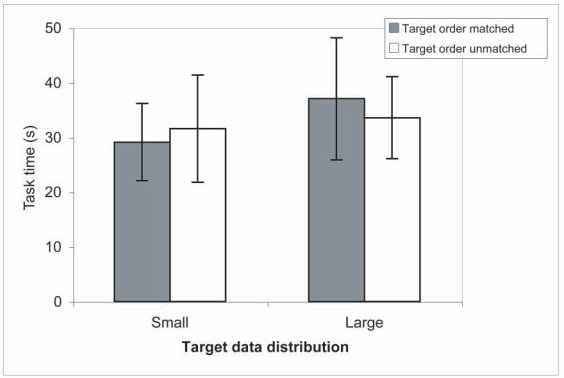

FIGURE 3 Mean task times (N = 28) for the linear display. Error bars show standard deviation.

Table 2: Post-Task Questions and Scores, Study 1

	Percentiles		
Questions	*25th*	*Mdn*	*75th*
Rating questions			
It is easy to find the necessary data using the linear display.	4	6	6
It is easy to find the necessary data using the hierarchy+elision display.	4	5	6
It is easy to match the statements using the linear display.	4	5	6
It is easy to match the statements using the hierarchy+elision display.	4	5	6
I prefer using the hierarchical+elision than the linear display in doing the study task.	3	4	6
Open-ended questions			
List the three most negative aspect(s) of the linear display when performing the study task.			
List the three most positive aspect(s) of the linear display when performing the study task.			
List the three most negative aspect(s) of the hierarchical display when performing the study task.			
List the three most positive aspect(s) of the hierarchical display when performing the study task.			

Note. Ratings on a 7-point scale, from 1 (*strongly disagree*) to 7 (*strongly agree*).

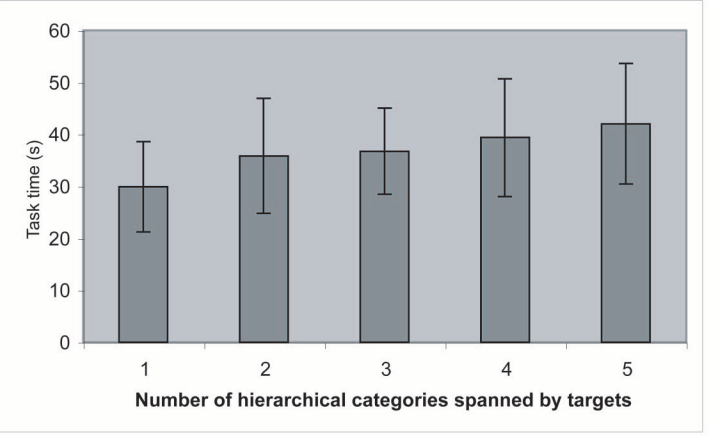

FIGURE 4 Mean task times ($N = 28$) for the hierarchical+elision display with target data found among one to five categories. Error bars show standard deviation.

mat, target distribution was significant, $F(4, 422) = 15.0$, $p < .0001$. Post hoc tests indicated that trials with target data in a single group were significantly faster than the rest, and those with data distributed among two adjacent categories were significantly faster than those with data distributed among five categories. Participants analyzed target distributions contained in a single group 28% faster than those contained in two or more groups. Distribution had no effect on accuracy, $\chi^2(4, N = 446) = 2.54$, $p = .64$. Unlike target distribution, target order did not significantly affect task time, $F(1, 83) = 0.34$, $p = .56$, or accuracy, $\chi^2(1, N = 54) = 0.72$, $p = .39$.

In the open-ended qualitative questions (Table 2), the most positive aspects of the hierarchical display were that the display (a) was visually clear and compact with fewer distracting items (13 reports), (b) presented a shorter list once the cate-

gory was determined (8 reports), and (c) was easier to remember when data fit into the same category (7 reports). The most negative aspects were (a) the need to understand the data name to classify into a category (20 reports), (b) the extra clicks required in opening and closing categories (15 reports), and (c) the initial hiding of items (5 reports).

Comparison of hierarchy+elision and linear displays. Comparisons of participant performance using the display formats can usefully be made only with respect to data distribution. The formats supported equivalent performance when the required data were contained in a single category for hierarchical displays (15% faster than average, 5.1 sec) and when they fit on a single screen for linear displays (13% faster than average, 4.7 sec). The difference between the two was not significant, $t(339) = 0.35$, $p = .73$. Combining the remaining distributions, performance with the linear display was 1% slower than average (0.3 sec), and performance with the hierarchical display was 9% slower than average (3.1 sec). The difference was significant, $t(569) = -2.89$, $p = .004$.

