Research Paper ■

Quantifying Visual Similarity in Clinical Iconic Graphics

PHILIP R. O. PAYNE, MA, JUSTIN B. STARREN, MD, PHD

Abstract **Objective:** The use of icons and other graphical components in user interfaces has become nearly ubiquitous. The interpretation of such icons is based on the assumption that different users perceive the shapes similarly. At the most basic level, different users must agree on which shapes are similar and which are different. If this similarity can be measured, it may be usable as the basis to design better icons.

Design: The purpose of this study was to evaluate a novel method for categorizing the visual similarity of graphical primitives, called Presentation Discovery, in the domain of mammography. Six domain experts were given 50 common textual mammography findings and asked to draw how they would represent those findings graphically. Nondomain experts sorted the resulting graphics into groups based on their visual characteristics. The resulting groups were then analyzed using traditional statistics and hypothesis discovery tools. Strength of agreement was evaluated using computational simulations of sorting behavior.

Measurements: Sorter agreement was measured at both the individual graphical and concept-group levels using a novel simulation-based method. "Consensus clusters" of graphics were derived using a hierarchical clustering algorithm.

Results: The multiple sorters were able to reliably group graphics into similar groups that strongly correlated with underlying domain concepts. Visual inspection of the resulting consensus clusters indicated that graphical primitives that could be informative in the design of icons were present.

Conclusion: The method described provides a rigorous alternative to intuitive design processes frequently employed in the design of icons and other graphical interface components.

■ **J Am Med Inform Assoc.** 2005;12:338–345. DOI 10.1197/jamia.M1628.

Background

Graphical user interface models such as those found in the Microsoft Windows™ or Mac OS™ operating systems have become ubiquitous in the modern computing environment. A key component of such interface models is the use of graphical elements, such as icons, as presentation models[1] for tasks and objects within a given context (for example, the folder and document icons commonly used to present file storage objects in the context of a computer's operating system). The usefulness of icons in the domain of health care computing applications has been evaluated in studies examining their effectiveness in conveying complex medical data or concepts, such as that found in medical records[2] and radiology reports.[3] However, despite a significant amount of literature describing the evaluation of existing icons and ways in which icons can be combined within graphical user interfaces,[4] there is a paucity of literature describing rigorous methodologies for the initial design of icons or their graphical components.

Affiliation of the authors: Department of Biomedical Informatics, Columbia University, New York, NY.

Supported in part by NLM training grant N01-LM07079.

Correspondence and reprints: Philip R. O. Payne, MA, Department of Biomedical Informatics, Columbia University, 622 West 168th Street, VC5, New York, NY 10025; e-mail: <philip.payne@dbmi.columbia.edu>.

Received for publication: 06/01/04; accepted for publication: 01/10/05.

The need for rigorous methods for the design or selection of icons varies with application domain. In some application domains, for example word processing, it is relatively easy to locate graphic designers with domain expertise and an intuitive grasp of the metaphor of reference (e.g., the desktop metaphor).[5] The health care environment is one of many that do not share this advantage. In addition, within the health care environment, there are multiple subdomains (e.g., administrative, clinical, research), each with unique metaphors. These subdomains, combined with the breadth of data elements and concepts that need to be modeled in health care applications, lead to significant challenges in the design of suitable iconic presentation models. As in other highly technical areas, it is difficult to find individuals who possess both interface design skills and an intuitive understanding of the given domains concepts. Finally, owing to the time-sensitive nature of many health care activities and the need for both rapid and accurate interpretation of critical data, ambiguity concerning the intended meaning of an icon or graphical element is highly undesirable. Because of these difficulties, the conventional intuitive approach to the design of icons[6] is far from ideal for the design of health care computing applications.

A more structured approach to the design of icons has been proposed, known as Presentation Discovery.[7] Presentation Discovery consists of four major steps:

1. The identification of target domain concepts for use in a presentation model
2. The elicitation of candidate graphical primitives that represent the selected domain concepts from domain experts

3. The categorical sorting of candidate graphical primitives into consensus clusters based on their visual characteristics
4. The extrapolation of representative prototype graphics from the consensus clusters.

This approach incorporates both the knowledge of domain experts during the generation of candidate graphical primitives and the personal constructs used by the targeted end-user community during the subsequent categorical sorting phase of the methodology. An important characteristic of Presentation Discovery, compared with other methods that have been proposed to evaluate iconic presentation models, is that it is intended for use in the early design phase of a software project (e.g., during initial prototype development and iterative refinement) to enable designers to select the best constituent members of an iconic presentation model.

This study focuses on steps 3 and 4 of the Presentation Discovery methodology, which supposes a level of consistency among experts during the elicitation phase of the process. This work extends the Presentation Discovery methodology by providing novel techniques to evaluate such consistency, thus mitigating potential problems that could result from an inconsistent body of prototypical graphics. In particular, it evaluates several of the premises that underlie Presentation Discovery. In doing so, the authors intend to address the following questions:

1. Do multiple sorters without domain expertise group graphics together in a similar manner?
2. Will the application of conventional hypothesis discovery tools to the sorted graphics yield clearly differentiable groupings of prototype graphics?
3. Will the resulting groups of graphics correlate back to the original domain concepts?

