UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

UNIVERSITE PARIS I PANTHÉON SORBONNE
UFR de SCIENCES ECONOMIQUES
Centre d'Economie de la Sorbonne

THÈSE
Pour l'obtention du titre de Docteur en Sciences Economiques

Soutenue publiquement par

Marine Hainguerlot

Le 21 décembre 2017

Probability distortion in clinical judgment: Field study and laboratory experiments

Sous la direction de:
Vincent de Gardelle, Chargé de Recherche CNRS-CES-PSE
Jean- Christophe Vergnaud, Directeur de Recherche CNRS- CES

Jury:
Ulrich Hoffrage, Professeur à l'Université de Lausanne (Rapporteur)
Olga Kostopoulou, Maître de Conférences à l'Imperial College of London
Pascal Mamassian, Directeur de Recherche CNRS- LSP (Rapporteur)
Jean- Marc Tallon, Directeur de Recherche CNRS- PSE
Ann Van den Bruel, Professeur Associé à l'Université d'Oxford

L’Université Paris I Panthéon-Sorbonne n’entend donner aucune approbation, ni improbation aux opinions émises dans cette thèse; elles doivent être considérées comme propres à leur auteur.

Remerciements

Je tiens d'abord à remercier particulièrement mes deux directeurs de thèse Jean- Christophe Vergnaud et Vincent de Gardelle pour leur soutien inconditionnel durant toute la thèse. Ils ont cru en moi et m'ont donné la chance de construire ce travail durant ces quatre années. Je me considère très chanceuse d'avoir pu réaliser cette thèse à leurs côtés. Je les remercie pour leur encadrement, leur disponibilité, et nos discussions passionnées à trois. J'espère avoir la chance de pouvoir poursuivre avec eux des travaux de recherche.

Ma gratitude va également aux membres de mon jury, Ulrich Hoffrage, Olga Kostopoulou, Pascal Mamassian, Jean- Marc Tallon et Ann Van den Bruel, qui ont accepté d'en faire partie et qui se sont déplacés, parfois de loin. Leurs remarques et critiques m'ont été d'une aide précieuse et ont grandement contribué à l'amélioration de mon travail.

Je remercie les chercheurs avec qui j'ai eu le plaisir de travailler Thibault Gajdos et Vincent Gajdos pour leur disponibilité et précieux conseils. Je remercie Thibault Gajdos de m'avoir intégré dans le groupe de travail sur la menace du stéréotype. Je remercie Vincent Gajdos de m'avoir donné l'opportunité de travailler sur les données de médecine. Je remercie aussi les chercheurs qui m'ont permis de faire mes premiers pas dans la recherche : Nicolas Soulié, Matthieu Manant, Hélène Hubert et Louis Lévy Garboua.

Merci aux personnes de la Maison des Sciences Economiques (MSE) qui m'ont permis de réaliser ma thèse dans les meilleures conditions : Joël, Loïc Sorel, Jean- Christophe Fouillé, Francesca Di Legge, Nathalie Louni et Leïla Sidali. Je remercie Maxim Frolov pour

l'organisation des expériences au LEEP. Je remercie Patricia Vornetti, Fabrice Le Lec et Fabrice Rossi qui m'ont confié leurs Tds pendant mes années d'ATER.

Je tiens à remercier les doctorants avec qui les journées à la MSE sont toujours très heureuses: Antoine Malézieux, Feryel Bentayeb, Irène Iodice, Sandra Daudignon, Pierre Aldama, Niels Boissonnet, Antoine Hémon, Rémi Yin, Justine Jouxtel, Matthieu Cassou, Antoine Prévet, Quentin Couanau, Hector Moreno, Clément Goulet, Mathilde Poulain et Matthieu Plonquet. Un grand merci à Justine et Rémi pour avoir relu l'introduction générale. Je remercie également les anciens doctorants pour leurs précieux conseils : Noémi Berlin, Marco Gazel, Chaimaa Yassine, Anna Bernard et Sébastien Massoni.

Je remercie Amalia pour sa disponibilité, son soutien et sa relecture des chapitres. Spéciale dédicace aux Cotos: Tom, Ophélie, Karl, Popo, Matoche, Quent, Tib, Bobo, Marion, Pierre, Pelletier, Nissou et Dany pour leur amitié. Merci à tous mes amis et proches : Amandine, Christophe, Jen, Amau, Shriman, Angie, Amma, Cristina, Marilou, Mathieu, Barbara, Cyrielle, Amir, Yanir, Wieder ,Jorge, Jacky, Catherine et bien d'autres. Un grand merci à la famille de Jean- Christophe, Val, Loïse, Anita, Arsène, Arnold, Denise et Daniel de m'avoir accueillie de nombreuses fois pour travailler sur la thèse chez eux dans le 19ème et cet été dans le Limousin. Je remercie également le conservatoire de Maisons- Alfort, en particulier Nicolas, Anton, Marie- Laure, Dominique et l'orchestre, pour les fins de semaine musicales.

Un grand merci à ma famille pour leur soutien, mon père Thierry, mes grands- parents, Françoise et Germain, ma sœur Charlène, mon frère Simon et Aurélie. Mes derniers remerciements vont à Avshalom pour sa gentillesse et son soutien.

A ma famille.

Notice

Except the general introduction and conclusion, all chapters of this thesis are self-containing research articles. Consequently, each chapter contains its corresponding literature. The bibliography only contains the corresponding literature of the general introduction and conclusion.

Table of contents

General Introduction

With the advanced development of machine learning technologies, one timely question is: should we replace human judgment by predictions from the machines to improve decision-making? This question is also pertinent for medical decision-making (Chen & Asch, 2017; Donnelly, 2017; Obermeyer & Emanuel, 2016). The purpose of this thesis is to investigate whether replacing physician judgment by statistical models can improve actual medical decision accuracy.

To compare the physician with the model, we will use probability distortion as a central measure. Probability distortion corresponds to the difference between the subjective probabilities of the man and the objective probabilities of the model. This thesis shall go beyond the question of measuring how the physician performs compared to the model. This thesis also studies the cognitive processes that may drive physicians' deviations from the model. Understanding the reasons behind probability distortion may help better address the "man versus model" debate in medical decision-making.

In the general introduction, we develop a theoretical framework of how physicians may process information. This framework allows us to derive the cognitive mechanisms through which the physician may depart from the model, and, as a result, how his subjective probabilities may deviate from the objective probabilities. This framework will be used as a theoretical ground to motivate the questions addressed in the thesis. First, we summarize the main findings from the literature comparing physician judgment with statistical models.

Related literature on physician versus statistical model

Prediction and Probability judgment

Comparisons of predictions from statistical models with physician predictions date back to the fifties with the pioneering work of Meehl (1954). It is typically acknowledged that simple linear algorithms can better diagnose conditions or predict outcomes than physicians (Dawes et al, 1989; Grove et al, 2000; Ægisdóttir et al, 2006).

However, comparisons between the probability estimates from physicians and the ones generated by statistical models have yielded mixed results on the superiority of models (for a review, see O'Hagan et al, 2006). On one hand, statistical models are found to be better calibrated: the probability estimates generated differ less from the actual frequencies. On the other hand, physicians' probability estimates seem to better discriminate between patients who do or do not have the condition. (e.g.: McClish & Powell, 1989).

The distinct strengths of statistical models and physicians are at the core of the "man versus model" debate: how much information is available and how is the information processed? Physicians may have more information but statistical models better process the information.

Statistical models' strength: Analytical judgment

Statistical models produce an analytical judgment derived from a set of assumptions specifying how to integrate the evidence. Models are known to be proficient at integrating the evidence consistently (Karelaia & Hogarth, 2008). Presented with the same set of medical evidence, statistical models may have better calibration than physicians because they optimally weight the evidence, they are not biased.

Physicians’ strength: Intuitive judgment

Physician judgment involves not only analytical processes but also intuitive processes (Greenhalgh, 2002; Stolper et al, 2011; Wooley & Kostopoulou, 2013) and their intuition might be valid information (Van den Bruel, 2012). Statistical models can only generate probability estimates based on the medical evidence that they are taught to use whereas physicians may have access to more information, in particular their intuitive judgment, to better discriminate the presence or absence of the disease.

Combination: statistical model + intuitive physician

As a result of their complementarity, it has been recommended to capitalize on their respective strengths by combining the statistical model with the physician’s intuitive judgment to improve the accuracy of probability judgment (Blattberg & Hoch, 1990; Yaniv & Hogarth, 1993; Whitecotton et al, 1998).

A theoretical framework of physician judgment

To address the debate “man versus model” in the thesis, we propose a theoretical framework that describes how physicians may process the information. Within this framework, we distinguish between the analytical and intuitive processes that may be involved in the physician judgment. We propose a Bayesian formalization that defines how physicians may form their clinical judgment by integrating their analytical and intuitive processes. This framework will be used as a theoretical ground to motivate the questions addressed in the thesis.

A model of physician information processing

During medical encounters, physicians process large amounts of information from their environment to make a diagnosis that will eventually guide their decision. Here, we adapt an information-processing model from Wickens et al. (2015) to the medical encounter situation. Below, we describe a simple four-stage model of physician information processing (see also **Figure 1**).

(1) *Inputs:* The physician collects information about the patient (diagnosis cues) in two ways. Some cues (x_i) are explicit in that they take a particular value. For example, the physician observes that the temperature of the patient is 37°C. The physician may also receive an internal signal (x) which corresponds to an impression about the presence or absence of the disease and that cannot be articulated easily to the diagnosis cues. For example, the physician may have the feeling that something is wrong with the patient without being able to explicit why.

(2) *Central processing*: As aided by her prior knowledge and experience, she retrieves from her long-term memory a meaningful interpretation about the association between information collected in the samples and the occurrence of the disease. For example, she considers that observing a temperature of 37°C is usually reassuring. On the other hand, she remembers that her feeling that something is wrong with the patient has already been a red flag with other patients. Overall, she holds in working memory two informative components: an analytical component (i.e. the evaluation of the explicit cues) and an intuitive component (i.e. the evaluation of her internal signal). She then integrates the analytical and intuitive components together to judge whether or not the patient has the disease.

(3) *Decision:* The physician decides to treat or not the patient as a function of her judgment about whether or not the patient may have the disease and the consequences associated to her decision.

(4) *Learning*: Finally, after observing the outcome of her decision the physician may update the meaningful interpretation she attributes to the information collected. She then stores the updated valuation into her long- term memory.

Note that these stages depend on several cognitive abilities: working memory, long-term memory, attention resources and effort. Below, we shortly describe their potential roles in the information processing model.

Working memory: Working memory is essential for holding and manipulating information in the short- term: “It is the temporary store that keeps information active while we are using it or until we use it.” (Wickens et al, 2004). Working memory is required to keep the analytical and intuitive components active to integrate them together. It is also necessary to update the value of the diagnosis cues on the basis of the information arriving from the outcome (i.e. the feedback).

Long- term memory: Long-term memory is responsible for storing and retrieving information about the value of the diagnosis cues in the long-term. It corresponds to the process of learning.

Attention resources and effort: In particular, selective attention is necessary to select which diagnosis cues to process (the ones with the highest perceived informative value) and which diagnosis cues to filter out. The ability to do several tasks at one time, to allocate the attention resources and effort to different tasks is also required.

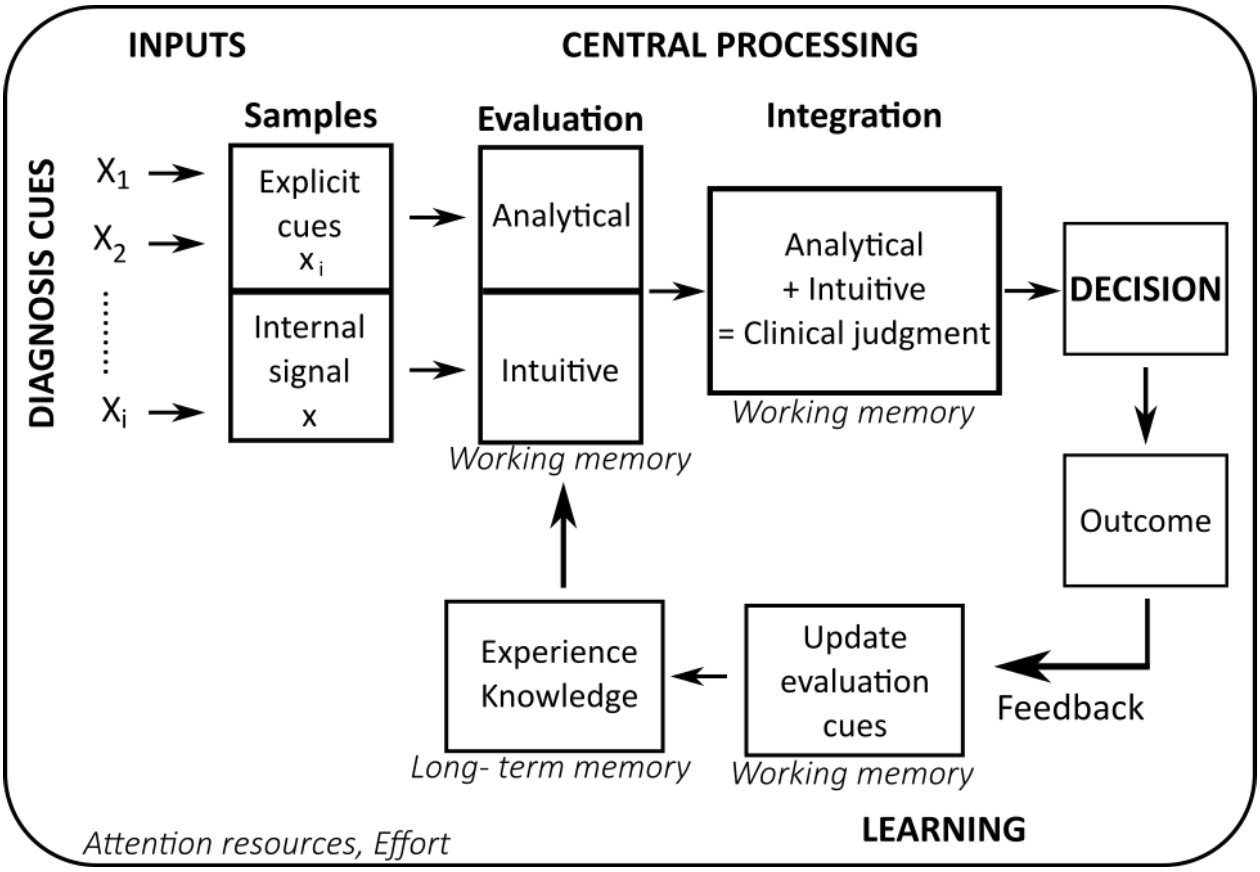


Figure 1: *A simple physician information- processing model*

Bayesian model of central information processing

We model how a physician forms her clinical judgment about whether or not the patient has a disease: $Y = 1$ or $Y = 0$. First, we define the evaluation of the explicit cues and the internal signal. Second, we model how the analytical and the intuitive components are integrated into the clinical judgment. The model is developed in log odds. For the sake of simplicity, we assume that the explicit cues x_i and the internal signal x are independent.

Samples of diagnosis cues X_i

The physician observes the values taken by the explicit cues (x_i) and the internal signal (x). For the sake of simplicity, we consider that the explicit cues can take 2 values: 1 or 0.

Knowledge about x_i and x

We consider that the physician holds in her long-term memory knowledge about the sensitivity $P^s(x_i = 1|Y = 1)$[1] and the specificity $P^s(x_i = 0|Y = 0)$ of the explicit cues x_i. Similarly, the physician stores in long-term memory the probability distribution of the internal signal conditional on the presence $P^s(x|Y = 1)$ or absence of the disease $P^s(x|Y = 0)$.

Evaluation of the explicit cues

According to this knowledge, she evaluates each cue x_i by calculating the weight of evidence (log odds) as follows: $W^s_{x_i=a} = \ln \frac{P^s(x_i=a|Y=1)}{P^s(x_i=a|Y=0)}$ where $a = 0{,}1$. The analytical component corresponds to the sum of the weight of evidence of x_i : $\sum_{i=1}^{k} W^s_{x_i}$.

Evaluation of the internal signal

[1] Where P^s corresponds to the subjective probability. We note P^o the objective probability.

We postulate that the physician also evaluates her internal signal by the weight of evidence as follows:

$$W_{x=a}^{s} = ln\left(\frac{P^{s}(x=a|Y=1)}{P^{s}(x=a|Y=0)}\right), \text{where } a \in \mathbb{R}$$

W_{x}^{s} corresponds to the intuitive component.

Integration of analytical and intuitive components

Finally, she forms her subjective probability by revising her prior belief on the disease $P_{prior}^{s}(Y = 1)$ through the following integration equation:

$$ln\frac{P^{s}(Y=1|X,x)}{P^{s}(Y=0|X,x)} = ln\frac{P_{prior}^{s}(Y=1)}{P_{prior}^{s}(Y=0)} + \alpha \sum_{i=1}^{k} W_{x_i}^{s} + \beta W_{x}^{s} \qquad \text{(eq1)}$$

Where the parameters α and β capture how the physician may distort the analytical and intuitive components respectively.

Physician versus ideal clinical judgment

Overall, the physician may suffer from several biases in the way she processes and integrates the information. First, she may misevaluate both the analytical and the intuitive components with respect to their objective values $W_{x_i}^{o}$ and W_{x}^{o}. Second, she may inaccurately integrate the analytical and/ or intuitive component(s).

Terminology

In the remaining of the thesis, we use the following terms:

- "Statistical model"[2]: judgment generated by a set of assumptions specifying how to best integrate the evidence.
- "Analytical man": part of the physician's judgment explained by an analytical integration of the evidence
- "Intuitive man": part of the physician's judgment explained by intuition
- Physician judgment: physician subjective probability estimate that a patient has the disease.
- Physician decision: physician decision to treat or not to treat.

Questions addressed in the thesis

Within this framework, the thesis addresses the following three questions:

- Does a biased analytical man make poorer decisions?
- Does combining the statistical model with the intuitive man improve decision?
- What are the factors that affect human information processing?

[2] Please note in chapter 3, we used the terminology "mechanical model" instead of "statistical model"

Does a biased analytical man make poorer decisions?

As described previously, physician judgment involves not only an analytical component but also an intuitive component. The extent to which physicians use analytical and intuitive processes may vary. Thus, even if it is well documented that physicians are biased in their analytical integration of the evidence, it may not have an impact on the quality of their decision. For example, physicians could rely on their intuition alone to make a decision. Also, their intuitive component could offset their biased analytical component.

In chapter 1, our goal is to evaluate, using medical data from the field, whether a biased analytical physician makes poorer decisions. Our set of medical data contains for each patient: information regarding the presence or absence of the disease, the available medical evidence, the physician's probability judgment that the patient has the disease and her treatment decision.

Within our theoretical framework, physician's judgment is composed of two components: an analytical part and an intuitive part. We need to separate these two components. Operationally, we propose to separate the physician's judgment into two components: the linear judgment and the residual judgment. We define the analytical component as a linear judgment that contains the part of the physician's judgment that is explained by a linear integration of the medical evidence. The intuitive component corresponds to the residual judgment which captures the part of the judgment that is not explained by a linear integration. It may capture physicians' intuition but also physicians' ability to integrate the evidence in a nonlinear way. To assess whether the physician is biased in his linear judgment, we compare the physician probability predicted by the linear judgment to the disease probability predicted by the linear model. We quantify bias in the analytical part as the distortion between the physician probability predicted by the linear judgment and the disease probability predicted by the linear model. Finally, we test whether probability distortion impairs the accuracy of medical decision.

Method

Figure 2 illustrates our method.

Lens Model approach

How good is the physician at integrating the available medical evidence compared to a statistical model? To answer this question, we use the Lens model approach (Brunswick, 1952; Goldberg, 1970). We consider that the physician judgment and the presence or absence of the disease being predicted can be modeled as two separate linear functions of cues available in the environment (Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975). The presence or absence of the disease is modeled as a linear function of a set of cues $X_i, i = 1,..,k$, as follows:

$$ln\frac{P^o(Y=1|X)}{P^o(Y=0|X)} = \beta_o^o + \sum_{i=1}^{k} \beta_i^o\, x_i \qquad \text{(eq2a)}$$

Similarly, physician probability judgment that a patient has the disease is modeled as a linear function of a set of cues $X_i, i = 1,..,k$, as follows:

$$ln\frac{P^s(Y=1)}{P^s(Y=0)} = \beta_o^s + \sum_{i=1}^{k} \beta_i^s\, x_i + \mu \qquad \text{(eq2b)}$$

where the error term μ corresponds to the “residual judgment”.

Interpretation of β_i^o versus β_i^s

The coefficient β_i corresponds to the log odds ratio for a given cue x_i.

Odds define the likelihood that the disease will occur:

$$Odds = \frac{P(Y = 1)}{P(Y = 0)}$$

The odds ratio corresponds to the odds that the disease will occur given that $x_i = 1$ compared to the odds that the disease will occur when $x_i = 0$. The odds ratio measures the association between the presence or absence of the disease and the cue x_i:

$$Odds\ Ratio = \frac{\frac{P(Y = 1|x_i = 1)}{P(Y = 0|x_i = 1)}}{\frac{P(Y = 1|x_i = 0)}{P(Y = 0|x_i = 0)}}$$

We can observe that the way the physician weights the cues in terms of log odds ratio β_i^s can differ in two distinct ways compared to β_i^o. First, she can over or under weight relevant cues (*over or under weighting bias*): $\frac{\boldsymbol{\beta_i^s}}{\boldsymbol{\beta_i^o}} < 1$ or $\frac{\boldsymbol{\beta_i^s}}{\boldsymbol{\beta_i^o}} > \mathbf{1}$. Second, she can weight irrelevant cues (*mis-weighting bias*): $\beta_i^s \neq \beta_i^o = 0$.

Probability distortion analysis: Linear judgment versus linear model

By applying the corresponding regression weights to the cues, we can estimate in log odds ($Lo(p) = \frac{p}{1-p}$) the physician probability predicted by the linear judgment ($Lo(\widehat{P^s})$) and the disease probability predicted by the linear model ($Lo(\widehat{P^o})$). To compare the linear judgment with the linear model, we plot $Lo(\widehat{P^s})$ versus $Lo(\widehat{P^o})$.

We use the general linear model to estimate a slope parameter of the distortion when probabilities are transformed in log odds, as follows: $Lo\left(\widehat{P^s}\right) = \beta_0 + \beta_1 Lo(\widehat{P^o})$[3] (Zhang & Maloney, 2012).

When β_1=1, there is no distortion in the slope. When $\beta_1 < 1$, physicians over-estimate small probabilities and under-estimate large probabilities. When $\beta_1 > 1$, physicians under-estimate small probabilities and over-estimate large probabilities.

Does a bias in probability distortion impair the accuracy of medical decision?

Empirically, we separate the dataset into 2 groups of patients based on the physician probability judgment: a high bias and a low bias group such that in the high bias group the distortion is greater compared to the low bias group. We investigate whether the accuracy of medical decision is impaired in the high bias group compared to the low bias group. To evaluate the accuracy of medical decision, we consider two measures: the sensitivity (i.e. the proportion of patients with the disease who receive treatment) and the specificity (i.e. the proportion of patients without the disease who do not receive treatment).

