Förderung Bayesianischen Denkens

Effekte verschiedener Baumdiagramme in unterschiedlichen Bayesianischen Situationen

Dissertation

zur Erlangung des akademischen Grades einer Doktorin der Didaktik der Naturwissenschaften „Dr. phil. nat." (doctor philosophiae naturalis) im Promotionsfach Didaktik der Mathematik der Fakultät Mathematik der Universität Regensburg

Vorgelegt von

Karin Binder

geboren in Ingolstadt

Einreichung
2018

Erstgutachter:	Prof. Dr. Stefan Krauss
Zweitgutachter:	Prof. Dr. Oliver Tepner
Drittgutachter:	Prof. Dr. Gerd Gigerenzer

Tag der mündlichen Prüfung: 19. Oktober 2018


Visualisierung Grafische Darstellung
Mammographie-Problem Vierfeldertafel
Baumdiagramm Stochastik Bayes Natürliche Häufigkeiten
Prozent Formel von Bayes Grafische Darstellung
Kognitive Illusion Häufigkeitsbaum Vierfeldertafel
Häufigkeitsbaum Prozent Visualisierung Prozent Formel von Bayes Stochastik
Natürliche Häufigkeiten Baumdiagramm Relative Häufigkeiten Prozent Kog
Mammographie-Problem Vierfeldertafel Häufigkeitsbaum Baumdiagramm
Grafische Darstellung Natürliche Häufigkeiten Prozent Wahrsche
Visualisier
Formel von Bayes Wahrscheinlichkeiten Bayes Stochastik Mammogr
Kognitive Illusion Grafische Darstellung
Natürliche Häufigkeiten
Häufigkeitsbaum

„Solving a problem simply means representing it so as to make the solution transparent."

Herbert A. Simon

Danksagung

[Die Danksagung wurde aus Gründen des Datenschutzes entfernt.]

Inhaltsverzeichnis

Liste der in der Dissertation zusammengefassten Publikationen[1]

Artikel 1

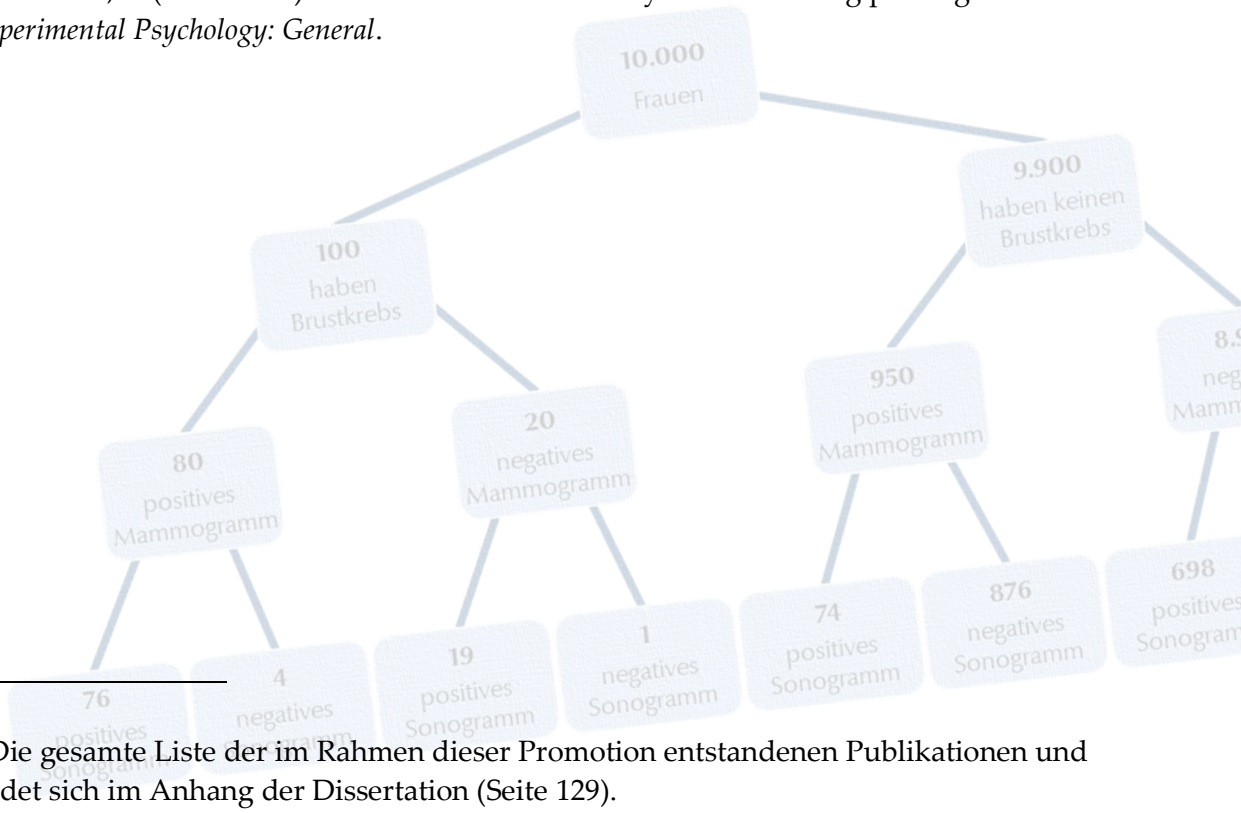
Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information – An empirical study on tree diagrams and 2 x 2 tables. *Frontiers in Psychology*, *6*(1186).
DOI: 10.3389/fpsyg.2015.01186

Artikel 2

Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (2018). Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making. *PlosONE*, *13*(3).
DOI: 10.1371/journal.pone.0195029.

Artikel 3

Binder, K., & Krauss, S. (submitted). Generalizations of the Bayesian reasoning paradigm. *Journal of Experimental Psychology: General*.



[1] **Hinweis:** Die gesamte Liste der im Rahmen dieser Promotion entstandenen Publikationen und Vorträge findet sich im Anhang der Dissertation (Seite 129).

Zusammenfassung

Das fehlerhafte Verknüpfen oder Interpretieren statistischer Informationen kann in der Medizin zu Überdiagnosen oder Überbehandlungen führen, schlimmstenfalls sogar zu Suizid, wenn einem positiven Testergebnis, das eine schwere Erkrankung indiziert, zu großes Vertrauen geschenkt wird. In der vorliegenden kumulativen Dissertation sollen Strategien analysiert werden, die Menschen helfen, bedingte Wahrscheinlichkeiten in Bayesianischen Aufgabenstellungen zu verstehen: 1. Natürliche Häufigkeiten und 2. Visualisierung mit Hilfe von Baumdiagrammen (und Vierfeldertafeln). Im ersten Artikel wird eine Studie mit 259 Schülerinnen und Schülern vorgestellt, in der typische schulische Visualisierungen Bayesianischer Aufgaben untersucht werden, nämlich Vierfeldertafeln und Baumdiagramme, die beide mit Wahrscheinlichkeiten oder natürlichen Häufigkeiten ausgefüllt werden können. Es zeigte sich, dass maximal 10% der Schülerinnen und Schüler der 11. Klassen in der Lage waren, Bayesianische Aufgaben korrekt zu lösen, wenn diese mit Wahrscheinlichkeiten und Wahrscheinlichkeitsvisualisierungen gegeben waren, obwohl gerade das Baumdiagramm mit Wahrscheinlichkeiten an den Ästen im Fokus des Mathematikunterrichts steht. Die größtenteils unbekannten Häufigkeitsbäume konnten die Schülerinnen und Schüler bei der Lösung hingegen deutlich besser unterstützen (Lösungsrate 45%).

Der zweite Artikel beschreibt zwei Studien mit Medizinstudierenden des Universitätsklinikums Regensburg, in denen Bayesianische medizinische Entscheidungsfindungsprozesse untersucht werden, die realitätsnah nicht nur *ein* diagnostisches Verfahren, sondern zwei berücksichtigen, um zu einer medizinischen Diagnose zu gelangen (z.B. Mammographie und Sonographie zur Diagnose einer Brustkrebserkrankung). In der ersten Studie wurde gezeigt, dass sowohl natürliche Häufigkeiten als auch Baumdiagramme mit natürlichen Häufigkeiten das Verständnis der Situationen auch im 2-Test-Fall unterstützen. Hierbei spielt es keine Rolle, ob die statistischen Informationen zusätzlich auch noch als Text geschildert werden oder ob diese lediglich aus dem Baumdiagramm entnommen werden können. In der zweiten Studie des Artikels wurden modifizierte Baumdiagramme untersucht, bei denen die beiden zur Lösungsfindung relevanten Äste entweder farblich markiert wurden oder sogar nur diese beiden Äste dargestellt wurden. Während der markierte Häufigkeitsbaum das Verständnis gegenüber einem normalen Häufigkeitsbaum nochmal deutlich verbesserte (67% vs. 47%), blieb die Lösungsrate beim „reduzierten Baumdiagramm“ bei 47%.

Der dritte Artikel beinhaltet eine ausführliche theoretische Analyse verschiedener Formulierungsmöglichkeiten der eben beschriebenen 2-Test-Fälle. Hierbei werden vier Eigenschaften vorgestellt, die die Formulierung statistischer Informationen erfüllen sollten, damit diese möglichst gut von Menschen verstanden werden. Anschließend wird eine Studie mit 123 Medizinstudierenden der Charité Berlin vorgestellt, in der neben dem 2-Test-Fall weitere Verallgemeinerungen Bayesianischer Standardaufgaben untersucht werden: Ein 3-Test-Fall, eine

Situation, in der zwei verschiedene Erkrankungen mit einem Test diagnostiziert werden können und eine Situation, in der drei verschiedene Testergebnisse (z.B. auch unklarer Befund) möglich sind. Während natürliche Häufigkeiten in allen vier verallgemeinerten Situationen das Verständnis entscheidend verbessern konnten, war die zusätzliche Darbietung von Häufigkeitsbäumen nur dann hilfreich, wenn es sich um 2- oder 3-Test-Fälle handelte, in denen die statistischen Informationen also mehrfach ineinander verschachtelt waren. Darüber hinaus wurden im dritten Artikel auch alternative Diagnosen untersucht (z.B. die Wahrscheinlichkeit einer Erkrankung nach einem positiven und einem negativen Testergebnis), bei denen gerade im 3-Test-Fall die Präsentation eines Häufigkeitsbaumes das Verständnis verbesserte.

Einleitung

In der heutigen Informationsgesellschaft werden wir täglich mit einer Fülle an statistischen Informationen, Daten, Fakten (manchmal auch „alternativen Fakten"), Diagrammen und Tabellen konfrontiert. Laut Walter Krämer (2008) ist „Prozent" sogar das häufigste Substantiv, das in deutschen Zeitungen gefunden werden kann. Unglücklicherweise leben wir jedoch auch in einer Gesellschaft, die nicht in der Lage ist, mit statistischen Informationen umzugehen und adäquate Schlussfolgerungen aus ihnen zu ziehen. Im schlimmsten Fall kann diese mangelnde Risikokompetenz zu schwerwiegenden Fehlurteilen führen, beispielsweise bei medizinischen Entscheidungsfindungsprozessen (siehe z.B. Gigerenzer & Gray, 2011; Wegwarth & Gigerenzer, 2013) oder bei juristischen Urteilen (siehe z.B. Schneps & Colmez, 2013; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; Satake & Murray, 2017; Barker, 2017).

Besonders häufig unterliegen Menschen Trugschlüssen in Bayesianischen Situationen. Diese Klasse an Aufgaben hat überdies eine hohe Relevanz im realen Umgang mit Risiken und wird bereits in Schulen (Artikel 1) und Universitäten unterrichtet – allerdings überwiegend mit kognitiv ungünstigen Lösungsstrategien.

Relevanz Bayesianischer Aufgaben

Bayesianische Schlussfolgerungen stellen neben anderen Entscheidungsfindungsprozessen (wie Heuristiken, siehe z.B. Gigerenzer, Todd, & ABC research group, 1999) eine Form statistischen Denkens dar, die in der Realität von großer Bedeutung ist. Menschen werden tatsächlich mit A-priori-Wahrscheinlichkeiten, Sensitivitäten und Falsch-Positiv-Raten konfrontiert und müssen hieraus eine A-posteriori-Wahrscheinlichkeit ableiten. Wenn Ärzte beispielsweise in der medizinischen Realität (Artikel 2, ab Seite 38; Artikel 3, ab Seite 67) nicht in der Lage sind, die statistischen Informationen adäquat miteinander zu verknüpfen, so führt dies sowohl zu Überdiagnosen und Übertherapien (Jorgensen & Gotzsche, 2009; Wegwarth & Gigerenzer, 2013) als auch zu psychischen Belastungen nach fälschlicherweise positiven Testergebnissen (Brewer, Salz, & Lillie, 2007; Salz, Richman, & Brewer, 2010; Wegwarth, 2018). Wenn medizinisches Personal in HIV-Beratungsstellen die Möglichkeit falsch-positiver Testergebnisse verschweigt, können schlimmstenfalls sogar Suizidfälle die Folge sein, wenn Patienten den Ergebnissen in einer Illusion der Gewissheit voll und ganz vertrauen (Stine, 1996).

Neben der Medizin sind Bayesianische Aufgaben auch für Manager (Hoffrage, Hafenbrädl, & Bouquet, 2015) oder im juristischen Kontext relevant (Hoffrage et al., 2000; Satake & Murray, 2017; Barker, 2017). Im letzteren Bereich führen verschiedene Klassen von Trugschlüssen, die mit bedingten Wahrscheinlichkeiten zusammenhängen, regelmäßig zu Fehlverurteilungen (z. B. die „prosecutor's fallacy" wie im Fall Sally Clark, Hill, 2004).

Bei „Bayesianischen Aufgaben" handelt es sich demnach um weit mehr als nur um reine Mathematikaufgaben ohne jeglichen Anwendungsbezug, die lediglich schulisch und universitär unterrichtet werden, um die dahinterliegenden mathematischen Konzepte vermitteln zu können.

Kognitiv ungünstige Strategien in bisheriger Ausbildung

All diese Fehlentscheidungen in den verschiedenen Professionen wären vermeidbar, wenn die Risikokompetenz (bestehend aus Gesundheitskompetenz, Finanzkompetenz und digitale Risikokompetenz, siehe z.B. Gigerenzer & Martignon, 2015) und der Umgang mit statistischen Informationen bereits möglichst früh gefördert würde, wie beispielsweise bereits ab dem Kindergarten oder in der Grundschule (Zhu & Gigerenzer, 2006, Gigerenzer & Martignon, 2015; Martignon & Krauss, 2007).

Zahlreiche Studien und eine aktuelle Meta-Analyse belegen mittlerweile nämlich, dass natürliche Häufigkeiten und Visualisierungen Menschen sehr gut dabei unterstützen, die korrekte Lösung bei Bayesianischen Aufgaben zu bestimmen (Gigerenzer & Hoffrage, 1995; McDowell & Jacobs, 2017; Siegrist & Keller, 2011).

Auch wenn Bayesianische Aufgabenstellungen in der Schule und vielen universitären Studiengängen gelehrt werden, erfolgt bislang in weiten Teilen jedoch noch keine adäquate Förderung der Risikokompetenz von Schülerinnen und Schülern oder Studierenden. Die meisten Lehrkonzepte greifen unglücklicherweise immer noch auf ungünstige kognitive Strategien zurück, nämlich auf die Formel von Bayes oder die Pfadregeln (auch Multiplikationssatz und Additionsstz genannt), basierend auf Wahrscheinlichkeiten (siehe auch Artikel 1, ab Seite 20). Nur zu selten finden sich in schulischen oder universitären Lehrbüchern (für eine Ausnahme siehe z.B. Sedlmeier & Köhlers, 2005) bereits Darstellungen und Erklärungen, die auf natürliche Häufigkeiten zurückgreifen.

Aufgrund der Relevanz Bayesianischen Denkens in realen Entscheidungsfindungsprozessen und der bisherigen mangelhaften Umsetzung bekannter Strategien zur Förderung Bayesianischen Denkens in deutschen Schulen und Universitäten, wurde in der vorliegenden Dissertation untersucht, welche Baumdiagramme für das Verständnis von Schülerinnen und Schülern oder Medizinstudierenden in Bayesianischen 1-Test-Fällen (Artikel 1, ab Seite 20), Bayesianischen 2-Test-Fällen (Artikel 2, ab Seite 38) und weiteren Bayesianischen Situationen (Artikel 3, ab Seite 67) besonders hilfreich sind. Gleichzeitig verdeutlichen die Artikel 1 bis 3 dabei auch, wie wenig wirksam die bisher häufig eingesetzten Wahrscheinlichkeitsbäume sind.

Auf Anregung der Mathematikdidaktik der Universität Regensburg wurden die gewonnenen Erkenntnisse aus dem ersten hier vorgestellten Artikel im neuen LehrplanPlus (Bayern) bereits umgesetzt und „Häufigkeitsbäume“ erstmals explizit im bayerischen Gymnasium eingeführt:

> Die Schülerinnen und Schüler…
> - verstehen, dass in Sachzusammenhängen (z. B. in der medizinischen Diagnostik) klar zwischen $P_B(A)$ und $P_A(B)$ unterschieden werden muss, und sind in der Lage, mithilfe von Vierfeldertafeln oder Baumdiagrammen – auch solchen, in denen sie Wahrscheinlichkeiten mithilfe von absoluten Häufigkeiten in den Feldern bzw. Knoten illustrieren – von der einen auf die andere bedingte Wahrscheinlichkeit zu schließen.
>
> *ISB (2018), LehrplanPlus Gymnasium Bayern, Mathematik 10. Klasse*

Überblick über die drei Artikel der kumulativen Promotion

Zwei der drei Artikel sind bereits in englischsprachigen Open Access-Journalen erschienen, der dritte wurde in einem englischsprachigen APA-Journal eingereicht: Der erste Artikel erschien 2015 im kognitionspsychologischen Journal *Frontiers in Psychology* in der Rubrik *Cognition* im Research Topic „Improving Bayesian Reasoning: What Works and Why?“, das von Gorka Navarrete und David Mandel herausgegeben wurde. Der zweite Artikel erschien 2018 im multidisziplinären Journal *PlosONE*, da die hier vorgestellten Ergebnisse für ein besonders breites Publikum interessant sind, wie beispielsweise für die Psychology, Medizin, Medizindidaktik und die Mathematikdidaktik. Der dritte Artikel wurde im *Journal of Experimental Psychology: General* eingereicht. Ein Überblick über die drei Artikel findet sich in Tabelle 1 und eine Zusammenfassung der Ergebnisse der drei Artikel in Tabelle 2 (Seite 120).

Im Frontiers-Artikel wurden bei einer Stichprobe von *N*=259 Gymnasiasten Bayesianische Standardaufgaben (1-Test-Fälle) untersucht, in denen – wie in der Schule üblich – *ein* binäres Kriterium mithilfe *eines* binären Prädiktors vorhergesagt werden soll. Bayesianische Standardaufgaben können als ein Grundbaustein der Risikokompetenz eines jeden Menschen angesehen werden, da jede Person in ihrem Leben auch beispielsweise als Patient verstehen sollte, was ein positives medizinisches Testergebnis tatsächlich bedeutet. Aus diesem Grund und weil Bayesianische Standardaufgaben bereits im schulischen Curriculum verankert sind, wurden als Versuchspersonen für den ersten Artikel Schülerinnen und Schüler ausgewählt.

In der medizinischen Realität treten jedoch sehr häufig Situationen auf, in denen *mehrere* Testergebnisse zugrunde liegen oder *mehrere* Symptome auf eine bestimmte Erkrankung hinweisen. Im PlosONE-Artikel standen deshalb bei einer Stichprobe von *N*=388 Regensburger Medizinstudierenden Bayesianische 2-Test-Fälle im Fokus und im Artikel, der bei JEP eingereicht wurde, zusätzlich noch bei einer Stichprobe von *N*=123 Medizinstudierenden der Charité Berlin Fälle mit mehreren Tests, Testergebnissen oder Krankheiten, die im medizinischen Alltag ebenfalls höchst relevant sind.

Tabelle 1. Überblick über die drei Artikel der Dissertationsschrift

Artikel		**Erster Artikel**	**Zweiter Artikel**		**Dritter Artikel**
Journal		Frontiers in Psychology	PlosONE		JEP: General
Autoren (Jahr)		Karin Binder, Stefan Krauss, & Georg Bruckmaier (2015)	Karin Binder, Stefan Krauss, Georg Bruckmaier & Jörg Marienhagen (2018)		Karin Binder & Stefan Krauss (eingereicht)
Titel		*Effects of visualizing statistical information – an empirical study on tree diagrams and 2 × 2 tables*	*Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making*		*Generalizations of the Bayesian reasoning paradigm*
Thema		1-Test-Fall	2-Test-Fall		Weitere Generalisierungen
Studienteilnehmer		256 Schülerinnen und Schüler der 11. Klasse	190 Medizinstudierende der Universität Regensburg	198	123 Medizinstudierende der Charité Berlin
Untersuchte Situationen		Bayesianische Standardaufgabe (1-Test-Fall)	Bayesianischer 2-Test-Fall		• 2-Test-Fall • 3-Test-Fall • 3 Hypothesen • 3 Testergebnisse
Design	**Faktor 1**	Format: Wahrscheinlichkeiten vs. natürliche Häufigkeiten (für alle vier Teilstudien)			
	Faktor 2	Visualisierung: keine Visualisierung vs. Vierfeldertafel vs. Baumdiagramm	Visualisierung: Nur Text vs. Text und Baum vs. Nur Baum	Visualisierung: vollständiger Baum vs. markierter Baum vs. reduzierter Baum	Visualisierung: Nur Text vs. Text und Baumdiagramm
	Faktor 3	Kontext: Brustkrebs vs. Persönlichkeitseigenschaft	Kontext: Brustkrebs vs. HIV	Kontext: Brustkrebs vs. HIV	Kontext: Brustkrebs vs. spezielle Erkrankung vs. Erkrankung A und B vs. Brustkrebs
Gesuchte Wahrscheinlich-keiten		• Wahrscheinlichkeit einer Erkrankung nach einem positiven Testergebnis • Wahrscheinlichkeit den Wirtschaftskurs zu wählen, wenn man karriere-orientiert ist	Wahrscheinlichkeit einer Erkrankung nach zwei positiven Testergebnissen		• Wahrscheinlichkeit einer Erkrankung nach positiven Testergebnissen • Wahrscheinlichkeit einer Erkrankung nach nicht ausschließlich positiven Testergebnissen

1-Test-Fall (Artikel 1, Frontiers in Psychology)

Inhaltliche Schwerpunktsetzung des Frontiers-Artikels

Der Artikel zum 1-Test-Fall trägt den Titel *Effects of visualizing statistical information – An empirical study on tree diagrams and 2 × 2 tables* und ist im August 2015 im englischsprachigen Onlinejournal *Frontiers in Psychology* erschienen.

Im Fokus des Artikels steht der schulische Stochastikunterricht. In einer empirischen Untersuchung mit N=259 Gymnasiasten wird der Frage nach der Wirksamkeit der zusätzlichen Darbietung einer Visualisierung (Baumdiagramm oder Vierfeldertafel, jeweils mit Wahrscheinlichkeiten oder natürlichen Häufigkeiten) bei Bayesianischen Aufgaben nachgegangen. Nach einer Darstellung der typischen Visualisierungen Bayesianischer Standardaufgaben[1] wird erläutert, warum sich Vierfeldertafeln und Baumdiagramme unterrichtlich besonders gut eignen. Diese beiden Visualisierungen werden daher auch tatsächlich typischerweise im Mathematikunterricht eingesetzt, um Aufgaben zu bedingten Wahrscheinlichkeiten zu lösen. Im schulischen Unterricht finden sich jedoch vor allem Vierfeldertafeln mit absoluten oder relativen Häufigkeiten und Baumdiagramme mit (bedingten) Wahrscheinlichkeiten an den Ästen des Baumes, allerdings in der Regel keine Baumdiagramme mit absoluten Häufigkeiten in den Knoten des Baumes.

Die Hypothesen des ersten Artikels lauten im Einzelnen:
Hypothese 1: Bayesianische Aufgaben mit natürlichen Häufigkeiten werden häufiger korrekt gelöst als Bayesianische Aufgaben mit Wahrscheinlichkeiten.
Hypothese 2: Die Darbietung eines Baumdiagramms oder einer Vierfeldertafel unterstützt das Verständnis bei Bayesianischen Aufgaben zusätzlich.

Die Ergebnisse der Studie zeigen, dass natürliche Häufigkeiten, Vierfeldertafeln mit natürlichen Häufigkeiten, aber auch die unterrichtlich bislang nicht beachteten *Häufigkeitsbäume* das Verständnis ganz deutlich unterstützen. Ebenso zentral ist jedoch das Ergebnis, wie wenig förderlich die typischerweise unterrichteten Wahrscheinlichkeitsvisualisierungen für die Schülerinnen und Schüler sind. Diese Erkenntnis führte schließlich (wie in der Einleitung bereits dargelegt) zur Implementation von Häufigkeitsbäumen in den bayerischen gymnasialen LehrplanPLUS (10. Klasse).

Aus den Ergebnissen lassen sich konkrete unterrichtliche Empfehlungen ableiten (siehe Diskussion des Artikels). Ein unterrichtliches Konzept zur Umsetzung der Strategien 1) natürliche Häufigkeiten und 2) Baumdiagramme bei Bayesianischen Aufgaben findet sich in Stochastik in der Schule (Binder, Krauss, & Wassner, 2018), wo das Konzept der Häufigkeitsbäume auf Häufigkeitsdoppelbäume erweitert wird.

[1] Forschung zu Bayesianischen Aufgaben fokussiert zumeist „Bayesianische Standardaufgaben“, also bezogen auf medizinische Kontexte den „1-Test-Fall“ (es liegt lediglich *ein* Testergebnis vor). Aus diesem Grund wird im ersten Artikel nicht explizit von „1-Test-Fällen“ oder „Bayesianischen Standardaufgaben“ gesprochen, weil jedem Leser klar ist, dass genau dieser Fall damit gemeint ist.

Artikel 1: Effects of visualizing statistical information

Effects of visualizing statistical information – an empirical study on tree diagrams and 2 × 2 tables

Karin Binder, Stefan Krauss & Georg Bruckmaier

Mathematics Education, Faculty of Mathematics, University of Regensburg, Regensburg, Germany

Abstract

In their research articles, scholars often use 2 × 2 tables or tree diagrams including natural frequencies in order to illustrate Bayesian reasoning situations to their peers. Interestingly, the effect of these visualizations on participants' performance has not been tested empirically so far (apart from explicit training studies). In the present article, we report on an empirical study (3 × 2 × 2 design) in which we systematically vary visualization (no visualization vs. 2 × 2 table vs. tree diagram) and information format (probabilities vs. natural frequencies) for two contexts (medical vs. economical context; not a factor of interest). Each of N = 259 participants (students of age 16–18) had to solve two typical Bayesian reasoning tasks ("mammography problem" and "economics problem"). The hypothesis is that 2 × 2 tables and tree diagrams – especially when natural frequencies are included – can foster insight into the notoriously difficult structure of Bayesian reasoning situations. In contrast to many other visualizations (e.g., icon arrays, Euler diagrams), 2 × 2 tables and tree diagrams have the advantage that they can be constructed easily. The implications of our findings for teaching Bayesian reasoning will be discussed.

Keywords: Bayesian reasoning, 2 × 2 table, natural sampling tree, natural frequencies, visual representation

Introduction

Bayes' formula is vitally important in many areas, such as in medicine or law. Unfortunately, both laymen and professionals have trouble understanding and combining statistical information effectively. The resulting misjudgments can have severe consequences, for example when juries must convict or acquit defendants based on probabilistic evidence in legal trials (Hoffrage et al., 2000; Krauss and Bruckmaier, 2014), or when physicians have to understand and to communicate what a positive test result really means, for example in a HIV or cancer test (Ellis et al., 2014). Consider, for instance, the classic mammography problem (adapted from Eddy, 1982; see also Gigerenzer and Hoffrage, 1995; Siegrist and Keller, 2011; Micallef et al., 2012; Garcia-Retamero and Hoffrage, 2013).

Mammography Problem (Probability Format):

> The probability of breast cancer is 1% for a woman who participates in routine screening. If a woman who participates in routine screening has breast cancer, the probability is 80% that she will have a positive test result. If a woman who participates in routine screening does not have breast cancer, the probability is 9.6% that she will have a positive test result. What is the probability that a woman who participates in routine screening and receives a positive test result has breast cancer?
>
> Answer: _____ %

According to Bayes' theorem, the resulting posterior probability P(B|M+) is:

$$P(B|M+) = \frac{P(M+|B) \cdot P(B)}{P(M+|B) \cdot P(B) + P(M+|\neg B) \cdot P(\neg B)}$$

$$= \frac{80\% \cdot 1\%}{80\% \cdot 1\% + 9.6\% \cdot 99\%} \approx 7.8\%$$

The correct result 7.8% is much lower than most people, including physicians, would expect (Eddy, 1982). Several studies show that medical doctors (Hoffrage and Gigerenzer, 1998; Garcia-Retamero and Hoffrage, 2013), patients (Garcia-Retamero and Hoffrage, 2013), legal professionals (Hoffrage et al., 2000), and students (Ellis et al., 2014) have difficulties with similar tasks. In order to help people to understand the situation, Gigerenzer and Hoffrage (1995) replaced the probabilities in Eddy's task by natural frequencies.

Mammography Problem (Natural Frequency Format):

> 100 out of 10,000 women who participate in routine screening have breast cancer. Out of 100 women who participate in routine screening and have breast cancer, 80 will have a positive result. Out of 9,900 women who participate in routine screening and

> have no breast cancer, 950 will also have a positive result. How many of the women who participate in routine screening and receive a positive test result have breast cancer?
>
> Answer: ____ out of ____

The percentage of correct responses increased from about 10–20% to about 50% in 15 different Bayesian reasoning tasks, including the mammography problem (Gigerenzer and Hoffrage, 1995). While the facilitating effect of natural frequencies is accepted by now, scholars differ in explaining this effect. Gigerenzer and Hoffrage (1995), for instance, argue that the human mind is evolutionarily adapted to the information format of natural frequencies ("ecological rationality") that result from a natural sampling process (Kleiter, 1994). Other theorists, however, claim that essentially the partitive information structure is responsible for the facilitating effect ("nested sets hypothesis" e.g., Girotto and Gonzalez, 2001; Sloman et al., 2003; Barbey and Sloman, 2007). Some scholars suggest that two different cognitive systems ("dual process theory" Sloman, 1996; Kahneman and Frederick, 2005; Barbey and Sloman, 2007) may be responsible for inferences with respect to the different information formats. While probability format triggers intuitive thinking according to system 1 ("associative system" see also Sloman, 1996), which may lead to base rate neglect, natural frequency format triggers deliberate reasoning according to system 2 ("rule based system"). Advocates of the dual process theory often support the nested sets hypothesis (e. g., Barbey and Sloman, 2007). For a discussion of the concept of natural frequencies see Gigerenzer and Hoffrage (1999), Lewis and Keren (1999), Mellers and McGraw (1999), Girotto and Gonzalez (2001, 2002), Hoffrage et al. (2002), Barbey and Sloman (2007), or Sirota et al. (2015a).

In fact, there are recommendations that natural frequencies should become part of the training for all medical students (Gigerenzer, 2013) and, moreover, should be part of elementary school curricula (Gigerenzer, 2014). Although the effect of numerical format (probabilities vs. natural frequencies) is quite substantial, it has to be noted that there is still potential for improvement ("only" approximately 50% correct solutions).

Another idea to improve insight into Bayesian reasoning situations is the additional representation of visual aids such as *Euler diagrams*, *icon arrays*, *frequency grids*, *unit squares*, *roulette wheel diagrams*, and *tree diagrams* (see Figure 1). According to the nested sets hypothesis, most of these visual aids represent the set-subset relation of the information. For an overview of possible visualizations see Paling (2003) or Spiegelhalter et al. (2011). Figure 1 shows some visual aids which have been tested empirically so far.

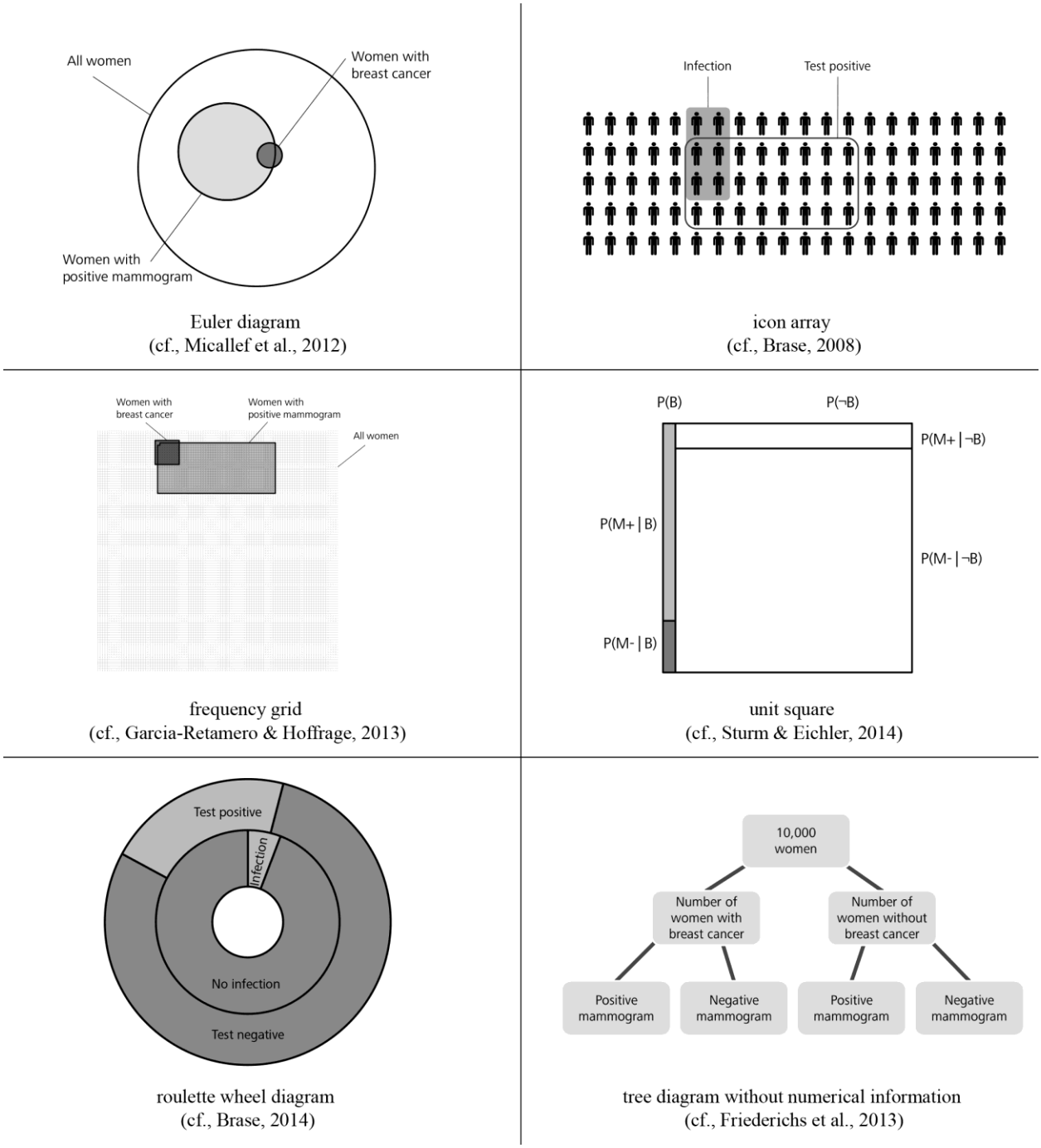


FIGURE 1. Risk communication via Euler diagram, icon array, frequency grid, unit square, roulette wheel diagram and tree diagram without numerical information.

Sloman et al. (2003), Brase (2008), Micallef et al. (2012), and Sirota et al. (2014b) investigated to what extent the presentation of *Euler diagrams* can boost performance in Bayesian reasoning tasks. They obtained different findings regarding the effectiveness of Euler diagrams, a result which potentially is affiliated to the various types of participants in their studies. *Icon arrays* (also called *pictographs*) are matrices of small figures that represent the given information. Within an array, some of the icons are shaped in a special form or are colored in order to show the number of figures that fulfill a special feature. Brase (2008, 2014) and Zikmund-Fisher et al. (2014)

recommend risk communication via icon arrays since their studies showed a positive influence of this visual aid (for a discussion of the concept of "iconicity" see, e.g., Sirota et al., 2014b). *Frequency grids* are close to icon arrays showing the overall number of persons in a large grid where particular subsets of persons are marked characteristically. Garcia-Retamero and Hoffrage (2013) found that both doctors' and patients' performance increased when frequency grids are provided (see also Garcia-Retamero et al., 2015). *Unit squares* (Bea, 1995; Sturm and Eichler, 2014) also mirror the statistical information geometrically and represent the different sets of the task. Bea (1995) recommends the visualization of information via a unit square since his research reveals substantial improvement in performance. *Roulette wheel diagrams* (Brase, 2014) summarize the information presented by two circles (inner and outer circle) which represent different subsets of the problem. However, the additional representation of a roulette wheel diagram causes only a very small or even no improvement in performance compared to versions without any visual aid (Brase, 2014). Friederichs et al. (2014) investigated *tree diagrams* without numerical values (except an imaginary sample size). In their studies, performance in probability versions with tree diagrams was similar to the performance in natural frequency versions without visualization.

Note that one can differentiate between two types of studies in general: On the one hand there are training studies where participants are explicitly instructed in how to create visual aids on their own, and consequently, how to combine the given numbers to arrive at the solution. The effect of this "teaching" then is investigated by presenting additional problems without visualizations (e.g., Sedlmeier and Gigerenzer, 2001; Ruscio, 2003; Sirota et al., 2015b). On the other hand there are studies – as in our study – where word problems are accompanied by visualizations (e.g., Brase, 2008; Garcia-Retamero and Hoffrage, 2013). Note that in the latter studies, it is not taught how to construct visualizations for fostering insight, and therefore, there is no prior instruction as to how the given numbers should be applied to infer the solution. The visualizations in this case rather illustrate the information of the given problem in parallel.

Interestingly, the beneficial effect of 2 × 2 tables and tree diagrams presently was investigated only in the context of training studies (e.g., Sedlmeier and Gigerenzer, 2001). This is astonishing since scholars commonly use tree diagrams (Kleiter, 1994; Gigerenzer and Hoffrage, 1995; Mandel, 2014; Navarrete et al., 2014) and 2 × 2 tables (Goodie and Fantino, 1996; Dougherty et al., 1999; Fiedler et al., 2000) containing numerical values in their research papers to represent Bayesian reasoning situations to their colleagues.