The experimental context is the development of icons that may be used to present concepts commonly found in textual mammography reports. It is important to note that the evaluation of these premises is limited by methods used to identify the domain concepts, including the corpus analysis techniques used, which rely on automated natural language processing and expert knowledge elicitation. Furthermore, the types of concepts modeled in the described study are limited to diagnoses and anatomic modifiers and therefore cannot project the effects of representing more complex procedural concepts. These limitations are discussed in greater detail later.

Methods

Generation of Candidate Graphical Primitives

Fifty typical mammography findings were used to generate the graphics in the original Presentation Discovery study and in this analysis. These findings were derived from the Columbia University Medical Center Clinical Data Repository (CDR), which contains more than ten years of mammography reports. These reports have been processed using the MedLEE Natural Language Processing System since 1994.[8,9] This allows reports to be retrieved based on specific findings. A query of the CDR yielded 10,000 mammography reports that had been coded using the MedLEE system. Reports containing negative findings were eliminated from the initial group, and 10% of the remaining reports were selected at random. From the randomly selected group of reports, duplicates or reports containing nondescriptive findings were eliminated, and the remaining reports were inspected manually, yielding a group of 105 commonly used findings. From these 105 findings, 50 were selected manually to include roughly equal portions of masses, calcifications, and other findings (Appendix 1, available as an online data supplement at www.jamia.org).[7]

Six radiologists were each given a survey booklet with 50 findings. Each page of the booklet contained one of the 50 selected textual findings, an anatomic outline of the breasts, and two check boxes allowing the subject to indicate whether the best presentation of the specific finding was a blank diagram or that the subject could not ascertain a way to draw the given finding. The radiologists were then instructed to draw what they considered the best graphic for the given finding or to check one of the provided boxes as necessary (Fig. 1).[7]

Sorting

Of the resulting 300 survey pages, those that did not contain a graphic were discarded; the remaining 238 were masked to prevent subsequent sorters from reading the textual finding associated with each drawing, and the order of the surveys was randomized. A group of six graduate students in the Columbia University Department of Biomedical Informatics was asked to sort the surveys into groups based on the visual similarity of the graphics, a process known as categorical sorting.[10] Nonradiologists were intentionally selected as sorters for two reasons. First, it was thought to be a more rigorous test of the method because the sorters would use only the appearance of the graphics rather than higher level domain knowledge to inform the sorting process. Second, in many instances, it may be difficult to recruit sufficient domain experts for both the elicitation and the sorting phases, making the use of nonexpert sorters desirable. Therefore, it is important to validate the ability of such nonexperts to reliably sort graphics for which they do not have expert-level domain knowledge. Three of the subjects were medical doctors. Of the remaining three subjects, two had backgrounds in the natural sciences, and one had a background in computer science. None of the subjects were radiologists. The resulting groups created by each subject were recorded into a 238×237 element matrix for later analysis. The matrix corresponds to the pair-wise grouping of each graphic with the other 237. Whenever a sorter placed two graphics in the same group, the corresponding cell was assigned a value of 1; if a pair were in different groups, the cell was assigned the value of 0.

Analysis

The matrices for all sorters were summed together into a single 238×237 element agreement matrix (Fig. 2, available as an online data supplement at www.jamia.org). Each cell had a value between 0 (corresponding to no sorters placing the two indicated graphics in a group together) and 6 (corresponding to all sorters placing the two indicated graphics in a group together). For the purposes of subsequent data analysis, those cells in the agreement matrix that remained unassigned (e.g., those having a weight of 0) were censored.

Using the cell values recorded in the agreement matrix, aggregate user agreement ($\overline{A}$) for all pairings of a single graphic

