Int. J. Human-Computer Studies 83 (2015) 94–113

Contents lists available at ScienceDirect

Int. J. Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Benefits of visualization in the Mammography Problem ☆

Azam Khan [a,b,*], Simon Breslav [b], Michael Glueck [b], Kasper Hornbæk [a]

[a] Department of Computer Science, University of Copenhagen. Njalsgade 128, DK-2300 Copenhagen, Denmark
[b] Autodesk Research, 210 King Street East, Toronto, Ontario, Canada M5A 1J7

A R T I C L E I N F O

Article history:
Received 15 December 2014
Received in revised form
15 June 2015
Accepted 3 July 2015
Communicated by Scott Bateman
Available online 15 July 2015

Keywords:
Bayesian reasoning
Decision making
Comparability criteria
Visualization
Crowdsourcing
Mammography Problem

A B S T R A C T

Trying to make a decision between two outcomes, when there is some level of uncertainty, is inherently difficult because it involves probabilistic reasoning. Previous studies have shown that most people do not correctly apply Bayesian inference to solve probabilistic problems for decision making under uncertainty. In an effort to improve decision making with Bayesian problems, previous work has studied supplementing the textual description of problems with visualizations, such as graphs and charts. However, results have been varied and generally indicate that visualization is not an effective technique. As these studies were performed over many years with a variety of goals and experimental conditions, we sought to re-evaluate the use of visualization as an aid in solving Bayesian problems. Many of these studies used the classic Mammography Problem with visualizations portraying the problem structure, the quantities involved, or the nested-set relations of the populations involved. We selected three representative visualizations from this work and developed two hybrid visualizations, combining structure types and frequency with structure. We also included a text-only baseline condition and a text-legend condition where all nested-set problem values were given to eliminate the need for participants to estimate or calculate values. Seven hundred participants evaluated these seven conditions on the classic Mammography Problem in a crowdsourcing system, where micro-interaction data was collected from the participants. Our analysis of the user input and of the results indicates that participants made use of the visualizations but that the visualizations did not help participants to perform more accurately. Overall, static visualizations do not seem to aid a majority of people in solving the Mammography Problem.

1. Introduction

Decision making can be simple when there are limited choices and all the available options are known. However, unknowns introduce probabilities and the need for statistical inference. One method of modelling statistical inference is the Bayes theorem. For many years Bayesian problems have been presented to subjects to test if people are rational when making decisions under uncertainty. However, the majority of people do not answer these problems correctly.

Bayesian problems have been studied for many years in the fields of medical decision making, human–computer interaction (HCI), and information visualization. To help people better understand the subtleties of these problems, visualizations of the problem structure or the quantities involved have been studied. As these studies were performed over a long period of time with a variety of goals and experimental conditions, the aim of the present paper is to re-evaluate the use of visualization as an aid in solving Bayesian problems.

Given the variety of visual properties employed in the visualizations of Bayesian problems in previous work, we sought to control more factors of design properties than has previously been done to better explain differences in performance. To this end, we developed *comparability criteria* to (a) help normalize the information content of the visualizations across experimental conditions and (b) develop the conditions for the experiment, including two novel visualization conditions. Also, following the recommendations made by a previous study on Bayesian visualization (Breslav et al., 2014), and other work in visual analytics (Segel and Heer, 2010) and bioinformatics (Turkay et al., 2014), we designed the *problem presentation* and recorded micro-interaction data to confirm the effectiveness of the way that the Bayesian problem was presented to users. We ran a controlled crowdsourcing experiment with 700 participants and we provide a detailed analysis together with a complete supplemental material report.

The benefits of these contributions are twofold. First, the work clearly shows the lack of benefit of static visualization in the

☆This paper has been recommended for acceptance by Scott Bateman.

* Corresponding author at: Autodesk Research, 210 King Street East, Toronto, Ontario, Canada M5A 1J7.

E-mail addresses: azam.khan@autodesk.com (A. Khan), simon.breslav@autodesk.com (S. Breslav), michael.glueck@autodesk.com (M. Glueck), kash@diku.dk (K. Hornbæk).

http://dx.doi.org/10.1016/j.ijhcs.2015.07.001
1071-5819/

Mammography Problem. Second, we propose a generalized methodology of visualization comparison which supports the comparison of distinct visual representations of the same underlying data. This is achieved by consideration of both the content and the structure of this underlying data. We use this methodology to produce distinct visual representations which do not differ in the level of information provided to a user, removing the potential confound that different visual representations provide participants with more or less information. Removing these confounds allows us to explore the effectiveness of different visual representations on a level playing field.

