A Multidisciplinary Approach to Designing and Evaluating Electronic Medical Record Portal Messages that Support Patient Self-Care

Daniel Morrow[a], **Mark Hasegawa-Johnson**[b], **Thomas Huang**[b], **William Schuh**[c], **Renato Ferreira Leitão Azevedo**[a], **Kuangxiao Gu**[b], **Yang Zhang**[b], **Bidisha Roy**[a], and **Rocio Garcia-Retamero**[d]

[a]University of Illinois at Urbana-Champaign, Department of Educational Psychology, Champaign, IL, United States

[b]University of Illinois at Urbana-Champaign, Department of Electrical and Computer Engineering, Urbana, IL, United States

[c]Carle Foundation Hospital, Urbana, IL, United States

[d]University of Granada, Department of Experimental Psychology, Granada, Spain

Abstract

We describe a project intended to improve the use of Electronic Medical Record (EMR) patient portal information by older adults with diverse numeracy and literacy abilities, so that portals can better support patient-centered care. Patient portals are intended to bridge patients and providers by ensuring patients have continuous access to their health information and services. However, they are underutilized, especially by older adults with low health literacy, because they often function more as information repositories than as tools to engage patients. We outline an interdisciplinary approach to designing and evaluating portal-based messages that convey clinical test results so as to support patient-centered care. We first describe a theory-based framework for designing effective messages for patients. This involves analyzing shortcomings of the standard portal message format (presenting numerical test results with little context to guide comprehension) and developing verbally, graphically, video- and computer agent-based formats that enhance context. The framework encompasses theories from cognitive and behavioral science (health literacy, fuzzy trace memory, behavior change) as well as computational/engineering approaches (e.g., image and speech processing models). We then describe an approach to evaluating whether the formats improve comprehension of and responses to the messages about test results, focusing on our methods. The approach combines quantitative (e.g., response accuracy, Likert scale responses) and qualitative (interview) measures, as well as experimental and individual difference methods in order to investigate which formats are more effective, and

Correspondence concerning this article should be sent to Dan Morrow, dgm@illinois.edu, Department of Educational Psychology, University of Illinois at Urbana-Champaign, Education Building, 1310 S. 6th St. Champaign, IL 618203.

whether some formats benefit some types of patients more than others. We also report the results of two pilot studies conducted as part of developing the message formats.

Graphical Abstract
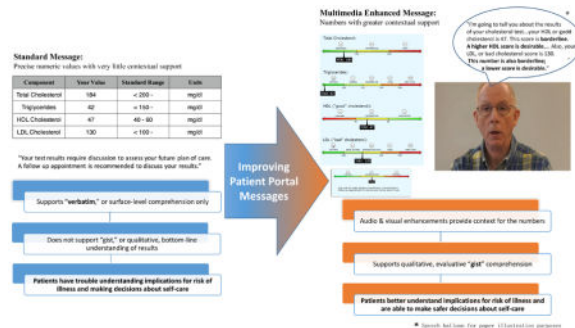
Keywords

Cognition; Learning; Patient Portal; Electronic Medical Record; Aging; Computer Agent

1. Introduction

Self-managing health often places heavy demands on patients, including understanding the risks and benefits of different treatments, test results, and lifestyle choices [1]. These tasks often require numeracy, or the ability to make sense of numbers [2], which is an important component of the health literacy needed to navigate health care systems [3]. Unfortunately, patients often struggle to grasp numerical concepts essential for understanding and communicating health-relevant information [2, 4, 1]. Patients with lower numeracy are more likely to misunderstand test results, undermining health decisions, behaviors, and outcomes [5]. Older adults, the most frequent consumers of health information because they often self-manage chronic illness, tend to have lower numeracy and literacy skills [6, 7, 8].

Patients increasingly receive or find health information through health information technology (HIT). For example, patient portals to Electronic Medical Record (EMR) systems are intended to bridge patients and providers by ensuring that patients have continuous access to their health information and services. Portals have expanded the delivery of numeric information to patients because of federal requirements for 'meaningful use' of EMRs [9], and there is some evidence that they improve patient outcomes [10]. However, they are underutilized [11], especially by older adults with lower health literacy [12, 13]. Portal use among older adults is predicted by limited numeracy, even when controlling for cognitive ability [14, 15].

Portals are also underutilized because they function more as information repositories than as tools to engage patients in self-care. Patients may also misunderstand portal-based information [14, 15]. For example, clinicians at our partner health organization have expressed concern that their patients can become confused about test results delivered through their EMR portal without prior discussion with their provider, and need to call to get

help understanding the results. In their view, the portal may sometimes increase rather than reduce clinical workload and impact the quality of care delivered to their patients. This concern is echoed in the literature on patient/provider communication. While there is general agreement that face-to-face communication between providers and patients is a key to effective patient education, there is also recognition of diminishing time for this communication, so that critical information is less likely to be presented. In addition, information that is presented may be misunderstood by patients [16]. For this and other reasons, patients often forget important information presented by their provider [17]. In short, while HIT has the potential to support patient/provider collaboration despite diminished opportunity for face-to-face communication, this potential has yet to be realized [9, 18].

Our project is intended to improve use of portal information by older adults with diverse numeracy and literacy abilities, so that portals can better support patient-centered care. The project has three phases. 1) Develop portal message formats for clinical test results and evaluate them in a simulated patient portal environment. 2) Develop a tool for clinicians to annotate test results for their patients, facilitating collaborative use of portals by providers and patients. 3) Evaluate use of the tool and analyze patient responses to the clinician-constructed messages in an actual portal system.

In this paper, we focus on Phase 1 goals and describe a multidisciplinary approach to designing and evaluating portal-based messages that convey clinical test results so as to support patient-centered care. We first outline a theory-based framework for designing effective message formats that convey numerical information to diverse patients through portals (Figure 1). The framework encompasses theories from cognitive and behavioral science as well as computational/engineering approaches. This cognitive informatics approach [19] helps explain why patient portals often fail to support patient/provider collaboration and guides our approach to designing and evaluating portal messages. We then describe how we evaluate whether the formats improve comprehension of messages about test results by older adults with diverse abilities. The approach combines quantitative (e.g., response accuracy, Likert scale responses) and qualitative (interview) measures, as well as experimental and individual difference methods in order to investigate not only which formats are more effective, but whether some formats benefit some types of patients more than others. In the present paper we describe the rationale for and methods involved in our approach, as well as results from two pilot studies that are part of developing the portal messages.

1.1 Designing Actionable Portal Messages

1.1.1 Theoretical framework—Numeric information such as test results is often presented in patient portals as a table of numbers with little context (e.g., information about the scale for the scores, or relation of the scores to risk for illness [15] (see Figure 2A). Therefore, patients, especially older adults with lower literacy and numeracy abilities, may struggle to make sense of their test results [21, 4]. Designing formats that help patients understand and use numeric information requires explaining how patients understand and respond to this kind of information [1]. The framework in Figure 1 integrates fuzzy trace

memory, health literacy, text processing, and behavior change theories in order to identify processes involved in understanding and responding to portal messages, as well as the patient abilities/resources that influence these processes.

According to fuzzy trace theory [4], people understand and remember information at multiple levels, as shown in the Comprehension box in Figure 1. In particular, this theory distinguishes detailed quantitative analysis using verbatim representations from intuitive reasoning based on simple gist representations of information, extending this distinction beyond linguistic information to numbers, pictures, graphs, events, and other forms of information [22].

The verbatim level captures the literal facts or "surface form" of information, preserving information about precise numeric values. However, making sense of numbers in order to act on them often requires interpreting the information in terms of goals and knowledge to also create a gist-based representation, which captures its meaning or interpretation (based on a person's age, education, and illness experience, among other factors known to affect meaning see Figure 1; also [21]). In other words, gist representations capture the bottom-line implications of the numbers for health, and therefore are less precise than verbatim representations. They are organized around more qualitative dimensions such as categorical or ordinal relations that are often affective and evaluative (good/bad; higher/lower).

People encode verbatim representations as well as multiple gist representations of the same information. When presented with various numbers (e.g., test results), people encode verbatim representations of the numbers and gist representations that capture the order of magnitudes of the different scores (e.g., whether they are increasing or decreasing and which magnitudes seem large or small; [22]). Accurate gist understanding of test results includes implications of the results for risk of illness, and therefore can support decision making about self-care. For instance, gist representations of cholesterol test results may capture ordinal values of risk for heart disease, such as lower/borderline/higher.