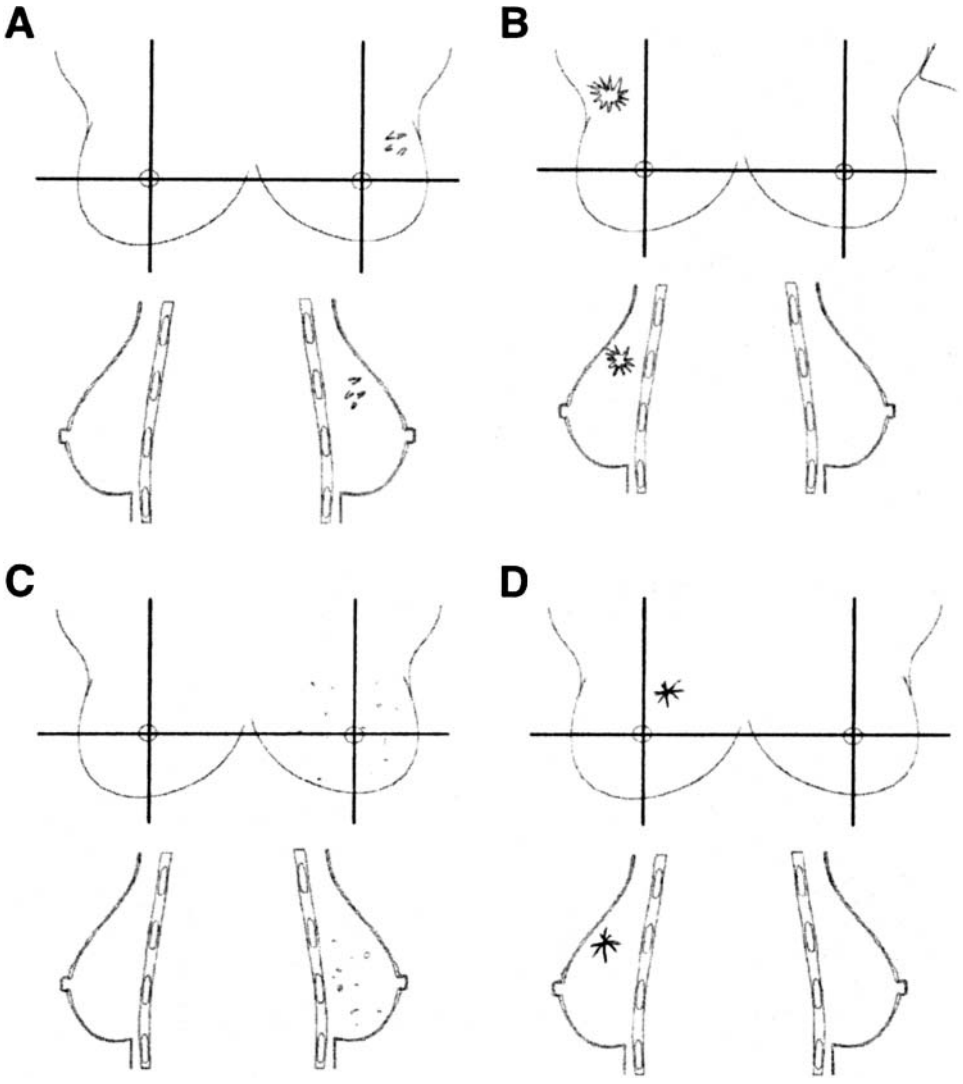

Figure 1. Example graphical prototypes generated by domain experts. For elicitation, each survey page contained the breast template shown and one textual finding. Each graphic corresponds to one of the following findings. **A:** "Multiple surgical clips are again noted in the outer upper quadrant of the left breast." **B:** "Malignant appearing lesion right breast ten o'clock." **C:** "There are scattered microcalcifications seen throughout the left breast. These calcifications most likely represent early dystrophic calcifications." **D:** "There is an 8-mm spiculated mass in the right breast at one o'clock that corresponds to an area of intense shadowing by ultrasound."

with the remaining 237 graphics (representing a single row or column of the agreement matrix) was calculated for each graphic using a simple average of non-zero cells:

$$\overline{A} = \frac{\sum_{i=1}^{n} s}{n}$$

Equation 1: Aggregate user agreement (where *s* is the score of any nonzero cell for the given row or column and *n* is the number of nonzero cells in that row or column). This yields the average agreement score for all possible pairings of a single graphic with all other graphics.

Thus, if a given row or column of the agreement matrix contains only scores of 6 or 0, the aggregate user agreement is 6. Similarly, if the sorters are evenly split and the given row or column of the agreement matrix contains only scores of 3 or 0, the aggregate user agreement is 3. If all the sorters agreed on one association, but only half assigned a second association (e.g., one cell is assigned a score of 6, one cell a score of 3, and multiple cells assigned scores of 0), the agreement would be 4.5.

Average group size ($\overline{G}$) for the set of unique groups of graphics created by the individual sorters was calculated using the following method:

$$\overline{G} = \frac{\sum_{i=1}^{j} S_j}{j}$$

Equation 2: Average group size (where *j* is the number of unique groups created by the sorter and *S* is the number of elements in a given unique group).

As described previously, for each unique textual finding, six radiologists were asked to draw representative graphics. Thus, sorters were presented with as many as six drawings representing the same domain concept. This group of drawings will be referred to as a concept group (e.g., the graphics represent a common textual finding). Concept-level agreement was analyzed by calculating how often graphics representing the same concept were grouped together by the sorters. This was computed by averaging the 16 unique combinations of the set of six graphics that comprised each such concept group.

Computational Simulation of Sorting Behavior

Because the sorters in this study were not given predefined groups into which the provided graphics were to be sorted, each observer created a different number of groups having a unique number of members. Such an outcome is common in studies using categorical sorting. Due to this variability in group size and composition, conventional statistical measures of interobserver agreement were not applicable.[11] A certain amount of apparent sorter agreement would occur by chance. To evaluate the level of actual versus random sorter agreement in the study data set, a computer simulation was constructed to emulate random sorting behavior. Summary statistics (such as aggregate agreement) could be computed on the simulated agreement matrix and compared with the observed values. The simulation could also be run multiple times to generate variance statistics. The authors hypothesize that the greater the difference between the observed and random matrices, the more likely it is that the observed agreement matrix represents the result of a reproducible, significant categorical organization of the objects under study by the sorters.

