Evaluation of a pictograph enhancement system for patient instruction: a recall study

Qing Zeng-Treitler1, Seneca Perri1,2, Carlos Nakamura1, Jinqiu Kuang1, Brent Hill1,2, Duy Duc An Bui1, Gregory J Stoddard3, Bruce E Bray1

Author Affiliations

Received 7 September 2013
Revised 21 April 2014
Accepted 27 May 2014
Published Online First 12 June 2014

Abstract

Objective We developed a novel computer application called Glyph that automatically converts text to sets of illustrations using natural language processing and computer graphics techniques to provide high quality pictographs for health communication. In this study, we evaluated the ability of the Glyph system to illustrate a set of actual patient instructions, and tested patient recall of the original and Glyph illustrated instructions.

Methods We used Glyph to illustrate 49 patient instructions representing 10 different discharge templates from the University of Utah Cardiology Service. 84 participants were recruited through convenience sampling. To test the recall of illustrated versus non-illustrated instructions, participants were asked to review and then recall a set questionnaires that contained five pictograph-enhanced and five non-pictograph-enhanced items.

Results The mean score without pictographs was 0.47 (SD 0.23), or 47% recall. With pictographs, this mean score increased to 0.52 (SD 0.22), or 52% recall. In a multivariable mixed effects linear regression model, this 0.05 mean increase was statistically significant (95% CI 0.03 to 0.06, p<0.001).

Discussion In our study, the presence of Glyph pictographs improved discharge instruction recall (p<0.001). Education, age, and English as first language were associated with better instruction recall and transcription.

Conclusions Automated illustration is a novel approach to improve the comprehension and recall of discharge instructions. Our results showed a statistically significant increase in recall with automated illustrations. Subjects with no-colleague education and younger subjects appeared to benefit more from the illustrations than others.

Background and significance

‘A picture is worth a thousand words’ is a well-known saying. Congruently, research in health communication has shown that pictographs can effectively improve patient comprehension and recall of instructions. In particular, it has been suggested that pictographs may improve the recall of conceptual knowledge and problem solving information for individuals who have little expertise in a particular subject.1–3 As such, pictographs have been developed for health communication. Kools et al showed that pictures significantly improved patient recall and performance when used to enhance text instructions for two asthma devices (inhaler chamber and peak flow meter).4 Morrow et al reported that patients answered questions quicker and more accurately when icons were added to medication instructions.5 Kripalani et al found that the use of illustrated medication schedules led to positive outcomes.6

While pictograph enhancement can improve health communication, pictograph development is typically a costly and time-consuming task. In addition, illustrations do not always lead to better comprehension and recall.7 Hwang et al showed that adding icons to medication instructions did not result in better patient comprehension.8 Mayer and Gallini reported that while particular categories of illustrations (parts and steps) improved recall performance of conceptual information, non-conceptual information was not improved.9 Indeed, the creation of useful pictographs requires graphic design skills and an understanding of patient needs and clinical context. Typically, illustrations also need to be tested and refined iteratively.

Objective

In order to provide high quality pictographs for health communication, we developed a novel computer application (called Glyph) that automatically converts text to
sets of illustrations using natural language processing and computer graphics techniques. In this study, we evaluated the ability of the Glyph system to illustrate a set of actual patient instructions. We then tested participant recall of the original and Glyph illustrated instructions.

Materials and methods

We developed an automated illustration system Glyph, which employs natural language processing and computer graphics techniques. In the Glyph system, data flows through a sequence of processing modules. The five main processing steps are: pre-processing, annotation, post-processing, image composition, and image rendering. The pre-processing step accommodates different text formats. The annotation step makes use of a terminology database and a medication database to recognize textual entities such as instruction concept, medication, weight, negation term, and frequency. In the image composition step, the system determines how individual images should be combined using a rule base. The rule base was developed based on a linguistic analysis of a large set of University of Utah Cardiology Service's discharge instructions (n=2000). Finally, in the image rendering step, the system generates an illustration for the text using pictures from a pictograph library with more than 800 pictographs. These pictographs were developed by a professional graphic designer in our team.