People generally prefer to operate on the fuzziest or least precise representation that they can to accomplish a task, such as making a decision [21, 22]. They begin with the simplest distinctions (e.g., categorical) and then move to more precise representations (e.g., ordinal and interval) as needed. Therefore, categorical or ordinal gist, or gist combined with verbatim representations, may be most effective for understanding implications of test results for risk, depending on task requirements [23, 21].

Gist is similar to the concept of situation model in text comprehension theory. Situation models represent the situations described by the text and are created by integrating text information with knowledge [24]. Both gist and situation model concepts emphasize that comprehension depends on knowledge and help explain how comprehension supports decision making and action.

The process-knowledge model of health literacy [25, 20] draws on theories of text comprehension to identify processes involved in understanding health information at verbatim and gist levels. These include word recognition and integrating the concepts associated with words into propositions. Understanding numeric information requires

similar processes [2, 4]. For example, understanding cholesterol scores involves identifying numeric values (verbatim representation) and mapping the values to categorical or ordinal risk categories (gist representation). These gist representations for the component scores may be integrated into an overall or global gist representation of risk associated with the message. Integrating the scores can be challenging, in part because they have different scales. The high-density lipoprotein (HDL) scale may confuse patients because it has an opposite polarity from the other scales (higher values indicate lower rather than higher risk [26]). Because gist comprehension depends on patients' goals and illness knowledge, information from the scores may be integrated with knowledge and plans, similar to the situation model constructed from text [20]. The challenge of understanding cholesterol scores is reflected in the finding that adults generally know little about their current or target cholesterol scores [27].

As seen in the left part of Figure 1, these processes in turn depend on several types of cognitive abilities or resources, as well as broader patient characteristics such as age and education. These general cognitive resources include processing capacity (e.g., working memory) and knowledge (of language and health-related concepts), as well as non-cognitive resources such as affect, which is especially likely to influence gist comprehension of risk for illness because gist is often organized in terms of threat and other evaluative dimensions [20, 28].

Aging influences risk comprehension processes because of age-related changes in the cognitive and affective resources needed for comprehension [6]. Processing capacity tends to decline with age, while general knowledge (e.g., about language) and domain-specific knowledge (e.g., about health) tend to grow with age-related increases in experience. These age-related resource limits and gains interact to influence health literacy [25] and comprehension of health information [29], such that high levels of knowledge can offset processing capacity limits on comprehension. Aging is also accompanied by increasing focus on affect and emotion, which influences understanding and decision making about health information [30, 20].

While the left side of the framework in Figure 1 identifies patient abilities that influence comprehension at verbatim and gist levels, the right side shows that perception of the risk associated with test results is influenced by comprehension, and therefore is likely to depend on the cognitive and affective resources that influence gist comprehension. Risk perception is also shaped by other factors such as beliefs about illness (e.g., How susceptible and vulnerable to the illness am I? [31, 32]).

According to theories of behavior change (e.g., the theory of reasoned action/planned behavior [33], the health-action process approach [34], common-sense model, [32]), risk perception in turn shapes attitudes toward actions that may mitigate perceived risk, such as taking medication or following exercise and diet recommendations. Behavioral attitudes are influenced by other factors as well, such as beliefs about whether the actions influence factors believed to be linked to the illness [31]. For example, patients may understand that their cholesterol scores indicate high risk for cardiovascular illness, but do not believe that exercise or diet reduce this risk. Finally, behavioral attitudes predict intention to perform the

behaviors, which predict performance of the behaviors (for review [35]). In short, the framework suggests comprehension of health information is an important influence on health-related decision making and behavior, so that it is critical to deliver health information in ways that support comprehension. This is especially crucial for patients with low health literacy and numeracy.

1.2 Improving comprehension and use of portal-based numeric information

The literature on risk perception suggests that people better extract gist that supports decision making and action if numbers are presented in a context that supports comprehension processes (e.g., mapping numbers to risk categories that guide appropriate affective/evaluative responses [23, 28]). Enriched contexts may also improve the ability to use verbatim numeric information when making decisions [23]. Our framework suggests that enriched context will especially benefit older adults with fewer cognitive resources by reducing comprehension demands on processing capacity and enabling use of prior knowledge. For example, research on age differences in memory for numeric information suggests that relying on gist may especially benefit older adults with lower health literacy because their verbatim memories are less robust [28, 36]. In addition, age differences in choice quality are reduced when decisions are based on affect or bottom-line valence [37]. However, we note that these claims have not been evaluated in the context of patient portals. Accurate gist understanding in turn may help calibrate older adults' risk perception based on information such as clinical test results, and therefore support their attitudes toward and intention to perform behaviors that address this risk.

Traditionally, patients (especially with lower health literacy) rely on providers to help them understand information such as their clinical test results. In face-to-face communication, providers can use both verbal (e.g., words that emphasize key concepts) and nonverbal cues (tone of voice, facial expressions) to indicate (1) what information is most important, and (2) appropriate affective and evaluative reactions to this information, which help patients understand the gist of their test results [38, 4]. Providers' nonverbal as well as verbal behaviors have been found to predict patient satisfaction and other outcomes [39, 40]. Unfortunately, providers do not consistently use these strategies because of limited communication time and inadequate training, so patients leave office visits without enough information or remembering little of what was presented [17]. Portals exacerbate this problem of limited provider support for patient comprehension because portal-based test results are usually delivered as a set of numbers with minimal context (e.g., about the scale of test score values), often in a table format (for example, see Figure 2A).

We are investigating whether formats designed to enhance the context of standard portal messages about cholesterol and diabetes screening test results improve gist comprehension and risk perception, and boost attitudes toward and intention to perform self-care behaviors. More specifically, we investigate verbally, graphically, and video-enhanced formats that provide increasing levels of support for gist-level understanding of the risk indicated by test results. They may also increase motivation to make decisions related to self-care behaviors that mitigate risk, although the display features that improve comprehension are not always the most effective for motivating actions [41].

The verbally enhanced format provides verbal cues often used in face-to-face communication to help patients understand the risk associated with the test scores (see Figure 2B). For example, labels for evaluative categories (low, borderline, or high risk) are added to the message to provide context for interpreting the specific scores, which should facilitate the process of mapping the score to the appropriate region of risk. Such labels promote emotional processing of the quantitative information [23] and may especially help older adults, who tend to have high levels of verbal ability and focus on emotional meaning [30]. In addition, categorical information is often processed with little effort, which may benefit older adults who experience declines in processing capacity [23].

The graphically enhanced condition provides graphical as well as verbal context for gist comprehension (see Figure 2C). The test scores are embedded in graphic representations of the scale for each score, with color coding reinforcing the verbal labels of each region of risk. The format emphasizes the ordering of the regions from lower to higher risk (or higher to lower risk for HDL), which should support ordinal gist understanding [41, 23]. Such graphics support gist understanding of numbers that is organized around evaluative and affective dimensions [42, 43, 4], and have been found to improve patients' understanding of their health parameters, including cholesterol and blood pressure [26].

In the most enhanced message condition, the same graphic display (with the labeled test scores) is accompanied by a video of a physician discussing the results and their implications for risk (Figure 2D). The physician provides high-level commentary about the results, using nonverbal cues (prosody, facial expressions) and verbal cues (the same risk category labels as in the other enhanced formats) to signal information relevance and guide affective interpretation, as in ideal face-to-face communication.

Based on our framework and the reviewed literature, we expect that the more enhanced messages will be better understood and more acceptable to patients (meeting their informational and affective needs) by providing context that helps guide understanding of specific test values, particularly at the gist level. Adding verbal labels for risk categories to the standard messages should help older adults integrate numeric information into a gist representation [8], and especially benefit those with lower health literacy by reducing the need for high levels of numeracy and processing capacity (e.g., working memory) to develop gist representations [23]. Evaluative labels also help older adults interpret and integrate numbers in terms of their affective significance, so that they can develop affectively as well as cognitively organize gist representations that support health decisions [23]. In the graphically enhanced condition, embedding the labeled scores in the graphic number lines for each scale helps convey quantitative information more efficiently than verbal and numeric formats do [44], so that patients better integrate risk information across test scores into a gist representation. These graphic formats may especially benefit older adults with lower health literacy by making key gist features more cognitively accessible [23, 4]. Finally, the multimedia video format may be most effective because the verbal and nonverbal cues reinforce each other [45, 46]. Differences in performance associated with numeracy and literacy skills may also be reduced, by decreasing the need for numeracy-based skills and encouraging the use of affect to understand risk.