This simulation was implemented by modeling the results of sorting as a forest of fully connected graphs, where each vertex represents a single graphic and the edges between vertices represent the placement of two graphics into the same group. An exhaustive description of the simulation is beyond the scope of this work. Initial studies quickly demonstrated that simply populating an agreement matrix with random values yielded nonsensical results. Further study confirmed that the simulation needed to accurately replicate the behavior of multiple sorters grouping items together; assigning initial edges or incrementing edge weights between vertices during multiple iterations (Fig. 3, available as an online data supplement at www.jamia.org). To model observed sorting behavior in a more equitable manner to the sorting behavior of actual subjects, a number of constraints were incorporated into the simulation, as enumerated below:

- *Group size*: Larger groups produce higher levels of random agreement, thus the simulation needed to account for the group sizes in the real-world data set.
- *Group size variance*: Since there was significant variability in observed group sizes, it was necessary to model this feature computationally. Therefore, the simulation creates groups of sizes that fall within the statistical distribution found in the observed sorting study under consideration.
- *Transitive closure*: The simulation was designed to prevent the creation of nonsensical edges (e.g., nonsymmetric edges). It is important to note that the creation of nonsensical edges in this context would be a violation of the transitive nature of the edgewise connectivity required to comparably represent an observed agreement matrix. As an example, given three graphics, symbolically represented as A, B, and C, if A and B are in the same group, and B and C are placed in the same group, then A and C must be in the same group as well.

To generate statistics that could be compared with the real-world data, the simulation was run 100 times using the same numbers of sorters and groups as occurred in the real-world data set. Each simulated data set was recorded and subjected to the same analysis as that performed on the study data set to generate predicted aggregate agreement scores for the simulated agreement matrices. From these 100 scores, a mean and standard deviation were computed. These were subsequently compared with the observed data set using conventional statistical significance tests. One hundred runs were selected as a convenient threshold. Pilot studies with as many as 1,000 iterations were also performed. Visual inspection of the results confirmed that the model was linear above 100 iterations.

Hypothesis Discovery Tools

The previous metrics are able to determine sorter agreement at the individual graphic or pairwise level but do not measure larger clusters or higher order relationships. To evaluate higher order relationships in the study data set, hypothesis discovery tools were used. The goal of hypothesis discovery tools is to provide an objective or semiobjective means of determining previously unknown or unrecognized classifications or groupings within a large data set. In this study, cluster analysis (a type of hypothesis discovery tool) was used to determine the underlying organization of the multiple groups of graphics created by the sorters. Cluster analysis can be generally described as "an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise" (http://www.statsoft.com/textbook/stcluan.html). Specifically, an unweighted pair group centroid hierarchical clustering algorithm was applied to the agreement matrix. Such hierarchical clustering algorithms are agglomerative, beginning with each graphic as an individual cluster and recursively aggregating clusters based on Euclidean distance measures derived from the pairwise comparison of the individual row or column that defines the sorting characteristics of a given cluster (representing either a single graphic or an aggregate sequence derived from previously clustered graphics).[12,13]

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2}$$

Equation 3: Euclidean distance measure between cluster i and cluster j, where each cluster is characterized by p parameters. The distance is calculated from the pairwise comparison of each matching set of parameters in the respective clusters.[13]

Subsequent analyses can then be performed using statistical measures applied to the proximity matrix (consisting of cluster elements and Euclidean distance measures) that results from the clustering algorithm. This type of algorithm was chosen because, unlike other similar hypothesis discovery tools, it does not require the preanalysis designation of cluster sizes (e.g., number of cluster members) and cluster instances (e.g., the number of resulting clusters). Given the variation in group size among the sorters, such a preanalysis designation would introduce undesirable bias into the data analysis process. The algorithm, as implemented in SAS Institute's JMP 5.0.1a statistics package running on Mac OS™ 10.3.2, produced the previously described proximity matrix, including intergraphic Euclidean distance measures that can be used as an indicator of cluster "tightness" (e.g., the higher the distance measure, the less closely related the elements of the cluster).[14] Cluster analysis was performed at both the individual graphic and concept group levels (e.g., each derived element in a cluster was the group of graphics that

represented a single textual finding). For each resulting cluster, the ratio of the number of unique findings associated with the graphics that comprise the cluster to the total number of graphics in the cluster was calculated. The clustering algorithm generated a dendrogram of the resulting clusters, which was used for visual inspection of the results (Fig. 4, available as an online data supplement at www.jamia.org). Dendrograms are a visualization of the hierarchical structure generated by the clustering algorithm, providing a "tree-like diagram that summarizes the process of clustering…similar cases are joined by links whose position in the diagram is determined by the level of similarity between the cases" (http://obelia.jde.aca.mmu.ac.uk/multivar/gloss.htm).

Data Visualization

To evaluate the agreement and proximity matrices for discernible patterns, data visualization tools were applied to the study data set. Prior to visualization, one axis of the agreement matrix was sorted based on the order of graphics in the dendrogram (above) and the other axis was sorted based on original domain concept. Using this sorted matrix, a heatmap was constructed, using HeatMap Builder (http://quertermous.stanford.edu/heatmap.htm) (Fig. 5). The heatmap was then inspected visually and selectively annotated to indicate patterns associated with discrete concept meanings of interest.