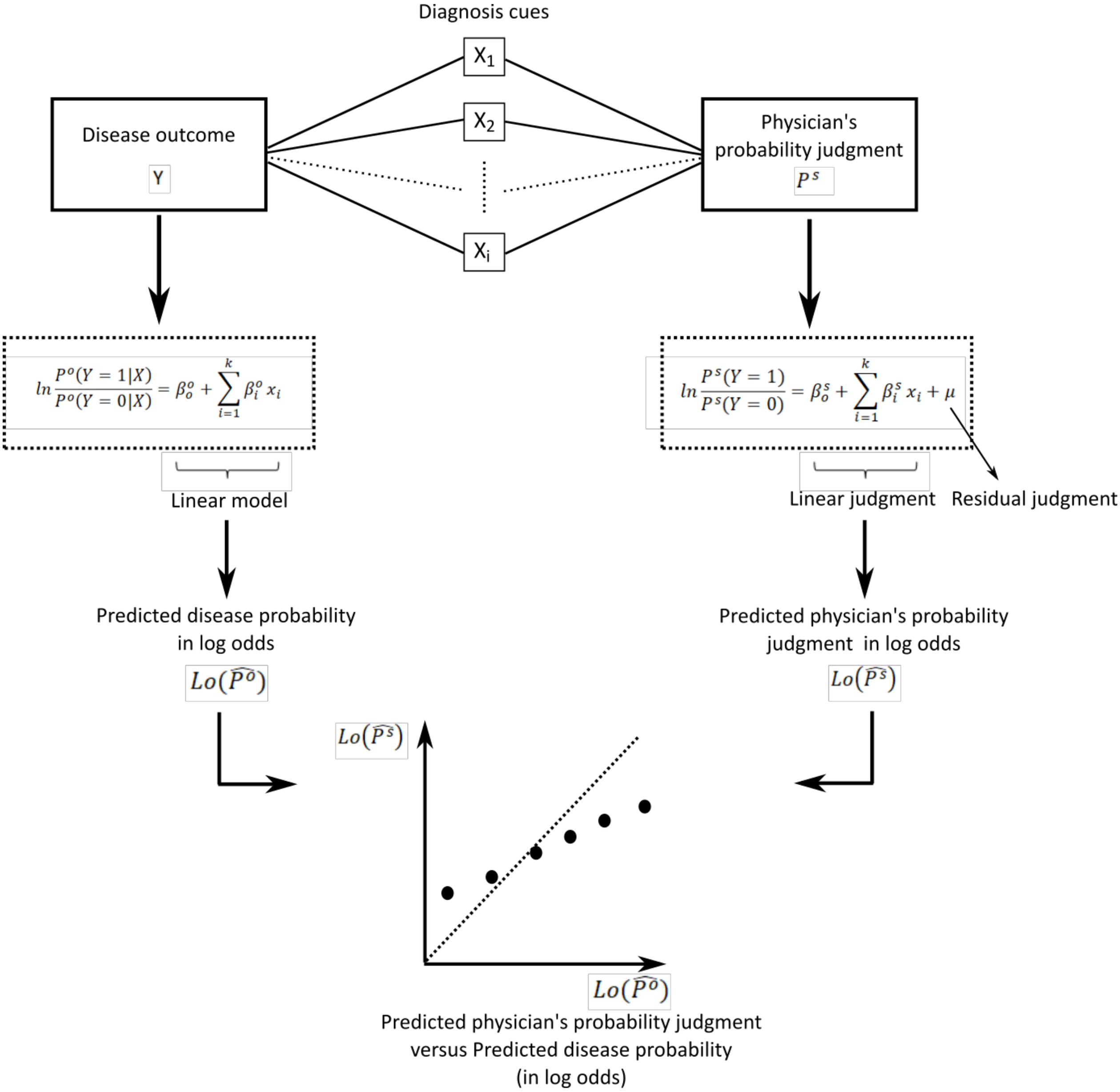
Data

We applied our method to the detection of bacterial infection in febrile infants younger than 3 months (N=1848) and assessed its impact on two dimensions of heath care (antibiotic treatment and hospital admission). Our data come from a prospective cohort study of the procalcitonin biomarker in the detection of bacterial infection in febrile infants younger than 3 months (Milcent et al., 2016). Physicians were asked to record the information they

[3] Please note that the coefficients β_0 and β_1 in this equation are different from the coefficients β_i^o and β_i^s in the Lens model approach.

collected about patients from their admission to their discharge. They recorded the demographic and neonatal data, medical history, physical examination, and clinical findings from ordered laboratory tests. Physicians were required to report their probability estimate that the infant had a bacterial infection, on a scale from 0 to 100% at two stages of the data collection: (i) at the end of the physical examination (pretest probability); and (ii) after receiving the clinical findings from the laboratory tests (posttest probability). Following the second estimate, they reported their decisions to hospitalize and to treat with antibiotics.

-



Diagnosis cues
X_1
X_2
X_i
Disease outcome
Y
Physician's probability judgment
P^s
$ln\frac{P^o(Y=1|X)}{P^o(Y=0|X)} = \beta_o^o + \sum_{i=1}^{k} \beta_i^o x_i$
Linear model
$ln\frac{P^s(Y=1)}{P^s(Y=0)} = \beta_o^s + \sum_{i=1}^{k} \beta_i^s x_i + \mu$
Linear judgment
Residual judgment
Predicted disease probability in log odds
$Lo(\widehat{P^o})$
Predicted physician's probability judgment in log odds
$Lo(\widehat{P^s})$
$Lo(\widehat{P^s})$
$Lo(\widehat{P^o})$
Predicted physician's probability judgment versus Predicted disease probability (in log odds)

Figure 2: *Diagram of the Lens model approach applied to judgment*

Remark about the relationship between the estimates β_i from the Lens model approach and the theoretical biases in evaluation and integration from the Bayesian framework

Note that the comparison between the estimates from the Lens model approach i.e. β_i^s (in equation 2b) versus β_i^o (in equation 2a) allows us to study whether physicians weight the evidence in a different way compared to the objective weights. However, and as it is described below, we cannot disentangle whether the difference in weights is related to a bias in integration and/ or a bias in evaluation of the evidence as formalized with the Bayesian framework (in equation 1).

β_i^s and β_i^o

In equation 2b, β_i^s corresponds to the difference *ceteris paribus* in the subjective probability (in log odds) between patients with $x_i = 1$ and patients with $x_i = 0$.

In equation 1, this difference is equal to $\alpha\left(W^S_{x_i=1} - W^S_{x_{i=0}}\right)$.

Thus we have: $\beta_i^s = \alpha\left(W^S_{x_i=1} - W^S_{x_{i=0}}\right)$.

In equation 2a, β_i^o is equal to $ln\left(\frac{\frac{P^O(Y=1|x_i=1)}{P^O(Y=0|x_i=1)}}{\frac{P^O(Y=1|x_i=0)}{P^O(Y=0|x_i=0)}}\right)$.

Note that by Bayes rule: $ln\left(\frac{\frac{P^O(Y=1|x_i=1)}{P^O(Y=0|x_i=1)}}{\frac{P^O(Y=1|x_i=0)}{P^O(Y=0|x_i=0)}}\right) = ln\left(\frac{\frac{P^O(x_i=1|Y=1)}{P^O(x_i=1|Y=0)}}{\frac{P^O(x_i=0|Y=1)}{P^O(x_i=0|Y=0)}}\right)$

where $ln\left(\frac{\frac{P^O(x_i=1|Y=1)}{P^O(x_i=1|Y=0)}}{\frac{P^O(x_i=0|Y=1)}{P^O(x_i=0|Y=0)}}\right) = ln\frac{sensitivity/(1-specificity)}{(1-sensitivity)/specificity} = W^O_{x_i=1} - W^O_{x_{i=0}}$.

The ratio $\frac{sensitivity/(1-specificity)}{(1-sensitivity)/specificity}$ is sometimes called the diagnostic odds ratio (see Glas et al, 2003).

Over/ under weighting bias ($\boldsymbol{\frac{\beta_i^s}{\beta_i^o}}$) *and mis-weighting bias* ($\beta_i^s \neq \beta_i^o = 0$)

The *over or under weighting bias* $\boldsymbol{\frac{\beta_i^s}{\beta_i^o}}$ corresponds theoretically to $\frac{\alpha\left(W^S_{x_i=1} - W^S_{x_{i=0}}\right)}{W^O_{x_i=1} - W^O_{x_{i=0}}}$. It can be the consequence of two biases in the analytical judgment:

- a bias in the knowledge of the diagnostic odds ratio: $W^S_{x_i=1} - W^S_{x_{i=0}} \neq W\ ^O_{x_i=1} - W^O_{x_{i=0}}$
- a bias in the integration of the evidence: $\alpha \neq 1$

The *mis-weighting bias* $\beta_i^s \neq \beta_i^o = 0$ corresponds theoretically to $\alpha\left(W_{x_i=1}^s - W_{x_{i=0}}^s\right) \neq W_{x_i=1}^o - W_{x_{i=0}}^o = 0$. It can be the consequence of a bias in the knowledge of the diagnostic odds ratio: $W_{x_i=1}^s - W_{x_{i=0}}^s \neq 0$.

Remark about the relationship between the estimates $\boldsymbol{\beta}_\mathbf{i}$ from the Lens model approach and the shape of the probability distortion

Note that the shape of the probability distortion in log odds between predicted physician's probability judgment and the predicted disease probability depends on the value of the corresponding estimates from the Lens model approach β_i^s (in equation 2b) and β_i^o (in equation 2a) . Chapter 2 discusses the shape of the distortion as a function of the estimates β_i .

Combining the statistical model with the intuitive man to improve decision?

In the theoretical framework, we documented that physicians may suffer from several biases in the way they process and integrate the information. First, they may misevaluate both the analytical and the intuitive components with respect to their objective values $W_{x_i}^{o}$ and W_{x}^{o}. Second, they may inaccurately integrate the analytical (α) and/ or intuitive (β) component(s). How can we replace physician judgment by statistical models to improve the quality of judgment? Operationally, we can replace the analytical man $(\alpha W_{x_i}^{s})$ by the statistical model $(W_{x_i}^{o})$. However, we do not observe the optimal value of intuition (W_{x}^{o}). Thus, the best we can do to improve physician judgment, would be to optimally combine the statistical model $(W_{x_i}^{o})$ with the intuitive man (W_{x}^{s}). The efficacy of this combined approach to improve the accuracy of judgment is well documented (Blattberg & Hoch, 1990). However, whether or not such combined statistical scores can improve the accuracy of actual decisions remains an open question. Indeed, some studies in the literature suggest that physicians' decisions do not fully depend on their judgment (Sorum et al., 2002; Beckstead, 2017). Physicians' decision departure from the judgment may be also relevant. In chapter 3, our goal is twofold: (1) to evaluate, using medical data from the field, whether combining the statistical model with the intuitive man can improve decision; (2) to assess, on the same dataset, whether physicians' actual decision deviate from their expected decision, that is to say the one produced by their judgment, and if so, whether this deviation constitutes relevant information that should be accounted for when designing a combined statistical score.

Our analysis is performed on the same set of medical data, previously described in chapter 1. To measure the statistical model and the intuitive man, we use the same identification approach than the one previously described in chapter 1 (i.e. the Lens model approach). We define physician expected decision to treat as a linear integration of diagnosis cues and their judgment. Deviation from expected decision to treat is the difference between the actual decision and the expected decision. We estimate two statistical scores that combine: (1) the statistical model and the intuitive man, (2) the statistical model, the intuitive man and the observed deviation from expected decision. Finally, we test whether these two combined statistical scores can improve actual decision accuracy.

Methods

Statistical score combining predicted disease probability and residual judgment

We estimate the combined statistical score as the best fitting model for predicting the presence or absence of the disease given the predicted disease probability in log odds ($Lo(\widehat{P^o})$) and the residual judgment (μ), as follows:

$$ln\frac{P\left(Y=1|Lo(\widehat{P^o}),\mu\right)}{P\left(Y=0|Lo(\widehat{P^o}),\mu\right)}=\alpha_0+\ \alpha_1 Lo(\widehat{P^o})+\alpha_2\mu$$

Statistical score combining predicted disease probability, residual judgment and residual decision

- Residual decision

Here, we describe our method to identify physician deviation from expected decision ("residual decision"). **Figure 3** summarizes the method.

First, following the same Lens model approach as described in section 1, we model the decision to treat (T) as a linear logistic function of the cues x_i and the physician probability judgment in log odds ($Lo(P^s)$), as follows:

$$ln\frac{P(T=1|X,Lo(P^s))\,)}{P(T=0|X,Lo(P^s))}=\beta_o^T+\sum_{i=1}^{k}\beta_i^T\,x_i+\beta_s^T Lo(P^s)$$

By applying the regression weights, we estimate in log odds the predicted probability of treatment ($Lo(\widehat{P^T})$).

Second, we measure residual decision (ω), as the difference between the actual decision to treat in log odds and $Lo(\widehat{P^T})$, as follows:

$$\begin{cases} \ln\left(\frac{0.99}{0.01}\right)-Lo(\widehat{P^T}), & if\ actual\ decision\ is\ to\ treat \\ \ln\left(\frac{0.01}{0.99}\right)-Lo(\widehat{P^T}), & if\ actual\ decision\ is\ to\ not\ treat \end{cases}$$

where we choose to attribute the value $ln\frac{0.99}{0.01}$ if actual decision is to treat and $ln\frac{0.01}{0.99}$ if actual decision is to not treat, to handle infinite values.

The residual decision contains the part of physicians' decision that is not explained by physicians' judgment and a linear integration of the cues.

- Statistical score

Finally, we estimate the combined statistical score as the best fitting model for predicting the presence or absence of the disease given the predicted disease probability in log odds, the residual judgment (μ) and the residual decision (ω) as follows:

$$ln\frac{P\big(Y=1|Lo(\widehat{P^o}),\mu,\omega\big)}{P\big(Y=0|Lo(\widehat{P^o}),\mu,\omega\big)}=\gamma_0+\ \gamma_1 Lo(\widehat{P^o})+\gamma_2\mu+\gamma_3\omega$$

Can a combined statistical score improve the accuracy of medical decision?

Empirically, we determine the decision threshold for each combined score that maximizes specificity (i.e. the proportion of patients without disease who do not receive treatment) under the constraint that sensitivity (i.e. the proportion of patients with disease who receive treatment) is equal to the sensitivity of the actual treatment decision. We can then compare the specificity obtained with the combined statistical score to the actual specificity.

Data

We apply our method to the dataset presented in the previous section.

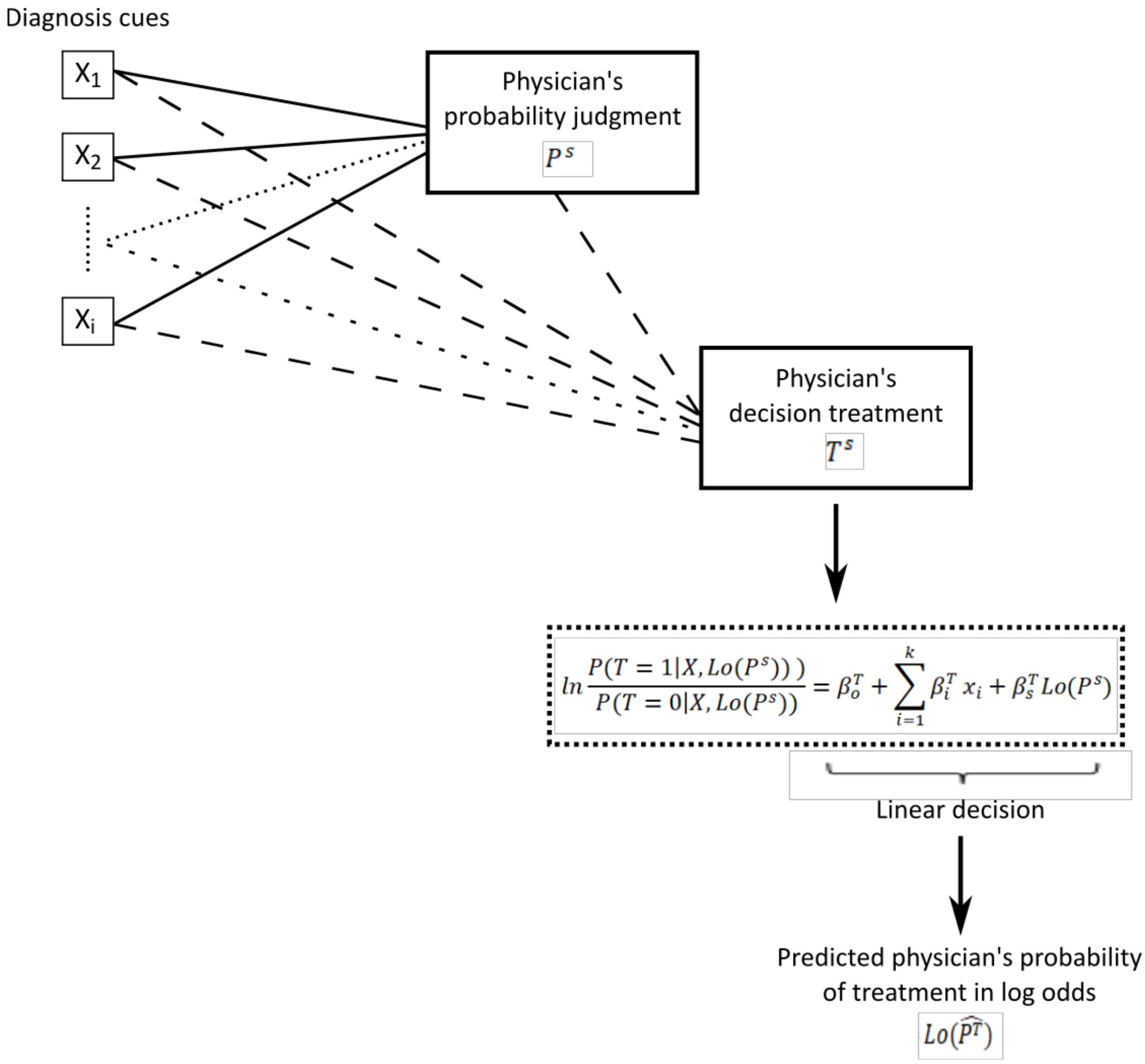

Diagnosis cues
X_1
X_2
X_i
Physician's probability judgment
P^s
Physician's decision treatment
T^s
$ln\frac{P(T = 1|X, Lo(P^s)))}{P(T = 0|X, Lo(P^s))} = \beta_o^T + \sum_{i=1}^{k} \beta_i^T x_i + \beta_s^T Lo(P^s)$
Linear decision
Predicted physician's probability of treatment in log odds
$Lo(\widehat{P^T})$

Figure 3: *Diagram of the Lens model approach applied to decision*

What are the factors that affect human information processing?

In the theoretical framework, we documented that physicians may suffer from several biases in the way they process and integrate the information. First, they may misevaluate both the analytical and the intuitive components with respect to their objective values $W_{x_i}^{o}$ and W_{x}^{o}. Second, they may inaccurately integrate the analytical (α) and/ or intuitive (β) component(s).

What are the factors that affect human information processing?

In chapters 4, 5 and 6, we investigate potential sources of misevaluation of the analytical component[4](chapter 4) and inaccurate integration of the analytical and intuitive components[5] ((chapters 5 and 6). The first factor that we study is working memory. It is well documented that the ability to maintain information in working memory is limited in time and capacity (for a review, see Wickens et al, 2015). The second factor that we study is the misevaluation of the intuitive component[6] as "intuition is sometimes marvelous and sometimes flawed" (Kahneman & Klein, 2009).

In chapter 4, we test whether people's ability to learn about the value of the analytical component, in the absence of external feedback, depends on the quality of their intuitive component. We reason that, in the absence of external feedback, the only source of information that may help people to learn about which diagnosis cues is relevant or not is their intuitive component. Furthermore, we test whether working memory is also necessary to learn in that situation.

[4] i.e. $W_{x_i}^{s} \neq W_{x_i}^{o}$

[5] i.e. $ln\frac{P_{prior}^{s}(Y=1)}{P_{prior}^{s}(Y=0)} + \alpha \sum_{i=1}^{k} W_{x_i}^{s} + \beta W_{x}^{s} \neq \ ln\frac{P_{prior}^{o}(Y=1)}{P_{prior}^{o}(Y=0)} + \sum_{i=1}^{k} W_{x_i}^{o} + W_{x}^{o}$

[6] i.e. $W_{x}^{s} \neq W_{x}^{o}$

In chapter 5, we investigate whether people' ability to integrate the analytical and intuitive components depends on the quality of their intuitive component. We consider that people may misevaluate their intuitive component, which would affect the quality of the integration process.

In chapter 6, we investigate whether people's ability to integrate the analytical and intuitive components depends on their working memory capacity, as we hypothesize that this capacity is required to manipulate information during this stage.

To measure the way people value and integrate the analytical component, we consider the simple situation where only one explicit cue is available. Valuation of the explicit cue is measured by a subjective report when objective value is not available. Integration of the explicit cue is measured by observing how it is used when objective value is known.

To measure the quality of the intuitive component, we propose to use confidence in one's own decision, when no explicit cue is available to make the decision, as a measure of the value one attributes to the intuitive component.

To test our hypotheses about the impact of these two factors, we ran two experiments with a simple perceptual decision task. In one experiment, participants had to learn the value of an explicit cue, in the absence of external feedback. In another experiment, participants were asked to integrate an explicit cue (whose informative value was provided to them) with the perceptual stimulus. For each experiment, we measured working memory and confidence in decision separately.

Method

In the following section, we develop the method to measure the quality of the intuitive component and the method to measure the ability to integrate the analytical and intuitive component.

Measuring the quality of the intuitive component

As previously described, to measure the quality of the intuitive component, we propose to use confidence in decision. Hereafter, we describe the measures used to assess the quality of confidence ratings (Fleming & Lau, 2014).

From confidence ratings, it is important to distinguish between bias and sensitivity. Bias in confidence, also called under- or over-confidence, corresponds to the tendency to give low or high confidence ratings. Sensitivity in confidence corresponds to the ability to distinguish between one's correct and incorrect responses. Figure 4, below, illustrates schematically the difference between sensitivity and bias. The blue and red distributions represent respectively the confidence ratings when the observer is correct and incorrect. For example, an observer can be good at discriminating between her correct and incorrect responses (high sensitivity) but display overconfidence overall (high bias).

We quantify the accuracy of the observer's intuitive component by measuring both dimensions: bias (overconfidence) and sensitivity (metacognitive sensitivity).

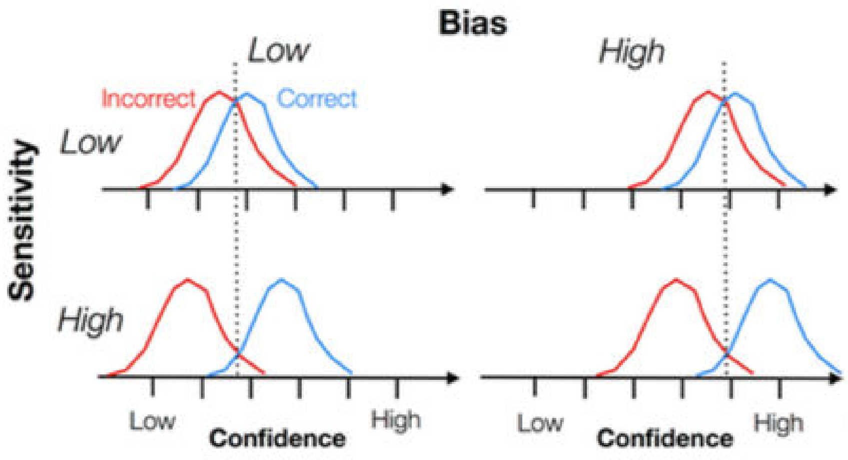


Figure 4: *Schematic representation showing the difference between confidence sensitivity and bias. From "How to measure metacognition" by Fleming, S. M., & Lau, H. C (2014). Frontiers in human neuroscience, 8.*

Measuring the ability to integrate the analytical and intuitive components

To assess the extent to which participants are able to integrate the analytical and intuitive components, we need an optimal benchmark. To determine how both components should be integrated ideally, we need to know the objective value of the intuitive component (i.e. the internal signal). We use Signal Detection Theory (SDT) (Green & Swets, 1966) which proposes a formalization of the internal signal. Hereafter, we present the SDT framework and our measure of integration.

Signal Detection Theory

Transcribed in our medical framework, the SDT model assumes that the internal signal of the observer follows a Gaussian distribution conditional on the presence of the disease ($Y = 1$) ($\sim\mathcal{N}({}^{+d'}/_{2},1)$) or the absence of the disease ($Y = 0$) ($\sim\mathcal{N}({}^{-d'}/_{2},1)$). The distance between the two Gaussian curves is equal to d', named sensitivity, which corresponds to the observer's ability to discriminate between disease present or absent. The observer sets a decision criterion c on her internal axis that determines above which level of internal signal she will answer “Y=1” (i.e. “Disease present”). **Figure 5** illustrates the SDT model.

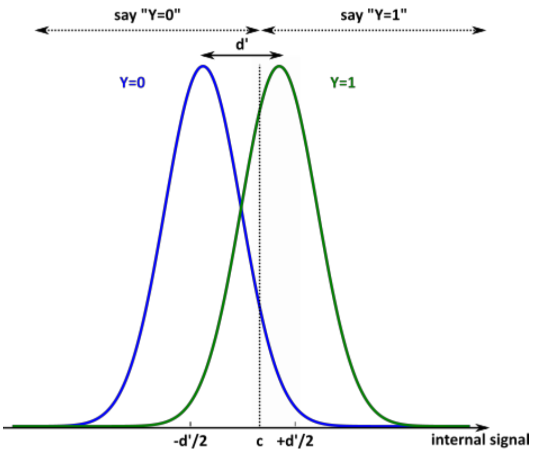


Figure 5: *Diagram of SDT model*

Following this model, the probability of the physician response depends on criterion c, sensitivity d' and the patient condition Y as described in the table below.