In the present paper we investigate how performance in Bayesian reasoning tasks can additionally be enhanced by providing 2 × 2 tables and tree diagrams containing numerical values. Since 2 × 2 tables and tree diagrams both can be equipped with natural frequencies or with probabilities we decided to test all four possible visualizations (compare Figure 2). Our hypotheses were:

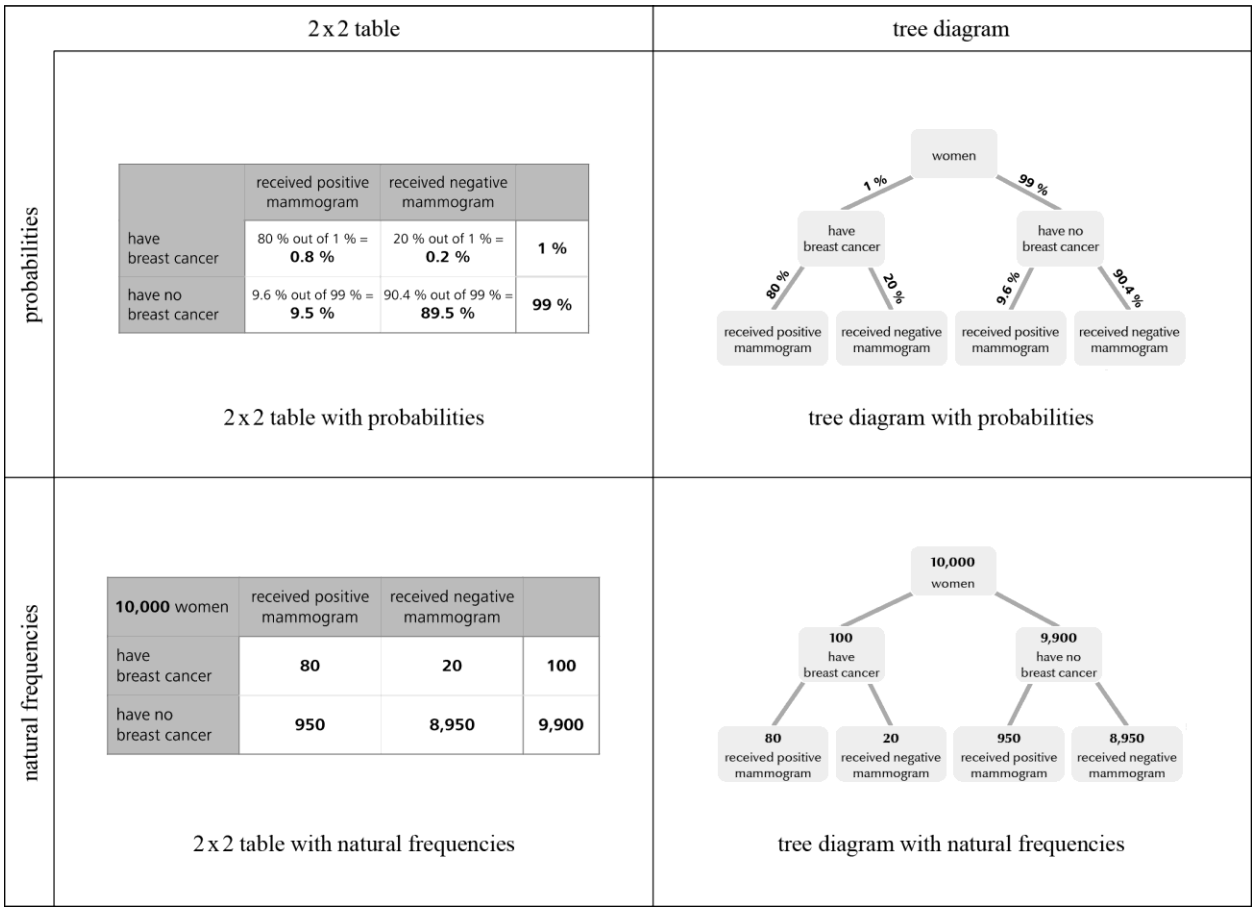


FIGURE 2. Four resulting visualizations of the respective information format (mammography problem).

- Hypothesis 1: Problems in which information is presented in natural frequencies are easier to solve than problems containing probabilities. This holds true when problems without visualization are compared (replication of previous studies) and when problems with visualizations are compared.

- Hypothesis 2: The additional presentation of visualizations of the numerical values (2 × 2 tables and tree diagrams) facilitates understanding. This holds for natural frequency and for probability versions as well.

We had no hypothesis as to which of both kinds of visualization is more beneficial. Furthermore we had no hypothesis on the effect of the problem context (we had chosen two problem contexts for mutual validation of our results; see Table 1).

TABLE 1. Design of the 12 tested problem version.

		Context	
		Mammography problem	**Economics problem**
Format	**Probabilities**	• No visualization • 2 x 2 table • Tree diagram	• No visualization • 2 x 2 table • Tree diagram
	Natural frequencies	• No visualization • 2 x 2 table • Tree diagram	• No visualization • 2 x 2 table • Tree diagram

Experimental Study

Design

In a paper-and-pencil questionnaire participants were presented with two Bayesian reasoning tasks, the mammography problem and a short version of the economics problem (Ajzen, 1977; for problem formulations see Table 2). The design of the study includes two factors of interest (visualization and format of information) and one factor which was not of interest (context), resulting in a 3 × 2 × 2 design:

- *Visualization:* no visualization vs. 2 × 2 table vs. tree diagram.
- *Format of statistical information*: probabilities vs. natural frequencies.
- *Context:* mammography problem vs. economics problem (not a factor of interest).

Each participant received one of the two problem contexts with probabilities and the other problem with natural frequencies. Thereby the order of context and information format was varied systematically. Furthermore, if in one of the two problems, for instance, a 2 × 2 table was added, in the other problem either no visualization or a tree diagram was presented. There were no time constraints for completing the questionnaire (participants required about 20 min for both tasks). In Table 1 the design, resulting in 12 tested versions, is illustrated, whereas in Table 2 the corresponding problem formulations are denoted.

TABLE 2. Problem formulations.

	Mammography problem		Economics problem	
	Probability version	**Natural frequency version**	**Probability version**	**Natural frequency version**
Cover story	Imagine you are a reporter for a women's magazine and you want to write an article about breast cancer. As a part of your research, you focuses on mammography as an indicator of breast cancer. You are especially interested in the question of what it means, when a woman has a positive result (which indicates breast cancer) in such a medical test. A physician explains the situation with the following information:		Imagine you are interested in the question, if career-oriented students are more likely to attend an economics course. Therefore the school psychological service evaluates the correlations of personality characteristics and choice of courses for you. The following information is available:	
Version	The probability of breast cancer is 1% for a woman who participates in routine screening. If a woman who participates in routine screening has breast cancer, the probability is 80% that she will have a positive test result. If a woman who participates in routine screening does not have breast cancer, the probability is 9.6% that she will have a positive test result.	100 out of 10,000 women who participate in routine screening have breast cancer. Out of 100 women who participate in routine screening and have breast cancer, 80 will have a positive result. Out of 9,900 women who participate in routine screening and have no breast cancer, 950 will also have a positive result.	The probability that a student attends the economics course is 32.5%. If a student attends the economics course, the probability that he is career oriented is 64%. If a student does not attend the economics course, the probability that he is still career oriented is 60%.	325 out of 1,000 students attend the economics course. Out of 325 students who attend the economics course, 208 are career-oriented. Out of 675 students who not attend the economics course, 405 are still career-oriented.
Visual aid	•No visualization, or •2 x 2 table (prob.), or •Tree diagram (prob.)	•No visualization, or •2 x 2 table (nat. freq.), or •Tree diagram (nat. freq.)	•No visualization, or •2 x 2 table (prob.), or •Tree diagram (prob.)	•No visualization, or •2 x 2 table (nat. freq.), or •Tree diagram (nat. freq.)
Question	What is the probability that a woman who participates in routine screening and receives a positive test result has breast cancer? Answer: _____%	How many of the women who participate in routine screening and receive a positive test result have breast cancer? Answer: ____ out of ____	What is the probability that a student attends the economics course if he is career-oriented? Answer: _____%	How many of the students who are career-oriented attend the economics course? Answer: ____ out of ____

The key factor under investigation in the present article is the effect of visualization. Note that in contrast to most visual aids tested so far (Figure 1) our visualizations explicitly contain numerical information. It is generally possible to equip both 2 × 2 tables and tree diagrams with natural frequencies or with probabilities, respectively (Figure 2). The construction rationale for the visualizations was to provide statistical information that is also reported in the typical problem formulations. However, to "complete" the tree diagrams some information must be added that is not mentioned in the problem formulation (the information "20%" and "90.4%" in the probability tree or "20" and "8,950" in the frequency tree, respectively). In order to mirror these numerical values in the 2 × 2 table containing natural frequencies, one (of two possible) marginal distribution has to be depicted (Figure 2). Most problematic is the construction of the 2 × 2 table

with probabilities. Such 2 × 2 tables usually contain conjoint probabilities, whereas Bayesian reasoning tasks contain conditional probabilities. The underlying relationship between both kinds of probabilities is included in the cells of the 2 × 2 tables (probabilities). It has to be noted that the 2 × 2 table (with conjoint probabilities), the 2 × 2 table (with natural frequencies) and the tree diagram (with probabilities) are part of the German school curriculum, whereas the tree diagram with natural frequencies ("natural frequency tree") is not.

Instrument

Each participant was presented two successive tasks which varied in terms of (1) visualization (no visualization vs. 2 × 2 table vs. tree diagram), (2) information format (probabilities vs. frequencies), and (3) problem context (mammography vs. economics problem). All versions begin with a cover story (see also Table 2); after that, one of three different kinds of visualization (including no visualization) was given (Figure 2). Finally, the question was provided in the same format as the information in the text.

The correct solution for the mammography problem is 80 out of 1,030 (about 7.8%), and for the economics problem 208 out of 613 (33.9%). Note that the corresponding algorithm to calculate the Bayesian posterior probability is identical for 2 × 2 tables concerning both information formats. However, the algorithm for computing P(B|M+) based on a tree diagram differs substantially with respect to both information formats.

A response has been classified as a correct "Bayesian answer" if the exact probability or frequency solution was provided, or the probability solution was rounded up or down to the next full percentage point (e.g., in the mammography problem the correct solution is 7.8%, therefore answers between 7 and 8% were classified as a correct solution; see also Gigerenzer and Hoffrage, 1995).

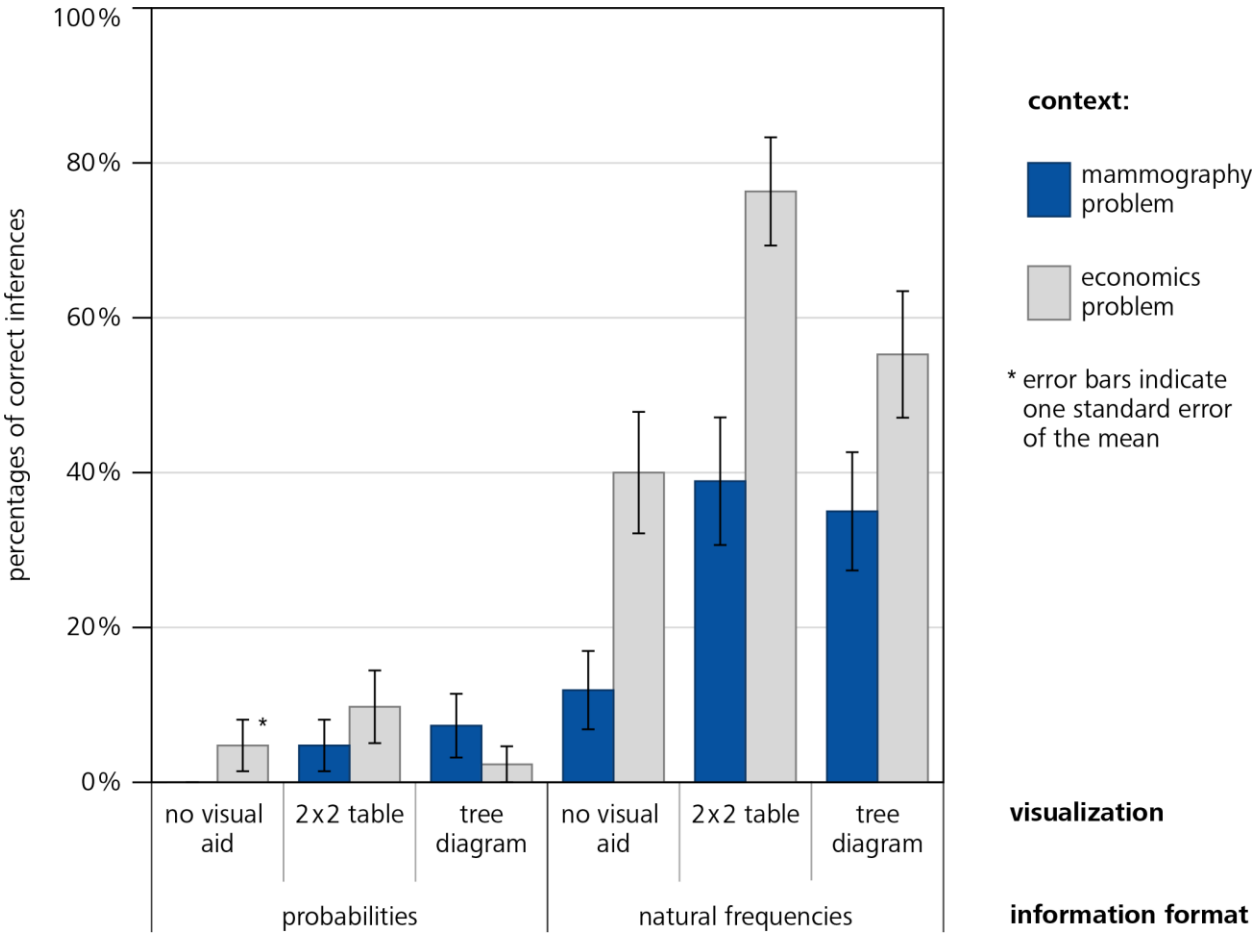
Participants

The participants were N = 259 German secondary school students age 16–18. Students were recruited from 12 different classes (grade 11) at two Bavarian Gymnasiums. Note that in Germany there are different kinds of secondary school tracks. In order to study at a university, the Gymnasium (academic track) must be pursued. All students were familiar with 2 × 2 tables and tree diagrams containing probabilities and with 2 × 2 tables containing frequencies but not with natural frequency trees.

The study was carried out in accordance with the University Research Ethics Standards. The principals of both schools approved conduction of the study (this is mandatory in Germany when testing school students). When conducting the study we did not collect personal data (our questionare did not include questions with regard to age, gender etc.). Students were informed that their participation was voluntary (two students refrained from participating) and anonymity was guaranteed. After the study participants were debriefed.

Results

Our study showed three important findings (Figure 3). First, students' performance was higher when information in the problems was presented in natural frequencies (42% correct inferences, averaged across context and visualization) instead of probabilities (5%), which supports our hypothesis 1. This finding holds when only problems *without* visualizations are compared (26% correct inferences in natural frequency versions vs. 2% correct inferences in probability versions, averaged across both contexts, which replicates previous findings, e.g., Gigerenzer and Hoffrage, 1995; Siegrist and Keller, 2011) and when problems *with* visualizations are compared (51% correct inferences in natural frequency versions vs. 6% correct inferences in probability versions, averaged across both contexts).



Note: Concerning the performance of 0% in the probability version of the economics problem (without visual aid), no standard error can be calculated.

FIGURE 3. Participants performance (error bars indicate the SE).

Second, the additional presentation of visualizations helps understanding (hypothesis 2): Averaging across all versions *with* visualization yields higher performance (28%) than averaging

across all versions *without* visualizations (14%). Note that this difference is much stronger in the natural frequency versions (51% vs. 26%, averaged across both contexts) than in the probability versions (6% vs. 2%, see Figure 2). The fact that probability visualizations only have very limited effect is irritating since these visual aids are frequently applied in statistical text books (see Discussion).

Furthermore, participants showed better performance in almost every version of the economics problem (30% correct inferences, averaged across format of information and visualization) compared to the respective versions of the mammography problem (16%). Possible reasons will be debated in Section "Discussion."

In order to analyze the impact of information format and visualization simultaneously we ran binary logistic regressions. Since we had no hypothesis on possible effects of problem context we performed two logistic regressions for the mammography problem and for the economics problem separately. The independent variables were visualization (only distinguishing between *no visualization* vs. *visualization*) and information format, respectively. The dependent variable was the correctness of the solution (1 – correct solution, 0 – incorrect solution). The results of the statistical analyses are illustrated in Table 3. For both contexts model 1 shows the impact of information format, whereas model 2 shows the impact of information format and visualization simultaneously.

TABLE 3. Results of binary logistic regression; independent variables: visualization and information format; dependent variable: correctness of solution.

	Dependent variable: correctness of solution			
	Mammography problem		**Economics problem**	
	Model 1	**Model 2**	**Model 1**	**Model 2**
Independent variable	EXP(B)	EXP(B)	EXP(B)	EXP(B)
Format of information	9.40***	10.44***	22.44***	24.73***
Visualization		4.99**		2.53*
R^2	0.19	0.27	0.41	0.44

EXP(B): Odds ratio (indicates how many times the odds of solving the task is higher when the independent variable is 1, as compared to the independence variable of 0);
R^2: Goodness of fit (according to Nagelkerke).
** significant at p = 0.05; ** significant at p = 0.01; *** significant at p = 0.001.*

In both problem contexts we found significant coefficients regarding information format (hypothesis 1) and visualization (vs. no visualization; hypothesis 2). Additional analyses revealed no statistical differences between 2 × 2 table and tree diagram in each information format. Although Figure 3 suggests a possible interaction of format and visualization the regression does

not yield a respective significant coefficient. Note that the seeming interaction between format and visualization may be due to the floor effect with respect to the probability versions. However, considering Figure 2 it becomes clear that visualizations of the numerical values in probability versions do not help substantially.

Discussion

According to general theories of information encoding and processing (e.g., Cognitive Load Theory, Sweller, 2003; Cognitive Theory of Multimedia Learning, Mayer, 2005), understanding of statistical information could be supported by presenting additional visual aids. In our study, participants' performance in two Bayesian reasoning tasks was higher when additionally 2 × 2 tables and tree diagrams containing natural frequencies were presented. However, when applying these visual aids for Bayesian inferences, the information format should be taken into account: both tools have only very limited effects when probabilities are included. Since in statistics text books and school curricula both probability visualizations – but not frequency trees – commonly are applied in order to foster insight, this finding is quite remarkable.

In general, our results are in line with the "frequentist hypothesis" (Gigerenzer and Hoffrage, 1995; Cosmides and Tooby, 1996) as well as the "nested sets hypothesis" (Barbey and Sloman, 2007). Regarding all problem versions, natural frequency versions resulted in higher performance levels compared to the respective probability versions. The low performance, however, in the natural frequency version of the mammography problem without visualization indicates only moderate statistical literacy in the participants of our study. Interestingly, the performance in the economics problem was much better than in the mammography problem under almost every condition. A possible reason might be the extreme base rate (1%) in the mammography problem which basically constitutes the cognitive illusion (in contrast, the result of the economics problem is no longer counterintuitive). Another reason might be that the context of the economics problem is more adapted to the living environment of young people (a strong dependency from the problem context was also found by Siegrist and Keller, 2011). The more complicated terminology or taxing cognitive capacity in the mammography problem could also account for the deviant effects in the different contexts (e.g., Lesage et al., 2013; Sirota et al., 2014a).

The need for tools for teaching statistics is repeatedly stressed (Gigerenzer, 2013, 2014; Navarrete et al., 2014). There are several teaching studies (Sedlmeier and Gigerenzer, 2001; Wassner, 2004; Mandel, 2015; Sirota et al., 2015b) where the solution process of a Bayesian reasoning problem is explained explicitly, e.g., with the help of visualizations, and the effect of teaching is investigated. For instance, it is even possible to advise students to imagine an arbitrary sample when given a probability version and then to construct a frequency table or tree diagram accordingly (by increasing the size of the arbitrary sample whole numbers always can be reached for each respective subset). Furthermore Hoffrage et al. (submitted, same issue) instructed participants to solve complex Bayesian reasoning problems (e.g., with more than one cue) by translating the given information in terms of probabilities into natural frequencies and to construct a corresponding tree diagram accordingly. Note again, that our study is not an explicit teaching

study; nevertheless our findings have pragmatic implications for teaching Bayesian reasoning. Our visualizations have the advantage that they can be constructed easily by teachers or students. In contrast, the diagrams in Figure 1 are complicated to produce, which is especially problematic when base rates are extreme. In the unit square, for instance, areas can become very small (in Figure 1 therefore a higher base rate of the disease was chosen). Similarly, concerning the icon array, more symbols would be required in the case of small or unmanageable proportions (such as 1.25 or 9.6%) thus entailing an enormous effort. Our frequency visualizations, which of course can be combined with other visualizations (for an integration of a natural frequency tree and an icon array see, e.g., Mossburger, unpublished manuscript), thus may be a helpful aid for fostering statistical understanding and for teaching statistics in schools.

Note that 2 × 2 tables and tree diagrams containing natural frequencies can not only aid in Bayesian reasoning problems, but can also illustrate situations with two dichotomous features in general. For instance, it is possible to justify and explain the rules for multiplication and addition of conditional probabilities with natural frequency trees very easily (Mossburger, unpublished manuscript). Since 2 × 2 tables and tree diagrams containing natural frequencies can be provided long before students have to solve Bayesian reasoning problems, these visual aids offer the opportunity to consider various types of problems over a long period of a school or university curriculum.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

We thank both reviewers for helpful comments and Robert DeHaney for editing of an earlier version of the manuscript. This work was supported by the German Research Foundation (DFG) within the funding program Open Access Publishing.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *J. Pers. Soc. Psychol 35*, 303–314. doi: 10.1037/00223514.35.5.303

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci. 30*, 241–297. doi: 10.1017/S0140525X07001653

Bea, W. (1995). *Stochastisches Denken [Stochastical Reasoning]*. Frankfurt am Main: Peter Lang.

Brase, G. L. (2008). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol. 23*, 369–381. doi: 10.1002/acp.1460

Brase, G. L. (2014). The power of representation and interpretation: doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. J. *Cogn. Psychol. 26*, 81–97. doi: 10.1080/20445911.2013.861840

Cosmides, L., and Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition 58*, 1–73. doi: 10.1016/0010-0277(95)00664-8

Dougherty, M. R., Gettys, C. F., and Ogden, E. E. (1999). MINERVA-DM: a memory processes model for judgments of likelihood. *Psychol. Rev. 106*, 180–209. doi: 10.1037/0033-295X.106.1.180

Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities," in *Judgment under Uncertainty: Heuristics and Biases*, eds D. Kahneman, P. Slovic, and A. Tversky (New York: Cambridge University Press), 249–267. doi: 10.1017/CBO9780511809477.019

Ellis, K. M., Cokely, E. T., Ghazal, S., and Garcia-Retamero, R. (2014). Do people understand their home HIV test results? Risk literacy and information search. *Proc. Hum. Fact. Ergon. Soc. Annu. Meet. 58*, 1323–1327. doi: 10.1177/1541931214581276

Fiedler, K., Brinkmann, B., Betsch, T., and Wild, B. (2000). A sampling approach to biases in conditional probability judgments: beyond base rate neglect and statistical format. *J. Exp. Psychol. 129*, 399–418. doi: 10.1037/0096-3445.129.3.399

Friederichs, H., Ligges, S., and Weissenstein, A. (2014). Using tree diagrams without numerical values in addition to relative numbers improves students' numeracy skills: a randomized study in medical education. *Med. Decis. Making 34*, 253–257. doi: 10.1177/0272989X13504499

Garcia-Retamero, R., Cokely, E. T., and Hoffrage, U. (2015). Visual aids improve diagnostic inferences and metacognitive judgment calibration. *Front. Psychol. 6:932*. doi: 10.3389/fpsyg.2015.00932

Garcia-Retamero, R., and Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Soc. Sci. Med. 83*, 27–33. doi: 10.1016/j.socscimed.2013.01.034

Gigerenzer, G. (2013). HIV screening: helping clinicians make sense of test results to patients. *BMJ 347*, f5151. doi: 10.1136/bmj.f5151

Gigerenzer, G. (2014). How I got started: teaching physicians and judges risk literacy. *Appl. Cogn. Psychol. 28*, 612–614. doi: 10.1002/acp.2980

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev. 102*, 684–704. doi: 10.1037/0033295X.102.4.684

Gigerenzer, G., and Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev. 106*, 425–430. doi: 10.1037/0033-295X.106.2.425

Girotto, V., and Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition 78*, 247–276. doi: 10.1016/S0010-0277(00)00133-5

Girotto, V., and Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning: rejoinder to Hoffrage, Gigerenzer, Krauss, and Martignon. *Cognition 84*, 353–359. doi: 10.1016/S0010-0277(02)00051-3

Goodie, A. S., and Fantino, E. (1996). Learning to commit or avoid the base-rate error. *Nature 380*, 247–249. doi: 10.1038/380247a0

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Acad. Med. 73*, 538–540. doi: 10.1097/00001888-199805000-00024

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition 84*, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science 290*, 2261–2262. doi: 10.1126/science.290.5500.2261

Kahneman, D., and Frederick, S. (2005). "A model of heuristic judgment," in *The Cambrigde Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morris (Cambridge: Cambridge University Press), 267–293.

Kleiter, G. D. (1994). "Natural sampling: rationality without base rates," in *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, eds G. H. Fischer and D. Laming (New York: Springer), 375–388. doi: 10.1007/978-1-4612-4308-3_27

Krauss, S., and Bruckmaier, G. (2014). "Eignet sich die Formel von Bayes für Gerichtsverfahren? [Is formula of Bayes appropriate for legal trials?]," in *Daten, Zufall und der Rest der Welt*, eds U. Sproesser, S. Wessolowski, and C. Wörn (Wiesbaden: Springer), 123–132. doi: 10.1007/978-3-658-04669-9_10

Lesage, E., Navarrete, G., and De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: the role of general cognitive resources. *Think. Reason. 19*, 27–53. doi: 10.1080/13546783.2012.713177

Lewis, C., and Keren, G. (1999). On the difficulties underlying Bayesian reasoning: a comment on Gigerenzer and Hoffrage. *Psychol. Rev. 106*, 411–416. doi: 10.1037/0033-295X.106.2.411

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Front. Psychol. 5:1144*. doi: 10.3389/fpsyg.2014.01144

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Front. Psychol. 6:387*. doi: 10.3389/fpsyg.2015.00387

Mayer, R. E. (2005). "Cognitive theory of multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, ed. R. E. Mayer (New York: Cambridge University Press), 31–48. doi: 10.1017/CBO9780511816819.004

Mellers, B. A., and McGraw, A. P. (1999). How to improve Bayesian reasoning: comment on Gigerenzer and Hoffrage (1995). *Psychol. Rev. 106*, 417–424. doi: 10.1037/0033-295X.106.2.417

Micallef, L., Dragicevic, P., and Fekete, J. (2012). Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. *IEEE Trans. Vis. Comput. Graph. 18*, 2536–2545. doi: 10.1109/TVCG.2012.199

Navarrete, G., Correia, R., and Froimovitch, D. (2014). Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. *Front. Psychol. 5:1272*. doi: 10.3389/fpsyg.2014.01272

Paling, J. (2003). Strategies to help patients understand risks. *BMJ 327*, 745–748. doi: 10.1136/bmj.327.7417.745

Ruscio, J. (2003). Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teach. Psychol. 30*, 325–328.

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen. 130*, 380–400. doi: 10.1037/0096-3445.130.3.380

Siegrist, M., and Keller, C. (2011). Natural frequencies and Bayesian reasoning: the impact of formal education and problem context. *J. Risk Res. 14*, 1039–1055. doi: 10.1080/13669877.2011.571786

Sirota, M., Juanchich, M., and Hagmayer, Y. (2014a). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonom. Bull. Rev. 21*, 198–204. doi: 10.3758/s13423-013-0464-6

Sirota, M., Kostovičová, L., and Juanchich, M. (2014b). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychonom. Bull. Rev. 21*, 961–968. doi: 10.3758/s13423-013-0555-4

Sirota, M., Kostovičová, L., and Vallée-Tourangeau, F. (2015a). Now you Bayes, now you don't: effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychon. Bull. Rev.* doi: 10.3758/s13423-015-0810-y [Epub ahead of print].

Sirota, M., Kostovičová, L., and Vallée-Tourangeau, F. (2015b). How to train your Bayesian: a problem-representation transfer rather than a format-representation shift explains training effects. *Q. J. Exp. Psychol. 68*, 1–9. doi: 10.1080/17470218.2014.972420

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychol. Bull. 119*, 3–22. doi: 10.1037/0033-2909.119.1.3

Sloman, S. A., Over, D., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process 91*, 296–309. doi: 10.1016/S0749-5978(03)00021-9

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science 333*, 1393–1400. doi: 10.1126/science.1191181

Sturm, A., and Eichler, A. (2014). "Students' beliefs about the benefit of statistical knowledge when perceiving information through daily media," in *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9)*, Flagstaff, AZ: Sustainability in Statistics Education, eds K. Makar, B. de Sousa, and R. Gould (Voorburg: International Statistical Institute.).

Sweller, J. (2003). Evolution of human cognitive architecture. *Psychol. Learn. Motiv. 43*, 215–266. doi: 10.1145/1404520.1404521

Wassner, C. (2004). *Förderung Bayesianischen Denkens – Kognitionspsychologische Grundlagen und didaktische Analysen [Promoting Bayesian Reasoning – Principles of Cognitive Psychology, and Didactical Analyses]*. Hildesheim: Franzbecker.

Zikmund-Fisher, B. J., Witteman, H. O., Dickson, M., Fuhrel-Forbis, A., Kahn, V. C., Exe, N. L., et al. (2014). Blocks, ovals, or people? Icon type affects risk perceptions and recall of pictographs. *Med. Decis. Making 34*, 443–453. doi: 10.1177/0272989X13511706

2-Test-Fall (Artikel 2, PlosONE)

Inhaltliche Schwerpunktsetzung des PlosONE-Artikels

Der zweite Artikel trägt den Titel *Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making* und ist 2018 im englischsprachigen Onlinejournal *PlosONE* erschienen.

Die Erkenntnisse aus dem Frontiers-Artikel werden nun in die Domäne Medizin übertragen. Im Fokus stehen daher medizinische 2-Test-Fälle, in der eine Diagnose aufgrund von zwei positiven Testergebnissen erfolgen soll, wie es in der medizinischen Realität durchaus üblich sein kann. Mit zwei Stichproben aus N=190 und N=198 Medizinstudierenden der Universität Regensburg wird der Frage nachgegangen, welche Baumdiagramme das Verständnis im Bayesianischen 2-Test-Fall unterstützen. Auf eine Untersuchung von Vierfeldertafeln wird im zweiten Artikel verzichtet, da es keine logische Erweiterung vom 1-Test-Fall auf den 2-Test-Fall bei Vierfeldertafeln gibt.

Nach einer Darstellung möglicher Visualisierungen des Bayesianischen 2-Test-Falls wird erläutert, warum sich Baumdiagramme für die medizinische Ausbildung besonders gut eignen. Anschließend werden zwei empirische Studien zur Wirksamkeit natürlicher Häufigkeiten und verschiedener Baumdiagramme vorgestellt.

Die Forschungsfragen des Artikels lauten im Einzelnen:
Forschungsfrage 1 (Studie 1 und 2): Unterstützen natürliche Häufigkeiten und Baumdiagramme das Verständnis im Bayesianischen 2-Test-Fall?
Forschungsfrage 2 (Studie 1): Werden Aufgaben besser gelöst, wenn die statistischen Informationen nur als Text, nur als Baum oder sowohl als Text als auch als Baumdiagramm dargestellt werden?
Forschungsfrage 3 (Studie 2): Ist es für das Verständnis hilfreich, die beiden aufgabenbezogenen Äste des Baumdiagramms zu markieren oder sogar die irrelevanten Äste des Baumdiagrammes wegzulassen?

Die Ergebnisse der Studie zeigen, dass natürliche Häufigkeiten und Häufigkeitsbäume das Verständnis deutlich unterstützen, während die Lösungsraten mit Wahrscheinlichkeiten oder Wahrscheinlichkeitsbäumen sehr niedrig bleiben (maximal 13% korrekte Lösungen). Dabei macht es bei den Häufigkeitsversionen keinen Unterschied, ob die statistischen Informationen zusätzlich noch textuell dargeboten werden oder ausschließlich mithilfe eines Baumdiagramms.

Die Markierung der beiden Äste des Baumdiagramms, die zur Beantwortung der Frage relevant sind, unterstützt das Verständnis zusätzlich, während das Abschneiden der irrelevanten Äste lediglich zur gleichen Performanz führt wie ein vollständiges Baumdiagramm.

Konkrete hochschuldidaktische Empfehlungen, die sich aus dem Artikel ableiten lassen, finden sich im Buch „Zeig mir mehr Biostatistik – Mehr Ideen und neues Material für einen guten Biometrie-Unterricht“ (Binder & Marienhagen, 2017).

Artikel 2: Visualizing the Bayesian 2-test case

Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making

Karin Binder[1], Stefan Krauss[1], Georg Bruckmaier[2] & Jörg Marienhagen[3]

[1] *Mathematics Education, Faculty of Mathematics, University of Regensburg, Regensburg, Germany*

[2] *Institute for Primary Education, School of Education, FHNW University of Applied Sciences and Arts, Northwestern Switzerland, Liestal, Switzerland*

[3] *University Hospital Regensburg, Regensburg, Germany*

Abstract

In medicine, diagnoses based on medical test results are probabilistic by nature. Unfortunately, cognitive illusions regarding the statistical meaning of test results are well documented among patients, medical students, and even physicians. There are two effective strategies that can foster insight into what is known as Bayesian reasoning situations: (1) translating the statistical information on the prevalence of a disease and the sensitivity and the false-alarm rate of a specific test for that disease from probabilities into natural frequencies, and (2) illustrating the statistical information with tree diagrams, for instance, or with other pictorial representation. So far, such strategies have only been empirically tested in combination for "1-test cases", where one binary hypothesis ("disease" vs. "no disease") has to be diagnosed based on one binary test result ("positive" vs. "negative"). However, in reality, often more than one medical test is conducted to derive a diagnosis. In two studies, we examined a total of 388 medical students from the University of Regensburg (Germany) with medical ª2-test scenariosº. Each student had to work on two problems: diagnosing breast cancer with mammography and sonography test results, and diagnosing HIV infection with the ELISA and Western Blot tests. In Study 1 (N = 190 participants), we systematically varied the presentation of statistical information ("only textual information" vs. "only tree diagram" vs. "text and tree diagram in combination"), whereas in Study 2 (N = 198 participants), we varied the kinds of tree diagrams ("complete tree" vs. "highlighted tree" vs. "pruned tree"). All versions were implemented in probability format (including probability trees) and in natural frequency format (including frequency trees). We found that natural frequency trees, especially when the question-related branches were highlighted, improved performance, but that none of the corresponding probabilistic visualizations did.

Introduction

Physicians, medical staff, and patients frequently have difficulty understanding what medical test results really mean. This is a serious issue because patients must often make tough decisions about specific medical treatments, for example after a positive test result from a routine screening [1]. Unfortunately, not only patients but also physicians and medical staff are often unable to combine and understand statistical information correctly. The resulting cognitive illusions can lead to an overestimation of the benefits of diagnostic methods or to an underestimation of the possible damage they could do [2,3]. For example, a positive HIV test result can lead to mental disorders or even suicide [4,5]. But what does an HIV test result really mean? Most counselors in the studies from Prinz et al. [6], Gigerenzer et al. [7], and Ellis and Brase [8] operate under an illusory belief that positive test results indicate certainty. But in fact, a positive HIV test result does not indicate the presence of HIV infection with absolute certainty [9].

Of course, the same applies to other medical diagnostic procedures. Another example is the mammography screening for breast cancer, which is very expensive and heavily promoted in many countries as necessary for every woman in a particular age group [10]. In the following, we call judgments based on a single medical test *1-test cases*.

The medical 1-test case

A study by Eddy [11] shows that even physicians are often unable to combine the statistical information of a breast cancer screening diagnosis in a 1-test case correctly. For instance, consider a situation in which breast cancer is diagnosed based on a mammogram (adapted from [11]):

Screening for breast cancer—1-test case (Probability Format):

> The probability of breast cancer is 1% for a woman of a particular age group who participates in a routine screening. If a woman who participates in a routine screening has breast cancer, the probability is 80% that she will have a positive mammogram. If a woman who participates in a routine screening does not have breast cancer, the probability is 9.6% that she will have a false-positive mammogram.
>
> What is the probability that a woman who participates in a routine screening and has a positive mammogram has breast cancer?

In the situation above, the *a priori probability* P(B) = 1% denotes the prevalence of the disease in a particular age group. The conditional probabilities P(M+|B) = 80% and P(M+|¬B) = 9.6% are called the *sensitivity* and the *false-alarm rate* of the mammography. In medicine, the *a posteriori probability* P(B|M+), which is the relevant one for patients, is called the *positive predictive value* of a medical test. The Bayes' theorem shows that the actual probability of breast cancer given a positive mammogram P(B|M+) is only about 7.8%.

$$P(B|M+) = \frac{P(M+|B) \cdot P(B)}{P(M+|B) \cdot P(B) + P(M+|\neg B) \cdot P(\neg B)}$$

$$= \frac{80\% \cdot 1\%}{80\% \cdot 1\% + 9.6\% \cdot 99\%} \approx 7.8\%$$

However, most physicians in Eddy`s study assumed this probability to be between 70% and 80%, far from the correct positive predictive value. A wide variety of empirical studies have shown that physicians, medical staff, and patients [12,13] have difficulties with problems of this kind. Furthermore, Bayesian reasoning problems are of relevance in many other domains, and the respective cognitive illusions are well documented among school students [14], university students [15], legal professionals [16], and managers [17].

Fortunately, there are two highly effective strategies for overcoming occurring cognitive illusions and helping people to understand statistical information—namely, natural frequencies and visualizations.

Strategy 1: Natural frequencies instead of probabilities

Rather than presenting all statistical information in the format of confusing conditional probabilities and percentages, one can provide natural frequencies as a means of describing Bayesian reasoning situations. In a seminal paper, Gigerenzer and Hoffrage [18] translate the numbers in the breast cancer screening problem into natural frequencies:

Screening for breast cancer—1-test case (Natural Frequency Format):

> 100 out of 10,000 women of a particular age group who participate in a routine screening have breast cancer. 80 out of 100 women who participate in a routine screening and have breast cancer will have a positive mammogram. 950 out of 9,900 women who participate in a routine screening and have no breast cancer will have a false-positive mammogram.
>
> How many of the women who participate in a routine screening and receive positive mammograms have breast cancer?

It is now easier to see that there are 80 + 950 women with positive mammograms, and that only 80 out of these 1,030 women actually have breast cancer, which again results in a positive predictive value of about 7.8%. With the natural frequency version significantly more people are able to make the correct inference [18,19], because one simply needs to calculate the proportion of women with breast cancer among those who have a positive mammogram.

For more than 20 years, natural frequencies have been a well-known tool for overcoming cognitive illusions in Bayesian reasoning situations, also with respect to slightly more complicated scenarios, such as the notorious Monty Hall problem [20]. More generally, frequency formulations (beyond natural frequencies) have also been able to reduce the so-called conjunction

fallacy (see, e.g., the Linda Problem [21,22]). With regard to Bayesian reasoning, there are myriad studies showing the enlightening properties of natural frequencies in a variety of domains: they help physicians in diagnostic inferences [12,13], patients in understanding these diagnoses [13], advanced law students in adequately evaluating legal indications [16], and managers and executives in management decisions [17], as well as university students [23] and secondary school students [14]. Even fourth graders are able to solve Bayesian reasoning tasks using natural frequencies [24].

A recently conducted meta-analysis from McDowell and Jacobs [25] reviews the results of 35 papers describing the impact of natural frequencies on decision-making processes and finds that the facilitating effect of natural frequencies is quite robust; the estimated average percentage correct for the probability versions of Bayesian reasoning tasks is 4%, while it is 24% for the corresponding natural frequency versions. Although there has been some discussion concerning the beneficial effect of natural frequencies [26,27], this effect has generally been recognized [25] and repeatedly replicated by now (for an exception see [28]), because they simplify the Bayesian calculation and more people are able to find the correct solution.

Strategy 2: Visualizing Bayesian reasoning tasks

There is another strategy for improving Bayesian reasoning in the 1-test case, namely, visualizing the statistical information. Some prominent visualizations that have been developed are *Euler diagrams* (e.g., [29–31]), *roulette-wheel diagrams* (e.g., [32,33]), *frequency grids* (e.g., [23,34,35]), *Eikosograms* (sometimes also called *unit squares* or *mosaic plots*; e.g., [36–39]), *icon arrays* (e.g., [32,40,41]), *2×2-tables* (e.g., [14,42]), and *tree diagrams* (e.g., [14,33,42–44]). For an overview of these visualizations, see [14], and for corresponding visualizations regarding the 2-test case, see Fig 1. With respect to the first strategy (natural frequencies), it must be noted that most visualizations do not contain any numbers (e.g., icon arrays, frequency grids, roulette-wheel diagrams or Euler diagrams) and therefore can illustrate natural frequency or probability versions as well.

Several of these visualizations have already been tested empirically (for an overview, see [14,45,46]). The previously mentioned meta-analysis [47] found that visualizations can also improve participant performance in Bayesian reasoning situations. The aggregate effect across various visualizations is an increase in correct inferences of about 23 percentage points. However, there is evidence that not all types of visualizations support people in their decision-making processes. With visualizations that contain numbers (i.e., tree diagrams or Eikosograms), the format of these numbers can make a difference in how participants understand the statistical information. For instance, it must be noted that in the 1-test case, only tree diagrams containing natural frequencies in the nodes, not tree diagrams with probabilities at the branches [14,23] or without any numerical information [43], significantly foster insight into Bayesian reasoning problems.