The questionnaire had five rating questions (Table 2). Overall, participants did not prefer either display (Question 5). There were also no significant differences in the answers to Questions 1 to 4 (Q1 vs. Q2 on the effectiveness of the displays on target location; Wilcoxon signed rank test, $Z = 148.5$, $N = 28$, $p = .97$; and Q3 vs. Q4 on the effectiveness of the displays on value matching, $Z = 75$, $N = 28$, $p = .94$).

4.3. Discussion

The major factor determining performance was the distribution of the target data on the display. Participants arrived at conclusions much more quickly when all the test results were well contained in the display. For these distributions, participants were similarly fast on both display formats. For the worst-case distributions, the hierarchy+elison display trials were slower than those of the linear display. This could be due to the more difficult conditions we tested in the hierarchical case where, at worst, all the data required for the task were in different categories. In contrast, data were only distributed among two screens in the worst-case scenario of the linear display. For both display types, target order had no effect on time or accuracy, even though items in mismatch order required "back-and-forth" target searches. We were surprised by this result and can only speculate that the relatively small number of targets (five) was not enough for the disruptive effects of scrolling to impact performance time. For the linear display, we found an interaction effect between target distribution and target order in task time, with a much larger difference between distributions with ordered targets (8 sec) than those with unordered targets (2.5 sec). This suggests that the effects of target distribution were masked when the targets were unsorted. Despite the lack of a main effect for target order, questionnaire comments suggest a negative effect. For the linear display, participants considered scrolling a major problem (23 reports), whereas for the hierarchy+elision display, opening and closing of categories was problematic (15 re-

ports). The absence of effects for accuracy is not surprising, given the high accuracy achieved by the participants.

The preference results agreed with the overall time results. Our participants did not prefer a display for the overall task, and they did not have a preferred display for the component tasks of locating targets or drawing conclusions. Answers to the open-ended questions suggest that preference for the display formats was dependent on target distributions, corroborating the time results.

5. STUDY 2—QUALITATIVE EVALUATION BY PHYSICIANS

5.1. Design

Participants. Personal contacts were used to recruit 7 physicians (3 women, 4 men). Although the sample was largely determined by the available contacts, informants were also chosen to reflect a range of clinical experience. Our sample case was from family medicine. Our participant physicians included 5 active family medicine clinicians and 2 psychiatrists. These last 2 would be considered inexperienced for our sample case, because its domain is outside their area of specialization. Our family physicians participants can be grouped as 2 experienced senior family physicians (S1, S2) and 3 less experienced physicians (1 intern [I] and 2 junior family physicians [J1, J2]). All participants considered themselves infrequent computer users, and only 4 had ever used a personal digital assistant, or PDA. Participants were not compensated for their participation in our study.

Study protocol and materials. Each session was conducted separately, and took 30 to 45 min. The sessions were not recorded electronically, but verbal comments by the participants and areas of confusion were noted. All participants were interviewed in the medical institutions where they worked, although they were not interviewed in their offices.

Participants were asked to diagnose a clinical case taken from a Web site for training family medicine interns (Goodfriend, 2005). The recommended diagnosis was diabetes, a common condition in family medicine. The written part of the case closely resembled hospital charts reporting patient cases. The "laboratory result" section of the clinical case was presented on a Hewlett-Packard iPAQ 3850 Pocket PC (Hewlett-Packard, Palo Alto, CA). Participants were encouraged to explore the software even after they had arrived at a diagnosis and to "think aloud" during their explorations.

Both display formats were available on the display (Figure 1). Participants could select between the two by tapping on a "mode" icon. To display the laboratory report using the hierarchy+elision technique, we needed to group the data. Although it initially seemed natural to adopt the grouping scheme used by local laboratories, a closer look revealed a potential problem. Most of the commonly ordered tests are subgroups of Chemistry, such as Chemistry→Liver function tests→Bilirubin and Chemistry→Renal function tests→BUN. Retaining this three-level structure

would increase navigation complexity. Furthermore, the three-level structure does not provide additional clinical information, because Chemistry is usually considered a miscellaneous category by physicians. We consequently limited the hierarchical structure to two levels. Test subgroups belonging to the Chemistry group were treated as main groups, whereas systemic tests (e.g., glucose) were retained in the Chemistry category. In addition to this modification, we modified the case to use Canadian rather than American units and reference ranges.

To reduce visual clutter and reinforce the hierarchy+elision structure, hue, luminance, and saturation were used to create an illusion of depth (van Laar, 2001). For the hierarchy+elision interface (Figure 1a, b), four perceptual layers were created: (a) group headings, (b) test names and values that fall outside the accepted range, (c) tests within the normal range, and (d) reference information for the tests. For the linear interface (Figure 1c), only the lower three perceptual layers were used.