		Patient condition	
		Disease	No disease
Physician response	Say “Y=1”	$P(\text{“Y} = 1\text{”}\vert Y = 1)$ $= 1 - F(-d'/_2 + c)$	$P(\text{“Y} = 1\text{”}\vert Y = 0)$ $= 1 - F(d'/_2 + c)$
	Say “Y=0”	$P(\text{“Y} = 0\text{”}\vert Y = 1)$ $= F(-d'/_2 + c)$	$P(\text{“Y} = 0\text{”}\vert Y = 0)$ $= F(d'/_2 + c)$

where $F(z) = \int_{t=-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt.$

SDT parameters estimates criterion c and sensitivity d'

The parameters c and d' can be inferred from the observer’s response by computing $P(\text{“Y} = 1\text{”}|Y = 1)$ and $P(\text{“Y} = 0\text{”}|Y = 0)$, as described below.

$$P(\text{“Y} = 1\text{”}|Y = 1) = 1 - F\left(-d'/_2 + c\right) = F\left(d'/_2 - c\right)$$

Then $d'/_2 - c = Z(P(\text{“Y} = 1\text{”}|Y = 1))$ where $Z(f) = F^{-1}(z)$

$$P(\text{“Y} = 0\text{”}|Y = 0) = F\left({d'}/{2} + c\right)$$

Then ${d'}/{2} + c = Z(P(\text{“Y} = 0\text{”}|Y = 0))$.

Thus

$$d' = Z\big(P(\text{“Y} = 1\text{”}|Y = 1)\big) + Z(P(\text{“Y} = 0\text{”}|Y = 0))$$

$$c = \frac{1}{2}\Big(Z\big(P(\text{“Y} = 0\text{”}|Y = 0)\big) - \ Z(P(\text{“Y} = 1\text{”}|Y = 1)\Big)$$

SDT application: the observer receives one explicit cue x_1 and an internal signal x

Here we derive the optimal integration of the explicit cue x_1 and the internal signal x.

Given the cue x_1, the internal signal x and the objective prior $P^o_{prior}(Y = 1)$, the Bayesian observer forms her posterior belief according to Bayes rule, as follows:

$$ln\frac{P^o(Y = 1|x_1, x)}{P^o(Y = 0|x_1, x)} = ln\frac{P^o_{prior}(Y = 1)}{P^o_{prior}(Y = 0)} + W^o_{x_1} + W^o_x$$

The Bayesian observer decides to respond "$Y = 1$" if $ln\frac{P^o(Y=1|x_1,x)}{P^o(Y=0|x_1,x)} > 0$

The optimal decision criterion $c^{opt}(X)$ is the value of x such that $ln\frac{P^o(Y=1|x_1,x)}{P^o(Y=0|x_1,x)} = 0$

Within the SDT framework, we can compute the objective weight of evidence of the internal signal x, as follows:

$$W^o_x = ln\left(\frac{P^o(x|Y = 1)}{P^o(x|Y = 0)}\right)$$

where $P^o(x|Y = 1) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(x-d'/2\right)^2}$ and $P^o(x|Y = 0) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(x+d'/2\right)^2}$

$$W^o_x = ln\left(\frac{e^{-\frac{1}{2}\left(x-d'/2\right)^2}}{e^{-\frac{1}{2}\left(x+d'/2\right)^2}}\right) = xd'$$

Thus

$$ln\frac{P^o(Y=1|x_1,x)}{P^o(Y=0|x_1,x)} = ln\frac{P^o_{prior}(Y=1)}{P^o_{prior}(Y=0)} + W^o_{x_1} + xd'$$

The optimal decision criterion is equal to:

$$x^* = -\frac{1}{d'}\left(ln\frac{P^o_{prior}(Y=1)}{P^o_{prior}(Y=0)} + W^o_{x_1}\right) \equiv c^{opt}$$

Criterion adjustment

We quantify the extent to which the observer is able to integrate the explicit cue x_1 (i.e. analytical component) and the internal signal x (i.e. intuitive component) with respect to the ideal integration as the observer's deviation from the optimal decision criterion: c^{opt} - c^{obs}.

Data

Experimental protocol

We ran two experiments (for details about the experimental protocol, see **Figure 6**). In each experiment, participants engaged in a simple perceptual decision task: they had to identify which of two sets presented on the computer screen contained more dots (see **Figure 7a**). Each participant completed two experimental sessions, 4 days apart. The order of the two sessions was counterbalanced across participants. In the cue learning experiment (N=65), participants engaged in a "cue learning" session and a "confidence" session. In the cue integration experiment (N=69), participants engaged in a "cue integration" session and a "confidence" session. Each session started with a working memory task. The data of the cue learning experiment are analyzed in chapter 4 and the data of the cue integration experiment are analyzed in chapters 5 and 6. Below we describe each session.

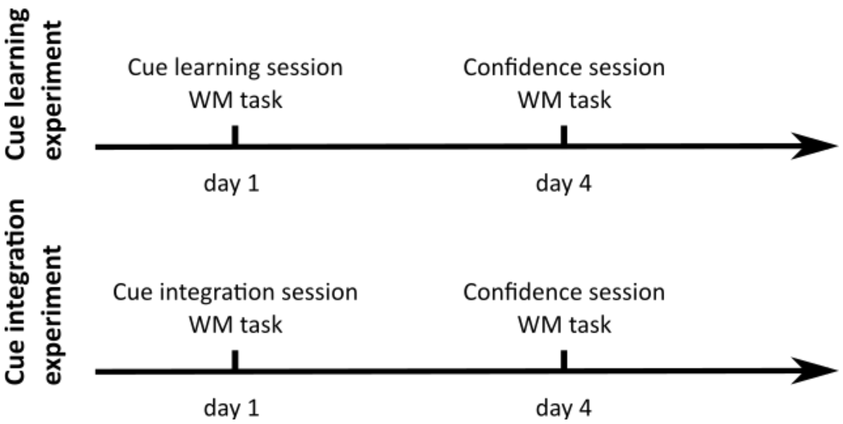


***Figure 6**: Cue learning experiment (N=65) and cue integration experiment (N=69). The order of the two sessions was counterbalanced across participants.*

"Cue learning" session: Each trial started with a central cue. The cue was a square, a circle or a triangle. One shape predicted the left category, one predicted the right category (both with probability p=0.75) and one provided no information about the forthcoming category. Participants were not informed about the associations between the cues and the probabilities of occurrence of a stimulus but they were informed that there were a 'left', a 'right' and a 'neutral' cue. At the beginning of the session, they were required to learn about the cue-stimulus associations, in order to optimize their decisions and they were told that at the end of the session, they had to report these associations (e.g.: Which shape was predictive of the left category?)

"Cue integration" session: Each trial started with a central cue presented for 250ms, before the fixation cross. The cue was either a triangle pointing to the left or the right side of the screen, indicating the correct response with 75% validity (cue condition), or a diamond providing no information (no-cue condition). Participants were fully informed of the meaning of these cues, and instructed to use both the stimulus and the cue to make the best possible decisions.

"Confidence" session: After each left or right response, participants had to indicate their subjective probability of success on a quantitative scale from 50% to 100% confident (see **Figure 7b**).

Working memory task: On each trial, participants received a sequence of letters and were required to report in the forward order the last letters (see **Figure 7c**).

Empirical tests

In chapter 4, we study whether participants' successful identification of the cue-stimulus associations reported in the cue learning session is related to their sensitivity in confidence (measured 4 days apart) and their working memory capacity (average measure across the two sessions).

In chapter 5, we assess whether participants' criterion adjustment measured in the cue integration session is related to their bias in confidence (measured 4 days apart).

In chapter 6, we investigate whether participants' criterion adjustment measured in the cue integration session is related to their working memory capacity (measured the same day).

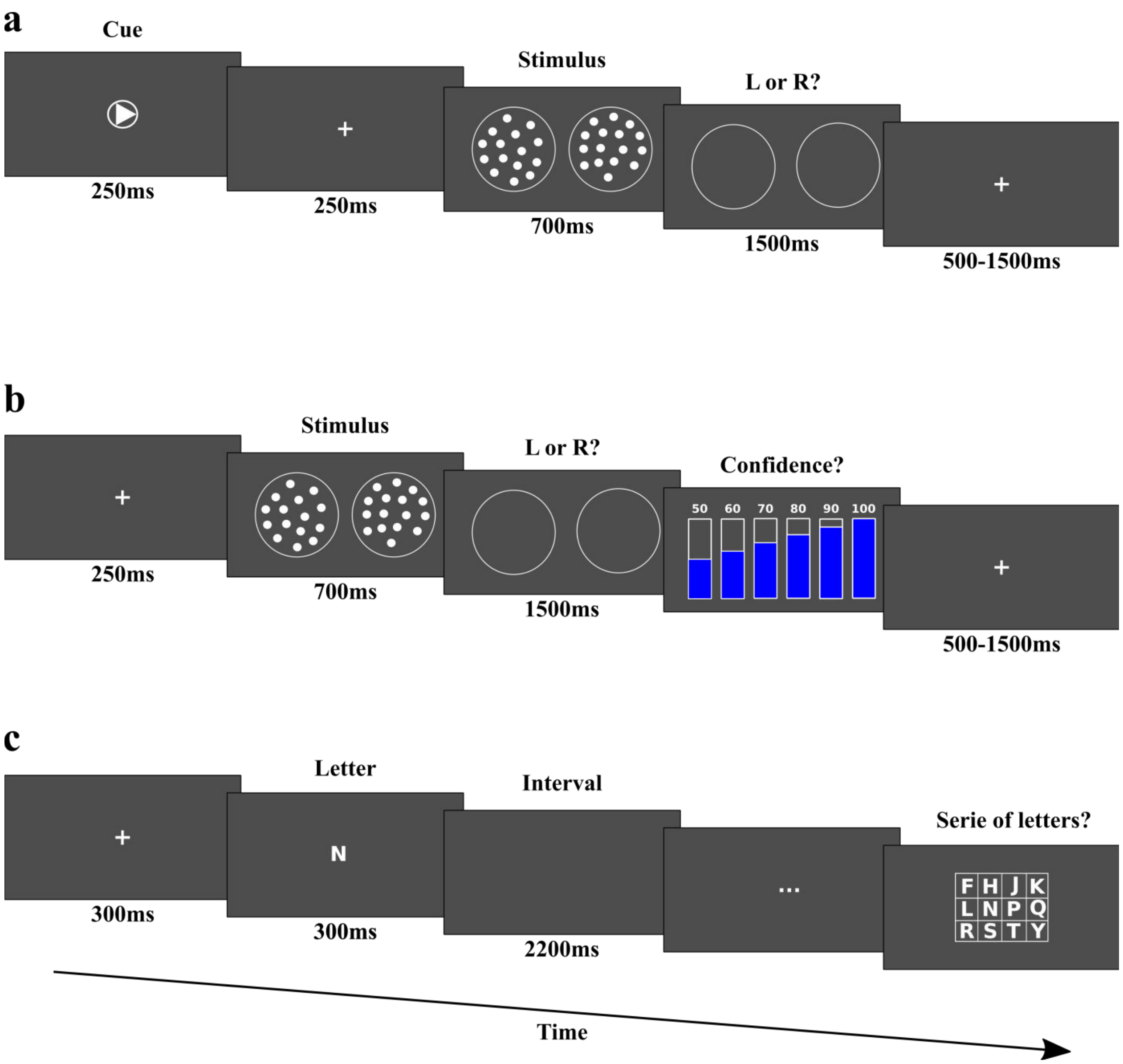


***Figure 7**: **(a)** "Cue learning" session /"Cue integration" session. **(b)** "Confidence" session. **(c)** Working memory task*

Part 1: Probability distortion in clinical judgment: field study

Chapter 1: Impact of probability distortion on medical decisions: a field study

Marine Hainguerlot[1], Vincent Gajdos[2,3], Jean-Christophe Vergnaud[1]

[1]Centre d'Economie de la Sorbonne CNRS UMR 8174, Paris, France.

[2]Department of Pediatrics, Antoine Béclère University Hospital, Assistance Publique-Hôpitaux de Paris.

[3]INSERM, CESP Centre for Research in Epidemiology and Population Health, Paris-Sud, Paris-Saclay University, Villejuif, France.

Abstract

Using data from actual medical practice, we investigated whether probability distortion was a source of suboptimal health care. We developed a method, based on the Lens model approach (Brunswick, 1952; Goldberg, 1970), to quantify probability distortion in actual clinical judgments. We applied our method to the detection of bacterial infection in febrile infants younger than 3 months (N=1848) and assessed its impact on two dimensions of heath care (antibiotic treatment and hospital admission). Overall, we found that physicians' probability estimates were distorted and followed a S-shaped function. They over-estimated small probabilities and under-estimated large probabilities. To assess the impact of probability distortion on medical decisions, we categorized physicians' distortion into a high and a low probability distortion group. Our data showed that the high distortion group provided more health care when the probability of a bacterial infection was small and less health care when the probability was high, compared to the low group. Importantly, we found that such distortion had implications on the accuracy of medical decisions. The specificity was significantly lower in the high bias group while the sensitivity did not differ across the two groups. Critically, the high bias group had a lower decision accuracy rate than the low group. Overall, our results suggested that probability distortion in clinical judgment might cause unnecessary health care.

Introduction

Estimating the probability that a patient has a disease is essential in the diagnostic process. Deviations from the accurate probability may lead to medical errors. Critically, a similar pattern of probability distortion has been observed in many domains (Zhang & Maloney, 2012). Small probabilities are over-estimated and large probabilities are under-estimated. Such S- shaped distortion may have major implications for medical decisions by leading to either unnecessary or insufficient health care. Various studies have documented that cognitive biases and heuristics may affect diagnostic accuracy in medical decision-making. However, most of the studies are based on hypothetical medical situations (Blumenthal-Barby & Krieger, 2014) and there is too little evidence of the impact on medical decision (Saposnik et al, 2016). To the best of our knowledge, probability distortion has not been studied in actual medical practice. In this article, we developed a method, based on the Lens model approach (Brunswick, 1952; Goldberg, 1970), to quantify probability distortion in actual clinical judgments and then assessed its impact on medical decisions.

Probability distortion is commonly measured as the difference between stated subjective probabilities and controlled objective probabilities (eg, see Herrle et al, 2011). However, measuring probability distortion in the field raises methodological issues because the flow of information is not controlled as opposed to hypothetical medical situations. The first concern is the computation of objective probabilities. During the diagnostic process, physicians have access to a massive flow of medical features including but not limited to the medical history, the physical examination, and the clinical findings from laboratory tests. Even though unadjusted diagnostic values of clinical and laboratory features are now synthesized in the literature (eg, see Van den Bruel et al, 2010; Van den Bruel et al, 2011), the values may not be available for all the features. Furthermore, medical features may not be conditionally independent (Holleman & Simel, 1997). Therefore, adjusted diagnostic values should be preferred to compute objective probabilities. A second concern is the use of stated subjective probabilities. The physician may take into account information that is not contained in the medical features. For example, it has been documented that physicians also rely on their clinical intuition (Van den Bruel et al, 2012; Woolley & Kostopoulou, 2013). Consequently, subjective probabilities may depart from objective probabilities if physicians use information that is outside the scope of the medical features used to compute objective probabilities.

Our method requires an exhaustive collection of data from the field during the diagnostic process and a large number of observations. It is necessary to collect the medical features available for each patient (medical history, physical examination and clinical findings) but also to ask physicians, at the end of the diagnostic, to state their probability estimate that the patient has the disease and finally to know whether or not the disease actually occurred. From this exhaustive collection, two sets of predictors can be identified by stepwise regression: the predictors of the presence or absence of the disease (objective set) and the predictors of the probability stated by the physician (subjective set). Next, objective and subjective probabilities can be estimated by applying the corresponding regression weights to the objective and subjective sets of predictors respectively. Probability distortion is computed as the difference between the estimated subjective and objective probabilities. We applied our method to the detection of bacterial infection in febrile infants younger than

3 months (N=1848) and assessed its impact on two dimensions of heath care (antibiotic treatment and hospital admission).

Methods

Study design, setting and participants

Our data come from a study of the procalcitonin biomarker in the detection of bacterial infection (BI) in febrile infants younger than 3 months (Milcent et al., 2016). They performed a prospective, multicenter, cohort study in 15 French pediatric emergency departments for a period of 30 months from October 1, 2008, through March 31, 2011. Infants were eligible in the study if they were older than 7 days and younger than 91 days, with temperatures of 38°C or higher (at home or on admission), without antibiotic treatment within the previous 48 hours, and without major comorbidities (immune deficiency, congenital abnormality, or chronic disease). More details are reported in the original study (see Milcent et al., 2016).

Data collection

Physicians were asked to record the information they collected about the patient from its admission to its discharge. They recorded the demographic and neonatal data, medical history, physical examination, and clinical findings from ordered laboratory tests. Physicians were required to report their probability estimate that the infant had a BI, on a scale from 0 to 100%, and an interval within which their probability fell, at two stages of the data collection: (i) at the end of the physical examination (pretest probability); and (ii) after receiving the clinical findings from the laboratory tests (posttest probability). Following the second estimate, they reported their decisions to hospitalize and to treat with antibiotics, along with a reason for using antibiotics (BI for sure, BI likely or precautionary principle).

Study size

Milcent et al. (2016) enrolled 2273 patients and 2047 infants were included in their final analysis. Furthermore, we restricted our sample to cases with available data regarding probability estimates and medical decisions and obtained a final sample of N = 1848 patients.

Outcome measure

The outcome measure included definite several bacterial infections and possible several bacterial infections categorized by the attending physician. Definite SBIs included bacteremia, bacterial meningitis, urinary tract infection, pneumonia, otitis, gastroenteritis, soft tissue infection, bone and joint infection. Possible SBIs included possible pneumonia and possible otitis. A committee of medical experts reviewed the diagnoses.

Predictor variables

We included all candidate predictor variables collected by physicians in the statistical analysis. Four categories of predictor variables were considered: (1) neonatal and demographic data, (2) medical history, (3) clinical examination, and (4) laboratory tests. We restricted laboratory tests to tests with clinical findings that were available when physicians were asked to report their second estimate, thus we excluded from our analysis blood culture, stool test and the procalcitonin assay. We also excluded additional samplings and additional radiography from the predictors.

Continuous variables were included as continuous and missing data were replaced by the average value computed on available data. Dichotomous variables were coded as +1 if the risk factor was present and -1 if the risk factor was absent; missing data were replaced by the value 0.

Principle for the statistical analysis

Our objective was to measure probability distortion resulting from how physicians processed the medical information collected. We wanted to compare how physicians subjectively integrated this information in their stated probabilities, to an optimal integration. To define optimal integration, we chose to construct a predictive model from our data and not to use odds ratio from the literature (Van den Bruel et al, 2010; Van den Bruel et al, 2011) or existing models (Irwin et al, 2017; Nijman et al, 2013). To carry on this analysis, we made the assumption that medical variables were statistically independent and thus by Bayes rule, we assumed that the log likelihood ratio of the bacterial infection conditional on the medical information was a linear sum of the log likelihood ratio of each medical variable *i* and of the log likelihood ratio of the prior as follows:

$$LLR(BI|medical\ information\ ) = LLR_0(BI) + \sum_i LLR_i \ \ (1)$$

Statistical analysis for predictors

To identify the statistically significant predictors of bacterial infection, we ran stepwise logistic regressions at two stages. A first regression identified predictors from neonatal and demographic data, medical history and physical examination (*objective pretest predictors*). A second regression identified predictors from laboratory tests while controlling for the predictors selected in the first step (*objective posttest predictors*).

Statistical analysis of clinical judgements

To quantify how physicians integrated the medical information into their subjective probabilities, we also assumed that they were forming their belief in a Bayesian way through the linear model (1). Our purpose was to infer from the stated probability estimates the subjective statistical values attributed to medical variables. Assuming that physicians shared some common prior and similar subjective values, we aggregated all data together and estimated one set of subjective values. Estimation was done by running OLS regressions with the stated probability written in a log-odds form as dependent variable. We used stepwise regression to identify the statistically significant predictors of subjective probabilities. Using pretest probability as a dependent variable, a first regression was run to identify predictors from neonatal and demographic data, medical history and physical examination (*subjective pretest predictors*). Using posttest probability, a second regression was run to identify predictors from laboratory tests while controlling for the predictors selected in the first step (*subjective posttest predictors*).

Objective and subjective odds ratio for predictor variables

From the stepwise regressions, we were able to identify the predictors of the bacterial infection (objective predictors) and the predictors of the probability stated by physicians (subjective predictors). First, we compared how physicians were subjectively weighting the predictors compared to the objective weights. To do so, we ran regressions on pooled list of objective and subjective predictors. We considered that subjective odds ratio could depart from objective odds ratio in two ways. First, physicians could over- or under-value significant predictors of BI (over- or under-evaluation bias). Second, physicians could attribute values to predictors that were not predictive of BI (misattribution bias). Next, objective and subjective posttest probabilities were estimated by applying the corresponding objective and subjective regression weights to the pooled list of pretest and posttest predictors.

Probability distortion analysis

We computed probability distortion as the difference between the estimated subjective and objective probabilities. To evaluate probability distortion, estimated subjective probabilities (ps) versus objective probabilities (po) were plotted on probability scales but also on log odds scales ($Lo(p) = \frac{p}{1-p}$). We used the general linear model to estimate a slope parameter and an intercept parameter of the distortion when probabilities were transformed in log odds ($Lo(ps) = \beta_0 + \ \beta_1 Lo(po)$) (Zhang & Maloney, 2012). The over- or under-evaluation bias corresponds to the slope parameter β_1. If the probability distortion follows a S- shaped pattern, we should observe a linear pattern in log odds form representation. When β_1=1, there is no distortion. When $\beta_1 < 1$, the probability distortion follows a S-shaped curve (i.e. overestimation of small probabilities and under- estimation of large probabilities). When $\beta_1 > 1$, the probability distortion follows an inverted S- shaped curve (i.e. under- estimation of small probabilities and over- estimation of large probabilities). The misattribution bias corresponds to the goodness of the fit (R^2). If R^2=1, there is no misattribution bias. On the other hand, the less the fit explains the variance, the greater should be the misattribution bias. Additionally, we used the predicted values ($Lo(\widehat{ps}) = \hat{\beta}_0 + \ \hat{\beta}_1 Lo(po)$) to plot the distortion in probability form with $\widehat{ps} = \frac{\exp(Lo(\widehat{ps}))}{(1+\exp(Lo(\widehat{ps})))}$. Finally, we considered that the difference between the posttest probability reported by physicians and the estimated posttest subjective probability, defined in the remainder of the article as “residuals”, contains the part of the physicians’ judgment that is not explained by a linear integration of the predictors. It may capture physicians’ intuition, physicians’ ability to interpret omitted features, or their ability to take into account predictors in a nonlinear way.

Impact on medical decisions

Critically, if physicians based their decisions on their subjective probabilities, by using for instance a cutoff value (i.e. a level of probability of the disease required) to decide when to treat by antibiotics and when to admit to the hospital, a S-shaped probability distortion should impact decisions. If the threshold is low, over-estimation of small probabilities should lead to excessive health care. If the threshold is high, under-estimation of large probabilities should lead to insufficient health care. To test our hypothesis, we classified physicians’ clinical judgments into two groups: a high probability distortion group and a low probability distortion group. The groups were constructed by median split on the value of the predicted subjective probabilities while controlling for the level of objective probabilities. More precisely, we classified in the high bias group subjective probabilities above the median split when objective probabilities were small and subjective probabilities below the median when objective probabilities were large. We defined small and large objective probabilities with respect to the prevalence of BI in our data. Objective probabilities were considered to be small when they were lower than the prevalence rate and to be large otherwise. When objective probabilities were small, the high bias group had higher subjective probabilities than the low bias group. Inversely, the high bias group had lower subjective probabilities when objective probabilities were large. Consequently, assuming that physicians in the two groups held a common threshold value, medical decisions might differ across the two groups even though the prevalence of BI was the same. If the threshold is low, the high bias group might provide health care more intensively than the low bias group. If the threshold is high, the high bias group might provide health care less intensively.

Impact on the accuracy of medical decisions

To evaluate the accuracy of medical decisions, we considered that health care (antibiotic treatment and hospital admission) was necessary only if the infant had a BI. We used three measures: sensitivity (i.e. the proportion of infants with BI who received heath care), specificity (i.e. the proportion of infants without BI who did not receive health care) and accuracy (i.e. the proportion of infants who correctly received health care). We investigated whether probability distortion had an impact on the accuracy of medical decisions by comparing the sensitivity, specificity and accuracy between the high and low bias probability distortion groups.

Results

Diagnoses

In our data (N=1848), the prevalence of BI was 18.1% (334/1848). Among the infants with definite BI (284/334), 257 had urinary tract infection. Among the possible BI (50/334), possible pneumonia was the most frequently diagnosed (35/50). Detailed data on diagnoses are reported in **Table 1**.