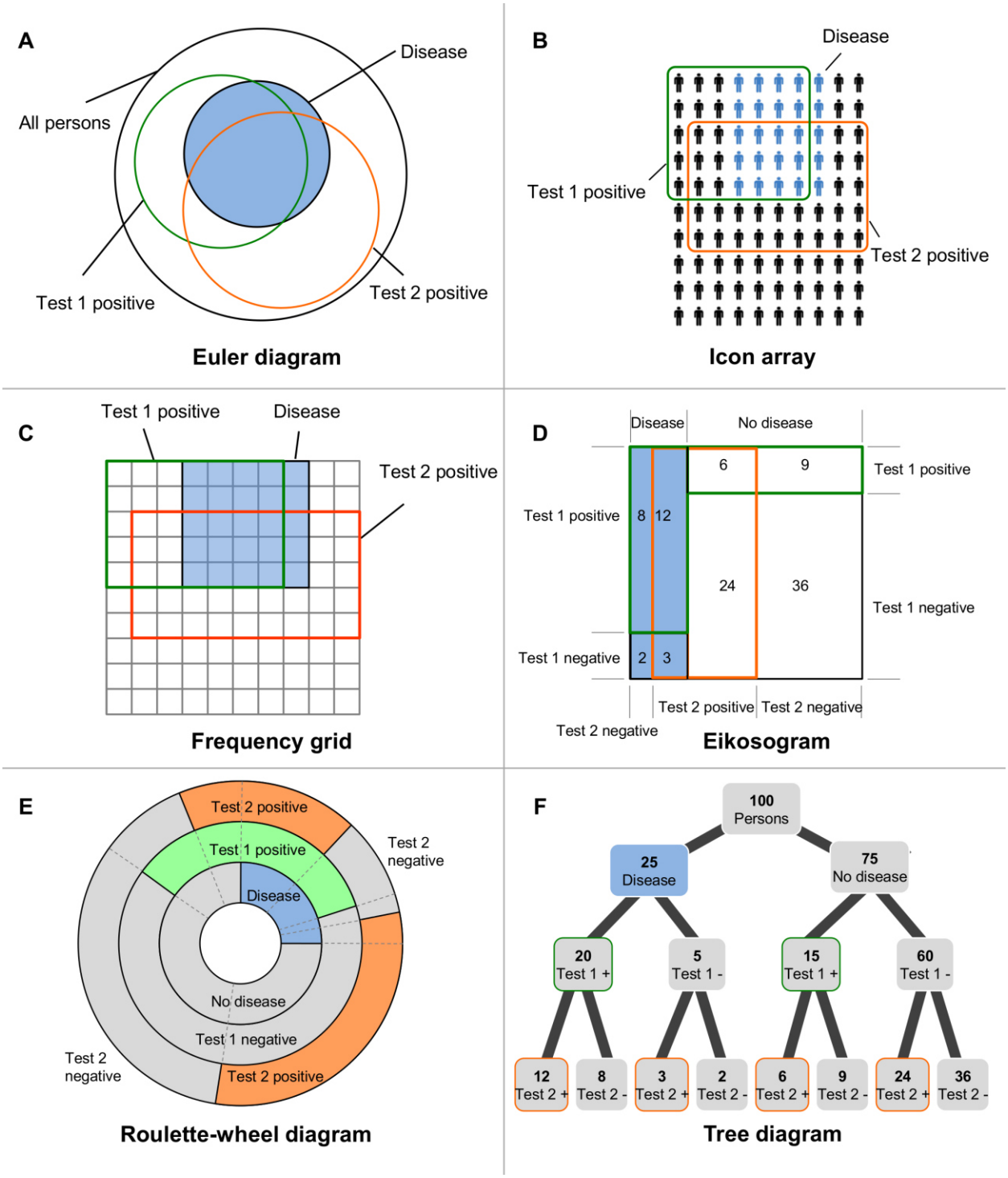


FIGURE 1. Six different types of visualization for the Bayesian 2-test case. (A) Euler diagram (B) Icon array (C) Frequency grid (D) Eikosogram (E) Roulette-wheel diagram, and (F) Tree diagram. Omitting the information on the second test in the different visualizations results in the corresponding visualization of the 1-test case.

The medical 2-test case

So far, empirical studies concerning visualizations of Bayesian reasoning situations are predominantly conducted with 1-test cases (for visualizing cases with non-binary hypotheses, see [32,33,48]). However, in many medical real-life applications, there is more than one medical test (or clinical symptom) available [49].

For instance, consider a situation in which breast cancer is diagnosed based on both a mammogram and a sonogram (adapted from [50,51]):

Screening for breast cancer—2-test case (Probability Format):

> The probability of breast cancer for a woman of a particular age group is 1%. The probability that a woman with breast cancer will have a positive mammogram is 80%. The probability that a woman with breast cancer will have a positive sonogram is 95%. The probability that a woman without breast cancer will have a false-positive mammogram is 9.6%. The probability that a woman without breast cancer will have a false-positive sonogram is 7.8%.
>
> What is the probability that a woman with a positive mammogram and a positive sonogram actually has breast cancer?

For alternative ways to present the statistical information of 2-test cases, for example by providing a combined sensitivity and a combined false-alarm rate, see the S1 Appendix. In the following we apply both natural frequencies and visualizations to situations where two medical test results are provided.

Strategy 1: Natural frequencies

Just as in the 1-test case, diagnoses based on two indicators can be formulated with natural frequencies instead of probabilities. Translating the 2-test case described into a natural frequency format yields:

Screening for breast cancer—2-test case (Natural Frequency Format):

> 100 out of 10,000 women of a particular age group have breast cancer. 80 out of 100 women with breast cancer have a positive mammogram. 76 out of 80 women with breast cancer and a positive mammogram have a positive sonogram. 950 out of 9,900 women without breast cancer have a false-positive mammogram. 74 out of 950 women without breast cancer but with a positive mammogram have a false-positive sonogram.
>
> How many of the women with a positive mammogram and a positive sonogram actually have breast cancer?

It has already been demonstrated empirically that the beneficial effect of natural frequencies is not limited to Bayesian 1-test cases but also holds for 2-test and even for 3-test cases [50,51].

Furthermore, Hoffrage et al. [51] successfully applied the natural frequency strategy to situations where either three hypotheses (e.g., disease A, disease B, or healthy) or three test results (e.g., positive, negative, or unclear test result) were provided. Yet as far as we know, only strategy 1, not strategy 2 (applying visualizations), has been investigated with regard to 2-test cases.

Strategy 2: Visualization

Generally, all visualizations of "simple" Bayesian reasoning problems (i.e., one binary hypothesis must be inferred from one binary cue) can be extended to visualizing medical 2-test cases (see Fig 1). It is not immediately obvious, however, which visualization is most helpful in 2-test cases. In the following we will point out why we chose to study tree diagrams.

Some general remarks on visualizing Bayesian reasoning problems

There are basically two possible applications of visualizations (regardless of the number of tests provided): (1) Visualizations can be *presented* to illustrate statistical information for physicians or patients. One can present visualizations *in addition* to textual information or *instead of* textual information. It is an open question as to which of these variants is most helpful for understanding the situation. (2) If no visualization is provided, problem solvers could *create visualizations on their own* in order to understand the situation. Here the question of which visualization can be produced with the least amount of effort arises.

Thus, it would be advantageous if the visualization were not only cognitively helpful but could also be constructed quickly simply using paper and pencil. Regarding Fig 1, producing Euler diagrams (Fig 1A), frequency grids (Fig 1C), Eikosograms (Fig 1D), and roulette-wheel diagrams (Fig 1E) all obviously require deliberate geometrical operations. Concerning Euler diagrams (Fig 1A) and roulette-wheel diagrams (Fig 1E), even areas of circles or circle sections have to be constructed. And with the icon array (Fig 1B), it is very tedious work to depict all of the figures (for N = 1,000 persons, 1,000 icons have to be charted). Furthermore, the geometrical nature of visualizations A-E (Fig 1) leads to the problem that extreme base rates (which are often responsible for cognitive illusions) are nearly impossible to depict. For example, in order to illustrate a base rate of 0.1%, visualizations such as A, C, D, and E (Fig 1) would contain unmanageably small areas, while icon arrays (Fig 1B) would require 1,000 symbols, thus all entailing enormous effort to produce these visualizations.

In contrast, the tree diagram (Fig 1F) can be produced with a simple paper-and-pencil-operation in a short amount of time. Because the tree diagram is the only non-geometrical visualization, even very small base rates can be illustrated simply by depicting the respective numbers. In addition, tree diagrams generally can be equipped with both (conditional) probabilities at the branches (a strategy that is predominantly implemented in teaching statistics in secondary schools and at universities) and also natural frequencies in the nodes. Fig 2 shows tree diagrams with respect to both information formats, depicting a medical 2-test case (diagnosing breast cancer based on a mammogram and a sonogram).

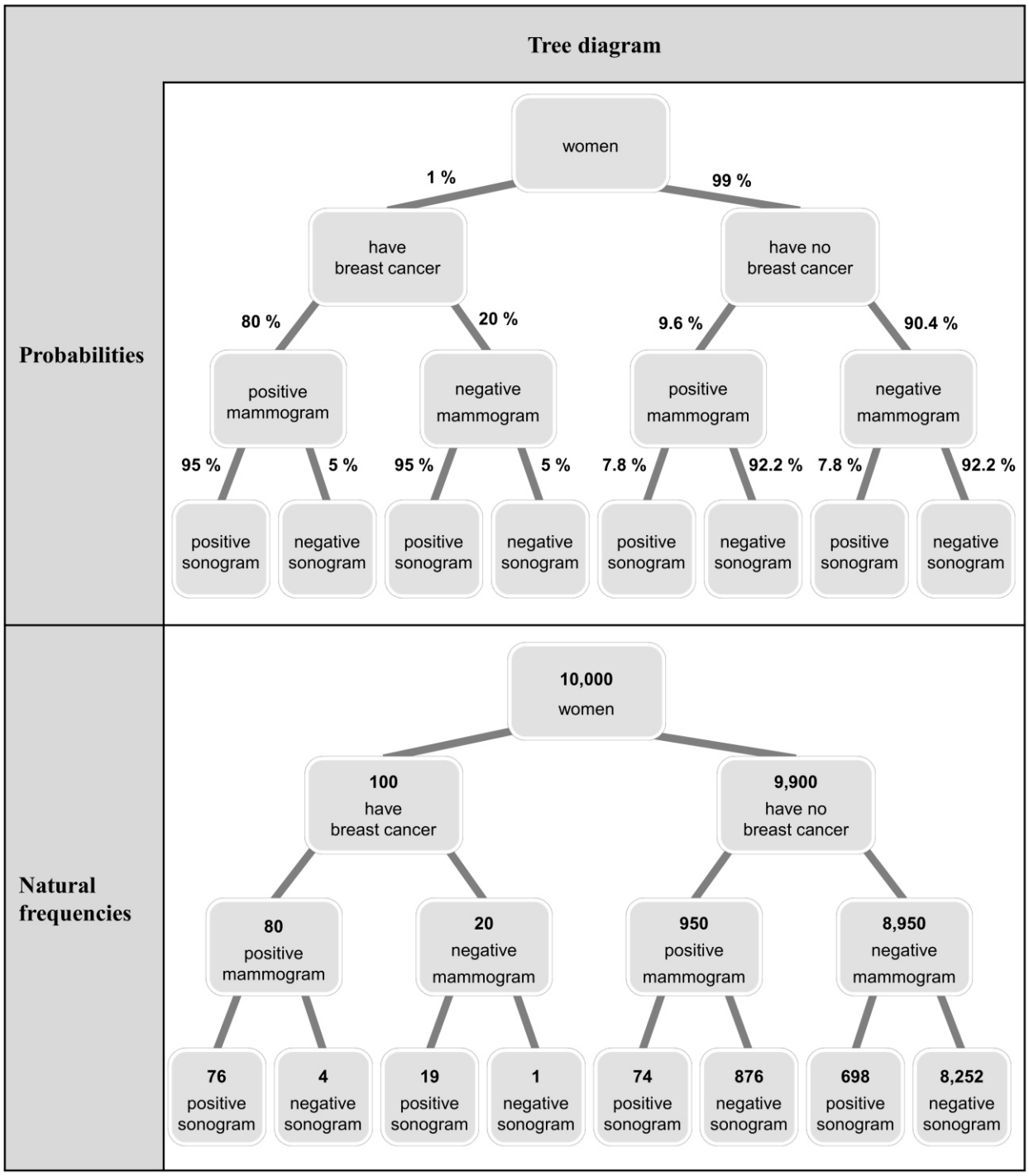


FIGURE 2. Probability and natural frequency tree of a 2-test case (implemented in studies 1 and 2).

Furthermore, there is another notable feature in tree diagrams that argues for choosing them for our empirical study: if the aim of the visualization is to illustrate the typical *conditional* probabilities of Bayesian reasoning tasks, the tree diagram is the only (!) possibility for visualizing the numbers of both the frequency format and the probability format. Let us explain this argument regarding the Eikosogram (Fig 1D), where the implemented numbers are frequencies (the sum of all numbers is 100). Of course these frequencies can be replaced by probabilities by simply adding the percent symbol after every number (if the sum, say N, were unequal to 100,

probabilities could be derived by dividing by N). Yet it has to be noted that these eight percentage points then display *conjoint probabilities* but not *conditional probabilities*, which are predominantly displayed in Bayesian reasoning tasks (compare the versions above).

Interestingly, there is no intuitive way to display conditional probabilities in any of the other diagrams because there is no branch (or similar prominent place) for them (where should conditional probabilities be placed in Fig 1A–Fig 1E?). Since the diagnostic information of medical tests is usually presented in terms of sensitivities and false-alarm rates (or specificities; see [52]), this is a significant problem, especially if the problem solver has to construct the visualization on his or her own. This feature, namely that all numbers of typical Bayesian diagnostic situations can be directly implemented into tree diagrams, is an enormous advantage with respect to *teaching* statistics.

In addition, with reference to tree diagrams, it would be possible (and should be examined in a future empirical study) to provide probabilities in the branches (which are dominant in teaching statistics) and absolute numbers in the nodes simultaneously and therefore to present not only both formats, but also conjoint and conditional information in one visualization. With respect to Fig 2 this would mean adding the probabilities to the branches of the natural frequency tree or vice versa.

Note that all the arguments presented in favor of tree diagrams hold for 1-test cases as well as 2-test cases (or, of course, for cases with even more tests involved). In the following, let us focus on two details regarding the tree diagrams in Fig 2.

Redundancy of information

It has to be noted that both the textual wording and the tree diagram already contain all of the information that is needed in order to solve Bayesian reasoning problems (given conditional independence; see S1 Appendix). Consequently, the question arises as to whether (a) only the wording, (b) only the tree diagram, or (c) both representations taken together best helps to solve the problem.

Cognitive Load Theory [53] and Cognitive Theory of Multimedia Learning [54] suggest that the representation of a textual wording in addition to a specific visualization might increase the extraneous cognitive load and thus might lead to poorer performance because of the redundancy principle [54]; however, the redundancy effect can reverse under certain conditions [55,56]. Similarly, results from a study of Micallef et al. [30] indicate that a visualization is only helpful when no (corresponding) textual information is additionally presented. In Study 1 we will address this issue of redundancy.

Diagrams contain more information than the textual wording

A closer comparison of the statistical information presented in the tree diagrams (Fig 2) and the textual wordings reveals that the tree diagram contains *more* information than the textual

wording. For example, while statistical information on persons with two negative test results are presented in the tree diagram, only statistical information on women with positive test results is provided in the text. Note that for the given question ("What is the probability of the disease given two positive test results?"), several branches of Fig 2 are indeed dispensable (for participant performance in alternative questions, see [57]). Thus it would be possible (a) to *highlight* both question-related branches or (b) even to *prune* the tree and only display those two relevant branches.

In cases (a) and (b), Cognitive Load Theory would suggest that according to the signaling principle, highlighting the relevant branches in the tree diagram (or even pruning the diagram by omitting the question-irrelevant branches) might improve participant performance [58,59]. However, the representation of unnecessary information could also increase the extraneous cognitive load; in that case, improved performance would be attained only with a pruned tree (since in a tree with highlighted branches the non-relevant branches would still be visible). Yet it has to be noted that only the full tree diagram allows the direct combining of numbers for *any* possible question that might be posed (e.g., "What is the probability of the disease given that test 1 is positive and test 2 is negative?" or vice versa). In Study 2 we focus on the issue of highlighting branches or pruning tree diagrams.

Research question

It should be noted that with respect to all three following research questions, we will compare probability versions (including probability trees) with natural frequency versions (including frequency trees).

1. What is the effect of visualizing statistical information with a tree diagram in a Bayesian 2-test case (Study 1 and Study 2)?
2. Is it easier to solve a purely textual version, a purely visual version, or a version that presents the text and the tree diagram simultaneously (Study 1)?
3. Does it help to highlight relevant branches or even prune irrelevant branches instead of simply presenting a full tree diagram (Study 2)?

Study 1

Method

Participants.

A total of 190 medical students (56 men, 133 women, one person who gave no answer) at different stages of their medical education at University Hospital Regensburg were recruited in 2016. Participants' ages ranged from 18 to 41 years ($M = 23.1$, $SD = 3.3$). All students were informed that their participation was voluntary, and that anonymity was guaranteed. Participants had given

their prior written consent to participating in the study. The Review Board of University Hospital Regensburg confirmed that, for this kind of study, no ethical approval would be necessary.

Design and materials.

A paper-and-pencil questionnaire contained two successive Bayesian 2-test tasks. We implemented a 3×2×2 design with the factors *presentation of information* (text only vs. tree only vs. text and tree), *information format* (probabilities vs. natural frequencies) and *context* (breast cancer screening problem vs. HIV testing problem) (see also Table 1 and section "Procedure").

TABLE 1. Design of the twelve resulting problem versions implemented (Study 1).

		Context	
		Breast cancer screening problem	**HIV testing problem**
Information format	**Probabilities**	**Presentation of information** • Text only • Tree only • Text and tree	**Presentation of information** • Text only • Tree only • Text and tree
	Natural frequencies	**Presentation of information** • Text only • Tree only • Text and tree	**Presentation of information** • Text only • Tree only • Text and tree

All versions began with a description of the medical situation (Table 2). After that, one of the six different presentations of information was provided. In the tree-only and text-and-tree versions, the tree diagrams of Fig 2 were implemented. Finally, the question was formulated in the same format as was used with the previous statistical information. The complete problem formulations can be seen in Table 2.

Procedure.

Each participant received one of the two problem contexts in probability format and the other problem context in natural frequency format, with the order of context and information format varied systematically. When one of the problems the participant worked on had a certain presentation of information (e.g., text only), the other problem contained one of the other remaining types of information presentation.

TABLE 2. Problem formulations for both contexts (breast cancer screening problem and HIV testing problem).

	Breast cancer screening problem		HIV testing problem	
	Probability version	**Natural frequency version**	**Probability version**	**Natural frequency version**
Medical situation	Imagine that you are a physician in a mammography screening center where women without symptoms are screened for breast cancer. In addition to mammograms, you frequently use sonograms as a supplementary medical test to detect breast cancer. At the moment, you are advising a woman who has no symptoms but who has received a positive result from her mammogram as well as a positive result from her sonogram. This woman wants to know what these results mean for her. For your answer, there is the following information available, which is based on a random sample of women who have all undergone a mammography and a sonography[1]:		Imagine that you are a physician in an AIDS information center. In addition to individual counseling interviews, your information center also provides HIV testing, for which two blood samples are taken: An ELISA test is conducted with the first blood sample. If the ELISA test is positive (indicating a possible HIV infection), a Western Blot test is ordered with the second blood sample. At the moment, you are advising a low-risk client who has received a positive result from the ELISA test as well as from the Western Blot test. This client wants to know what these results mean for him. For your answer, there is the following information available, which is based on a random sample of low-risk persons who have all undergone both the ELISA and the Western Blot test[1]:	
Presentation of information	• *Text only* • *Tree only* • *Text and tree*	• *Text only* • *Tree only* • *Text and tree*	• *Text only* • *Tree only* • *Text and tree*	• *Text only* • *Tree only* • *Text and tree*
Text	The probability of breast cancer for a woman with no symptoms is 1%. The probability that a woman with breast cancer will have a positive mammogram is 80%. The probability that a woman with breast cancer will have a positive sonogram is 95%. The probability that a woman without breast cancer will have a false-positive mammogram is 9.6%. The probability that a woman without breast cancer will have a false-positive sonogram is 7.8%. [1] Footnote: Assume for your calculations that the results of both tests are (statistically) independent for women with breast cancer as well as for women without breast cancer.	100 out of 10,000 women with no symptoms will have breast cancer. 80 out of 100 women with breast cancer will have a positive mammogram. 76 out of 80 women with breast cancer and a positive mammogram will have a positive sonogram. 950 out of 9,900 women without breast cancer will have a false-positive mammogram. 74 out of 950 women without breast cancer but with a positive mammogram will have a false-positive sonogram. [1] Footnote: Assume for your calculations that the results of both tests are (statistically) independent for women with breast cancer as well as for women without breast cancer.	The probability of an HIV infection for a low-risk client is 0.01%. The probability that an HIV-infected client will have a positive ELISA test result is 99.9%. The probability that an HIV-infected client will have a positive Western Blot test result is 99.8%. The probability that a client without HIV infection will have a false-positive ELISA test result is 0.4%. The probability that a client without HIV infection will have a false-positive Western Blot test result is 0.1%. [1] Footnote: Assume for your calculations that the results of both tests are (statistically) independent for HIV-infected clients as well as for clients who are not HIV-infected.	100 out of 1,000,000 low-risk clients are HIV-infected. 100 out of 100 HIV-infected clients will have a positive ELISA test result. 100 out of 100 HIV-infected clients with a positive ELISA test result will have a positive Western Blot test result. 4,000 out of 999,900 clients without an HIV infection will have a false-positive ELISA test result. 4 out of 4,000 clients without an HIV infection but with a positive ELISA test result will have a false-positive Western Blot test result. [1] Footnote: Assume for your calculations that the results of both tests are (statistically) independent for HIV-infected clients as well as for clients who are not HIV-infected.
Tree diagram	Probability tree (in the tree-only and in the text-and-tree version)	Natural frequency tree (in the tree-only and in the text-and-tree version)	Probability tree (in the tree-only and in the text-and-tree version)	Natural frequency tree (in the tree-only and in the text-and-tree version)
Question	What is the probability that a woman with both positive mammogram and positive sonogram actually has breast cancer? Answer: _____	How many of the women with both positive mammogram and positive sonogram actually have breast cancer? Answer: ____ out of ____	What is the probability that a client with both positive ELISA test and positive Western Blot test results is actually HIV-infected? Answer: _____	How many of the clients with both positive ELISA test and positive Western Blot test results are actually HIV-infected? Answer: ____ out of ____

Solutions of the problems.

The solution for the breast cancer screening problem is 76 out of 150, or about 50.7%. Note that the positive predictive value of about 50% corresponds to the actual values for women who participate in breast cancer screenings and receive positive results from a mammography as well as another non-invasive clarification (according to the latest evaluation report of the German Cooperative Association for Mammography [60]). For the HIV testing problem, the solution is 100 out of 104, or about 96.2%. Following Prinz et al. [6], the HIV testing problem uses a combined sensitivity (99.7%) and a combined specificity (99.9996%) of the ELISA test and the Western Blot test, resulting in a positive predictive value of about 96% when a prevalence of 0.01% is assumed (see also [61]).

It should be noted that in the medical 2-test case, the problem of conditional independence arises (see Footnote 1 in the Text section of Table 2). Readers interested in details concerning this issue can find more information in the S1 Appendix.

Coding.

In accordance with Gigerenzer and Hoffrage [18], we classified a response elicited from a probability version as correct if it was the exact Bayesian solution or rounded to the next whole percentage point above or below (i.e., in the breast cancer screening problem, all solutions between 50% and 51%, and in the HIV testing problem, all solutions between 96% and 97% were classified as correct). In the natural frequency versions, responses were classified as correct only if both numbers (e.g., in the breast cancer screening solution of "76 out of 150", both the 76 and the 150) were denoted correctly (a very conservative criterion regarding the natural frequency version; see also [28]).

Administration.

Students were examined in larger groups during their university lecture sessions. Trained administrators guaranteed a quiet atmosphere and professional supervision of the study. Students sitting next to each other always worked on different versions. Pocket calculators were distributed and students were allowed to use them at any point during the test. There were no time constraints for completing the questionnaire. Participants needed on average about 30 minutes total for both tasks.

Results

Study 1 yielded two important findings (Fig 3). First, students performed better when statistical information was presented in natural frequencies (36% correct inferences across context and presentation) rather than as probability versions (5% correct inferences across context and presentation). This finding holds true for both contexts and for all three presentation formats. Second, the addition of a tree diagram leads to higher performance rates (again holding true across all versions and conditions). One exception is the weaker performance observed with the

probability format in the HIV testing problem, which went from 6% with the text-only version to 0% with the text-and-tree version. However, we refrained from statistically comparing performance rates in probability versions because of the low achievement in all of these versions. Interestingly, in the natural frequency format, performance did not differ between tree-only and text-and-tree versions (i.e., when a frequency tree is provided, the additional text is neither harmful nor helpful).

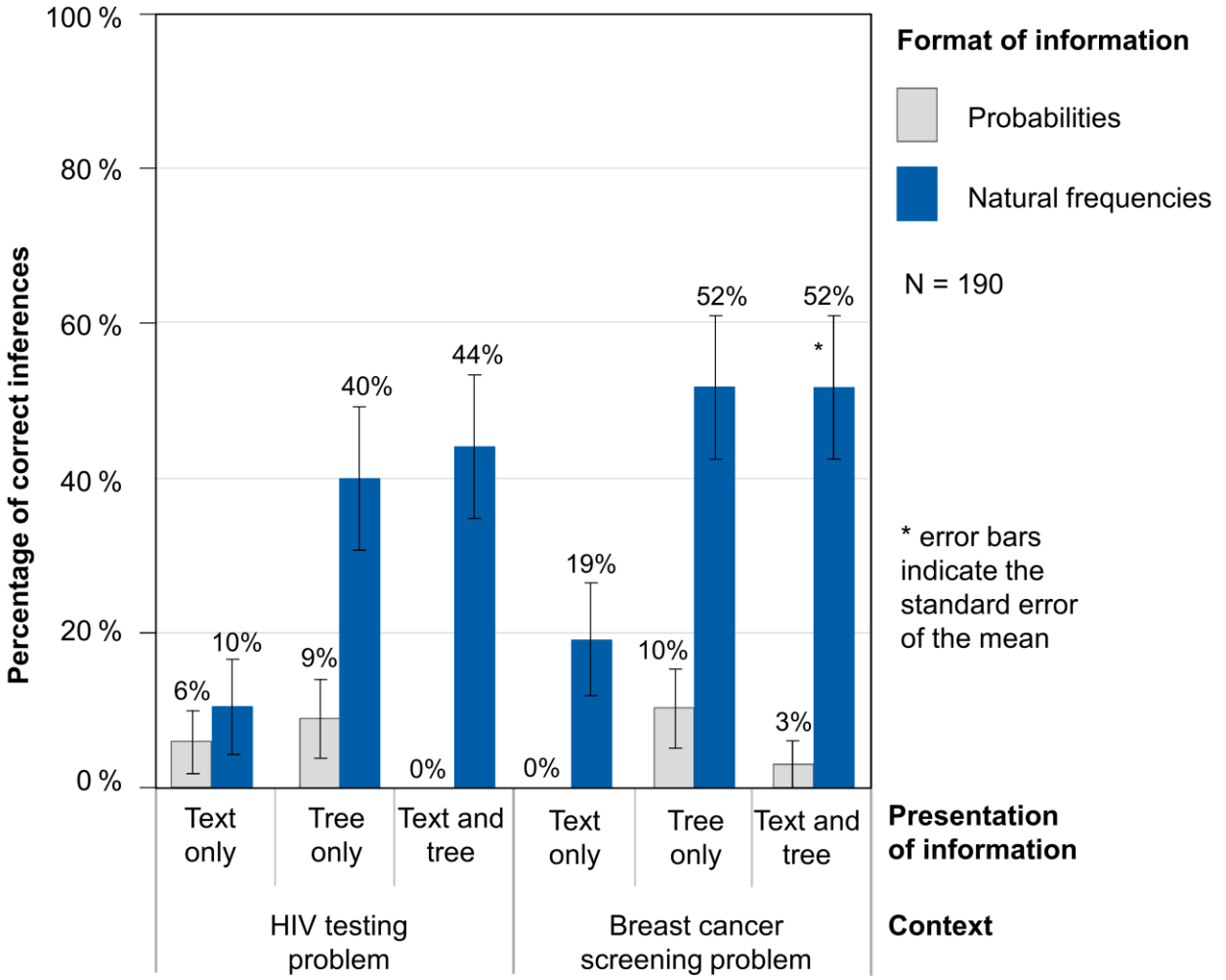


FIGURE 3. Percentages of correct inferences in Study 1.

Note that the advantage of "tree versions" (i.e., tree-only versions or text-and-tree versions) over text-only versions is much stronger with respect to natural frequency trees (47% vs. 15% correct inferences across both contexts) than it is with respect to probability trees (6% vs. 3%; see Fig 3). However, the weaker results obtained with probability trees could be instructive, since probabilities and probability trees are frequently used in statistical textbooks in both secondary schools and universities. Furthermore, participants performed descriptively slightly better in almost every version of the breast cancer screening problem (23% correct inferences) than in the respective versions of the HIV testing problem (18% correct inferences).

Since probability trees obviously do not foster insight within Bayesian reasoning situations, we will concentrate in the following on the results of the natural frequency versions. In order to analyze the effect of tree diagrams in natural frequency versions, we ran a generalized linear mixed model with a logit link function. In this model we specified the text-and-tree version as the reference version and included the possible explanatory factors "omitting tree" (i.e., text-only version) and "omitting text" (i.e., tree-only version) to predict the probability of a correct inference.

According to the results of the generalized linear mixed model, the probability of solving the text-and-tree version was 47.7% (unstandardized regression coefficient: $b_0 = -0.09$). The (unstandardized) regression coefficient for omitting the tree was significant ($b_1 = -1.68$, $SE = 0.44$, $z = -3.84$, $p < 0.001$), suggesting that the probability for solving the text-only version is reduced to only 14.5%. In contrast, omitting the text (i.e., using the tree-only version) leads to a non-significant regression coefficient ($b_2 = -0.07$, $SE = 0.35$, $z = -0.19$, $p = 0.85$), which implies that the probability of solving tree-only versions (46.0%) does not differ significantly from the probability of solving text-and-tree versions.

A closer inspection of the data revealed an additional effect of student high school's grade point average (the German *Abiturnote*). However, implementing grade point average in the generalized linear mixed model did not change the presented results substantially (omitting the tree diagram was still a significant factor and omitting the text was still non-significant). In order to exclude possible transfer effects (learning from the first task for the second task), we also implemented the position number of the task as an additional factor in the generalized linear mixed model. However, it turned out that participants performed even slightly (but not significantly) better if a particular task was located at the first position, which allowed us to exclude a possible transfer effect. Notably, when a tree diagram was provided, several participants marked the branches relevant to the question, which leads directly to Study 2.

Study 2

In the second study, we aimed to increase participant performance even more by providing different kinds of tree diagrams, that is, by highlighting the question-related branches in a special color or by pruning all branches but question-related ones. The three different tree diagrams that were implemented with respect to the breast cancer screening problem are shown in Fig 4. The respective probability versions of these tree diagrams were also tested in Study 2, of course.

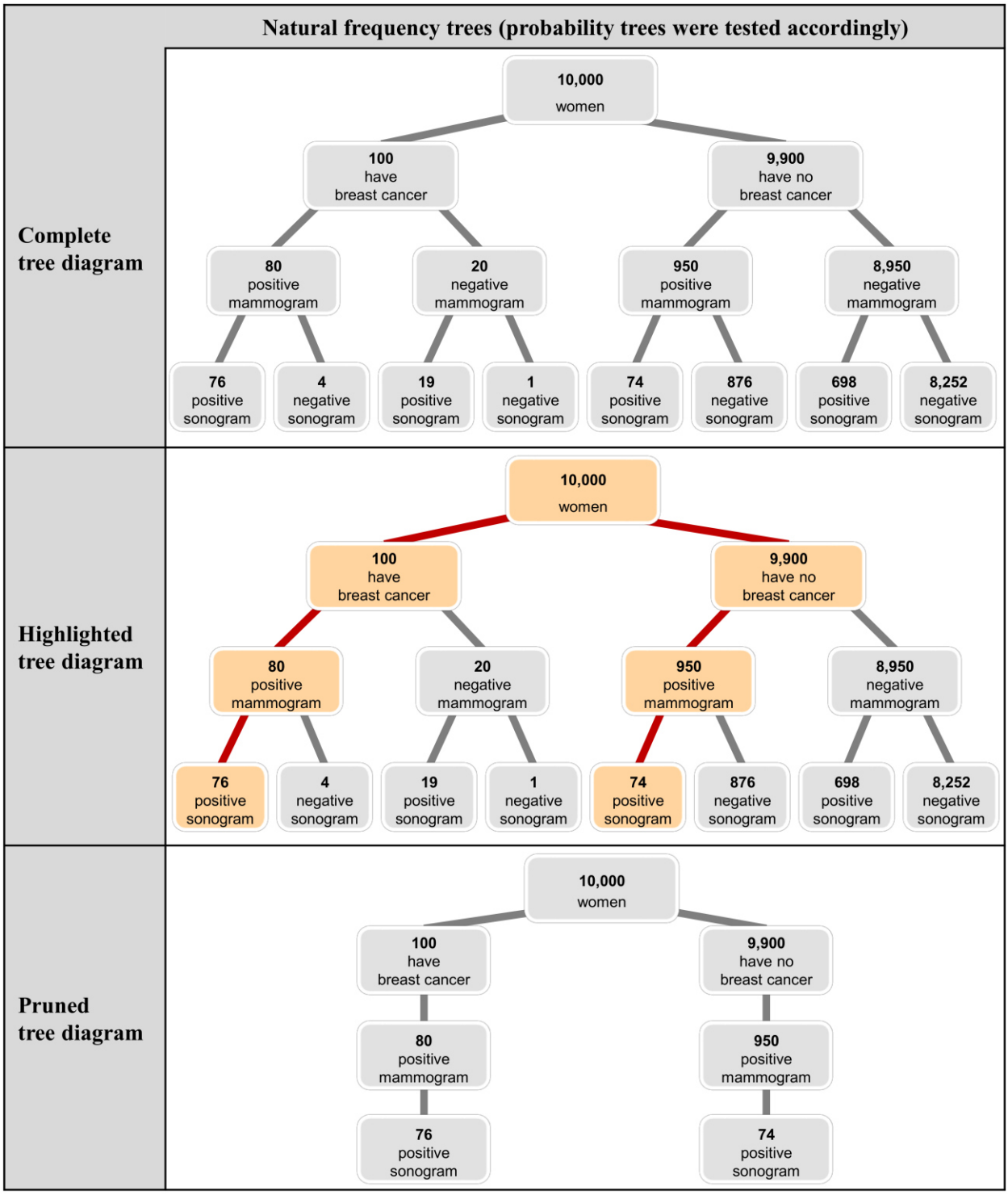


FIGURE 4. Three different tree diagrams with natural frequencies for the breast cancer screening problem (implemented in Study 2).

Method

Participants.

In all, 198 medical students (65 men, 133 women) at different stages of their medical education were recruited in 2016 from University Hospital Regensburg. Students who participated in Study 1 were excluded from taking part in Study 2. Participants' ages ranged from 18 to 38 years (M = 21.7, SD = 3.3). Again, all students were informed that their participation was voluntary and that anonymity was guaranteed. Participants had given their prior written consent to participating in the study. The Review Board of University Hospital Regensburg confirmed that no ethical approval would be necessary.

Design and materials.

A paper-and-pencil-questionnaire contained two successive Bayesian tasks (both 2-test cases). We used the same medical contexts (breast cancer screening and HIV testing) as in Study 1 in order to enable comparisons between Study 1 and Study 2. We implemented a 3×2×2 design with the factors *kind of tree diagram* (complete tree vs. highlighted tree vs. pruned tree), *information format* (probabilities vs. natural frequencies), and *context* (breast cancer screening problem vs. HIV testing problem) (see Table 3 and section "Procedure and administration").

TABLE 3. Design of the twelve resulting problem versions implemented (Study 2).

		Context	
		Breast cancer screening problem	**HIV testing problem**
Information format	**Probabilities**	**Kind of tree diagram** • Complete tree • Highlighted tree • Pruned tree	**Kind of tree diagram** • Complete tree • Highlighted tree • Pruned tree
	Natural frequencies	**Kind of tree diagram** • Complete tree • Highlighted tree • Pruned tree	**Kind of tree diagram** • Complete tree • Highlighted tree • Pruned tree

In light of the results obtained in Study 1, it had to be decided whether or not the statistical information should be presented in text form as well. Because the text-and-tree version produced the strongest student performance in Study 1, we decided to use this version in Study 2 as well in order to be conservative when investigating the beneficial effects of highlighting and pruning tree diagrams.

All versions began with the same medical situations used in Study 1. After the statistical information was provided, one of the three different kinds of tree diagrams was presented. Finally, the question was provided in the same format as the information in the text. Note that the complete-tree versions in Study 2 were identical to the text-and-tree version in Study 1.

Procedure and administration.

As in Study 1, each participant received one of the two problem contexts in probability format and the other in natural frequency format, again with the order of problem context and information format varying systematically. This time, the two problems each participant worked on had two out of the three different kinds of tree diagrams: *complete tree*, *highlighted tree* and *pruned tree*. For further details of the study administration, see Study 1.

Solutions of the problems and coding.

Since Study 1 and Study 2 did not differ in these two aspects, the respective solution and coding can be taken from Study 1. Again, readers interested in the issue of conditional independence can consult S1 Appendix.

Results

Study 2 produced three important findings (Fig 5). First, as in Study 1, student performance was substantially stronger when the statistical information in the problem was presented in natural frequencies (54% correct inferences across contexts and kinds of tree diagram) rather than probabilities (7% correct inferences). Because probability trees in Study 2 also did not constitute helpful visualizations (the maximum was 13% correct solutions; see Fig 5), we concentrate on natural frequency trees here again. Second, highlighting the two relevant branches of natural frequency trees leads to the highest performance rates, namely 67% (across contexts) as compared to 47% with the complete tree (not highlighted). Third, the use of a pruned tree does not improve Bayesian reasoning more than the use of a complete tree (both performance rates were 47% across contexts).

In order to analyze the effect of different kinds of tree diagrams in natural frequency versions, we again ran a generalized linear mixed model with a logit link function. In this model we specified the complete-tree version as the reference version (this version is identical to the text-and-tree version in Study 1) and included the possible explanatory factors *highlighting tree* and *pruning tree* to predict the probability of a correct inference.

According to the results of the generalized linear mixed model, the probability of solving the complete-tree version was 46.9% (unstandardized regression coefficient: $b_0 = -0.13$). The (unstandardized) regression coefficient for the highlighted tree was significant ($b_1 = 0.82$, $SE = 0.36$, $z = 2.26$, $p = 0.02$), suggesting that the probability for solving a version with a highlighted tree increased to even 66.7%. In contrast, pruning the irrelevant branches of the tree diagrams leads to a non-significant regression coefficient ($b_2 = 0.01$, $SE = 0.35$, $z = 0.02$, $p = 0.98$), which implies a probability for solving the task of 47.1% (comparable to the complete-tree version). In summary, highlighting the relevant branches and simultaneously presenting the complete situation can foster insight.