With this study design, we can explore trade-offs between the extent of message enhancement and gains in patient response. This is clinically important because if minimal enhancement substantially boosts comprehension, these messages may be most cost-effective for many patients and applications.

However, the video format would be difficult to implement in actual patient portals because different videos would be required for each patient and set of test results. Therefore, we are developing a computer-based agent (CA), or 'virtual provider' to deliver health information in portal environments. Ideally, the CA emulates best practices for face-to-face communication but is also generative, with capability of delivering a wide range of health information to diverse patients. In our portal format evaluation studies, this CA-enhanced format will be identical to the video condition, except the video is replaced by a CA being developed using the video as a template.

Our CA builds on progress in developing CAs for human-computer systems. CAs with realistic facial expressions, gestures, and other relational cues improve learning in tutoring systems compared to text-based systems [47, 48, 49], in part by engendering social responses. [8] While CAs are often used and evaluated in education settings, evaluation of applications in health care are less frequent, although Bickmore and colleagues have developed CAs that support a variety of patient goals and have helped diverse patients to follow self-care recommendations compared to text-based or lower fidelity interfaces ([50, 51, 52], also see [53]). CAs in EMR portals may especially help patients with lower numeracy and literacy because, as in face-to-face communication, they can provide commentary with nonverbal cues (prosody, facial expressions) and verbal cues (risk category labels) to guide gist comprehension.

2. Methodology for Evaluating Enhanced EMR Portal Message Formats

We are using a multi-method approach to evaluate EMR patient portal message formats. This approach combines quantitative and qualitative measures, as well as experimental and individual difference methods in order to investigate not only which formats are more effective, but whether some formats benefit some patients more than others. As part of developing the portal messages, pilot studies are conducted to help finalize the design of the message formats, leading to more formal scenario-based experimental studies that evaluate whether enhanced formats improve patient memory for and use of self-care information compared to the standard format. The formal studies first focus on the impact of video-enhanced format and then compare CA-enhanced to video-enhanced formats. In this section of the paper, we first describe the scenario-based method and measures that will be used in the formal experiments to evaluate the message formats, and then report results from two pilot studies. In later papers we will report results of the formal evaluation experiments.

2.1 Evaluation Study Scenarios

The study scenarios were developed in collaboration with two physicians from our partner health care organization. Each scenario contains a patient profile and a message describing cholesterol or HbA1c diabetes screening test results for that patient. Patient profiles are included because risk level depends on patient-specific risk factors (e.g., coronary artery

disease, hypertension, family history of heart disease) as well as the specific test results. Diabetes as well as cholesterol test results are evaluated because type 2 diabetes is a common age-related chronic illness that is often co-morbid with cholesterol-related illness. The diabetes messages describe a percent score on one test (plasma glucose concentration, A1C), while the cholesterol messages describe more complex patterns of scores on multiple tests (total cholesterol, triglycerides, high density lipoproteins (HDL), and low density lipoproteins (LDL)) that suggest lower, borderline, or higher risk for cardio-vascular illness. To help patients understand the overall risk associated with each message, the message ends with a summary of the overall risk for heart disease associated with the test scores. For the enhanced messages, the summary conveys the gist risk category.

2.2 Message Formats

The messages describing the test results in the scenarios were presented in one of the four formats.

2.2.1 Standard—Messages in the standard format condition are similar to the format used in our partner health organization's EMR portal. Test results are presented as a table of numbers, with minimal information about the scale of each score (see Figure 2A for an example of cholesterol test results; a similar format was used for diabetes screening results). The table is followed by a summary statement about the overall level of risk associated with the test results, which is modeled after practices in our partner health organization.

2.2.2 Verbally enhanced—This format adds verbal support for comprehension. The standard format is enhanced with verbal cues related to the risk associated with the test scores (Figure 2B). In addition to more information about the regions of risk associated with the scale for each score, labels for evaluative categories (low, borderline, high risk levels) are added to provide context for interpreting the specific numbers in relation to the appropriate risk category. The verbally enhanced summary statement labels the overall level of risk associated with the results.

2.2.3 Graphically enhanced—This condition provides graphical as well as verbal context for categorical or ordinal gist comprehension. The test scores are embedded within graphic representations of the scale for each of the four component cholesterol scores, with color coding and facial expression icons reinforcing the verbal labels of risk regions, which are ordered into a scale (Figure 2C). The summary statement is also a graphic scale with color-coded lower, borderline and higher risk regions, with the appropriate region marked to indicate the overall risk level associated with the test results.

2.2.4 Video enhanced—In this condition, the same graphics (with verbal labels and the test scores) are accompanied by the video of the physician providing commentary about the test results using nonverbal cues (prosody, facial expressions) and verbal cues as in ideal face-to-face communication (Figure 2D). The videos were recorded by a retired physician who used a script that not only indicated what to say, but which information should be emphasized during the discussion. To help link this verbal commentary with the relevant information in the graphic display, the corresponding part of the graphic loomed as each test

score was discussed by the physician, providing a dynamic attentional cue to help participants integrate information from the video and the graphic display. Message presentation time was equated across format conditions for each scenario, and corresponded to the duration of the physician video for that scenario.

2.3 Measures for Scenario-based Evaluation Studies

Selection of measures is guided by our framework. Measures of participant responses to the portal messages are included to evaluate the impact of message formats, and measures of participant individual differences are included to explore whether the impact of formats depends on older adults' abilities.

2.3.1 Responses to portal messages—We developed measures of verbatim and gist memory for the patient test results in the scenarios. Verbatim memory for the individual component scores in the message is probed by asking participants to recall the exact numeric value of the component scores from their cholesterol or diabetes screening test; for example, *"Kathleen's LDL score was = _____ (numerical score)"*.

Gist comprehension of risk associated with both the individual test scores (component-level gist) and with the overall message (global gist) are also probed. Component-level gist is measured by asking participants to remember the component scores at a gist level relevant to risk associated with the score. They are asked whether scores are "High/Good," "Borderline," or "Low/Bad" (ordinal risk [4]). Overall gist memory is measured by asking participants to consider all of the patient's results and indicate the overall level of risk. For example, "Considering Jennifer's cholesterol test results, her overall level of risk indicated by the set of results in the message was" (ordinal level gist: low/borderline/high [4]). Overall gist is measured both before and after the summary statement. For both verbatim and gist questions, participants are encouraged to answer as if they were the patient in each scenario.

We also measure participant responses to the risk associated with the described test results, which may reflect motivation to act so as to mitigate that risk. Perceived seriousness of the risk is measured by asking participants about the patient's likelihood of developing heart disease if nothing is done to reduce his/her cholesterol levels (9-point scale, 1=very unlikely; 9=very likely) [54]. Affective reactions are also measured by asking how participants feel when reading the test results. On 9-point scales, (1=not at all; 9 =very much), participants indicate to what extent they feel assured, calm, cheerful, happy, hopeful, relaxed, and relieved (positive adjectives). On 9-point scales (1=not at all, 9 =very much), they also indicate the extent to which they feel anxious, afraid, discouraged, disturbed, sad, troubled, and worried (negative adjectives) [54]. 3

Attitude towards taking medications is measured by asking participants to rate on a nine-point scale (1=not at all; 9=very much) how favorable they would feel about taking medications for lowering cholesterol if prescribed by the doctor. In addition, intention to perform self-care activities related to reducing risk of cardio-vascular disease (e.g., making follow-up appointments, following diet, exercise, medication recommendations) is measured. Participants rate the following (1=I have no intention of doing this; 9=I am certain that I would do this): If they were the patient in the scenario, (1) how likely would they be to

take medication to reduce cholesterol if prescribed by the doctor; (2) how likely would they be to change their diet if prescribed; and (3) how likely would they be to increase their level of exercise if prescribed (adapted from [54]). Although actual behavior will not be measured in the studies, intention to perform health behaviors has been found to predict actual behavior [35].

