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, M5A 1J7, Canada*

Abstract

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.

Keywords: Bayesian reasoning, decision making, comparability criteria, visualization, crowdsourcing, 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 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) to 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)),

*Corresponding author
Email addresses: azam.khan@autodesk.com (Azam Khan), simon.breslav@autodesk.com (Simon Breslav), michael.glueck@autodesk.com (Michael Glueck), kash@diku.dk (Kasper Hornbæk)

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 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 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).

Group	Nested-Set Equation	Value	Outcome
Got positive mammography	d	103	positive
Have breast cancer	h	10	true
Have breast cancer & got positive mammography	$h \wedge d$	8	true-positive
Have breast cancer & 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 & got positive mammography	$\neg h \wedge d$	95	false-positive
Do not have breast cancer & got negative mammography	$\neg h \wedge \neg d$	895	true-negative
Entire Population	$\neg d \wedge d$	1000	negative & positive
Entire Population	$h \wedge \neg h$	1000	cancer & no cancer

Table 1: Extracted and derived values from the Mammography Problem. Blue cells are values extracted from the problem text and yellow cells indicate derived values. 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.

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 & got positive mammography = 8)? Therefore the correct answer is 8 out of 103 women.

To successfully answer this question, it seems that some knowledge about sets and nested-set relations will be needed. There may also be some cultural knowledge needed such as the definition of the word 'mammography' and understanding that, even though the word 'positive' has a connotation of good fortune, the term 'positive mammography' is in fact an unfortunate result.

Previous work has identified several opportunities for error in considering how to answer the question posed above. Specific difficulties have been examined such as the format of numeric data and the typical use of opaque percentages (e.g., "7.8%") or probability formats over a more transparent frequency format (Gigerenzer and Hoffrage, 1995), such as "8 out of 103".

The complex interplay between the three Bayesian parameters, namely sensitivity (true-positive rate), specificity (true-negative rate), and prevalence (number of cases of the condition in the population), has also been studied. For example, Cole (1989) found that subjects would significantly overestimate the effect of sensitivity or underestimate the effect of specificity when the problem was presented in this manner.

An understanding of the procedure to correctly apply Bayesian reasoning has also been investigated (Gigerenzer and Hoffrage, 1995). Other difficulties include base rate (prevalence) neglect or misjudging scale (Spiegelhalter et al., 2011), and mistakenly equating $P(A|B)$ with $P(B|A)$, that is, the probability of A given B is incorrectly thought to be equivalent to the probability of B given A (Casscells et al., 1978).

Unfortunately, explicitly providing Bayes theorem would require a lengthy mathematical explanation. The theorem states that the probability of A given B is equivalent to the probability of B given A multiplied by the probability of A, divided by the probability of B (see Equation 1).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Understanding even this basic formulation is challenging. When making an important decision, knowing how to apply and interpret this mathematical formula containing conditional probabilities with several interacting parameters is inherently difficult (Cole and Davidson, 1989). Yet this scenario is surprisingly commonplace in medical decision making, given laboratory

test results (Cole, 1989), when seeking informed consent, understanding the risks and benefits of participating in a clinical trial (Schapira et al., 2008), or for making financial decisions about investments without guaranteed returns (Spiegelhalter et al., 2011).

In making critical decisions in the medical domain, physicians and health care workers will presumably spend time with patients to achieve 100% understanding to ensure that all patients fully understand their test results and their implications. However, this is doubtful as physicians have also been shown to have poor understanding of Bayesian inference and so, may not be able to help patients in making informed medical decisions (Gigerenzer et al., 2007).

As an alternative to using significant educational efforts to help individuals to understand and apply Bayesian reasoning, researchers have proposed many kinds of visualizations of Bayesian problems including hierarchical trees, Euler diagrams, frequency grids and more. However, experimental results have generally been poor with recent work showing very low accuracy levels of only 5.0% from participants when solving the Mammography Problem (Micallef et al., 2012). To better understand why 95.0% of people do not correctly solve the Mammography problem, we examine previous studies and the visualizations they employed.

3. Visualizing the Mammography Problem

Visualizations of the Mammography Problem, used in studies over the past 25 years, are shown in Figure 1 together with untested and novel visualizations arranged to indicate commonalities between neighbouring visualizations. They are not designed to explain Bayes theorem in general but attempt to show this particular problem in a way that may help people to reason in accordance with Bayes' theorem to support better decision making.

We classify these depictions into three types of visual representation: *branching* (Figure 2a), *nested-set relations* (Figure 2b), and *frequency* (Figure 2c). Shown in Figure 2 is a group of 100 individuals that is made up of two subgroups with 10 members in one subgroup and 90 in the other. Figure 2a shows branching which uses node-link diagrams to represent the branching structures of trees. When reading the diagram in a top-down fashion, we call the splitting of a group into two distinct subgroups a *Branch* style. When two subgroups come together to form a larger group, we call this a *Join* style. Figure 2b shows the Nested style which spatially emphasizes containment, or the group membership of nested-set relations, by positioning circles representing subgroups within each other. We use the term *nested* as a short-form for *nested-set relation*. Figure 2c shows the Frequency style which spatially emphasizes the scale of the number of individuals in each subgroup. Here, each individual is represented explicitly as an icon or a glyph. Shading, together with a legend, is used to differentiate subgroup membership.

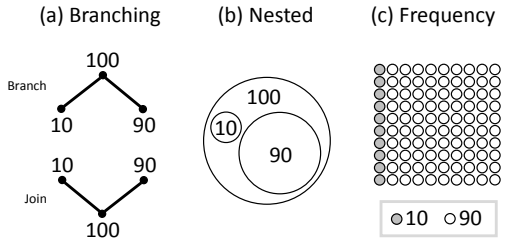


Figure 2: Branching, Nested, and Frequency diagram styles emphasizing specific characteristics of the data as spatial structures.

In Figure 1, we place the three visualization types (Branching, Nested, Frequency) at the three corners of a design space. Between these, we place hybrid visualizations that combine features of these extremes. We now describe each visualization shown in Figure 1.

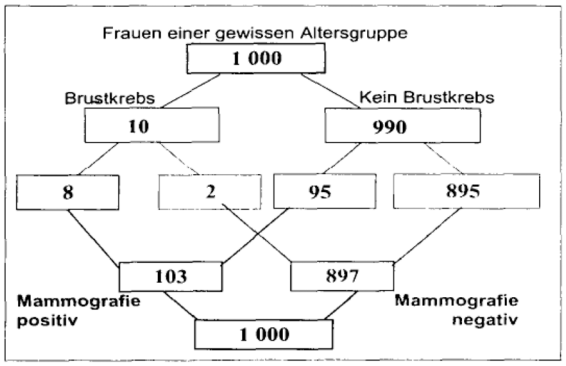


Figure 3: Double-tree visualization from Wassner et al. (2004).

3.1. Visualization Designs

Double-tree (Figure 1a and Figure 3): In Wassner et al. (2004), a double-tree representation is used to convey all of the sizes of the nested-sets for the Mammography Problem (as described in Table 1). The double-tree fully captures the double branching structure of a Bayesian problem, including the re-classification of members (shown as branches joining in the lower half of the diagram), and includes the complete set of all nine numeric values needed for Bayesian inference, avoiding the need for any kind of arithmetic calculation or estimation on the part of the participant.

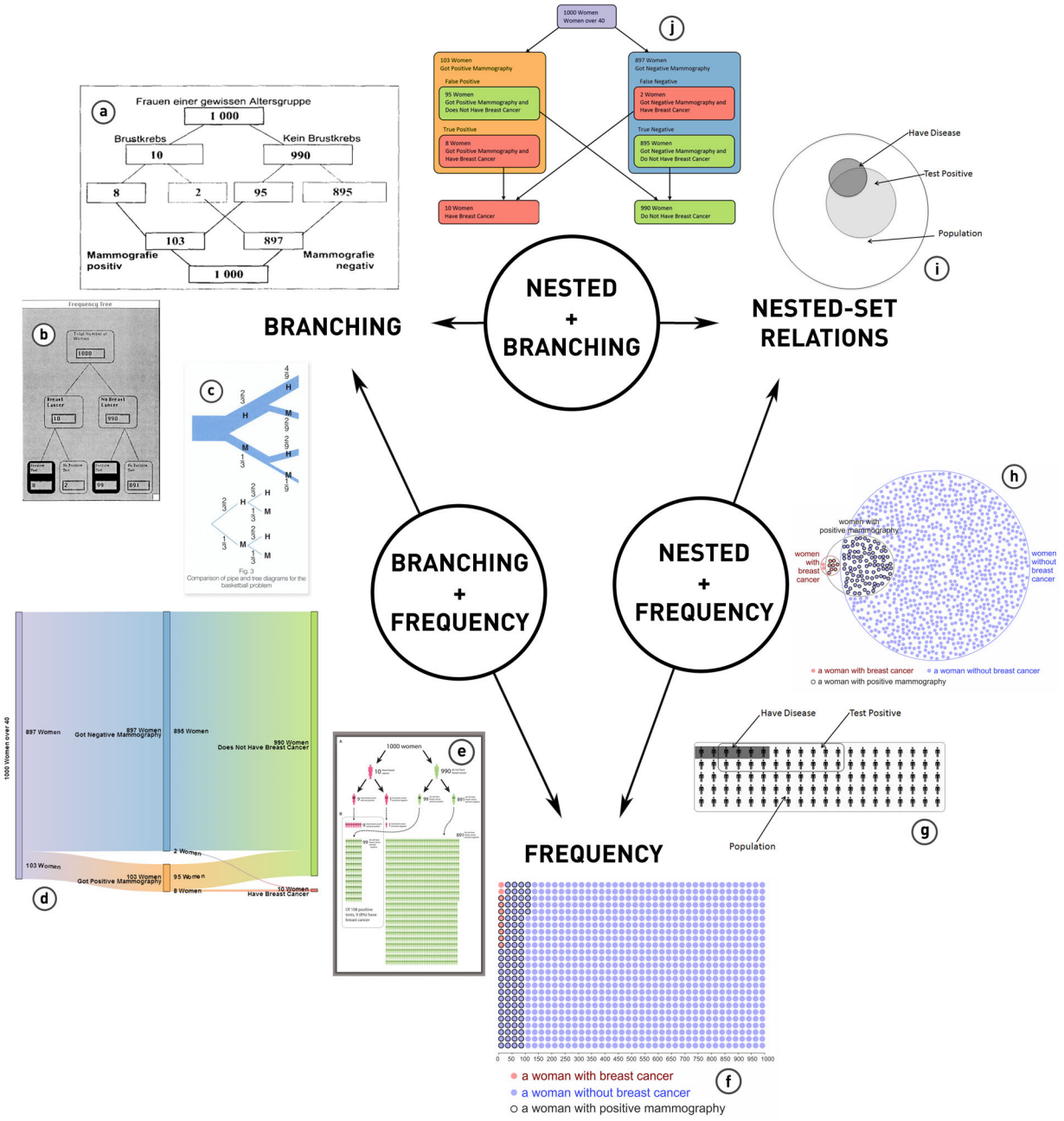


Figure 1: Design space of Bayesian visualizations: primary techniques at extents with hybrid combinations on edges. Visualizations are grouped to emphasize branching structure, frequency (scale of the subsets), and nested-set relations (emphasized by using properties of Euler diagrams). (a), (b), and (c) emphasize Branching characteristics, (d) and (e) exhibit a mix of Branching and Frequency, (f) primarily shows Frequency, (g) and (h) exhibit a mix of Frequency and Nested characteristics, (i) emphasizes Nested-set relations, and (j) is a mix of Nested and Branching styles. See Section 3 for detailed descriptions of the designs.