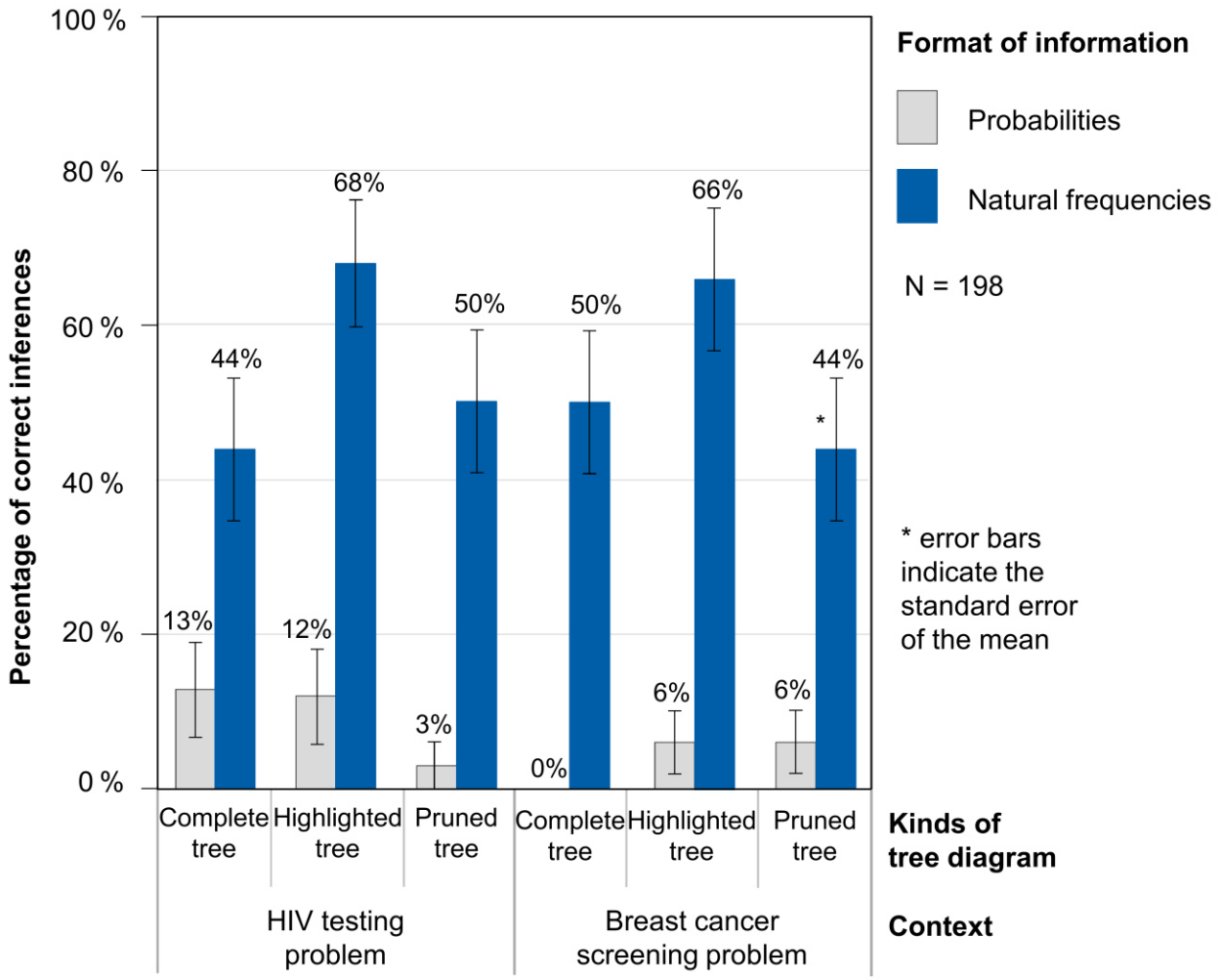


FIGURE 5. Percentages of correct inferences in Study 2.

Moreover, in Study 2 there was an additional effect of the position number of the solved tasks. All versions placed in the first position were again solved better than the identical tasks placed in the second position. In contrast to Study 1, this factor was even significant. However, implementing the position number of the task in the generalized linear mixed model did not change the presented results substantially (highlighting the tree diagram was still a significant factor and pruning the tree was still non-significant). Therefore, we can again exclude transfer effects in Study 2. Whereas in Study 1 grade point average but not position number had a significant effect, the opposite was the case in Study 2 (however, both effects did not affect the main results). Furthermore, *context* (which was not a factor of interest) did not change the results substantially in either study.

Because the text-and-tree versions of Study 1 were identical to the complete-tree versions of Study 2 (each for both contexts and both formats as well), and since the performance of the participants in these versions was comparable in both studies, we assume similar competencies in both subsamples. Therefore, it seems reasonable to compare performances between the two single studies.

Recommendations for fostering insight

Taken together, Studies 1 and 2 suggest three strategies that can be recommended to stimulate insight within Bayesian reasoning situations: (1) replace probabilities by natural frequencies, (2) create a natural frequency tree, and (3) highlight the two question-relevant branches in the natural frequency tree.

General discussion

Both studies (1) replicated earlier findings that—regardless of visualization—natural frequency versions could be solved much more easily than probability versions (e.g., for 1-test cases see [18,19,47,62], and for 2-test cases see [51]). The new results demonstrated that (2) natural frequency trees but not probability trees were substantially helpful and that (3) highlighting the question-related branches in a natural frequency tree can additionally improve performance, but pruning the tree does not.

Since in all implemented probability versions participant performance ranged from 0% to only 13% (across both studies), probability tree diagrams clearly do not qualify as visualizations that stimulate great insight within Bayesian reasoning situations. Because the focus of the present article is not the reinvestigation of format effects (the probability versions only served as control versions) but the boosting of participant performance, we will concentrate in the remaining discussion on natural frequency trees.

Considering the Cognitive Load Theory [53] and the Cognitive Theory of Multimedia Learning [54], two results here are remarkable: (1) text-and-tree versions and tree-only versions (Study 1) can both be solved with similarly little effort, and (2) pruning irrelevant branches (Study 2) does not help participants, probably because the situation as a whole is no longer fully presented. Neither finding supports the hypothesis that the extraneous cognitive load is increased by (a) presenting text and tree simultaneously or (b) presenting information that is not directly relevant to the question at hand. Yet highlighting the question-related branches (while still showing the irrelevant branches) was of greatest help for participants in Bayesian reasoning situations, therefore supporting the signaling principle with respect to frequency trees [58,59].

Thus, highlighted natural frequency trees are the best recommendation for teaching statistics (in secondary schools and at universities) and for communicating risks (e.g., in the medical domain). With respect to medical decision making, understanding the meaning of medical test results is crucial for medical students and physicians as well as for patients, because it can reduce the possible harms of overdiagnosis and overtreatment but can also reduce the danger of serious diseases being overlooked.

Frequency trees can easily be constructed and, if need be, also extended to situations with more than one hypothesis (e.g., several possible diseases), to non-binary test results or symptoms (e.g., unclear test results or symptoms), or to situations where even more than two tests (or symptoms)

are provided [51]. Furthermore, besides the described *causal trees* (first split the sample into patients with the disease and without the disease and then split these two nodes into sets with respect to the test result), *diagnostic trees* including natural frequencies can be constructed (i.e. first split the sample with respect to the test result and then with respect to the disease) [63–65].

Interestingly, in both studies, performance did not depend on the students' level of medical education, which indicates that statistical education is not sufficiently implemented in the training of medical students. However, it has to be noted that we did not run a training study, and thus our results suggest that natural frequency trees are effective even in the absence of prior instruction. Consequently, natural frequency trees can be directly used by patients and physicians and hence should be implemented in medical textbooks and in statistics education materials for prospective physicians, thus making this helpful communication tool available to both physicians and patients.

Acknowlegdements

The authors would like to thank all participating students for their contribution, Robert DeHaney and Francis Lorie for editing the manuscript, Sven Hilbert for his statistical advice and Editor (Gerd Gigerenzer), Ulrich Hoffrage and an anonymous reviewer for their helpful comments. This publication was supported by the German Research Foundation (DFG) within the funding program Open Access Publishing.

References

1. Operskalski JT, Barbey AK. Risk literacy in medical decision-making. Science. 2016; 352: 413–414. pmid:27102467

2. Wegwarth O, Gigerenzer G. Overdiagnosis and overtreatment. Evaluation of what physicians tell their patients about screening harms. JAMA Intern Med. 2013; 173: 2086–2087. pmid:24145597

3. Jorgensen KJ, Gotzsche PC. Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. BMJ. 2009; 339: b2587. pmid:19589821

4. Bhattacharya R, Barton S, Catalan J. When good news is bad news. Psychological impact of false positive diagnosis of HIV. AIDS Care. 2008; 20: 560–564. pmid:18484325

5. Stine GJ. Acquired immune deficiency syndrome. Biological, medical, social, and legal issues. Englewood Cliff, NJ: Prentice Hall; 1996.

6. Prinz R, Feufel M, Gigerenzer G, Wegwarth O. What counselors tell low-risk clients about HIV test performance. Current HIV research. 2015; 13: 369–380. pmid:26149159

7. Gigerenzer G, Hoffrage U, Ebert A. AIDS Counselling for low-risk clients. AIDS Care. 1998; 10: 197–211. pmid:9625903

8. Ellis KM, Brase GL. Communicating HIV results to low-risk individuals. Still hazy after all these years. Current HIV research. 2015: 381–390. pmid:26149160

9. Reimer L, Mottice S, Schable C, Sullivan P, Nakashima A, Rayfield M, et al. Absence of detectable antibody in a patient infected with human immunodeficiency virus. Clinical Infectious Diseases. 1997; 25: 98–100. pmid:9243042

10. Gigerenzer G, Gray JAM. Launching the century of the patient. In: Gigerenzer G, Gray JAM, editors. Better doctors, better patients, better decisions. Envisioning health care 2020. Cambridge, Mass.: MIT; 2011. pp. 3–28.

11. Eddy DM. Probabilistic reasoning in clinical medicine. Problems and opportunities. In: Kahneman D, Slovic P, Tversky A, editors. Judgment under Uncertainty: Heuristics and Biases. New York: Cambridge University Press; 1982. pp. 249–267.

12. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. Acad. Med. 1998; 73: 538–540. pmid:9609869

13. Garcia-Retamero R, Hoffrage U. Visual representation of statistical information improves diagnostic inferences in doctors and their patients. Soc. Sci. Med. 2013; 83: 27–33. pmid:23465201

14. Binder K, Krauss S, Bruckmaier G. Effects of visualizing statistical information. An empirical study on tree diagrams and 2 × 2 tables. Front Psychol. 2015; 6. pmid:26379569

15. Ellis KM, Cokely ET, Ghazal S, Garcia-Retamero R. Do people understand their home HIV test results? Risk literacy and information search. Proc. Hum. Fact. Ergon. Soc. Annu. Meet. 2014; 58: 1323–1327.

16. Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. Communicating statistical information. Science. 2000; 290: 2261–2262. pmid:11188724

17. Hoffrage U, Hafenbrädl S, Bouquet C. Natural frequencies facilitate diagnostic inferences of managers. Front Psychol. 2015; 6: 642. pmid:26157397

18. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. Psychol. Rev. 1995; 102: 684–704.

19. Siegrist M, Keller C. Natural frequencies and Bayesian reasoning. The impact of formal education and problem context. J. Risk Res. 2011; 14: 1039–1055.

20. Krauss S, Wang XT. The psychology of the monty hall problem. Discovering psychological mechanism for solving a tenacious brain teaser. J. Exp. Psychol. Gen. 2003; 132: 3–22. pmid:12656295

21. Hertwig R, Gigerenzer G. The 'conjunction fallacy' revisited. How intelligent inferences look like reasoning errors. J. Behav. Decis. Making. 1999; 12: 275–305.

22. Tversky A, Kahneman D. Extensional versus intuitive reasoning. The conjunction fallacy in probability judgment. Psychological Review. 1983; 90: 293–315.

23. Sedlmeier P, Gigerenzer G. Teaching Bayesian reasoning in less than two hours. J. Exp. Psychol. Gen. 2001; 130: 380–400. pmid:11561916

24. Zhu L, Gigerenzer G. Children can solve Bayesian problems. The role of representation in mental computation. Cognition. 2006; 98: 287–308. pmid:16399266

25. McDowell M, Jacobs P. Meta-Analysis of the Effect of Natural Frequencies on Bayesian Reasoning. Psychol Bull. 2017; 143: 1273–1312. pmid:29048176

26. Barbey AK, Sloman SA. Base-rate respect. From ecological rationality to dual processes. Behav Brain Sci. 2007; 30: 241–297. pmid:17963533

27. Sirota M, Kostovičová L, Vallée-Tourangeau F. Now you Bayes, now you don't. Effects of set-problem and frequency-format mental representations on statistical reasoning. Psychon Bull Rev. 2015. pmid:25711182

28. Pighin S, Gonzalez M, Savadori L, Girotto V. Natural frequencies do not foster public understanding of medical test results. Med Decis Making. 2016; 36: 686–691. pmid:27034447

29. Sloman SA, Over D, Slovak L, Stibel JM. Frequency illusions and other fallacies. Organ Behav Hum Decis Process. 2003; 91: 296–309.

30. Micallef L, Dragicevic P, Fekete J-D. Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. IEEE Trans. Vis. Comput. Graph. 2012; 18: 2536–2545. pmid:26357162

31. Sirota M, Kostovičová L, Juanchich M. The effect of iconicity of visual displays on statistical reasoning. Evidence in favor of the null hypothesis. Psychon Bull Rev. 2014; 21: 961–968. pmid:24307248

32. Brase GL. The power of representation and interpretation. Doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. J. Cogn. Psychol. 2014; 26: 81–97.

33. Yamagishi K. Facilitating normative judgments of conditional probability. Frequency or nested sets. Exp Psychol. 2003; 50: 97–106. pmid:12693194

34. Garcia-Retamero R, Cokely ET, Hoffrage U. Visual aids improve diagnostic inferences and metacognitive judgment calibration. Front Psychol. 2015; 6. pmid:26236247

35. Cosmides L, Tooby J. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition. 1996; 58: 1–73.

36. Böcherer-Linder K, Eichler A. The impact of visualizing nested sets. An empirical study on tree diagrams and unit squares. Front Psychol. 2016; 7: 2026. pmid:28123371

37. Oldford RW, Cherry WH. Picturing probability. The poverty of venn diagrams, the richness of eikosograms.

38. Pfannkuch M, Budgett S. Reasoning from an Eikosogram. An exploratory study. Int. J. Res. Undergrad. Math. Ed. 2016.

39. Talboy AN, Schneider SL. Improving accuracy on Bayesian inference problems using a brief tutorial. J. Behav. Dec. Making. 2016; 30: 373–388.

40. Brase GL. Pictorial representations in statistical reasoning. Appl. Cogn. Psychol. 2008; 23: 369–381.

41. Zikmund-Fisher BJ, Witteman HO, Dickson M, Fuhrel-Forbis A, Kahn VC, Exe NL, et al. Blocks, ovals, or people? Icon type affects risk perceptions and recall of pictographs. Med Decis Making. 2014; 34: 443–453. pmid:24246564

42. Steckelberg A, Balgenorth A, Berger J, Mühlhauser I. Explaining computation of predictive values: 2 x 2 table versus frequency tree. A randomized controlled trial [ISRCTN74278823]. BMC Med Educ. 2004; 4: 13. pmid:15301689

43. Friederichs H, Ligges S, Weissenstein A. Using tree diagrams without numerical values in addition to relative numbers improves students' numeracy skills: A randomized study in medical education. Med Decis Making. 2014; 34: 253–257. pmid:24085290

44. Budgett S, Pfannkuch M, Franklin C. Building conceptual understanding of probability models. Visualizing chance. In: Hirsch CR, McDuffie AR, editors. Annual Perspectives in Mathematics Education 2016. Mathematical Modeling and Modeling Mathematics. Reston, VA: Natl Coun Teachers Math; 2016. pp. 37–49.

45. Khan A, Breslav S, Glueck M, Hornbæk K. Benefits of visualization in the mammography problem. International Journal of Human-Computer Studies. 2015; 83: 94–113.

46. Spiegelhalter D, Pearson M, Short I. Visualizing uncertainty about the future. Science. 2011; 333: 1393–1400. pmid:21903802

47. McDowell M, Jacobs P. Meta-Analysis of the effect of natural frequencies on Bayesian reasoning. 2017.

48. Johnson ED, Tubau E. Words, numbers, & numeracy. Diminishing individual differences in Bayesian reasoning. Learning and Individual Differences. 2013; 28: 34–40.

49. Garcia-Retamero R, Hoffrage U, Dieckmann A. When one cue is not enough: combining fast and frugal heuristics with compound cue processing. Q J Exp Psychol (Hove). 2007; 60: 1197–1215. pmid:17676553

50. Krauss S, Martignon L, Hoffrage U. Simplifying Bayesian Inference: The General Case. In: Nea Magnani, editor. Model-based Reasoning in Scientific Discovery; 1999. pp. 165–179.

51. Hoffrage U, Krauss S, Martignon L, Gigerenzer G. Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. Front Psychol. 2015; 6: 1473. pmid:26528197

52. McGee SR. Evidence-based physical diagnosis. 3rd ed. Philadelphia: Elsevier/Saunders; 2012.

53. Sweller J. Evolution of human cognitive architecture. Psychology of Learning and Motivation. 2003; 43: 215–266.

54. Mayer RE. Cognitiv theory of multimedia learning. In: Mayer RE, editor. The Cambridge handbook of multimedia learning. Cambridge, U.K., New York: Cambridge University Press; 2005. pp. 31–48.

55. Sweller J. The redundancy principle in multimedia learning. In: Mayer RE, editor. The Cambridge handbook of multimedia learning. Cambridge, U.K., New York: Cambridge University Press; 2005. pp. 159–167.

56. Rey GD, Buchwald F. The expertise reversal effect. Cognitive load and motivational explanations. J Exp Psychol Appl. 2011; 17: 33–48. pmid:21443379

57. Binder K, Krauss S. Generalizations of the Bayesian reasoning paradigm. submitted.

58. Mautone PD, Mayer RE. Signaling as a cognitive guide in multimedia learning. Journal of Educational Psychology. 2001; 93: 377–389.

59. Mayer RE. Applying the science of learning: evidence-based principles for the design of multimedia instruction. Am Psychol. 2008; 63: 760–769. pmid:19014238

60. Jahresbericht Evaluation 2013. Deutsches Mammographie-Screening-Programm. Kooperationsgemeinschaft Mammographie. [German mammography screening programme—cooperative association mammography]; 2016.

61. Chou R, Huffman LH, Fu R, Smits AK, Korthuis PT. Screening for HIV. A review of the evidence for the U.S. preventive services task force. Ann Intern Med. 2005; 143: 55. pmid:15998755

62. Mandel DR, Navarrete G. Editorial: Improving Bayesian reasoning: What works and why. Front Psychol. 2015; 6: 1872. pmid:26696936

63. Woike JK, Hoffrage U, Martignon L. Integrating and Testing Natural Frequencies, Naïve Bayes, and Fast-and-Frugal Trees. Decision. 2017.

64. Martignon L, Vitouch O, Takezawa M, Forster MR. Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. In: Hardman D, Macchi L, editors. Thinking: Psychological perspectives on reasoning, judgment and decision making: Wiley; 2003.

65. Wu CM, Meder B, Filimon F, Nelson JD. Asking Better Questions: How Presentation Formats Influence Information Search. J Exp Psychol Learn Mem Cogn. 2017. pmid:28318286

Supporting information

S1 Appendix. Conditional independence

Alternative presentations of the 2-test case

An alternative way to present statistical information in a 2-test case were providing the sensitivity or the specificity of the whole testing procedure (e.g., a combined sensitivity and a combined specificity), which would reduce it to a 1-test case. However, in the literature on evidence-based medicine, sensitivities or specificities are usually only provided for a single test, and there is little available information about combined sensitivities or combined specificities because, in most medical test procedures, there is no standard test sequence that is followed consistently (further possibilities to extent the 1-test case to the 2-test case can be found in [1]).

If sensitivities and specificities are presented for each single test in the 2-test case, the question of the conditional independence of the involved tests arises. Because in probability theory conditional independence is based on the simpler situation of a statistical independence of two events, we first address the latter concept.

Statistical independence of two events

Note that the two events A and B are (statistically) independent if and only if $P(A \cap B)=P(A)\cdot P(B)$ or, equivalently, $P(A|B)=P(A)$. This means that the occurrence of one event does not affect the probability of the occurrence of the other event.

Also note that in the 1-test case, disease and test results (both considered events) are obviously statistically dependent, which is why this is not an issue in the 1-test case.

Conditional independence of two events given a third event

For three events the question of conditional independence arises. The two events A and B are conditionally independent *given* C if and only if

$P(A \cap B|C)=P(A|C)\cdot P(B|C)$

This is equivalent to

$P(A|B \cap C)=P(A|C)=P(A|\neg B \cap C)$

and to

$P(B|A \cap C)=P(B|C)=P(B|\neg A \cap C)$

Thus conditional independence of the events A and B given C means that the probability of A given C is not affected by the simultaneous occurrence of the event B (or *not* B). The same holds true if A and B are interchanged.

Note that it is possible that the events A and B are conditionally independent given C, but not conditionally independent given the complementary event *not* C.

Conditional independence in medical Bayesian 2-test cases

In our study, the question of the conditional independence of the two test results occurs both in the case of healthy people and in the case of people with a certain disease *D*. The conditional independence of the two test results

(1) given that the person has the disease (*D*) means

$P(T_2+|T_1+\cap D)=P(T_2+|D)$

(2) given that the person does not have the disease means

$P(T_2+|T_1+\cap\neg D)=P(T_2+|\neg D)$

Consequently, depicting the original sensitivity of the second test directly at the lower branch of a probability tree is only possible in the case of the conditional independence of both test results involved for people with the disease

Medical reality

In real-life medical situations, it cannot be assumed as a matter of course that two successive medical tests are conditionally independent both for individuals with the disease and individuals without the disease: For instance, a false-positive HIV test result might occur because of a chronic hepatitis B infection. Imagine that a person who is not HIV infected but who is infected with hepatitis B receives a (false!) positive test result from an HIV test because of the hepatitis B. If a second HIV test is then conducted, the probability that the second test will also be positive is higher than if there has been no first positive test result, because the reason for the false-positive result also exists in the second test.

As a consequence, when both tests applied are the same (e.g., a second ELISA test given after a positive ELISA test result), or are at least based on similar medical methods, conditional dependence should be assumed. However, the less similar the two test procedures are, the more likely they are not to affect each other (e.g., see [2]).

Concerning the contexts implemented in our study, the medical students we observed would assume the "independence" of both test procedures (probably without being able to provide a formula or distinguish between "independence" and "conditional independence"). Interestingly, the conditional independence of medical tests seems not to be considered an important issue in the field of medicine today.

Experimental decisions for our study

How should the natural frequency version look?

According to the natural sampling paradigm [3], the textual information in our frequency versions should match the sampling process that physicians utilize in real-life situations, namely performing one test after the other in a typical sequence. Thus the absolute numbers sampled represent the sensitivity and the false-alarm rate in an ecologically valid way. These obtained

absolute numbers are perfectly represented by the sequential partitioning of patients in the textual natural frequency versions and in the tree diagram as well. Therefore, both our 2-test case textual formulation and the tree diagrams with natural frequencies grow out of the idea of the natural frequency approach of Gigerenzer and Hoffrage [4].

How should the probability version look?

We declined to present the sensitivity and the false alarm rate of the second test in the textual formulation as conditional upon the result of the first test for the following reasons.

In a pilot study, we tried to implement the information, for example, presenting the sensitivity of the sonography as conditional upon the mammography test result ("The probability that a woman with breast cancer will have a positive sonogram, given that she has already had a positive mammogram, is 95%" as compared to "The probability that a woman with breast cancer will have a positive sonogram is 95%"). However, this led to even weaker performance because participants seemed to assume conditional independence anyway. For them, the supplement was misleading because stressing this condition suggests that the sensitivity of the second test would differ for different results of the first test. And if an event occurs with the same probability when given "A" as when given "not A", what is the point of mentioning "given A"?

In addition, there is another theoretical reason for not presenting combined sensitivities or sensitivities bases on previous test results: In evidence-based medicine, sensitivities, false alarm-rates, and specificities are almost never provided as conditional on other tests. As mentioned above, it is virtually impossible to find statistical information on sensitivities or specificities presented as conditional on previous test results. Therefore, our wording choice strengthens the ecological validity for the examined medical students.

However, in order to be mathematically correct, we added footnote 1 in all versions implemented in both studies (see Table 3).

References

1. Binder K, Krauss S. Generalizations of the Bayesian reasoning paradigm. submitted.

2. Shen Y, Wu D, Zelen M. Testing the Independence of Two Diagnostic Tests. Biometrics. 2001; 57: 1009–1017. doi: 10.1111/j.0006-341X.2001.01009.x.

3. Kleiter GD. Natural sampling: Rationality without base rates. In: Fischer GH, Laming, D. R. J, editors. Contributions to Mathematical Psychology, Psychometrics, and Methodology. New York: Springer; 1994. pp. 375–388.

4. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. Psychol. Rev. 1995; 102: 684–704. doi: 10.1037/0033295X.102.4.684.

Weitere Generalisierungen (Artikel 3, JEP: General)

Inhaltliche Schwerpunktsetzung des JEP-Artikels

Der dritte Artikel trägt den Titel *Generalizations of the Bayesian reasoning paradigm* und wurde kürzlich im englischsprachigen APA-Journal *Journal of Experimental Psychology: General* eingereicht.

Ebenso wie bereits im PlosONE-Artikel wurden die Erkenntnisse aus dem Frontiers-Artikel in die Domäne Medizin übertragen. Neben den ausführlich analysierten medizinischen 2-Test-Fällen, die im PlosONE-Artikel fokussiert wurden, werden nun zusätzlich noch weitere Diagnosesituationen betrachtet, die im medizinischen Alltag von Relevanz sind: Situationen, in denen drei Testergebnisse vorliegen; Situationen, in denen ein Test neben einem positiven und negativen Befund auch einen unklaren Befund liefern kann, und Situationen, in denen zwei verschiedene Erkrankungen mit demselben Test erkannt (aber nicht unterschieden) werden können.

In der bisherigen Forschung zu Bayesianischem Denken in der Medizin werden zumeist Fragen nach dem *positiven prädiktiven Wert* gestellt, also nach der Wahrscheinlichkeit erkrankt zu sein, wenn das Testergebnis positiv ist. Aber natürlich stellt sich beispielsweise auch ein Patient mit einem negativen Testergebnis die Frage, was dieses Testergebnis nun für ihn bedeutet. Derartige „alternative Diagnosen“ werden in diesem Artikel ebenfalls untersucht.

Die Forschungsfragen des Artikels lauten im Einzelnen:

Forschungsfrage 1: Unterstützen natürliche Häufigkeiten und Baumdiagramme das Verständnis für den positiven prädiktiven Wert im Bayesianischen 2-Test-Fall, im 3-Test-Fall, im Fall mit drei Testergebnissen, im Fall mit drei Hypothesen?
Forschungsfrage 2: Unterstützen natürliche Häufigkeiten und Baumdiagramme das Verständnis bei Fragen nach alternativen Diagnosen (z.B. die Wahrscheinlichkeit einer Erkrankung nach einem positiven und einem negativen Testergebnis) im Bayesianischen 2-Test-Fall, im 3-Test-Fall, im Fall mit drei Testergebnissen, im Fall mit drei Hypothesen?

Die Ergebnisse der Studie zeigen, dass natürliche Häufigkeiten und Häufigkeitsbäume das Verständnis im Bayesianischen 2-Test-Fall und im Bayesianischen 3-Test-Fall unterstützen, während in den Fällen mit drei Testergebnissen oder drei Hypothesen lediglich Häufigkeiten, nicht aber Häufigkeitsbäume hilfreich sind. Die Performanz bei den Wahrscheinlichkeitsvarianten bleibt durchgehend unter 20% und damit deutlich hinter den Varianten mit natürlichen Häufigkeiten. Für alternative Diagnosen ist die Darstellung eines Häufigkeitsbaumes dann besonders hilfreich, wenn die Frage stark von den dargebotenen Informationen abweicht.

Artikel 3: Generalizations of the Bayesian reasoning paradigm

Generalizations of the Bayesian reasoning paradigm

Karin Binder[1] & Stefan Krauss[1]

[1] *Mathematics Education, Faculty of Mathematics, University of Regensburg, Regensburg, Germany*

Abstract

In the present article we analyze three generalizations of the Standard Bayesian Reasoning Paradigm ("SBRP") in which one binary criterion (e.g., being ill or being healthy) has to be inferred from one binary predictor (e.g., a positive or negative medical test result). However, most decision-making processes, for instance in medical reality, involve more than one predictor (generalization 1) or more than one predictor or criterion value (generalization 2). It has to be noted that in such complex situations, not only are inferences based on positive test results relevant, but also inferences based on alternative test configurations (i.e., that involve negative or unclear test results; generalization 3). In an empirical study, 123 medical students from the Charité Berlin had to work on a 2-test and a 3-test scenario relating to generalization 1 and on one scenario with three criterion values and one with three predictor values relating to generalization 2. Including the additional questions that were posed in all four scenarios, a total of 1,353 nonstandard Bayesian inferences were analyzed. By systematically varying the factors *information format* (probabilities vs. natural frequencies) and *visualization* (no tree diagram vs. tree diagram) we found that natural frequencies were helpful with respect to all three generalizations, while tree diagrams were especially helpful with respect to generalization 1, when multiple-nested information had to be integrated. Specifically, when alternative questions (generalization 3) were posed in the 3-test scenario—and thus the present test configuration deviated from the information typically provided—natural frequency trees revealed their full potential.

Keywords: Bayesian reasoning, natural frequencies, natural frequency tree, medical decision-making

Probabilistic learning by evidence can be modeled by the Bayes formula, which describes the reevaluation ("updating") of the probability of a hypothesis in light of new data. Since Bayesian reasoning situations are of great relevance for various professions, such as medicine or law (Fenton, Neil, & Berger, 2016; Operskalski & Barbey, 2016; Prakken, 2014), extensive research addresses laypeople's and experts' abilities with this type of reasoning (Mandel & Navarrete, 2015; Barbey & Sloman, 2007), repeatedly documenting people's difficulties in dealing with statistical information (Pohl, 2017; Tversky & Kahneman, 1992). If someone judging a situation is actually a medical or a legal expert, misjudgments can have severe consequences for patients or clients (Operskalski & Barbey, 2016; Schneps & Colmez, 2013; Stine, 1996). As a result, several strategies for overcoming these difficulties have been proposed, such as the translation of statistical information into natural frequencies or rendering statistical information visually (see, e.g., the meta-analysis from McDowell & Jacobs, 2017). However, it must be noted that almost all empirical research on Bayesian reasoning addresses a very special case, where one binary criterion (e.g., having or not having a certain disease) must be inferred by one single binary predictor (e.g., a positive or negative test result). In the following, we call this kind of situation the *Standard Bayesian Reasoning Paradigm (SBRP)*, or specifically in the medical context, a *1-test scenario*. Figure 1 (disregarding the sections highlighted in gray) illustrates the SBRP of the famous mammography problem and both named strategies simultaneously.

Yet in reality, situations are often not quite so simple. For many diseases, there are instead standard procedures of applying a cascade of tests to confirm or reject an initial suspicion of the disease's presence (Karadawi et al., 2016). Similarly, typically more than one indicator is cited by a judge as an explanation for this or her verdict in a legal case because items of evidence in court are also not absolutely certain (see Saks & Koehler, 2005). Thus although in the research for this article we used experimental stimuli and participants from the medical domain, the arguments that follow hold equally for various professions and for experts and laypeople alike. We will now flesh out two different ways to generalize the SBRP.

Generalization 1. Consider, for instance, the usual procedure regarding routine screening for breast cancer, which is very expensive and heavily promoted in many countries for any woman in a specific age group (Gigerenzer & Gray, 2011). A positive mammogram does not definitively lead to a breast cancer diagnosis; rather, a second noninvasive test, such as a sonogram, is routinely applied (Ohuchi et al., 2016). Extending the scope of Bayesian reasoning by increasing the number of medical tests, we specify the first possible way of generalizing the SBRP (see middle column in Figure 2, read from top to bottom).

Generalization 2. Furthermore, in medical reality, test results may in fact be unclear (e.g., an ambiguous mammogram), or test results might be indicative of more than one disease (e.g., there are tests for diagnosing type 1 and type 2 diabetes that cannot distinguish between both diseases, Atkinson, Eisenbarth, & Michels, 2014). Thus another way of generalization (generalization 2)

would be to remain with the 1-test scenario but to increase the number of test or criterion values (see first line in Figure 2, read from the middle to the sides).

SBRP and 2-test case (with additional marked parts)

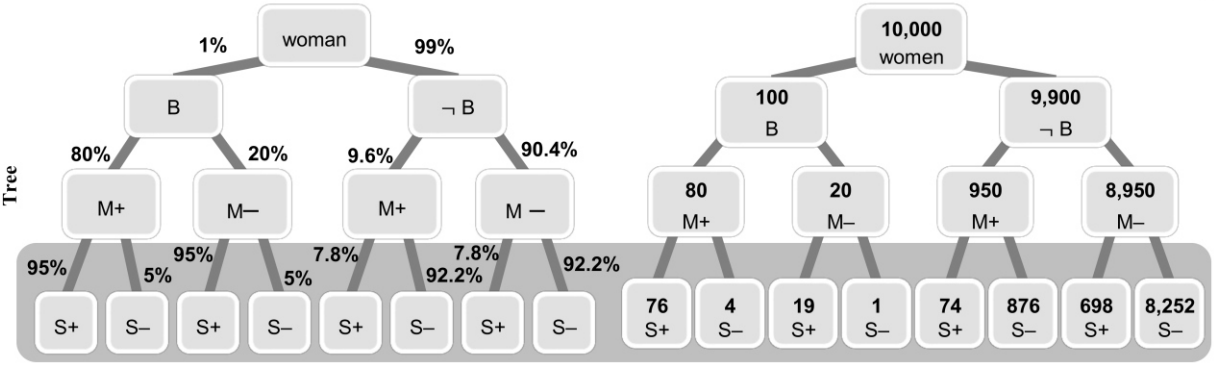
	Probability version	Natural frequency version
Medical situation	Imagine that you are a physician in a mammography screening center, where women without symptoms are screened for breast cancer. In addition to mammograms you frequently use a sonogram as a supplementary medical test to detect breast cancer. At the moment, you advise a woman with no symptoms who has had a positive mammogram as well as a positive sonogram. This woman wants to know what this means for her. For your answer, there is the following information available, which is based on a random sample of women who have all had a mammogram and a sonogram:	
Statistical information	The probability of breast cancer (B) for a woman without any symptoms is 1%.	100 out of 10,000 women without any symptoms have breast cancer (B).
	The probability that a woman with breast cancer will get a positive mammogram (M+) is 80%.	80 out of 100 women with breast cancer get a positive mammogram (M+).
	The probability that a woman with breast cancer and a positive mammogram will get a positive sonogram (S+) is 95%.	76 out of 80 women with breast cancer and a positive mammogram get a positive sonogram (S+).
	The probability that a woman without breast cancer wrongly will get a positive mammogram is 9.6%.	950 out of 9,900 women without breast cancer wrongly get a positive mammogram.
	The probability that a woman without breast cancer and with a positive mammogram wrongly will get a positive sonogram is 7.8%.	74 out of 950 without breast cancer and with a positive mammogram wrongly get a positive sonogram.
Tree		
Question	What is the probability that a woman with a positive mammogram and a positive sonogram actually has breast cancer?	How many of the women with a positive mammogram and a positive sonogram actually have breast cancer?
Solution	1-test case $\frac{80\%\cdot 1\%}{80\%\cdot 1\%+9.6\%\cdot 99\%} = 7.8\%$	$\frac{80}{80+76} = 7.8\%$
	2-test case $\frac{95\%\cdot 80\%\cdot 1\%}{95\%\cdot 80\%\cdot 1\%+7.8\%\cdot 9.6\%\cdot 99\%} = 50.7\%$	$\frac{76}{76+74} = 50.7\%$

FIGURE 1. A typical SBRP (without highlighted sections) and a corresponding 2-test case (including gray highlighted sections) in probability format (left) and natural frequency format (right), both illustrated by tree diagrams.

Generalization 3. There is a third way of generalization, however (generalization 3). The typical question posed in (medical) SBRPs is one regarding the probability of the disease (D) given a positive test result (T+), which is called the *positive predictive value* P(D | T+), although three alternative inferences might be relevant in 1-test scenarios, namely P(¬D | T+), P(D | T–) and

P(¬D | T–), with ¬D denoting the absence of the disease and T– denoting a negative test result. It has to be noted that based on the typical wording of medical SBRPs (where the sensitivity and the false-alarm rate of a specific medical test are provided), all four of these probabilities can be derived with comparable ease, so that considering these alternative inferences is not a difficult issue in the research on SBRP (for an exception see Armstrong, Spaniol, & Persaud, 2018). The more medical tests and the more possible test and criterion values involved, however, the more test configurations (e.g., test 1 positive, test 2 negative, test 3 unclear) there can be, and in consequence, the more conceivable questions that markedly deviate from the positive predictive value. Importantly, the complexity of the algorithm for answering such different questions can vary substantially, depending on the "fit" of the statistical information provided to the question posed (see subsection Generalization 3).

In sum, physicians and patients must often make momentous decisions about specific medical treatments and surgeries based upon multiple factors, including one or more medical test or symptom, ambivalent test results, and even negative test results (Karadawi et al., 2016; Vecchio, 1966). Medical high-stakes decisions, in particular, are only very rarely based on just an SBRP.

Rationale

Since both the SBRP itself and the natural frequency approach are documented and discussed at length (Barbey & Sloman, 2007; Brase, 2008a; Gigerenzer & Hoffrage, 1995; McDowell & Jacobs, 2017; Pighin, Gonzalez, Savadori, & Girotto, 2016; Siegrist & Keller, 2011; Sirota, Juanchich, & Hagmayer, 2014), we will not reopen these discussions for standard cases. For the sake of clarity, we consider the ecological rationality framework (Gigerenzer & Hoffrage, 1995; Todd & Gigerenzer, 2007) and the nested-set hypothesis (Girotto & Gonzalez, 2001) as two sides of the one coin and explicitly acknowledge the importance of both of them as there are more commonalities than divergences between the two (for further discussion see, e.g., McDowell & Jacobs, 2017; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002; McDowell & Jacobs, 2017; Sirota et al., 2014; Sloman, Over, Slovak, & Stibel, 2003). In fact, there have even been recent calls for theory integration (Brase & Hill, 2015; Johnson & Tubau, 2015; McNair, 2015).

We rather seek to make progress by contributing to the new field of nonstandard Bayesian reasoning paradigms (which are called "complex cases" by Hoffrage, Krauss, Martignon, & Gigerenzer, 2015). We especially focus on tree diagrams as visualization tools because of their potential to be broadened for use in *all* generalizations (generalization 1, generalization 2, and generalization 3).

In this article, we theoretically and empirically address the three aforementioned generalizations of the SBRP: First, we extend the number of medical tests (generalization 1). Second, we extend the number of predictor or criterion values (generalization 2). Third, we vary the configuration of predictor or criterion values in medical situations by asking not just for the probability of, say, being ill given that all tests conducted have come out positive, but also for being ill given that one

test has come out negative and two tests have come out positive (generalization 3). In our empirical approach, a sample of medical students had to work on four different nonstandard diagnosing situations (two following the structure of generalization 1 and two following the structure of generalization 2), with each situation containing questions on more than one specific test configuration (generalization 3). With respect to all three generalizations, we systematically varied two different strategies that actually fostered insight into the SBRP, namely 1) presenting natural frequencies instead of probabilities (Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer, 1998; McDowell & Jacobs, 2017; Siegrist & Keller, 2011); and 2) visualization of the information by tree diagrams (Binder, Krauss, & Bruckmaier, 2015; Sedlmeier & Gigerenzer, 2001).

Our study extends the work of Krauss, Martignon, and Hoffrage (1999), Hoffrage et al. (2015), and Binder, Krauss, Bruckmaier, and Marienhagen (2018), who already addressed the effect of natural frequencies in these kinds of "complex" Bayesian situations, but neither implemented tree diagrams nor addressed alternative test configurations systematically. Finally, in the General discussion, we provide an overview of all related studies conducted thus far (Table 10) and compare their respective results on nonstandard Bayesian reasoning. The distinguishing elements of this paper are that it presents for the first time 1) a detailed theoretical explication of various ways to extend an SBRP to a 2-test scenario (including a discussion of the notorious issue of conditional independence); 2) the systematic investigation of the additional impact of tree diagrams (both probability and natural frequency trees) in non-SBRPs over and above format effects (probabilities vs. natural frequencies); and 3) the theoretical and empirical consideration of inferences addressing alternative test configurations.