We also examine trust in message information and source by measuring participants' acceptance of and response to the risk associated with the message. Participants indicate the extent to which they consider the information conveyed in the message as useful, on a 9-point scale ranging from 1 (not at all useful) to 9 (very useful; adapted from [54]). They also indicate whether they would want more information than provided in the message, and if yes, they identify the source (e.g., the provider of the message, other health care providers, internet), and the reason for wanting more information (to clarify the message, learn more about the topic, verify appropriateness of the reported test results, which possibly indicates lack of trust in the provider) [55].

2.3.2 Individual difference measures—Selection of individual difference measures is guided by the process-knowledge model of health literacy [25] (see Figure 1). Health literacy is measured by the Short Test Of Functional Health Literacy for Adults (STOFHLA), which requires reading brief health care documents such as a consent form and completing Cloze items in which a word is deleted from the text so that participants choose the best of 4 options to complete the sentence [56]. Performance on this test predicts recall of self-care information among older adults [29]. Numeracy is measured by the Berlin Numeracy Test, a brief, valid, measure of statistical numeracy and risk literacy [57] and by the Subjective Numeracy Scale [42] because older adults often overestimate their numeracy ability, influencing their risk assessment [15].

The broader cognitive abilities that influence health literacy are also measured. Processing capacity is measured by the Letter and Pattern Comparison tests [58]. Performance on these mental processing speed tests decline with age and predicts health literacy [59] and recall of health information [29]. General knowledge, or verbal ability, is measured by the ETS-KFT Advanced Vocabulary test [60] and the Author Recognition Test [61]. Performance on these measures predicts performance on health literacy measures [25] and recall of health information [29]. Finally, demographic variables (e.g., age, education) are also measured.

2.4 Scenario Study Procedure

Older adults with diverse abilities will view the scenarios (equal number for cholesterol and diabetes test results that indicate lower, borderline, and higher risk for illness for hypothetical patients). For each scenario, they first read the patient profile, followed by the test result message for that patient. All messages for a participant are presented in one of the four message formats, depending on the condition the participant is randomly assigned to. After each message, participants respond to questions as if they are the scenario patient. Message memory, risk perception, affective responses, behavioral attitudes, behavioral intentions, and message satisfaction are probed.

2.5. Video Enhanced Message Pilot Study

To finalize development of the video-enhanced messages before conducting the formal evaluation studies, we conducted a pilot study to investigate whether the video messages themselves (i.e., presented without the graphic displays) were easy to understand and would prompt appropriate affective responses to the information [62]. It was important to validate the video messages because the physician serves as the template for developing the CA to be used in the CA-enhanced message format condition. We were most interested in whether older adults would be able to identify the level of risk associated with the test results, based solely on the physician's verbal and nonverbal cues in the video messages.

Both quantitative measures (the gist memory, affective response, and message satisfaction measures to be used in the formal scenario-based evaluation studies) and qualitative measures (participant interviews about videos) were included in the pilot study. Twelve older adults (average age 77 years, range 65–89 years, 67% females) who were native English speakers viewed 24 video messages that conveyed cholesterol and diabetes clinical tests results (equal numbers of cholesterol and diabetes messages describing test results indicating lower, borderline, and higher risk). After each message, participants pretended to be the patient and responded to questions that probed gist memory, affective reaction to the test results, and satisfaction with the message. Thus, the study also piloted the procedure and some of the measures to be used in the formal experiments. In addition, we evaluated participants' satisfaction with the messages by asking them whether the way the information was delivered matched the information provided *("I think the way the doctor conveyed the information matched what he said"*, 9-point scale: strongly disagree to strongly agree).

Table 1 shows that participants considered both the cholesterol and the diabetes messages to be very appropriate. Participant responses were highly consistent across the messages (see Table 1). Participants also considered the information delivered in both the cholesterol and diabetes video messages as very useful, and again, responses were consistent across messages.

Finally, participants were generally able to understand the gist of the video messages describing the cholesterol and diabetes test results [62]. Their affective responses to the messages were also appropriate to the level of risk associated with the test scores: As the risk increased, positive affect decreased and negative affect increased, as shown in Figure 3. These results suggest that participants developed a gist representation organized in terms of affective as well as cognitive dimensions, which may be important for motivating healthy behavior.

The findings from the pilot study helped validate the scenario-based evaluation method for the formal studies. Moreover, because the study included more video messages at each level of risk (lower, borderline and higher) than needed for the formal studies, we were able to select those messages that were most accurately understood and identified as most effective for the formal studies. In addition, the pilot validated the videos as the template for the CA-based format.

2.6. Developing CA-Enhanced Portal Messages

The first message evaluation experiment should not only provide evidence that enhanced formats improve memory for and responses to portal messages, but identify which enhanced formats are most effective: The video-enhanced format should be most effective because it provides the most elaborate context for interpreting the numeric test results. We will follow-up these results by comparing the CA-enhanced format to the video-enhanced format, using the same procedure and measures but replacing the video-recorded physician with the CA.

The CA is being developed from the video physician template. Our goal is a life-like CA that encourages patients to take a 'social stance' toward the agent, which should prime social and cognitive processes that support learning [48, 8]. Characteristics such as lip-synching with speech, realistic facial expressions, and appropriate eye gaze are important for giving CAs a lifelike appearance [63]. The CA is being developed in three phases (the first two concurrently): The CA's visual appearance is generated from a frontal face image of the physician; the audio (synthesized voice) of the CA is generated using a speech synthesizer; and the visual and audio components are then synchronized.

2.6.1 Developing the CA visual appearance—A 2D frontal face image of the physician in neutral expression is used to generate the CA's 3D appearance, using an algorithm that localize the feature points on the face and fit a 3D morphable generic model to the image (for details, [64]). This approach of reconstructing a 3D from a 2D facial image is very fast compared to creating a 3D scan and is highly versatile, making the algorithm suitable for large scale CA database generation and testing, and thus feasible for tailoring use of CAs for different goals, users and contexts.

Once generated, the 3D CA is animated to simulate a real person talking, so that the CA makes appropriate lip movements that match spoken phonemes, as well as facial expressions in order to match the input text with emotion markers. In our implementation, the CA is controlled by Facial Action Parameters (FAP). For example, the "stretch lip corner" FAP specifies how much the lip corner of the 3D CA is stretched to represent smiling or certain lip shapes during speech. Different FAPs are combined to control the whole face to match the CA appearance with the speech and emotion context.

The lip movements during speech are mainly affected by speech content and emotion. To integrate these two factors, we define a template for each lip movement under each expression separately. During the synthesis process, all lip movements are predefined and the best fit template selected. Finally, we construct a set of keyframes, where the CA has expressions and lip shapes matching the target viseme and emotion marker, by selecting the appropriate templates and interpolating between those keyframes to generate the talking head (see Figure 4).

2.6.2 Developing the CA synthetic voice—A physician from our partner health organization first recorded about eight hours of speech, including the scripts for the video messages for the message evaluation experiments as well as other text needed to develop the CA. The resulting audiovisual recordings of the physician's face were segmented using two automatic speech recognition toolkits: HTK [65] and Kaldi [66]. Therefore, the baseline

audio speech synthesizers were developed using first a unit selection concatenation of audiovisual segments from this recorded corpus [67].

The speech in the corpus was segmented into phone units and a synthetic utterance produced by selecting and concatenating these units that constitute the target utterances, maximizing continuity and naturalness. To capture emotional nuance, each sentence was assigned positive, neutral and negative labels, based on the valence of the medical test results in that sentence. To synthesize an utterance with a specific emotional valence, the synthesizer placed a preference on the phone units with the matched valence label.

To informally test intelligibility of the emotional variation, the same utterance, announcing a borderline result, was synthesized with the three different valences. Six naïve participants matched the valence to the utterance. Their accuracy was worse than chance, indicating they could not determine the valence of the utterances signaled exclusively by the origin of the audio clips (in a concatenative synthesizer), without corresponding changes to the utterance content. It is possible that the emotional nuance was overwhelmed by degraded naturalness of the utterance, or unit-selection was unable to capture emotional nuances given the limited corpus.