The false-positive/true-positive and false-negative/true-negative symmetry of the problem is directly represented in a visual way. However, note that the size of the sets is drawn as a box around the number, and does not reflect the size of the set. That is, the frequency information in the problem is represented numerically but not graphically in this type of visualization.

Tree (Figure 1b and Figure 4): A common visual representation for probability is the simple hierarchical tree (Sedlmeier, 1997; Sedlmeier and Gigerenzer, 2001; Dolan and Iadarola, 2008). The nodes in the tree may contain probability values between 0.0 and 1.0 or natural frequency values, between 0 and the size of the total population. Natural frequency whole numbers were preferred by Gigerenzer et al. (2007) and Kurz-Milcke et al. (2008), among others, arguing that natural frequencies constitute a proper representation of uncertainty.

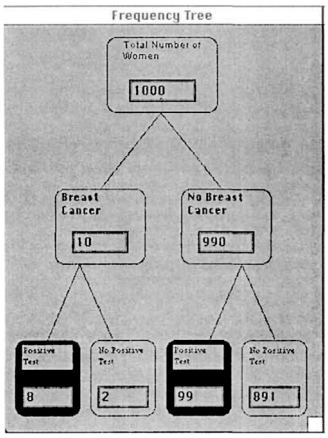


Figure 4: Tree visualization from Sedlmeier (1997).

Cole and Davidson (1989) suggested that using trees to represent probability is limiting and not well understood by students. Furthermore, as the tree is a subset of the double-tree, several of the complete set of values are not directly shown, implying that calculation or estimation is needed, limiting the usefulness of this representation. As in the case of the double-tree, frequency information is only represented numerically, not spatially.

Pipe Diagram (Figure 1c and Figure 5): To better visually represent the proportion, or probability, of one event over another, Konold (1996) suggested a pipe branching metaphor with node labels in the centre of the links and numeric weights on the branches. Frequency information was visually represented by widening the pipes for larger values. While this was shown for joint probabilities, a similar approach could be used for conditional probability. Like the tree, the pipe diagram presents a subset of the full problem information, all of which is available directly in the double-tree.

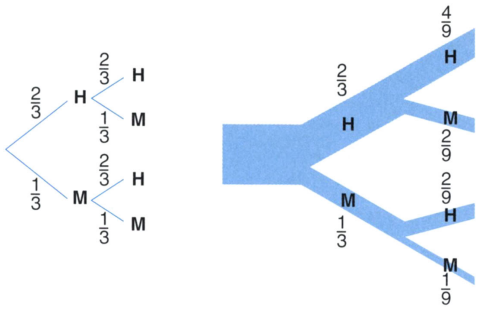


Figure 5: Comparison of tree (left) and pipe (right) diagrams, oriented left-to-right, from Konold (1996).

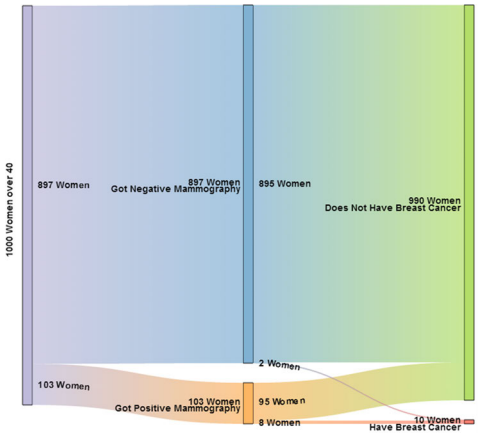


Figure 6: A Sankey diagram, proposed here as a previously untested visualization type for the Mammography Problem. The structure is similar to the Double-tree diagram and the frequency representation is similar to the Frequency Grid diagram.

Sankey Diagram (Figure 1d and Figure 6): The Sankey Diagram was developed to visually describe flows of quantities as they are transformed by a number of processes (Schmidt, 2008). Like the pipe diagram, the Sankey diagram conveys the scale of the values spatially with wider paths. However, this unique diagram type combines both double-branching and double-joining (both branches and joins in a left-to-right reading order) together with frequency information. It can therefore, when subsets are labelled, represent all of the information needed for solving a Bayesian problem removing the need for estimation or calculation. In terms of branching, the Sankey is equivalent to the double-tree, but oriented right-to-left instead of top-down.

To our knowledge, the Sankey Diagram has not previously been evaluated as a visualization of Bayesian problems. However, we propose that this hybrid rep-

resentation of both graphical branching structure and spatial frequency information, represented by the width of the branches, is a promising visualization type for Bayesian problems and was included in our experiment described later.

Hybrid-tree and Icon Array (Figure 1e and Figure 7): As mentioned above, poor accuracy in study results are often interpreted as participants misjudging scale, or base rate (prevalence) neglect (Cole, 1989). To better draw the subjects' attention to the prevalence rate, Spiegelhalter et al. (2011) presented a hybrid diagram combining the typical tree representation with large groups of icons explicitly showing the exact number of members of each subset, or each branch.

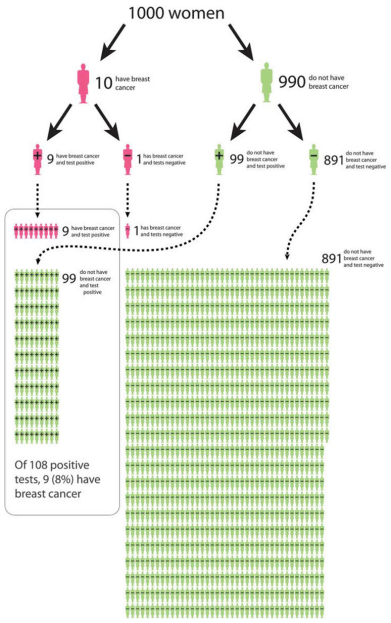


Figure 7: A hybrid-tree and icon array diagram, from Spiegelhalter et al. (2011).

While this diagram is visually pleasing, it mixes its representation of scale as both an icon array of human figures in some cases, and larger or smaller human figures in other cases. Descriptive text explains each subset but only one of the branches is rejoined resulting in a partial double-tree. The join is indicated by an outline around the group members, which could also be considered as containment, as in the nested style.

Frequency Grid (Figure 1f and Figure 8): The frequency grid, also known as an icon array, forgoes representations of structural information, such as branching or containment, in favour of a complete enumeration of every individual in the population where each one is visually associated with a specific subgroup, differentiated by icon fill and border colors. An accompanying legend provides the mapping between a visual element and its associated subgroup (Cole and Davidson, 1989; Sedlmeier and Gigerenzer, 2001; Dolan and Iadarola, 2008; Brase, 2009). As this type of visualization favors frequency completely, over any visual representation of branching or nested-set relations, we place this diagram type at a corner of the design space. Stone et al. (1997) showed that glyph styles do not effect performance and Müller et al. (2014) shows that glyphs can be used to encode a significant amount of information.

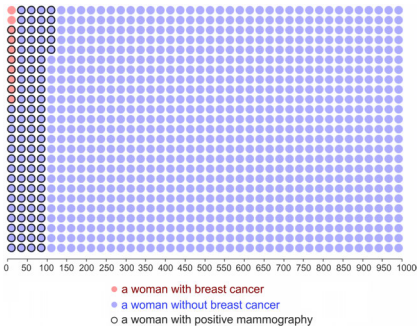


Figure 8: The frequency grid diagram, from Micallef et al. (2012).

Frequency Set Diagram (Figure 1g and Figure 9): This hybrid representation (Brase, 2009) combines the frequency grid with Venn-like outlined subsets, similar to the group of icons in Figure 1e. Labels with arrows identify the groups and their nested-set relations.

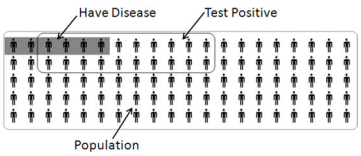


Figure 9: The frequency set diagram, from Brase (2009).

Area-Proportional Euler Diagram (Figure 1h and Figure 10): To help participants estimate relative set sizes, an area-proportional Euler diagram is used (Micallef et al., 2012). In this case, a hybrid Euler-frequency grid representation is used to reinforce the sense that each set contains a number of individuals represented by the icons. Also, this visual frequency information is organized spatially as in the Euler diagram conveying nested-set relations. Note that both a legend and set labels are used to convey set membership, but numeric values are not shown.

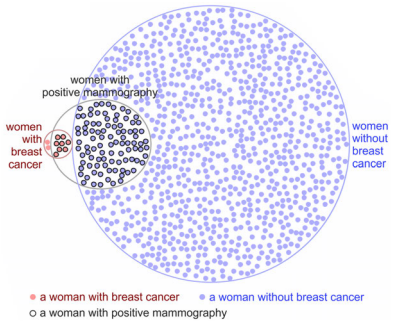


Figure 10: Area-proportional Euler diagram, from Micallef et al. (2012).

Euler Diagram (Figure 1i and Figure 11): In Brase (2009), a traditional Euler diagram is used. This representation is not area-proportional nor does it convey frequency in other ways. This type of diagram favours the representation of nested-set relations over specific quantities or branching structure and so, is placed at a corner of our design space.

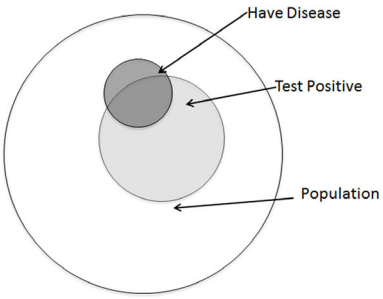


Figure 11: The Euler diagram, from Brase (2009).

Flowchart (Figure 1j and Figure 12): In surveying existing visualizations, we have presented representations that exclusively present branching, frequency, or nested-set relations information. While hybrid visualizations have been developed that mix branching with frequency (Sankey 1d), and frequency with nested-set relations (Area-proportional Euler 1h), no hybrid visualizations combining branching with nested-set relations has been developed to our knowledge. We therefore designed such a hybrid visualization of Bayesian problems that we call a Flowchart. This visualization includes the complete set of information available in the double-tree diagram but encloses subset regions in nested outlined areas and uses color and arrows to convey set membership. We also included this diagram type in our experiment described later.

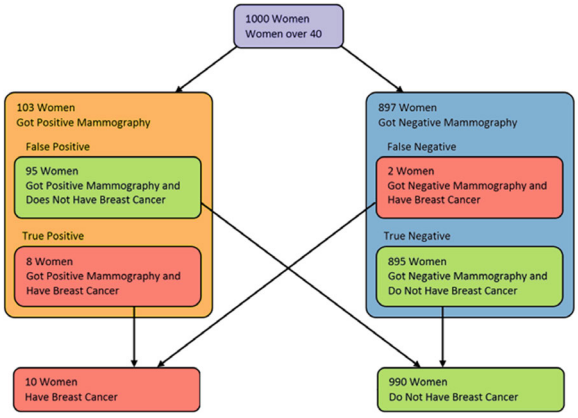


Figure 12: A novel Flowchart diagram, proposed here as a previously untested visualization type. This hybrid diagram emphasizes the Double-tree structure together with nested-set relations as in the Euler diagram.

3.2. Design Space of Visualizations

We organize these graphical representations into three non-exclusive categories: branching, frequency, and nested-set relations (see Figure 1). Branching structures (Figures 1a, 1b, 1c, 1d, 1e, 1j) can be used to convey how sets of values breakdown into their constituent parts or combine to form resultant sets. Frequency can be encoded numerically (Figures 1a, 1b, 1d, 1j), or graphically (Figures 1c, 1d). Also, frequency is emphasized in Figures 1e, 1f, 1g, 1h through the use of glyphs representing each individual in the set, to better convey how many people are in each set. Finally, nested-set relations are visually represented in Figures 1g, 1h, 1i, 1j through outlined overlapping regions.