Generalizations of the SBRP

All three generalizations implemented can be illustrated by the structure of tree diagrams (Figure 2). Tree diagrams were chosen because they are appropriate for theoretical explication as well as empirical implementation. From a theoretical viewpoint, tree diagrams generally play an important role in theory formation regarding machine, but also human, multiple-cue decision-making (see e.g., Bröder, 2000; Garcia-Retamero, Hoffrage, & Dieckmann, 2007; Woike, Hoffrage, & Martignon, 2017). Similar kinds of trees are, for instance, diagnostic trees (Woike et al., 2017; Wu, Meder, Filimon, & Nelson, 2017), and fast and frugal trees (see, e.g., the famous tree of Green and Mehr on diagnosing heart attacks, Green & Mehr, 1997; or Jenny, Pachur, Lloyd Williams, Becker, & Margraf, 2013; Gigerenzer, 2008; Gigerenzer, Czerlinski, & Martignon, 1999). Our following theoretical explications based on tree diagrams can therefore be integrated within a superordinate research paradigm (see General discussion). From an empirical standpoint, tree diagrams were chosen because they are extremely helpful in the SBRP (Binder et al., 2015) and seem to be of use in more complicated situations as well (Yamagishi, 2003; Binder et al., 2018). It is interestingly to note that scholars presenting research articles on Bayesian reasoning to their peers often use tree diagrams rather than Euler diagrams, roulette-wheel diagrams, or other

diagrams in order to illustrate specific situations for their readers (e.g., Gigerenzer & Hoffrage, 1995; Kleiter, 1994; Mandel, 2014; Navarrete, Correia, & Froimovitch, 2014).

Two important observations with respect to *teaching and learning* Bayesian reasoning have to be noted: First, not only can tree diagrams be extended intuitively in any required direction, but they can also be constructed easily, and because of their "branch style", even very small base rates can be implemented numerically without difficulty (for different "visualization styles", see Khan, Breslav, Glueck, & Hornbæk, 2015). In contrast, visualizations following a "nested style" (e.g., Euler diagrams, see Micallef, Dragicevic, & Fekete, 2012) are based on comparisons of geometrical areas and thus require some effort to construct for new situations, and small base rates may not be "visible" at all (a comparison of different visualizations of the SBRP can be found in Binder et al., 2015). Finally, visualizations following a "frequency style" (e.g., icon arrays, see Brase, 2008b; Micallef et al., 2012) require that one denote one icon for each person (or object) involved and thus, for example, in the situation in Figure 1, 10,000 icons would have to be drawn (for combinations of styles, see, e.g., Spiegelhalter, Pearson, & Short, 2011). This easy adaptability to new circumstances makes tree diagrams a very effective tool for risk education and communication.

The second observation—which is also very relevant to teaching and learning—is that probabilities and natural frequencies can theoretically be placed *simultaneously* in a tree diagram (i.e., frequencies in the nodes and probabilities on the branches). In contrast, visualizations constructed in the frequency style (e.g., icon arrays) or in the nested style (e.g., Euler diagrams) do not have two such demarcated locations for placing each format or often do not involve numbers at all. Another type of representation, which is predominantly used in schools and universities—the 2×2-table—contains conjoint instead of conditional probabilities, and, again the cells can carry either natural frequencies or percentages (see Talboy & Schneider, 2016; Steckelberg, Balgenorth, Berger, & Mühlhauser, 2004; Fiedler, Brinkmann, Betsch, & Wild, 2000).

In Figure 2, generalization 1 is represented by the height of the tree (from top to bottom), generalization 2 by the width of the tree (from the center to the sides), and generalization 3 by considering alternative nodes within a tree diagram (e.g., those that deviate from exclusively positive test results). Note that the tree diagrams presented to our participants had concrete probabilities at the branches or corresponding absolute frequencies in the nodes and were each labelled with specific tests or diseases (for the tree diagrams actually implemented, see Figures 4, 5, 6, and 7).

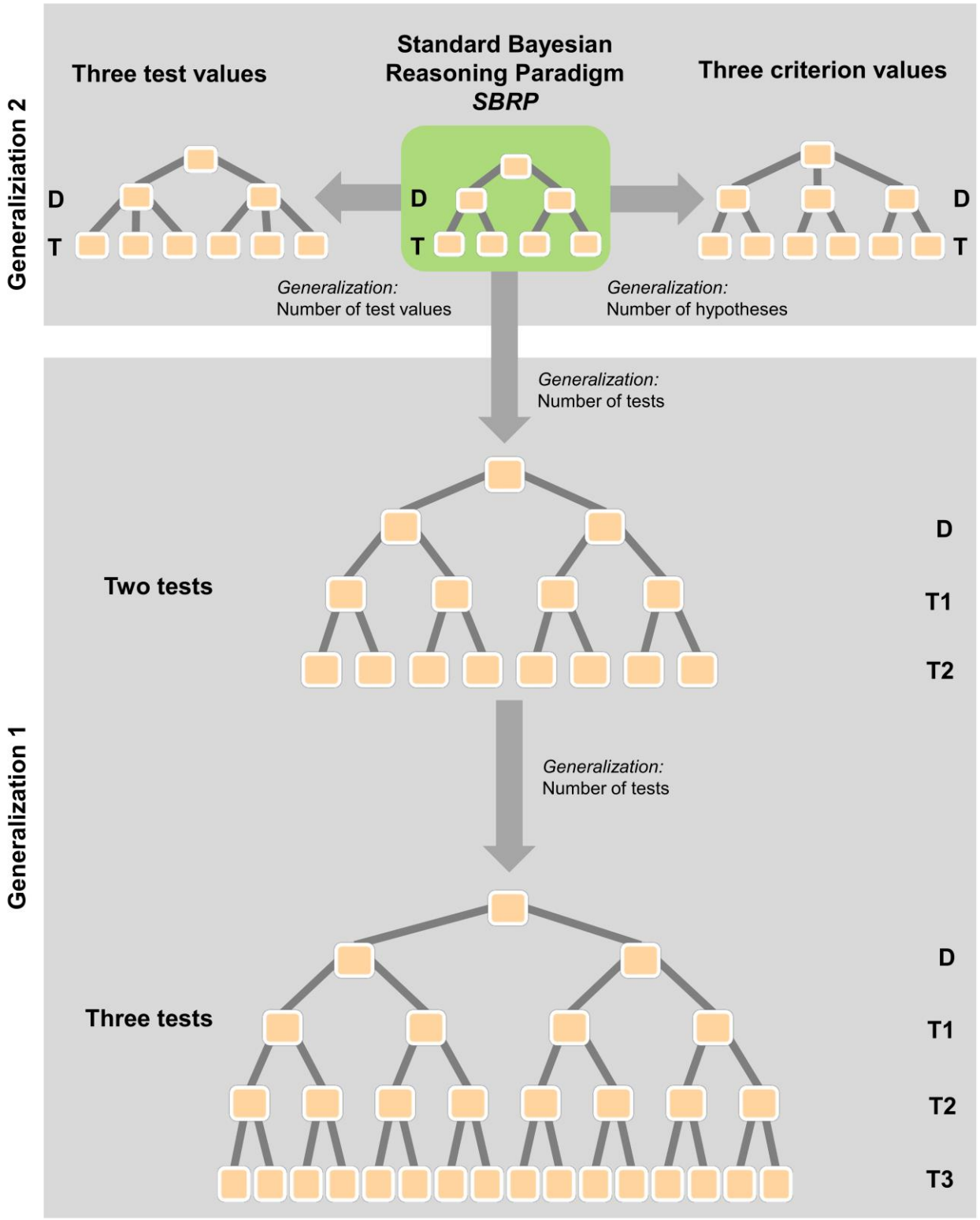


FIGURE 2. Generalizations of Bayesian standard tasks. Generalization 1: Middle column from top to bottom; Generalization 2: Top box from the middle to the sides; Generalization 3: Taking alternative nodes (e.g., concerning "not positive" test results) of a tree into account. D = diagnose, T_i = test result of test i.

Generalization 1: Extending the number of tests

The extension of the number of tests is obviously the most fundamental generalization. Since Krauss et al. (1999) and Hoffrage et al. (2015) only briefly referred to theoretical issues (such as conditional independency) or alternative ways to extend the SBRP to 2-test-scenarios (deviating from Figure 1), we will now discuss both issues in detail in order to outline possible future research. We close the section on generalization 1 with a short outlook addressing 3-test-cases.

In the following theoretical analyses, D denotes a certain disease and T_i+ means that test i has come out positive (with $i = 1,\ldots,n$ indicating the number of the tests). Note that the straightforward "generalization" of the question typically posed in SBRPs, namely the one for the *positive predictive value* P(D | T+), would be to ask for $P(D \mid \cap T_i+)$. This probability of being ill given that all tests applied come out positive we will refer to as the *"general" positive predictive value* (for the explicit distinction of the general positive predictive value and inferences regarding other possible test configurations, see generalization 3). The subsequent theoretical discussion, however, holds regardless of the specific question finally posed.

Generalization 1a: 2-test scenarios

In Figure 1, one possible way to extend the mammography SBRP to a 2-test scenario (sections highlighted in gray) is illustrated in probabilities (left side) and in natural frequencies (right side), and respective tree diagrams are presented. It is important to note that this constitutes just *one* possible description of a medical 2-test scenario. In this section, we will explicate five alternative ways to formulate and visualize the extension from 1-test to 2-test cases, holding the number of test values (positive or negative) and criterion values (having or not having the disease) constant. We did not implement all five alternatives experimentally, but we rather applied detailed theoretical analyses on these alternatives (based on natural frequency trees) in order to theoretically justify which of the five alternatives to choose for our experimental study.

In 2-test scenarios, in general, the question of the *conditional independence* of both applied test results generally comes into play. In the following, we assume conditional independence (see also the concept of "Naïve Bayes," Martignon, Vitouch, Takezawa, & Forster, 2003; Woike et al., 2017) but return to the issue later in detail.

Note that already in typical SBRP wording (e.g., see unhighlighted text in Figure 1), only statistical information for *one* special test outcome (namely M+, i.e., a positive mammogram) is provided, and therefore only two branches are textually addressed (namely $B \rightarrow M+$ and $\neg B \rightarrow M+$, with B and ¬B indicate the presence and absence of breast cancer, respectively). Usually the prevalence P(B) is given, but not P(¬B); the sensitivity P(M+ | B), but not the false negative rate P(M– | B); and the false-positive rate P(M+ | ¬B), but not the specificity P(M– | ¬B). However, because the probabilities of event and complementary event add up to 1, in SBRPs the missing pieces of information can be deduced without too much difficulty. In order to mirror (and therefore generalize) typical SBRP wording, the probability (and natural frequency)

wording of 2-test scenarios can also be formulated *along two corresponding branches*. In Figure 1, for example, the respective branches that are addressed by the depicted wording are B → M+ → S+ and ¬B → M+ → S+.

When only the (probability or frequency) wording is provided, the assumption of conditional independence in that case would allow one to derive the statistical information concerning all alternative branches and consequently to mathematically derive answers to all possible questions (for details, see section Conditional Independence). This analysis yields two implications: First, all possible wording that is formulated along two corresponding branches allows one to construct and also complete probability and natural frequency tree diagrams in the case of conditional independence and thus are mathematically equivalent to each other. Second, this work of completing is already done in tree diagrams (since each number on a branch or in a node has its concrete counterpart in the neighboring branch or node). Thus the four representations in Figure 1 (probability wording, natural frequency wording, probability tree, natural frequency tree) are mathematically equivalent in the sense that along alternative corresponding branches, a textual problem principally (implicitly) carry the identical statistical information (given conditional independence).

Thus we will specify "tree diagram" as "frequency tree" (Figure 3), but respective probability trees can easily be derived in any of the cases by simply depicting the proportion "upper node divided by lower node" at the corresponding branch. Figure 3 illustrates five basic possibilities for generalizing the SBRP of the breast cancer screening problem to a 2-test scenario (taking the sonogram as the second test). Tree A represents the method of extension that was also chosen for our empirical study (the original 2-test scenario that was implemented is displayed in Table 6). In addition, Figure 3 illustrates alternative extensions (B–E), all based on the same prevalence of breast cancer and the same sensitivities and false-positive rates of mammography (T_1) and the sonography (T_2). Regarding each of the tree diagrams A–E, we will discuss the following four salient features (see Table 1) that—we postulate—give insight into the situation: 1) Do all "nested sets" refer to *one* overarching sample? 2) Is the reality of *sequentially* applying medical tests acknowledged? 3) Are all single sensitivities and false-positive rates of both included tests "conserved" (i.e., represented by a simple ratio a/b with a and b being numbers of the frequency tree)? And 4) Can all questions regarding all four possible test configurations ($T_1+ \cap T_2+$, $T_1+ \cap T_2-$, $T_1- \cap T_2+$, $T_1- \cap T_2-$) be directly answered by forming a ratio of the form a / (a + b)?

In tree A (Figure 3) all statistical information refers to one overarching sample of women and represents sequential testing of this sample. Note that both features correspond to "natural sampling processes" (Fiedler & Juslin, 2006; Kleiter, 1994) and conform to the original definition of natural frequencies (Gigerenzer & Hoffrage, 1995; Hoffrage et al., 2002). Furthermore, in tree A, each single sensitivity and specificity of both tests is "conserved" (by a simple division). Finally, respective inferences regarding each possible test configuration (e.g., a negative sonogram after a positive mammogram) can be deduced by a ratio of the form a / (a + b). Thus tree A fulfills features 1–4 simultaneously (see Table 1).

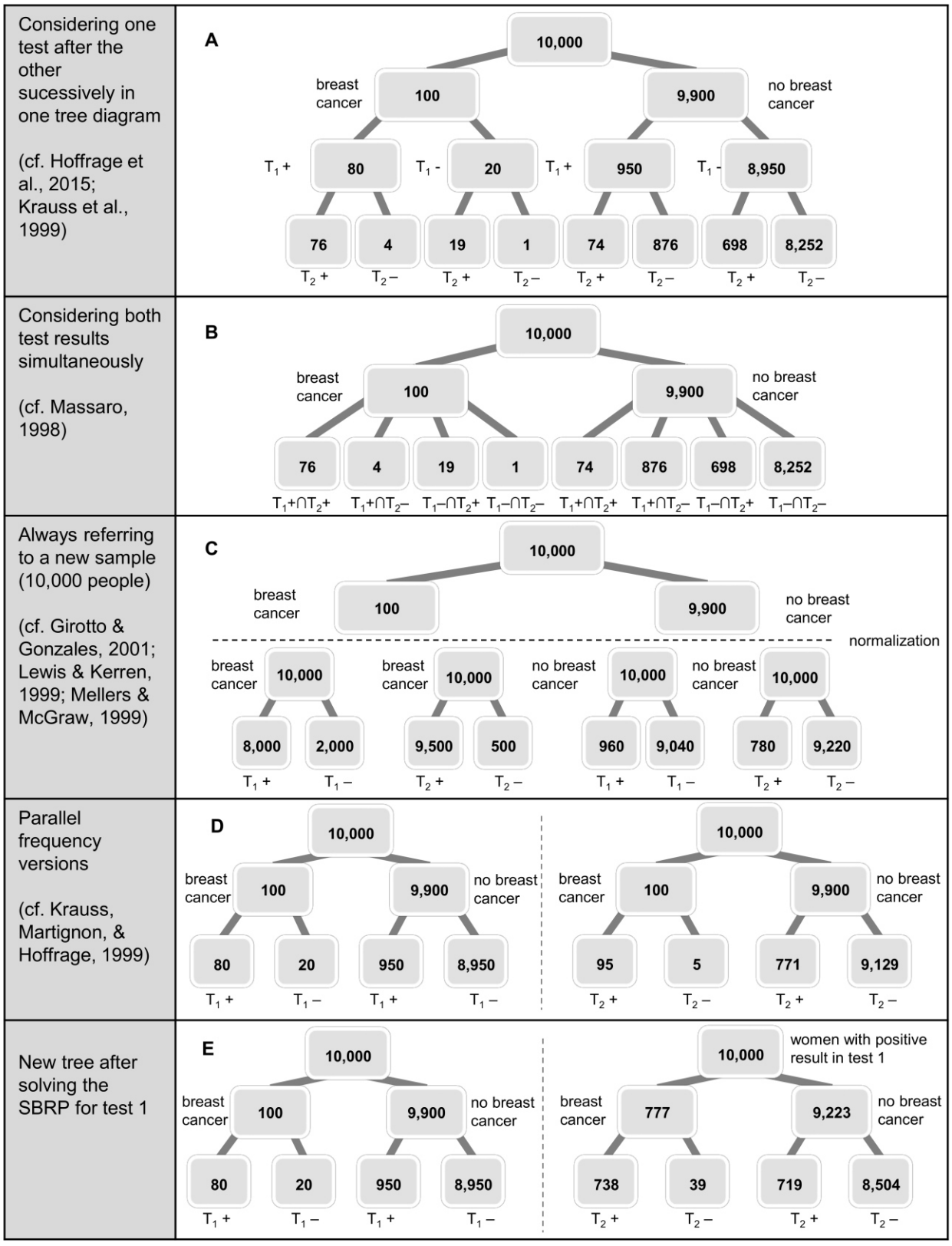


FIGURE 3. Five different options for extending the SBRP to the 2-test case.

TABLE 1. Four features of natural frequencies trees.

Type of tree	1 all nested sets refer to an overarching sample	2 tree(s) represent(s) sequential testing	3 single sensitivities and specificities are conserved by a ratio of type a / b *	4 questions regarding all four test configurations can be answered by a ratio of type a / (a + b)*
A	yes	yes	yes	yes
B	yes	no	no	yes
C	no	no	yes	no
D	no	no	yes	no
E	no	yes	yes	no

Note: * a, b are numbers in nodes of a frequency tree, while a / b and a / (a + b) correspond to the frequentistic expressions "a out of b" and "a out of a + b," respectively.

Another possibility for describing the 2-test scenario is simply to consider all four finally possible test configurations without taking the actual order of application of both tests into account (tree B). Massaro (1998), speaking of such a tree diagram, states that "a frequency algorithm will not work [for two (or more) tests]" (p. 178). However, we speculate that respective frequency versions might work better than Massaro assumes because all numbers still refer to one overarching sample (which we consider the most important feature), and a corresponding textual frequency wording can easily be imagined by simply replacing an expression like "gets a positive mammogram" (in the SBRP) with "gets two positive test results," etc. (in this case, the 2-test scenario even mimics the structure of a 1-test scenario; see also Prinz, Feufel, Gigerenzer, & Wegwarth, 2015). Furthermore, all possible questions concerning alternative configurations can be answered with the same algorithm—a / (a + b) (see Table 1). Yet in tree B, the single sensitivities and specificities are no longer conserved (by a / b), but must be reconstructed using slightly more demanding operations (which is possible in the case of conditional independence, for instance $P(T_1+ \mid B) = (76 + 4) / 100$). However, because in evidence-based medicine the combined sensitivities and specificities of multiple tests are usually not provided, and tree B does not reflect the reality of sequential testing, we did not choose this representation for our study.

The third diagram (tree C) refers to a frequency concept that repeatedly relates each given percentage to a new sample (e.g., 10,000 women), which still allows the formulation of a textual problem based on expressions like "100 out of 10,000." Regarding the SBRP, for example, Girotto and Gonzalez (2001), Lewis and Keren (1999), and Mellers and McGraw (1999) found that participants performed poorly with respective frequency wordings. Yet such formulations neither refer to one overarching sample nor represent sequential testing of patients and thus cannot grow out of from natural sampling (see Table 1). As a consequence, it is no longer possible to answer questions regarding all four test configurations by simply calculating a / (a + b), and

for that reason, we also did not choose such versions for our study. However, in diagram C, each single sensitivity and false-positive rate was still conserved by a / b.

The fourth possibility (tree D) is to "double" the natural frequency tree for the 1-test case. However, respective frequency wordings would represent information that was sampled by two different physicians, with one only applying test 1 and the other only applying test 2 (or, alternatively, by one physician, applying just one of both tests for each new patient). Although the single sensitivities and specificities are clearly conserved (by a / b), the given numbers neither stem from one overarching sample nor represent sequential testing (see Table 1). Consequently, tree diagram D would only allow one to answer questions regarding both of the "included" 1-test scenarios by forming a ratio of the form a / (a + b), while the relevant questions referring to a combination of both tests (regardless of the specific test configuration) require cognitively more demanding operations (e.g., first forming $95 / 100 \cdot 80 = 76$ and $771 / 9{,}900 \cdot 950 = 74$ before dividing 76 by (76 + 74)). Therefore, tree D only fulfills feature 3. Frequency wordings on 2-test scenarios formulated in the structure of tree D (but without the trees themselves being presented) were called "parallel frequency format" by Krauss et al. (1999), and it turned out that only 14.6% of participants could solve such versions.

The last diagram (tree E) also does not correspond to natural sampling but is constructed by first reducing the 2-test case to a 1-test case. Solving the 2-test case in this way means solving two 1-test cases in succession. Although tree diagram E represents sequential testing and perfectly preserves all single sensitivities and specificities, we did not test corresponding frequency wording, because we specifically consider the relatedness of all frequencies to one overarching sample as crucial to fostering understanding. Moreover, based upon tree diagram E (or related wordings), questions referring to configurations with regard to a negative mammogram cannot be answered by simply forming a / (a + b). Note that the situation would change if the second sample consisted of 1,030 women (instead of 10,000) because then feature 1 (an overarching sample) would be fulfilled but not feature 4 (because questions including a negative mammogram still cannot be answered by forming a / (a + b)).

In sum, although all diagrams (A–E) and even all conceivable respective textual problems would generally allow—given statistical independence—one to derive answers to all relevant questions (there are basically eight in a 2-test scenario), the cognitive effort of such diagnostic inferences substantially varies between diagrams A to E. The same is true with respect to the reconstruction of the sensitivities and specificities of both underlying tests. Another question to consider (that is not depicted in Table 1) would be whether both underlying SBRPs can also be answered by forming a ratio a / (a + b). Note that concerning tree A, this is only the case for the SBRP regarding the mammogram. To assess $P(B \mid T_2+)$ is slightly more complicated because it requires taking into account four nodes of the tree by

$$\frac{76 + 19}{(76 + 19) + (74 + 698)}.$$

Obviously, only statistical information structured according to tree D would allow one to answer questions regarding *both* of the SBRPs by simply building a / (a + b), but inferences regarding two test results can no longer be obtained in this way. Since, for example, in routine screening for breast cancer, the mammogram is usually fixed as the first test (Kooperationsgemeinschaft Mammographie GbR [cooperative association mammography], 2016), a physician does not have to diagnose breast cancer based solely on a sonogram. The (relevant) SBRP regarding the *first* test in the sequence (the mammogram) can be answered by trees A, D, and E.

It can be concluded that *only* tree A reflects the four important features (Table 1) that we believe to be beneficial for fostering insight into SBRPs as well as non-SBRPs. In addition, it will become clear from our following explication on the issue of conditional independence that diagram A is the only tree that allows for directly establishing conditional independence or dependence of test results by simply comparing two ratios of the form a / b.

Conditional independence

In 2-test cases, the question of conditional independence arises.[1] The two test results T_1+ and T_2+ are conditionally independent given the existence of disease (D) if and only if $P(T_1+ \cap T_2+ \mid D)$ equals $P(T_1+ \mid D) \cdot P(T_2+ \mid D)$. This is equivalent to $P(T_2+ \mid T_1+ \cap D) = P(T_2+ \mid D) = P(T_2+ \mid T_1- \cap D)$ and to $P(T_1+ \mid T_2+ \cap D) = P(T_1+ \mid D) = P(T_1+ \mid T_2- \cap D)$ (see also, e.g., Dawid, 1979; Jarecki, Meder, & Nelson, 2017; Pearl, 2008). This is to say, the conditional independence of two test results when the disease is present means that the probability of one test result (given the disease) is not affected by the other test result. Note that it is possible for test results T_1+ and T_2+ to be conditionally independent given the presence of a disease, while not being conditionally independent in the absence of a disease.

We will again illustrate this issue of the conditional independence by referring to tree diagrams, first beginning with probability trees (see, e.g., Figure 1). In the case of conditional dependence of test results, a probability tree (if it represents sequential testing) would carry conditional probabilities at the lowest branches that deviate from the original basic sensitivities and specificities of the sonogram. With respect to our empirical study, it is important to note that depicting the original (unconditioned) sensitivity of the sonogram at the lowest branches is only possible in the case of the conditional independence of both involved test results given B, that is, if the *conditioned* sensitivity of the sonogram $P(S+ \mid M+ \cap B)$ equals the (unconditioned) sensitivity of the sonogram $P(S+ \mid B)$.

In order to guarantee the validity of an inference based on a purely textual probability wording of a 2-test scenario (without a tree diagram), there are basically two options: Either the sensitivities and false-alarm rates of the second test must be verbally communicated as conditional on the first test results (e.g., "The probability that a woman with breast cancer and a positive mammogram will get a positive sonogram is 95%," see Figure 1). Or, if only the basic sensitivities and false-alarm rates of the medical tests are available, conditional independence must in addition be stated explicitly (see, e.g., footnote 1 in Table 6).

With respect to natural frequency trees (or natural frequency wording in general), it is important to note that according to the concept of "natural sampling," the resulting frequencies at the lowest level (e.g., concerning tree A) would in *any case* allow one to answer questions concerning all test configurations *regardless of potential conditional dependencies*. This is the case because, given that both tests were indeed conditionally dependent (given B or not B), the physician would nevertheless *automatically* sample the correct natural frequencies because the mutual dependency would be reflected by them at the lowest level. Thus based on his/her experience, the physician would come to correct judgments without ever having heard about the concept of conditional independence (but see the problem of "profile memorization" with an increasing number of tests, e.g., Woike et al., 2017).

Let us now consider the remaining tree diagrams B–E. Again, although tree B does not represent sequential testing (Table 1), all possible questions can be answered regardless of potential conditional dependencies. In contrast, tree diagrams C and D (that consist of "unconnected" tree diagrams) do not allow one to answer questions based on a combination of both test results without a statement on conditional independence, since both trees do not contain branches where conditioned sensitivities and specificities might be located. Based on diagram E, only questions including a positive first test result can be answered without a statement on conditional independence, whereas answering, for example, $P(B \mid T1- \cap T2+)$ would require such a statement.

Interestingly, the only tree diagram that allows one to establish conditional independence *at a glance* would be the probability tree in Figure 1 (bottom left) because one might simply check the probabilities $P(S+ \mid M+ \cap B)$ and $P(S+ \mid M- \cap B)$ depicted at the respective branches to establish whether they are equal (given equality, both test results are conditionally independent given B). Note that it is slightly more difficult to establish conditional independence in the corresponding natural frequency tree (tree A or Figure 1, bottom right) because "76 out of 80" has to be compared with "19 out of 20." In tree B, conditional independence can only be derived by using a more complex algorithm, and thus tree diagrams C–E do not allow one to establish conditional independence at all. Although we had to restrict our study to one of extensions A–E (because of several further factors of interest), it is our hope that Figure 3 and Table 1 may outline fields for future theoretical and empirical research.

Before we turn to medical reality, it should be noted that a commonly applied simplification of the general problem used by Bayesian statistics and Bayesian modelers is to assume that evidence (e.g., tests, or cues in general) is conditionally independent given a specific criterion (called "Naïve Bayes," see Jarecki et al., 2017; Woike et al., 2017; Martignon et al., 2003). In medical reality, however, it *cannot* be automatically assumed that two successive medical tests are conditionally independent for patients with (or without) the disease (e.g., Fryback, 1978). For instance, a false-positive HIV test result might occur because of a chronic hepatitis B infection (Lee, Park, & Kang, 2013). Imagine that a person who is not HIV infected but instead infected with hepatitis B receives a (false) positive test result from an HIV test because of the hepatitis B. If now a second HIV test is conducted, the probability that this second test will also be positive is

higher than if there had been no first positive test result. In such cases, where both tests applied are the same (for example a second ELISA test after a positive ELISA test result) or at least based on similar medical procedures, conditional dependence should rather be assumed. However, the less similar two medical tests are, the more likely that the two test results are not additionally affected by the same background variable (e.g., an hepatitis B infection; see, e.g., Shen, Wu, & Zelen, 2001). Astoundingly, the conditional independence of two medical tests seems not to be an important issue in the field of medicine. Based on intensive research of the literature and consultation with several experts from the university hospital, we believe that this issue may be either underestimated or ignored because of the complicated underlying statistical structure. Consequently, in the literature on evidence-based medicine (McGee, 2012), the overwhelming majority of sensitivities and specificities are provided for single tests only, and there is little information about combined sensitivities or specificities available. This has to be taken into account when trying to build ecologically valid experimental environments for medical experts (see Materials).

Generalization 1b: 3-test scenarios

Of course, the idea of generalization 1a can be extended and, theoretically, a third test added. For instance, if the mammogram and the sonogram are both positive, the next step might be a biopsy (Berg et al., 2012). Figure 2 (below) illustrates a corresponding tree that extends the idea of diagram A (Figure 3) further. Note that all arguments concerning Figure 3 (A–E) principally also hold for the 3-test case, and the theoretical deliberations above can be transferred to situations with three or even more medical tests. Regarding 3-test scenarios, the lowest level of tree A (or B or C, respectively, Figure 3) would contain $2^4 = 16$ nodes (instead of $2^3 = 8$ in the 2-test case), and the lowest level of diagrams D and E would contain $3 \cdot 4 = 12$ nodes.

Obviously, the number of test configurations increases exponentially with the number of medical tests involved, and it becomes cognitively more problematic to store and memorize all of these "profiles" (Martignon & Schmitt, 1999; Kipling, 1999; Woike et al., 2017). Here is where alternative theoretical approaches on probabilistic decision-making come into play, concerning both humans (Martignon et al., 2003; Gigerenzer & Goldstein, 1999; Mousavi & Gigerenzer, 2014) and machines (Martignon & Schmitt, 1999; Lewis, 1998). The full-frequency tree diagram of the 3-test scenario (Figure 6), however, perfectly illustrates what happens when three medical tests with different sensitivities and false-positive rates are applied because it displays—regardless of whether a physician is actually able to memorize these frequencies in an authentic situation—all possible configurations in one diagram (for alternative probabilistic approaches, see General discussion).

Empirical studies concerning generalization 1 (a and b)

Natural frequency and probability versions according to 2- test and 3-test scenarios have already been tested empirically—although, without tree diagrams presented or alternative questions posed—by Hoffrage et al. (2015), who found that natural frequencies could help people in their decision-making processes in both scenarios (see also Krauss et al., 1999). In Binder et al. (2018),

tree A of Figure 3 was varied (by either the pruning or highlighting of branches), but only in a 2-test scenario and without considering alternative questions. All studies that have examined scenarios that follow generalizations 1 and 2 are summarized in the General discussion (Table 10).

Generalization 2: Extending the number of predictor or criterion values

A further means of generalizing the SBRP is represented in Figure 2 in the upper block. While on the left the number of predictor values is extended (e.g., three medical test values including an unclear test result), on the right the number of criterion values is extended (e.g., two different diseases vs. being healthy). In the following we address both generalizations, one after the other.

Generalization 2a: Three predictor values

It is imaginable that in a breast cancer screening, for example, test results might be unclear (e.g., Partik et al., 2001). The trees on the left side of Figure 4 can serve as a model for such situations (unclear test results in addition to positive and negative ones). However, it might also describe a scenario in which two physicians have different opinions about a mammogram (i.e., one considers it to be positive and the other negative; see, e.g., Elmore, Wells, Lee, Howard, & Feinstein, 1994). In medicine the latter aspect is broadly discussed under the concept "second opinion" (Schrader, Zengerling, Hakenberg, & Protzel, 2016). Note that with respect to unclear test results, there is no specific term such as sensitivity or specificity to describe P(test unclear | disease) appropriately.

Generalization 2b: Three criterion values

Even more categories are possible, not only with respect to the number of potential test outcomes but also with respect to the number of potential hypotheses. For example, a medical test may be indicative of more than one disease (see trees on the right side of Figure 4). There are tests for diagnosing type 1 and type 2 diabetes, for instance, that cannot distinguish between the two (e.g., Atkinson, Eisenbarth, & Michels, 2014).

Empirical studies concerning generalization 2 (a and b)

Hoffrage et al. (2015) found that in pure textual problems, participants could profit more from natural frequencies with respect to generalization 2a than with respect to generalization 2b, but did not discuss this empirical difference. In the present study, we intend to a) replicate this difference but furthermore to explain it theoretically, b) implement probability and frequency trees as visual aids in such situations, and c) pose additional alternative questions in addition.

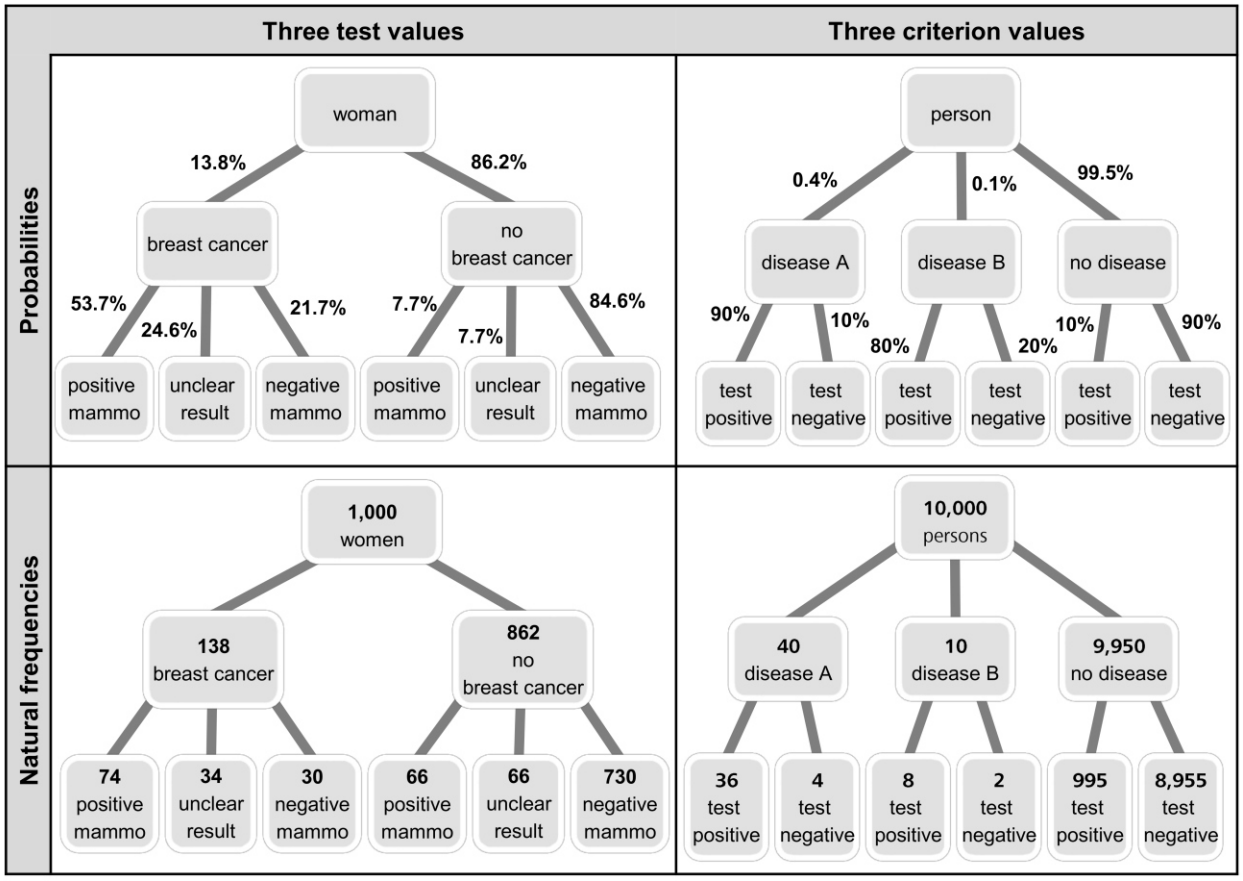


FIGURE 4. On the left: tree diagrams (with probabilities or natural frequencies) for a scenario with three possible test values (positive, unclear, and negative). On the right: tree diagrams (with probabilities or natural frequencies) for a scenario with three possible hypotheses (disease A, disease B, or no disease). These four visualizations were also applied in our empirical approach (see Materials).

It has to be noted that a special case of the structure of Figure 4 (right trees) was empirically examined in several studies by Brase (2014), Johnson and Tubau (2013), Moro, Bodanza, and Freidin (2011), and Yamagishi (2003). For instance, Brase (2014) examined a problem where three different weather conditions (rainy, sunny, and mixed) were forecast by barometric pressure (high vs. low). When it was sunny, the barometric pressure was always high, and when it was rainy, the barometric pressure was always low. A mixed day was associated with low pressure only half of the time. Thus only one of the three hypotheses (mixed day) is associated with uncertainty, whereas the other two hypotheses (i.e., rainy or sunny day) always lead to a definitive outcome (i.e., high or low barometric pressure). Johnson and Tubau (2013) found that natural frequencies were helpful for participants in this situations. Brase (2014) analyzed in similar types of scenarios the effect of representing roulette-wheel diagrams, realistic pictographs, or abstract icon arrays in addition to textual information and found a positive effect with all three visualizations over and above natural frequencies. Moro, Bodanza, and Freidin (2011) examined a similar scenario in the context of a gemstone problem, also finding a frequency facilitating effect and an improvement with the use of an unusual visualization (structured

brackets without numbers). Similarly, Yamagishi (2003) examined visualizations of the gemstone problem (which even included tree diagrams), finding that (probability and frequency) tree diagrams support decision-making processes in these specific type of cases. It must be noted that the famous Monty Hall problem (which is discussed by, e.g., Baratgin, 2015 and Krauss & Wang, 2003) is slightly more complicated and has the described structure only at first glance. None of these problems is mathematically or cognitively equivalent to the medical scenario displayed in Figure 4, but all do constitute special cases containing three criterion values (e.g., with some probabilities fixed to 0% or 100%).

Generalization 3: Extending the number of questions

In the SBRP only two test configurations, T+ and T–, and therefore only questions concerning a positive or a negative test result, can be asked, that is, the typical positive predictive value P(D | T+), but also P(¬D | T+), P(D | T–) as well as P(¬D | T–). While P(¬D | T+) can easily be derived by 1 – P(D | T+), the inferences regarding negative test results require one to first consider the false negative rate P(T– | D), which is *1 – sensitivity*, and the specificity P(T– | ¬D), which is *1 – false-positive rate*. In research on the SBRP, inferences based on negative test results are very seldom assessed empirically, but such an assessment should not be much more cognitively demanding than a positive predictive value.

The more tests applied, or the more test or criterion values possible, however, the more different test configurations that can occur. In the case of generalization 1, the number of potential inferences increases exponentially (with n binary tests, 2^n test configurations can be realized). Feature 4 (Table 1) acknowledges the option for four different test configurations in 2-test scenarios yielding eight possible questions to be stated (including events and complementary events), and in cases where three tests are conducted, 16 probabilities can become relevant, and so on. Note that each possible inference—regardless of the number of binary tests—is based on exactly two corresponding branches of a respective tree diagram (e.g., Figure 2 below). For example, for assessing $P(D \mid T_1+ \cap T_2- \cap T_3+)$ or the probability of the complementary event $P(\neg D \mid T_1+ \cap T_2- \cap T_3+)$, respectively, only the branches $D \rightarrow T_1+ \rightarrow T_2- \rightarrow T_3+$ and $\neg D \rightarrow T_1+ \rightarrow T_2- \rightarrow T_3+$ have to be considered. Because in evidence-based medicine, and in typical SBRP wording as well, information on T+ is usually provided, namely the sensitivity P(T+ | D) and the false-positive rate P(T+ | ¬D), we will retain and generalize this information structure in our empirical approach (see Materials). Alternatively, the information presented could be "adjusted" each time to the question finally posted (e.g., providing $P(T_1+ \mid D)$, $P(T_2- \mid D)$, $P(T_3+ \mid D)$). This, however, would reduce the ecological validity of the medical scenario. Moreover, adjusting the information to an alternative question both mathematically and cognitively would reduce it to a basic case, which is not exactly the SBRP but a kind of twin of the SBRP.