Results

The six sorters generated 169 unique groups of graphics. The groups created by the individual sorters ranged in size from 6 to 68 graphics. Average group size ($\overline{G}$) was found to be 8.46 and aggregate observer agreement ($\overline{A}$) was 4.07 with a standard deviation of 0.65. Of particular interest was the distribution of group size when comparing sorters with a medical background who consistently created fewer, larger groups and sorters with a nonmedical background who created a larger number of smaller groups, especially as was the case with the sorter whose background was in computer science (Fig. 6, available as an online data supplement at www.jamia.org). Compared with the observed sorter agreement, the computer model yielded a predicted aggregate agreement of 0.42 with a standard deviation of 0.17. The magnitude of this difference (5.6 standard deviations), in the absence of conventional statistical measures for such data, demonstrates that agreement among the sorters was highly significant (Fig. 7, available as an online data supplement at www.jamia.org). While visualization of agreement at the individual graphic level does indicate the existence of a number of "near plateaus," the number of these phenomena approximates the number of sorters, and therefore the authors have concluded that it would be premature to postulate that this demonstrates the existence of a true, reproducible phenomenon.

Consensus clustering yielded 87 unique clusters of graphics, with an average cluster size of 4.4 elements (e.g., graphics) and an average distance between cluster elements of 2.84 (Fig. 8, available as an online data supplement at www.jamia.org). The average ratio of unique findings to total elements in a given cluster was found to be 1.56. The resulting consensus clusters generated by the complete hierarchical clustering algorithm were inspected visually and found to contain usable groups of prototypical graphical elements (Fig. 9, available as an online data supplement at www.jamia.org). These clusters consistently exhibited "tight" intergraphic distances, which were reflected when visually comparing the graphics. A concept-group aggregate agreement score was produced by calculating the mean aggregate agreement for all six graphics associated with each original textual finding. The analysis at the concept-group level yielded an observed aggregate agreement ($\overline{A}$) of 4.16. Concept-group level aggregate agreement predicted from the simulation was 1.02 with a standard deviation 0.20 (Fig. 10, available as an online data supplement at www.jamia.org), indicating a high level of statistical significance. Cluster analysis at the concept-group level yielded 15 clusters, with an average cluster size of 4.27 unique findings and an average distance between findings of 5.33.

The relationship between the clusters of graphics and the underlying concepts can be explored through the use of data visualization tools.[13–15] The heatmap visualization used in this study (Fig. 5) is sorted along the vertical axis based on the cluster order generated by the hierarchical clustering of individual graphics (as is reflected in the resulting dendrogram). The heatmap is sorted along the horizontal axis by mean cluster order for the graphics associated with the previously defined concept groups. As a result, if the sorters always sorted graphics associated with different findings into different groups, the heatmap would show a single diagonal one cell wide. The more that graphics derived from different findings are sorted into similar groups, the greater are the width and "fuzziness" of the diagonal. It is clear from the high level of structure in the heatmap that specific meanings are strongly associated with certain graphics. This also shows that some findings demonstrated more overlap with other findings and associated graphics than others. The implications of this are discussed later.

Discussion

The results of this study support the assumptions underlying the use of Presentation Discovery as a method for developing the graphical primitives of an iconic presentation model. Specifically, the following two assumptions were found to be true:

- Nondomain experts reliably and consistently grouped graphics into similar groups. Aggregate sorter agreement ($\overline{A}$) was found to be extremely high at both the individual graphic (~67.8%) and concept-group (~69.3%) levels, especially in comparison with the predicted values generated by the computer model.
- Application of clustering algorithms to the sorted groups produced consensus clusters that were similar enough to have produced prototypical iconic primitives.

The purpose of this study was to evaluate components of a proposed method for the prospective selection of graphical primitives for use in the development of new icons. At the most basic level, this approach presumes that such a presentation model would provide acceptable usability in the human-computer interaction context. For this to be the case, end users must be able to interpret the meaning of such icons efficiently and with a minimum of ambiguity. As stated earlier, there is a paucity of literature that addresses this type of prospective selection process. However, several studies have retrospectively evaluated the ability of end users to interpret icons similarly,[2,3] lending support to such an argument. The results