The input of Glyph is free text medical instruction and the output is a set of images. Figure 1 shows an example of how a discharge instruction is processed and illustrated step-by-step. The Glyph system was tested iteratively using batches of randomly selected instructions with 50–100 per batch. The development team reviewed the illustrations and provided feedbacks. According to the feedback, the system was refined and retested on the same batch before proceeding to the next batch and iteration.

Pictograph generation and content review

We used Glyph to illustrate 49 patient instructions contained in 10 different discharge templates (e.g., congestive heart failure, ablation, thoracotomy) from the University of Utah Cardiovascular Service. Each of the 49 instructions was broken down into unique semantic units, or unique pieces of instruction for the patient to understand or act on, resulting in 160 final units. For example, the instruction 'Weigh yourself every morning, on the same scale. Wear the same clothes each time' was divided into four semantic units: (1) weigh yourself; (2) every morning; (3) on the same scale; and (4) wear the same clothes each time. In the case of conjunctive conditional sentences, the conjunction was contained within the semantic unit. For example, the instruction, ‘Call your doctor if you experience fainting, dizziness, or a racing heart rate’, was divided into four units: (1) call your doctor; (2) if you experience fainting; (3) if you experience dizziness; and (4) if you experience a racing heart rate. The breakdown of semantic units was done to facilitate precision of the annotation process and coding. This process was manually conducted by one of the researchers. While Glyph also parses the text syntactically and semantically, this manual annotation of semantic units is necessary because Glyph’s processing is driven by the illustration task. For instance, the phrase ‘avoid driving’ is regarded as one semantic unit, as it is a single instruction, but Glyph parsed the phrase into two parts: a negation for ‘avoid’, and ‘driving’ for illustration.

A nurse, a graphic designer, and a software engineer were asked to review the Glyph-created pictographs in terms of completeness and accuracy. Among the 160 semantic units, the reviewers determined 66.2% to be correctly represented, 20.2% were not represented, and 13.5% were incorrectly represented. Seven instructions containing the 32 unrepresented or incorrectly represented semantic units were removed from further evaluation, for a final inclusion of 42 discharge
instructions for use in the study. Figure 2 is an example of a correctly represented Glyph illustration.

![Figure 2](image)

**Figure 2**
A correctly represented Glyph illustration of the instruction ‘Caffeine can make your heart beat irregularly, and alcohol can make your body hold on to extra water’.

An overview of the study design is illustrated in figure 3.

![Figure 3](image)

**Figure 3**
Study design.

**Recall testing: sample and setting**
Recall testing was performed with non-clinical participants through convenience sampling. Inclusion criteria included age ≥21 years and able to speak, read, and write in English. Exclusion criteria included visual, cognitive, language, or other impairments that would prevent full participation and previous or current work as a physician, nurse, pharmacist, or other clinical position using patient discharge instructions. Eighty-four participants were recruited (see table 1).

![Table 1](image)

**Table 1**
Study participant demographics (n=84)

**Discharge instruction study forms**
To test the recall of illustrated versus non-illustrated instructions using participating subjects, a set of discharge instruction study forms was created using the 42 previously described instructions. Each study form contained 10 discharge instructions, which were randomized at three levels: (1) random selection from the set of 42 discharge instructions; (2) presence of pictograph enhancement; and (3) the sequence in which the instructions were presented on the study form. Appendix A available online displays an example study form.