Each of the three primary visualizations represent one dominant property. The Double-tree directly represents branching graphically. Indirectly, frequency information is conveyed numerically and nested-set structures may be inferred by understanding the parent-child relationships embedded in the double-tree. The Euler Diagram directly represents the nested-set relations of a Bayesian problem graphically. Indirectly, frequency information may be added by labels or in a legend. However, branching is not conveyed directly or indirectly. The Frequency Grid directly represents the number of individuals in each set graphically. Indirectly, nested-set relations may be inferred through color and glyph edge treatment. However, as with the Euler Diagram, branching is not conveyed at all.

Hybrid visualizations, that mix primary visualization types, are shown along the edges of the design space (see Figure 1). We introduce the use of a Sankey Diagram as a branching-with-frequency hybrid and we de-

veloped a novel Flowchart diagram as a nested-with-branching hybrid visualization. The existing Proportional Euler diagram conveys a nested-with-frequency hybrid visualization.

We next discuss the challenge of fairly comparing these visualizations, given their diversity of information content and design features, and we put forth principles to better judge the validity or compatibility of a given comparison.

4. Comparability Criteria

We seek to compare visualizations from our design space mentioned above to explore the role that major visual features may play in the performance of participants trying to solve the Mammography Problem. To guide the design of the visualizations being compared, we introduce three control measures to help ensure that experimental results reflect the visualization features unique to each condition. By increasing the number of visualization elements with a common visual treatment, and increasing the number of common informational elements, we hope to increase the relation between performance differences and the design sources of the differences.

We define three criteria to be met when comparing a set of visualizations: completeness of information, consistency of visual encoding, and consistency of presentation.

4.1. Completeness of Information

To compare two visualizations, they should contain the same level of information. If they do not, assuming that it is possible to reconstruct the absent data items, the missing information must be derived or calculated by the person using the visualization. This situation makes it difficult to compare the performance of the visualizations when one of them calls for estimation or calculation which could consume a significant amount of time from the person using the visualization. To avoid this confound, we analyse the information content of the visualizations to ensure that we are testing the visualization and not the ability of a participant to perform arithmetic operations.

Beginning with the textual description of the Mammography Problem, we see there are five numeric pieces of information given: 1000, 10, 8, 990, and 95. However, as shown in Table 1, four additional values must be derived to properly answer the possible probability questions. As shown in Figure 13, the tree visualization (Figure 1b) shows two additional values, 2 and 895, but is still missing two pieces of information. The Euler Diagram (Figure 1i) is also missing two pieces of information but shows 103 instead of 990.

Visualization	1000	10	8	990	95	2	895	897	103	
Double-tree	join	branch	branch	branch	branch	branch	branch	join	join	numeric
Flowchart	join	join	nested	join	nested	nested	nested	branch	branch	
Sankey	join	join	branch	join	branch	branch	branch	branch	branch	
Frequency	num	count	count	count	count	count	count	count	count	
Icon Euler	count	count	count	count	count	count	count	count	count	
Euler	num	num	num		num	num	num		num	
Tree	num	num	num	num	num	num	num			
Text-only	num	num	num	num	num					

Figure 13: Methods of representation used in several visualizations. ”num” indicates a numeric representation while ”count” indicates a set of icons/glyphs are used to represent the quantity. The top three rows are also numeric but are further sub-labelled as ”join” or ”branch” when they are in nodes of a tree structure, or as ”nested” when the numeric value is embedded in a larger shape.

Figure 13 summarizes which visualizations contain which key pieces of information for the Mammography Problem. The first five visualizations listed all contain the complete set of information. Double-tree, Flowchart, and Sankey all contain the numeric values directly embedded in the visualizations. The Frequency Grid and Area-Proportional Euler, with frequency icons, do not directly show the numeric values but show a one-to-one mapping of the values to circular icons. This makes it possible for a person using the visualization to count all the icons of a specific set or subset to obtain the numeric value. The one exception is that the Frequency Grid visualization contains a ruler indicating the size of the complete population (1000). As discussed in Section 5, these two visualizations could be extended with a legend containing the complete set of numeric values, similar to the content of Table 1, so that the visualizations are not testing the ability of the participants to count icons and may be more fairly compared to the other visualizations that meet the completeness of information criteria. Finally, as shown in Figure 13, the Double-tree, Flowchart, and Sankey all represent some form of branching structures that contain numeric values but they may be further differentiated by describing the branching or nesting property around the value. We have labeled the values as “join”, “branch”, or “nested” but other schemes could be used as well. This further highlights the unique aspects and differences in these visualizations even though they all contain the same numeric information. That is, these properties capture some of the information unique to each visualization.

4.2. Consistency of Visual Encoding

Numeric data can be visually encoded into graphical elements (Cleveland, 1994). For example, a series of values can be shown as a line graph, bar chart, or a number of points or glyphs. When comparing two visualizations, it may be difficult to attribute differences in performance to specific visual encodings. To reduce this difficulty, visual encodings should be made as consistent as possible between the visualizations so that any difference in performance may be clearly attributed to the small number of essential differences in visual encoding choices.

As shown in Figure 14, the visualization design space can be arranged by methods of visual encoding. For example, if we compare the Sankey Diagram and the Double-tree, both represent branching graphically and indicate nesting indirectly through the parent-child property inherent in branching structures. However, the Double-tree conveys frequency information numerically whereas the Sankey Diagram graphically indicates scale by having larger sets consume more area than smaller sets.

Visualization	Frequency	Branch	Nested
Frequency Grid	icons	n/a	graphical
Sankey Diagram	relative	graphical	parent-child
Double-tree	numerical	graphical	parent-child
Flowchart	numerical	graphical	graphical
Euler	n/a	n/a	graphical
Icon Euler	icons	n/a	graphical

Figure 14: Methods of visual encoding used in several visualizations organized by visual technique. The dominant visual encodings used by each visualization are shown on blue while indirect representations of a visualization property is shown on grey. Note hybrid visualizations combine two dominant techniques (on blue), e.g. Sankey, Flowchart, and Icon Euler.

4.3. Consistency of Presentation

Given two visualizations, a valid comparison is only possible if the visual presentation is consistent between them. Graphical elements, including icons, glyphs, colors, fonts, text sizes, and shapes, should be consistent across visualizations to minimize graphical variability. For example, this can help control against confounds arising from perception differences between users. Moreover, the users environment when viewing visualizations may be quite varied, and should be taken into account. For example, a layout displayed at one screen resolution may require no scrolling, while the same layout displayed at a different resolution may require scrolling. Breslav et al. (2014) showed that scrolling has a high cost in performance when answering the Mammography Problem. Efforts should be made to minimize these variations. Our techniques to better achieve consistency of presentation are described in Section 6.

4.4. Summary

We propose that comparability criteria be used throughout the study design process. We adopt this approach here to guide the visualization design choices for the experimental conditions, to support hypothesis generation to be based on the information content and visual encodings used, and to assist with interpreting experimental results by comparing differences in measures to differences noted in the designs. A comparability analysis can also be used to eliminate candidate visualizations when they are inherently incomplete, such as the Euler Diagram and the Tree, or to suggest the need for new visualization designs to complete a design space, as we do with the Flowchart and Sankey Diagram.

5. Designing Comparable Visualizations

Using the comparability criteria defined above, we identified candidate visualizations for our study from the spectrum of set-based, frequency-based, and branching-based visual representations. We ensured consistency of information by adding missing information where necessary. We increased the consistency of visual encoding by including explicit legends when counts were displayed non-numerically. We did not include legends in cases where counts were already represented numerically, as we consider these numbers to be an integrated, or implicit legend. Finally, we ensured consistency of presentation by using the same font family and size, and selected a qualitative colour palette from colorbrewer.org (Harrower and Brewer, 2003). In addition to visualizations, we designed a comparable non-visual representation of the complete set of information needed, provided as a text-only legend (see Section 5.2). The range of designs considered are summarized in Section 3, and the details of each visualization are discussed in the following sections.

5.1. Text-only

The text of the classic Mammography Problem, as given in Section 2, consists of two paragraphs; a problem statement and the problem question. This text is consistent between all of the experimental conditions that we use.

5.2. *Text-legend*

Notably, the text of the classic Mammography Problem does not actually meet our comparability criteria. To address this problem, we include all of the information contained in the Double-tree diagram as a set of facts, similar to an unlabelled legend (see Figure 15). This satisfies both the completeness of information criteria and eliminates the need for any type of arithmetic to be done by the participant.

While this can clearly help the participant in answering the question, the extra information and/or the way it is presented may also confuse the participant.

Got positive mammography (103 women)
Have breast cancer (10 women)
Have breast cancer & got positive mammography (8 women)
Have breast cancer & got negative mammography (2 women)
Got negative mammography (897 women)
Do not have breast cancer (990 women)
Do not have breast cancer & got positive mammography (95 women)
Do not have breast cancer & got negative mammography (895 women)

Figure 15: Unlabelled legend shown in Text-legend condition of our study to satisfy the completeness of information criteria.

5.3. *Area-Proportional Euler Diagram*

To fulfil the completeness of information criteria using the area-proportional Euler diagram (Micallef et al., 2012), we include the complete legend, as shown in Figure 15, but with icon tags to indicate set membership (see Figure 16). This expands upon the short legend (without numeric values) used in Micallef et al. (2012) (see Figure 1g). This approach also explicitly defines the sets so that in-place set labels are not needed. This may be helpful since the set label placement did not explicitly define which group of icons were included, leaving the meaning open to interpretation.

5.4. *Frequency Grid*

The Frequency Grid (Micallef et al., 2012) is similar to the Euler diagram with respect to the legend, and so, the same legend can be used in this case (see Figure 17). Again, the legend labels define the meaning of the sets. Note that the elements in the complete legend are logically arranged into two groups of four with disease indicators together with the disease status and their subcases. Again, while this design gives completeness of information, the complexity of the legend may have adverse effects on performance as well. Overall, this follows the recommendations of Breslav et al. (2014) to

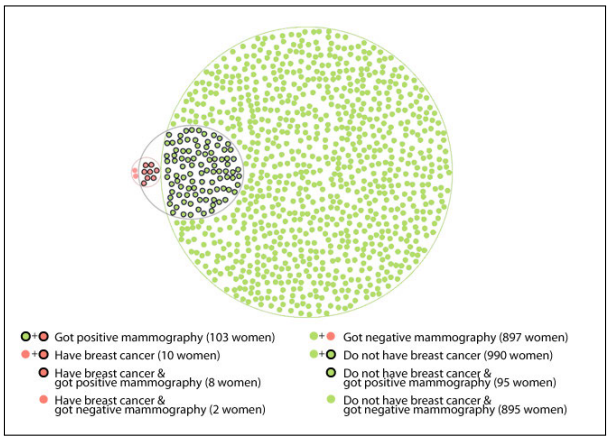


Figure 16: Area-proportional Euler diagram with complete legend.

avoid participants from counting the icons to help answer the question, which was a behaviour observed in both Micallef et al. (2012) and Breslav et al. (2014).

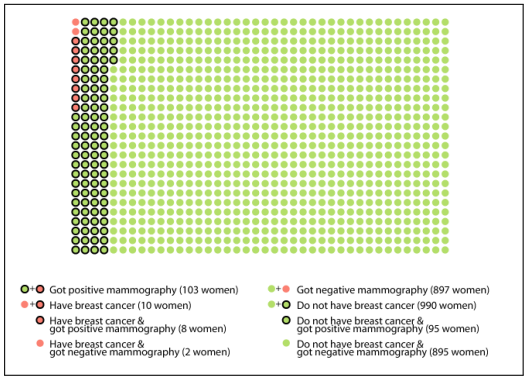


Figure 17: Frequency Grid with complete legend.