With regard to generalization 2 (Figure 4), again, not simply the positive predictive value P(D | T+) can be addressed. Given that three test results are possible (positive, unclear, negative), P(D | T unclear) and P(D | T–) might also be relevant (or the probabilities of their three respective

complementary events). Notably, with three criterion values (disease A, disease B, healthy), the positive predictive value is not clearly defined. For instance, one might consider P(disease A | T+) and P(disease B | T+) two different positive predictive values, but of course the probability of being ill at all, that is, P(disease A ∪ disease B | T+), might also be designated positive predictive value. However, the respective inferences based on a negative test result here might be of interest as well (again including the probability of the complementary events).

In sum, in all of these generalizations (1a, 1b, 2a, and 2b), not only one can assess a positive predictive value, but one can also pose different questions based on alternative test configurations (or regarding different diseases). Consequently, by empirically addressing generalization 3, one enters new research fields in the following ways: a) by posing alternative questions based on typically available sensitivities and false-alarm rates, b) by examining the effect of natural frequencies, and c) by examining the effect of visualizations (tree diagrams), the latter two especially in situations where the information structure does not perfectly match the question posed.

Our empirical approach

For examining non-SBRPs, our rationale was to implement the following guidelines:

1) Extending the positive predictive value P(D | T+) to P(D | ∩Ti+) in generalization 1.

2) Generalizing the typically available information by also presenting sensitivities and false-positive rates of further tests (according to evidence-based medicine) in generalization 1.

3) Presenting sensitivities and false-positive rates of tests *unconditioned* to each other in probability versions (according to evidence-based medicine), also adding a comment on conditional independence in order to guarantee the validity of the Bayesian answer.

4) Presenting natural frequencies according to the paradigm of natural sampling, again adding a comment on conditional independence for comparability.

Research questions and hypotheses

Research questions

The research questions of our study are summarized in Table 2. For each of the three generalizations explicated above, we examine the effect of natural frequencies (as compared to probabilities) and the effect of presenting tree diagrams (as compared to pure textual presentation of information). In the following we will address generalizations 1–3, one after the other, and derive concrete hypotheses from acknowledging the complexity of the cognitive operations required.

TABLE 2. Research questions.

	Generalization 1 Extending the number of tests		**Generalization 2** Extending the number of test or criterion values		**Generalization 3** Extending the number of questions in generalization			
	a) 2tests	b) 3 tests	a) 3 test values	b) 3 criterion values	1 a)	1 b)	2 a)	2 b)
Research question I	Are natural frequencies effective?		Are natural frequencies effective?		Are natural frequencies effective (based on the typical information structure)?			
Research question II	Are (additional) tree diagrams effective?		Are (additional) tree diagrams effective?		Are (additional) tree diagrams effective?			

Hypotheses regarding generalization 1: Extending the number of tests

Table 3 illustrates the required cognitive operations for computing the generalized positive predictive value $P(D \mid \cap T_i+)$ concerning statistical information given in natural frequency or probability wording. Our implemented 2-test and 3-test scenarios are depicted in Table 6 and their respective tree diagrams in Figures 5 and 6. According to Hoffrage et al. (2015) and McDowell and Jacobs (2017), we expect natural frequencies to foster insight into the 2- (1a) and 3-test scenarios (1b) as well. However, even though the algorithm for solving a natural frequency task is identical for 1-test, 2-test, and 3-test cases (a / (a + b), see Table 3), we assume a slight decline in performance with increasing numbers of tests because the wording becomes more complicated. Following the results of Binder et al. (2018) on 2-test scenarios, we assume frequency tree diagrams to also be helpful in 3-test scenarios, since such diagrams make the required numbers a and b more stand out.

In probability wording each additional test involves two new pieces of information, namely an additional sensitivity and an additional false-positive rate (see Table 3). Note that in the probability format, *all* pieces of information must indeed be cognitively integrated in a solution algorithm. Therefore, the performance on probability wording should also decline with increasing numbers of tests. However, we assume probability trees to be (moderately) helpful because displaying the full situation might also foster insight into probability versions.

TABLE 3. Pieces of information to integrate in order to assess the "generalized positive predictive value" P(D | ∩T_i+) in generalization 1.

	Natural frequencies	Probabilities
1a: 2-test case	"a out of (a + b)" ≙ $\frac{a}{a+b}$	$\frac{p \cdot s_1 \cdot s_2}{p \cdot s_1 \cdot s_2 + (1-p) \cdot f_1 \cdot f_2}$
1b: 3-test case	"a out of (a + b)" ≙ $\frac{a}{a+b}$	$\frac{p \cdot s_1 \cdot s_2 \cdot s_3}{p \cdot s_1 \cdot s_2 \cdot s_3 + (1-p) \cdot f_1 \cdot f_2 \cdot f_3}$

Note: a and b are absolute numbers (given in frequency versions). p: prevalence; s_i: sensitivity of test i; f_i: false-alarm rate of test i (given in probability versions)

Hypotheses regarding generalization 2: Extending the number of predictor or criterion values

In a scenario with three test values (2a), the solution algorithm for the positive predictive value P(D | T+) in both information formats is identical to an SBRP (Table 4). Therefore we assume effects of natural frequencies and frequency trees to be similar to those of in typical SBRPs. For convenience, in scenario 2b with three criterion values (disease A, disease B, healthy), we denote P(disease A | T+) as the positive predictive value and further assume that the single sensitivities and false-alarm rates with respect to *both* diseases are given. The solution algorithm for the positive predictive value P(disease A | T+), however, differs from that in all situations discussed so far (SBRP, generalizations 1a, 1b, 2a) since here the denominator consists of a sum of three numbers (Table 4), for which reason we expect a slightly worse performance in 2b than in 2a in general.

Natural frequencies and tree diagrams should be helpful in both 2a and 2b. However, tree diagrams may only be of limited additional help here because no sequential subsetting needs to be displayed. In the following, we will call this (alternative) hypothesis on the beneficial effect of tree diagrams concerning their degree of iterative nestedness "iteration-hypothesis."

TABLE 4. Cognitive operations required for the "positive predictive value" in generalization 2.

	Natural frequencies	Probabilities
2a: 3 test values	"a out of (a + b)" $\triangleq \frac{a}{a+b}$	$\frac{p \cdot s}{p \cdot s + (1-p) \cdot f}$
2b: 3 hypotheses	"a out of (a + b + c)" $\triangleq \frac{a}{a+b+c}$	$\frac{p_A \cdot s_A}{p_A \cdot s_A + p_B \cdot s_B + (1-p_A-p_B) \cdot f}$

Note: a, b, and c are absolute numbers (provided in natural frequency versions)
p, p_A, p_B: prevalence of diseases; s, s_A, s_B: sensitivities; f: false-positive rate (provided in probability versions)

Hypotheses regarding generalization 3: Extending the number of questions

With respect to generalizations 1 and 2, many varying inferences are possible. The alternative questions actually provided to our participants are displayed in Tables 6 and 7 (see questions 2 and 3 below).

Let us first consider alternative questions in 2- and 3-test scenarios: Note that even if conditional independence is guaranteed, much cognitive work has to be done (when no tree diagram is given in addition to text) in order to transfer the statistical information to other test configurations, for example, deriving probabilities (or natural frequencies) of complementary events up to the lowest level of the two branches in question.

We hypothesize that an inference based on a textual problem for a multiple-test scenario should become more difficult to achieve the more the final question deviates from the information provided (based on positive test results), and that a natural frequency format does not help in this type of situation per se. However, we expect tree diagrams to have a large effect with regard to alternative questions (especially natural frequency trees). This is because tree diagrams always present the full picture, and therefore no new information has to be derived and all inferences can immediately be built by a / (a + b), with a and b being elements of the tree diagram but not necessarily of the wording. Thus the "deviation" of the question posed from the given information no longer matters.

Interestingly, the situation is different in generalization 2: Here the textual descriptions of scenarios with three predictor or criterion values *must* of necessity already display all relevant information (with the exception of simple complementary events) because otherwise P(D | T unclear) or $P(D_A \mid T+)$ cannot be derived at all. Thus all arbitrary questions posed in these situations are symmetrical in a mathematical sense and therefore should be comparably easy. Consequently, only natural frequencies should be helpful, and—because no iterative subsetting is needed—the additional presentation of a tree diagram might only have limited advantage here.

Method

Materials

A paper-and-pencil-questionnaire contained four successive nonstandard Bayesian scenarios, each addressing one of generalizations 1a, 1b, 2a, and 2b. In all four scenarios, aside from the general positive predictive value, further inferences based on alternative test configurations were added according to generalization 3. Table 5 displays the implemented 4×2×2 design with factor 1 being *generalization* (1a: 2-test scenario; 1b: 3-test scenario; 2a: three test values; 2b: three criterion values), factor 2: *information format* (probabilities vs. natural frequencies), and factor 3: *visualization* (no tree vs. tree).

According to Table 5, four versions of each scenario were implemented. In each problem version, after a short description of the medical situation, the statistical information (e.g., prevalence, sensitivities, and false-alarm rates, etc.) is presented, eventually followed by a tree diagram in versions with a visualization. Finally, questions—also on the alternative test configurations—are formulated in the same format as the statistical information presented earlier. The wording of the original problems is depicted in Table 6 (generalization 1) and Table 7 (generalization 2), each with the alternative questions posed included below (generalization 3). Since there has already been some research regarding the 2-test scenario (Hoffrage et al., 2015; Binder et al., 2018), we declined to ask a third question because of the time constraints involved in a one-hour session. The presented tree diagrams (in the visualization conditions) are depicted in Figures 4, 5, and 6.

Each participant received each of the four medical scenarios (2-test scenario, 3-test scenario, three test values, three criterion values) only once (factor 1). Of these four problems, two were formulated in probabilities and the other two in natural frequencies (factor 2). One of the probability versions and one of the natural frequency versions were presented without a corresponding tree, and the remaining two problems were equipped with an additional tree diagram (factor 3) so that each participant saw exactly one probability and one frequency tree. The order of type of generalization, information format, and visualization was varied systematically to avoid effects of presentation order.

TABLE 5. Overview of the 16 resulting problem versions implemented.

Factor 1: Generalization		Situation (context)	**Factor 2:** Information format	**Factor 3:** Visualization	"Pos. pred. value" Generalization 3 (alternative questions)		
					Question 1	Question 2	Question 3
1a	2-test scenario	Diagnosis of breast cancer B by mammogram M (pos./neg.) and sonogram S (pos./neg.)	prob.	no tree	$P(B \mid M+ \cap S+)$	$P(B \mid M+ \cap S-)$	--
				prob. tree			
			nat. freq.	no tree	$P(B \mid M+ \cap S+)$ *	$P(B \mid M+ \cap S-)$ *	--
				nat. freq. tree			
1b	3-test scenario	Diagnosis of unspecified complaint D by three unspecified tests, T_1, T_2, T_3 (each pos./neg.)	prob.	no tree	$P(D \mid T_1+ \cap T_2+ \cap T_3+)$	$P(D \mid T_1+ \cap T_2- \cap T_3+)$	$P(D \mid T_1- \cap T_2+ \cap T_3-)$
				prob. tree			
			nat. freq.	no tree	$P(D \mid T_1+ \cap T_2+ \cap T_3+)$ *	$P(D \mid T_1+ \cap T_2- \cap T_3+)$ *	$P(D \mid T_1- \cap T_2+ \cap T_3-)$ *
				nat. freq. tree			
2a	1-test scenario with 3 test values	Diagnosis of breast cancer B by mammogram (positive, unclear, negative)	prob.	no tree	$P(B \mid M+)$	$P(B \mid M \text{ unclear})$	$P(B \mid M-)$
				prob. tree			
			nat. freq.	no tree	$P(B \mid M+)$ *	$P(B \mid M \text{ unclear})$ *	$P(B \mid M-)$ *
				nat. freq. tree			
2b	1-test scenario with 3 criterion values	Diagnosis of disease A, disease B vs. healthy by unspecified test (pos./neg.)	prob.	no tree	$P(A \mid T+)$	$P(B \mid T+)$	$P(A \cup B \mid T+)$
				prob. tree			
			nat. freq.	no tree	$P(A \mid T+)$ *	$P(B \mid T+)$ *	$P(A \cup B \mid T+)$ *
				nat. freq. tree			

Note: * Questions formulated in natural frequencies

TABLE 6. Problem wording for the two situations with respect to generalization 1 (including additional questions concerning generalization 3).

	Generalization 1a: 2-test scenario		Generalization 1b: 3-test scenario	
	Probability version	**Natural frequency version**	**Probability version**	**Natural frequency version**
Medical situation	Imagine that you are a physician in a mammography screening center where women without symptoms are screened for breast cancer. In addition to mammograms, you frequently use a sonogram as a supplementary medical test to detect breast cancer.[1] […] The following information is available:		Imagine a serious illness that you usually diagnose with the help of three different medical tests (test 1, test 2, and test 3).[1] The following information is available:	
Statistical information	The probability of breast cancer for a woman without any symptoms is 1%. The probability that a woman with breast cancer will get a positive mammogram is 80%. The probability that a woman with breast cancer will get a positive sonogram is 95%. The probability that a woman without breast cancer will wrongly get a positive mammogram is 9.6%. The probability that a woman without breast cancer will wrongly get a positive sonogram is 7.8%.	100 out of 10,000 women without any symptoms have breast cancer. 80 out of 100 women with breast cancer get a positive mammogram. 76 out of 80 women with breast cancer and a positive mammogram get a positive sonogram. 950 out of 9,900 women without breast cancer wrongly get a positive mammogram. 74 out of 950 without breast cancer and with a positive mammogram wrongly get a positive sonogram.	The probability of having a special complaint is 2%. If a person has the complaint, the probability that she or he will receive a positive test result in test 1 is 80%. If a person has the complaint, the probability that she or he will receive a positive test result in test 2 is 95%. If a person has the complaint, the probability that she or he will receive a positive test result in test 3 is 75%. If a person does not have the complaint, the probability that she or he will nevertheless receive a positive test result in test 1 is 25%. If a person does not have the complaint, the probability that she or he will nevertheless receive a positive test result in test 2 is 10%. If a person does not have the complaint, the probability that she or he will nevertheless receive a positive test result in test 3 is 20%.	200 out of 10,000 persons have a special complaint. 160 out of every 200 persons who have the complaint receive a positive result in test 1. 152 out of 160 persons who have the complaint and have received a positive result in test 1 will also receive a positive result in test 2. 114 out of 152 persons who have the complaint and have received positive results in tests 1 and 2 will also receive a positive result in test 3. 2,450 out of 9,800 persons who do not have the complaint nevertheless receive a positive result in test 1. 245 out of 2,450 persons who do not have the complaint and have received a positive result in test 1 will nevertheless also receive a positive result in test 2. 49 out of 245 persons who do not have the complaint and have received positive results in test 1 and 2 will nevertheless also receive a positive result in test 3.
	[1] Footnote: Assume for your calculations that the results of both tests are (statistically) independent for women with breast cancer as well as for women without breast cancer.	[1] Footnote: Assume for your calculations that the results of both tests are (statistically) independent for women with breast cancer as well as for women without breast cancer.	[1] Footnote: Assume for your calculations that the results of all three tests are (statistically) independent for patients with the disease as well as for patients without the disease.	[1] Footnote: Assume for your calculations that the results of all three tests are (statistically) independent for patients with the disease as well as for patients without the disease.
Visualization	• no tree diagram, or • tree diagram (prob.)	• no tree diagram, or • tree diagram (nat. freq.)	• no tree diagram, or • tree diagram (prob.)	• no tree diagram, or • tree diagram (nat. freq.)
Question 1	What is the probability that a woman with a positive mammogram and a positive sonogram actually has breast cancer?	How many of the women with a positive mammogram and a positive sonogram actually have breast cancer?	What is the probability that a person actually has the complaint if all three tests are positive?	How many of the persons with the three positive test results actually have the complaint?
Question 2	What is the probability that a woman with a positive mammogram and a negative sonogram actually has breast cancer?	How many of the women with a positive mammogram and a negative sonogram actually have breast cancer?	What is the probability that a person actually has the complaint if test 2 is positive but tests 1 and 3 are negative?	How many of the persons with a positive test 2 but with negative tests 1 and 3 actually have the complaint?
Question 3	---	---	What is the probability that a person actually has the complaint if test 2 is negative but tests 1 and 3 are positive?	How many of the persons with a negative test 2 but with positive tests 1 and 3 actually have the complaint?

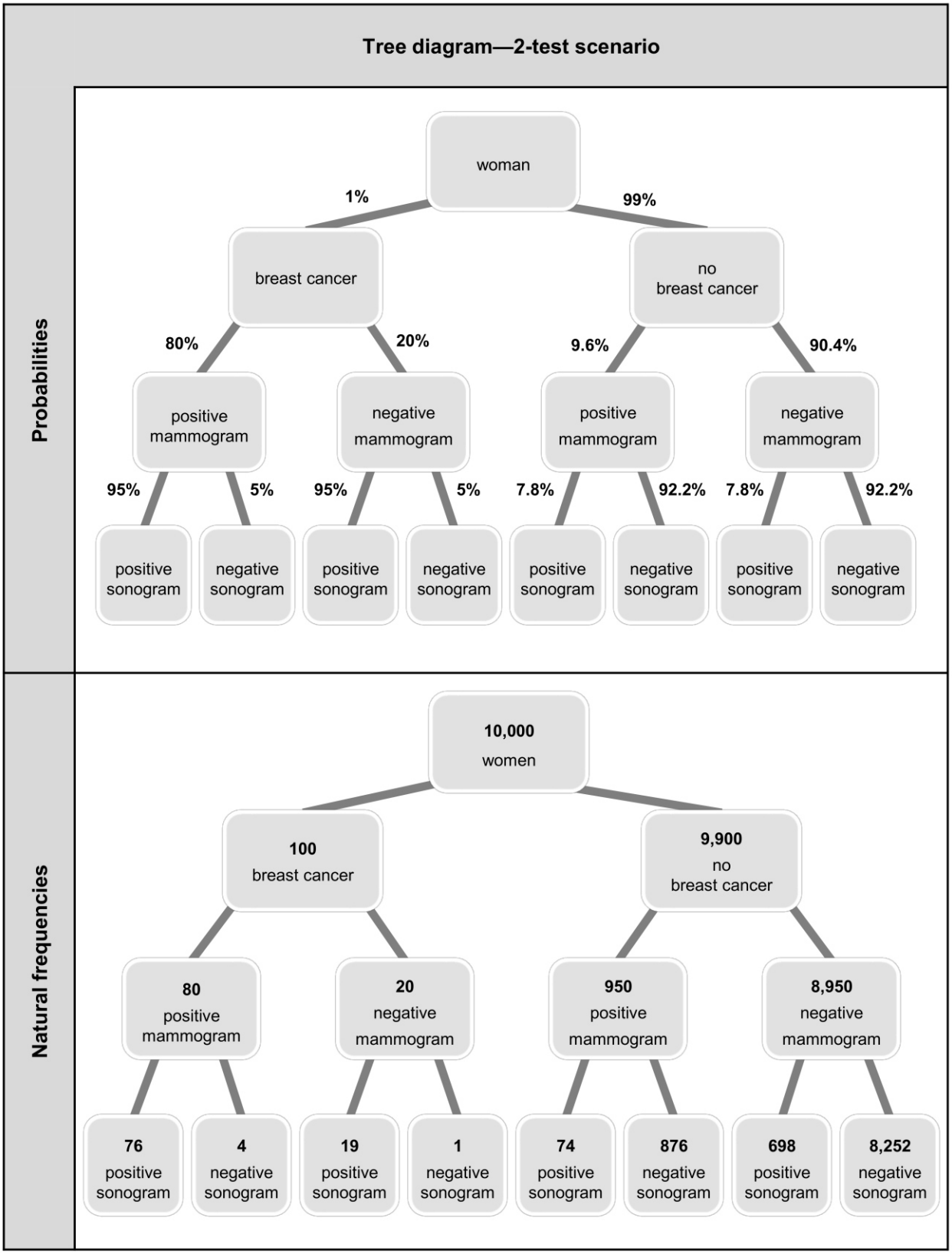


FIGURE 5. Probability tree and natural frequency tree for the implemented 2-test scenario (generalization 1a).

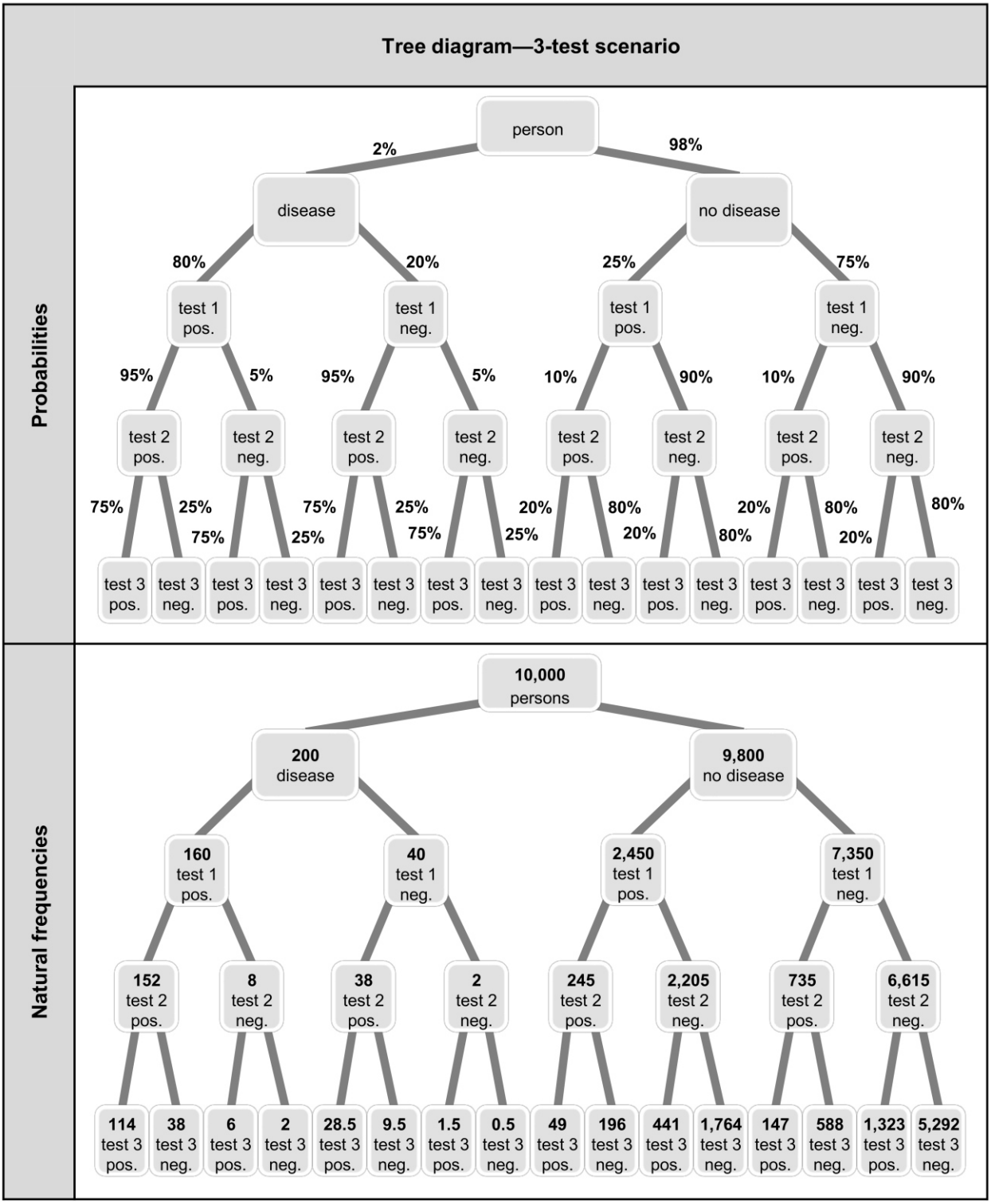


FIGURE 6. Probability tree and natural frequency tree for the implemented 3-test scenario (generalization 1b).

Remarks on footnotes in Table 6 about conditional independence

As mentioned above, in evidence-based medicine, sensitivities and specificities are nearly always provided unconditioned to each other, and information on conditioned sensitivities or false-positive rates is extremely unusual. In order to formulate wordings with respect to generalization 1 that are as ecologically valid as possible, we followed this tradition in our probability wording. However, to enable the mathematical solution (see, e.g., Table 8), we added a footnote confirming the conditional independence of the tests involved for women with and without the disease. Note that, in order to make sure that the paradigm of natural sampling and the sequential nature of medical test procedures is appropriately accounted for, natural frequency versions *automatically* present the sensitivity and false-alarm rate of the sonogram conditioned on the mammogram and thus must deviate from the (ecologically valid) probability wording.

This was also why our first pilot versions in probability format were formulated—for the sake of parallelism—in terms of (unusual) conditional sensitivities (e.g., "The probability that a woman with breast cancer will get a positive sonogram, *given that she has already had a positive mammogram*, is 95%," see Figure 1) and false-alarm rates. Yet, interestingly, implementing such formulations in a pilot study led to cognitive confusions (indeed, none of the participants of in the pilot study could solve such a task). For instance, a participant asked: "Why does the sonogram depend on a mammogram" (which in fact it does not—given B—in our 2-test scenario). Thus on the one hand, the addition of "given that she has already had a positive mammogram" might be cognitively misleading for participants because stressing this condition suggests that the sensitivity of the second test would differ given different results from the first test. On the other hand, *not mentioning* the condition logically implies "regardless of the result of the first test," anyway, and this is exactly equivalent to the situation of conditional independence (this analysis holds for women with and without breast cancer as well; see previous theoretical section on conditional independence). In consequence, a version that is formulated as "unconditioned" but does have a footnote stating conditional independence (Table 6) can mathematically and especially cognitively better describe the conditional independence of medical tests than any attempt to explicitly formulate conditional sensitivities and false-alarm rates.

Note that when a tree diagram was also given to participants, the conditional independence was documented anyway because at the lowest level, the two corresponding probabilities (or the two proportions in natural frequency trees) in each case are equal. However, the footnote on conditional independence was presented in each version, that is, in natural frequency versions as well as in versions with tree diagrams (for the sake of comparability). Finally, a physician and a statistician (an expert on medical biometry) from the university hospital[2] checked the formulation of the breast cancer screening scenario in order to ensure the ecological validity of the problem formulations. With respect to generalization 2, the issue of conditional independence is not relevant because only one medical test is included (Table 7).

TABLE 7. Problem wording of the scenarios with respect to generalization 2 (including additional questions concerning generalization 3 below).

	Generalization 2a: 3 test values		Generalization 2b: 3 hypotheses	
	Probability version	**Natural frequency version**	**Probability version**	**Natural frequency version**
Medical situation	Imagine that you are a radiologist and an expert at using mammography for the early detection of breast cancer. In your practice you frequently treat women with a dominant gene mutation for breast cancer. At the moment, you advise a woman with a dominant gene mutation, who has had a positive test result in mammography (positive mammogram). This woman wants to know what this means for her. For your answer, there is the following information available, which is based on a random sample of women, who all have a dominant gene mutation and have had a mammogram:		Imagine that you are a physician and that you frequently examine for whether a person has a particular disease A or a particular disease B. You run a medical test for diagnosing disease A or disease B—but this test cannot discriminate between diseases A and B. Furthermore, diseases A and B never occur at the same time. At the moment, you are treating a person who has no symptoms and you examine to see if the person has disease A or B. This person receives a positive test result and wants to know what this means for him or her. To formulate your answer, there is the following information available, which is based on a random sample of persons with no symptoms who have all had the medical test in question:	
Statistical information	The probability of breast cancer is 13.8% for a woman with a dominant gene mutation. If a woman with a dominant gene mutation actually has breast cancer, the probability is: - 53.7% that she will get a positive mammogram, - 24.6% that she will get an unclear result, and - 21.7% that she will wrongly get a negative mammogram. If a woman with a dominant gene mutation does not have breast cancer, the probability is: - 7.7% that she will wrongly get a positive mammogram, - 7.7% that she will get an unclear result, and - 84.6% that she will get a negative mammogram.	138 out of every 1,000 women with a dominant gene mutation have breast cancer. Out of every 138 women with a dominant gene mutation who actually have breast cancer: - 74 will get a positive mammogram, - 34 will get an unclear result, and - 30 will wrongly get a negative mammogram. Out of every 862 women with a dominant gene mutation who do not have breast cancer: - 66 will wrongly get a positive mammogram, - 66 will get an unclear result, and - 730 will get a negative mammogram.	The probability of having disease A is 0.4% for a person with no symptoms. If a person has disease A, the probability is 90% that she or he will receive a positive test result. The probability of having disease B is 0.1% for a person with no symptoms. If a person has disease B, the probability is 80% that she or he will receive a positive test result. If a person has neither disease A nor disease B, the probability is 10% that she or he will nevertheless receive a positive test result.	40 out of every 10,000 persons with no symptoms have disease A. 36 out of every 40 persons who have disease A receive a positive test result. 10 out of every 10,000 persons with no symptoms have disease B. 8 out of every 10 persons who have disease B receive a positive test result. 995 out of every 9,950 persons who have neither disease A nor disease B receive a positive test result.
Visualization	• no tree diagram, or • tree diagram (prob.)	• no tree diagram, or • tree diagram (nat. freq.)	• no tree diagram, or • tree diagram (prob.)	• no tree diagram, or • tree diagram (nat. freq.)
Question 1	What is the probability that a woman with a dominant gene mutation actually has breast cancer, given that she has had a positive mammogram?	How many of the women with a dominant gene mutation who receive a positive mammogram do you expect to actually have breast cancer?	What is the probability that a person with no symptoms who receives a positive test result, actually suffers from disease A?	How many of the persons with no symptoms who receive a positive test result actually suffer from disease A?
Question 2	What is the probability that a woman with a dominant gene mutation actually has breast cancer, given that she has had an unclear mammogram?	How many of the women with a dominant gene mutation who receive an unclear mammogram do you expect to actually have breast cancer?	What is the probability that a person with no symptoms who receives a positive test result, actually suffers from disease B?	How many of the persons with no symptoms who receive a positive test result actually suffer from disease B?
Question 3	What is the probability that a woman with a dominant gene mutation actually has breast cancer, given that she has had a negative mammogram?	How many of the women with a dominant gene mutation who receive a negative mammogram do you expect to actually have breast cancer?	What is the probability that a person with no symptoms who receives a positive test result, actually suffers from disease A or from disease B?	How many of the persons with no symptoms who receive a positive test result actually suffer from disease A or from disease B?

Solution of the problems and coding

Table 8 shows the solutions of the presented problems and the details of the coding procedure. In the 2-test scenario, the solution was adapted to the actual values for women who have participated in breast cancer screenings and got a positive mammogram and have also undergone another noninvasive clarification, according to the latest evaluation report of the German cooperative association for mammography (Kooperationsgemeinschaft Mammographie GbR [cooperative association mammography], 2016).

TABLE 8. Solution of the 11 different problems and coding.

		Solution		Coded as correct		
Generalization	Question	Probabilities	Natural frequencies	Probabilities	Natural frequencies**	
1a 2 tests	1) P(B | M+ ∩ S+)*	50.6%	76 out of 150	[50%–51%]	76	150
	2) P(B | M+ ∩ S–)	0.45%	4 out of 880	[0.4%–0.5%]	4	880
1b 3 tests	1) P(D | T_1+ ∩ T_2+ ∩ T_3+)	69.9%	114 out of 163	[69%–70%]	114	163
	2) P(D | T_1+ ∩ T_2– ∩ T_3+)	1.6%	9.5 out of 597.5	[1%–2%]	9.5	597.5
	3) P(D | T_1– ∩ T_2+ ∩ T_3–)	1.3%	6 out of 447	[1%–2%]	6	447
2a 3 test values	1) P(B | M+)	52.7%	74 out of 140	[52%–53%]	74	140
	2) P(B | M unclear)	33.8%	34 out of 100	[33%–34%]	34	100
	3) P(B | M–)	3.9%	30 out of 760	[3%–4%]	30	760
2b 3 criterion values	1) P(A | T+)	3.5%	36 out of 1,039	[3%–4%]	36	1,039
	2) P(B | T+)	0.77%	8 out of 1,039	[0.7%–0.8%]	8	1,039
	3) P(A ∪ B | T+)	4.2%	44 out of 1,039	[4%–5%]	44	1,039

Note: * P(B | M+ ∩ S+) denotes the probability question as well as the natural frequency question

** both included absolute numbers must be denoted correctly

In accordance with Gigerenzer and Hoffrage (1995), we classified a response in the probability versions as correct if it was the exact Bayesian solution or rounded to the next percentage level above or below, that is, in the problem version with two tests, all solutions between 50% and 51% were classified as correct, and in the problem with three criterion values, all solutions between 0.7% and 0.8% were classified as correct (Table 8). In natural frequency versions, responses were classified as correct only if both included numbers (e.g., “76” and “150”) were denoted correctly (i.e., we applied a very conservative criterion regarding the natural frequency versions [see also Pighin et al., 2016]).

Participants and Administration

Our sample was based on effect sizes observed in previous studies using a similar design (Binder et al., 2018, McDowell & Jacobs, 2017), which suggest a format effect close to 100% power (95% CI, 96.4% to 100%) with a sample size of about N=120 students. In 2016, we recruited 123 medical students (42 male, 81 female) from the University Hospital Charité Berlin in 2016. All of these 123 students gave answers to 11 non-SBRP tasks (i.e., one judgment according to each line in Table 8). Thus 123 · 11 = 1,353 Bayesian inferences were collected altogether. The students were at different stages of their medical education and participants' ages ranged from 18 to 46 (M = 23.21, SD = 4.52). All students were informed that their participation was voluntary and that anonymity was guaranteed. The study was approved by the ethics committee of the Max Planck Institute for Human Development, Berlin, and participants gave their written informed consent to participating in the study.

The medical students were tested in groups of about 20 persons (on average) at the University Hospital Charité. Trained administrators guaranteed a quiet atmosphere and a professional supervision of the study. Students sitting next to each other always worked on different scenarios. Pocket calculators were distributed and their use allowed at any point during testing. There were no time constraints for completing the tasks. Participants required on average about 60 minutes for all 11 inferences. Medical students received € 15 for participating in the study.

Results

The performance of the medical students is depicted in Figure 7 and in Table 9. First, we present the results specifically regarding each generalization, one after the other, and next, we discuss them as a whole picture. Finally (in the General discussion), we relate our results to similar studies on non-SBRPs (Table 10). Following McDowell, Galesic, and Gigerenzer (2018) and Fiedler et al. (2000), we apply—in addition to the commonly used percentages of Bayesian solutions—the mean absolute deviation from the Bayesian solution of the problem as an alternative scoring metric. Results according to that metric will be discussed for all generalizations taken together in the section General discussion. For inferential statistics we applied one-tailed significance tests in accordance with both of our directional hypotheses on performances (natural frequencies > probabilities; visualization > no visualization).

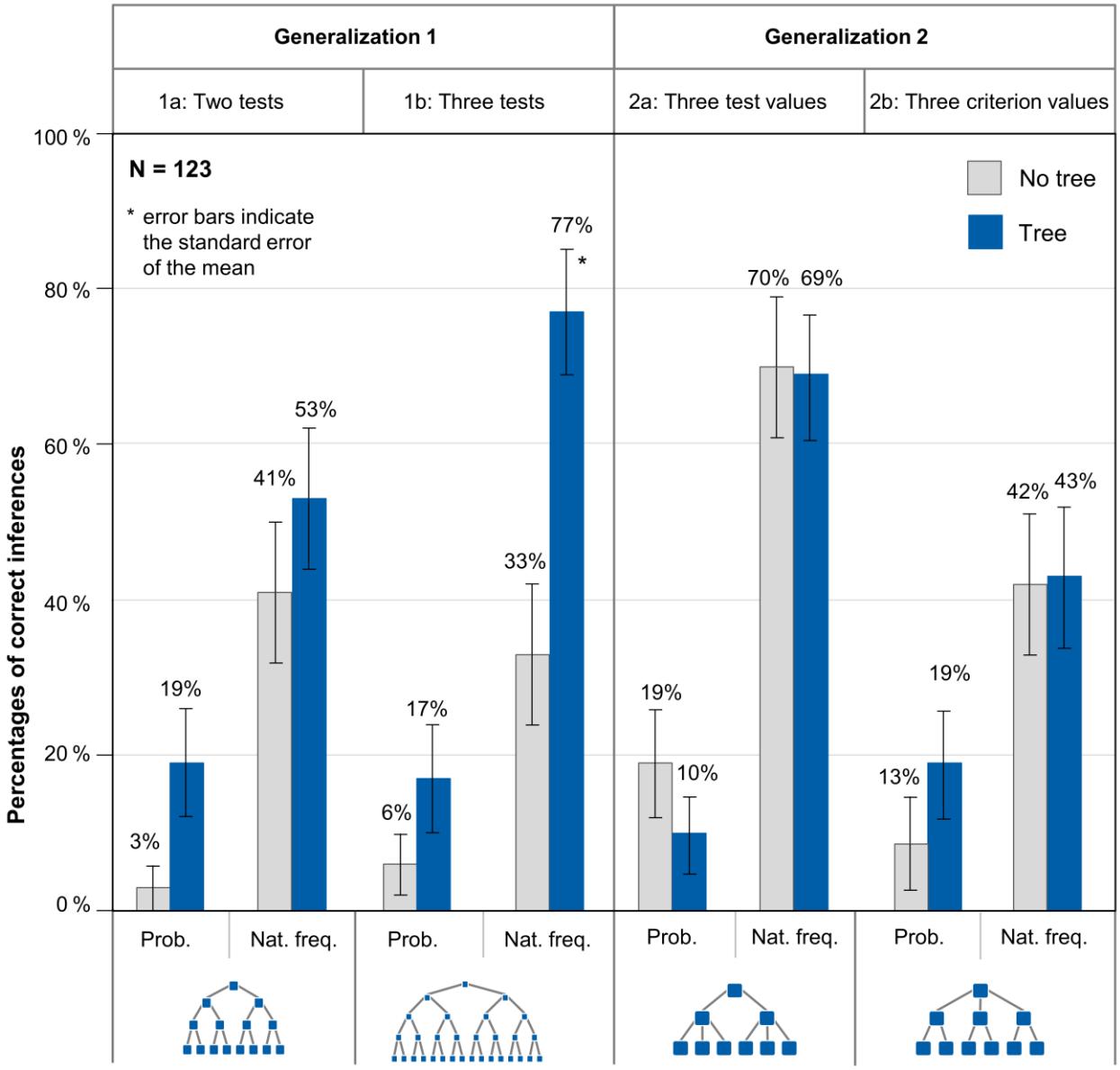


FIGURE 7. Participants' performance in generalizations 1 and 2.

TABLE 9. Participants' performance (N=29-33 for each task) in deriving the positive predictive value (gray highlighted lines) and answering alternative questions (unhighlighted lines) by percentages of correct Bayesian responses and Absolute Deviation of Responses from the Correct Solution[4] and results of the generalized linear mixed models.

Generalization	Scenario	Question	% Bayesian responses (Mean absolute deviation†)				Results of generalized linear mixed model		
			Probabilities		Natural frequencies		Intercept	Natural frequencies (unstandardized) beta	Tree diagram (unstandardized) beta
			No tree	Tree	No tree	Tree			
1a	2 tests	1) P(B | M+∩S+)††	3% (35%)	19% (32%)	41% (23%)	53% (18%)	-2.60***	2.05***	0.86*
3		2) P(B | M+∩S–)	7% (31%)	19% (11%)	34% (16%)	53% (5%)	-2.45***	1.73**	0.91*
1b	3 tests	1) P(D | T_1+∩T_2+∩T_3+)	6% (32%)	17% (26%)	33% (16%)	77% (9%)	-3.15***	2.56***	1.70***
3		2) P(D | T_1+∩T_2–∩T_3+)	9% (25%)	17% (28%)	13% (17%)	68% (4%)	-3.18***	1.74***	1.94***
3		3) P(D | T_1–∩T_2+∩T_3–)	9% (21%)	23% (20%)	13% (19%)	65% (9%)	-2.90***	1.37**	1.95***
2a	3 test values	1) P(B | M+)	19% (38%)	10% (30%)	70% (29%)	69% (23%)	-1.63**	2.62**	-0.32
3		2) P(B | M unclear)	26% (24%)	17% (21%)	77% (19%)	72% (23%)	-1.13*	2.40**	-0.39
3		3) P(B | M–)	23% (49%)	7% (31%)	80% (26%)	66% (16%)	-1.41	2.99*	-1.00*
2b	3 criterion values	1) P(A | T+)	13% (21%)	19% (21%)	42% (4%)	43% (8%)	-1.75***	1.36**	0.18
3		2) P(B | T+)	17% (17%)	16% (16%)	39% (3%)	43% (6%)	-1.70***	1.29**	0.09
3		3) P(A∪B | T+)	13% (7%)	13% (17%)	45% (3%)	47% (3%)	-1.92***	1.75***	0.02

Note: * p<0.05, ** p < 0.01, *** p<0.001

† Participants who did not provide an answer (fewer than 5% across all problems) were excluded.