**Subject enrollment and procedure**
During the study enrollment process, recruiters described the study inclusion and exclusion criteria to potential subjects, and explained each step of the study. This included informing participants that they would be given a set of discharge instructions to review for 10 min, asked to fill out a demographic form, and then recall and write down all of the instructions they could remember on a blank response form provided to them. Individuals agreeing to participate were given a written consent form that additionally explained the procedure. Hence, while the stepwise procedure of the study was fully explained, intent of the timing to fill out the demographic form to be used as a distraction technique was not disclosed.
After consenting to participate, subjects were given a set of cardiology discharge instructions that contained five pictograph-enhanced and five non-pictograph-enhanced items. Participants were monitored and given up to 10 min to read the 10 instructions; some participants required less time. On completion of the review, each was asked to fill out a demographic survey to fulfill the intent of the distraction technique. Following survey completion, they were asked to recall the 10 earlier instructions and given up to 10 min to write them down.

**Scoring**

The participant-recalled instructions were transcribed. One point was given for each semantic unit recalled in each instruction. Reviewers approximated the thoroughness represented in each semantic unit to inform the scoring of a point. Each semantic unit was scored as a whole unit; therefore, conjunctions or other grammatical units were not specifically counted. The number of semantic units in each instruction differed; therefore, the final rating for each instruction was the calculated proportion of semantic units recalled. Scores on individual questions ranged from 0 to 1. An example of the scoring scale is shown in Table 2.

Three investigators scored the recall instructions for the first 27 study subjects, assigning a score of proportion of semantic units recalled. The inter-rater reliability, as measured with the interclass correlation coefficient (ICC) was very high (ICC=0.94, 95% CI 0.93 to 0.95). For these first 27 study subjects, disagreements were resolved through consensus to arrive at the final score for analysis. Given that the high ICC revealed that raters were scoring similarly, the remaining 57 study subjects were divided equally among the three raters, with only one rater per study participant.

**Statistical methods**

Each participant was given a discharge instruction study form containing n=5 instructions with a pictograph, and n=5 instructions without a pictograph. The score assigned to each instruction was the proportion of semantic units in the instruction recalled correctly. The formula for the SD assumes that all instructions are independent, as if they are from different participants. To compute a valid SD that could be reported, the five scores from the five instructions without pictographs were converted to an average (mean) for each participant, and similarly for the five instructions with a pictograph. This reduced the data to one score per participant without pictographs and one score with pictographs. From this summary measure, a SD was computed, and reported to describe the variability between participants.

For all other statistical analyses, all 10 scores per participant were kept in the data. Three analyses were performed: (1) primary analysis: test if pictographs improve instruction recall; (2) secondary analysis: test for demographic variable influence on instruction recall; and (3) secondary analysis: investigate if demographic variables modify the effect of pictographs.

For analysis (1), to test for a pictograph effect on instruction recall, a mixed effects linear regression model was fitted. This model accounted for the lack of independence in the data due to having 10 observations (10 instructions) for each participant. The same instruction was not used by the same participant both with and without a pictograph, so a paired analysis at the specific instruction level was not possible. Instead, each participant was given five instructions with a pictograph and five different instructions without a pictograph. So, for each participant, there were 10 values of the outcome variable instruction recall, measured as the proportion of semantic units recalled, and 10 values of the predictor variable pictograph (five scores of ‘1 pictograph’ and five scores of ‘0 pictograph’). With 10 observations clustered within each participant, the model provided a ‘within participant’ analysis at the participant level, with each participant compared directly to him/herself in a pictograph and no-pictograph condition. Since participants acted as their own control, confounding was eliminated in this comparison. That is, age, education, and so on, including unmeasured variables, remained constant within the same participant when recalling instructions with and without pictographs. So, these variables could not alter the regression coefficient for pictograph, even if they were added to the model, which was confirmed with the analysis (2) model described next.