We next focused on pinpoint modification of F0 contours correlated with specific syntactic and semantic information in the text. Differential linear regressions were used to learn the difference between F0 (pitch) of any given word with versus without pitch accent, and texts were synthesized with pitch accents added in appropriate positions [68]. Next, this approach was generalized to learn differences in pitch contour between the speech of two a monotone and a dynamic speaker. The hidden F0 targets underlying observed F0 measurements in each syllable were estimated and learned using Xu's "Parallel Encoding of Target Approximation" (PENTA) model [69]. Mapping between PENTA targets of the monotone and dynamic speakers was learned and the speech of the monotone speaker was processed to have a pitch contour better matching that of the dynamic speaker. We are leveraging this approach to guide post-processing of utterances produced by the Vocalware speech synthesizer [70] for the CA to be used in the formal studies.

2.6.3 Synchronizing Video and Audio—Finally, the synthesized face animation and speech audio components are synchronized. The audio and video are aligned via keyframes, where each keyframe corresponds to one viseme. The input text (the CA messages) are first converted into visemes (phoneme for audio synthesis). For each viseme, a keyframe is generated with a visual template (defining the lip shape and expression) associated with it. For the video synthesis, the gaps between each keyframe are filled by interpolation to generate a smooth lip shape transition. For audio synthesis, the selected units are simply concatenated to generate the speech audio.

To summarize, the CA system takes a 2D picture and text script with emotion markers as input. The emotion markers indicate the desired emotion for each sentence. Finally, the CA system generates a 3D talking head as output, as shown in Figure 5.

2.6.4 Audio Message Pilot Study—To help refine the CA, we conducted a second pilot to examine the impact of synthesized versus natural voice on older adults' memory for clinical test results. The results of this pilot study also help to determine the voice to be used in the CA condition in the message evaluation studies. As in the video message pilot study, both quantitative and qualitative measures were used.

Twenty-four older adults (mean age=76; mean years of education=15) listened to a subset of the cholesterol messages that described results indicating lower, borderline, or higher risk. The messages were delivered either using the physician's voice from the video-enhanced messages (natural voice condition) or a synthesized voice (using the Vocalware speech synthesizer [70]). In both the natural and synthesized voice conditions, participants viewed the test results presented in standard format as they listened to the message. After each message, gist memory, perceived risk, and message satisfaction were measured. The participants then listened to natural and synthetic voice versions of the same message and indicated which voice (natural or synthesized) they preferred and why. We expected minimal differences in gist memory for the two types of voice messages (although there is some evidence that older adults better remember natural versus synthesized voice messages compared to younger adults [12]), even if participants prefer the natural voice messages [47].

Gist memory was accurate for both conditions, with no difference between the two (Natural voice = 88%, Synthesized = 90% correct; $t(23) < 1.0$; Cohen d= .08; power=.05). Perceived risk (Natural voice = 6.6, Synthesized = 6.5; $t(23) < 1.0$; Cohen d= .05; power=.04) and message satisfaction (1= least satisfied; 9= most satisfied; Natural voice = 6.3, Synthesized = 6.3; $t(23) < 1.0$; Cohen d= .01; power=.03) were also similar for the two conditions. Moreover, when participants directly compared natural and synthetic versions of the same message, there was not a reliable difference in preference for the two voices (natural=58%; synthetic=42%, binomial $p > .50$). Participants who preferred the natural voice mentioned that this voice sounded more human, conversational, and empathetic compared to the synthetic voice. On the other hand, those preferring the synthetic voice mentioned that this voice was clearer and attracted attention. In addition, 25% of participants mentioned that they did not have a preference although required to choose an option.

Therefore, even though there was a numeric trend in favor of the natural voice in a direct comparison with the synthetic voice, participants were satisfied with and accurately remembered and responded to the risk information in both types of messages. Based on these findings, as well as the fact that synthesized voice would be more feasible than recorded speech for large-scale implementation in patient portals, we decided to use the synthesized voice in the message evaluation experiments.

3. Discussion

3.1 Conclusions

We described a multi-phase project intended to improve the use of patient portals to EMR systems by patients with diverse cognitive and literacy abilities. The project should help realize the potential of portals to support patient-centered care, especially patient/provider

collaboration [17]. In this paper we focused on our approach for accomplishing phase 1 goals: Develop portal message formats for clinical test results and evaluate whether they improve the ability of patients to understand and use numeric information in a simulated patient portal environment. We described a multidisciplinary approach and methodology to design and evaluate portal-based messages guided by a framework that integrates cognitive and behavioral science theories (health literacy, text processing, fuzzy trace memory, behavior change) with computational and engineering approaches. The framework guides a multi-method approach to evaluating theory-based predictions about the impact of message formats on patient responses. These methods include experimental and individual difference techniques to evaluate messages and investigate whether some formats benefit some patients more than others (e.g., enhanced formats especially benefit adults with lower numeracy and literacy abilities), using both quantitative measures of message memory (based on fuzzy trace theory) and intention to perform behaviors in response to the messages (based on behavior change theories), and more qualitative measures.

More generally, the project demonstrates the value of 'use-inspired theoretical research' [71]: Theory-based approaches to address important practical problems such as supporting patients' use of health information technology intended to improve the quality and safety of health care. Theory plays a vital role in guiding the development and evaluation of effective solutions. At the same time, putting theory to work in this way can advance theory. For example, the project has required integration of theories of comprehension and memory with behavioral change in order to adequately address issues related to improving patient use of numeric information (clinical test results), which will spur development of more comprehensive theories rooted in cognition, individual differences (e.g., health literacy), and behavior change. Similarly, while previous studies have demonstrated speech synthesis with modulated emotion and affect, almost all such studies have used extremely small test corpora. The present research reveals challenges to scaling up such results to robust use of synthesized voice as part of CA applications in health contexts, even as it suggests the value of doing so.

3.2. Future Directions

Our general approach of combining informal pilot studies using qualitative as well as quantitative measures that guide formal experimental evaluation allows systematic exploration of the most effective message formats before committing to larger-scale studies in actual portal systems. For example, it is possible that the video or CA messages will be less effective than other enhanced formats because the demands of integrating information from the video and the graphic display may overload older adults' visual attention [72]. Similarly, in the pedagogical agent literature, there is some evidence that the visual presence of CAs distract attention from the primary learning tasks ([73], but for a different view see [48]). If this is the case, we will evaluate alternative formats, such as eliminating the visual component of the video and testing whether adding just the audio (physician's spoken commentary) to the graphic display provides useful verbal context without the cognitive cost of integrating multiple visual sources of information. This finding would be consistent with evidence that voice-only CAs can be more effective than CAs with image as well as voice [72].

We also plan to explore a range of CAs before advancing to later project phases. For example, providing the option of choosing from a set of CAs that vary in gender, age, and race may improve patients' response to the CA. There is also debate about the appropriate level of realism for CAs. Adding a CA to a lesson generally improves learning [48], especially if the CA has a human appearance (animated vs static) that produces a 'social stance' to the CA that primes learning [74, 47, 8]). However, increasing CA realism may have drawbacks. For example, in the 'uncanny valley' effect, increasing realism until the CA appears almost, but not quite human, can produce negative responses [75]. The photo-realistic CA in our project, may fall into the 'uncanny valley', such that less realistic versions of the CA will be more effective (although abstract agents that do not resemble people can reduce trust [55]). These small-scale controlled studies using simple task will set the stage to explore CA benefits for more complex patient self-care applications. These include helping patients understand the conditional probability of an illness given a positive test result and population base rates, or conveying the risks and benefits of taking medication in the context of a CA-based 'medication adviser' for older adults at home.

Systematically scaling up the complexity of patient-centered applications for CAs will help optimize CA design for actual patient portal systems. Implementing CA portal systems have been proposed (e.g. [76]) but pose many challenges, including easy to use interfaces for clinicians that fit their work flow and supporting the use of multiple CAs for a wide range of patient needs and goals. Meeting these challenges will require elaboration of how motivational and behavioral change theories interact with the cognitive components of our framework, as well as developing tighter links between the cognitive and computational/ engineering science components of the framework. To sum up, while the success of our specific project goal of improving patient portal messages awaits the results of the evaluation study, we believe that the general framework and approach described in this paper is important in itself. It can serve as a guide for researchers and practitioners from behavioral science, engineering, medicine, and other disciplines to integrate their different theoretical and methodological perspectives in order to meet the complex challenges involved in designing and implementing patient-centered systems.