†† P(B | M+∩S+) denotes the probability question and the natural frequency question as well.

Generalization 1

1a: 2-test scenario

The probability versions of the 2-test scenario (screening for breast cancer by mammogram and sonogram) were solved by 3% without the use of a probability tree and 19% with the use of a probability tree. The performance rate in natural frequency versions was substantially higher; the rate was 41% without the use of a natural frequency tree and 53% with the use of a natural frequency tree. While natural frequencies could increase the performance significantly (both with and without the use of tree diagrams), presenting a tree had a more moderate effect in both information formats (Figure 7).

In order to statistically compare the effects of the information format and the types of visualization, we estimated a generalized linear mixed model with a logit link function to predict performance for the 2-test scenario (Table 9, line one). In this model, we specified the probability version without a tree diagram as the reference category and included the possible explanatory factors "natural frequencies" and "tree diagram" via dummy coding.

The (unstandardized) regression coefficient for natural frequencies was significant ($b_1 = 2.05$, $SE = 0.64$, $z = 3.22$, $p < 0.001$), and presenting a corresponding tree diagram also led to a significant regression coefficient ($b_2 = 0.86$, $SE = 0.50$, $z = 1.71$, $p = 0.04$). Thus in the 2-test scenario, especially natural frequencies but also tree diagrams were helpful for solving the task.

1b: 3-test scenario

A similar pattern appears for the 3-test scenario (Figure 7). The probability versions of the 3-test scenario were solved by 6% without a probability tree and 17% with a probability tree. Performance in the natural frequency versions was higher, 33% without a tree diagram and 77% with a natural frequency tree. Since we assumed that performance in natural frequency 2-test and 3-test scenarios would be similar (because in both scenarios, only two numbers have to be integrated; see Table 3), the much higher score of 77% in the 3-test scenario was clearly unexpected. Presenting tree diagrams did in general increase participants' performance (especially, of course, in the natural frequency format).

In a generalized linear mixed model, analogous to the 2-test case, the (unstandardized) regression coefficients, both for natural frequencies ($b_1 = 2.56$, $SE = 0.53$, $z = 4.88$, $p < 0.001$) and for presenting a tree diagram ($b_2 = 1.70$, $SE = 0.49$, $z = 3.46$, $p < 0.001$), were significant and positive for the 3-test scenario.

Summary of generalization 1

In 2- and 3-test scenarios, medical students would be able to profit—according to our hypothesis—from the use of natural frequencies and tree diagrams. While the positive effect of tree diagrams in the 2-test scenario was moderate, it was especially pronounced in the 3-test scenario. Those results would support an "iteration hypothesis," which assumes that frequency

tree diagrams are especially helpful when "multiple-nested information" has to be displayed. In Table 10 (in General discussion) the results of all experimental studies that examine positive predictive values in non-SBRPs (with or without visualization) conducted thus far are summarized.

Finally, it has to be noted that in both generalization 1a and 1b, we did not implement an interaction term in our generalized linear mixed model because we had no hypotheses about interactions between the factors *information format* and *visualization*. However, implementing the interaction term in our model yields no significant interaction effects.

Generalization 2

2a: Three test values

The probability versions of the scenario including three test values (positive, unclear, or negative) were solved by 19% without a probability tree and 10% with a probability tree (see Figure 7 or Table 9). The performance in natural frequency versions was substantially higher, namely 70% without a frequency tree and 69% with a frequency tree. Thus while natural frequencies also helped in the scenarios with three test values, presenting tree diagrams did *not* increase performance (again giving support to the "iteration hypothesis").

In the respective generalized linear mixed model, we again included the possible explanatory factors "natural frequencies" and "tree diagram" via dummy coding (Table 9). Just as in generalizations 1a and b, the (unstandardized) regression coefficient for natural frequencies was significant ($b_1 = 2.62$, $SE = 0.96$, $z = 0.72$, $p < 0.01$). As already evident from the descriptive performances, presenting an additional tree diagram obviously does not even lead to a significant (or even positive) regression coefficient ($b_2 = -0.32$, $SE = 0.46$, $z = -0.69$, $p = 0.24$).

2b: Three criterion values

In the scenario with three criterion values, the probability versions were solved by 13% without a probability tree and 19% with a probability tree. The performance in natural frequency versions was substantially higher, namely 42% without a natural frequency tree and 43% with a natural frequency tree. As with the results in 2a, presenting a frequency tree did *not* increase participants' performance over pure textual formulations in natural frequencies.

The generalized linear mixed model consequently revealed a significant (unstandardized) regression coefficient for natural frequencies ($b_1 = 1.36$, $SE = 0.49$, $z = 2.77$, $p < 0.01$) but not for presenting an additional tree diagram ($b_2 = 0.18$, $SE = 0.42$, $z = 0.43$, $p = 0.33$).

Summary of generalization 2

Two details have to be noted regarding the results from generalizations 2a and 2b. First, and in contrast to results from generalization 1, natural frequencies but not the presentation of natural

frequency trees were helpful for problem solvers. However, in the subsection "Order effects", it will be demonstrated that the use of natural frequencies *without* natural frequency trees led to an especially high performance when a natural frequency tree had been presented in a previous task (which in general held for all scenarios). Second, the natural frequency versions with three test values were easier to solve than the natural frequency versions with three criterion values. Let us now address the first issue.

Concerning that issue, a potential explanation of those results might be that natural frequency trees are especially helpful with multiple-nested information ("iteration hypothesis"). Tree diagrams that are expanded in width, however, do not display sequentially nested information and so do not relay substantially different information from what is already available in the pure textual problem.

Hoffrage et al. (2015), who investigated such scenarios (but without presenting tree diagrams), also found a difference in natural frequency versions (59% with three test values and 46% with three criterion values) but did not provide any thoughts on potential causes. The difference that occurred in our study (70% vs. 42%), however, is much more pronounced, which seems to hint at a robust effect with respect to the differential difficulties between generalizations 2a and 2b. Note that this difference fits perfectly into our prior analysis on cognitive complexity: While for the natural frequency answer in scenario 2b three numbers have to be integrated, in all other scenarios (1a, 1b, and 2a), only two numbers have to be combined.

Again, with respect to generalizations 2a and 2b, we did not implement interaction terms in our generalized linear mixed models because we had no hypotheses about interactions between the factors *information format* and *visualization*. However, implementing the interaction term in our model, yields no significant interaction effects.

Generalization 3

We now turn to participant performance on the alternative questions posed in 1a, 1b, 2a, and 2b. These questions, referring to alternative test configurations, are displayed in Tables 5, 6, and 7 (below). The results can be found in Table 9 (unhighlighted lines). Notably, performance on alternative questions does not differ substantially from performance on respective "positive predictive values" (highlighted lines) with one exception: In the 3-test scenario there is a decline in correct responses in the natural frequency versions presented without a tree diagram (13% on both alternative questions as compared to 33% on the positive predictive value), while performance remains more stable when a frequency tree is provided (68%/65% vs. 77%).

Note that to answer the alternative question posed in the 2-test scenario $P(B \mid M+ \cap S-)$, only minor additional cognitive effort is required, namely calculating two simple differences ($80 - 76 = 4$ and $950 - 74 = 876$). We speculate that the respective performance rate would also be lower in 2-test scenarios, involving a negative mammogram, a situation which might be addressed in future research. In contrast, concerning alternative questions 2 and 3 in the 3-test

scenario (P(D | T1+ ∩ T2– ∩ T3+) and P(D | T1– ∩ T2+ ∩ T3–)), no longer would all numbers that have to be combined for the natural frequency solution be found in the textual problem, but they would appear in the frequency tree. This is why presenting a tree diagram might have increased participants' performance in natural frequency versions from 13% (the worst performance from all of the natural frequency versions) up to 68%/65%. This performance difference of over 50% again supports the "iteration hypothesis", namely that frequency trees are most powerful when multiple-nested information has to be integrated, and in this case, especially when this information is not available in the wording at all.

In contrast, to answer alternative questions 2 and 3 in the scenarios with three test or criterion values, the natural frequencies needed can be found in the natural frequency wording itself (and the tree diagram does not provide any additional information). Consequently, performance on questions 2 and 3 is similar to performance on question 1, and the (unstandardized) regression coefficients for the four generalized linear mixed models (see Table 9) reveal comparable significant effects with natural frequencies but no positive effects with tree diagrams.

Mean absolute deviation of the correct responses

Following the recommendation of McDowell et al. (2018), we provide in Table 9 the mean absolute deviation from the correct responses as an alternative metric of the correct Bayesian responses regarding each of the 11 questions. Participants who did not provide an answer were excluded (fewer than 5% across all problems).

The analysis yields that participant estimates were closer to the correct solution in natural frequency versions without a tree than in probability versions without a tree across all problems and all considered questions. This result is in accordance with McDowell et al. (2018), who reanalyzed 21 SBRPs using the mean absolute deviation. Furthermore, in 10 out of 11 questions, participant estimates were closer to the correct solution in versions with a natural frequency tree than in versions with a probability tree. In most cases the mean absolute deviation of the correct solution is lower when a tree diagram is provided. Altogether, the alternative metric supports all results obtained and conclusions drawn based on the traditional percentages of Bayesian solutions.

Covariates in generalizations 1–3

Indicators for education

Actual level of medical education ("Semesterzahl") and grade point average (German "Abiturnote," from high school) were collected from all participants, and both variables were implemented as potential predictors in the 11 generalized linear mixed models. It turned out that the level of medical education can significantly predict the probability of solving a task only in the situation with three test values for all stated questions (i.e., the question for the positive predictive value, but also the two alternative questions). However, implementing this factor in

the three respective generalized linear mixed models did not change the results on the presentation of natural frequencies (still significant) or tree diagrams (still not significant). In all other non-SBRPs, neither medical education nor grade point average served as a predictor of participant performance in any task.

Order effects

In an empirical study implementing successive tasks, possible transfer effects might, of course, be of relevance. Although we controlled experimentally for such effects by systematic variation of the order of the tasks, it is still fruitful to take a detailed look at this issue. Concerning probability versions, there was no effect of accumulated learning—it did not matter whether a task was presented as the first, second, third, or fourth task, regardless of whether a probability tree was available or not.

Corresponding analyses of the natural frequency versions revealed a remarkable finding: A comparison of performance on natural frequency versions *without frequency trees* that were positioned as the first or second task with performance when the problem was positioned as the third or fourth task yields significant differences. Whereas across all four scenarios only 32% of the participants who worked on a natural frequency version without a tree early on in the session (first or second task) could solve the problem correctly, 62% of participants were able to solve the same problem when it was positioned toward the end of the session (third or fourth task). It has to be noted that because of our systematic variation of factors, all participants who had to solve a natural frequency version without a tree toward the end of the session had previously seen a natural frequency tree in a prior scenario. Thus there does seem to be a learning effect from the presence of a natural frequency tree in an earlier version on subsequent natural frequency versions without a tree diagram. This finding might partially explain the similar performance levels in natural frequency versions both with and without tree diagrams, especially in generalization 2 (including the alternative question). There was no learning effect on any subsequent task from probability trees or pure natural frequency versions.

Summary and general discussion

In the present article, we theoretically structured possible generalizations of the SBRP (Standard Bayesian Reasoning Paradigm) by extending the number of medical tests (generalization 1), extending the number of test or criterion values (generalization 2), and extending the scope of questions that might be posed regarding alternative test configurations (generalization 3).

For exemplifying the possible variability of information presentation in non-SBRPs, we considered five different extensions from the 1-test to the 2-test scenario theoretically. This detailed analysis was illustrated using frequency trees and addressing potentially helpful features of these kinds of diagrams, namely whether 1) all nested sets refer to one overarching sample; 2) tree(s) represent(s) sequential testing; 3) single sensitivities and specificities are conserved; and 4) questions regarding all four test configurations can be answered. In the

framework of this discussion, the issue of *conditional* independence was addressed in detail, including the remarkable finding that explicitly formulating conditional sensitivities or false-positive rates in probability versions might in fact lead to new misunderstandings, which can be avoided by stating the conditional independence of test results for healthy people and also for ill people (e.g., in a footnote *outside* the sensitivity and false-positive rate stated in the problem wording). Following the paradigm of natural sampling (Kleiter, 1994; Fiedler & Juslin, 2006), natural frequencies automatically carry information on potential conditional independencies.

Tree diagrams were chosen (out of a diversity of competing visualizations) for theoretical explication as well as empirical implementation because they a) can be applied very flexibly to all imaginable generalizations of the SBRP (Figures 2 and 3); b) play an important role in theory formation in the field of judgment and decision-making in general (Martignon et al., 2003; Woike et al., 2017); c) helped participants in previous Bayesian reasoning studies (Binder et al., 2015; Binder et al., 2018; Steckelberg et al., 2004; Yamagishi, 2003); d) can represent all possible test configurations and alternative questions (by each of the two corresponding branches); e) can be constructed easily with paper and pencil (even when extreme base rates are given); and f) are thus perfectly applicable to medical education (and potential further training studies).

In an empirical study, we examined the effect of natural frequencies and of presenting additional tree diagrams on nonstandard Bayesian reasoning tasks. In a 4×2×2 design, four different medical scenarios were implemented: two medical tests (generalization 1a), three medical tests (generalization 1b), three possible test values (generalization 2a), and three possible criterion values (generalization 2b). In all scenarios, performance on alternative questions was also assessed (generalization 3). Each of the 123 medical students recruited from the Charité Berlin had to solve 11 Bayesian reasoning tasks (participants needed about one hour to complete all tasks), producing a total of 1,353 Bayesian inferences to be analyzed.

Statistical analyses revealed that natural frequencies could help with respect to all of these generalizations, while tree diagrams were specifically helpful in generalization 1 (regarding both the generalized positive predictive value and the alternative questions). The conclusion drawn was that tree diagrams (especially natural frequency tree diagrams) are helpful particularly if a) they represent repeated nested settings (there is less successive nesting in the scenario with three test or criterion values), and b) the question posed deviates from the information typically provided (i.e., sensitivities and false-alarm rates, which are both based on positive test results). When comparing the effects of natural frequencies and of tree diagrams, possible transfer effects have to be taken into account. A post hoc analysis revealed that (across all scenarios) natural frequency versions *without tree diagrams* can be solved much better when a frequency tree diagram has been presented in a previous task, thus slightly reducing the effect of natural frequencies in non-SBRPs. In contrast, natural frequency *trees* were of help regardless of task position, in that way revealing a (previously hidden) beneficial effect also present in generalizations 2: In order to foster *immediate* insight, natural frequencies in combination with frequency trees are the means of choice in nonstandard situations.

In Table 10 our results (concerning generalized positive predictive values) are related to previous studies that investigated similar generalizations of the SBRP paradigm. Four studies thus far have empirically addressed comparable scenarios, the majority of them using medical students as participants (only Krauss et al., 1999, tested students from different fields). Thus far, none of these studies implemented tree diagrams or examined alternative questions systematically in all generalizations. However, there are two exceptions, each of which concerned special cases: Binder et al. (2018) had already implemented tree diagrams, but only in a 2-test scenario (generalization 1a). Hoffrage et al. (2015) posed three different questions, but only in a scenario with three criterion values (generalization 2b), and then reported an averaged performance across these different questions (therefore, the performance on alternative questions obtained in the present study cannot be compared to other studies, which is why we did not include them in Table 10).

Table 10, which is based on a total of 1,606 Bayesian inferences, displays the performance of participants in each study and aggregates the corresponding performance of participants across studies (averages at the bottom of the table are weighted according to the number of respective participants). For the sake of clearer communication, we use the column labels 1-22 (see the bottom line in Table 10) in the following. First, aggregated performance on probability versions (uneven column numbers) varies between 0% and 19% across all scenarios. This also holds for every single study, regardless of whether a probability tree was presented or not. Because of the small variance (0%-19%) between conditions in probability format, we concentrate instead on performance on natural frequency versions (even numbers in Table 10), which was the format of interest in the present study. The average performance in frequency versions in all scenarios ranges from 34% up to 77% for the positive predictive value in non-SBRPs. Note that comparing performance across studies for natural frequency versions (even column numbers) revealed a fairly consistent pattern, with perhaps the exception of the 15% in the first line of column 2, thus confirming the aggregation across studies for this format.

Let us first consider 2-test scenarios. Binder et al. (2018) implemented slightly manipulated tree diagrams in Study 2, finding that *highlighting the tree* (i.e., marking the branches relevant to the question while leaving the non-question-related branches unmarked) leads to the highest performance in all 2-test scenarios, but *pruning the tree* (i.e., only showing the question-related branches and cutting the others) does not help, because the full situation is then no longer visible. Consequently, future research might establish whether highlighting trees might also work in generalizations 1b, 2a, and 2b. However, it must be noted that any highlighting can only fit to one certain posed question (e.g., the positive predictive value) and would not then fit questions based on alternative test configurations (of course, pruning a tree would probably be even worse for addressing alternative questions). The power of frequency trees obviously increases when the number of tests and consequently the "nestedness" of information is protracted (generalization 1b, column 14).

TABLE 10. Participants' performance in four different experiments concerning non-SBRPs (N=1,606 Bayesian inferences).

Study Nr.	Gen.	1a										1b				2a				2b			
	Scenario	2 tests										3 tests				3 test values				3 criterion values			
	Text Visualization	Text only		Tree only		Text & tree		Text & Pruned tree		Text & High. tree		Text only		Text & tree		Text only		Text & tree		Text only		Text & tree	
	Format	p	f	p	f	p	f	p	f	p	f	p	f	p	f	p	f	p	f	p	f	p	f
1	Binder et al. (2018), Study 1 N=190	3%	15%	9%	46%	2%	48%																
2	Binder et al. (2018), Study 2 N=198					6%	47%	4%	47%	9%	67%												
3	Hoffrage et al. (2015) N=64	0%	34%									3%	38%			0%	56%			8%	38%		
4	Krauss et al. (1999) N= 123	12%	54%																				
5	Binder & Krauss (present manuscript) N=123	3%	41%			19%	53%					6%	33%	17%	77%	19%	70%	10%	69%	13%	42%	19%	43%
Σ N		166	167	64	63	160	159	67	68	65	66	64	62	30	31	63	62	30	32	62	63	32	30
Weighted average		5%	33%	9%	46%	7%	48%	4%	47%	9%	67%	5%	36%	17%	77%	9%	63%	10%	69%	10%	40%	19%	43%
Column number		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*

Note: p: probability format, f: frequency format

Concerning generalization 2, our purely textual natural frequency versions (columns 16 and 20) replicated the decrease from 2a (three test values) to 2b (three criterion values) of Hoffrage et al. (2015). We could provide a theoretical explanation of this decrease by considering the pieces of information that had to be integrated in a respective solution algorithm (Table 4). The seemingly missing effect of frequency trees might at least partially be attributed to the order effects in our study, and this assumption is in line with the slightly worse performance documented in Hoffrage et al. (2015), where none of the participants had seen a frequency tree earlier.

With respect to alternative questions (generalization 3), we have opened a new research paradigm (and for that reason, our current results cannot be compared with other results in Table 10) that might be investigated in a more systematic way in the future. It has to be noted that not only can alternative *questions* be posed, but alternative *decision strategies* (apart from the Bayesian one) can also be applied in situations with multiple predictors (e.g., the Take the Best heuristic, Gigerenzer & Goldstein, 1999; Gigerenzer, Todd, & ABC research group, 1999; and for an overview, see Woike et al., 2017). In Martignon and Krauss (2003) and Hall et al. (2014), the first hints what may happen when people switch from Bayesian strategies to other non-Bayesian strategies appear.

The present study demonstrates the beneficial effects of natural frequencies in all considered generalizations and of natural frequency trees in generalizations 1 and 3. The advantages of frequency trees based on natural frequencies as tools for teaching and training are that: a) they can be constructed easily with paper and pencil (because they use concrete numbers instead of the geometrical areas that many other types of visualizations do); b) even extreme base rates can be illustrated (which again is problematic for geometrical visualizations); c) they present the "full picture," meaning that all test configurations are displayed so that they can be seen at a glance; d) corresponding conditional probabilities can be displayed on the branches as well (2×2 tables, for instance, can only display *conjoint* probabilities); e) *both formats* can thus be displayed simultaneously in *one* diagram (i.e., the probability trees traditionally used in textbooks can simply be *complemented*); f) they help *immediately* without prior instruction; and g) there seems to be a *learning effect* with natural frequency trees (see also Study 2 in Hoffrage et al., 2015).

In the light of this empirical and theoretical evidence, natural frequencies as well as natural frequency trees (in SBRPs and in non-SBRPs) should be a part of medical education for every aspiring physician, in addition to the other relevant topics regarding risk competence, such as using heuristics (Raab & Gigerenzer, 2015; Wegwarth, Gaissmaier, & Gigerenzer, 2009); problems of 5-year survival rates (Welch, Schwartz, & Woloshin, 2000; Gigerenzer & Wegwarth, 2013); relative vs. absolute change in risks (Gigerenzer, 2009; Schwartz & Meslin, 2008); finding clinically relevant evidence in the literature (e.g., with the PICO model, Akobeng, 2005; Keller, Feufel, Kendel, Spies, & Gigerenzer, 2017); and risk communication (Spiegelhalter & Gage, 2015; Slovic, Monahan, & MacGregor, 2000; Trevena et al., 2013; Prinz, 2017), including the weighing of benefits and risks of screening programs (e.g., with fact boxes, Schwartz, Woloshin, & Welch, 2007; Schwartz, Woloshin, & Welch, 2009; McDowell, Rebitschek, Gigerenzer, & Wegwarth, 2016) or the consumption of a specific medicine or a certain food (e.g., Habs et al., 2017).

Footnotes

[1] In probability theory, conditional independence regarding three events is based on the easier case of statistical independence regarding two events. In the SBRP, disease and test results (both considered events) are obviously statistically dependent, which is why this is not yet a relevant issue in Bayesian reasoning research (but, e.g., see Fiedler, 2000).

[2] We thank Jörg Marienhagen from the University Hospital in Regensburg for reviewing the problem formulation of the breast cancer screening scenario.

References

Akobeng, A. K. (2005). Principles of evidence based medicine. *Archives of disease in childhood*, *90*(8), 837–840. https://doi.org/10.1136/adc.2005.071761

Armstrong, B., Spaniol, J., & Persaud, N. (2018). Does exposure to simulated patient cases improve accuracy of clinicians' predictive value estimates of diagnostic test results? A within-subjects experiment at St Michael's Hospital, Toronto, Canada. *BMJ open*, *8*(2), e019241. https://doi.org/10.1136/bmjopen-2017-019241

Atkinson, M. A., Eisenbarth, G. S., & Michels, A. W. (2014). Type 1 diabetes. *The Lancet*, *383*(9911), 69–82. https://doi.org/10.1016/S0140-6736(13)60591-7

Baratgin, J. (2015). Rationality, the Bayesian standpoint, and the Monty-Hall problem. *Frontiers in Psychology*, *6*, 1168. https://doi.org/10.3389/fpsyg.2015.01168

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *The Behavioral and brain sciences*, *30*(3), 241–297. https://doi.org/10.1017/S0140525X07001653

Berg, W. A., Zhang, Z., Lehrer, D., Jong, R. A., Pisano, E. D., Barr, R. G.,. . . Gabrielli, G. (2012). Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA*, *307*(13), 1394–1404. https://doi.org/10.1001/jama.2012.388

Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information: An empirical study on tree diagrams and 2×2 tables. *Frontiers in Psychology*, *6*(1186). https://doi.org/10.3389/fpsyg.2015.01186

Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (2018). Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making. *PlosONE*, *13*(3). https://doi.org/10.1371/journal.pone.0195029

Brase, G. L. (2008a). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, *15*(2), 284–289. https://doi.org/10.3758/PBR.15.2.284

Brase, G. L. (2008b). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, *23*(3), 369–381. https://doi.org/10.1002/acp.1460

Brase, G. L. (2014). The power of representation and interpretation: Doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *Journal of Cognitive Psychology*, *26*(1), 81–97. https://doi.org/10.1080/20445911.2013.861840

Brase, G. L., & Hill, W. T. (2015). Good fences make for good neighbors but bad science: A review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, *6*, 340. https://doi.org/10.3389/fpsyg.2015.00340

Bröder, A. (2000). Assessing the empirical validity of the "Take-the-best" heuristic as a model of human probabilistic inference. *Journal of experimental psychology: Learning, memory, and cognition*, *26*(5), 1332–1346. https://doi.org/10.1037/0278-7393.26.5.1332

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society / B*, *41*(1), 1–31.

Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., & Feinstein, A. R. (1994). Variability in radiologists' interpretations of mammograms. *The New England journal of medicine*, *331*(22), 1493–1499. https://doi.org/10.1056/NEJM199412013312206

Fenton, N., Neil, M., & Berger, D. (2016). Bayes and the Law. *Annual review of statistics and its application*, *3*, 51–77. https://doi.org/10.1146/annurev-statistics-041715-033428

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659–676. https://doi.org/10.1037//0033-295X.107.4.659

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General*, *129*(3), 399–418. https://doi.org/10.1037/0096-3445.129.3.399

Fiedler, K., & Juslin, P. (2006). *Information sampling and adaptive cognition*. Cambridge, New York: Cambridge University Press.

Fryback, D. G. (1978). Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Computers and Biomedical Research*, *11*(5), 423–434. https://doi.org/10.1016/0010-4809(78)90001-0

Garcia-Retamero, R., Hoffrage, U., & Dieckmann, A. (2007). When one cue is not enough: combining fast and frugal heuristics with compound cue processing. *Quarterly journal of experimental psychology*, *60*(9), 1197–1215. https://doi.org/10.1080/17470210600937528

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The Take the Best heuristic. In G. Gigerenzer, P. Todd, & ABC research group (Eds.), *Simple heuristics that make us smart*. Oxford University Press.

Gigerenzer, G., & Gray, J. A. M. (2011). Launching the century of the patient. In G. Gigerenzer & J. A. M. Gray (Eds.). *Better doctors, better patients, better decisions. Envisioning health care 2020* (pp. 3–28). Cambridge, Mass.: MIT.

Gigerenzer, G. (2008). Fast and Frugal Heuristics: The Tools of Bounded Rationality. In D. J. Koehler (Ed.), *Blackwell Handbook of Judgment and Decision Making* (2004th ed., pp. 62–88). Chichester: Wiley. https://doi.org/10.1002/9780470752937.ch4

Gigerenzer, G. (2009). Making sense of health statistics. *Bulletin of the World Health Organization*, *87*(8), 567. https://doi.org/10.2471/BLT.09.069872

Gigerenzer, G., Czerlinski, J., & Martignon, L. (1999). How Good are Fast and Frugal Heuristics? In J. Shanteau, B. A. Mellers, & D. A. Schum (Eds.), *Decision Science and Technology. Reflections on the Contributions of Ward Edwards* (pp. 81–103). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4615-5089-1_6

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033295X.102.4.684

Gigerenzer, G., Todd, P., & ABC research group (Eds.). (1999). *Simple heuristics that make us smart*: Oxford University Press.

Gigerenzer, G., & Wegwarth, O. (2013). Five year survival rates can mislead. *BMJ (Clinical research ed.)*, *346*, f548. https://doi.org/10.1136/bmj.f548

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition*, *78*(3), 247–276. https://doi.org/10.1016/S0010-0277(00)00133-5

Green, L., & Mehr, D. R. (1997). What alters physicians' decisions to admit to the coronary care unit? *The Journal of family practice*, *45*(3), 219–226.

Habs, M., Binder, K., Krauss, S., Müller, K., Ernst, B., Valentini, L., & Koller, M. (2017). A Balanced Risk-Benefit Analysis to Determine Human Risks Associated with Pyrrolizidine Alkaloids (PA)-The Case of Tea and Herbal Infusions. *Nutrients*, *9*(7). https://doi.org/10.3390/nu9070717

Hall, S., Phang, S. H., Schaefer, J. P., Ghali, W., Wright, B., & McLaughlin, K. (2014). Estimation of post-test probabilities by residents: Bayesian reasoning versus heuristics? *Advances in health sciences education: theory and practice*, *19*(3), 393–402. https://doi.org/10.1007/s10459-013-9485-1

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*(5), 538–540. https://doi.org/10.1097/00001888-199805000-00024

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, *84*(3), 343–352. https://doi.org/10.1016/S0010-0277(02)00050-1

Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Frontiers in psychology*, *6*, 1473. https://doi.org/10.3389/fpsyg.2015.01473

Jarecki, J. B., Meder, B., & Nelson, J. D. (2017). Naïve and Robust: Class-Conditional Independence in Human Classification Learning. *Cognitive Science*, *14*(3), 471. https://doi.org/10.1111/cogs.12496

Jenny, M. A., Pachur, T., Lloyd Williams, S., Becker, E., & Margraf, J. (2013). Simple rules for detecting depression. *Journal of Applied Research in Memory and Cognition*, *2*(3), 149–157. https://doi.org/10.1016/j.jarmac.2013.06.001

Johnson, E. D., & Tubau, E. (2013). Words, numbers, & numeracy: Diminishing individual differences in Bayesian reasoning. *Learning and Individual Differences*, *28*, 34–40. https://doi.org/10.1016/j.lindif.2013.09.004

Johnson, E. D., & Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00938

Karadawi, N., O'Kane, G., Gallagher, D., Finn, S., Muldoon, C., & Mulligan, N. (2016). Cascade testing following universal immunohistochemistry for mismatch repair protiens. *Annals of Oncology*, *27*(suppl_6). https://doi.org/10.1093/annonc/mdw370.111

Keller, N., Feufel, M. A., Kendel, F., Spies, C. D., & Gigerenzer, G. (2017). Training medical students how to extract, assess and communicate evidence from an article. *Medical education*, *51*(11), 1162–1163. https://doi.org/10.1111/medu.13444

Khan, A., Breslav, S., Glueck, M., & Hornbæk, K. (2015). Benefits of visualization in the mammography problem. *International Journal of Human-Computer Studies*, *83*, 94–113. https://doi.org/10.1016/j.ijhcs.2015.07.001

Kipling, R. (1999). Benchmarks for fast and frugal heuristics. In G. Gigerenzer, P. Todd, & ABC research group (Eds.), *Simple heuristics that make us smart*. Oxford University Press.

Kleiter, G. D. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer & Laming, D. R. J (Eds.), *Contributions to Mathematical Psychology, Psychometrics, and Methodology* (pp. 375–388). New York: Springer. https://doi.org/10.1007/978-1-4612-4308-3_27

Kooperationsgemeinschaft Mammographie GbR. (2016). *Jahresbericht Evaluation 2013 – Deutsches Mammographie-Screening-Programm.* [German mammography screening programme – cooperative association mammography].

Krauss, S., Martignon, L., & Hoffrage, U. (1999). Simplifying Bayesian Inference: The General Case. In N. e. a. Magnani (Ed.), *Model-based Reasoning in Scientific Discovery* (pp. 165–179). https://doi.org/10.1007/978-1-4615-4813-3_11

Krauss, S., & Wang, X. T. (2003). The psychology of the monty hall problem: Discovering psychological mechanism for solving a tenacious brain teaser. *Journal of Experimental Psychology: General*, *132*(1), 3–22.

Lee, K., Park, H.-D., & Kang, E.-S. (2013). Reduction of the HIV seroconversion window period and false positive rate by using ADVIA Centaur HIV antigen/antibody combo assay. *Annals of laboratory medicine*, *33*(6), 420–425. https://doi.org/10.3343/alm.2013.33.6.420

Lewis, C., & Keren, G. (1999). On the difficulties underlying Bayesian reasoning: a comment on Gigerenzer and Hoffrage. *Psychological Review*, *106*(2), 411–416. https://doi.org/10.1037/0033-295X.106.2.411

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec & C. Rouveirol (Eds.), Lecture Notes in Computer Science: Vol. 1398. Machine learning. ECML-98: *10th european conference on machine learning.* Chenmitz, Germany, April 21-23, 1998 : proceedings (Vol. 1398, pp. 4–15). Berlin, Heidelberg: Springer. https://doi.org/10.1007/BFb0026666

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Frontiers in Psychology*, *5*(1144). https://doi.org/10.3389/fpsyg.2014.01144

Mandel, D. R., & Navarrete, G. (2015). Editorial: Improving Bayesian reasoning: What works and why? *Frontiers in psychology*, *6*, 1872. https://doi.org/10.3389/fpsyg.2015.01872

Martignon, L., & Krauss, S. (2003). Can l'homme eclaire be fast and frugal? Reconciling Bayesianism and bounded rationality. In S. L. Schneider & J. Shanteau (Eds.), Cambridge series on judgment and decision making. *Emerging perspectives on judgment and decision research* (pp. 108–122). Cambridge: Cambridge University Press.

Martignon, L., & Schmitt, M. (1999). Simplicity and Robustness of Fast and Frugal Heuristics. *Minds and Machines*, *9*(4), 565–593. https://doi.org/10.1023/A:1008313020307

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. R. (2003). Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment and decision making*. Wiley.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge: MIT Press.

McDowell, M., Galesic, M., & Gigerenzer, G. (2018). Natural Frequencies Do Foster Public Understanding of Medical Tests: Comment on Pighin, Gonzalez, Savadori and Girotto (2016). *Medical decision making: an international journal of the Society for Medical Decision Making*, 272989X18754508. https://doi.org/10.1177/0272989X18754508

McDowell, M., & Jacobs, P. (2017). Meta-Analysis of the Effect of Natural Frequencies on Bayesian Reasoning. *Psychological bulletin*, *143*, 1273–1312. https://doi.org/10.1037/bul0000126

McDowell, M., Rebitschek, F. G., Gigerenzer, G., & Wegwarth, O. (2016). A Simple Tool for Communicating the Benefits and Harms of Health Interventions. *MDM Policy & Practice*, *1*(1), 238146831666536. https://doi.org/10.1177/2381468316665365

McGee, S. R. (2012). *Evidence-based physical diagnosis* (3rd ed.). Philadelphia: Elsevier/Saunders.

McNair, S. J. (2015). Beyond the status-quo: Research on Bayesian reasoning must develop in both theory and method. *Frontiers in Psychology*, *6*, 97. https://doi.org/10.3389/fpsyg.2015.00097

Mellers, B. A., & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer und Hoffrage (1995). *Psychological Review*, *106*(2), 417–424. https://doi.org/10.1037/0033-295X.106.2.417

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. https://doi.org/10.1109/TVCG.2012.199

Moro, R., Bodanza, G. A., & Freidin, E. (2011). Sets or frequencies? How to help people solve conditional probability problems. *Journal of Cognitive Psychology*, *23*(7), 843–857. https://doi.org/10.1080/20445911.2011.579072

Mousavi, S., & Gigerenzer, G. (2014). Risk, uncertainty, and heuristics. *Journal of Business Research*, *67*(8), 1671–1678. https://doi.org/10.1016/j.jbusres.2014.02.013

Navarrete, G., Correia, R., & Froimovitch, D. (2014). Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. *Frontiers in Psychology*, *5*(1272). https://doi.org/10.3389/fpsyg.2014.01272

Ohuchi, N., Suzuki, A., Sobue, T., Kawai, M., Yamamoto, S., Zheng, Y.-F.,. . . Ishida, T. (2016). Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): A randomised controlled trial. *The Lancet*, *387*(10016), 341–348. https://doi.org/10.1016/S0140-6736(15)00774-6

Operskalski, J. T., & Barbey, A. K. (2016). Risk literacy in medical decision-making. *Science* (New York, N.Y.), 352(6284), 413–414. https://doi.org/10.1126/science.aaf7966

Partik, B., Mallek, R., Rudas, M., Pokieser, P., Wunderbaldinger, P., & Helbich, T. H. (2001). Maligne und benigne Erkrankungen der Brust bei 41 männlichen Patienten: Mammographie und Sonographie mit histopathologischer Korrelation [Malignant and benign diseases of the breast in 41 male patients: mammography, sonography and pathohistological correlations]. *RoFo: Fortschritte auf dem Gebiete der Röntgenstrahlen und der Nuklearmedizin*, *173*(11), 1012–1018. https://doi.org/10.1055/s-2001-18336

Pearl, J. (2008). Probabilistic reasoning in intelligent systems: Networks of plausible inference (Rev. 2. print., rev. & updated.). *The Morgan Kaufmann series in representation and reasoning*: Books. San Francisco, CA: Kaufmann.

Pighin, S., Gonzalez, M., Savadori, L., & Girotto, V. (2016). Natural frequencies do not foster public understanding of medical test results. *Medical decision making: an international journal of the Society for Medical Decision Making*, *36*(6), 686–691. https://doi.org/10.1177/0272989X16640785

Pohl, R. (2017). *Cognitive illusions: Intriguing phenomena in judgement, thinking and memory* (2nd edition). Abingdon, Oxon, New York, NY: Routledge.

Prakken, H. (2014). On direct and indirect probabilistic reasoning in legal proof. *Law, Probability and Risk*, *13*(3-4), 327–337. https://doi.org/10.1093/lpr/mgu013

Prinz, R. (2017). Risk Comprehension and Communication - Insights from current grievances and ways to resolve them. *European journal of public health*, *27*(3). https://doi.org/10.1093/eurpub/ckx187.518

Prinz, R., Feufel, M., Gigerenzer, G., & Wegwarth, O. (2015). What counselors tell low-risk clients about HIV test performance. *Current HIV research*, *13*(5), 369–380. https://doi.org/10.2174/1570162X13666150511125200

Raab, M., & Gigerenzer, G. (2015). The power of simplicity: A fast-and-frugal heuristics approach to performance science. *Frontiers in Psychology*, *6*, 1672. https://doi.org/10.3389/fpsyg.2015.01672

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science* (New York, N.Y.), 309(5736), 892–895. https://doi.org/10.1126/science.1111565

Schneps, L., & Colmez, C. (2013). *Wahrscheinlich Mord: Mathematik im Zeugenstand*. München: Hanser.

Schrader, M., Zengerling, F., Hakenberg, O. W., & Protzel, C. (2016). Nationale Zweitmeinungsnetzwerke Hodentumoren und Peniskarzinom: Zwei "Evidenz-Pipelines" [German national second-opinion network for testicular cancer and penile carcinoma : Two sources for evidence-based information]. *Der Urologe. Ausg. A*, *55*(9), 1192–1198. https://doi.org/10.1007/s00120-016-0201-7

Schwartz, L. M., Woloshin, S., & Welch, H. G. (2007). The drug facts box: Providing consumers with simple tabular data on drug benefit and harm. *Medical decision making: an international journal of the Society for Medical Decision Making*, *27*(5), 655–662. https://doi.org/10.1177/0272989X07306786

Schwartz, L. M., Woloshin, S., & Welch, H. G. (2009). Using a Drug Facts Box to Communicate Drug Benefits and Harms: Two Randomized Trials. *Annals of Internal Medicine*, *150*(8), 516. https://doi.org/10.7326/0003-4819-150-8-200904210-00106

Schwartz, P. H., & Meslin, E. M. (2008). The ethics of information: Absolute risk reduction and patient understanding of screening. *Journal of general internal medicine*, *23*(6), 867–870. https://doi.org/10.1007/s11606-008-0616-y

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*(3), 380–400. https://doi.org/10.1037//0096-3445.130.3.380

Shen, Y., Wu, D., & Zelen, M. (2001). Testing the Independence of Two Diagnostic Tests. *Biometrics*, *57*(4), 1009–1017. https://doi.org/10.1111/j.0006-341X.2001.01009.x

Siegrist, M., & Keller, C. (2011). Natural frequencies and Bayesian reasoning: The impact of formal education and problem context. *Journal of Risk Research*, *14*(9), 1039–1055. https://doi.org/10.1080/13669877.2011.571786

Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonomic bulletin & review*, *21*(1), 198–204. https://doi.org/10.3758/s13423-013-0464-6

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational behavior and human decision processes*, *91*(2), 296–309. https://doi.org/10.1016/S0749-5978(03)00021-9

Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior*, *24*(3), 271–296. https://doi.org/10.1023/A:1005595519944

Spiegelhalter, D., & Gage, J. (2015). What can education learn from real-world communication of risk and uncertainty? *The mathematics enthusiast*. (12).