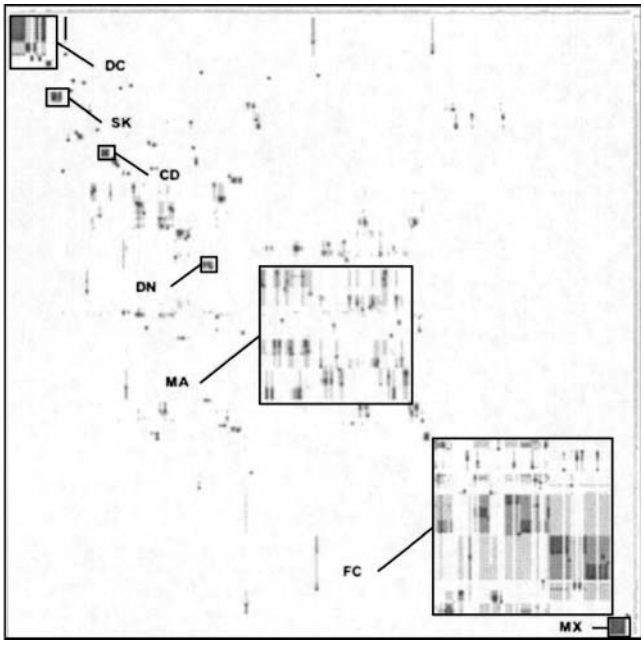

Figure 5. Heatmap of concept-group clusters ordered by concept-group cluster index on the X axis (e.g., ordered by an average hierarchical clustering algorithm location index for each member of the concept-group) and hierarchical clustering index on the Y axis (e.g., the hierarchical clustering algorithm location index assigned to each graphic when reading from left to right in the resulting dendrogram). Each cell in the heatmap is of equal size, and the intensity of cell shading indicates the level of agreement observed between the two graphics as indicated by the X and Y coordinates of the cell. This visualization demonstrates the existence of overlapping concept meanings for some graphics. Selected clusters of interest are annotated with the type of concept-group finding most frequently present in the given region. DC = diffuse calcification; SK = skin thickening; CD = cystic disease; DN = density; MA = masses; FC = focal.

of these studies have led to the conclusion that in many instances, the speed and accuracy of interpretation of iconic models is superior to that demonstrated with alternatives such as text, tabular data, or traditional graphics (e.g., histograms or pie charts).[16] Such studies have focused almost exclusively on retrospective analysis of the usability of existing iconic presentation models rather that examining the qualities of such graphics prospectively, as was done in this study. This differentiation is critical to the underlying premises of Presentation Discovery, in that the method is intended for use in a domain in which no preexisting body of graphical primitives exists, which allows for the integration of both expert knowledge and subsequent hypothesis discovery tools, as described in the preceding discussion. However, such retrospective studies have generated a number of theoretical constructs that are helpful in providing context to the results of this study.

One such construct postulates that iconic presentation models derive usability from the manner in which end users interpret a superclass of interface objects known as "metaphor graphics."[17] The theory behind the use of "metaphor graphics" is that such presentation models recontextualize data from the digital form, allowing users to use inherent cognitive strengths in the areas of pattern matching and mental modeling to interpret and act on the data being presented. The expressiveness of a given "metaphor graphic" (e.g., the ability to convey a discrete datum or concept) relies on the level of mapping between its physical form and its intended function. This mapping can be classified as belonging to one of four classes[18]:

- *Resemblance*, in which the icon represents an image analogous to the intended concept
- *Exemplar*, in which the icon serves as an example of a class of concepts
- *Symbolic*, in which the icon conveys a reference to a level of abstraction higher than that of the intended concept
- *Arbitrary*, in which the meaning of the icon must be learned by the user

In general, resemblance and exemplar metaphors are believed to be more intuitive than symbolic or arbitrary metaphors.[19] The graphics created in this study appear to be predominantly of the first two types. However, a significant limitation to the metaphor graphic construct is that it does not provide a method for creating such graphic but rather only proposes a means of evaluation.

An alternative construct is derived from the field of semiotics, a discipline that examines the design, interpretation, and derivation of meaning for "signs."[20] A "sign" is defined as an object that is intended to represent a concept to the intended end user. The semiotic construct defines three components involved in the interpretation of "signs": an object, which is the concept to be represented; the representamen, which is the sign used to represent the object; and the interpretant, which is the concept as understood by the intended end user. The relationship between these three components, known as semiosis, has been systematically represented as the Pericean Triad (Fig. 11, available as an online data supplement at www.jamia.org). Implicit in the Pericean Triad are three key subrelationships[5]:

- The relationship between the object and the representamen, known as the "representation" relation
- The relationship between the representamen and the interpretant, known as the "interpretation" relation
- The relationship between the interpretant and the object, known as the "effect" relation.

The overall effectiveness of a sign is dependent on the strength of these three relationships. Weaknesses between any of the relationships in the triad may lead to an erroneous effect relation. Most studies of iconic presentation models focus on the effect relation, and to a lesser extent, the interpretation relation. Little attention is paid to the representation relation, thus omitting a key component of the triad. This is in large part a consequence of the fact that most applications of semiotics to icon design have been limited to studies of existing icons in which the representation relation is often taken as a given. Presentation Discovery complements semiotic-based evaluation methods because it provides a way to address the representation relation prospectively.

A third theoretical construct related to Presentation Discovery is Kelley's Personal Construct Theory.[21] This theory proposes that people contextualize information through categorization and that they are able to reliably communicate these categories to other individuals. The categorical sorting technique used in Presentation Discovery is a variant known

as an "all in one sort," in which the sorter performs a single sort of all objects to be sorted based on either sorter, as is the case in the Presentation Discovery process, or investigator-specified criteria. The use of the all in one sort is well suited to situations in which the study designers are interested in the quantification of inherent categories within the body of items to be sorted (usually through the application of hypothesis discovery or statistical tools).[10] In the case of Presentation Discovery, this process is used to quantify relationships between newly discovered graphical primitives as perceived by potential end users of the resulting presentation model under study. Thus, Presentation Discovery can be viewed as an operationalization of Personal Construct Theory.