Table 1: Bacterial Infections
Value are numbers (percentage)

	N	No	(%)
BI	1848	334	(0.18)
Definite BI	1848	284	(0.15)
Bacteremia	284	10	(0.04)
Bacterial meningitis	284	7	(0.02)
Urinary tract infection	284	257	(0.90)
Pneumonia	284	0	(0.00)
Otitis	284	4	(0.01)
Gastroenteritis	284	3	(0.01)
Soft tissue infection	284	1	(0.00)
Bone and joint infection	284	0	(0.00)
Other definite BI	284	2	(0.01)
Possible BI	1848	50	(0.03)
Possible pneumonia	50	35	(0.70)
Possible otitis	50	11	(0.22)
Other possible BI	50	4	(0.08)

N: Total number of available data
Abbreviation: BI, bacterial infection

Clinical judgments and decisions

The statistics concerning clinical judgments (clinical appearance, pretest probability, posttest probability and probability intervals) as well as medical decisions (hospital admission and antibiotic treatment) are reported in **Table 2**. Overall, physicians' probability estimates that infants had BI were pessimistic compared to the prevalence of BI observed in our data. The mean of the pretest probability was 27.5% and the mean of the posttest probability was 25.43% whereas we observed a rate of BI of 18.1%. Posttest probabilities were more dispersed (sd = 27.9) than posttest probabilities (sd = 25.43) thus probability estimates were more precise (closer to 0% or 100%) after receiving the clinical findings from the laboratory tests. Antibiotic treatment was administered to 41% of the infants and 73% of the infants were admitted to the hospital.

Table 2: Judgments and medical decisions
Values are numbers (percentage)
unless stated otherwise by*: mean (sd)

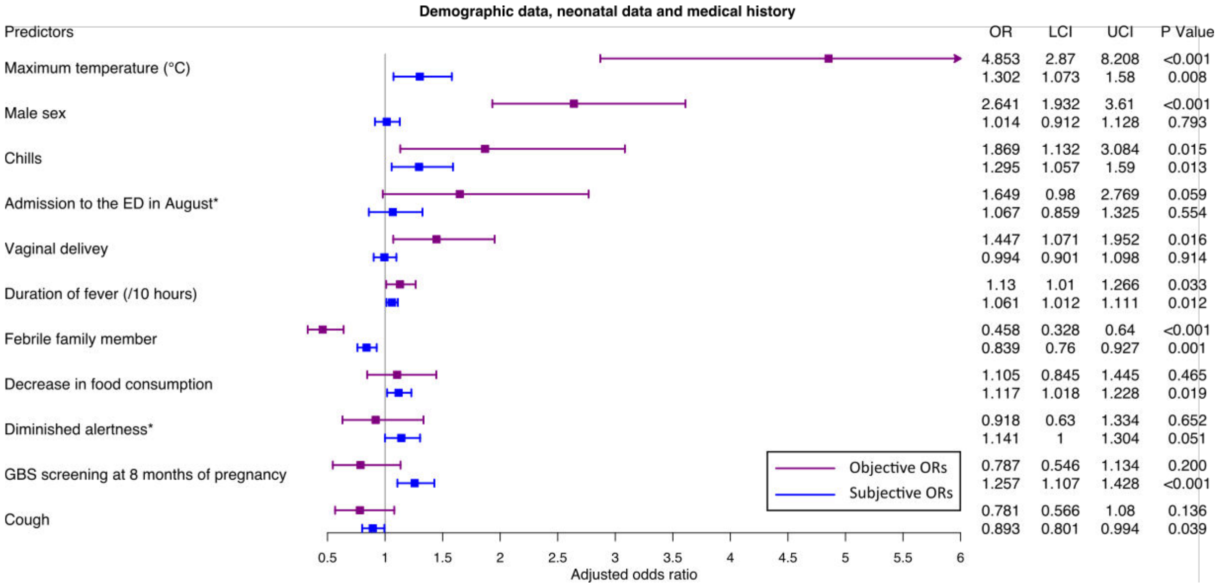
	N	No	(%)
Clinical appearance, well	1820	196	(0.11)
Clinical appearance, minimally ill	1820	1065	(0.59)
Clinical appearance, moderately ill	1820	500	(0.27)
Clinical appearance, very ill	1820	59	(0.03)
Pretest probability*	1848	27.47	(20.88)
Lower bound pretest probability*	1848	17.21	(18.66)
Upper bound pretest probability*	1848	40.85	(24.05)
Posttest probability*	1848	25.43	(27.93)
Lower bound posttest probability*	1848	18.30	(26.21)
Upper bound posttest probability*	1848	34.17	(29.55)
Hospital admission	1848	1340	(0.73)
Antibiotic treatment	1848	766	(0.41)
Reason for antibiotic, BI for sure	766	217	(0.28)
Reason for antibiotic, BI likely	766	200	(0.26)
Reason for antibiotic, PP	766	182	(0.24)
Reason for antibiotic, None	766	167	(0.22)

N: Total number of available data
Abbreviation: PP, precautionary principle

Objective and subjective odds ratio for predictor variables

The list of predictor variables that we considered and descriptive statistics are reported in Supplementary Results (**Table S1**). **Figures 1**, **2** and **3** show the comparison of objective odds ratios (OR_o) and subjective odds ratios (OR_s) for the predictors identified from the stepwise regressions. Overall, subjective odds ratios departed from objective odds ratios in two ways. First, physicians undervalued significant predictors of BI. When the odds ratios of the predictors of BI were above the value one, the subjective odds ratios were lower than the objective odds ratios (i.e. $1 < OR_s < OR_o$). When the odds ratios of the predictors of BI were below the value one, the subjective odds ratios were greater than the objective odds ratios (i.e. $OR_o < OR_s < 1$). Second, physicians attributed predictive value to predictors that were actually not predictive of BI in our data. The same patterns of under-valuation and misattribution were observed for the demographic data, neonatal data and medical history (**Figure 1**), clinical examination (**Figure 2**) and laboratory tests (**Figure 3**)

In **Figure 1**, some medical variables (maximum temperature, male sex, chills, duration of fever and febrile member family) were common predictors of BI and stated pretest probabilities but were systematically under-valued subjectively. Moreover, physicians attributed predictive values to predictors that were not statistically significant in the objective analysis (food consumption, diminished alertness, GBS screening at 8 months of pregnancy and cough). We also noted that admission to the ED in August and vaginal delivery were predictive of BI but were not taken into account by physicians. In **Figure 2**, only rhinitis and rash were objective predictors that physicians undervalued while they attributed predictive values to many other variables from the clinical examination. Concerning laboratory tests (**Figure 3**), physicians undervalued the alveolar consolidation, urine analysis positive n°1, urine analysis positive n°2, fibrinogen, granulocyte neutrophile, C-reactive protein and pulmonary hyperinflation while they misattributed value to capillary blood glucose, WBC count in CSF, lymphocyte, monocyte and interstitial syndrome.



Predictors	OR	LCI	UCI	P Value
Maximum temperature (°C)	4.853 1.302	2.87 1.073	8.208 1.58	<0.001 0.008
Male sex	2.641 1.014	1.932 0.912	3.61 1.128	<0.001 0.793
Chills	1.869 1.295	1.132 1.057	3.084 1.59	0.015 0.013
Admission to the ED in August*	1.649 1.067	0.98 0.859	2.769 1.325	0.059 0.554
Vaginal delivey	1.447 0.994	1.071 0.901	1.952 1.098	0.016 0.914
Duration of fever (/10 hours)	1.13 1.061	1.01 1.012	1.266 1.111	0.033 0.012
Febrile family member	0.458 0.839	0.328 0.76	0.64 0.927	<0.001 0.001
Decrease in food consumption	1.105 1.117	0.845 1.018	1.445 1.228	0.465 0.019
Diminished alertness*	0.918 1.141	0.63 1	1.334 1.304	0.652 0.051
GBS screening at 8 months of pregnancy	0.787 1.257	0.546 1.107	1.134 1.428	0.200 <0.001
Cough	0.781 0.893	0.566 0.801	1.08 0.994	0.136 0.039

Figure 1: *Forest plot of objective and subjective adjusted odds ratio for predictors from demographic data, neonatal data and medical history. * The p-values of the predictors admission to the ED in August and diminished alertness became greater than 5% once we included the objective and subjective predictors selected in the first step.*

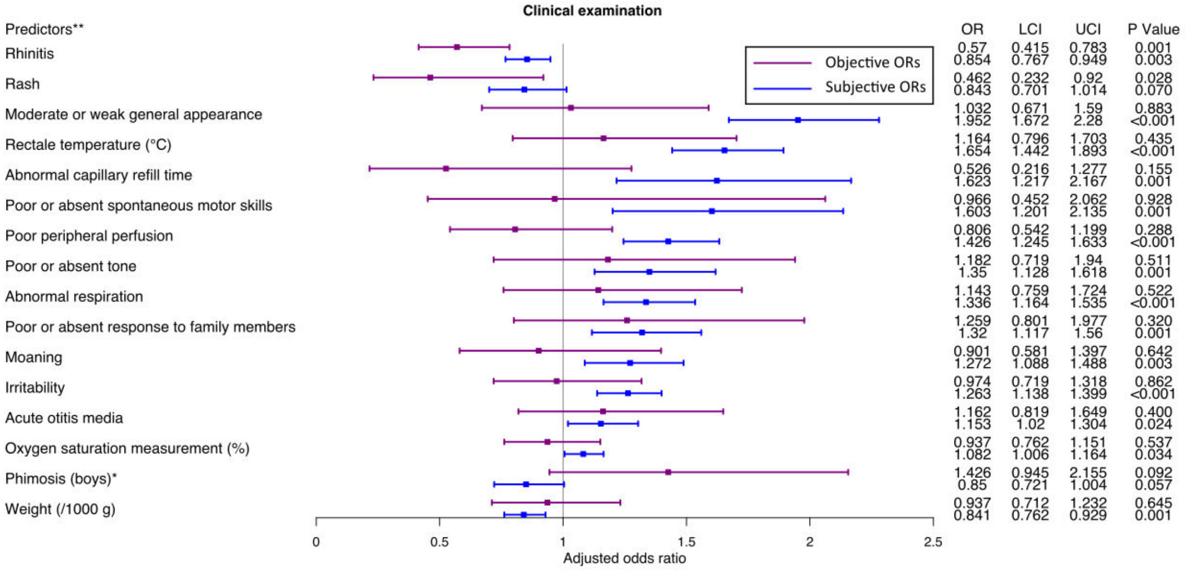


Figure 2: *Forest plot of objective and subjective adjusted odds ratio for predictors from clinical examination. * The p-value of the predictor phimosis (boys) became greater than 5% once we included the objective and subjective predictors selected in the first step. **We excluded from the plot the predictor lymphadenopathy (Objective values: OR= 10.883, LCI= 0.941, UCI= 125.709, p-value= 0.056; Subjective values: OR= 4.490, LCI= 1.631, UCI= 12.355, p-value= 0.004)*

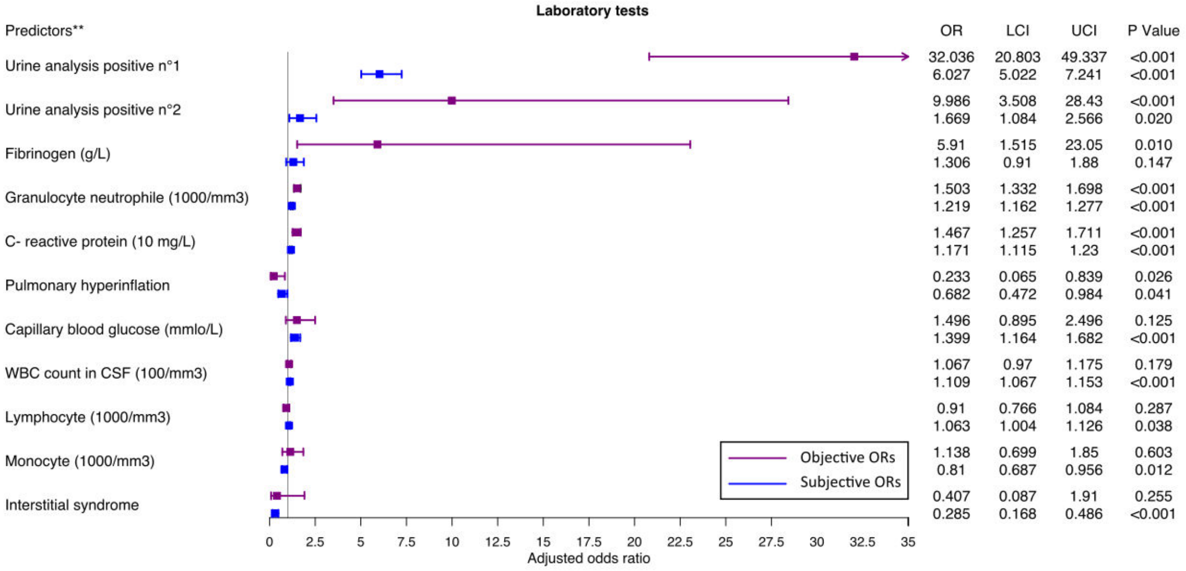


Figure 3: *Forest plot of objective and subjective adjusted odds ratios for predictors from laboratory tests. **We excluded from the plot the predictor alveolar consolidation (Objective*

values: OR= 59.044, LCI= 26.843, UCI= 129.892, p-value<0.001; Subjective values: OR= 4.109, LCI= 2.853, UCI= 5.919, p-value<0.001)

Probability distortion

First, we confirmed that our measure of probability distortion captured a distortion that was not predictive of BI. We ran a logistic regression with objective probability, probability distortion and residuals (all expressed in log odds form) as explanatory variables to predict BI. We found that probability distortion was not predictive (b=-0.022, p = 0.876) of BI while objective probability (b=1.021, p < 0.001) and residuals (b=0.483, p < 0.001) were significant predictors. The significance of the residuals term provides support for the hypothesis that physicians had access to relevant information to detect BI outside the scope of the statistical model. Next, to illustrate probability distortion in our data**,** estimated subjective probabilities were plotted against objective probabilities in **Figure 4**. The distortion is reported with probabilities expressed on log odds scale in **Figure 4a** and probabilities without transformation in **Figure 4b**. The fit of the data ($Lo(\widehat{ps}) = -0.3608 + \ 0.4201 * Lo(po)$)) showed that subjective and objective probabilities followed a linear pattern in log odds form representation (**Figure 4a**) with observations above the 45° degree line for low objective probabilities and below the 45° degree line for large objective probabilities. The slope of the linear was less than 1 (b=0.42, p <0.001). Overall, physicians over-estimated small probabilities and under-estimated large probabilities indicating that they did not take into account the predictors as much as they should have (under-evaluation bias). The linear fit accounted for 72.6% of the variance suggesting that misattribution bias was substantial in the data. The S- shaped pattern when expressed in probability form is plotted in **Figure 4b**.

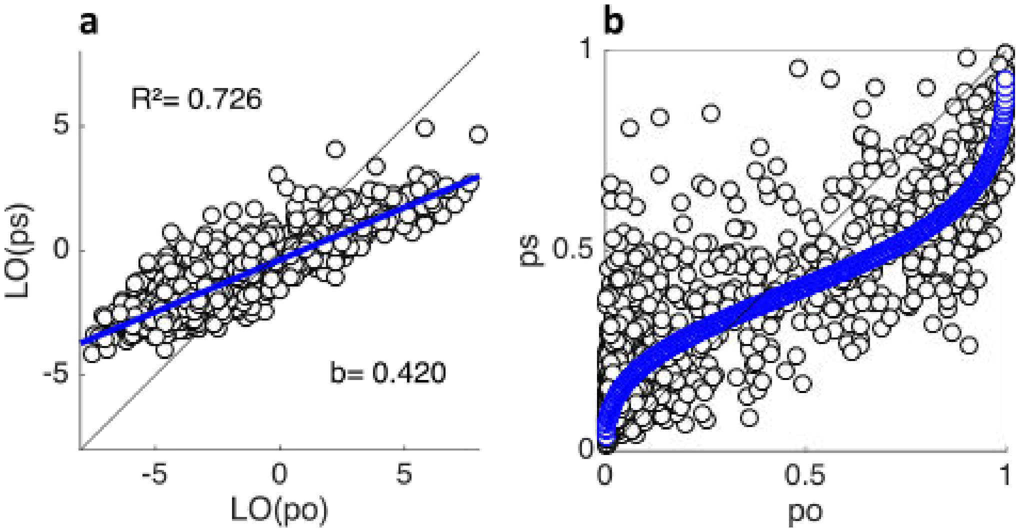


Figure 4: ***(a)*** *Estimated subjective probabilities versus objective probabilities on log odds scale. The blue line is the linear fit.* ***(b)*** *Subjective probabilities versus objective probabilities.*

Impact on medical decisions

The classification of the high and low probability distortion groups is presented in **Figure 5**. The slope of the linear fit was closer to one for the low bias group (b=0.539) than for the high bias group (b=0.289), which confirmed that physicians' estimates were more distorted in the high bias group than in the low bias group.

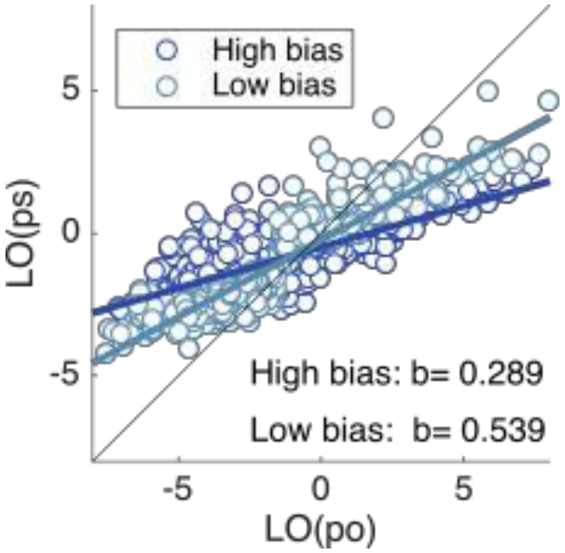


Figure 5: *Estimated subjective probabilities versus objective probabilities on log odds scale by distinguishing the high and low probability distortion groups. The dark and light blue lines are the best linear fits for the high and low bias probability distortion groups respectively.*

Proportion tests in **Table 3** show the differences in health care decisions across the two groups. Importantly, the prevalence of BI was not statistically different across the two groups. When objective probabilities were small, the rates of antibiotic treatment and hospital admission were significantly higher in the high bias group compared to the low bias group. When objective probabilities were large, the rates of antibiotic treatment and hospital admission were significantly lower in the high bias group compared to the low bias group. These results were in line with our predictions concerning the impact of probability distortion on decisions but we did not expect to observe an impact for both small and large probabilities as we assumed that physicians used only one threshold value to make decisions. The results suggested that physicians might use different threshold values.

Table 3: Proportion test comparisons of bacterial Infection rate, antibiotic treatment rate and hospital admission rate (for small and large objective probabilities) by low and high bias group

	Low bias		High bias		Proportion test 2 tailed		
	N	Mean	N	Mean	Proportion difference	z	p-value
BI	955	0.183	893	0.178	0.005	0.290	0.7718
Antibiotic treatment							
Small probabilities	733	0.202	680	0.379	-0.178	-7.367	<0.001
Large probabilities	222	0.892	213	0.761	0.131	3.625	<0.001
Hospital admission							
Small probabilities	733	0.565	680	0.803	-0.238	-9.584	<0.001
Large probabilities	222	0.946	213	0.798	0.148	4.638	<0.001

Impact on the accuracy of medical decisions

Table 4 reports proportion test comparisons of sensitivity, specificity and accuracy rates between the two groups for health care (antibiotic treatment and hospital admission). Importantly, we found that specificity was significantly lower in the high bias group. The sensitivity did not differ among the two groups. Overall, the high bias group had a lower decision accuracy rate. Critically, our results suggest that probability distortion in clinical judgment might impact negatively the accuracy of medical decisions. In particular, it might cause unnecessary health care that is providing antibiotic treatment and admitting to the hospital infants without BI detected.

Table 4: Proportion test comparisons of sensitivity rate, specificity rate and decision accuracy rate (for antibiotic treatment and hospital admission) by low and high bias group

	Low bias		High bias		Proportion test 2 tailed		
	N	Mean	N	Mean	Proportion difference	z	p-value
Antibiotic treatment							
Sensitivity	175	0.977	159	0.987	-0.010	-0.706	0.480
Specificity	780	0.776	734	0.642	0.134	5.745	<0.001
Accuracy	955	0.813	893	0.703	0.109	5.497	<0.001
Hospital admission							
Sensitivity	175	0.920	159	0.862	0.058	1.718	0.086
Specificity	780	0.406	734	0.211	0.195	8.196	<0.001
Accuracy	955	0.501	893	0.327	0.174	7.562	<0.001

Discussion

We report evidence from a large prospective study that physicians' probability estimates that infants have BI were distorted. Our analysis revealed that physicians under-valued medical evidence and also attributed values to factors that were not predictive of BI. Overall, physicians over-estimated small probabilities and under-estimated large probabilities (S-shaped probability distortion). Critically, we found that such distortion in clinical judgment might cause unnecessary health care regarding antibiotic treatment and hospital admission.

Probability distortion

A possible explanation for the under-evaluation of medical evidence is that physicians might insufficiently integrate the information provided. Laboratory experiments have reported that observers integrate probabilistic information but not to the extent that they should (Ackerman & Landy, 2015; Lusted, 1976). Interestingly, it has been suggested that one potential source of insufficient integration may be over-confidence in one's own judgment (Daniel et al, 1998; Kubovy, 1972; Odean, 1998). To explore this source, we tested whether over-confidence in pretest probabilities explained probability distortion in posttest probabilities. We measured confidence in pretest probabilities by using the size of the interval in probability estimates (Moore & Healy, 2008). By median split on the value of the interval, while controlling for the level of objective probabilities, we separated our data in a high (N=888) and a low (N=960) interval group. In the high interval group, we observed that the interval contained the objective probability on 58.56% of the clinical cases and only on 28.44% of the cases in the low interval group. Intervals were too narrow in the high interval group, suggesting that physicians were excessively certain (over-precision) about the accuracy of their probability. We tested whether probability distortion was related to over-precision. In an OLS regression, with probabilities expressed in the log odds form, we found that posttest probability was explained both by objective probability (b=0.38, p<0.001) and an interaction term between objective probability and the interval group (dummy equal to one for the high interval group, and 0 for the low interval group) (b=0.086, p<0.001). Thus the slope of the linear fit between subjective and objective probabilities was significantly higher for the low interval group. This result suggests that physicians' overconfidence may contribute to probability distortion.

A common explanation for under-evaluation and misattribution of medical evidence is that physicians might not value the evidence in accordance with our objective estimates.

Both biases might come from a discrepancy between physicians' knowledge and our estimates. In particular, we found that clinical factors identified as predictors of BI in the literature (for a review, see Van den Bruel et al, 2010) such as rapid breathing and poor peripheral perfusion, or recommended by the Yale Observation Scale (e.g.: moaning, irritability, poor or absent response to family members and poor or absent spontaneous motor skills) were taken into account by physicians whereas these factors had no objective predictive value in our study.

Impact of probability distortion on medical decisions

We investigated an explanation for the detrimental effect of probability distortion on decision. We reasoned that decision might be negatively impacted by probability distortion through clinical judgment. In particular, physicians might rely on their clinical judgment to make decisions by, for instance, applying a decision threshold on their judgment (i.e. a level of probability of the disease required) above which they will provide health care. In that case, the sensibility and specificity of the decision should correspond to a point on the ROC (Receiver Operating Characteristic) curve to detect BI for the clinical judgment. As a result, if physicians rely on their clinical judgment and that probability distortion impacts negatively their judgment, then the quality of decisions might be impaired. Critically, we observed in our data that the AUC (Area Under the Curve) of the ROC curve to detect BI for stated subjective probabilities was significantly lower in the high bias group (AUC high bias group: 0.802 (95% CI: 0.760 - 0.844) vs. AUC low bias group: 0.895 (95% CI: 0.864 - 0.926)). This empirical result provided support for our intuitive explanation. Nevertheless, at a theoretical level, it is not obvious that probability distortion should negatively impact the diagnostic accuracy (AUC) of subjective probabilities. We investigated how and when probability distortion should impair diagnostic accuracy (for more details, see the proof in the Supplementary Results). We argued that probability distortion in the medical evidence is not a sufficient condition. Instead, we showed that the detrimental effect might occur when physicians have to combine multiple sources of information. In such situation, probability distortion in the medical evidence should bias the aggregation of the information, which should eventually affect diagnostic accuracy. Interestingly, we believe that our data might provide support for the proof. Indeed, as mentioned above, we found that the diagnostic accuracy for stated subjective probabilities was significantly lower in the high bias group. On the other hand, for estimated subjective probabilities, the difference between the two groups in AUCs of the ROC curve to detect BI was reduced (AUC high bias group: 0.8802 (95% CI: 0.85001 - 0.91041) vs. AUC low bias group: 0.9135 (95% CI: 0.88373 - 0.94323)). **Figure S1** in the Supplementary Results summarizes the AUC of the ROC curves to detect BI for stated and estimated subjective probabilities, per low and high bias group. By definition, estimated subjective probabilities correspond to the evaluation of medical evidence whereas stated probabilities combine both estimated probabilities and physicians' private information not captured by our model. Our proof provides one possible explanation for the observation that probability distortion had as substantial impact on stated subjective probabilities whereas the effect on estimated subjective probabilities was quite small.