5.5. *Sankey Diagram*

In considering the primary visual features used in conditional probability problems, we found that the Sankey diagram (Schmidt, 2008) could convey both branching (as does the Double-tree) and frequency (with one-dimensional proportional branch thickness similar to the Pipes representation, see Figure 1c). As this diagram type does not lend itself to frequency represented as icons, labels are included together with frequency values, obviating the need for a legend (see Figure 6).

Note that in reading the Sankey diagram in the English left-to-right convention, this representation also conveys a temporal aspect where branches split or join to update, or re-categorize, individuals. In fact, Bayes theorem is typically described in temporal terms: updating our beliefs when given new information. Specifically, Bayes theorem expresses precisely how a prior

probability becomes a revised probability (or the posterior probability) when new evidence is considered. Despite the high compatibility of this visual representation with Bayes theorem, the authors are not aware of any previous studies using this diagram type.

As this is the initial evaluation of this diagram type, there may be many variations worthy of testing beyond its application to the Mammography problem at hand, including Bayesian problems with different levels of sensitivity, specificity, and prevalence.

5.6. Flowchart Hybrid

To explore the benefits of branching and nested-set representation together, we developed a Flowchart style hybrid diagram that explicitly shows the subsets for the test outcomes, positive mammography and negative mammography, including the Double-tree crossings (see Figure 12). This diagram supports the completeness of information criteria with all of the information implicitly embedded in the diagram. Also, some explanatory text is included within the positive and negative mammography groups indicating the meaning of the sub-cases as being false or true positives, or false or true negatives. However, it does differ slightly from the Double-tree diagram in that the joining of the bottom two nodes (Have Breast Cancer and Does Not have Breast Cancer) is omitted as it subjectively seemed to indicate that there was a second group of 1000 Women.

5.7. Double-tree

In augmenting the Double-tree design to meet the comparability criteria, we placed the legend labels and values inside the nodes of the tree (see Figure 18). We maintained the same colour scheme and, although fairly subtle, we used the node outline colour to convey the nested-set membership that is explicitly shown in the Flowchart hybrid and in the Sankey diagram color transitions.

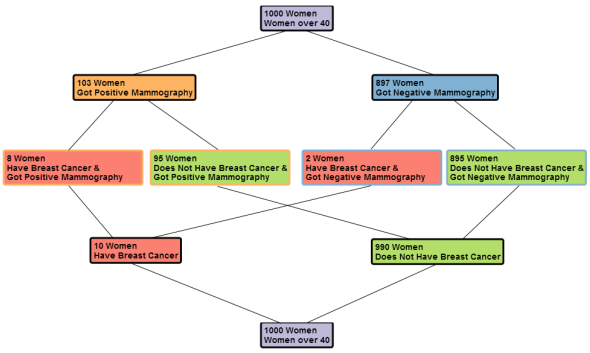


Figure 18: Double-tree emphasizing the problem structure.

6. Page Design

Crowdsourcing using Amazon Mechanical Turk (MTurk) has previously been used to test visualizations of Bayesian reasoning (Micallef et al., 2012; Ottley et al., 2012; Breslav et al., 2014). One advantage of crowdsourcing is the ability to easily perform online testing of design variations (Heer and Bostock, 2010). However, performing a crowdsourcing experiment also adds additional difficulty to meet our Consistency of Presentation criterion, since we cannot directly control the users experimental environment. Breslav et al. (2014) present a number of recommendations when designing webpages for use in crowdsourcing experiments. These recommendations consider the complications introduced by running an experiment remotely, such as varying user contexts, browsers, window sizes and screen resolutions. Below, we outline how we addressed each of these to meet our Consistency of Presentation criteria.

6.1. User Context and Browser

While we cannot control the users environment while carrying out the experiment, we can take steps to enforce base restrictions to maximize consistency. The user context is inferred from the browser reported User Agent, which contains information such as the platform and browser.

First, we exclude participants on mobile platforms from participating. This reduces the likelihood that participation will occur while in a mobile context, where attention may be less focused. Also, it reduces the variability of display resolution, which is addressed in a later section.

Second, we restrict browsers to recent versions of Chrome ($\geq$ 14), Firefox ($\geq$ 4), and Internet Explorer ($\geq$ 9). This simplifies addressing cross-browser scripting and layout compatibility issues. To test the success of these restrictions and cross-browser visual consistency, we use the BrowserStack.com web service which renders screenshots of a webpage using different platforms and browsers. While we have found some minor discrepancies in font sizes across browsers and platforms, none of the issues were significant enough to cause concern about the validity of the results.

6.2. Screen Resolution and Window Size

Screen resolution and window size vary tremendously between MTurk participants. As discovered by Breslav et al. (2014), the need for repeated scrolling can

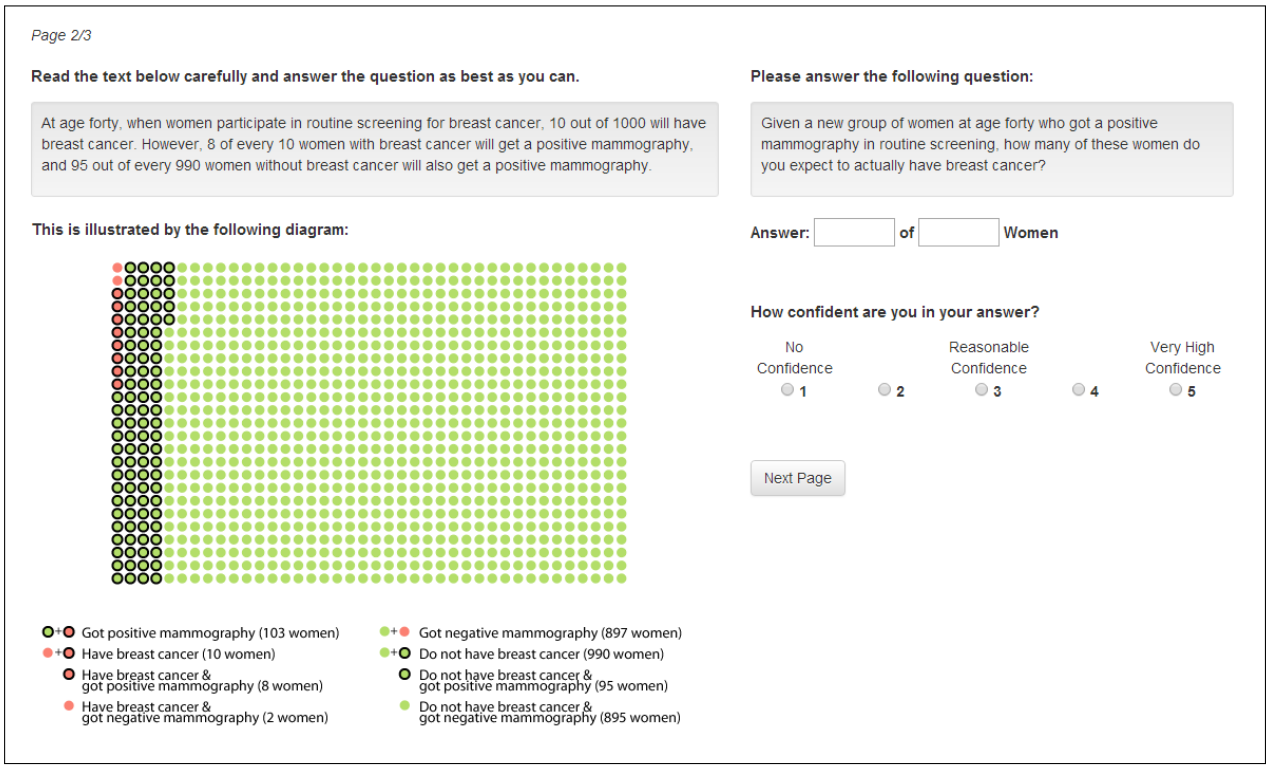


Figure 19: Two-column page layout of the experiment.

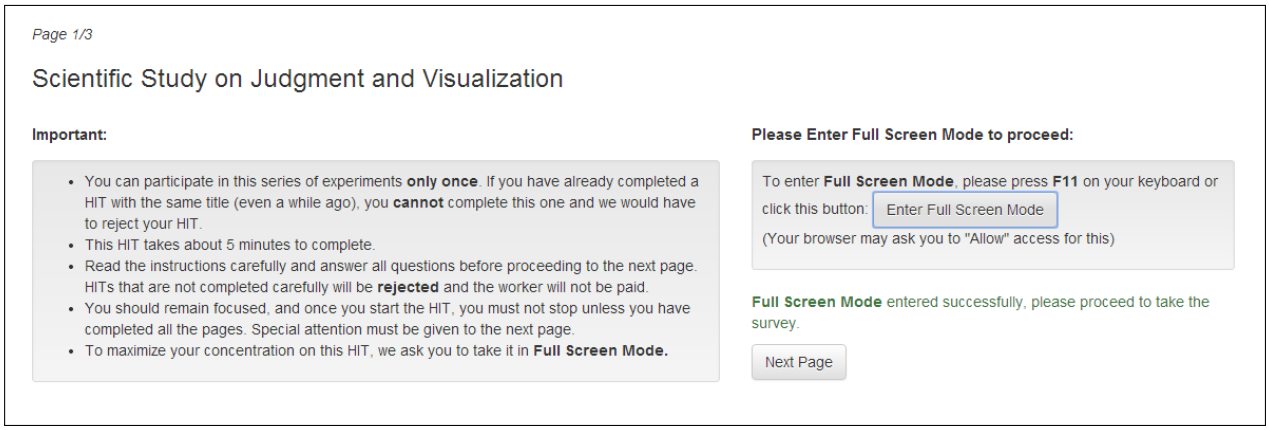


Figure 20: Instruction page of the survey for Full Screen mode.

dominate the participant input if the visualization displayed is too large or the web page design is not an appropriate shape. Thus, it was very important to us that we eliminate scrolling, minimizing split-attention as a confounding factor to performance.

We first ran a pilot study on MTurk with 404 users to sample the range of screen resolutions. The three most common screen resolutions were 1366x768 (132/404, 33%), 1920x1080 (39/404, 10%), 1280x800 (37/404, 9%) pixels. We also binned the observations by canonical ranges of screen width to adjust for aspect ratio differences. Only 2% (9/404) of respondents had a resolution of less than 1024px in width. A vast majority, 87% (352/404) had resolutions between 1024 and 1680 pixels in width, inclusive. The remaining 11% (43/404) of respondents had high-definitions displays of 1920px in width or greater. The vast majority of respondents used widescreen aspect ratios: 16:9 (243/404, 60%) and 16:10 (93/404, 23%). We did not observe any computers with portrait-oriented displays.

Taking these observations into account, we specifically design a page layout for landscape oriented widescreen displays with a minimum width of 1024 pixels. We excluded participants with resolutions less than 1024px in width to avoid designing another layout for 2% of potential participants. Since the most aggressive widescreen aspect ratio (16:9) can be displayed in less extreme (16:10) and non-widescreen aspect ratios (4:3 or 5:4) without clipping, we designed our page layout for an aspect ratio of 16:9.

Based on these observations, we designed a page layout for our experiment that eliminates the need to scroll. To ensure that windows were all the same size, and to minimize distraction from other open windows, we forced users to be in full-screen mode while participating in the experiment.