5.2. Results

Overall, the handheld displays were both accepted and considered to be effective. All physicians concluded that the patient was diabetic, although the psychiatrists expressed less confidence in this conclusion, as they had less clinical experience with the condition. Despite their relative inexperience with computer interfaces, all the physicians understood the hierarchy+elision technique quickly and could interact with the system without problems. All found the hierarchy+elision display familiar and commented that the interface resembled current paper reports. Initially, some had reservations about the modified categorization due to the extraction of organ-specific tests from the Chemistry category. However, after noting the size of the display and the number of tests that can potentially appear in the Chemistry category, these physicians preferred to have the specific tests extracted out of Chemistry.

Several physicians commented favorably on the hierarchy+elision display's ability to focus on interesting laboratory results. One family physician mentioned that the elision technique helped reduce the need for scrolling, a task she found difficult on the iPAQ, whereas another family physician said that with the hierarchy+elision display, there would be "no need to hunt through pages for recent results."

There were distinct differences in display preference depending on level of family medicine experience. The physicians less experienced in the case domain (I, J1, J2, and the psychiatrists) used the hierarchy+elision display to make their decision and used the linear display only when they were prompted to explore it. In fact, J2 did not believe the linear view would be useful at all (as it was not provided by the existing paper system), whereas J1 commented that the linear view would be handy only when looking for specific test results. Some of these physicians seemed to be gathering information while reading the report ("So his HBA1C is also up … ").

The two senior family medicine clinicians, S1 and S2, behaved quite differently. After some initial exploration of the hierarchy+elision view, both switched to the linear view and used it exclusively, calling out test names to search while looking at the results ("Now, how is his HBA1C?"). In fact, S2 was annoyed when one of the tests he was looking for was not included in the report.

5.3. Discussion

It appears that different physician populations use the display quite differently. Domain experts S1 and S2 developed a list of tests to look for after reading the clinical case. This indicates that they developed a schema during the clinical investigations that served as the basis for their medical reasoning. The laboratory results were merely pieces to be added to that schema. In short, they did not need the display as an external aid in decision making (although our sample size of two experts is too small for definitive conclusions).

These observations are in accordance with results in medical cognition, where domain experts are believed to develop a highly structured schema before reading laboratory reports (Patel et al., 1994; Patel et al., 1997). These observations also confirm our hypothesis that the linear format would provide better support for the clinical reasoning of domain experts who would not necessarily benefit from grouping of results. Even though the linear format is not used by current paper systems, its unfamiliarity may not inhibit its acceptance by experienced clinicians. Indeed, they might even adopt it enthusiastically.

6. CONCLUSION

The results of our studies indicate that neither the linear display nor the hierarchy+elision display is inherently better for displaying lists. Rather, the nature of the task and the subjective preference of the target users determine the most suitable display format. If the task requires data that can be predominantly grouped within a small number of hierarchical groups, the hierarchy+elision display would be preferred provided the user's classifications of the items match the hierarchy. All our study participants appreciated the focus the hierarchical display afforded, and our less experienced physician participants also appreciated the external aid the organization provided. On the other hand, our participants noted that the hierarchical display required mental effort to categorize the targets and physical effort to open and close categories. Some of the quantitative study participants considered this additional effort disruptive to an already cognitively intensive task. This may explain the two experienced physicians' preference for the linear format when working on the clinical problem, as theories of medical cognition suggest that they would be less likely to find the external organization of the hierarchical display useful in their task. This trade-off is also indicated by the quantitative study participants' lack of overall preference for one format, as the protocol exposed them to each format under its best and worst conditions. In future studies, we suggest that preference data be separately gathered for each condition of each format, rather than the overall preference for format across all tasks and distributions, to obtain more precise measurements of participants' subjective preferences.

The choice of display format is therefore based on the interaction of data distribution and user expertise. Our quantitative results suggest that linear display and hierarchy+elision display provide equivalently fast and accurate performances for clinical decision making. Speed was strongly influenced by the distribution of the

target items on the display, which seems to favor the hierarchy+elision display due to the nature of the clinical decision-making task. Our qualitative results, combined with prior results on medical cognition, suggest that both display formats are appropriate for small-screen display of medical data for clinical decisions. In fact, even though traditional paper reports do not offer a linear list of laboratory results, our 2 experienced physician participants seemed enthusiastic in its initial use, and the unfamiliarity of the format did not seem to hinder its acceptance. We recommend providing both display types.

Our quantitative study looked at a task and a data type targeted for the medical laboratory report system. Our data was a list of 20 to 30 items that can be clustered into five groups with similar number of items per group. To generalize our results, it would therefore be interesting to study how the number of groups and the total number of items per group affect the choice of display type. In fact, some of our quantitative study participants speculated that the hierarchy+elision displays would be useful for large numbers of results, but the linear display would be preferred if there were too many categories. As in menu design, the breadth/depth trade-off may prove to be important in displaying lists on small screens.