Even though we found that probability distortion impaired specificity, the quality of the decision for antibiotic treatment remained high in our data with an excellent sensitivity (98.20% (95% CI: 96,78%-99,63%) and a good specificity (71.07% (95% CI: 68,79%-73,35%)). We compared the quality of the actual decision with potential aids that could minimize the impact of probability distortion. We investigated whether guidelines, biomarkers and mechanical aids could improve the quality of the decision. First, guidelines recommend antibiotic treatment for all young febrile infants except the infants older than 1 month, showing good clinical appearance and negative results to urine analysis, WBC, CSF analysis and chest radiography. Applied to our data, a sensitivity of 92.51% (95% CI: 89.69%-95.34%) and a specificity of 47.16% (95% CI: 44,65% - 49,67%) would have been obtained with the guideline. Regarding biomarkers, we considered the procalcitonin and the C-reactive protein and used decision threshold suggested in the literature. With a threshold at 0.3 ng/ml for the procalcitonin, a sensitivity of 68.04% (95% CI: 62.90%-73.18%) and a specificity of 83.37% (95% CI: 81.42%-85.32%)) would have been reached. A threshold at 20 mg/L for the C-reactive protein would have produced a sensitivity of 84.84% (95% CI: 80.61%-89.06%) and a specificity of 53.52% (95% CI: 48.94%- 58.11%)). Finally, we reasoned that if physicians have distorted probability estimates, it would be best to replace their clinical judgment by a mechanical aid which is unbiased, by definition. We considered the objective probabilities estimated in our data as a potential aid. With a probability threshold at 4.84% for our estimated objective probabilities, we obtained a sensitivity of 93.11% (95% CI: 90.40%-95.83%) and a specificity of 71.07% (95% CI: 68.79%-73.35%). Thus, overall, despite probability distortion, actual decision could outperform the potential aids that we considered. As a result and importantly, to minimize the impact of probability distortion it might not be the best to replace physicians.

Limitations

Our study has several limitations. First, to define probability distortion we considered that physicians integrated medical information collected in a linear way. This assumption may not be accurate. Physicians might take into account the information available differently. In particular, they could integrate the evidence by giving different weights to the predictors depending on the presence or absence of other predictors. This limit has two implications. The discrepancy between stated subjective probability and estimated subjective probability, which we defined as "residuals", may also contain information from the data collected that was not captured by our linear assumption. Moreover, physicians could integrate the information more optimally than our linear specification. As a result, our measure may inaccurately capture a distortion. The comparison of the AUC of the ROC curve to detect BI for objective probabilities predicted by our linear assumption 0.937 (95%CI: 0.920- 0.954) with the AUC of the ROC curve to detect BI for stated subjective probabilities 0.855 (95%CI: 0.829- 0.881) showed that our linear assumption better predicted BI. This result provided support for the relevance of our assumption. Second, it may be argued that probability distortion is due to physicians' inability to report correctly their estimates on a probability scale. To discuss this issue, we compared the pretest probability reported by physicians with another subjective report commonly used in medical practice that is the clinical appearance. They were asked to report these two subjective reports at the end of the clinical examination. They classified the clinical appearance of the patient as well, minimally, moderately or very ill. A Spearman correlation confirmed that pretest-probability was significantly correlated with the clinical appearance (N=1820, r= 0.705, p <0.001). In addition, the comparison of the AUC of the ROC curves to detect BI for pretest probability 0.640 (95%CI: 0.606- 0.674) and for clinical appearance 0.580 (95%CI: 0.549- 0.612) showed that physicians' reports were more predictive of BI with the probability scale thus physicians were able to take advantage of the scale. Finally, when calculating probability distortion we used all the information collected that was available even when predictors had few observations. Indeed, in an attempt to be exhaustive about the information that physicians used to report their subjective probabilities we decided to include clinical findings from laboratory tests even when such tests were not ordered by many physicians. However, by including clinical findings with few observations we might have inaccurately estimated the odds ratios. To assess the robustness of our findings, estimation of the probability distortion on a restricted set of predictors with limited missing observations is in progress.

Conclusion

Our field study revealed that physicians had distorted probability estimates that infants had bacterial infection. Importantly, we found that such distortion impaired the specificity of medical decision on two dimensions: antibiotic treatment and hospital admission. Statistical aids for febrile infants at risk of bacterial infection should be developed to minimize the impact of probability distortion on medical decisions. Overall, our results suggest that probability distortion in clinical judgment might cause unnecessary health care. Further studies should investigate whether this phenomenon is also observed for other medical diagnostic problems.

References

Ackermann, J. F., & Landy, M. S. (2015). Suboptimal decision criteria are predicted by subjectively weighted probabilities and rewards. *Attention, Perception, & Psychophysics*, *77*(2), 638-658.

Blumenthal-Barby, J. S., & Krieger, H. (2015). Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. *Medical Decision Making*, 35(4), 539-557.

Brunswik, E. (1952). The conceptual framework of psychology. *Psychological Bulletin*, *49*(6), 654-656.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. *The Journal of Finance*, *53*(6), 1839-1885.

Dawson, N. V., & Arkes, H. R. (1987). Systematic errors in medical decision making. *Journal of General Internal Medicine*, *2*(3), 183-187.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin*, *73*(6), 422.

Herrle, S. R., Corbett Jr, E. C., Fagan, M. J., Moore, C. G., & Elnicki, D. M. (2011). Bayes' theorem and the physical examination: probability assessment and diagnostic decision-making. *Academic medicine: journal of the Association of American Medical Colleges*, *86*(5), 618.

Holleman, D. R., & Simel, D. L. (1997). Quantitative assessments from the clinical examination. *Journal of General Internal Medicine*, *12*(3), 165-171.

Irwin, A. D., Grant, A., Williams, R., Kolamunnage-Dona, R., Drew, R. J., Paulus, S., ... & Appelbe, D. (2017). Predicting Risk of Serious Bacterial Infections in Febrile Children in the Emergency Department. *Pediatrics*, 140(2), e20162853.

Kubovy, M. (1977). A possible basis for conservatism in signal detection and probabilistic categorization tasks. *Attention, Perception, & Psychophysics*, *22*(3), 277-281.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of experimental psychology: Human learning and memory*, *4*(6), 551.

Lusted, L. B. (1976). Clinical decision making. In D. Dombal & J. Grevy (Eds.), Decision making and medical care. Amsterdam: North Holland.

Milcent, K., Faesch, S., Gras-Le Guen, C., Dubos, F., Poulalhon, C., Badier, I., ... & Nissack, G. (2016). Use of procalcitonin assays to predict serious bacterial infection in young febrile infants. *JAMA pediatrics*, 170(1), 62-69.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502.

Nijman, R. G., Vergouwe, Y., Thompson, M., Van Veen, M., Van Meurs, A. H., Van Der Lei, J., ... & Oostenbrink, R. (2013). Clinical prediction model to aid emergency doctors managing febrile children at risk of serious bacterial infections: diagnostic study. *British Medical Journal*, 346, f1706.

Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. *The Journal of Finance*, *53*(6), 1887-1934.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of experimental psychology*, *72*(3), 346.

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. *BMC medical informatics and decision making*, *16*(1), 138.

Simel, D.; Rennie, D. (2008). The Rational Clinical Examination: Evidence-Based Clinical Diagnosis. New York, NY: McGraw-Hill.

Van den Bruel, A., Thompson, M. J., Haj-Hassan, T., Stevens, R., Moll, H., Lakhanpaul, M., & Mant, D. (2011). Diagnostic value of laboratory tests in identifying serious infections in febrile children: systematic review. *British Medical Journal*, 342, d3082.

Van den Bruel, A., Haj-Hassan, T., Thompson, M., Buntinx, F., Mant, D., & European Research Network on Recognising Serious Infection investigators. (2010). Diagnostic value of clinical features at presentation to identify serious infection in children in developed countries: a systematic review. *The Lancet*, 375(9717), 834-845.

Woolley, A., & Kostopoulou, O. (2013). Clinical intuition in family medicine: more than first impressions. *The annals of family medicine*, *11*(1), 60-66.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, 6.

Supplementary materials

Supplementary results

Proof

We investigated how and when probability distortion could impair diagnostic accuracy. We argued that probability distortion in the medical evidence is not a sufficient condition. Instead, we showed that the detrimental effect might occur when physicians have to combine multiple sources of information. In such situation, probability distortion in the medical evidence should bias the aggregation of the information, which should eventually affect diagnostic accuracy.

We develop a Bayesian model according to which physicians revise their prior $p(BI)$ using two independent signals. The first signal is a signal e taking values in E and corresponding to the medical predictors. The second signal o takes values in O and captures some other sources of information. The signals are noisy and are defined by objective probabilities $p_{obj}(e,o|h)$ for $h = BI, \neg BI$. We assume that signals are independent: $p_{obj}(e,o|h) = p_{obj}(e|h) * p_{obj}(o|h)$.

Physicians hold subjective probabilities $p_{subj}(e|h)$ and $p_{subj}(o|h)$. Written in a log likelihood form, belief revision follows Bayes rule:

$$\log\left(\frac{p(I|e,o)}{p(\neg I|e,o)}\right) = \log\left(\frac{p(I)}{p(\neg I)}\right) + \log\left(\frac{p_{subj}(e|I)}{p_{subj}(e|\neg I)}\right) + \log\left(\frac{p_{subj}(o|I)}{p_{subj}(o|\neg I)}\right)$$

We call $w_e = \log\left(\frac{p(e|I)}{p(e|\neg I)}\right)$ the weight of evidence for signal e and $w_o = \log\left(\frac{p(o|I)}{p(o|\neg I)}\right)$ the weight of evidence for signal o.

We assume that physicians' subjective beliefs are correct for signal o but not for signal e. In particular, they display probability distortion with an evidence ratio ER defined as:

$$ER(e) = \frac{w_e^{subj}}{w_e^{obj}}$$

Which is such that $\forall\ e\colon 0 \leq ER(e) \leq 1$

To measure how well posterior beliefs predict BI we consider the area under the ROC curve. The ROC curve C is defined by using posterior beliefs as a classifier:

for all $x \in \mathbb{R}$, define $F(x) \subset E \times O$ as $F(x) = \{(e,o)|w_e^{subj} + w_o > x\}$ and define the point $p(x) \in C$ as the point of coordinates (a,b) such that a $= 1 - p(E \times O \backslash F(x)|\neg BI)$ (i.e: 1-specificity) and b $= p(F(x)|BI)$ (i.e: sensibility).

To understand why probability distortion is not a sufficient condition per se to impact AUC, consider a simple situation where the signal *o* is absent and where the evidence ratio has a constant value: $\forall\ e\colon ER(e) = c > 0$. It is clear that the ROC curve remains the same whatever the value of c since c does not modify how the values e are classified.

But in the general case with two signals, probability distortion might lead to misperceive the total value of the evidence. For example, consider that objectively the clinical case (e_1, o_1) is more severe than the clinical case (e_2, o_2) because $w_{e_1} + w_{o_1} > w_{e_2} + w_{o_2}$. However, the physician might consider that it is less severe if $c.w_{e_1} + w_{o_1} < c.w_{e_2} + w_{o_2}$. The following proof is based on a generalization of this illustrative example. We now show that with two distinct sources of information, posterior beliefs accuracy to predict BI should decrease with increasing probability distortion.

Proof: Consider 2 physicians A and B with the same prior and sets of evidence and such that A has more probability distortion than B. Let $x \in \mathbb{R}$, and define respectively $F_A(x) = \{(e,o)|w_{e,A}^{subj} + w_{o,A} > x\}$ and $F_B(x) = \{(e,o)|w_{e,B}^{subj} + w_{o,B} > x\}$. Let x and x' be such that $p(F_A(x)|BI) \geq p(F_B(x')|BI)$, then let us show that $1 - p(E \times O \backslash F_A(x)|\neg BI) \geq 1 - p(E \times O \backslash F_B(x')|\neg BI)$. This means that if at the respective classifying value x and x', the sensitivity for agent A is higher, then necessary his specificity is lower. This observation implies that the ROC curve for physician A lies necessary under the ROC curve for physician B.

Define $G = F_A(x)\backslash F_B(x')$ and $H = F_B(x')\backslash F_A(x)$. Thus $p(F_A(x)|BI) - p(F_B(x')|BI) = p(G|BI) - p(H|BI)$ while $p(E \times O\backslash F_B(x')|\neg BI) - p(E \times O\backslash F_A(x)|\neg BI) = p(G|\neg BI) - p(H|\neg BI)$. By definition:

$$G = \{(e,o)|w_{e,A}^{subj} + w_{o,A} > x \ and \ w_{e,B}^{subj} + w_{o,B} < x'\}$$

$$H = \{(e,o)|w_{e,A}^{subj} + w_{o,A} < x \ and \ w_{e,A}^{subj} + w_{o,A} > x'\}$$

Since $\forall\ e,\ 0 \leq \frac{w_{e,A}^{subj}}{w_e^{obj}} \leq \frac{w_{e,B}^{subj}}{w_e^{obj}} \leq 1$, then $\forall\ (e,o) \in \mathrm{G}$ and $\forall\ (e',o') \in \mathrm{H}$, $w_e^{obj} + w_o < w_{e'}^{obj} + w_{o'}$. Therefore, $\log\left(\frac{p(G|BI)}{p(G|\neg BI)}\right) < \log\left(\frac{p(H|BI)}{p(H|\neg BI)}\right)$. So if $p(G|BI) - p(H|BI) > 0$, then since $\log\left(\frac{p(G|BI)}{p(H|BI)}\right) < \log\left(\frac{p(G|\neg BI)}{p(G|\neg BI)}\right)$, we also have $p(G|\neg BI) - p(H|\neg BI) > 0$ which completes the proof.

The hypothesis that there is no probability distortion on the second source of information is certainly unrealistic. We think that the result still holds if we relax this hypothesis.

Table S1: Predictor variables
Values are numbers (percentage)
Unless stated otherwise by*: mean (sd)

	N	No	(%)
Demographic and neonatal data			
>30 days	1848	1483	(0.80)
Male sex	1848	1100	(0.60)
GBS screening at 8 months of pregnancy	1199	973	(0.81)
Detection of GBS	864	149	(0.17)
Premature rupture of membranes >12 hours	1729	181	(0.10)
Vaginal delivey	1813	1300	(0.72)
Maternal fever >38°C during labor &/ delivery	1788	77	(0.04)
NPS	1335	385	(0.29)
Detection of bacteria in NPS	347	68	(0.20)
Length of pregnancy (weeks of amenorrhea)*	1768	38.98	(1.53)
Neonatal fever	1811	37	(0.02)
Neonatal parenteral antibiotic treatment	1802	55	(0.03)
Neonatal oral antibiotic treatment	1801	15	(0.01)
Medical history			
Admission to the ED in August	1848	81	(0.04)
Duration of fever (hours)*	1795	13.96	(20.45)
Maximum temperature*	1835	38.65	(0.50)
Fever during the first 48 hours after vaccination	1837	65	(0.04)
Breast feeding at the admission	1827	843	(0.46)
Decrease in food consumption	1842	889	(0.48)
Diminished alertness	1843	292	(0.16)
Deterioration of general conditions	1842	567	(0.31)
Hypotonia, hyporesponsiveness	1841	400	(0.22)
Chills	1831	91	(0.05)
Another family member has fever	1812	513	(0.28)
Cough	1843	702	(0.38)
Breathing difficulty	1845	253	(0.14)
Rhinitis	1844	877	(0.48)
Vomiting	1843	299	(0.16)
Diarrhea	1842	259	(0.14)
Clinical examination			
Rectale temperature (°C)*	1843	37.98	(0.71)
Heart rate (/min)*	1802	158.59	(21.31)
Respiratory rate (/min)*	1351	43.68	(12.12)
Weight (g)*	1837	4833.03	(929.69)
Oxygen saturation measurement (%)*	1187	98.99	(1.55)
Systolic blood pressure (mmHg)*	751	94.34	(15.32)
Diastolic blood pressure (mmHg)*	747	55.40	(12.72)
Abnormal respiration	1839	256	(0.14)
Poor peripheral perfusion	1844	316	(0.17)

Irritability	1842	873	(0.47)
Weak or absent cry	1842	134	(0.07)
Poor or absent response to family members	1846	225	(0.12)
Moaning	1841	216	(0.12)
Poor or absent tone	1846	172	(0.09)
Poor or absent spontaneous motor skills	1846	59	(0.03)
Poor or absent eye contact	1842	101	(0.05)
Moderate or weak general appearance	1844	366	(0.20)
Abnormal lung auscultation	1846	207	(0.11)
Abnormal capillary refill time	1838	49	(0.03)
Abnormal anterior fontanelle	1844	35	(0.02)
Clinical signs of dehydration	1845	17	(0.01)
Rash	1843	112	(0.06)
Joint anomaly	1846	2	(0.00)
Erythemathous throat	1831	139	(0.08)
Rhinitis	1843	804	(0.44)
Acute otitis media	1840	284	(0.15)
Lymphadenopathy	1846	3	(0.00)
Diarrhea	1845	214	(0.12)
Vomiting	1844	167	(0.09)
Hepatomegaly and/ or splenomegaly	1846	14	(0.01)
Phimosis (boys)	1014	140	(0.14)
Circumcision (boys)	1022	31	(0.03)
Laboratory tests			
WBC (/mmm3)*	1833	10917.78	(5200.71)
Lymphocyte (/mmm3)*	1799	4662.20	(2558.21)
Myelaemia	1786	73	(0.04)
C- reactive protein (mg/L)*	731	43.86	(46.42)
Capillary blood glucose (mmlo/L)*	548	5.18	(1.32)
Granulocyte neutrophile (/mmm3)*	1803	4604.60	(3371.77)
Monocyte (/mmm3)*	1772	1382.41	(943.26)
Platelets (/mmm3)*	1812	406096.64	(123983.97)
Fibrinogen (g/L)*	113	3.97	(1.46)
Blood lactate (mmlo/L)*	48	2.91	(1.46)
Blood culture	1848	1068	(0.58)
CSF analysis	1637	1030	(0.63)
WBC count in CSF (/mmm3)*	1148	73.08	(423.91)
RBC count in CSF (/mmm3)*	1101	2523.97	(24754.70)
CSF Gram stain	1219	7	(0.01)
UDT n°1	1827	1499	(0.82)
CBEU n°1	1845	1189	(0.64)
Nitrites in UDT n°1	1476	141	(0.10)
WBC in UDT n°1	1489	380	(0.26)
WBC count in CBEU n°1 (/ml)*	1147	61830409.26	(7.79e+08)
RBC count in CBEU n°1 (/ml)*	1106	1.82e+08	(6.01e+09)
CBEU Gram stain n°1	999	322	(0.32)

Urine analysis positive n°1	1741	335	(0.19)
UDT n°2	1480	58	(0.04)
CBEU n°2	1486	205	(0.14)
Nitrites in UDT n°2	54	6	(0.11)
WBC in UDT n°2	57	24	(0.42)
WBC count in CBEU n°2 (/ml)*	177	730378.64	(3866341.69)
RBC count in CBEU n°2 (/ml)*	165	6192698.24	(77846569.44)
CBEU Gram stain n°2	157	60	(0.38)
Urine analysis positive n°2	168	62	(0.37)
NBS	1794	21	(0.01)
Cells in the NBS	12	6	(0.50)
Granulocyte in the NBS	13	6	(0.46)
NBS Gram stain	13	5	(0.38)
Nasopharyngeal search for viral infection	1835	556	(0.30)
Detection of a virus in nasopharyngeal search	286	79	(0.28)
Stool test	1820	125	(0.07)
Virological analysis of stool sample	1806	167	(0.09)
Detection of a virus in stool sample	52	11	(0.21)
Additional sampling n°1	1848	102	(0.06)
Additional sampling n°2	1848	14	(0.01)
Additional sampling n°3	1848	7	(0.00)
Chest radiography	1832	1194	(0.65)
Pulmonary hyperinflation	1194	66	(0.06)
Alveolar consolidation	1194	56	(0.05)
Pleural effusion	1194	0	(0.00)
Bronchial syndrome	1194	214	(0.18)
Interstitial syndrome	1194	6	(0.01)
Atelectasis	1194	12	(0.01)
Additional radiography n°1	1848	35	(0.02)
Additional radiography n°2	1848	2	(0.00)
Additional radiography n°3	1848	0	(0.00)

N: Total number of available data

Abbreviations: CBEU, cyto- bacteriological examination of the urines; CSF, cerebrospinal fluid; GBS, group B streptococcus; NBS, nasopharyngeal bacterial sampling; NPS, neonatal peripheral samplings; RBC, red blood cell; UDT, urine dipstick test; WBC, white blood cell.

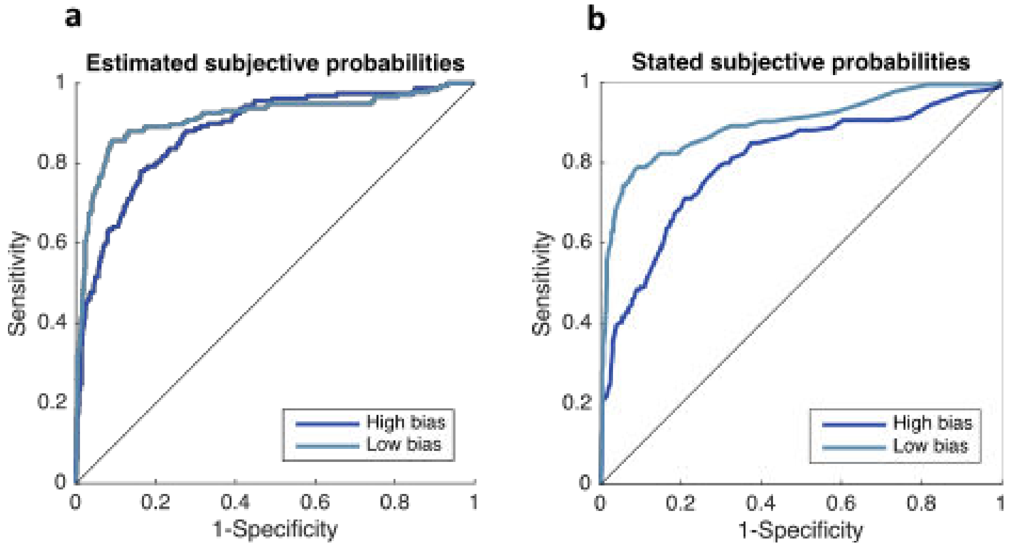


Figure S1: *Receiver operating characteristic curves to detect BI by high and low bias group for* ***(a)*** *estimated subjective probabilities;* ***(b)*** *stated subjective probabilities.*

Chapter 2: Mathematics for probability distortion analysis

Marine Hainguerlot[1] and Jean-Christophe Vergnaud[1]

[1]Centre d'Economie de la Sorbonne CNRS UMR 8174, Paris, France.

Abstract

In the previous chapter, we presented and used several tools to analyze clinical judgment and decision making. We compared the linear judgment with the linear model in order to measure how evidence was integrated. We represented probability distortion through the comparison of the predicted probabilities of the linear judgment versus the predicted probabilities of the linear model. We also run ROC curve analyses. Overall, all these analytical tools are related and share some mathematical links that we did not explicit. Thus it could be enlightening to explain these mathematical links to better understand how biases in evidence integration transfer to distorted probabilities and to the Area under the ROC curve. We will explicit these links in the second part of this chapter. First, we would like to start this chapter with some discussions about probability distortion. In chapter 1, probability distortion was central in our analysis but we were not very explicit about how our approach is connected to the literature on probability distortion. Thus, we first relate our approach to other approaches.

Probability distortion in judgment: a single representation for different phenomena

A common S- shaped pattern…

In judgment, humans need to estimate the true probability of an event. It is typically observed that humans distort the true probability estimate of an event in a non- linear way. Their subjective probabilities systematically deviate from the objective probabilities in an S-shaped pattern (i.e. underestimation of small probabilities and overestimation of large probabilities) or an inverted S- shaped pattern (i.e. overestimation of small probabilities and underestimation of large probabilities).

Similar patterns of S- shaped distortion have been found with tasks of frequency estimation and confidence ratings (for a review, see Zhang & Maloney, 2012). To model probability distortion, a linear transformation of the log odds of probability can well capture the S-shaped pattern with two parameters (Gonzalez & Wu, 1999), as follows:

$$\log\left(\frac{p_s}{1-p_s}\right) = \beta_0 + \beta_1 \log\left(\frac{p_o}{1-p_o}\right)$$

where p_s is the subjective probability and p_o is the objective probability.

As demonstrated by Gonzalez & Wu (1999), p_s can be rewritten as:

$$p_s = \frac{\delta {p_o}^{\beta_1}}{\delta {p_o}^{\beta_1} + (1-p_o)^{\beta_1}}$$

where $\delta = \exp(\beta_0)$.