6.3. Page Layout

We used the Bootstrap (getbootstrap.com) front-end framework to build the webpages. Bootstrap is designed to automatically scale content appropriately given the platform and size of a browser window. This follows an ethos of responsive web design. Images can also be automatically scaled in a similar manner. However, this relies on the browser to resample images, and we have found the quality to be unreliable: sometimes introducing resampling artefacts that render text in the image illegible. Given the vast majority of resolutions in our pilot study fall between 1024 and 1680 pixels in width, we target two ranges of resolutions: small ($<$ 1280px) and large ($\geq$ 1280px). As it is not possible to determine the physical monitor size, our assumption is that larger monitors will have larger resolutions. Therefore, when also factoring in the widescreen aspect ratio, we used a small layout at 1000x560 pixels and a large layout at 1200x675 pixels. Bootstrap handles switching font and image sizes between page layouts based on the size of the browser window. This ensures that the page content will be legible and will likely be an appropriate size for the physical scale of the users monitor.

We use a two column layout to fit the text and visualization of the question on the same page, avoiding the need to scroll (see Figure 19). The problem statement and visualization are displayed in the left column, while the question and the input fields for the answer are displayed in the right column. Given the prevalence of widescreen displays seen in our pilot study, this layout choice maximizes use of screen space, making it possible to avoid scrolling. Our cross-browser BrowserStack.com tests showed that the page content fit within the window without the need for scrollbars and that content is legible across the sizing policy.

6.4. Full Screen Mode

To be able to consistently rely of the full use of the available screen space, Full Screen mode of the browser was enforced. An additional important justification for using Full Screen mode was to help participants to focus on the task, reducing distraction, since everything else is hidden. Note however that notifications (e.g., email popups) may still appear above the browser window.

To implement Full Screen mode, a participant is required to press the F11 key on their keyboard, or alternatively press a button on the initial test page (see Figure 20) which calls the browser JavaScript API to enter Full Screen mode. If the participant exits Full Screen mode in subsequent pages of the survey, a modal popup is shown, requiring the user to return back into Full Screen mode.

In some browsers, submission of the form automatically triggers the browser to exit Full Screen mode. To maintain a more seamless flow from page to page of the survey, we submit answers of each page using AJAX requests, which does pose a requirement on our participants to have a modern browser that supports Javascript and AJAX (Holzinger et al. (2010)). This approach helps ensure that Full Screen mode is maintained, since some browsers exit Full Screen mode on a browser refresh. In fact, the submission of each page is separated into two AJAX requests. We first synchronously submit just the answer to the question, validating the input before letting the participant proceed to the next page, and then asynchronously submitting all the detailed user interaction data (e.g., mouse movements, clicks, key presses) to a central database (Breslav et al., 2014), while the participant proceeds and begins to interact with the next page of the survey. This ensures short loading times between each page and an efficient use of participants time, since potentially large payloads of mouse movements are being uploaded from the previous page.

7. Experiment

We conducted a controlled experiment to evaluate user performance across five different visual representations, together with two textual representations, of the Mammography Problem. All of the designs are described in Section 5 and adhered to our proposed comparability criteria to help ensure that visual representations contain complete information and consistent visual encodings and presentation.

The experiment was carried out using MTurk, a crowd-sourcing service. Each MTurk worker completed

a single trial. In this trial, the participant was introduced to the Mammography Problem, shown one of our visual representations, asked to answer the question, and finally asked to rate their confidence in their answer. A final page captured demographics information. After the survey was completed, the participant received a code to submit to the MTurk website indicating that the trial had been completed and to receive credit for taking the survey. Participants were compensated $1.00 USD for their participation. The qualification requirement for the study included a Human Intelligence Tasks (HIT) Approval Rate greater than or equal to 95% and Number of HITs Approved greater than or equal to 50.

7.1. Design

Similar to previous work (Micallef et al., 2012; Breslav et al., 2014; Brase, 2009) our study used a between-subjects design. Each participant completed a single trial to control against learning effects and fatigue. The format and wording of the Mammography Problem is similar to previous work, and can be found in Section 2. As we were re-evaluating many common visualizations under novel conditions, we wanted a large number of participants to resolve a clear outcome. Therefore, we kept our survey active until at least 100 participants per condition have completed the survey. The survey was conducted for all seven conditions as described in the seven subsections of Section 4: Text-only, Text-legend, Area-proportional Euler Diagram (Euler), Frequency Grid, Sankey Diagram (Sankey), Flowchart Hybrid (Flowchart), and Double-tree. Each condition was run as a separate MTurk HIT. MTurk workers were restricted to only one trial and only one of the HITs. This was done by keeping track of all worker IDs (which can be extracted from the URL using javascript) that finish the survey and denying access to workers who have completed any previous HITs. Since Micallef et al. (2012) reported no effect of subject's numeracy and spatial abilities in answering various questions, we did not include any such tests (Ekstrom et al. (1976)) in our survey.

7.2. Hypotheses

With the guidance of our design space organization and our proposed comparability criteria, we hypothesize that:

H1 The Text-only condition, which requires computation, will have lower accuracy than conditions that do not require computation and only test comprehension of the problem.

H2 Visualizations with embedded numeric information (Double tree, Sankey, Flowchart) will outperform visualizations with explicit legends (Text-legend, Frequency grid, Euler) since embedding minimizes a split-attention effect (Chandler and Sweller, 1992).

H3 Hybrid visualizations (Sankey, Flowchart, Euler) will outperform other visualizations because they have two dominant visual encodings.

7.3. Variables

The two textual representations of the Mammography Problem (Text-only and Text-legend) together with the five visual representations (Euler, Frequency Grid, Sankey, Flowchart, and Double-tree) were included in the experiment. Thus, our independent variable is:

- VIZ $\in$ { Text-only, Text-legend, Frequency grid, Euler, Double tree, Sankey, Flowchart }

Our dependent variables were:

- EXACT $\in \{true, false\}$, when the numerator = 8 and the denominator = 103.
- $\text{EXACT}_N \in \{true, false\}$, when the numerator = 8.
- $\text{EXACT}_D \in \{true, false\}$, when the denominator = 103.
- BIAS, the difference between the subjects answer and the exact answer, computed as a log ratio: $log_{10}\left(\frac{entered\ answer}{correct\ answer}\right)$ (Micallef et al., 2012).
- ERROR, the absolute value of BIAS (Micallef et al., 2012).
- TIME, the time taken to solve the problem.
- CONF $\in$ [1..5], the subjects confidence in his/her answer. (Brase, 2009).

As suggested by Micallef et al. (2012), BIAS and ERROR was used as a more detailed metric of accuracy than just an exact answer count EXACT answer. The notion of an exact answer is included as it is possible for both the numerator and denominator to be incorrect but result in a correct probability when they are combined through division. A BIAS of 0.0 indicates the correct answer, a negative bias represents an underestimation of the answer, and a positive bias represents an overestimation. ERROR is calculated as the absolute value of the BIAS and provides an overall distance of the participants response to the correct answer and gives a more detailed metric of accuracy than just an exact answer count.

We also captured the following information: amount of training (prior experience with probabilistic problem solving on a 5-point Likert scale), the highest level of education that the participant has completed, colour blindness, and occupation (Standard Occupational Classification – Major Groups (U.S. Department of Labor, 2010) augmented with one additional group for "Student, Trainee"). The experiment was instrumented using the Mimic system (Breslav et al., 2014), a toolkit for capturing detailed interaction logs (e.g. mouse movements, mouse clicks, key presses, etc.) from web-based experiments. Specifically, in our analysis we used key press events and mouse movement events.

7.4. Participants

The participants consisted of 700 workers from the MTurk service. The majority of participants were male (429, 61%). Less than 3% of participants reported colour blindness (13, 2%) or reported not knowing (6, 1%). The majority of participants had completed a college level or higher education: undergraduate degree (358, 51%), graduate degree (69, 10%). The top five reported occupations were "unemployed, retired, homemaker" (133, 19%), "computer and mathematical occupations" (68, 10%), "student, trainee" (64, 9%), "arts, design, entertainment, sports, and media occupations" (57, 8%), and "education, training, and library occupations" (55, 7%). A majority of participants (547, 78%) reported little to no prior experience with probabilistic reasoning: "very little or none" (346, 49%), "a little" (201, 29%), "moderate" (117, 17%), "quite a bit", (29, 4%), and "a lot" (7, 1%). We found that the small layout was used by 8% (56) of participants.

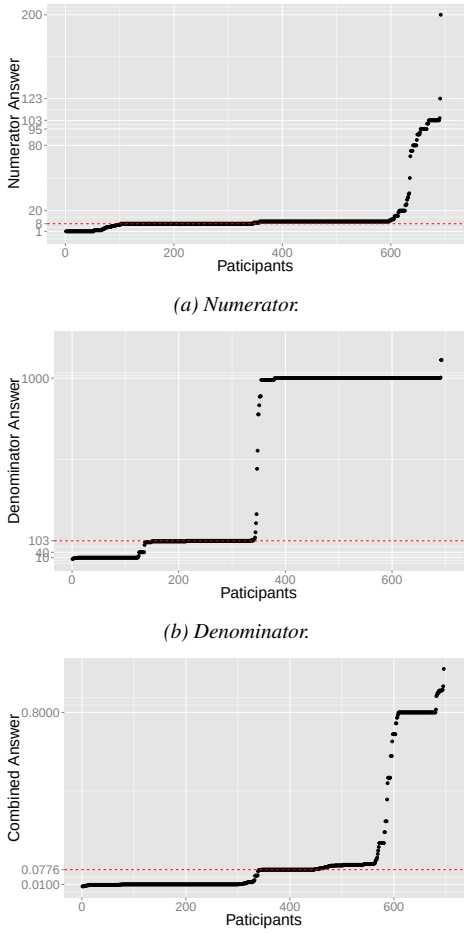
7.5. Data Validation

Since the study used crowdsourcing, we implemented rigorous data validation to ensure usable data was captured. We restricted input in answer fields to numeric values. Likert scales were represented by radio button fields. Prior to submitting the form, we checked that no empty values were being submitted. Any violation would result in a warning being presented to the user to ensure valid data was supplied. Although we implemented a number of data validation precautions, one user in the Text-legend condition submitted an answer of 0 for the denominator. Using the detailed input capture and review capabilities of Mimic (Breslav et al., 2014), we can see that it was likely an error when attempting to enter a denominator value of "1000" since the user actually typed in three zeroes, not just one. This record was excluded from the analysis, and so, the total number of participants used in the following section is $N = 699$.

8. Results

8.1. Answers

As can be seen in Figure 21, the top three answers for the numerator were 8 (245 participants, 35%), 10 (239 participants, 34%), and 1 (52 participants, 7%). The correct answer being 8 is the overall top choice. The top three answers for the denominator are 1000 (311 participants, 44%), 103 (120 participants, 17%), and 10 (113 participants, 16%). The correct answer being 103, appears significantly less frequently than a 1000.



(a) Numerator.

(b) Denominator.

(c) Combined probability.

Figure 21: For the Numerator, most participants entered either 8 or 10. For the Denominator, most entered 1000, 103, or 10. The resulting Combined probability is generally underestimated.