In addition to the theoretical constructs, which may serve to inform the design of iconic presentation models, a number of practical issues are regularly reported on in literature describing icon design methodologies. Commonly described difficulties in the development of iconic languages include the problems of expressiveness and ambiguity.[6,17] One of the judgments that icon designers must make is deciding the number of discrete concepts in a domain that can be reliably represented by icons. Although there are no concrete rules for this decision, the tools demonstrated in this study can be used in two ways to address this problem. First, the number of consensus clusters provides an insight regarding the number of domain graphics that are easily differentiable. For example, the existence of 15 unique concept-group clusters of graphics in contrast to 50 textual findings used to generate the graphics is highly suggestive that it will be impossible to create an icon set that reliably captures all the distinctions among the original 50 findings. Anecdotally, the original Presentation Discovery analysis, using slightly different methods, resulted in 16 different prototypical iconic primitives.[7] Given the similarity of some concepts in the set of original mammography findings, such as "suspicious calcification" versus "suspicious microcalcification," this is not surprising. Furthermore, visualization of the consensus clusters created by the sorters, as seen in the previously described heatmap (Fig. 5), provides insight into which concepts will be easiest to differentiate. It can be seen in the heatmap that some pairs of initial findings show almost identical patterns of associated graphics, while others show no overlap. The pattern observed in the heatmap indicates that masses and calcifications will be easiest to differentiate, while distinctions within these groups will be harder to discern. The pattern further suggests that some groups of graphics, such as the central mass of graphics attributed to masses, are more heterogeneous. Such a finding is to be expected and leads to the hypothesis that for some constructs, multiple candidate graphics will be generated by the Presentation Discovery methodology. One potential explanation for this phenomenon is the varying artistic abilities of the domain experts who initially provided the candidate graphics used in the study. An open question that results from these findings is whether metrics may be applicable to the described data types to determine the maximum number of discernible graphics in this type of study or if such a process will remain reliant on manual inspection and knowledge-based heuristics.

This work lies at the interface between Knowledge Acquisition (KA)[10,22–26] and Requirements Analysis (RA).[27] Traditionally, the definition of the internal concepts and logic of an application have been separated from the design of the visual presentation. The former (i.e., KA) involves expert system developers and focuses on determining the ways in which domain experts conceptualize and reason about a domain. The application of KA methodologies in the biomedical domain has been reported on in a number of instances, usually in the context of defining and structuring clinical guidelines and knowledge-based heuristics.[23,28,29] Numerous KA techniques exist, for example, interviews, formal and informal usability analysis, repertory grid analysis, laddering, and various sorting techniques (e.g., card sorts, categorical sorts, and picture sorts).[10,26,30,31] All these techniques rely on the ability of domain experts or targeted end users to categorize concepts from the domain in a reliable manner. Personal construct theory suggests that this ability is derived from the inherent capacity of humans to contextualize and make sense of their surroundings through a process of creating and communicating categories.[21] In contrast, the design of the visual presentation and work flow of an application (i.e., RA) has traditionally involved interface designers using different techniques and assumptions.[27] Like KA, RA often involves the elicitation of information from domain experts. However, this separation is not absolute, and techniques conventionally associated with KA have been used to address purely visual aspects of user interfaces.[10,22,24] In fact, the growing field of Interaction Design[27,32] is grounded on the belief that the visual presentations and underlying conceptual models must be more tightly integrated. To our knowledge, Presentation Discovery is the first application in a biomedical domain of techniques, normally associated with KA, to the purely visual presentation aspects of an application.

Limitations

The described study is limited due to the use of a data set derived from a single discipline (e.g., radiology reports) and the nature of the findings used to generate the set of candidate graphical primitives, which represented only objects (nouns) and not actions (verbs) or modifiers (adjectives). Although the graphics came from a specific domain, the methods used in this study are domain independent. In particular, the original textual findings were identified through text mining of a large corpus using a natural language-processing technique to minimize researcher-induced bias. However, this approach may not be applicable to all domains owing to the scarcity of accurate natural language-processing systems, and therefore it may be necessary to use other methods, such as expert consensus or manual review of the source data, to develop the initial concept list. The subsequent methods used were selected to minimize reliance on domain-specific characteristics or a priori knowledge of the data set. Therefore, the methodology could be applied to a wide variety of health care domains in which a similar finding-at-location descriptive structure is used to represent narrative information. It is also important to note that during the second phase of the Presentation Discovery process, nonexperts were used to sort the study graphics; this approach was used to replicate real-world circumstances, where it may be difficult to recruit domain experts for both phase one and two. However, it could be posited that the use of domain experts for sorting might yield different results. Direct comparisons between expert and nonexpert categorizations of graphical

primitives were beyond the scope of the present study. The role of domain expertise in the sorting phase of Presentation Discovery process remains an open question,[33,34] and one of ongoing research.

Conclusion

The objective of this study was to explore some of the assumptions underlying Presentation Discovery. The results have demonstrated that these assumptions hold true in the context of icons intended to represent objects in one domain, specifically concluding that multiple sorters could reliably group such graphics together into similar groups and that through the use of hypothesis discovery tools, higher order groups could be identified. The applicability of these methods to other concept classes, such as processes, remains an open question. Furthermore, this study has provided two contributions to the development of human–computer interfaces. First, it demonstrated that nonexperts in a domain can reliably sort domain-specific graphics. Second, it demonstrated a novel method for the measurement of such agreement through the development of a new simulation-based statistical method for assessing user agreement for categorical sorting. These statistical methods will have application to many other interrater agreement problems in which the number and size of groups is not uniform.