For analysis (2), to test for demographic variable influence on instruction recall, demographic variables were added to the analysis (1) mixed effects linear regression model. Thinking of this model as a multilevel model, the individual instructions were at level 1, the lower level. The pictograph variable could vary at level 1, and so was a level 1 predictor variable. The participant was at level 2, the higher level, as were the demographic variables. These were: education (high school or less vs some college), age (≥40 years vs ≤39 years), first language (English vs other language), gender, and race (Caucasian vs other race). At level 2, the demographic variables remained constant for each specific participant, but they varied between participants. So, the demographic variable comparisons were a ‘between participant’ comparison. Only eight subjects (9.5%) were of Hispanic ethnicity, and 22% of subjects had missing data for this variable, so the Hispanic variable was not used in this model. The referent group for each variable was selected to be the category with the greatest frequency, to provide the most reliable model.

For analysis (3), to investigate if demographic variables modify the effect of pictographs, a subgroup analysis was done. This approach was used instead of adding demographic-pictograph interaction terms to the analysis (2) model. If interaction terms were added to the analysis (2) model, overfitting would have been introduced, where overfitting is the situation of unreliable associations being observed due to having too many predictor terms in the model for the given number of participants. In these subgroup analyses, the analysis (1) model was fitted to a specific subgroup defined by one value of a single demographic predictor variable. For example, one model was fitted with pictograph as the predictor variable in the subgroup of participants whose first language was English, and a second model was fitted in the subgroup of participants whose first language was not English. These subgroup analyses were done with one demographic variable at a time, since using two or more demographic variables simultaneously would have produced subgroup sample sizes too small to provide reliable estimates of the effect of pictographs.

**Results**
Mean difference in scores
The mean score without pictographs was 0.47 (SD 0.23), or 47% recall. With pictographs, this mean score increased to 0.52 (SD 0.22), or 52% recall. In a univariable mixed effects linear regression model, this 0.05 mean increase was statistically significant (95% CI 0.03 to 0.06, p<0.001).

Participants acted as their own control, so the univariable model was sufficient to control for confounding. To see what other factors were associated with recall, a multivariable mixed effects linear regression model was computed (table 3). Whereas pictographs improved recall by 0.05, or an absolute 5% on average, a first language other than English decreased it by 16%, having no college education decreased it by 16%, and being ≥40 years old decreased it by 11%. Race and gender were not significant predictors.

Figure 4 illustrates the mean improvement for illustrated versus not illustrated instructions, in subgroups from univariable mixed effects linear regression models (with the pictograph as the only predictor variable and the model limited to the subgroup of study subjects).

On further examination of the pictographs’ impact in different subpopulations, we found that some populations benefit from illustrations more than do others (figure 3). As previously described, the mean increase is 0.05 overall (i.e., a relative 11% improvement). The largest increase we saw with the subpopulation is 0.2 (in the no-college group), which is four times the overall increase. In contrast, the subpopulation of age ≥40 years saw a decrease of 0.05. Potential reasons for the subpopulation differences are discussed below.

Discussion
To the best of our knowledge, this is the first formal evaluation of automated pictograph creation for health content. A number of prior studies have shown that illustrations could improve patient comprehension and recall of clinical instructions. However, not all illustrations are effective in improving recall and comprehension. Thus, evaluating Glyph illustrations is a highly relevant endeavor to improve discharge instructions.

In our study, the presence of Glyph pictographs improved discharge instruction recall (p<0.001). Education, age and English as first language were associated with better instruction recall and transcription. Non-native English speakers had lower scores overall, which may be in part due to the requirement of the exercise to write responses, possibly reflecting the challenge of writing in English rather than failure to recall instructions.

The absolute amount of improvement of 0.05 (or relative 11% improvement) from automated illustration is smaller when compared with the 0.11 absolute improvement in our prior study (or relative 25% improvement) using manual illustrations. The differences, however, may not be attributable to the automation of illustration. Our prior study had a much smaller sample size (n=13) and a much larger number of instructions for each participant to recall (n=34). The participants in the prior study were also presented the illustrated and non-illustrated instructions in two different sets, while in this study, the sets were mixed.