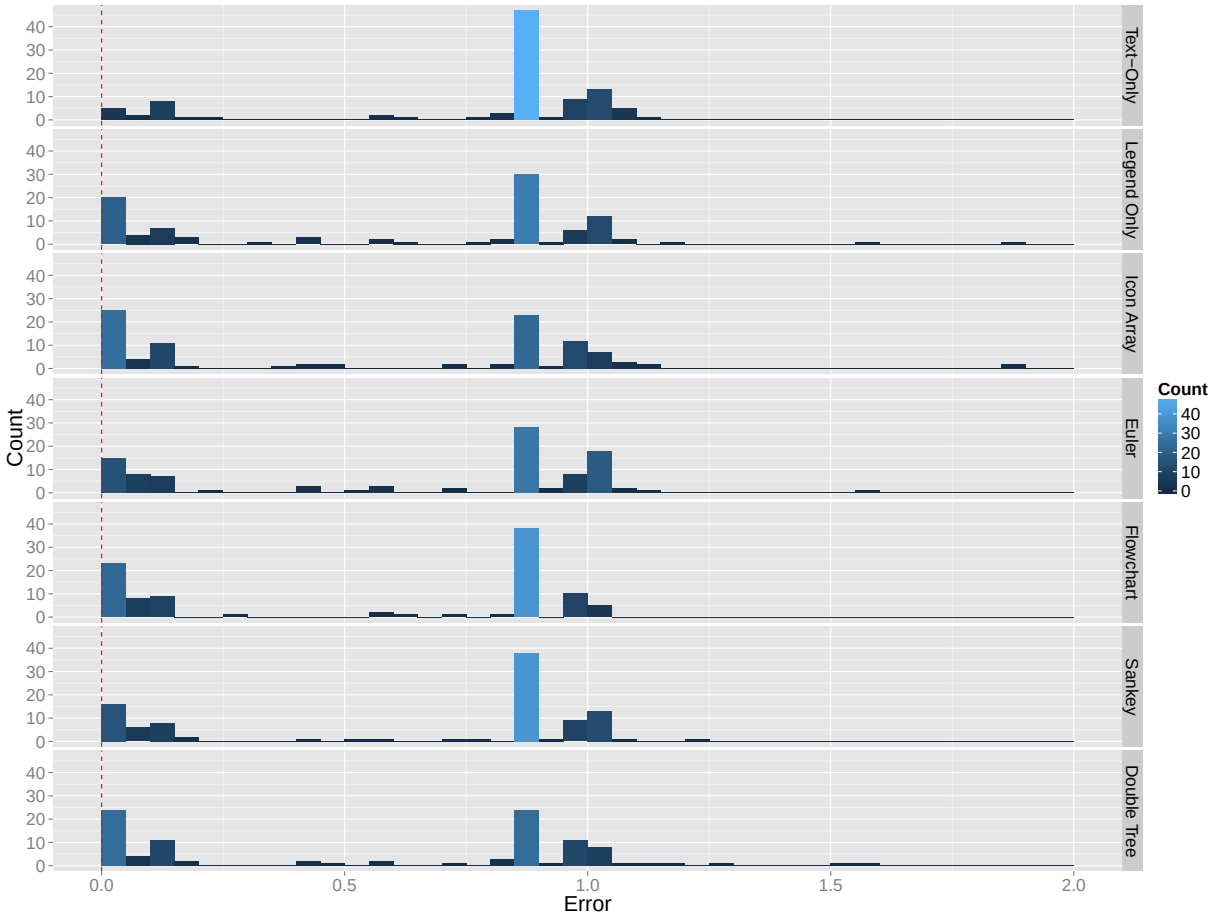


Figure 22: Error distributions for all the conditions shown with Histogram plot.

8.2. *Accuracy*

Exact answers are summarized in Figure 23. We can observe a significant difference between Text-only and the other conditions, with Text-only achieving 4% Exact answers, while others range from 10% to 20%. In all cases, low denominator performance (Exact_D) consistently lowered all scores. The difference between Exact_D and Exact suggests that there were occasions where participants correctly entered a denominator, but not the numerator. We can see that the Double-tree and the Frequency Grid achieved the highest Exact answer count with 20% exact answers, and 22% correct denominator. Flowchart and Text-Legend are not far behind with 18% and 17% of Exact answers. Our Euler condition uses the visualization from previous work (Micallef et al., 2012), who for a similar sample size (N=120 for Micallef et al. (2012), and N=100 in our case) only received 5% of Exact answers, likely due to the need for scrolling, while in our case, 13% of participants answered correctly in the Euler condition. However, Text-only performed similarly to previous work, 4% in our case, compared to 3.3% in Micallef et al. (2012). As the answers are not normally distributed, we use Kruskal-Wallis, a non-parametric one-way analysis of variance (ANOVA) test, to calculate differences between groups. There was a statistically significant difference between Exact answers across the conditions (Kruskal-Wallis, $H(6) = 17.03$, $p < 0.01$). Thus, we believe that the Completeness of Information criterion had a large effect on correct responses.

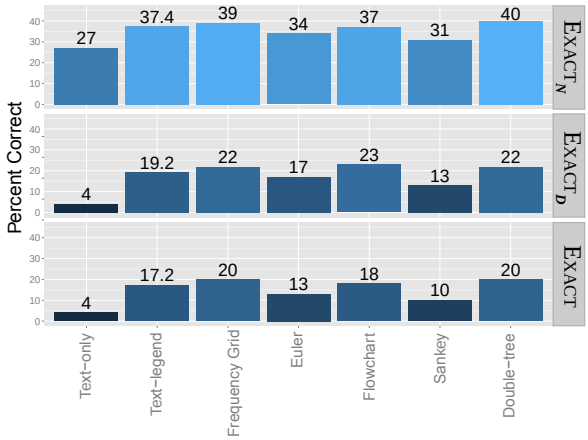


Figure 23: Exact answers (Numerator, Denominator, and cases where participants entered both the correct Numerator and the correct Denominator) for all conditions (%).

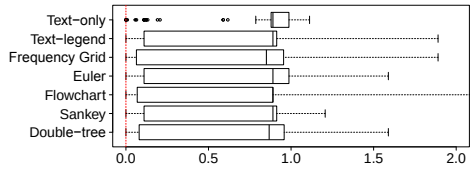


Figure 24: Error distributions for all the conditions shown with Boxplot graph.

A more detailed metric for measure of accuracy than just an exact answer count is the ERROR metric. As can be seen from the histogram in Figure 22 and boxplot in Figure 24, the median errors for all conditions are quite high. The overall median ERROR is 0.89 (M = 0.63, SD = 0.43) consistent with results in Micallef et al. (2012). The standard deviation for Text-only is low, 0.33 but is 0.45 for the other conditions. The differences between the conditions are statistically significant (Kruskal-Wallis, $H(6) = 19.16$, $p < .005$). However, without the Text-only condition there is no statistically significant difference (Kruskal-Wallis, $H(5) = 6.2$, $p = 0.29$). For further comparison of the Text-only condition with all of the visualization conditions, we also calculated effect size (Cohen's d) for the ERROR of each pair of conditions: Text-legend shows a small effect ($d = 0.36$), Frequency Grid shows a medium effect($d = 0.52$), Euler shows a small effect ($d = 0.35$), Flowchart shows a medium effect($d = 0.54$), Sankey shows a medium effect($d = 0.37$), and Double-tree shows a medium effect($d = 0.52$). Comparing Legend-only with other visualization conditions shows no effect (Frequency Grid $d = 0.13$, Euler $d = 0.03$, Flowchart $d = 0.14$, and Sankey $d = 0.02$). In summary, there is a difference between the Text-only condition and the other visualization conditions, but no difference between the Text-legend condition and the visualizations.

8.3. Bias

Systematic BIAS in the answers reveal misunderstandings or mistakes that are commonly made. Figure 26 shows a histogram of the distribution of the BIAS for each condition of the overall answer ratio. We can observe that the biases are not normally distributed. Also, we can see a stronger negative bias, indicating underestimation of the overall probability. This systematic error is consistent with previous findings (Micallef et al., 2012; Breslav et al., 2014). The median biases were -0.8903 (M = -0.35, SD = 0.77) for Text-only and -0.6 (M = -0.36, SD = 0.63) for Frequency Grid, while for other conditions medians are at zero. See Figure 25 for a Boxplot of the bias. The means were: Text-only (M = -0.35, SD= 0.77), Text-legend (M = -0.2, SD = 0.76), Frequency Grid (M = -0.22, SD = 0.7), Euler (M = -0.08, SD = 0.77), Flowchart (M = -0.36, SD = 0.63), Sankey (M = -0.19, SD = 0.74), Double-tree (M = -0.27, SD = 0.68). Similar to Breslav et al. (2014), there is no statistically significant difference between these biases across the different conditions (Kruskal-Wallis, $H(6) = 10.5$, $p = 0.1$). Looking at the effect sizes of BIAS of the Text-only and other conditions, tells a similar story, finding only small effect sizes (Text-legend $d = 0.18$, Frequency Grid $d = 0.17$, Euler $d = 0.36$, Flowchart $d = 0.01$, Sankey $d = 0.21$, and Double-tree $d = 0.11$).

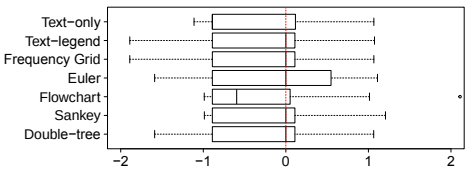


Figure 25: Bias boxplot indicating an overall answer underestimation.

8.4. Effect of Visualization Types

The H1 hypothesis, that the Text-only condition will result in the least accurate answers, was confirmed by statistically significant differences and medium effect sizes in the ERROR metric, but was not confirmed by the BIAS metric, which exhibited no statistically significant difference between conditions. Both H2 and H3 were not confirmed. The H2 hypothesis, that visualizations with embedded numeric information would outperform visualizations with explicit legend was not confirmed and there was no statistically significant difference between error of these groups (Kruskal-Wallis,

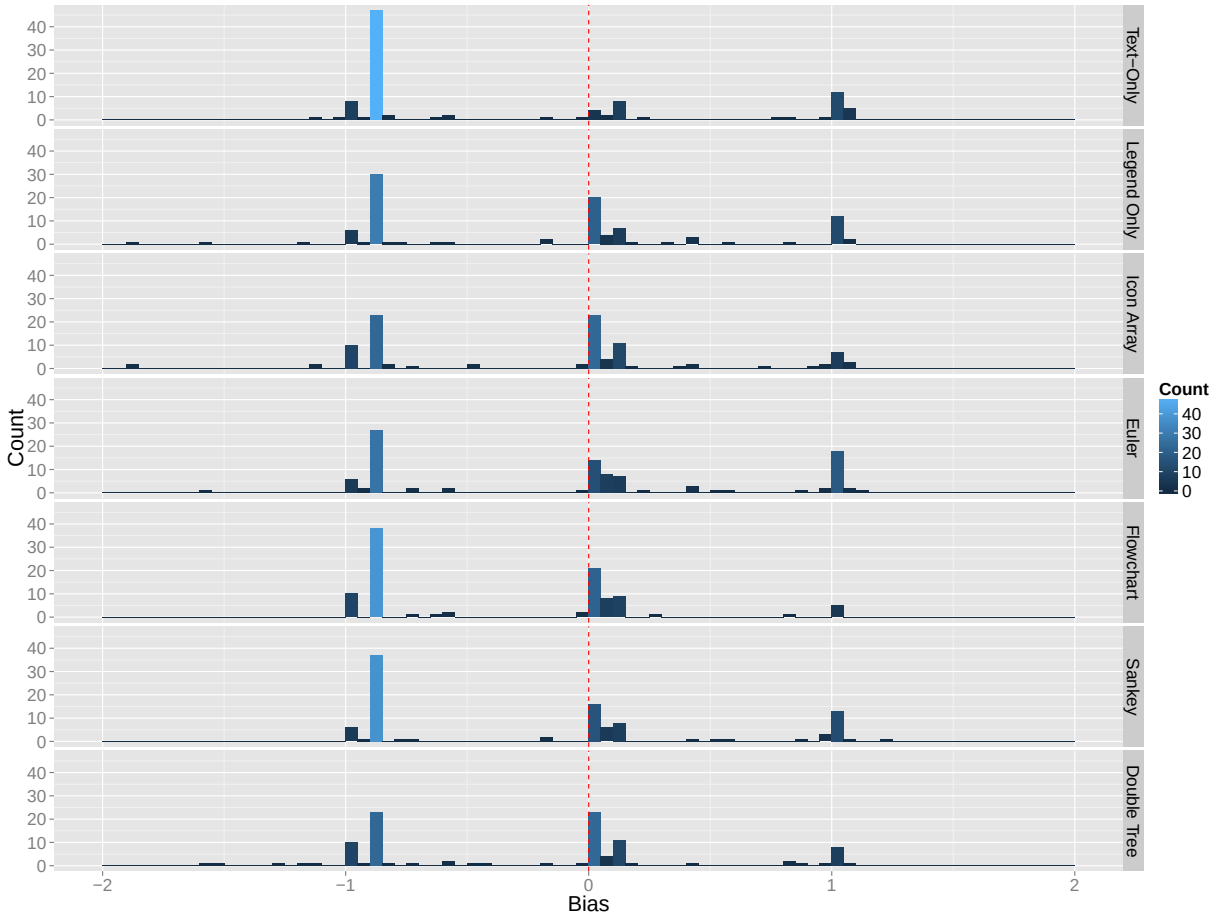


Figure 26: Histogram of the biases in the overall answer for each condition.