The first panel of **Figure 1** shows an example of inverted S shaped distortion from frequency estimation. The second panel of **Figure 1** plots the linear log odds transformation of the same data. This example is taken from the study of Zhang & Maloney (2012) who re- plotted data from Attneave (1953) where participants had to estimate the relative frequencies of occurrence of English letters in text. Attneave found that the frequency of rare letters was overestimated whereas the frequency of common letters was underestimated.

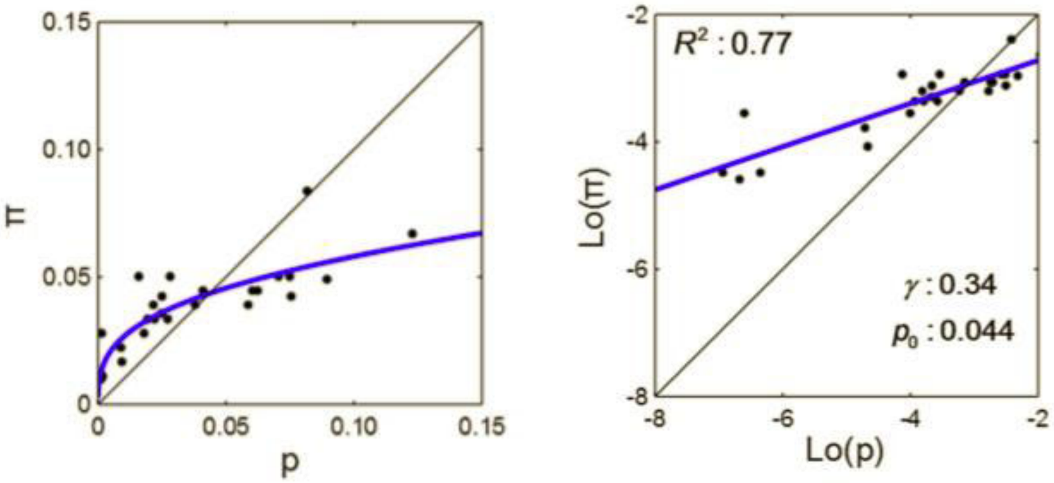


Figure 1: (First panel) *Estimated relative frequencies* (π) *versus actual frequencies* (p) *from Attneave (1953).* ***(Second panel)*** *Log odds of estimated relative frequencies* $lo(\pi)$ *versus log odds of actual frequencies* $lo(p)$ *from Attneave (1953). From "Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition" by Zhang, H., & Maloney, L. T. (2012). Frontiers in Neuroscience, 6.*

The parameter β_1 is the slope of the linear transformation, it controls for the curvature of the probability function. The parameter β_0 is the intercept which controls for the elevation of the probability function. When β_1=1 and β_0=0, there is no distortion. When β_1>1, the distortion is S- shaped. When β_1<1, the distortion is inverted S- shaped. **Figure 2** shows how the parameters β_1 and β_0 respectively affect the curvature and the elevation. This illustration is taken from Gonzalez & Wu (1999).

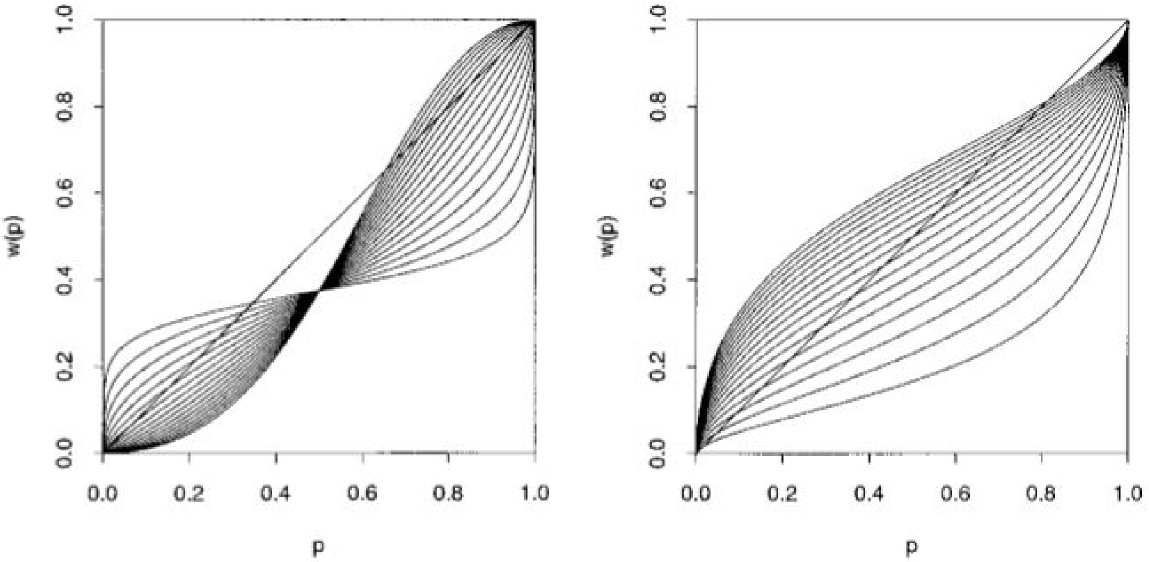


Figure 2: *Effect of β_0 on the elevation and β_1 on the curvature of the curve.* ***(First panel)*** *fixes β_0 at $ln(0.6)$ and varies β_1 between 0.2 and 1.8.* ***(Second panel)*** *fixes β_1 at 0.6 and varies β_0 between $ln(0.2)$ and $ln(1.8)$. From "On the shape of the probability weighting function" by Gonzalez, R., & Wu, G. (1999). Cognitive psychology, 38(1), 129-166.*

…but different phenomena

In the judgment literature, probability distortion has been studied through the comparison of subjective probabilities with objective probabilities. But, depending on the task, two parallel approaches coexist. In confidence ratings tasks, participants are asked to report their probability of success (ie subjective probability). Then an operationalized probability can be computed as the average success rate for a given level of confidence (Lichtenstein et al, 1977) and a calibration curve is plotted. In frequency estimation and belief revision tasks, objective probabilities are computed independently of subjective probabilities. For instance, in tasks where subjects are asked to revise an initial prior after the observation of data, an objective probability can be calculated by means of Bayes rule (Edwards, 1968). Overall, in confidence rating tasks subjective probabilities are more extreme than the associated relative frequencies (S shaped) while in revision of opinion and frequency estimation tasks, subjective probabilities are less extreme than objective probabilities (inverted S Shaped). Our results from chapter 1 are consistent with the later literature.

Probability distortion seems to follow different patterns according to the nature of the judgment task. Interestingly, Erev, Wallsten and Budescu (1994) questioned this conclusion by running simultaneously the two analyses. For the same set of data, they observed both an S shaped pattern with the calibration curve analysis and an inverted S shaped pattern when plotting subjective probability against objective probability. They concluded that S shaped or inverted S shaped effects could be artifacts of the methods used to analyze the data.

Probability distortion in our data

In chapter 1, we found an inverted S- shaped pattern when plotting predicted subjective probabilities ("linear judgment") versus predicted objective probabilities ("linear model") (see First panel of **Figure 3**). In this section, we report the calibration analysis of our data (see Second panel of **Figure 3**). First, we can observe that the linear model is well calibrated. We still observe an inverted S shaped pattern for the linear judgment. Subjective probabilities are always too high whatever their value. Overall, for the same set of data, an inverted S shape pattern was observed with the calibration curve analysis or the method that we used in chapter 1. The pattern seems to be robust in our data.

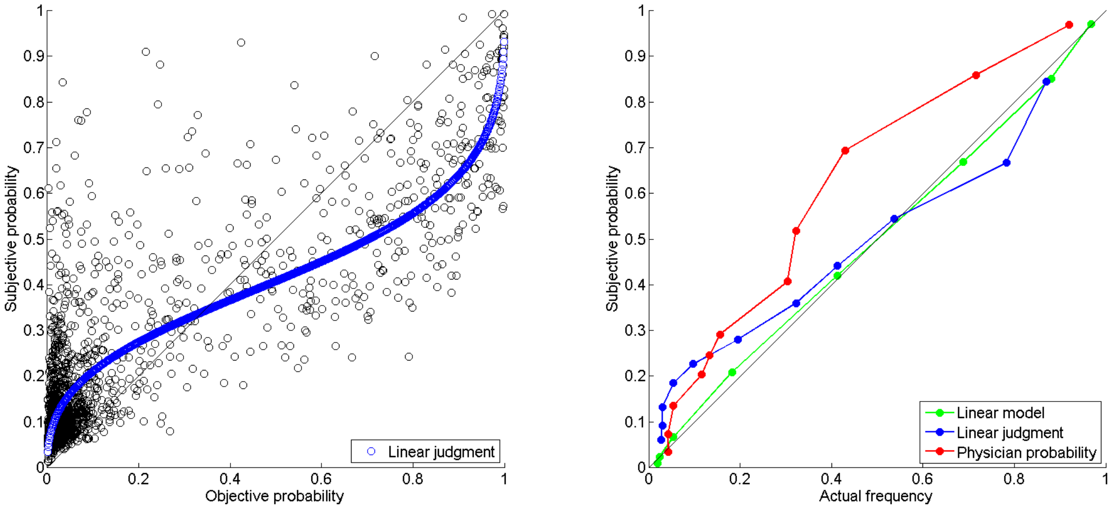


***Figure 3: (First panel)** Probabilities predicted by linear judgment versus probabilities predicted by linear model. **(Second panel)** Probabilities predicted by linear model (green line), probabilities predicted by linear judgment (blue line), physician probability judgment (red line) versus actual frequencies.*

Miscellaneous

It is important to note that this thesis focuses on probability distortion in judgment. A similar pattern of distortion has also been observed in decision under risk or uncertainty. In Prospect Theory (Kahneman & Tversky, 1979) and rank dependent expected utility (Quiggin, 1982), it is assumed that the objective probabilities provided to the decision-maker are transformed into decision weights through a nonlinear weighting function. It is typically observed that decision weights follow an inverted S shaped pattern (Wu & Gonzalez, 1996). Importantly, one needs to recognize that the meaning of the distortion is different. In tasks of judgment, probability distortion corresponds to a distorted estimate of the true probability; whereas in tasks of decision, it corresponds to a distorted weight of the subjective probability. Interestingly, in settings where subjects are not informed about the probabilities attached to the lotteries but learn them by experience in a sequence of repeated choices, an S-shaped was observed for the weighting function (Hertwig & Erev, 2009). In experience based decision, these decision weights also capture a judgmental aspect related to the perception of frequencies.

Mathematical links between integration of evidence, probability distortion and Area under the ROC curve

We model the situation of a physician who has to detect and treat infection in a population at risk. We first state the stochastic model of the environment and then analyze how the physician forms subjective probability about the health state of a patient according to the medical information she collected.

Model of the environment[7]

We model a population at risk of bacterial infection through a simple Bayesian model. The population is composed of patients who may be infected $H = h$ or not infected $H = \bar{h}$ according to some prior probability $P^o_{prior}(h)$.

For each patient, some cues X_i are randomly drawn according to conditional probabilities $P^o(X_i|H)$. We suppose that these cues take two values $X_i = x_i^+, x_i^-$[8] and by convention, we state that $P^o(x_i^+|h) \geq P^o\left(x_i^+|\bar{h}\right)$.

Ideal and subjective integration of evidence

We now take the point of view of the physician who needs to form her beliefs on one patient. To apply Bayes rule, we now work in log odds.

[7] In this theoretical section, we present the model of the environment (model of the physician was presented in the introduction). We also use different mathematical notations.

[8] In the introduction, we used the notation $x_i = 0,1$.

First let us examine how an ideal Bayesian observer would calculate her posterior probability $P^o(H|X_1, \dots, X_i)$. Let denote $lp = \log\left(\frac{P^o_{prior}(h)}{1-P^o_{prior}(h)}\right)$ the log odds of the prior. To evaluate the evidence provided by the cues, the ideal Bayesian observer will calculate the weight of evidence[9] W_i.

With $W_i^+ = \log\left(\frac{P^o(x_i^+|h)}{P^o(x_i^+|\bar{h})}\right)$ if the observer receives the cue x_i^+ and $W_i^- = \log\left(\frac{P^o(x_i^-|h)}{P^o(x_i^-|\bar{h})}\right)$ if the observer receives the cue x_i^-.

A cue is relevant if $P^o(x_i^+|h) > P^o(x_i^+|\bar{h})$ => $W_i^+ > 0 > W_i^-$.

A cue is irrelevant if $P^o(x_i^+|h) = P^o(x_i^+|\bar{h})$ => $W_i^+ = W_i^- = 0$.

Let denote $lo = \log\left(\frac{P^o(h|X_1,\dots,X_i)}{P^o(\bar{h}|X_1,\dots,X_i)}\right)$ the log odds of the posterior. In log odds terms, the Bayes rule is written as:

$$lo = lp + \sum_i W_i$$

The physician may differ from the ideal Bayesian observer by the values she used in her integration of the evidence. Thus, the log odds of the posterior subjective probability can be written as[10]:

$$ls = \tilde{l}p + \sum_i \breve{W}_i$$

Deriving the mathematical links between integration of evidence and probability distortion

[9] In the introduction we used the notation $W^o_{x_i}$.

[10] This is a simplified version of the model of the physician defined in the introduction (eq1). We do not consider intuition and do not distinguish between weight of evidence and the parameter of integration. Thus $\breve{W}_i$ corresponds to $\alpha W^s_{x_i}$ in eq 1.

In this section, we want to express the linear transformation of the log odds of probability that is commonly used in the probability distortion literature with respect to our model of integration of the evidence. As mentioned earlier, the transformation is typically expressed with two parameters β_0 for the elevation and β_1 for the slope, as follows:

$$ls = \beta_0 + \beta_1 \, lo$$

Elevation corresponds to what extent subjective probabilities are higher/ lower than objective probabilities. The slope corresponds to how sensitive subjective probabilities are to a change in objective probabilities.

Hereafter, we shall identify the values of β_0 and β_1 with respect to our model: How does integration of evidence produce probability distortion?

We consider the relationship between lo and ls for a dataset generated by the environment and the physician.

One relevant cue example

Let's consider a very simple case with only one relevant cue X_1 in the environment. Then the physician's subjective probabilities only take two values, one when the cue is positive:

$$ls^+ = \tilde{l}p + \breve{W}_1^+$$

and one when the cue is negative:

$$ls^- = \tilde{l}p + \breve{W}_1^-$$

The corresponding objective probabilities are $lo^+ = lp + W_1^-$ and $lo^- = lp + W_1^-$.

In our dataset, we observe two points (lo^+, ls^+) and (lo^-, ls^-) corresponding respectively to x_1^+ and x_1^-. **Figure 3** illustrates this example.

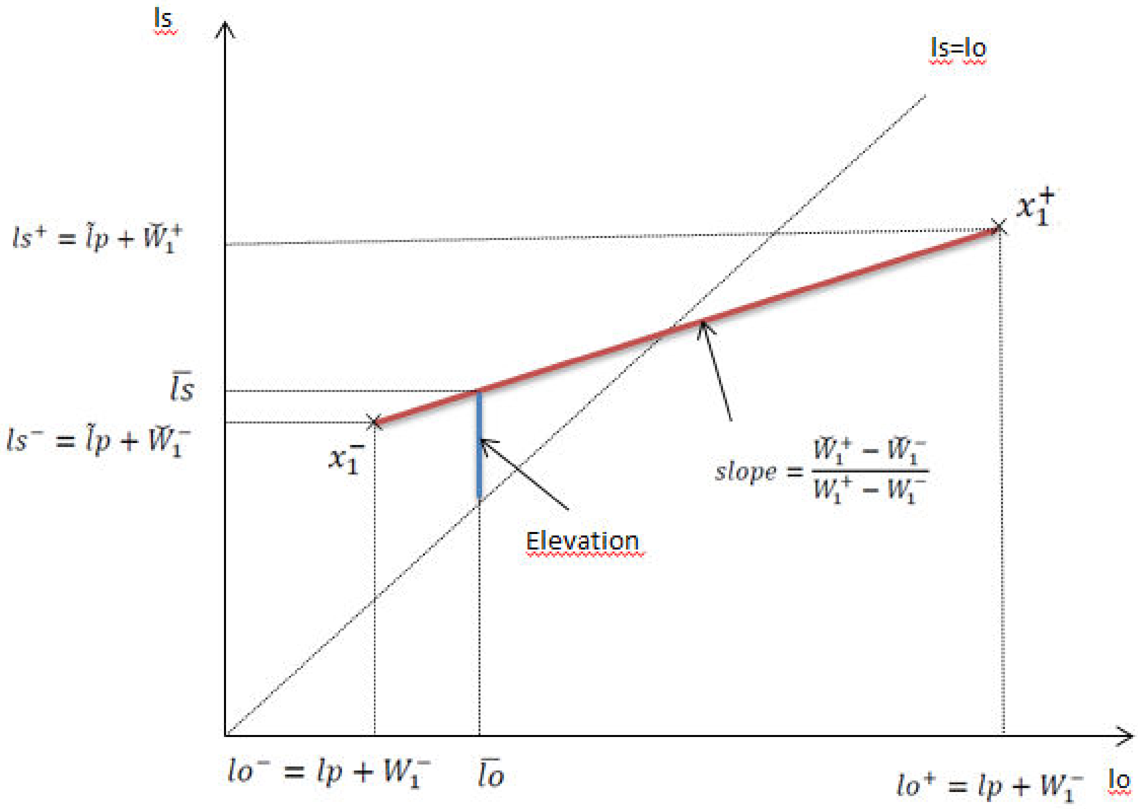

ls
ls=lo
$ls^+ = \tilde{l}p + \tilde{W}_1^+$
x_1^+
$\overline{ls}$
$ls^- = \tilde{l}p + \tilde{W}_1^-$
x_1^-
$slope = \frac{\tilde{W}_1^+ - \tilde{W}_1^-}{W_1^+ - W_1^-}$
Elevation
$lo^- = lp + W_1^-$
$\overline{lo}$
$lo^+ = lp + W_1^-$
lo

***Figure 3:** Diagram of the log odds of the objective probabilities (lo) versus the log odds of the subjective probabilities (ls) for one relevant cue X_1.*

The probability of observing x_1^+ is:

$$P^o(x_1^+) = P^o(x_1^+|h)p(h)\text{+ } P^o\left(x_1^+\middle|\bar{h}\right)p(\bar{h})$$

Within our framework, we propose to define elevation as the difference between the mean subjective probability $\overline{ls} = P^s(x_1^+)ls^+ + P^s(x_1^-)ls^-$ and the mean objective probability $\overline{lo} = P^o(x_1^+)lo^+ + P^o(x_1^-)lo^-$.

$$elevation = \overline{ls} - \overline{lo}$$

This difference can be rewritten as follows:

$$elevation = \tilde{l}p - lp + (P^o(x_1^+)\check{W}_1^+\text{+ } P^o(x_1^-)\check{W}_1^-) - (P^o(x_1^+)W_1^+ + P^o(x_1^-)W_1^-$$

Two elements contribute to the elevation:

- a prior bias $\tilde{l}p - lp$: the difference between the subjective prior $\tilde{l}p$ and the objective prior lp
- a mean weight of evidence bias: $(P^o(x_1^+)\breve{W}_1^+ + P^o(x_1^-)\breve{W}_1^-) - (P^o(x_1^+)W_1^+ + P^o(x_1^-)W_1^-)$

The slope is equal to the ratio $\frac{ls^+ - ls^-}{lo^+ - lo^-}$, which can be rewritten as:

$$slope = \frac{\breve{W}_1^+ - \breve{W}_1^-}{W_1^+ - W_1^-}$$

Note that $W_1^+ - W_1^-$ is the log of the odds ratio. Therefore, the slope corresponds to the relative value of the log of the subjective odds ratio with respect to the log of the objective odds ratio (see introduction). If the physicians under weight the cue, as we found in our data, then the slope is below 1 leading to an inverted S-shaped.

Note that the slope is also equal to $\frac{ls - \overline{ls}}{lo - \overline{lo}}$. Thus, we can write the probability distortion as:

$$ls = \overline{lo} + elevation + slope(lo - \overline{lo})$$

Two relevant cues example

How does this formula generalize to more than one relevant cue?

We shall demonstrate that the relationship between ls and lo will not be linear anymore as long as the relative value of the log of the subjective odds ratio with respect to the log of the objective odds ratio $\frac{\breve{W}_i^+ - \breve{W}_i^-}{W_i^+ - W_i^-}$ differ between the cues X_i.

Figure 4 illustrates the non-linear relationship for two cues X_1 and X_2 with $\frac{\breve{W}_1^+ - \breve{W}_1^-}{W_1^+ - W_1^-} < \frac{\breve{W}_2^+ - \breve{W}_2^-}{W_2^+ - W_2^-}$. There will be four cues value combinations: (x_1^+, x_2^+) (x_1^+, x_2^-), (x_1^-, x_2^+) and (x_1^-, x_2^-). The four points in our dataset form a parallelogram.

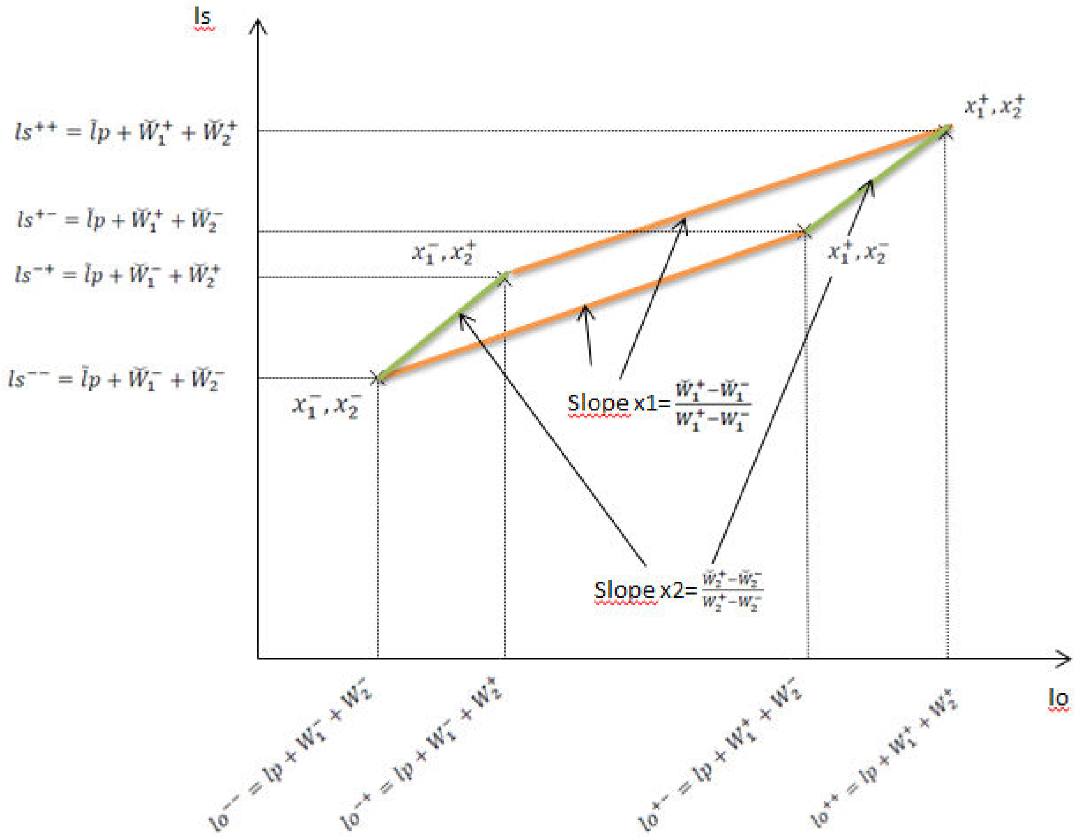

ls
$ls^{++} = \breve{l}p + \breve{W}_1^+ + \breve{W}_2^+$
$ls^{+-} = \breve{l}p + \breve{W}_1^+ + \breve{W}_2^-$
$ls^{-+} = \breve{l}p + \breve{W}_1^- + \breve{W}_2^+$
$ls^{--} = \breve{l}p + \breve{W}_1^- + \breve{W}_2^-$
x_1^+, x_2^+
x_1^-, x_2^+
x_1^+, x_2^-
x_1^-, x_2^-
Slope x1= $\frac{\breve{W}_1^+ - \breve{W}_1^-}{W_1^+ - W_1^-}$
Slope x2= $\frac{\breve{W}_2^+ - \breve{W}_2^-}{W_2^+ - W_2^-}$
lo
$lo^{--} = lp + W_1^- + W_2^-$
$lo^{-+} = lp + W_1^- + W_2^+$
$lo^{+-} = lp + W_1^+ + W_2^-$
$lo^{++} = lp + W_1^+ + W_2^+$

Figure 4: *Diagram of the log odds of the objective probabilities* (lo) *versus the log odds of the subjective probabilities* (ls) *for two relevant cues* X_1 *and* X_2 *with* $\frac{\breve{W}_1^+ - \breve{W}_1^-}{W_1^+ - W_1^-} < \frac{\breve{W}_2^+ - \breve{W}_2^-}{W_2^+ - W_2^-}$.