References ■

1. Starren J, Johnson SB. An object-oriented taxonomy of medical data presentations. J Am Med Inform Assoc. 2000;7:1–20.
2. Litt HI, Schmidt DF. Application of the Visual Chart in an Ambulatory OB-GYN Clinic. Annu Symp Comput Appl Med Care. 1995:1004.
3. Abad-Mota S, Kulikowski C, Gong L, Stevenson S, Mezrich R, Tria A, et al. Iconic reporting: a new way of communicating radiologic findings. Annu Symp Comput Appl Med Care. 1995;915.
4. Chang SK, Polese G, Orefice S, Tucci M. A methodology and interactive environment for iconic language design. Int J Hum Comput Stud. 1994;683–716.
5. Barr P, Noble J, Biddle R, Icons R Icons. In: Biddle R, Thomas B, (eds). Australasian User Interface Conference; 2003; Adelaide, Australia. Sydney, Australia: Australian Computer Society; 2003:25–32.
6. Horton WK. The icon book: visual symbols for computer systems and documentation. New York: John Wiley & Sons; 1994.
7. Starren J. From multimodal sublanguages to medical data presentations. New York: Columbia University; 1997.
8. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med. 1995; 122:681–8.
9. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. J Am Med Inform Assoc. 1999;6:143–50.
10. Rugg G, McGeorge P. The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. Expert Syst. 1997;14: 80–93.
11. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002;35:99–110.
12. Kim JH, Kohane IS, Ohno-Machado L. Visualization and evaluation of clusters for exploratory analysis of gene expression data. J Biomed Inform. 2002;35:25–36.
13. Everitt B, Landau S, Leese M. Cluster analysis. 4th ed. New York: Oxford University Press; 2001.
14. SAS. JMP. In. 5.0.1a ed: SAS; 1999.
15. Tufte ER. The visual display of quantitative information. 2nd ed. Cheshire (CT): Graphics Press; 2001.
16. Elting LS, Bodey GP. Is a picture worth a thousand medical words? A randomized trial of reporting formats for medical research data. Methods Inf Med. 1991;30:145–50.
17. Cole WG. Metaphor Graphics & Visual Analogy For Medical Data. Annu Symp Comput Appl Med Care. 1987.
18. Rogers Y. Icons at the interface; their usefulness. Interact Comput. 1989;1:105–7.
19. Hutchins E. Metaphors in interface design. In: Taylor MM, Neel F, Bouwhuis DG, (eds). The structure of multimodal dialogue. New York, NY: Elsevier Science, 1989, pp 11–28.
20. Leite JC. A Semiotic-based framework to user interface design. In: Proceedings of the Second Nordic Conference on Human-Computer Interaction–NordiCHI 2002; Aarhus, Denmark. New York, NY: ACM Press; 2002. pp 263–6.
21. Kelly GA. The psychology of personal constructs. New York: Norton; 1955.
22. Boy GA. The group elicitation method for participatory design and usability testing. Interactions 1997;4:27–33.
23. Ewing G, Freer Y, Logie R, et al. Role and experience determine decision support interface requirements in a neonatal intensive care environment. J Biomed Inform. 2003;36:240–9.
24. Girgensohn A, Shipman FM. Supporting knowledge acquisition by end users: tools and representations. In: 1992. pp 340–8.
25. Jain H, Vitharana P, Zahedi F. An assessment model for requirements identification in component-based software development. The Database for Advances in Information Systems. 2003;34: 48–63.
26. Zaff BS, McNeese MD, Snyder DE. Capturing multiple perspectives: a user-centered approach to knowledge and design acquisition. Knowledge Acquisition. 1993;5:79–116.
27. Preece J. Interaction design: beyond human-computer interaction. Hoboken (NJ): John Wiley & Sons; 2002.
28. Noy NF, Crubezy M, Fergerson RW, Knublauch H, Tu SW, Vendetti J, et al. Protege-2000: an open-source ontology-development and knowledge-acquisition environment. Proc AMIA Annu Symp. 2003;953.
29. Musen MA. Dimensions of knowledge sharing and reuse. Comput Biomed Res. 1992;25:435–67.
30. Nwana HS, Bench-Capon TJM. Domain-driven knowledge modelling for knowledge acquisition. Knowledge Acquisition. 1994;6:243–70.
31. Wood LE. Semi-structured interviewing for user-centered design. Interactions. 1997;Mar/Apr:48–61.
32. Cooper A, Reimann R. About Face 2.0: The essentials of interaction design. Indianapolis: Wiley Publishing; 2003.
33. Patel VL, Glaser R, Arocha JF. Cognition and expertise: acquisition of medical competence. Clin Invest Med. 2000;23:256–60.
34. Patel VL, Arocha JF, Kaufman DR. A primer on aspects of cognition for medical informatics. J Am Med Inform Assoc. 2001; 8:324–43.