$H(1) = 0.9841, p = 0.3212$). The H3 hypotheses, that the hybrid visualizations (Flowchart, Sankey, Euler) would be the best performers, was also not confirmed. In terms of exact answers, the Frequency Grid and Double-tree designs were tied as the best visualizations. However, it should be noted that they were only marginally better than the Text-Legend condition.

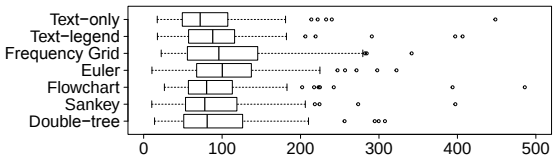
8.5. Confidence

The median for the Conf was 4 on a 5-point Likert scale ($M = 3.54$, $SD = 1.07$) indicating an overall high confidence in the answer. There is no statistically significant differences in reported confidence across the conditions (Kruskal-Wallis, $H(6) = 10.08$, $p = 0.12$), consistent with previous work (Brase, 2009). To find correlation between dependent variables, we calculate Spearman's rank correlation coefficient, a non-parametric measure of statistical dependence between two variables commonly used with ordinal data. There was no strong correlation between confidence and error (Spearmans rank, $r_s = -0.09$, $p = 0.02$), but a weak correlation between confidence and bias was noted (Spearmans rank, $r_s = -0.11$, $p < 0.005$).

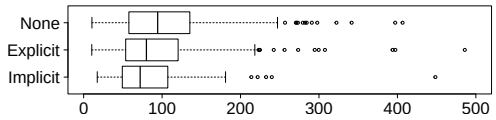
8.6. Time

The median completion time for the main mammography question was 83.94 seconds ($M = 100.4$, $SD = 68.64$) significantly lower than previous work (Micallef et al., 2012), where the median was 123 seconds. We did find some low correlation between time and error (Spearmans rank, $r_s = -0.15$, $p < 0.001$) and some correlation between Time and Exact answer (Spearmans rank, $r_s = 0.24$, $p < 0.001$). No significant differences between conditions were found (Kruskal-Wallis, $H(6) = 12.4$, $p = 0.05$). As seen in Figure 27, Frequency Grid and Euler took a bit longer than the other conditions, but not much. There was a statistically significant difference between legend types, (Kruskal-Wallis, $H(2) = 10.07$, $p < 0.01$), as can be seen in Figure 27b, where visualizations with implicit legend,

meaning numeric data was embedded in the visualization (Flowchart, Sankey,Double-tree), took less time.



(a) Per Condition.



(b) Per legend type.

Figure 27: Time spent on mammography question (seconds).

8.7. Training

Brase (2009) collected data about prior training in statistics as yes/no answers and excluded participants who reported that they had training. We collected more detail about training with a 5-point Likert scale. We did not exclude any participants from the analysis based on the prior training as there was no correlation between Error and Training (Spearmans rank, $r_s = -0.06$, $p = 0.11$) and a only weak correlation between Exact answer and training (Spearman's rank, $r_s = 0.107$, $p < 0.005$). However, we found that training is correlated to confidence (Spearman's rank, $r_s = 0.22$, $p < 0.0001$). Training also did not seem to relate to Time taken to answer the question (Spearmans rank, $r_s = -0.06$, $p = 0.132$).

8.8. Diagram Use

To convey a sense of the participants use of the visualizations, we used mouse movement data. For example, see Figure 28 and Figure 29 for the Euler condition with a Heat Map visualization of the mouse movements from all the participants (Figure 28) and the mouse movements from participants who answered the question correctly (Figure 29). To get a more quantitative comparison of the diagram use for different visualizations, we counted the number of mouse events (see Figure 30) that occurred within the HTML containing the visualization. We have observed that the Sankey Diagram stands out as being used less than other visualizations, however the difference is not statistically significant (Kruskal-Wallis, $H(5) = 9.371$, $p = 0.095$). In general, there was a weak correlation between diagram use and having a correct answer (Spearman's rank, $r_s = 0.11$, $p < 0.01$).

On the qualitative side of the analysis, examining the heat-maps confirms that the page design successfully prevented excessive scrolling, as most of the mouse movements are centered on the document. Also, the heat-maps reveal that people are counting elements in the visualizations, for example in the Figure 29 you can see a cluster of mouse movements around some groups of the Euler diagram. Another observation drawn from Figure 29 is that participants who got the question correct spent more time examining the correct numbers in the legend of the diagram.

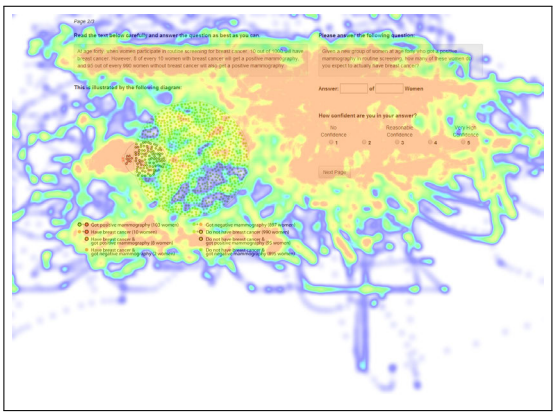


Figure 28: Heat Map Visualization of the Area-proportional Euler condition. All answers.

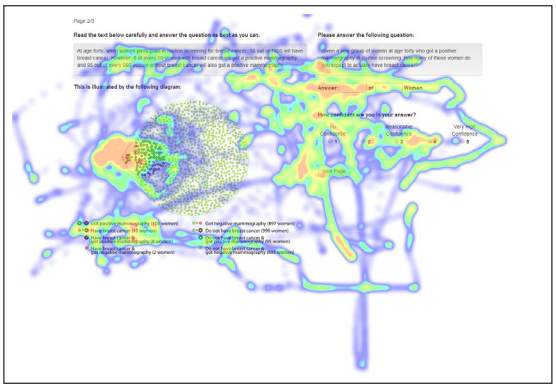


Figure 29: Heat Map Visualization of the Euler-proportional condition. Correct answers.

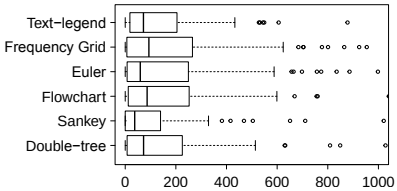


Figure 30: Diagram Use Boxplot (Mouse Move Events).

8.9. Comments

On the last page of the survey, participants were asked to "Please provide any comments or concerns." These comments were coded by one of the authors to identify general themes of responses. See the supplementary material for a full list of comments and tags added, a summary of which can be seen in the Tag Cloud in Figure 31. The most common tags used during the codification process were "none" (57, 23%), "cognitive process" (37, 15%), "thank you" (27, 11%), "praise" (21, 8%), "confused" (21, 8%), "graph" (20, 8%), "question" (9, 4%), "personal" (7, 3%), "interesting" (7, 3%), and "easy" (7, 3%). A number of people reported using the visualization within which a few commented on the experience (e.g. "This chart was incredibly confusing," referring to the Sankey Diagram). Since areas of confusion were of interest, co-occurrence of other tags with the tag "confused" was counted. We found that five comments with the tag "confused" did not specify a reason, nine comments specified the reason to be the question, five comments attributed their confusion to the graph, and two suggested that the medical terms caused confusion.

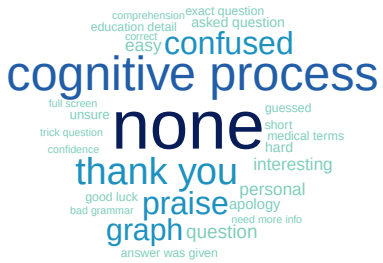


Figure 31: Tags used during codification of the participant's comments. The size and color intensity corresponds to frequency of the tag.

8.10. Effect Size

Several of our conditions (Text-only, Frequency Grid, Euler) are similar to those in the experiments of Micallef et al. (2012) Thus, it makes sense to compare the effects of visualization that we see to those that Micallef et al. (2012) saw. Hornbæ k et al. (2014) also recommended effect size comparisons as a way to make comparisons across replications more content rich. To compare the effect sizes, we first calculated an effect size of the manipulations in Micallef et al. (2012) from the available log files, finding a small effect $d = 0.11$ for the frequency grid (condition V3 in Micallef et al. (2012) experiment 1) and a small effect $d = 0.1$ for the Euler array (condition V4 in Micallef et al. (2012) experiment 2) over the text only condition (Micallef et al. (2012) V0). In our study, we find a medium effect size of $d = 0.5$ (frequency grid vs. text-only) and a small effect of Euler ($d = 0.33$) over the text-only condition. Both effects are much higher than the earlier study.

9. Discussion

The most interesting effect is the difference in performance in text-only and text-legend conditions. This highlights that it is not the representation of the problem as a visualization or as plain-text, but rather the information contained therein that has the largest effect on performance. This speaks to the information completeness criteria which, in our study, obviated the need for estimation or calculation.

We speculate the poor performance of participants in the text-only condition reflects that participants had difficulty extrapolating all of the necessary information to correctly answer the problem. This is similar to the text-only performance reported in Micallef et al. (2012) Participants performed markedly better using our visualizations designed with the information completeness criteria.

Surprisingly, performance with the text-legend condition did not significantly differ from any of the visualizations. In addition, the underlying representation of the visualization also had little effect on performance. This suggests to us that differences in structure alone do not help participants better understand the information provided. Our top performing conditions included explicit and implicit legends, and employed styles including frequency, branching, branching + nested-sets, and none. This suggests that no single stylistic approach provides a clear benefit.

The majority of prior work in Bayesian visualization has focussed on evaluating different underlying representations. Based on the results of the current work, we believe this structure plays less of a role than providing sufficient information to the user. Nevertheless, future work is still needed to explore the effect of different visualization properties. For example, one interesting property of visualizations such as the Euler Diagram or the Frequency Grid is that they do not emphasize a particular direction of reasoning, while the Flowchart or Sankey Diagram are meant to be read top-to-bottom or left-to-right. To examine such effects, future studies could try asking multiple Bayesian questions for a given visualization, requiring participants to find multiple different probabilities, such as $P(A|B)$ and $P(B|A)$. A study employing eye-tracking could also indicate how visualizations are read and if visualization orientation may relate to performance. Also, further work could explore

whether the addition of an explicit legend would benefit visualizations that already have implicit legend information.