As an example, the slope between the two points (x_1^+, x_2^+) and (x_1^-, x_2^+) is equal to $\frac{\breve{W}_1^+ - \breve{W}_1^-}{W_1^+ - W_1^-}$ while the slope between the two points (x_1^+, x_2^+) and (x_1^+, x_2^-) is equal to $\frac{\breve{W}_2^+ - \breve{W}_2^-}{W_2^+ - W_2^-}$. The two ratios are different thus the relationship between ls and lo cannot be linear.

More generally, we do not expect to observe that the data set will be organized on a line. The slope of a linear fit of the data will nevertheless be related to the ratio $\frac{\breve{W}_i^+ - \breve{W}_i^-}{W_i^+ - W_i^-}$ for each cue X_i. The lower these ratios are, the smaller should be the slope of the linear fit.

On the other hand, the generalization of the elevation term is straightforward. It is equal to the prior bias plus the sum of the mean weight of evidence bias of all the cues X_i.

Two cues example including one irrelevant cue

Let's consider a situation where in addition to the relevant cue X_1, there exists a second cue X_2 but this cue is irrelevant. **Figure 5** illustrates this example. The physician will observe four cue patterns with the corresponding subjective probabilities:

$$ls^{++} = \tilde{l}p + \breve{W}_1^{+} + \breve{W}_2^{+}$$

$$ls^{+-} = \tilde{l}p + \breve{W}_1^{+} + \breve{W}_2^{-}$$

$$ls^{-+} = \tilde{l}p + \breve{W}_1^{-} + \breve{W}_2^{+}$$

$$ls^{--} = \tilde{l}p + \breve{W}_1^{-} + \breve{W}_2^{-}$$

But the objective probabilities take only two values:

$$lo^{++} = lo^{+-} = lp + W_1^{+}$$

$$lo^{-+} = lo^{--} = lp + W_1^{-}$$

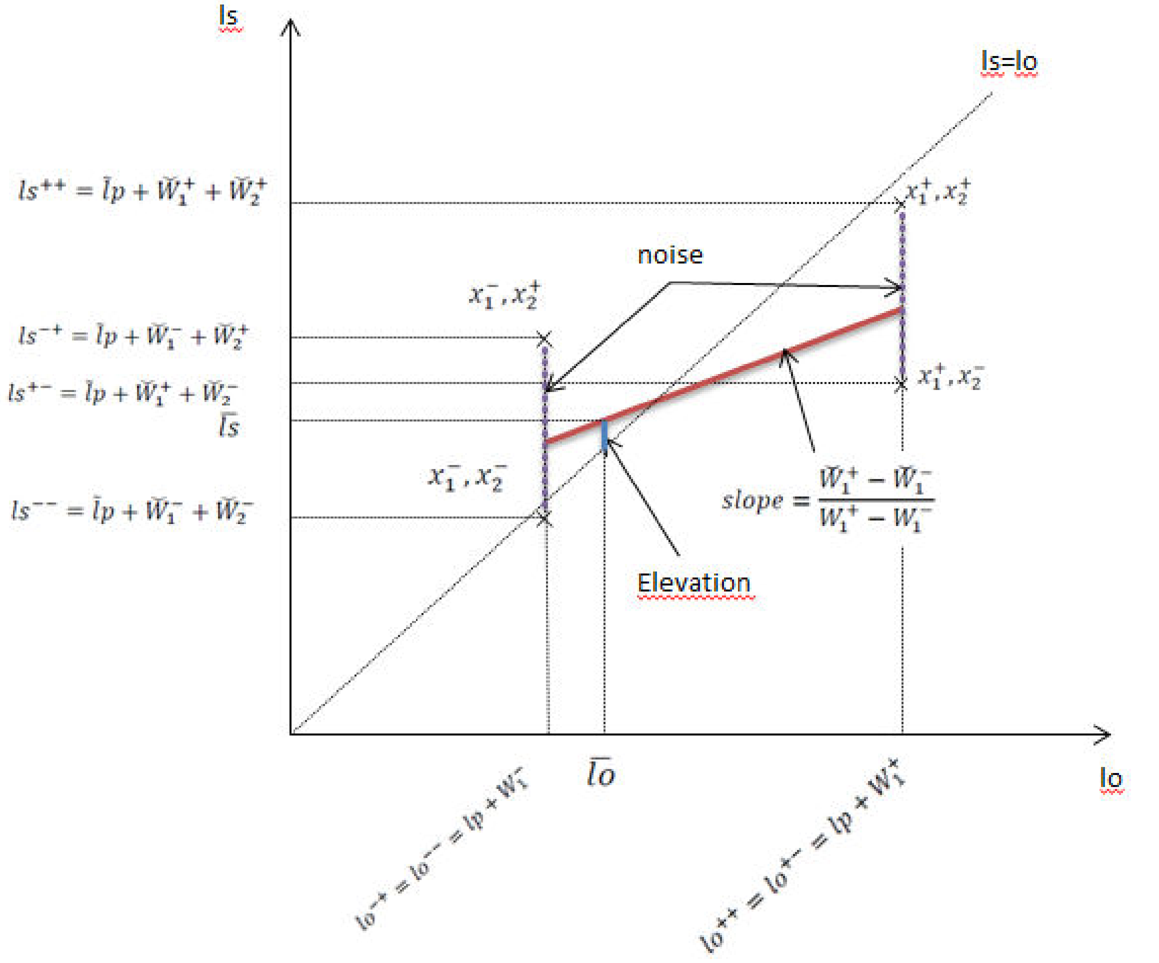

ls
ls=lo
$ls^{++} = \bar{l}p + \bar{W}_1^+ + \bar{W}_2^+$
x_1^+, x_2^+
noise
x_1^-, x_2^+
$ls^{-+} = \bar{l}p + \bar{W}_1^- + \bar{W}_2^+$
$ls^{+-} = \bar{l}p + \bar{W}_1^+ + \bar{W}_2^-$
x_1^+, x_2^-
$\overline{ls}$
x_1^-, x_2^-
$slope = \frac{\bar{W}_1^+ - \bar{W}_1^-}{W_1^+ - W_1^-}$
$ls^{--} = \bar{l}p + \bar{W}_1^- + \bar{W}_2^-$
Elevation
$\overline{lo}$
lo
$lo^{-+} = lo^{--} = lp + W_1^-$
$lo^{++} = lo^{+-} = lp + W_1^+$

Figure 5: *Diagram of the log odds of the objective probabilities* (lo) *versus the log odds of the subjective probabilities* (ls) *for one relevant cues* X_1 *and one irrelevant cue* X_2.

As can be observed in the figure, the data set (lo, ls) is not aligned, since we have two different subjective probabilities for the same objective probability.

In this section, we introduce a new component to describe the relationship between ls and lo when the physician is taking into account irrelevant cues. We call this component the *noise.*

We describe the probability distortion by including the noise as follows:

$$ls = \overline{lo} + elevation + slope(lo - \overline{lo}) + noise$$

where:

$$noise = \begin{cases} with\, probability\ P^o(x_2^+),\ noise = \ \breve{W}_2^+ - (P^o(x_2^+)\breve{W}_2^+ +\ P^o(x_2^-)\breve{W}_2^-) \\ with\, probability\ P^o(x_2^-),\ noise = \ \breve{W}_2^- - (P^o(x_2^+)\breve{W}_2^+ +\ P^o(x_2^-)\breve{W}_2^-) \end{cases}$$

The expected value of the noise is null.

Thus, if the physician attributes weight of evidence to an irrelevant cue, this mis weighting bias induces variation in subjective probabilities unrelated to variation in objective probabilities.

Probability distortion and ROC curve

In a detection problem, the physician has to estimate the probability that the patient has the disease or not. To classify patients in the "disease" or "no disease" group, we can apply a threshold on the physician judgment. Given the threshold we can construct the table below that describes the four possible outcomes in a detection problem.

		Patient condition	
		Disease	No disease
Physician judgment	Above threshold	True positive	False positive
	Below threshold	False negative	True negative

The ROC curve plots the true positive rate (or sensitivity) $\left(\frac{Number\ of\ True\ positive}{Number\ of\ Disease}\right)$ against the false positive rate (1-specificity) $\left(\frac{Number\ of\ False\ positive}{Number\ of\ No\ Disease}\right)$ at various thresholds. The AUC (Area under the ROC curve) measures the ability of the physician in detecting the presence or absence of the disease. When the AUC is equal to one, the physician has perfect detection ability; whereas when the AUC is equal to 0.5, the physician's ability is null.

What kind of probability distortion would impact the ROC curve?

Intuitively, we may think that any distortion bias in physician judgment could impair the AUC. But, it is not obvious that probability distortion should negatively impact the diagnostic accuracy (AUC) (as was mentioned in the appendix of Chapter 1). In this section, we show that what matters is how physician probabilities rank the patients in terms of risk of the disease compared to how objective probabilities rank the patients.

Let's examine a situation with two cues to understand this point. Let's suppose that the objective probabilities in log odds are as follows:

$$lo^{--} = lp + W_1^- + W_2^- < lo^{-+} = lp + W_1^- + W_2^+ < lo^{+-} = lp + W_1^+ + W_2^- < lo^{++} = lp + W_1^+ + W_2^+$$

Denote $po = \exp(lo)/(1+\exp(lo))$ the objective probability in absolute value. Thus:

$$po^{--} < po^{-+} < po^{+-} < po^{++}$$

We now develop the method to plot the ROC curve for objective probabilities (see **Figure 6**).

For $\delta > po^{++}$, none of the patients is classified in the "disease" group. Thus *sensitivity* and *1 - specificity* are null (which corresponds to the origin (0,0) on the ROC curve).

For $\epsilon\ ]po^{+-}; po^{++}\ ]$, only the patients with cues x_1^+, x_2^+ are classified into the « disease » group. Thus: *sensitivity* is equal to: $\frac{Number\ of\ patients\ with\ disease\ and\ cues\ x_1^+,x_2^+}{Number\ of\ patients\ with\ disease} = \frac{P(x_1^+,x_2^+)po^{++}}{p(h)}$ while *1 -specificity* is equal to: $\frac{Number\ of\ patients\ without\ disease\ and\ cues\ x_1^+,x_2^+}{Number\ of\ patients\ without\ disease}$ $= \frac{P(x_1^+,x_2^+)(1-po^{++})}{p(\bar{h})}$. Let's call (++) this point in the (sensitivity,1-specificty) representation.

The slope between the point (++) and the origin (0,0) is equal to: $\frac{sensitivity\ at{+}{+}}{1-specificity\ at{+}{+}} =$ $\frac{po^{++}/(1-po^{++})}{p(h)/p(\bar{h})}$.

For $\delta\ \epsilon\ ]po^{-+}; po^{+-}\ ]$, patients with the cues x_1^+, x_2^+ or x_1^+, x_2^- are classified into the "disease group". Similarly, we can obtain a second point (++,+-). The slope between (++,+-) and (++) is equal to $\frac{po^{+-}\ /(1-po^{+-})}{p(h)/p(\bar{h})}$.

Similarly, we can obtain the two last points in the ROC curve: (++,+-,-+) and the point (1,1).

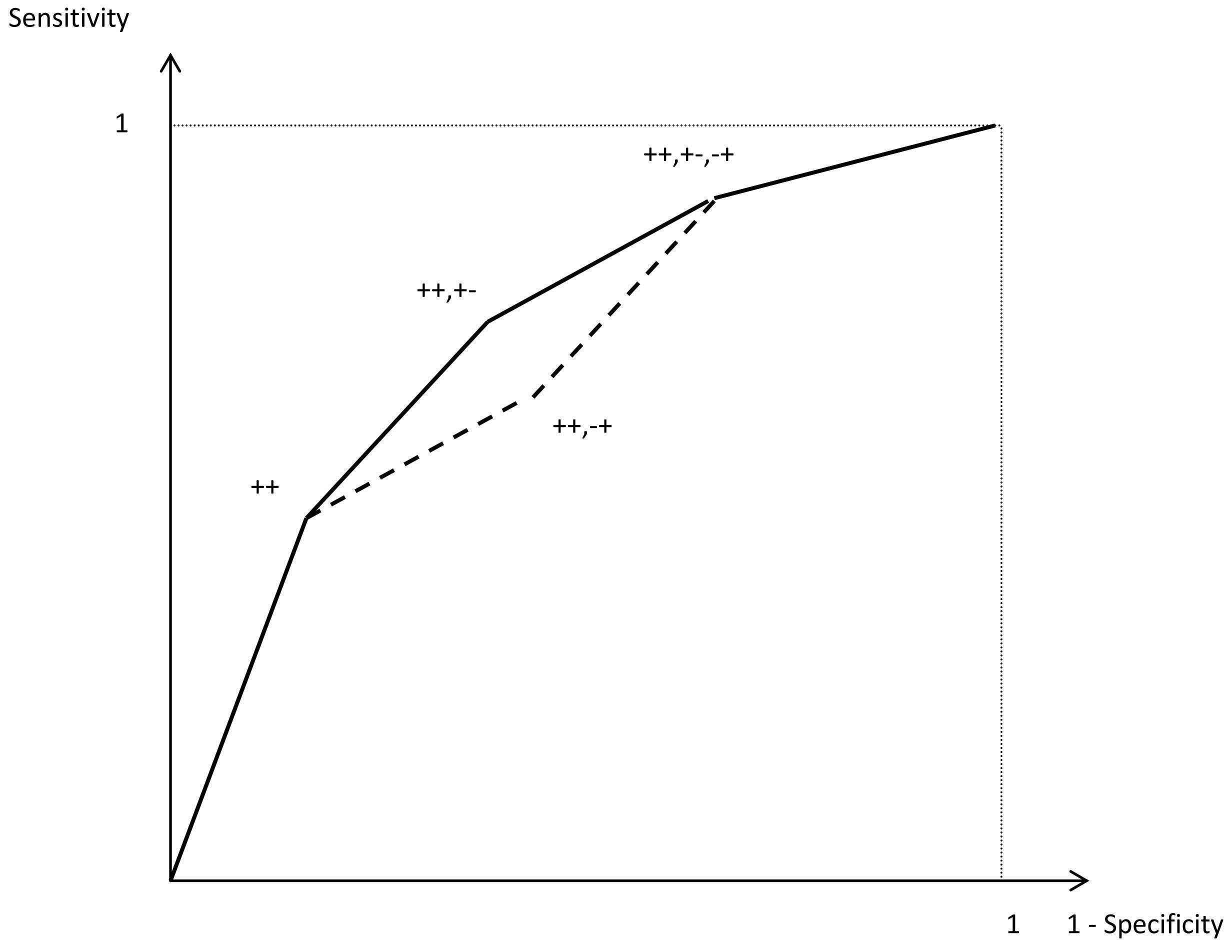


Figure 6: *ROC curves for objective probabilities (solid line) and subjective probabilities (dashed line)*

Similarly, we can plot the ROC curve for subjective probabilities. If subjective probabilities follow the same order than the objective probabilities then the ROC curve will be the same. Suppose now that the order in the subjective probabilities differs as follows:

$$ps^{--} < ps^{+-} < ps^{-+} < ps^{++}$$

According to this ranking, the first group of patients classified in the "disease" group are the ones with cues x_1^+, x_2^+. The second group of patients classified are the ones with cues x_1^+, x_2^+ or x_1^-, x_2^+. The same method is applied to classify the third and fourth groups of patients. Importantly, we observe that the second group of patients classified on the basis of subjective probabilities differs from the classification obtained with objective probabilities.

In terms of ROC curves, the second point for subjective probabilities (++,-+) will be different from the second point for objective probabilities (++,+-).

Note that the corresponding slope for subjective probabilities between the second point (++,-+) and (++) is equal to $\frac{po^{-+}/(1-po^{+-)}}{p(h)/p(\bar{h})}$. Critically, this slope is lower than the slope for objective probabilities between the second point (++,+-) and (++). Consequently, the subjective ROC curve is below the objective ROC curve, which translates into a lower AUC for subjective probabilities.

What is the source of a distortion in the ranking of probabilities?

Remind that, in the previous illustration, the objective probabilities were ranked as follows:

$$po^{--} < po^{-+} < po^{+-} < po^{++}$$

On the other hand, the subjective probabilities had the following order:

$$ps^{--} < ps^{+-} < ps^{-+} < ps^{++}$$

In terms of weights of evidence, the order of the probabilities correspond to:

$$po^{-+} < po^{+-} <=> W_1^+ - W_1^- > W_2^+ - W_2^-$$

$$ps^{+-} < ps^{-+} <=> \breve{W}_1^+ - \breve{W}_1^- < \breve{W}_2^+ - \breve{W}_2^-$$

Thus while the odds ratio for cue 1 is higher than for cue 2, the physician perceives the inverse.

This result suggests that if the physician misperceives the relative value of the cues in terms of odds ratio, his AUC will be impaired. Misattribution of evidence might be a source of such misperception.

References

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology*, *46*(2), 81.

Edwards, W. (1968). Conservatism in human information processing. *Formal representation of human judgment*, (pp 17-52), New-York: Wiley.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological review*, *101*(3), 519-527.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology*, *38*(1), 129-166.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, *13*(12), 517-523.

Kahneman, Daniel and Amos Tversky. (1979). "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica* 47, 263–291.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In *Decision making and change in human affairs* (pp. 275-324). Springer Netherlands.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323-343.

Wu, G., & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management science*, *42*(12), 1676-1690.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, 6.

Chapter 3: Statistical aids to improve medical decision accuracy

Marine Hainguerlot[1], Vincent Gajdos[2,3], Jean-Christophe Vergnaud[1]

[1]Centre d'Economie de la Sorbonne CNRS UMR 8174, Paris, France.

[2]Department of Pediatrics, Antoine Béclère University Hospital, Assistance Publique-Hôpitaux de Paris.

[3]INSERM, CESP Centre for Research in Epidemiology and Population Health, Paris-Sud, Paris-Saclay University, Villejuif, France.

Abstract

We developed statistical aids for physicians suffering from probability distortion in their clinical judgment to improve the accuracy of their medical decisions. We considered two main statistical aids. First, a combined mechanical model with physicians’ intuition was proposed to capitalize on physicians’ strength (their intuition) while compensating for their weakness (suboptimal weight of the medical evidence). Second, we included physicians’ deviation from expected decision, considering that it might be valid information, in the previous combined model. Our statistical aids were developed using the Lens model approach (Brunswick, 1952; Goldberg, 1970). We applied our statistical aids to the detection of bacterial infection in febrile infants younger than 3 months (N=1848) and assessed their efficiency in improving antibiotic treatment decisions. Overall, we found that a mechanical aid combined with human intuition could improve diagnostic accuracy but not decision accuracy. To improve physicians’ actual decision, it was necessary to include in the combined model physicians’ deviation from expected decision.

Introduction

In a previous study using data from actual medical practice, we found that physicians' probability estimates that infants have BI were distorted (Hainguerlot et al, 2017). Physicians over-estimated small probabilities and under-estimated large probabilities. Critically, we found that such distortion in clinical judgment might cause unnecessary antibiotic treatment. The aim of this study is to develop statistical aids that could improve medical decision accuracy by eliminating probability distortion in clinical judgment.

Mechanical models do not suffer from probability distortion since they optimally combine different sources of information (Dawes et al., 1989). Thus, replacing physicians' judgment by a mechanical model may be considered to be an efficient aid. However, it has been recommended to not replace human judgment when humans have intuition (Blattberg & Hoch, 1990). A robust finding is that a combination of a mechanical model with human intuition outperforms the two inputs in isolation when humans have valid intuition. Importantly, it has been documented that physicians rely on their clinical intuition (Van den Bruel et al, 2012; Woolley & Kostopoulou, 2013) and that their intuition might be even more predictive than clinical features (Van den Bruel et al, 2010). Therefore, it would be best to develop a statistical aid that combines a mechanical model with physicians' intuition. Such combined aid would capitalize on the physician' strength (his intuition) while compensating for his weakness (suboptimal weight of the medical evidence).

Nonetheless, whether or not a combined mechanical model with human intuition can improve decision accuracy remains an open question. The aid would generate a clinical prediction that infants have BI. A decision threshold on the clinical prediction could be determined such that physicians would be advised to treat with antibiotics when the clinical prediction exceeds the threshold. If physicians similarly apply a decision threshold on their clinical judgment (Pauker & Kassirer, 1980), then the combined aid should improve decision accuracy as long as the diagnostic accuracy of the clinical prediction outperforms physicians' judgments. However and to the best of our knowledge, the extent to which physicians use a decision threshold in practice is not known. Some studies have shown that non-clinical features such as the patient's personal characteristics and characteristics of the clinical practice were influencing the decision (Hajjaj et al, 2010; McKinlay et al, 1996), suggesting that physicians departed from a threshold model. Furthermore, a recent study revealed that the cohesion between clinical judgment and medical decision varied among physicians (Beckstead, 2017). Physicians' departure from the threshold model might be relevant. In that case, it would be best to develop a statistical aid that also takes into account the deviation from the expected decision.

Furthermore, humans are able to accurately evaluate their decisions. They report judgments of confidence in their decision that correlate with performance and they are capable to detect their errors (Yeung & Summerfield, 2012). If physicians were able to reflect on the accuracy of their subjective reports (clinical judgment and medical decision), it would be best to develop a statistical aid that takes advantage of physicians' metacognitive reports.

As a result, we investigated two main statistical aids: (1.1) a combined mechanical model with physicians' intuition ($\boldsymbol{M_1}$) and (2.1) the combined model ($\boldsymbol{M_1}$) with physicians' deviation from expected decision ($\boldsymbol{M_2}$). We also considered the previous aids augmented with physicians' metacognitive reports: (1.2) the model ($\boldsymbol{M_1}$) augmented with physicians' confidence in their clinical judgment ($\boldsymbol{M_1^*}$) and (2.2) the model ($\boldsymbol{M_2}$) augmented with physicians' error detection ($\boldsymbol{M_2^*}$).

Critically, combined mechanical models could improve diagnostic accuracy through two mechanisms: the elimination of probability distortion and the optimal weight of human components (intuition and deviation). Thus, to identify the extent to which the elimination of probability distortion improved decision accuracy, we used combined judgmental bootstrapped models as a benchmark. We reasoned that combined judgmental bootstrapped models could only improve diagnostic accuracy by optimally weighting the human components. Consequently, the difference in gain in accuracy between combined mechanical and combined bootstrapped models should capture the gain from eliminating probability distortion. As a benchmark, the following models were developed: (1.1) a combined judgmental bootstrapped model with physicians' intuition ($\boldsymbol{B_1}$), (1.2) the model ($\boldsymbol{B_1}$) augmented with physicians' confidence in their clinical judgment ($\boldsymbol{B_1^*}$), (2.1) the combined model ($\boldsymbol{B_1}$) with physicians' deviation from expected decision ($\boldsymbol{B_2}$), and (2.2) the model ($\boldsymbol{B_2}$) augmented with physicians' error detection ($\boldsymbol{B_2^*}$).

Our statistical aids were developed using the Lens model approach (Brunswick, 1952; Goldberg, 1970) to estimate the mechanical and the judgmental bootstrapped models but also to isolate physicians' intuition and physicians' deviation from expected decision. We applied our statistical aids to the detection of bacterial infection in febrile infants younger than 3 months (N=1848) and assessed their efficiency in improving antibiotic treatment decisions.

Methods

Data

Study design, setting and participants

Our data come from a study of the procalcitonin biomarker in the detection of bacterial infection (BI) in febrile infants younger than 3 months (Milcent et al., 2016). They performed a prospective, multicenter, cohort study in 15 French pediatric emergency departments for a period of 30 months from October 1, 2008, through March 31, 2011. Infants were eligible in the study if they were older than 7 days and younger than 91 days, with temperatures of 38°C or higher (at home or on admission), without antibiotic treatment within the previous 48 hours, and without major comorbidities (immune deficiency, congenital abnormality, or chronic disease). More details are reported in the original study (see Milcent et al., 2016).

Data collection

Physicians were asked to record the information they collected about the patient from its admission to its discharge. They recorded the demographic and neonatal data, medical history, physical examination, and clinical findings from ordered laboratory tests. Physicians were required to report their probability estimate that the infant had a BI, on a scale from 0 to 100%, and an interval within which their probability fell after receiving the clinical findings from the laboratory tests (posttest probability). Following this estimate, they reported their decisions to treat with antibiotics, along with a reason for using antibiotics (BI for sure, BI likely or precautionary principle).

Study size

Milcent et al. (2016) enrolled 2273 patients and 2047 infants were included in their final analysis. Furthermore, we restricted our sample to cases with available data regarding probability estimates and medical decisions and obtained a final sample of N = 1848 patients.