Acknowledgments

References

1. Garcia-Retamero R, Cokely ET. Communicating health risks with visual aids. Current Directions in Psychological Science. 2013; 22(5):392–399.

2. Peters E. Beyond Comprehension The Role of Numeracy in Judgments and Decisions. Current Directions in Psychological Science. 2012; 21(1):31–35.

3. Nielsen-Bohlman, L., Panzer, AM., Kindig, DA. Health literacy: A prescription to end confusion. Washington, DC: The National Academies Press; 2004.

4. Reyna VF, Nelson W, Han P, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. Psychological Bulletin. 2009; 135(6):943–973. [PubMed: 19883143]

5. Apter AJ, Paasche-Orlow MK, Remillard JT, Bennett IM, Ben-Joseph EP, Batista RM, et al. Numeracy and communication with patients: They are counting on us. Journal of General Internal Medicine. 2008; 23(12):2117–2124. [PubMed: 18830764]

6. Finucane ML, Slovic P, Hibbard JH, Peters E, Mertz C, MacGregor DG. Aging and decision-making competence: An analysis of comprehension and consistency skills in older versus younger adults considering health-plan options. Journal of Behavioral Decision Making. 2002; 15(2):141–164.

7. Galesic M, Garcia-Retamero R. Statistical numeracy for health: A cross-cultural comparison with probabilistic national samples. Archives of Internal Medicine. 2010; 170(5):462–468. [PubMed: 20212183]

8. Reeves, B., Nass, C. How people treat computers, television, and new media like real people and places. Cambridge, UK: CSLI Publications and Cambridge University Press; 1996. p. 19-36.

9. Institute of Medicine. Health IT and Patient Safety: Building Safer Systems for Better Care. Washington DC: The National Academies Press; 2012.

10. Ammenwerth E, Schnell-Inderst P, Hoerbst A. The impact of electronic patient portals on patient care: a systematic review of controlled trials. Journal of medical Internet research. 2012; 14(6):e162. [PubMed: 23183044]

11. Griffin A, Skinner A, Thornhill J, Weinberger M. Patient Portals. Applied Clinical Informatics. 2016; 7(2):489–501. [PubMed: 27437056]

12. Roring RW, Hines FG, Charness N. Age differences in identifying words in synthetic speech. Human Factors: The Journal of the Human Factors and Ergonomics Society. 2007; 49(1):25–31.

13. Sarkar U, Karter AJ, Liu JY, Adler NE, Nguyen R, López A, Schillinger D. The literacy divide: health literacy and the use of an internet-based patient portal in an integrated health system—results from the Diabetes Study of Northern California (DISTANCE). Journal of Health Communication. 2010; 15(S2):183–196. [PubMed: 20845203]

14. Sharit J, Lisigurski M, Andrade AD, Karanam C, Nazi KM, Lewis JR, Ruiz JG. The roles of health literacy, numeracy, and graph literacy on the usability of the VA's personal health record by veterans. Journal of Usability Studies. 2014; 9(4):173–193.

15. Taha J, Czaja S, Sharit J, Morrow D. Factors Affecting the Usage of a Personal Health Record (PHR) to Manage Health. Psychology and Aging. 2013; 28(4):1124–1139. [PubMed: 24364414]

16. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient education and counseling. 2009; 74(3):295–301. [PubMed: 19150199]

17. Kessels RP. Patients' memory for medical information. Journal of the Royal Society of Medicine. 2003; 96(5):219–222. [PubMed: 12724430]

18. Stead, WW., Lin, HS., editors. Computational technology for effective health care: Immediate steps and strategic directions. Washington, DC: National Academies Press; 2009.

19. Patel VL, Kannampallil TG. Cognitive informatics in biomedicine and healthcare. Journal of biomedical informatics. 2015; 53:3–14. [PubMed: 25541081]

20. Morrow, DG., Chin, J. Health literacy and health decision making among older adults. In: Hess, T.Strough, J., Lockenhoff, C., editors. Aging and Decision-Making: Empirical and Applied Perspectives. London: Elsevier; 2015. p. 261-282.

21. Reyna VF. A theory of medical decision making and health: Fuzzy-trace theory. Medical Decision Making. 2008; 28:829– 833. [PubMed: 19020341]

22. Reyna VF, Brainerd CJ. Fuzzy-trace theory: An interim synthesis. Learning and Individual Differences. 1995; 7(1):1–75.

23. Peters E, Dieckmann NF, Västfjäll D, Mertz CK, Slovic P, Hibbard JH. Bringing meaning to numbers: The impact of evaluative categories on decisions. Journal of Experimental Psychology: Applied. 2009; 15(3):213–227. [PubMed: 19751072]

24. Kintsch, W. Comprehension: A paradigm for cognition. New York, NY: Cambridge University Press; 1998.

25. Chin J, Morrow DG, Stine-Morrow EAL, Conner-Garcia T, Graumlich JF, Murray MD. The process-knowledge model of health literacy: Evidence from a componential analysis of two

commonly used measures. Journal of Health Communication. 2011; 16:222–241. [PubMed: 21951254]

26. Douglas SE, Caldwell BS. Design and validation of an Individual Health Report (IHR). International Journal of Industrial Ergonomics. 2011; 41(4):352–359.

27. Cheng S, Lichtman JH, Amatruda JM, Smith GL, Mattera JA, Roumanis SA, Krumholz HM. Knowledge of cholesterol levels and targets in patients with coronary artery disease. Preventive cardiology. 2005; 8(1):11–17. [PubMed: 15722689]

28. Reyna, V. Across the life span. In: Fischoff, B.Brewer, N., Downs, J., editors. Communicating Risks and Benefits: A Users Guide. Washington, DC: FDA; 2011. p. 111-120.

29. Chin J, Madison A, Stine-Morrow EAL, Gao X, Graumlich JF, Murray MD, Conner-Garcia T, Morrow DG. Cognition and health literacy in older adults' recall of self-care information. The Gerontologist. 2015:gnv091.

30. Charles ST, Carstensen LL. Social and emotional aging. Annual review of psychology. 2010; 61:383–409.

31. Brewer NT, Chapman GB, Gibbons FX, Gerrard M, McCaul KD, Weinstein ND. Meta-analysis of the relationship between risk perception and health behavior: the example of vaccination. Health Psychology. 2007; 26(2):136. [PubMed: 17385964]

32. Meyer D, Leventhal H, Gutmann M. Common-sense models of illness: the example of hypertension. Health Psychology. 1985; 4(2):115–135. DOI: 10.1037/0278-6133.4.2.115 [PubMed: 4018002]

33. Ajzen I, Fishbein M. Attitude-behavior relations: A theoretical analysis and review of empirical research. Psychological bulletin. 1977; 84(5):888.

34. Schwarzer, R., Luszczynska, A., Ziegelmann, JP., Scholz, U., Lippke, S. Social-cognitive predictors of physical exercise adherence: three longitudinal studies in rehabilitation. Vol. 27. American Psychological Association; 2008. p. S54

35. Webb TL, Sheeran P. Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. Psychological Bulletin. 2006; 132(2):249. [PubMed: 16536643]

36. Tanius BE, Wood S, Hanoch Y, Rice T. Aging and choice: Applications to Medicare Part D. Judgment and Decision Making. 2009; 4:92–101.

37. Mikels JA, Loeckenhoff CE, Maglio SJ, Goldstein MK, Garber A, Carstensen LL. Following your heart or your head: Focusing on emotions versus information differentially influences the decisions of younger and older adults. Journal of Experimental Psychology: Applied. 2010; 16(1):87–95. [PubMed: 20350046]

38. Henneman L, Marteau TM, Timmermans DR. Clinical geneticists' and genetic counselors' views on the communication of genetic risks: A qualitative study. Patient education and counseling. 2008; 73(1):42–49. [PubMed: 18583088]

39. Ambady N, Koo J, Rosenthal R, Winograd C. Physical therapists' nonverbal communication predicts geriatric patients' health outcomes. Psychology and Aging. 2002; 17(3):443–452. [PubMed: 12243386]

40. Roter DL, Frankel RM, Hall JA, Sluyter D. The expression of emotion through nonverbal behavior in medical visits. Journal of General Internal Medicine. 2006; 21(S1):S28–S34.