Further, we suggest future research focus on methods of helping users to interactively navigate the space of possible solutions, thereby supporting users directly in understanding how the data values interact and which values are relevant to the problem question. For example, participants showed great difficulty in answering the denominator correctly which suggests that they were unable to select an appropriate subset of the population. While such interactivity and animation is a very promising direction for this work, many new factors of difficulty could be introduced by the extension of the static graphs, (Tversky et al., 2002), and so should be considered carefully.

Overall, 20% is still a very low score and further research is needed to determine why 80% of people fail to find the correct answer for the Mammography Problem. We found that recording the micro-interactions of all the trials to be extremely helpful to us. In watching many hours of playback of the user sessions, we feel that participants were genuinely trying to answer the question but are missing several insights. First, as can be seen in Figure 23, approximately half of the participants who achieved the correct numerator failed to select the correct denominator, instead, often choosing the entire population (1000 women). While this may be an indication of base-rate neglect, as has been proposed by others, by looking directly at the answers, we believe the denominator issue may indicate ambiguity as the issue. That is, the question "out of" may be interpreted as asking for the superset, namely the total number of people in the population. If the question was interpreted in this way, then many of the participants who answered "1000" believed that they answered correctly, not that they misunderstood the question or neglected the base rate. Fixing this issue would still only bring us to 40% indicating that a significantly different approach is still needed. We believe that experiential learning is needed through interactivity to help the majority of people to be able to correctly answer the Mammography Problem.

10. Conclusion

Bayesian inference is an important yet difficult part of critical decision making. Visualizations may help people to apply Bayesian inference to solve problems involving uncertainty but previous work has not revealed insights into why some visualizations are more or less helpful than others. We propose that they have not closely examined the relation between the structure and information content of Bayesian problems and the visual encoding and information available in the graphical representations used.

We proposed comparability criteria to help organize a Bayesian problem visualization design space as a methodology to develop specific visualizations that represent the key features of Bayesian problems, that is, frequency, nested-set relations, and branching. After reviewing a broad set of existing visualizations that have been studied for Bayesian problems, we proposed that hybrid visualizations, that convey two of the three key features, may be more effective than primary visualizations that only convey one key feature.

To complete the visualization design space, we proposed two additional hybrid visualizations: the Sankey Diagram and a novel Flowchart design. We ran a crowdsourcing experiment with 700 people testing primary and hybrid visualizations as well as text-only conditions with and without complete information. Going beyond aggregate statistics, we recorded detailed micro-interactions from all participants. This revealed new insights based on examining the numerator and denominator separately, specifically that the textual formulation of the request for an answer in a generic frequency format (e.g., "X out of Y") may be too ambiguous leading to a superset answer for the denominator. This indication differs from the results interpretation in previous work where base-rate neglect was blamed as the cause of poor performance of participants. The most surprising outcome was that controlling for the completeness of information, that is, by ensuring that all information was directly available in all conditions (except text-only), did not lead to greatly higher performance. Moreover, three visualizations performed slightly better than this baseline while two hybrid visualizations were worse. As our analysis of the interaction data reveals that a good degree of attention was paid to the visualizations presented, the need for some further investigation of visual complexity and cognitive load may be indicated.

In summary, we improved the overall performance in solving the Mammography Problem using Mechanical Turk compared to previous work but not by a large margin. By providing complete information in each condition, and by designing the page layout to avoid scrolling, significant improvements in accuracy were measured. Furthermore, by adhering to our proposed completeness of information criteria, we avoided the need for either estimation or calculation to answer the question, in all conditions (except Text-only). These methodological and design improvements also significantly reduced the time taken to answer. Having tested a full design

space, it is surprising that significantly different visualization designs did not perform very differently from each other. While these considerations and criteria were developed for our specific study, we believe they apply more broadly to crowdsourcing experiments and visualization comparison in general.

Given the critical use of Bayesian reasoning in both personal and professional decision making under uncertainty, improvement in performance is still needed. However, our results suggest that major improvements will not come from better static visualizations. We therefore look to problem expression and animated visualizations for performance improvements in future work.

11. Acknowledgements

This work has been supported in part by the Danish Council for Strategic Research, grant 10-092316.

References

Brase, G. L., Apr. 2009. Pictorial representations in statistical reasoning. Applied Cognitive Psychology 23 (3), 369–381.
URL http://doi.wiley.com/10.1002/acp.1460

Breslav, S., Khan, A., Hornbæ k, K., 2014. Mimic: visual analytics of online micro-interactions. In: Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces - AVI '14. ACM Press, New York, New York, USA, pp. 245–252.
URL http://dl.acm.org/citation.cfm?doid=2598153.2598168

Casscells, W., Schoenberger, A., Graboys, T. B., Nov. 1978. Interpretation by physicians of clinical laboratory results. The New England journal of medicine 299 (18), 999–1001.
URL http://www.ncbi.nlm.nih.gov/pubmed/692627

Chandler, P., Sweller, J., Jun. 1992. The split-attention effect as a factor in the design of instruction. British Journal of Educational Psychology 62 (2), 233–246.
URL http://doi.wiley.com/10.1111/j.2044-8279.1992.tb01017.xhttp://onlinelibrary.wiley.com/doi/10.1111/j.2044-8279.1992.tb01017.x/full

Cleveland, W. S., 1994. The Elements of Graphing Data. AT&T Bell Laboratories.
URL http://books.google.com/books/about/The_Elements_of_Graphing_Data.html?id=KMsZAQAAIAAJ&pgis=1

Cole, W., Davidson, J., 1989. Graphic Representation Can Lead To Fast and Accurate Bayesian Reasoning. Symposium on Computer Application in Medical Care, 227–231.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245778/

Cole, W. G., Mar. 1989. Understanding Bayesian reasoning via graphical displays. In: Proceedings of the SIGCHI conference on on Human Factors in Computing Systems. ACM Press, New York, New York, USA, pp. 381–386.
URL http://portal.acm.org/citation.cfm?doid=67449.67522

Dolan, J. G., Iadarola, S., Jan. 2008. Risk communication formats for low probability events: an exploratory study of patient preferences. BMC medical informatics and decision making 8, 14.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2330036/

Eddy, D., 1982. Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities. Judgment Under Uncertainty: Heuristics and Biases, 249–267.
URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Probabilistic+reasoning+in+clinical+medicine:+Problems+and+opportunities#0http://philpapers.org/rec/EDDPRI

Ekstrom, R. B., French, J. W., Harman, H. H., 1976. Manual for Kit of Factor-Referenced Cognitive Tests (August).

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., Woloshin, S., Nov. 2007. Helping Doctors and Patients Make Sense of Health Statistics. Psychological Science in the Public Interest 8 (2), 53–96.
URL http://psi.sagepub.com/content/8/2/53.short

Gigerenzer, G., Hoffrage, U., 1995. How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review 102 (4), 684–704.
URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.102.4.684

Harrower, M., Brewer, C. A., Jun. 2003. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. The Cartographic Journal 40 (1), 27–37.
URL http://www.maneyonline.com/doi/abs/10.1179/000870403235002042

Heer, J., Bostock, M., 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In: Proceedings of the 28th international conference on Human factors in computing systems - CHI '10. ACM Press, New York, New York, USA, p. 203.
URL http://dl.acm.org/citation.cfm?id=1753357

Holzinger, A., Mayr, S., Slany, W., Debevc, M., 2010. The influence of AJAX on Web usability. In: e-Business (ICE-B), Proceedings of the 2010 International Conference on. IEEE, pp. 1–4.

Hornbæ k, K., Sander, S. r. S., Bargas-Avila, J. A., Grue Simonsen, J., 2014. Is once enough? In: Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14. ACM Press, New York, New York, USA, pp. 3523–3532.
URL http://dl.acm.org/citation.cfm?doid=2556288.2557004

Konold, C., 1996. Representing Probabilities with Pipe Diagrams. The Mathematics Teacher 89 (5), 378–382.
URL http://www.jstor.org/stable/27969794

Kurz-Milcke, E., Gigerenzer, G., Martignon, L., Apr. 2008. Transparency in risk communication: graphical and analog tools. Annals of the New York Academy of Sciences 1128, 18–28.
URL http://www.ncbi.nlm.nih.gov/pubmed/18469211

Micallef, L., Dragicevic, P., Fekete, J.-D., 2012. Assessing the Effect of Visualizations on Bayesian Reasoning through Crowdsourcing. IEEE Transactions on Visualization and Computer Graphics 18 (12), 2536–2545.
URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6327259

Müller, H., Reihs, R., Zatloukal, K., Holzinger, A., 2014. Analysis of biomedical data with multilevel glyphs. BMC Bioinformatics 15 Suppl 6 (Suppl 6), S5–S5.
URL http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=25079119&site=ehost-live

Ottley, A., Metevier, B., Han, P., Chang, R., 2012. Visually Communicating Bayesian Statistics to Laypersons, 1–11.
URL http://www.cs.tufts.edu/tech_reports/reports/

2012-02/report.pdfhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Visually+Communicating+Bayesian+Statistics+to+Laypersons#0

Schapira, M. M., Nattinger, A. B., McHorney, C. A., 2008. Frequency or probability? A qualitative study of risk communication formats used in health care. Medical decision making : an international journal of the Society for Medical Decision Making 21 (6), 459–67.
URL http://www.ncbi.nlm.nih.gov/pubmed/11760103

Schmidt, M., Feb. 2008. The Sankey Diagram in Energy and Material Flow Management. Journal of Industrial Ecology 12 (1), 82–94.
URL http://doi.wiley.com/10.1111/j.1530-9290.2008.00004.x

Sedlmeier, P., 1997. BasicBayes: A tutor system for simple Bayesian inference. Behavior Research Methods, Instruments, & Computers 29 (3), 328–336.

Sedlmeier, P., Gigerenzer, G., Sep. 2001. Teaching Bayesian reasoning in less than two hours. Journal of experimental psychology. General 130 (3), 380–400.
URL http://www.ncbi.nlm.nih.gov/pubmed/11561916

Segel, E., Heer, J., 2010. Narrative Visualization : Telling Stories with Data (March).

Spiegelhalter, D., Pearson, M., Short, I., Sep. 2011. Visualizing uncertainty about the future. Science (New York, N.Y.) 333 (6048), 1393–400.
URL http://www.ncbi.nlm.nih.gov/pubmed/21903802

Stone, E. R., Yates, J. F., Parker, A. M., 1997. Effects of numerical and graphical displays on professed risk-taking behavior. Journal of Experimental Psychology: Applied 4, 243.

Turkay, C., Jeanquartier, F., Holzinger, A., Hauser, H., 2014. On Computationally-Enhanced Visual Analysis of Heterogeneous Data and Its Application in Biomedical Informatics. Interactive Knowledge Discovery and Data Mining in Biomedical Informatics (Lecture Notes in Computer Science), 117–140.

Tversky, B., Morrison, J. B., Betrancourt, M., 2002. Animation : can it facilitate ? International journal of human-computer studies, 247–262.

U.S. Department of Labor, 2010. Standard Occupational Classification. No. January 2009. Bureau of Labor Statistics, Standard Occupational Classification Policy Committee.

Wassner, C., Martignon, L., Biehler, R., 2004. Bayesianisches Denken in der Schule. Unterrichtswissenschaft 32 (1), 58–96.
URL http://www.pedocs.de/volltexte/2013/5808/

Wong, B. W., Xu, K., Holzinger, A., 2011. Interactive visualization for information analysis in medical diagnosis., 109–120.
URL http://eprints.mdx.ac.uk/8410/