On the clinical side, our current qualitative study is only a preliminary screen of possible factors determining acceptance and preference of the two display types by the medical community. We wish to test our prototype in clinical contexts that incorporate other potentially important factors such as the different clinical goals for new visits and follow-ups or the compatibility of the displays with physicians' current workflows. In addition we would like to do a more complete study of the effect of physicians' clinical experience on display usage and preference.

REFERENCES

Altshuler, C. H. (1994). Data utilization, not data acquisition, is the main problem. *Clinical Chemistry, 40,* 1616–1620.

Bonini, P., Plebani, M., Ceriotti, F., & Rubboli, F. (2002). Errors in laboratory medicine. *Clinical Chemistry, 48,* 691–698.

Buyukkokten, O., Molina, H. G., & Paepcke, A. (2001). Accordion summarization for end-game browsing on PDAs and cellular phones. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 213–220.

Buyukkokten, O., Molina, H. G., Paepcke, A., & Winograd, T. (2000). Power Browser: Efficient Web browsing for PDAs. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 430–437.

Chen, E. S., Mendonça, E. A., McKnight, L. K., Stetson, P. D., Lei, J., & Cimino, J. J. (2004). PalmCIS: A wireless handheld application for satisfying clinician information needs. *Journal of American Medical Informatics Association, 11,* 19–28.

Furnas, G. W. (1997). Effective view navigation. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems,* 367–374.

Goodfriend, T. (2005). *PCC Case 3—Woman with high blood pressure. Diagnostics: What objective tests should now be done?* Retrieved January 22, 2005 from http://www.fammed.wisc.edu/pcc/curr/hypertension/case3/intro.html

Jones, M., Buchanan, G., & Mohd-Nasir, N. (1999). Evaluation of WebTwig—A site outliner for handheld web access. *Proceedings of International Symposium on Handheld and Ubiquitous Computing,* 343–345.

Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K., & Buchanan, G. (1999). Improving web interaction on small displays. *Proceedings of the WWW8 Conference*, *May 11-14,* 51–59.

Jones, S., Jones, M., & Deo, S. (2004). Using keyphrases as search result surrogates on small screen devices. *Personal and Ubiquitous Computing, 8,* 55–68.

Kamba, T., Elson, S., Harpold, T., Stamper, T., & Piyawadee, N. (1996). Using small screen space more effectively. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 383–390.

Kiger, J. I. (1984). The depth/breadth tradeoff in the design of menu-driven user interfaces. *International Journal of Man Machine Studies, 20,* 201–213.

Lam, H. (2004). *Displaying medical laboratory reports on small screens*. Unpublished master's thesis, Simon Fraser University, Burnaby, Canada.

Mayer, M., Wilkinson, I., Heikkinen, R., Ørntoft, T., & Magid, E. (1998). Improved laboratory test selection and enhanced perception of test results as tools for cost-effective medicine. *Clinical Chemistry Laboratory Medicine, 36,* 683–690.

Miller, D. P. (1981). The depth/breadth tradeoff in hierarchical computer menus. *Proceedings of the Human Factors Society 25th Annual Meeting*, 296–300.

Nygren, E., Wyatt, J. C., & Wright, P. (1988). Helping clinicians to find data and avoid delays. *Lancet, 352,* 1462–1466.

Patel, V. L., Arocha, J., & Kaufman, R. (1994). Diagnostic reasoning and medical expertise. *The Psychology of Learning and Motivation, 31,* 187–252.

Patel, V. L., Groen, C. J., & Patel, Y. C. (1997). Cognitive aspects of clinical performance during patient workup: The role of medical expertise. *Advances in Health Sciences Education in Theory and Practice, 2,* 95–114.

Tullis, T. S. (1997). Screen design. In M. Helander, T. K. Landauer, & P. Prabbu (Eds.), *Handbook of human–computer interaction* (2nd ed., pp. 503–531). Amsterdam: Elsevier Science BV.

van Laar, D. L. (2001). Psychological and cartographic principles for the production of visual layering effects in computer displays. *Displays, 22,* 125–135.

Watters, C., Duffy, J., & Duffy, K. (2003). Using large tables on small display devices. *International Journal of Human-Computer Studies, 58,* 21–37.

Wright, P., Jansen, C., & Wyatt, J. (1998). How to limit clinical errors in interpretation of data. *Lancet, 352,* 1539–1543.

Ye, N., & Salvendy, G. (1994). Quantitative and qualitative differences between experts and novices in chunking computer software knowledge. *International Journal of Human-Computer Interaction, 6,* 105–118.