41. Ancker JS, Senathirajah Y, Kukafka R, Starren JB. Design features of graphs in health risk communication: A systematic review. Journal of the American Medical Informatics Association. 2006; 13(6):608–618. [PubMed: 16929039]

42. Fagerlin A, Zikmund-Fisher BJ, Ubel PA, Jankovic A, Derry HA, Smith DM. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. Medical Decision Making. 2007; 27(5):672–680. [PubMed: 17641137]

43. Garcia-Retamero R, Galesic M. Communicating treatment risk reduction to people with low numeracy skills: A cross-cultural comparison. American journal of public health. 2009; 99(12): 2196–2202. [PubMed: 19833983]

44. Larkin J, Simon H. Why a diagram is (sometimes) worth ten thousand words. Cognitive Science. 1987; 11(1):65–100.

45. Pachman M, Ke F. Environmental support hypothesis in designing multimedia training for older adults: is less always more? Computers & Education. 2012; 58(1):100–110.

46. Van Gerven PWM, Paas F, Van Merriënboer JJG, Hendriks M, Schmidt HG. The efficiency of multimedia learning into old age. British Journal of Educational Psychology. 2003; 73(4):489–505. [PubMed: 14713374]

47. Mayer RE, DaPra CS. An embodiment effect in computer-based learning with animated pedagogical agents. Journal of Experimental Psychology: Applied. 2012; 18(3):239–252. [PubMed: 22642688]

48. Moreno, R. Multimedia learning with animated pedagogical agents. In: Mayer, R., editor. The Cambridge handbook of multimedia learning. New York, NY: Cambridge University Press; 2005. p. 507-523.

49. Schroeder NL, Adesope OO, Gilbert RB. How effective are pedagogical agents for learning? A meta-analytic review. Journal of Educational Computing Research. 2013; 49(1):1–39.

50. Bickmore T, Gruber A, Picard R. Establishing the computer–patient working alliance in automated health behavior change interventions. Patient Education and Counseling. 2005; 59(1):21–30. [PubMed: 16198215]

51. Bickmore T, Pfeifer L, Byron D, Forsythe S, Henault L, Jack B, et al. Usability of conversational agents by patients with inadequate health literacy: Evidence from two clinical trials. Journal of Health Communication. 2010; 15(2):197–210. [PubMed: 20845204]

52. Bickmore T, Pfeifer L, Paasche-Orlow M. Using Computer Agents to Explain Medical Documents to Patients with Low Health Literacy. Patient Education and Counseling. 2009; 75(3):315–320. [PubMed: 19297116]

53. Wolfe CR, Reyna VF, Widmer CL, Cedillos EM, Fisher CR, Brust-Renck PG, Weil AM. Efficacy of a Web-Based Intelligent Tutoring System for Communicating Genetic Risk of Breast Cancer: A Fuzzy-Trace Theory Approach. Medical Decision Making. 2015; 35(1):46–59. [PubMed: 24829276]

54. Garcia-Retamero R, Cokely ET. Effective communication of risks to young adults: Using message framing and visual aids to increase condom use and STD screening. Journal of Experimental Psychology-Applied. 2011; 17(3):270. [PubMed: 21942316]

55. Pak R, Fink N, Price M, Bass B, Sturre L. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. Ergonomics. 2012; 55(9):1059–1072. [PubMed: 22799560]

56. Baker DW, Williams MV, Parker RM, Gazmararian JA, Nurss J. Development of a brief test to measure functional health literacy. Patient Education and Counseling. 1999; 38:33–42. [PubMed: 14528569]

57. Cokely ET, Galesic M, Schulz E, Ghazal S, Garcia-Retamero R. Measuring risk literacy: The Berlin numeracy test. Judgment and Decision Making. 2012; 7(1):25–47.

58. Salthouse TA, Babcock RL. Decomposing adult age differences in working memory. Developmental psychology. 1991; 27(5):763.

59. Levinthal BR, Morrow DG, Tu W, Wu J, Young J, Weiner M, Murray MD. Cognition and health literacy in patients with hypertension. Journal of General Internal Medicine. 2008; 23(8):1172–1176. [PubMed: 18459011]

60. Ekstrom, RB., French, JW., Harman, HH., Dermen, D. Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational testing service; 1976. p. 109-113.

61. Stanovich KE, West RF, Harrison MR. Knowledge growth and maintenance across the life span: The role of print exposure. Developmental Psychology. 1995; 31(5):811.

62. Azevedo RFL, Morrow D, Hasegawa-Johnson M, Gu K, Soberal D, Huang T, Schuh W, Garcia-Retamero R. Improving Patient Comprehension of Numeric Health Information. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2015; 59(10):488–492.

63. Gulz A, Haake M. Design of animated pedagogical agents—A look at their look. International Journal of Human-Computer Studies. 2006; 64(4):322–339.

64. Tang, H., Hu, Y., Fu, Y., Hasegawa-Johnson, M., Huang, TS. Real-time conversion from a single 2D face image to a 3D text-driven emotive audio-visual avatar. 2008 IEEE International Conference on Multimedia and Expo; 2008 Jun. p. 1205-1208.

65. Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, … Valtchev V. The HTK book. Cambridge university engineering department. 2002; 3:175.

66. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … Silovsky, J. The Kaldi speech recognition toolkit. IEEE 2011 workshop on automatic speech recognition and understanding; IEEE Signal Processing Society; 2011. (No. EPFL-CONF-192584)

67. Dutoit, T. Springer Handbook of Speech Processing. Springer; Berlin Heidelberg: 2008. Corpus-based speech synthesis; p. 437-456.

68. Zhang Y, Mysore G, Berthouzoz F, Hasegawa-Johnson M. Analysis of Prosody Increment Induced by Pitch Accents for Automatic Emphasis Correction. Proc Speech Prosody. 2016:79–83.

69. Xu Y, Prom-On S. Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. Speech Communication. 2014; 57:181–208.

70. Vocalware Speech Synthesizer [Software]. Available from https://www.vocalware.com/#&panel11&panel1

71. Stokes, DE. Pasteur's quadrant: Basic science and technological innovation. Washington, DC: Brookings Institution Press; 1997.

72. Mayer RE, Moreno R. Nine ways to reduce cognitive load in multimedia learning. Educational psychologist. 2003; 38(1):43–52.

73. Choi S, Clark RE. Cognitive and affective benefits of an animated pedagogical agent for learning English as a second language. Journal of educational computing research. 2006; 34(4):441–466.

74. Lusk MM, Atkinson RK. Animated pedagogical agents: Does their degree of embodiment impact learning from static or animated worked examples? Applied cognitive psychology. 2007; 21(6): 747–764.

75. Mori M. Bukimi no tani [The uncanny valley]. Energy. 1970; 7(4):33–35. (Translated by Karl F. MacDorman and Takashi Minato in 2005) within Appendix B for the paper Androids as an Experimental Apparatus: Why is there an uncanny and can we exploit it?. *In Proceedings of the CogSci-2005 Workshop: Toward Social Mechanisms of Android Science* (pp. 106–118).

76. Tang C, Lorenzi N, Harle CA, Zhou X, Chen Y. Interactive systems for patient-centered care to enhance patient engagement. Journal of the American Medical Informatics Association. 2016; 23(1):2–4. [PubMed: 26912537]

Highlights

1. Improve the use of Electronic Medical Record (EMR) portal information by older adults with diverse numeracy and literacy abilities (support patient-centered care).