This study shows the value of capturing and studying micro-interactions and the value of disaggregating the analysis of the two key parts of the user input in Bayesian problems (numerator and denominator). The results point to the need to address confusion about both the question and the visualization. This could be achieved through a better correspondence between the question and the visualization, which could perhaps be presented using more compelling or engaging techniques such as animated or interactive visualizations (Wong et al., 2011), to help increase accuracy rates for this important class of problems.

We first describe the Mammography Problem in detail and show how it represents Bayesian problems. We then survey the visualizations that have been studied for this problem and extract a design space that we will use in a later section. Based on lessons learned in previous work, we present several criteria to consider when performing experiments to compare visualizations, especially in a crowdsourcing environment. Taking both the visualization design space and comparability criteria into account, we present the visualizations we designed for a controlled online experiment. To ensure as much consistency as possible in the experimental environment of the participants, we discuss the presentation design as a critical control factor that has not been discussed in previous works that have employed crowdsourcing. Finally, we present a controlled experiment and report on the results. We conclude with a discussion on the value of collecting and examining micro-interaction data to help directly answer questions that could previously only be answered indirectly.

2. The Mammography Problem

Bayesian problems can be presented in many different ways but always have the same structure. For example, if the problem uses a medical test as its scenario, two pieces of information are given. First, the number of people who receive a positive or negative test result is stated. Second, the number of people who actually have the condition, for which the test is being performed, is stated. The subject is then asked to answer one of four possible conditional probability questions.

To better compare results between experiments, a canonical Bayesian problem called the Mammography Problem, concerning probabilistic diagnosis, evolved from Casscells et al. (1978) and Eddy (1982). This problem is often used in decision making studies and consists of two parts, a problem statement, containing the two pieces of information mentioned above, and a problem question. One textual representation of the problem is

At age forty, when women participate in routine screening for breast cancer, 10 out of 1000 will have breast cancer. However, 8 of every 10 women with breast cancer will get a positive mammography, and 95 out of every 990 women without breast cancer will also get a positive mammography.

Given a new group of women at age forty who got a positive mammography in routine screening, how many of these women do you expect to actually have breast cancer?

From the information given in the textual problem statement, a number of values can be extracted and derived, from which many problem questions can be answered, including the question posed above. First, we see that the whole population is 1000 women and that there seem to be some implicit assumptions. For example, by definition it seems that a mammography test is either positive or negative and that a women either has breast cancer or does not have breast cancer. This latter statement is actually supported in the problem statement in that 10 women have cancer and 990 women do not have cancer. Of the 10 women with breast cancer, 8 women will get a positive mammography (a true-positive result), implying that 2 women with breast cancer will get a negative mammography (false-negative). Finally, the problem states that of the 990 women without cancer, 95 women will still get a positive mammography even though they do not have breast cancer (false-positive). Since $990-95=895$, this implies that 895 women who do not have breast cancer will correctly get a negative mammography (true-negative). We can also calculate the total number of women that got a positive mammography as $8+95=103$ women. And lastly, as $1000-103=897$, this implies that, in total, 897 women got a negative mammography. We summarize these values in Table 1. The first column describes the Group of women in question, and the second column shows the Nested-set Equation defining that Group. The value column shows the number of women in each group and has a blue background if the number is extracted directly from the question text but has a yellow background if the number of women is derived from the extracted numbers using a simple calculation.

We can now answer the posed question: given a new group of women at age forty who got a positive mammography in routine screening (got positive mammography$=$103), how many of these women do you expect to actually have breast cancer (have breast cancer and got positive mammography$=$8)? Therefore the correct answer is 8 out of 103 women.

Table 1
Extracted and derived values from the Mammography Problem. Values are extracted from the problem text and an asterisk indicates a derived value. Using the notation of Gigerenzer and Hoffrage (1995), d is *data* obtained from the mammography test and h is the *hypothesis* or outcome of cancer.

Group	Nested-set equation	Value	Outcome
Got positive mammography	d	103*	Positive
Have breast cancer	h	10	True
Have breast cancer and got positive mammography	$h \wedge d$	8	True-positive
Have breast cancer and got negative mammography	$h \wedge \neg d$	2*	False-negative
Got negative mammography	$\neg d$	897*	Negative
Do not have breast cancer	$\neg h$	990	False
Do not have breast cancer and got positive mammography	$\neg h \wedge d$	95	False-positive
Do not have breast cancer and got negative mammography	$\neg h \wedge \neg d$	895*	True-negative
Entire population	$\neg d \wedge d$	1000	Negative and positive
Entire population	$h \wedge \neg h$	1000	Cancer and no cancer