Participants with both college and no-college education recalled more illustrated than non-illustrated instructions. Similarly, participants who spoke English as either their primary or secondary language recalled more illustrated than non-illustrated instructions. On the other hand, younger subjects as a group (≤40 years old)
appeared to have benefited from illustrations while the older subpopulation (≥40 years old) did not. We hypothesize that this may be due to the fact that younger generations are more accustomed to visual communication methods, such as icons employed by graphic interfaces of computer applications. The fact that the age ≥40 group recalled slightly fewer discharge instructions when enhanced with pictographs was surprising to the researchers. This group of participants may have been historically less accustomed to visual communication methods, and the added material may have been confusing or perceived as extraneous information to assimilate. In future studies we will further examine this phenomenon in detail.

The no-college population showed a fourfold increase above the overall recall increase from illustrated instructions. This is consistent with literature evidence that illustration is particularly beneficial to the low health literacy population. Along the same line of reasoning, non-native-English speakers should have benefited more than the native speakers, though we did not see such an effect. We believe this is partially due to the requirement of written recall. A few participants commented to us that they remember the meaning of the instruction, but could not recall the specific English words. This is an issue we will address in future studies.

While not all instructions are illustrateable, and quality of illustration is a subjective measure, our results further suggest that partial and imperfect illustration could still improve participant recall. Varying completeness and quality of illustrations (determined by human review) still significantly improved recall. Overall, our study showed that automated illustration is feasible and is safe with adequate human review in place.

Limitations

There were several limitations of this study. First, the instructions used were not specific to the participants. We must further test the safety of the illustrations before proceeding to a clinical trial, which will involve patient-specific instructions in actual discharge settings. Second, a small number of the participants did not recall any of the instructions regardless of illustration, or recalled every instruction regardless of illustration. Each of these may have related to individual aptitude, knowledge, and life experience rather than presence of illustration. Further, while the illustrations were randomized on three levels, the number of semantic units in each instruction varied. Therefore, while each subject was presented with an equal number of illustrated and original instructions, the semantic content was not necessarily equal across all instructions. Finally, different semantic units of the discharge instructions possessed varying clinical significance. For example, some instructions contained explanations for recommended actions, which is less clinically significant than simply knowing the recommended action. In this study, we scored each semantic unit equally. While this method enabled a more nuanced approach to scoring than a simple count of recalled words could provide, the scoring method presented a limitation that will be addressed in future studies.

Recall is often used as a proxy for comprehension and is directly impacted by recognition. In other words, some illustrations may not have been recognized or misrecognized for cultural or other personal experience reasons, which is a confounding factor in this study. We conducted a separate study on recognition of a large set of pictographs, which will inform future continuation of our pictograph enhancement inquiry. We are currently utilizing the results of this and multiple additional studies to refine the Glyph system in preparation for a future clinical trial.

Conclusion

We conducted an evaluation of the automated illustration of discharge instructions using 84 non-patient volunteers. Automated illustration is a novel approach to improve the comprehension and recall of discharge instructions. Our results showed a statistically significant increase in recall with automated illustrations, with subjects having no college education and under the age of 40 appearing to benefit more from the illustrations than others. Future work is needed to improve the recall, comprehension, and presentation of illustrated patient instructions for all demographic and cultural groups.

Footnotes

Contributors OZ-T oversaw study design, analysis, and manuscript preparation. SP led manuscript preparation and participated in data collection, coding, and analysis. CN led study design. JK and BH coordinated data collection and coding. DDAB implemented the Glyph system and participated in data analysis. GJS performed statistical analysis. BEB participated in study design, analysis, and manuscript preparation.

Funding This work was supported by the National Institute of Health, grant # R01 LM009966-03.

Competing interests None.

Ethics approval University of Utah IRB.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The data of the study are available on request.

References


Related Article

Highlights:
Focusing on the patient: mHealth, social media, electronic health records, and decision support systems

Lucila Ohno-Machado


[Extract] [Full text] [PDF]