2. Describe a multi-disciplinary approach to designing and evaluating portal-based messages.
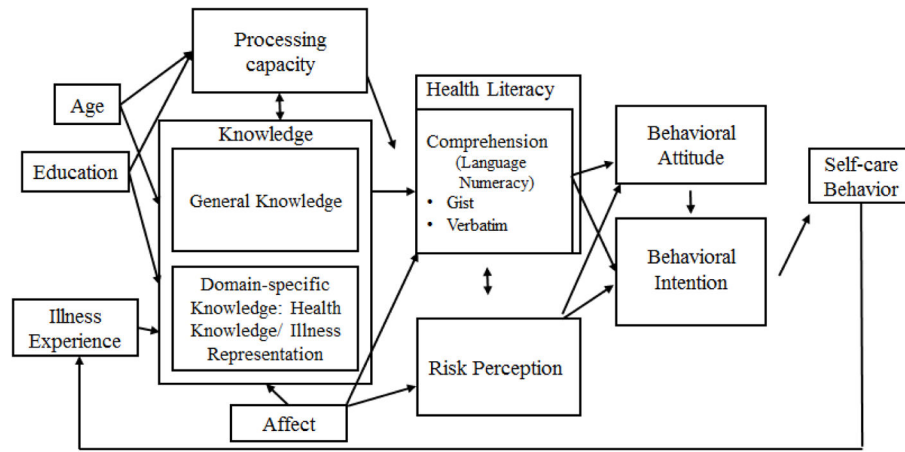
Figure 1.
Framework guiding the design and evaluation of portal messages (adapted from [20])

Component	Your Value	Standard Range	Units
Total Cholesterol	184	< 200 -	mg/dl
Triglycerides	42	< 150 -	mg/dl
HDL Cholesterol	47	40 - 60	mg/dl
LDL Cholesterol	130	< 100 -	mg/dl

"Your test results require discussion to assess your future plan of care. A follow up appointment is recommended to discuss your results."

Figure 2A

Component	Your Value (mg/dl)	Range of Scores (mg/dl)
Total Cholesterol	184 **Desirable**	< 200 (Desirable) 200-240 (Borderline) >240 (High)
Triglycerides	42 **Optimal**	< 100 (Optimal) 101-149 (Normal) 150-199 (Borderline) 200-499 (High) >500 (Very High)
HDL Cholesterol	47 **Borderline**	< 40 (Low/Bad) 41-59 (Borderline) >60 (High/Good)
LDL Cholesterol	130 **Borderline**	< 100 (Optimal) 101-129 (Near Optimal) 130-159 (Borderline) 160-189 (High) >190 (Very High)

"Your risk for heart disease is borderline. I recommend a follow up appointment to discuss your future plan of care."
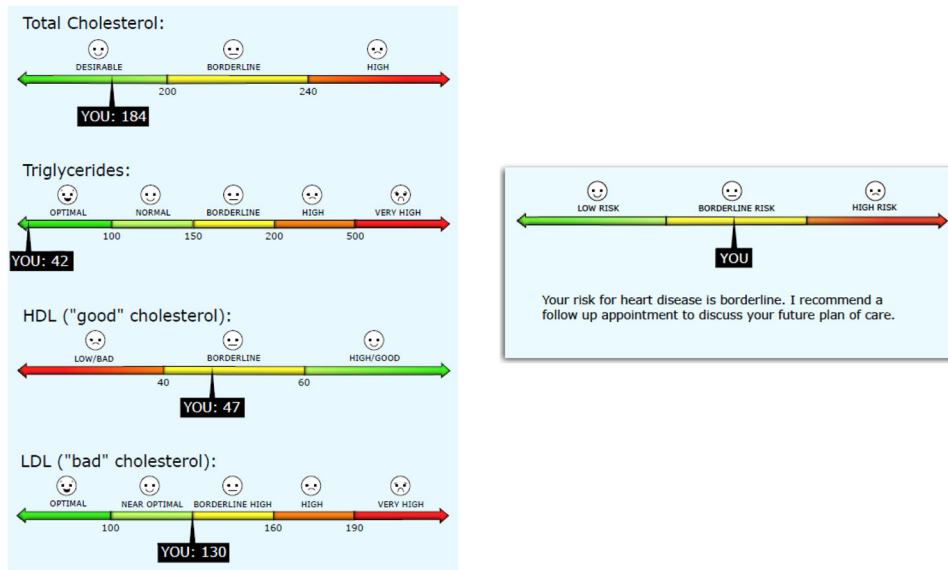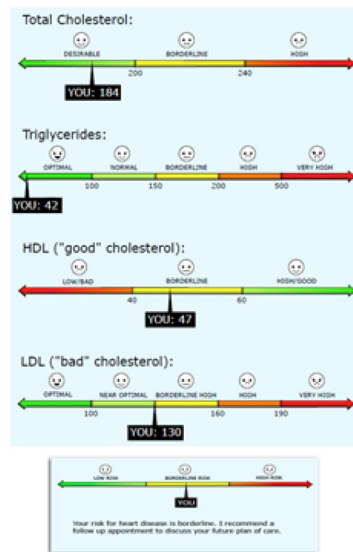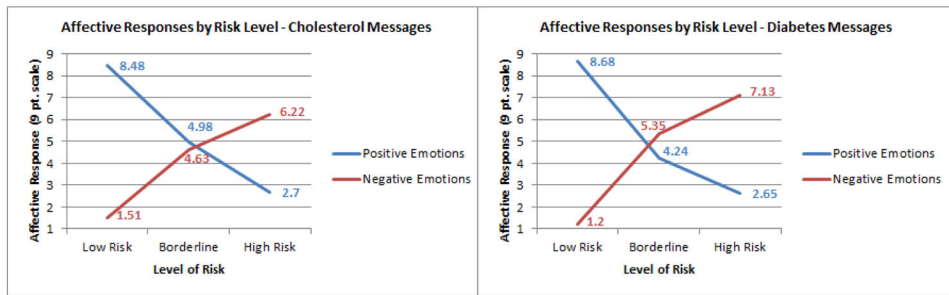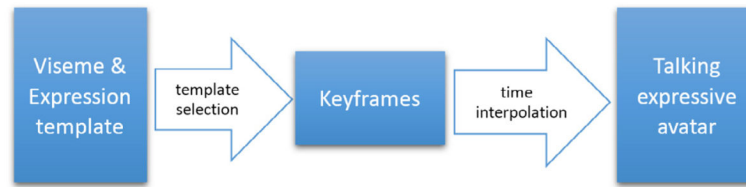
Figure 2B

Figure 2C

Figure 2.
Figure 2A. Standard Message Format
Figure 2B. Verbally Enhanced Message Format
Figure 2C. Graphically Enhanced Message Format
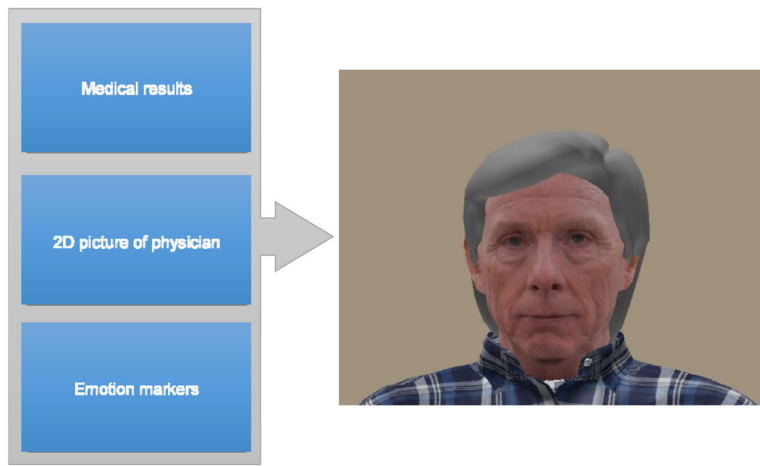Figure 2D. Video Enhanced Message Format [1]
[1]Speech balloon is for illustration purposes only. Participants in the study hear the speech.

Figure 3.
Affective Responses to Low, Borderline, and High Risk Cholesterol and Diabetes Video
Messages (Adapted from [62])

Figure 4.
Generating Expressive CA from visual and emotion information (adapted from [64]).

Figure 5.
Computer-based Agent (CA) Interface

Table 1

Mean (std) and reliability estimates for participant responses to video messages (adapted from [62])

		n# scenarios	Cronbach's alpha	Mean	SD
Cholesterol	Q1) I think the way the doctor conveyed the information matched what he said.	12	0.93	8.46	1.31
	Q2) How useful do you think was the information conveyed in this message?	12	0.95	8.45	1.15
Diabetes	Q1) I think the way the doctor conveyed the information matched what he said	12	0.90	8.60	0.87
	Q2) How useful do you think was the information conveyed in this message?	12	0.95	8.57	0.94