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*(6048), 1393–1400. https://doi.org/10.1126/science.1191181

Steckelberg, A., Balgenorth, A., Berger, J., & Mühlhauser, I. (2004). Explaining computation of predictive values: 2 x 2 table versus frequency tree. A randomized controlled trial [ISRCTN74278823]. *BMC medical education*, *4*, 13. https://doi.org/10.1186/1472-6920-4-13

Stine, G. J. (1996). *Acquired immune deficiency syndrome: Biological, medical, social, and legal issues*. Englewood Cliff, NJ: Prentice Hall.

Talboy, A. N., & Schneider, S. L. (2016). Improving accuracy on Bayesian inference problems using a brief tutorial. *Journal of Behavioral Decision Making*, *30*(2), 373–388. https://doi.org/10.1002/bdm.1949

Todd, P. M., & Gigerenzer, G. (2007). Environments That Make Us Smart. *Current Directions in Psychological Science*, *16*(3), 167–171. https://doi.org/10.1111/j.1467-8721.2007.00497.x

Trevena, L. J., Zikmund-Fisher, B. J., Edwards, A., Gaissmaier, W., Galesic, M., Han, P. K. J.,. . . Woloshin, S. (2013). Presenting quantitative information about decision outcomes: A risk communication primer for patient decision aid developers. *BMC medical informatics and decision making*, *13*(2), S7. https://doi.org/10.1186/1472-6947-13-S2-S7

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

Vecchio, T. J. (1966). Predictive value of a single diagnostic test in unselected populations. *The New England journal of medicine*, *274*(21), 1171–1173. https://doi.org/10.1056/NEJM196605262742104

Wegwarth, O., Gaissmaier, W., & Gigerenzer, G. (2009). Smart strategies for doctors and doctors-in-training: Heuristics in medicine. *Medical education*, *43*(8), 721–728. https://doi.org/10.1111/j.1365-2923.2009.03359.x

Welch, H. G., Schwartz, L. M., & Woloshin, S. (2000). Are Increasing 5-Year Survival Rates Evidence of Success Against Cancer? *JAMA*, *283*(22), 2975. https://doi.org/10.1001/jama.283.22.2975

Woike, J. K., Hoffrage, U., & Martignon, L. (2017). Integrating and Testing Natural Frequencies, Naïve Bayes, and Fast-and-Frugal Trees. *Decision*. Advance online publication. https://doi.org/10.1037/dec0000086

Wu, C. M., Meder, B., Filimon, F., & Nelson, J. D. (2017). Asking Better Questions: How Presentation Formats Influence Information Search. *Journal of experimental psychology: Learning, memory, and cognition*. Advance online publication. https://doi.org/10.1037/xlm0000374

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? *Experimental psychology*, *50*(2), 97–106. https://doi.org/10.1026//1618-3169.50.2.97

Diskussion

Übersicht über die erzielten Ergebnisse der drei Artikel

Tabelle 2 stellt die zentralen Ergebnisse (d.h. die Performanzen der Versuchspersonen) der drei Artikel im Überblick dar. Insgesamt zeigt sich ein sehr einheitliches Bild, das die positiven Effekte natürlicher Häufigkeiten und der Visualisierung durch Baumdiagramme – vor allem durch Häufigkeitsbäume (und Vierfeldertafeln mit absoluten Häufigkeiten) – belegt. Egal ob Schülerinnen und Schüler Bayesianische 1-Test-Fälle bearbeiteten oder Medizinstudierende 2-Test-Fälle oder 3-Test-Fälle, natürliche Häufigkeiten und Häufigkeitsbäume unterstützen Menschen in ihren Entscheidungsfindungsprozessen ganz deutlich. Für Situationen mit drei Testergebnissen oder drei Hypothesen ist hingegen die Präsentation der statistischen Informationen in Form von natürlichen Häufigkeiten entscheidend. Wenn diese auch noch visualisiert werden, verbessert dies das Verständnis der Studienteilnehmer nicht zusätzlich (aber siehe Diskussion des dritten Artikels zu Lerneffekten, ab Seite 104).

Da sowohl im zweiten als auch im dritten Artikel Medizinstudierende die Aufgaben bearbeiteten, ist ein Vergleich der in beiden Artikeln erzielten Ergebnisse besonders interessant. Dieser Vergleich findet sich ausführlich in der Diskussion des dritten Artikels, der nicht nur die Ergebnisse aus Artikel 2 und 3 miteinander in Beziehung setzt, sondern diese auch mit bisherigen Erkenntnissen aus früheren Studien mit Medizinstudierenden in Verbindung bringt. Die Performanz der Versuchspersonen bei alternativen Fragestellungen (also Fragen nach der Wahrscheinlichkeit einer Erkrankung, die nicht nur positive Testergebnisse beinhalteten, sondern z.B. auch negative Testergebnisse oder unklare Befunde) wird in Tabelle 2 nicht dargestellt, da diese im dritten Artikel erstmals fokussiert wurden und aufgrund dieser Originalität nicht mit früheren Studien in Verbindung gebracht werden können. Inwiefern sich die Lösungsraten der Studierenden beim positiven Vorhersagewert von denen bei alternativen Fragen unterscheiden, wird allerdings ebenfalls ausführlich in der Diskussion des dritten Artikels dargelegt.

Die Ergebnisse des ersten Artikels lassen sich dagegen nur bedingt mit den Befunden der anderen beiden Artikel in Verbindung bringen, da für den ersten Artikel keine Medizinstudierenden, sondern Schülerinnen und Schüler als Versuchspersonen rekrutiert wurden. Die Vergleichbarkeit der verschiedenen Stichproben ist hierbei in zweierlei Hinsicht eingeschränkt: 1. Aus früheren Studien ist bekannt, dass medizinische Experten Bayesianische Aufgaben zu medizinischen Kontexten etwas besser lösen können als medizinische Laien (siehe z.B. McDowell & Jacobs, 2018). 2. Es konnte in früheren Studien bereits gezeigt werden, dass Menschen mit höherer Rechenfertigkeit („Numeracy“, siehe z.B. McDowell & Jacobs, 2017) deutlich höhere Lösungsraten erzielen als solche mit niedrigerer Rechenfertigkeit. Inwiefern sich die beiden Stichproben hinsichtlich der unterschiedlichen Rechenfertigkeiten unterschieden haben, wurde jedoch nicht empirisch überprüft (allerdings wurde die Abiturnote erfasst). Die Schülerinnen und Schüler, die im ersten Artikel sowohl einen medizinischen als auch einen nichtmedizinischen

Kontext bearbeiten sollten, erzielten im medizinischen Kontext auch deutlich niedrigere Lösungsraten als im nichtmedizinischen Kontext (siehe auch Ergebnisse und Diskussion von Artikel 1, ab Seite 31). Ein Vergleich der Ergebnisse aus dem ersten Artikel mit Ergebnissen aus früheren Studien zu Bayesianischen Standardaufgaben findet sich allerdings in der Diskussion zum ersten Artikel und inzwischen auch in der Meta-Analyse zum Effekt natürlicher Häufigkeiten von McDowell und Jacobs (2017), in der neben 34 anderen Studien auch die Ergebnisse dieses ersten Artikels aufgenommen wurden.

Eine weitere Steigerung der Lösungsraten kann durch das Markieren der Äste des Baumdiagramms erreicht werden, die für die Lösungsfindung relevant sind (siehe Artikel 2), während ein Abschneiden der irrelevanten Äste keinen zusätzlichen Effekt hat im Vergleich zur Präsentation eines vollständigen Baumdiagramms. Inwiefern die Markierung der lösungsrelevanten Äste auch bei 3-Test-Fällen und Situationen mit drei Hypothesen oder drei Testergebnissen zielführend ist, kann Gegenstand zukünftiger Forschungsarbeiten sein. Eine weitere spannende Forschungsfrage für künftige Analysen ist sicherlich die Zusammenführung der beiden Paradigmen „Bayesian reasoning“ und „Fast and Frugal decision making“ (siehe z.B. Martignon & Krauss, 2003 bzw. Woike, Hoffrage, & Martignon, 2017) und somit die systematische Untersuchung des Übergangs von Bayesianischem zu heuristischem Denken.

Tabelle 2. Überblick über die Ergebnisse der drei Artikel

Wahrscheinlichkeiten

Häufigkeiten

		1-Test-Fall			2-Test-Fall					3-Test-Fall		3 Testergebnisse		3 Hypothesen	
		Nur Text	Text und Vierf.-Tafel	Text und vollst. Baum	Nur Text	Nur vollst. Baum	Text und vollst. Baum	Text und red. Baum	Text und mark. Baum	Nur Text	Text und vollst. Baum	Nur Text	Text und vollst. Baum	Nur Text	Text und vollst. Baum
Artikel 1 Binder, Krauss & Bruckmaier (2015)	*N*=259 Schülerinnen und Schüler Regensburg	2%	7%	5%											
		26%	58%	45%											
Artikel 2 Binder, Krauss, Bruckmaier & Marienhagen (2018)	**Studie 1** *N*=190 Medizinstudierende Regensburg				3%	9%	2%								
					15%	46%	48%								
	Studie 2 *N*=198 Medizinstudierende Regensburg						6%	4%	9%						
							47%	47%	67%						
Artikel 3 Binder & Krauss (eingereicht)	*N*=123 Medizinstudierende Berlin				3%		19%			6%	17%	19%	10%	13%	19%
					41%		53%			33%	77%	70%	69%	42%	43%

Bedeutung der Ergebnisse für den schulischen Unterricht und die universitäre Lehre

Der in Artikel 1 vorgestellte Beleg, dass Baumdiagramme und Vierfeldertafeln mit Wahrscheinlichkeiten das Verständnis der bayerischen Schülerinnen und Schüler nicht unterstützen, während die Häufigkeitsbäume hingegen sehr hilfreich waren, obwohl sie zuvor nicht unterrichtet wurden, führte zur Implementierung von Häufigkeitsbäumen in den neuen gymnasialen LehrplanPlus (ISB, 2018, LehrplanPlus Gymnasium Bayern, Mathematik 10. Klasse). Es bleibt zu hoffen, dass das Konzept der natürlichen Häufigkeiten sowie deren visuelle Darstellung durch Häufigkeitsbäume (wie z.B. in Binder et al., 2018 vorgeschlagen) in naher Zukunft flächendeckend in der schulischen und universitären Ausbildung eingesetzt werden, um die Risikokompetenz unserer Gesellschaft stärken zu können.

Die Übertragbarkeit der Ergebnisse auf medizinische Kontexte (Artikel 2 und 3) mit mehreren Testergebnissen (oder Symptomen) oder auf solche, in denen neben einem positiven und negativen Testergebnis beispielsweise auch ein unklarer Befund vorliegen kann oder in denen der Test zur Erkennung mehrerer verschiedener Erkrankungen eingesetzt wird, macht Baumdiagramme mit natürlichen Häufigkeiten zu einem besonders mächtigen Werkzeug. Die medizinische Fakultät des Universitätsklinikums Regensburg ist daher bestrebt, Häufigkeitsbäume möglichst bald in die universitäre Lehre für Medizinstudierende einzubinden. Konkrete Lehrempfehlungen lassen sich auch aus dem in Artikel 2 untersuchten und vielversprechenden Markieren der lösungsrelevanten Äste des Baumdiagramms ableiten. Da typische falsche Antworten bei Bayesianischen Aufgaben lediglich den Ast der kranken Frauen mit positiven Testergebnissen berücksichtigen und die zusätzliche Markierung des Astes der gesunden Frauen, die fälschlicherweise positive Testergebnisse erhalten, zu einer deutlichen Verbesserung im Verständnis der Aufgabe geführt hat, sollte dieser Ast im schulischen Unterricht und der medizinischen Lehre deutlich mehr in den Vordergrund gerückt werden (siehe auch Binder & Marienhagen, 2017).

Das Ziel soll nicht sein, dass Ärztinnen und Ärzte in ihrem beruflichen Alltag immer und überall Baumdiagramme mit Häufigkeiten erstellen, um korrekte Inferenzen bilden zu können (siehe auch Woike et al., 2017); vielmehr ermöglicht der Umgang mit Häufigkeitsbäumen im Medizinstudium und in Fortbildungen für praktizierende Ärztinnen und Ärzte, dass diese ein Gefühl dafür entwickeln können, welchen Einfluss die Prävalenz einer Erkrankung, aber auch die Sensitivität und Falsch-Positiv-Rate eines Tests auf den positiven Vorhersagewert hat (was auch im Rahmen eines DFG-Antrags, der in Kooperation mit Mathematik-Didaktikern in Heidelberg und Kassel eingereicht wird, in einer Trainingsstudie mit Medizinern und Juristen untersucht werden soll). Dies beugt dem weit verbreiteten Fehler vor, von einer sehr hohen Sensitivität und/oder einer niedrigen Falsch-Positiv-Rate eines medizinischen Tests geblendet zu werden, ohne die Prävalenz zu berücksichtigen (auch Basisratenvernachlässigung genannt).

Bayesianische Aufgaben (auch solche, die über die Bayesianischen Standardaufgaben hinausgehen) sollten demnach im schulischen Unterricht und in der medizinischen Ausbildung

mithilfe von natürlichen Häufigkeiten und Häufigkeitsbaumdiagrammen unterrichtet werden, neben weiteren Themen zur Stärkung der Risikokompetenz wie der Problematik mit 5- und 10-Jahresüberlebensraten (Gigerenzer & Wegwarth, 2013; Wegwarth, Wagner, & Gigerenzer, 2017), dem heuristischen Denken (siehe z.B. Gigerenzer et al., 1999, Dieckmann & Krauss, 2005) und dem Unterschied zwischen relativen und absoluten Risiko- oder Chancenveränderungen (Hoffrage et al., 2000; Gigerenzer, Wegwarth, & Feufel, 2010). Letzteres ist nun ebenfalls fest im schulischen Curriculum des bayerischen Gymnasiums verankert, was sehr zu begrüßen ist, wenn man bedenkt, wie häufig man in den Medien mit irreführenden relativen Veränderungen von Anteilswerten konfrontiert wird (siehe z.B. Bauer, Gigerenzer, & Krämer, 2014) und wie leicht und verständlich man derartige Sachverhalte beispielsweise mithilfe von Faktenboxen oder visualisierten Faktenboxen darstellen kann (Schwartz, Woloshin, & Welch, 2007; Schwartz, Woloshin, & Welch, 2009; McDowell, Rebitschek, Gigerenzer, & Wegwarth, 2016; Habs et al., 2017).

Gerd Bosbach und Jürgen Korff (2017) postulieren in Ihrem Buch „Die Zahlentrickser“, dass es positive Effekte auf die Gesellschaft hätte, wenn diese besser mit statistischen Informationen umgehen könnte: Denn „wenn Zahlentricks auffliegen und öffentlich kritisiert werden, steigt die Qualität des Zahlenmaterials und zugleich das demokratische Niveau politischer und wirtschaftlicher Debatten“ (Bosbach & Korff, 2017, Seite 12). In diesem Sinne wäre eine möglichst frühe Ausbildung im statistischen Denken und der Kommunikation von Risiken und Chancen wünschenswert (siehe auch Gigerenzer, 2013, Kapitel 12). Die vorliegende Arbeit beleuchtet daher mit der Thematik „Bayesianisches Denken“ einen ganz zentralen und vielfach beforschten Baustein der Risikokompetenz, der neben anderen Formen der Risikokompetenz flächendeckend und möglichst früh in der schulischen Ausbildung und der universitären Lehre gefördert werden sollte.

Literatur

Barker, M. J. (2017). Connecting Applied and Theoretical Bayesian Epistemology: Data Relevance, Pragmatics, and the Legal Case of Sally Clark. *Journal of Applied Philosophy*, *34*(2), 242–262. https://doi.org/10.1111/japp.12181

Bauer, T. K., Gigerenzer, G., & Krämer, W. (2014). *Warum dick nicht doof macht und Genmais nicht tötet: Über Risiken und Nebenwirkungen der Unstatistik*. Frankfurt am Main: Campus-Verlag.

Binder, K., Krauss, S., & Wassner, C. (2018). Der Häufigkeitsdoppelbaum als didaktisch hilfreiches Werkzeug von der Unterstufe bis zum Abitur. *Stochastik in der Schule*, *38*(1), 2–11.

Binder, K., & Marienhagen, J. (2017). Bayes'sches Denken – Schritt für Schritt: Mit Häufigkeiten und Baumdiagrammen Einsichten in komplexe Probleme ermöglichen. In R. Vonthein, I. Burkholder, R. Muche, & G. Rauch (Eds.), *Zeig mir mehr Biostatistik! Mehr Ideen und neues Material für einen guten Biometrie-Unterricht* (pp. 87–99). Berlin, Heidelberg: Springer Berlin Heidelberg.

Bosbach, G., & Korff, J. J. (2017). *Die Zahlentrickser: Das Märchen von den aussterbenden Deutschen und andere Statistiklügen*. München: Heyne.

Brewer, N. T., Salz, T., & Lillie, S. E. (2007). Systematic review: The long-term effects of false-positive mammograms. *Annals of internal medicine*, *146*(7), 502–510.

Dieckmann, A., & Krauss, S. (2005). Wenn weniger Wissen mehr sein kann. *Zeitschrift für Erziehungswissenschaft*, *8*(2), 187–201. https://doi.org/10.1007/s11618-005-0133-2

Gigerenzer, G. (2013). *Risiko: Wie man die richtigen Entscheidungen trifft* (6. Aufl.). München: Bertelsmann.

Gigerenzer, G., & Gray, J. A. M. (Eds.). (2011). *Better doctors, better patients, better decisions: Envisioning health care 2020*. Cambridge, Mass.: MIT.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033295X.102.4.684

Gigerenzer, G., & Martignon, L. (2015). Risikokompetenz in der Schule lernen. *Lernen und Lernstörungen*, *4*(2), 91–98. https://doi.org/10.1024/2235-0977/a000098

Gigerenzer, G., Todd, P., & ABC research group (Eds.). (1999). *Simple heuristics that make us smart*: Oxford University Press.

Gigerenzer, G., & Wegwarth, O. (2013). Five year survival rates can mislead. *BMJ (Clinical research ed.)*, *346*, f548. https://doi.org/10.1136/bmj.f548

Gigerenzer, G., Wegwarth, O., & Feufel, M. (2010). Misleading communication of risk. *BMJ (Clinical research ed.)*, *341*, c4830. https://doi.org/10.1136/bmj.c4830

Habs, M., Binder, K., Krauss, S., Müller, K., Ernst, B., Valentini, L., & Koller, M. (2017). A Balanced Risk-Benefit Analysis to Determine Human Risks Associated with Pyrrolizidine

Alkaloids (PA) – The Case of Tea and Herbal Infusions. *Nutrients*, *9*(7). https://doi.org/10.3390/nu9070717

Hill, R. (2004). Multiple sudden infant deaths -- coincidence or beyond coincidence? *Paediatric and perinatal epidemiology*, *18*(5), 320–326. https://doi.org/10.1111/j.1365-3016.2004.00560.x

Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in psychology*, *6*, 642. https://doi.org/10.3389/fpsyg.2015.00642

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science*, *290*(5500), 2261–2262. https://doi.org/10.1126/science.290.5500.2261

Jorgensen, K. J., & Gotzsche, P. C. (2009). Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. *BMJ (Clinical research ed.)*, *339*, b2587. https://doi.org/10.1136/bmj.b2587

Krämer, W. (2008). *So lügt man mit Statistik* (10. Aufl., Vol. 3038). München: Piper.

Martignon, L., & Krauss, S. (2003). Can l'homme eclaire be fast and frugal? Reconciling Bayesianism and bounded rationality. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 108–122). Cambridge: Cambridge University Press.

Martignon, L., & Krauss, S. (2007). Gezinkte und ungezinkte Würfel, Magnetplättchen und Tinkercubes: Materialien für eine Grundschulstochastik zum Anfassen. *Stochastik in der Schule*, *27*(3), 16–27.

McDowell, M., & Jacobs, P. (2017). Meta-Analysis of the Effect of Natural Frequencies on Bayesian Reasoning. *Psychological bulletin*, *143*, 1273–1312. https://doi.org/10.1037/bul0000126

McDowell, M., Rebitschek, F. G., Gigerenzer, G., & Wegwarth, O. (2016). A Simple Tool for Communicating the Benefits and Harms of Health Interventions. *MDM Policy & Practice*, *1*(1), 238146831666536. https://doi.org/10.1177/2381468316665365

Salz, T., Richman, A. R., & Brewer, N. T. (2010). Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. *Psycho-oncology*, *19*(10), 1026–1034. https://doi.org/10.1002/pon.1676

Satake, E., & Murray, A. V. (2017). Teaching an Application of Bayes' Rule for Legal Decision-Making: Measuring the Strength of Evidence. *Journal of Statistics Education*, *22*(1). https://doi.org/10.1080/10691898.2014.11889692

Schneps, L., & Colmez, C. (2013). *Wahrscheinlich Mord: Mathematik im Zeugenstand*. München: Hanser.

Schwartz, L. M., Woloshin, S., & Welch, H. G. (2007). The drug facts box: Providing consumers with simple tabular data on drug benefit and harm. *Medical decision making: an international journal of the Society for Medical Decision Making*, *27*(5), 655–662. https://doi.org/10.1177/0272989X07306786

Schwartz, L. M., Woloshin, S., & Welch, H. G. (2009). Using a Drug Facts Box to Communicate Drug Benefits and Harms. *Annals of Internal Medicine*, *150*(8), 516. https://doi.org/10.7326/0003-4819-150-8-200904210-00106

Sedlmeier, P., & Köhlers, D. (2005). *Wahrscheinlichkeiten im Alltag: Statistik ohne Formeln; mit einem Trainingsprogramm auf CD-ROM* (1. Aufl., [Nachdr.]). Braunschweig: Westermann.

Siegrist, M., & Keller, C. (2011). Natural frequencies and Bayesian reasoning: The impact of formal education and problem context. *Journal of Risk Research*, *14*(9), 1039–1055. https://doi.org/10.1080/13669877.2011.571786

Staatsinstitut für Schulqualität und Bildungsforschung (ISB). (2018). LehrplanPlus Gymnasium Bayern Mathematik 10. Klasse. Retrieved from http://www.lehrplanplus.bayern.de/fachlehrplan/gymnasium/10/mathematik

Stine, G. J. (1996). *Acquired immune deficiency syndrome: Biological, medical, social, and legal issues*. Englewood Cliff, NJ: Prentice Hall.

Wegwarth, O. (2018). Brustkrebsfrüherkennung – Nutzen und Risiken richtig kommunizieren. *Der Gynäkologe*, *289*(11), 1414. https://doi.org/10.1007/s00129-018-4199-3

Wegwarth, O., & Gigerenzer, G. (2013). Overdiagnosis and overtreatment: Evaluation of what physicians tell their patients about screening harms. *JAMA internal medicine*, *173*(22), 2086–2087. https://doi.org/10.1001/jamainternmed.2013.10363

Wegwarth, O., Wagner, G. G., & Gigerenzer, G. (2017). Can facts trump unconditional trust? Evidence-based information halves the influence of physicians' non-evidence-based cancer screening recommendations. *PlosONE*, *12*(8), e0183024. https://doi.org/10.1371/journal.pone.0183024

Woike, J. K., Hoffrage, U., & Martignon, L. (2017). Integrating and Testing Natural Frequencies, Naïve Bayes, and Fast-and-Frugal Trees. *Decision.* Advance online publication. https://doi.org/10.1037/dec0000086

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, *98*(3), 287–308. https://doi.org/10.1016/j.cognition.2004.12.003

Anhang

Darlegung des eigenen Anteils

Alle drei Artikel sind unter der Autorenschaft weiterer Koautoren entstanden, denen ich an dieser Stelle nochmal ganz herzlich meinen Dank für die wertvolle und fruchtbare Zusammenarbeit aussprechen möchte. Welche Anteile der jeweiligen Artikel mir zugeschrieben werden und welche den Mitautoren, soll an dieser Stelle detailliert dargelegt werden.

Artikel 1 wurde von mir in Mitautorenschaft von Stefan Krauss und Georg Bruckmaier verfasst. Das Design der Studie war meine Idee und wurde in Zusammenarbeit mit Stefan Krauss und Georg Bruckmaier weiterentwickelt. Ich übernahm die Auswahl der Aufgaben für die Erhebungsinstrumente, die Rekrutierung der Klassen, die Organisation, Durchführung und Auswertung der Daten und verfasste den Artikel. Stefan Krauss und Georg Bruckmaier wirkten an der Weiterentwicklung des Artikels sowie der Endredaktion mit.

Artikel 2 ist unter der Mitautorenschaft von Stefan Krauss, Georg Bruckmaier und Jörg Marienhagen entstanden. Meine Studienidee wurde zusammen mit allen drei Koautoren weiterentwickelt, so dass schließlich die zwei vorliegenden Teilstudien durchgeführt werden sollten. Jörg Marienhagen und auch Georg Bruckmaier standen mir bei der Rekrutierung der Regensburger Medizinstudierenden unterstützend zur Seite und waren zum Teil auch bei den Erhebungen anwesend. Die Auswertung der Daten und der Erstentwurf des vorliegenden Artikels erfolgte von mir. Die Weiterentwicklung des Artikels wurde von Stefan Krauss unterstützt. Die Endredaktion erfolgte gemeinsam mit allen drei Mitautoren.

Artikel 3 wurde von mir in Mitautorenschaft von Stefan Krauss verfasst. Das Studiendesign war meine Idee und wurde in Zusammenarbeit mit Stefan Krauss weiterentwickelt. Die Organisation, Rekrutierung der Medizinstudierenden, Durchführung und Auswertung der Studie erfolgte durch mich, wie auch das Verfassen des vorliegenden Artikels. Stefan Krauss wirkte an der Weiterentwicklung des Artikels mit.

Alle Publikationen und Vorträge

PUBLIKATIONEN (* mit Peer Review)

* Weber, P., Binder, K., & Krauss, S. (2018). Why can only 24% solve Bayesian reasoning problems in natural frequencies? Frequency phobia in spite of probability blindness. *Frontiers in Psychology*, *9*(1833).

* Hilbert, S., Bruckmaier, G., Binder, K., Krauss, S., & Bühner, M. (2018). Prediction of elementary mathematics grades by cognitive abilities. *European Journal of Psychology of Education*, *33*(4), 1-19.

Binder, K., & Vogel, M. (2018). Prä-Bayes'sche Verhältnisse. *mathematik lehren*, *209*, 13-17.

Binder, K., Krauss, S, Bruckmaier, G., & Marienhagen, J. (2018). T(h)ree steps to improve Bayesian reasoning. In, *Proceedings of the 10th International Conference on Teachings Statistics (ICOTS-10)*. Kyoto, Japan.

Weber, P., Binder, K., & Krauss, S. (2018). Frequency phobia in spite of probability blindness. In, *Proceedings of the 10th International Conference on Teachings Statistics (ICOTS-10)*. Kyoto, Japan.

* Binder, K., Krauss, S., & Wassner, C. (2018). Der Häufigkeitsdoppelbaum als didaktisch hilfreiches Werkzeug von der Unterstufe bis zum Abitur. *Stochastik in der Schule*, *38*(1), 2-11.

* Habs, M., Binder, K., Krauss, S., Müller, K., Ernst, B., Valentini, L. & Koller, M. (2018). A Balanced Risk-Benefit Analysis to Determine Human Risks Associated with Pyrrolizidine Alkaloids (PA)—The Case of Herbal Medicinal Products Containing St. John's Wort Extracts (SJW). *Nutrients*, *10*(7).

* Binder, K., Krauss, S., Bruckmaier, G. & Marienhagen, J. (2018). Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making. *PlosONE*, *13*(3). Zweiter Artikel der Dissertation.

Binder, K., & Krauss, S. (2018). Bayesianische Aufgaben mit mehreren Testergebnissen – Wann sind Baumdiagramme in komplexeren medizinischen Entscheidungsfindungsprozessen hilfreich? *Beiträge zum Mathematikunterricht 2018*. Münster: WTM.

Weber, P., Binder, K., & Krauss, S. (2018). Natürliche Häufigkeiten zur Lösung Bayesianischer Aufgaben – Systematische Vermeidung statt effektiver Nutzung. *Beiträge zum Mathematikunterricht 2018*. Münster: WTM.

Binder, K., Krauss, S., Hilbert, S., & Blum, W. (2018). Diagnostische Kompetenz von Lehrkräften in COACTIV und deren Auswirkung auf Unterrichtsqualität und den Lernzuwachs von Schülerinnen und Schülern. *Beiträge zum Mathematikunterricht 2018*. Münster: WTM.

Binder, K., Krauss, S., Hilbert, S., Brunner, M., Anders, Y., & Kunter, M. (2018). Diagnostic Skills of Mathematics Teachers in the COACTIV study. In T. Leuders, K. Philipp & J. Leuders (Eds.) *Diagnostic Competence of Mathematics Teachers – Unpacking a complex construct in teacher education and teacher practice* (pp. 33-53). Cham: Springer.

Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (2017). Visualisierung des Bayesianischen 2-Test-Falls. In Institut für Mathematik der Universität Potsdam (Hrsg.), *Beiträge zum Mathematikunterricht 2017* (S. 99-102). Münster: WTM.

Weber, P., & Binder, K. (2017). Häufigkeitsphobie trotz Wahrscheinlichkeitsblindheit. In Institut für Mathematik der Universität Potsdam (Hrsg.), *Beiträge zum Mathematikunterricht 2017* (S. 1435-1436). Münster: WTM.

Binder, K., & Marienhagen, J. (2017). Bayes'sches Denken - Schritt für Schritt: Mit Häufigkeiten und Baumdiagrammen Einsichten in komplexe Probleme ermöglichen. In R. Vonthein, I. Burkholder, R. Muche & G. Rauch (Hrsg.), *Zeig mir mehr Biostatistik* (S. 87-99). Heidelberg: Springer.

* Habs, M., Binder, K., Krauss, S., Müller, K., Ernst, B., Valentini, L., & Koller, M. (2017). A balanced risk-benefit analysis to determine human risks associated with Pyrrolizidine Alkoloids (PA) – The case of tea and herbal infusions. *Nutrients*, *9*(717).

Binder, K., Krauss, S., & Bruckmaier, G. (2016). Visualisierung komplexer Bayesianischer Aufgaben. In Institut für Mathematik und Informatik Heidelberg (Hrsg.), *Beiträge zum Mathematikunterricht 2016* (S. 1469-1472). Münster: WTM.

Bruckmaier, G., Binder, K., & Krauss, S. (2016). Numerische Darstellungsarten statistischer Informationen. In E.-M. Plackner & N. von Schroeders (Hrsg.), *Daten und Zufall. MaMut – Materialien für den Mathematikunterricht*, *3* (S. 47-76). Hildesheim: Franzbecker.

* Binder, K., Krauss, S. & Bruckmaier, G. (2015). Effects of visualizing statistical information – An empirical study on tree diagrams and 2 x 2 tables. *Frontiers in Psychology*, *6*(1186). Erster Artikel der Dissertation.

Binder, K., Krauss, S., & Bruckmaier, G. (2015). Welche Visualisierung unterstützt Bayesianisches Denken? In F. Caluori, H. Linneweber-Lammerskitten & C. Streit (Hrsg.), *Beiträge zum Mathematikunterricht 2015* (S. 160-163). Münster: WTM.

Binder, K. (2014). Bayesianische Inferenz – Kognitionspsychologische Grundlagen und didaktische Interventionen. In J. Roth & J. Ames (Hrsg.), *Beiträge zum Mathematikunterricht 2014* (S. 193-196). Münster: WTM.

VORTRÄGE

2018

Krauss, S., Bruckmaier, G., Blum, W., & Binder, K. (October 2018). Aspects of modeling in the COACTIV-Study. Autumn School at the University of Würzburg, Germany.

Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (July 2018). T(h)ree steps to improve Bayesian reasoning. 10th International Conference on Teachings Statistics (ICOTS-10). Kyoto, Japan.

Weber, P., Binder, K., & Krauss, S. (July 2018). Frequency phobia in spite of probability blindness. 10th International Conference on Teachings Statistics (ICOTS-10). Kyoto, Japan.

Binder, K. (Juni 2018, eingeladen). Die Bedeutung medizinischer Testergebnisse verstehen – Effekte des Informationsformates und von Visualisierungen auf das Verständnis bedingter Wahrscheinlichkeiten. Wissenschaftliches Kolloquium des IDMI, Münster.

Schulte, M., Hanisch, A., & Binder, K. (März 2018). "Minus ist, wenn man viel hatte und jetzt weniger hat" – Sprachsensibler Mathematikunterricht als Herausforderung für die Primarstufe. 5. Thementag Theorie-Praxis 2018 - Inklusion, Mehrsprachigkeit, Sprache im Fach, Regensburg.

Binder, K., & Krauss, S. (März 2018). Bayesianische Aufgaben mit mehreren Testergebnissen – Wann sind Baumdiagramme in komplexeren medizinischen Entscheidungsfindungsprozessen hilfreich? 52. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM), Paderborn.

Binder, K., Krauss, S., Hilbert, S., & Blum, W. (März 2018). Diagnostische Kompetenz von Lehrkräften in COACTIV und deren Auswirkung auf Unterrichtsqualität und den Lernzuwachs von Schülerinnen und Schülern. 52. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM), Paderborn.

Weber, P., Binder, K., & Krauss, S. (März 2018). Natürliche Häufigkeiten zur Lösung Bayesianischer Aufgaben – Systematische Vermeidung statt effektiver Nutzung. 52. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM), Paderborn.

Binder, K., Krauss, S., & Hilbert, S. (Februar 2018). Facetten diagnostischer Fähigkeiten in der COACTIV-Studie: Neue Analysen zu Zusammenhängen und zur prädiktiven Validität für Schülerleistungen. 6. Tagung der Gesellschaft für empirische Bildungsforschung (GEBF), Basel.

Rank, A., Deml, I., Wildemann, A., Schilcher, A., Krauss, S., Lenske, G., Merkert, A., Schulte, M., Binder, K., & Bien-Miller, L. (Februar 2018). Evaluation im Primarbereich – Erfahrungen mit Sprachförderung im Projekt „Eva-Prim“. 6. Tagung der Gesellschaft für empirische Bildungsforschung (GEBF), Basel.

Deml, I., Rank, A., Wildemann, A., Schilcher, A., Krauss, S., Lenske, G., Merkert, A., Schulte, M., Binder, K., & Bien-Miller, L. (Februar 2018). Sprachförderung professionalisieren – zum Zusammenhang von fachdidaktischem Wissen und Schülerleistung. 6. Tagung der Gesellschaft für empirische Bildungsforschung (GEBF), Basel.

2016-2017

Binder, K., Krauss, S., & Hilbert, S. (Dezember 2017). Diagnosefähigkeiten von Mathematiklehrkräften in COACTIV. cosima-Gruppenkolloquium an der LMU München, München.

Binder, K., Marienhagen, J., & Krauss, S. (November 2017). Medizinische Diagnosen mit Häufigkeitsdoppelbäumen verstehen. AG Lehre und Didaktik der Biometrie, Hannover.

Krauss, S., Binder, K., & Hilbert, S. (Oktober 2017). Diagnostische Fähigkeiten in der COACTIV-Studie – Vertiefte Analysen. Herbsttagung des Arbeitskreises für empirische Bildungsforschung in der Gesellschaft für Didaktik der Mathematik (GDM), Hannover.

Binder, K., Roth, C., Hartmann, L., & Kesdoğan, D. (Oktober 2017). Digitale Themen digital unterrichten - Ein VHB-Kurs zum Thema "IT-Sicherheit" (Poster). Tag der digitalen Lehre 2017, G.R.I.P.S.-Team des Rechenzentrums und Team des Verbundprojekts der bayerischen Universitäten ProfiLehrePlus (PLP), Regensburg.

Binder, K., Marienhagen, J., Krauss, S., & Bruckmaier, G. (September 2017). Bayesianisches diagnostisches Denken mit gezielten Visualisierungen unterstützen. Gemeinsame Jahrestagung der Gesellschaft für medizinische Ausbildung (GMA) und des Arbeitskreises für die Weiterentwicklung der Lehre in der Zahnmedizin (AKWLZ), Münster.

Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (März 2017). Visualisierung des Bayesianischen 2-Test-Falls. 51. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM), Potsdam.

Weber, P., & Binder, K. (März 2017). Häufigkeitsphobie trotz Wahrscheinlichkeitsblindheit (Poster). 51. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM), Potsdam.

Binder, K., Krauss, S., & Moßburger, M. (Februar 2017). Häufigkeitsdoppelbäume – Von der Unterstufe bis zum Abitur. Lehrerfortbildung "Daten und Zufall: Einführung in den LehrplanPlus und neue didaktische Ansätze" des Bayerischen Philologenverbandes, des Gymnasiums Herzogenaurachs, des Hans-Leinberger-Gymnasiums Landshut und der Mathematikdidaktik der Universität Regensburg, Regensburg.

Krauss, S., & Binder, K. (Februar 2017). Bäume und "Doppelbäume" mit absoluten Häufigkeiten. Lehrerfortbildung "Daten und Zufall: Einführung in den LehrplanPlus und neue didaktische Ansätze" des Bayerischen Philologenverbandes, des Gymnasiums Herzogenaurachs, des Hans-Leinberger-Gymnasiums Landshut und der Mathematikdidaktik der Universität Regensburg, Regensburg.

Binder, K. (Januar 2017). Der t-Test. Workshop für Schülerinnen und Schüler des Goethe-Gymnasiums, Regensburg.

Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (September 2016). T(h)ree steps to improve Bayesian reasoning. Talk am Max-Planck-Institut für Bildungsforschung, Berlin.

Binder, K., Krauss, S., & Bruckmaier, G. (July, 2016). Visualization of complex Bayesian tasks. 13th International Congress on Mathematical Education (ICME), Hamburg, Germany. Poster presentation.

Binder, K., Krauss, S., & Bruckmaier, G. (März 2016). Visualisierung komplexer Bayesianischer Aufgaben. 50. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM), Heidelberg.

2014-2015

Bruckmaier, G., & Binder, K. (Oktober 2015). Zahlen in der Statistik - Neues vom Minenfeld. Lehrerfortbildung der Mathematikdidaktik der Universität Regensburg, Regensburg.

Binder, K. (September 2015). Vierfeldertafeln und Baumdiagramme im Stochastikunterricht – Eine Studie zu Bayesianischen Aufgaben. Doktorandenkolloquium der Gesellschaft für Didaktik der Mathematik (GDM), Würzburg.

Bruckmaier, G., Binder, K., & Krauss, S. (Juli 2015). Zum Einfluss von Visualisierungen und Informationsformaten bei Bayesianischen Aufgaben. Forschungskolloquium des IMBF der Pädagogischen Hochschule, Freiburg.

Binder, K., Krauss, S., & Bruckmaier, G. (Juni 2015). Visual representation of natural frequencies improves Bayesian reasoning. 9th Annual International Conference on Mathematics & Statistics: Education & Applications (ATINER), Athens.

Binder, K., Krauss, S., & Bruckmaier, G. (Februar 2015). Welche Visualisierung unterstützt Bayesianisches Denken? 49. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM), Basel.

Bruckmaier, G., Binder, K., & Krauss, S. (Dezember 2014). Warum sich Wahrscheinlichkeiten und Prozentangaben oft unserer Intuition widersetzen - und was man dagegen tun kann. Lehrerfortbildung MaMut (Materialien für den Mathematikunterricht) der Friedrich-Alexander-Universität Erlangen-Nürnberg, Nürnberg.

Binder, K. (März 2014). Bayesianische Inferenz – Kognitionspsychologische Grundlagen und didaktische Interventionen. 48. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM), Koblenz