Outcome measure

The outcome measure included definite several bacterial infections and possible several bacterial infections categorized by the attending physician. Definite SBIs included bacteremia, bacterial meningitis, urinary tract infection, pneumonia, otitis, gastroenteritis, soft tissue infection, bone and joint infection. Possible SBIs included possible pneumonia and possible otitis. A committee of medical experts reviewed the diagnoses.

Predictor variables

We included all candidate predictor variables collected by physicians in the statistical analysis. Four categories of predictor variables were considered: (1) neonatal and demographic data, (2) medical history, (3) clinical examination, and (4) laboratory tests. We restricted laboratory tests to tests with clinical findings that were available when physicians were asked to report their second estimate, thus we excluded from our analysis blood culture, stool test and the procalcitonin assay. We also excluded additional samplings and additional radiography from the predictors.

Continuous variables were included as continuous and missing data were replaced by the average value computed on available data. Dichotomous variables were coded as +1 if the risk factor was present and -1 if the risk factor was absent; missing data were replaced by the value 0.

Statistical aids development

Combined models

We considered various combined models, which are summarized in **Table 1**. First, we considered a simple mechanical model ($\boldsymbol{M_0}$). Second, we combined the mechanical model with physicians' intuition from their clinical judgment ($\boldsymbol{M_1}$), referred to as "residual judgment" in the rest of the article. Finally, we proposed to add to the previous combined model physicians' deviation from expected decision ($\boldsymbol{M_2}$), called "residual decision" in the subsequent analyses. Additionally, we developed augmented models with metacognitive reports by weighting the residuals as a function of metacognitive reports. First, we considered that the size of the probability interval might be interpreted as a measure of physicians' confidence in their clinical judgment. Second, reasons for using antibiotic might be interpreted as a measure of physicians' awareness of the accuracy of their decisions ("error detection"). We added interaction terms in our combined model: an interaction between judgment residual and confidence ($\boldsymbol{M_1^*}$) and an interaction between decision residual and error detection ($\boldsymbol{M_2^*}$).

To compare the diagnostic and decision accuracies of the combined mechanical models ($\boldsymbol{M_i}$), we used as a benchmark similar combined models with judgmental bootstrapped ($\boldsymbol{B_i}$). Finally, all the models were optimal combinations. The optimal weights were estimated by running a logistic regression with BI as a dependent variable.

Table 1: Combined models

	Mechanical	Judgmental bootstrapped
Pre-judgment	$\boldsymbol{M_0}$: Mechanical model	
Post-judgment	$\boldsymbol{M_1}$: M0 + residual judgment	$\boldsymbol{B_0}$: Judgmental bootstrapped
	$\boldsymbol{M_1^*}$: M1 + confidence	$\boldsymbol{B_1}$: B0 + residual judgment
		$\boldsymbol{B_1^*}$: B1 + confidence
Post-decision	$\boldsymbol{M_2}$: M1 + residual decision	$\boldsymbol{B_2}$: B1 + residual decision
	$\boldsymbol{M_2^*}$: M2+ error detection	$\boldsymbol{B_2^*}$: B2+ error detection

Lens model approach

Mechanical model:

The mechanical model ($\boldsymbol{M_0}$) is defined as the best fitting model for predicting BI using only statistically significant predictors of BI. We modeled the presence or absence of BI by a linear logistic model. The log likelihood ratio of the bacterial infection conditional on the predictor variables X is a weighted sum of the predictors:

$$ln\frac{p(BI|X)}{1-p(BI|X)} = \beta_o^o + \sum_{i=1}^{n} \beta_i^o\, x_i \qquad \textbf{(1)}$$

To be parsimonious in the number of predictors included, we identified the statistically significant predictors i of bacterial infection by running a stepwise logistic regression on all the predictors collected by physicians. Given the estimated parameters $\widehat{\beta^o}$, the mechanical model predicted objective probability po as follows:

$$ln\frac{\widehat{po}}{1-\widehat{po}} = \widehat{\beta_o^o} + \sum_{i=1}^{n} \widehat{\beta_i^o}\, x_i \qquad \textbf{(2)}$$

Judgmental bootstrapped:

The model of physicians‘judgment (i.e. judgmental bootstrapped model ($\boldsymbol{B_0}$)) is constructed by regressing posttest probabilities (ps) onto the list of predictors:

$$ln\frac{ps}{1-ps} = \beta_o^s + \sum_{j=1}^{m} \beta_j^s\, x_j + \mu \qquad \textbf{(3)}$$

We also restricted our analysis to statistically significant predictors j determined by running a stepwise OLS regression on all the predictors with the probability reported in log odds form as the dependent variable.

Given the estimated subjective weights $\widehat{\beta^s}$, we defined the judgmental bootstrapped model ($\boldsymbol{B_0}$) as follows:

$$ln\frac{\widehat{ps}}{1-\widehat{ps}} = \widehat{\beta_o^s} + \sum_{j=1}^{m} \widehat{\beta_j^s}\, x_j \qquad \textbf{(4)}$$

We defined the residual judgment (μ) as the difference between ps (in log odds form) and equation **(4)**. The residual judgment contains the part of the physicians’ judgment that is not explained by a linear integration of the predictors. It may capture physicians’ intuition, physicians’ ability to interpret omitted features or to take into account predictors in a nonlinear way.

Decision model:

Following the same Lens model approach, we modeled the decision to treat with antibiotics (d) as a linear logistic function of the predictors and the stated subjective probability:

$$ln\frac{p(d|X,ps)}{1-p(d|X,ps)} = \beta_o^d + \sum_{k=1}^{p} \beta_k^d\, x_k + \beta_s^d ln\frac{ps}{1-ps} \qquad \textbf{(5)}$$

Statistically significant predictors k were identified by running a stepwise logistic regression. Given the estimated subjective weights $\widehat{\beta^d}$, we defined the decision model as follows:

$$ln\frac{p(\widehat{d|X,ps})}{1-p\left(\hat{d}|X,ps\right)} = \widehat{\beta_o^d} + \sum_{k=1}^{p}\widehat{\beta_k^d}\,x_k + \widehat{\beta_s^d}\,ln\frac{ps}{1-ps} \qquad \textbf{(6)}$$

We defined the residual decision (ω) as the difference between observed antibiotic treatment d (in log odds form) and equation **(6)** as follows**:**

$$\begin{cases} \ln\left(\frac{0.99}{0.01}\right) - ln\frac{p(\widehat{d|X,ps})}{1-p\left(\hat{d}|X,ps\right)}, & if\ actual\ decision\ is\ to\ treat \\ \ln\left(\frac{0.01}{0.99}\right) - ln\frac{p(\widehat{d|X,ps})}{1-p\left(\hat{d}|X,ps\right)}, & if\ actual\ decision\ is\ to\ not\ treat \end{cases}$$

where we choose to attribute the value $ln\frac{0.99}{0.01}$ if actual decision is to treat and $ln\frac{0.01}{0.99}$ if actual decision is to not treat, to handle infinite values.

The residual decision contains the part of physicians' decision that is not explained by physicians' judgment and a linear integration of the cues.

Human reports performance

To develop our statistical aids, we only included human reports with predictive accuracy. For both residual judgment and residual decision, we evaluated their predictive accuracy by estimating the AUC ROC curve for the detection of BI.

Regarding confidence, we separated our data in a high and a low interval group by a median split on the value of the interval while controlling for the level of objective probabilities; and we compared the AUC ROC curve for the detection of BI for posttest probabilities across the two groups. If confidence is predictive, the low interval group should have a greater AUC ROC curve. Finally, for each reason reported conditional on antibiotic treatment (BI for sure, BI likely or precautionary principle), we computed the false- alarm rate (proportion of infants without BI who received antibiotic treatment). If error detection is predictive, the false-alarm rate should decrease with physician's certainty.

Statistical aids performance

Diagnostic accuracy

For each combined model, we reported the area under the ROC curve (AUC) for the detection of BI. To assess the incremental value of nested models, we computed the Bayesian information criterion (BIC). The model with the lowest BIC should be preferred and a BIC difference of more than 10 would provide strong evidence in favor of the model with the lowest BIC.

Decision accuracy

For each combined model, we determined the decision threshold that maximized specificity (i.e. the proportion of infants without BI who did not receive antibiotic treatment) under the constraint that sensitivity (i.e. the proportion of infants with BI who received antibiotic treatment) was equal to the actual sensitivity. We used proportion tests to compare specificity reached by the combined models with actual specificity.

Probability distortion classification

To evaluate whether our combined mechanical aids could improve medical decision accuracy by eliminating probability distortion, we performed our analysis by classifying physicians into two groups: a high probability distortion group and a low probability distortion group. Combined models were estimated separately for each group and their performance was then compared. We defined probability distortion as the difference between objective probabilities $\widehat{po}$ (mechanical model M0, see **equation (2)**) and estimated subjective probabilities $\widehat{ps}$ (judgmental bootstrapped model B0, see **equation (4)**). The low and high bias groups were constructed by median split on the value of the estimated subjective probabilities $\widehat{ps}$ while controlling for the level of objective probabilities $\widehat{po}$. More precisely, we classified in the high bias group subjective probabilities above the median split when objective probabilities were small (i.e. below the prevalence rate of BI) and subjective probabilities below the median when objective probabilities were large (i.e. above the prevalence rate of BI). When objective probabilities were small, the high bias group had higher subjective probabilities than the low bias group. Inversely, the high bias group had lower subjective probabilities when objective probabilities were large. To confirm that physicians' estimates were more distorted in the high bias group than in the low bias group, we estimated separately a linear fit of $\widehat{ps}$ on $\widehat{po}$, expressed on log odds scale. The slope of the linear fit should be closer to one for the low bias group. Given our classification, we expected that the AUC ROC curve for the detection of BI for the mechanical model (M0) should be greater than the AUC ROC curve for the detection of BI for the judgmental bootstrapped model (B0) for the high bias group, whereas it should be similar in the low bias group. Furthermore, we expected to replicate previous findings regarding the impact of probability distortion on diagnostic and decision accuracies (Hainguerlot et al, 2017). We should observe for the high bias group a lower AUC ROC curve for the detection of BI for stated posttest probabilities and a lower specificity of antibiotic treatment, compared to the low bias group.

Results

Descriptive statistics

Descriptive statistics about the outcome measure (**Tables S1)**, judgment and decision **(Table S2**) and details about the classification high and low bias group (**Figure S1**) are reported in the Supplementary Results. Overall, the physicians' estimates were more distorted in the high bias group (N=892) than in the low bias group (N=956). Importantly, the low bias group outperformed the high bias regarding diagnostic and decision accuracies. First, the AUC of the ROC curve to detect BI for stated subjective probabilities was significantly higher in the low bias group (AUC low bias group: 0.910 (95% CI: 0.882 -0.938) vs. AUC high bias group 0.780 (95% CI: 0.735 -0.824)). Second, the specificity was significantly lower in the high bias group while the sensitivity did not differ across the two groups. Overall, the high bias group had a lower accuracy rate. **Table 2** presents the proportion test comparisons of sensitivity, specificity and decision accuracy rates between the two groups.

Table 2: Proportion test comparisons of sensitivity rate, specificity rate and decision accuracy rate for antibiotic treatment by low and high bias group

	Low bias		High bias		Proportion test 2 tailed		
	N	Mean	N	Mean	Proportion difference	z	pvalue
Sensitivity	176	0.983	158	0.981	0,002	0.13	0.8939
Specificity	780	0.779	734	0.638	0,142	6.08	<.0001
Accuracy	956	0.817	892	0.698	0.119	5.96	<.0001

Statistical aids development

Lens models

The list of predictor variables that we considered and descriptive statistics are reported in Supplementary Materials (**Table S3**). The predictors identified from the stepwise regressions are presented in **Table 3** for demographic data, neonatal data and medical history, **Table 4** for clinical examination and **Table 5** for laboratory tests. In each tables, **columns (1)**, **(2)** and **(3)** report the adjusted odds ratios of the predictors of BI, stated subjective probability and decision respectively. The stepwise regression identified only seven statistically significant predictors of BI (**column (1)**). Among demographic data, neonatal data and medical history, only sex and temperature were significant. No clinical predictors were selected. Five predictors from the laboratory tests -(urine analysis n°1, urine analysis n°2, GN, alveolar consolidation and CRP) were found to explain BI. Some predictors of BI were also predictors of the stated subjective probability (**column (2)**). In particular, the five predictors from the laboratory tests were also taken into account by physicians in their probability estimates but were systematically under- valued. Overall, the physicians used more predictors to form their probability estimates. Numerous predictors from the clinical examination influenced the physician' probability estimates but were not predictors of the BI. Interestingly, we found that to make their decision to treat with antibiotics, physicians were taking into account, in addition to their stated subjective probability, neonatal, medical history and laboratory predictors (**column (3)**). In particular, they considered predictors that were not taken into account in their probability estimates such as, for example, the maximum temperature and the duration of the fever. They also attributed more or less weight to predictors previously taken into account. For example, they gave a greater weight to a positive result from the urine analysis.

Table 3: Demographic, neonatal and medical history predictors of BI, probability and decision

	(1) BI		(2) Probability		(3) Decision	
	AOR	(95%CI)	AOR	(95%CI)	AOR	(95%CI)
Male sex	2.34	(1.57-3.47)	1.23	(1.08-1.40)		
Maximum temperature (°C)	2.16	(1.06-4.40)			2.01	(1.19-3.40)
GBS			1.40	(1.17-1.68)	1.48	(1.03-2.13)
Cough			0.84	(0.71-0.98)		
Neonatal fever			1.51	(1.00-2.27)		
Chills			1.44	(1.07-1.92)		
Admission to the ED in August			1.41	(1.03-1.92)		
>30 days					0.61	(0.44-0.83)
Duration of fever (10 hours)					1.15	(1.01-1.31)

Abbreviation: AOR, adjusted odds ratios; GBS, group B streptococcus screening at 8 months of pregnancy; ED, emergency department

Table 4: Clinical predictors of BI, probability and decision

	(1) BI		(2) Probability		(3) Decision	
	AOR	(95%CI)	AOR	(95%CI)	AOR	(95%CI)
Rectale temperature (°C)			1.24	(1.02-1.51)		
Heart rate (/min)			1.01	(1.00-1.02)		
Respiratory rate (/min)			1.02	(1.01-1.03)		
Weight (/1000g)			0.78	(0.68-0.90)		
OSM (%)			1.21	(1.09-1.34)		
Moderate or weak GP			1.30	(1.06-1.59)		
Abnormal respiration			1.31	(1.06-1.62)		
Rhinitis			0.74	(0.64-0.86)		
Abnormal capillary refill time			2.00	(1.33-3.00)		
Moaning			1.33	(1.07-1.67)		
Poor or absent tone			1.35	(1.06-1.73)		
Poor or absent eye contact			1.88	(1.37-2.58)		

Abbreviation: AOR, adjusted odds ratios; GP, general appearance;
OSM, Oxygen saturation measurement

Table 5: Laboratory predictors of BI, probability and decision

	(1) BI		(2) Probability		(3) Decision	
	AOR	(95%CI)	AOR	(95%CI)	AOR	(95%CI)
UA positive n°1	28.41	(18.97-42.55)	6.04	(5.05-7.23)	4.74	(3.17-7.08)
UA positive n°2	10.21	(3.70-28.20)	1.79	(1.16-2.74)	5.43	(2.05-14.36)
GN (1000/mm3)	1.54	(1.39-1.70)	1.23	(1.17-1.29)	1.26	(1.16-1.38)
A. consolidation	41.05	(20.01-84.20)	4.25	(2.97-6.09)	2.34	(1.14-4.80)
CRP (10 mg/L)	1.50	(1.29-1.74)	1.18	(1.13-1.24)		
WBC in CSF *			1.11	(1.07-1.15)	1.16	(1.02-1.31)
CBG (mmlo/L)			1.38	(1.15-1.65)		
Lymphocyte (1000/mm3)			1.07	(1.01-1.13)		
Cells in the NBS			5.51	(1.10-27.55)		
Monocyte**			0.81	(0.69-0.95)		
CSF Gram stain					0.09	(0.05-0.16)

Abbreviation: A, alveolar; AOR, adjusted odds ratios; CBG, capillary blood glucose; CSF, cerebrospinal fluid; CRP, C-reactive protein; GN, granulocyte neutrophile; NBS, nasopharyngeal bacterial sampling; UA, urine analysis, WBC, white blood cell. *(100/mm3); **(1000/mm3)

Descriptive statistics about the predicted values and the residuals obtained from the Lens model approach regarding the mechanical model, the judgemental bootstrapped model and the decision model are reported in **Table 6**.

Table 6: Predicted values and residuals from Lens models

	N	**Mean**	**(Sd)**
Predicted objective probability ($\boldsymbol{M_0}$)	1848	0.18	(0.29)
Predicted subjective probability ($\boldsymbol{B_0}$)	1848	0.23	(0.22)
Predicted probability of antibiotic treatment	1848	0.41	(0.32)
Residuals judgment	1848	0.02	(0.20)
Residuals decision	1848	0.00	(0.37)

Furthermore, the AUC ROC curve to detect BI for the mechanical model ($\boldsymbol{M_0}$) was equal to 0.93 (95%CI: 0.911-0.950) and for the judgmental bootstrapped model ($\boldsymbol{B_0}$) equal to 0.895 (95%CI: 0.872 – 0.917).

Human reports performance

The AUC ROC curve for the detection of BI revealed that the residual judgment (AUC: 0.622 (95%CI: 0.585- 0.660) and the residual decision (0.739 (95%CI: 0.715- 0.762) were both significant predictors of BI. On the other hand, the probability intervals were not related to the accuracy of stated probability. The AUC ROC curve for the detection of BI for stated subjective probability was not significantly greater in the low interval group (0.836 (95%CI: 0.801- 0.871)) compared to the high interval group (0.868 (95%CI: 0.83- 0.906)). Finally, reasons given for antibiotic treatment were related to the accuracy of the decision. The false-alarm rate increased as the physician reported that he was not sure about whether or not the infant had a BI (see **Table 7**). Consequently, residual judgment, residual decision and reasons for antibiotic treatment were used to develop statistical aids. On the other hand, we excluded the probability intervals from subsequent analyses.

Table 7: False alarm rate as a function of reason given by physicians for antibiotic treatment

	BI for sure	BI likely	No reason	Precautionary principle
False- alarm rate	9.7%	60%	78.4%.	91.2%

Summary

Table 8 reports the AUC and **Figure 1** plots the ROC curves to detect BI of the different elements that we considered to develop our statistical aids, for the high and low bias groups, compared to the stated subjective probability ($\boldsymbol{S}$). As expected from our classification, the mechanical model ($\boldsymbol{M_0}$) outperformed the bootstrapped model ($\boldsymbol{B_0}$) only in the high bias group since it eliminated probability distortion. Furthermore, in the high bias group, the bootstrapped model ($\boldsymbol{B_0}$) outperformed stated subjective probability, suggesting that the high bias group also had variance not predictive of BI in $\boldsymbol{S}$**.** Residual judgment was similar in the two groups. The AUC ROC curve for the detection of BI for the error detection was significantly higher in the low bias group.

Table 8: Summary of AUC of ROC curve to detect BI for individual elements

	Low bias		High bias	
	AUC	(95%IC)	AUC	95%IC
Stated subjective probability (***S***)	0.910	(0.882-0.938)	0.780	(0.735-0.824)
Judgmental bootstrapped ($\boldsymbol{B_0}$)	0.911	(0.881-0.941)	0.872	(0.842-0.902)
Mechanical model ($\boldsymbol{M_0}$)	0.924	(0.895-0.952)	0.937	(0.914-0.960)
Residual judgment	0.624	(0.572-0.676)	0.620	(0.564-0.675)
Residual decision	0.771	(0.740-0.803)	0.714	(0.679-0.749)
Error detection	0.959	(0.945- 0.973)	0.930	(0.910- 0.950)

Abbreviations: AUC, area under the curve for the receiver operating characteristic curve

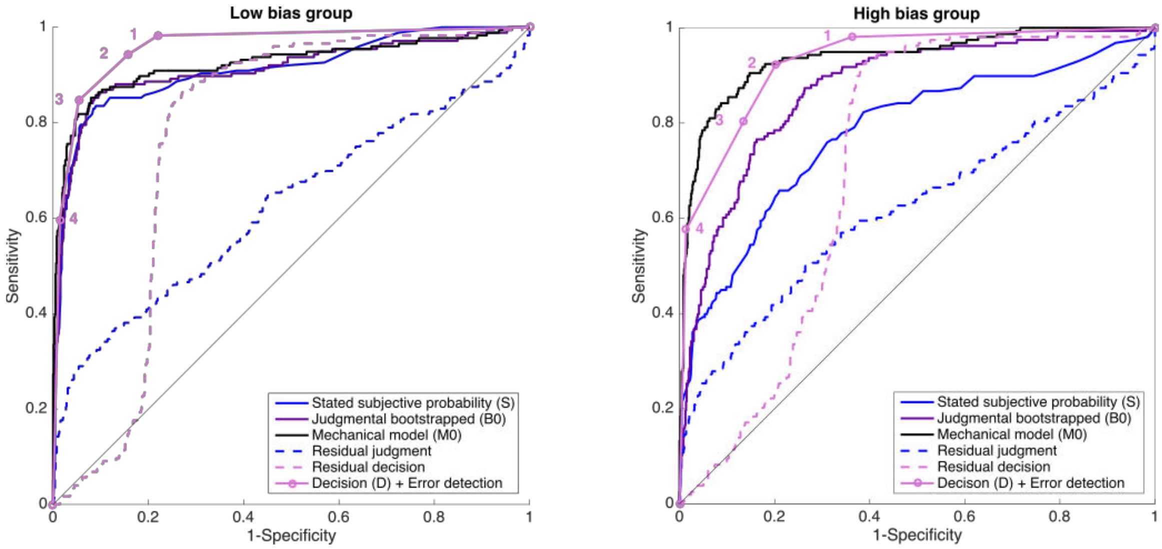


Figure 1: *Receiver operating characteristic curves to detect BI for stated subjective probability (**S**), judgmental bootstrapped ($\boldsymbol{B_0}$), mechanical model ($\boldsymbol{M_0}$), residual judgment, residual decision and reasons for antibiotic treatment (decision + error detection) for the low and high bias group. Four cut-off points are reported for the ROC curve decision + error detection with: (1) all antibiotic treatment are classified as positive; (2) all antibiotic treatment except the ones with precautionary principle are classified as positive; (3) only antibiotic treatment with BI for sure and BI likely are classified as positive; (4) only antibiotic treatment with BI for sure are classified as positive.*

Statistical aids performance

Diagnostic accuracy performance

The performance (AUC and BIC) of the combined models is reported in **Table 9** for the low and high bias group. Overall, the combined mechanical models ($\boldsymbol{M_i}$) outperformed the stated subjective probability (***S***). We observed an increase in AUC and a decrease in BIC with the successive inclusion of additional human reports (from model 0 to model 3). The differences in BIC were greater than 10, except from $\boldsymbol{M_2}$ to $\boldsymbol{M_2^*}$. Thus, introducing human reports greatly increased the diagnostic accuracy of the models.

Then, to assess the extent to which the mechanical model improved decision accuracy through the elimination of probability distortion, we compared it with the bootstrapped model. Even though the combined bootstrapped model $\boldsymbol{B_1}$ slightly outperformed $\boldsymbol{B_0}$ in both groups, the difference was not statistically significant. This result suggested that the optimal weighting of human intuition was not a substantial improvement. Importantly, comparing $\boldsymbol{M_1}$ with $\boldsymbol{B_1}$**,** we observed that the AUC of the combined mechanical model $\boldsymbol{M_1}$ significantly outperformed the bootstrapped model $\boldsymbol{B_1}$ only for the high bias group. Thus, the elimination of probability distortion significantly improved diagnostic accuracy in the high bias group.

Table 9: Summary of AUC and fit results for combined models (bootstrapped ($\boldsymbol{B_i}$) and mechanical ($\boldsymbol{M_i}$) models) compared to stated subjective probability ($\boldsymbol{S}$)

	Low bias			High bias		
	AUC	(95%IC)	BIC	AUC	(95%IC)	BIC
$\boldsymbol{S}$	0.910	(0.882-0.938)	493.992	0.780	(0.735-0.824)	675.696
Models						
$\boldsymbol{B_0}$	0.911	(0.881-0.941)	466.164	0.872	(0.842-0.902)	574.821
$\boldsymbol{B_1}$	0.931	(0.905-0.956)	422.773	0.885	(0.857-0.912)	551.734
$\boldsymbol{B_2}$	0.962	(0.947-0.978)	365.825	0.919	(0.901-0.938)	485.742
$\boldsymbol{B_2^*}$	0.968	(0.953-0.982)	356.001	0.942	(0.926-0.957)	423.893
$\boldsymbol{M_0}$	0.924	(0.895-0.952)	430.572	0.937	(0.914-0.960)	390.232
$\boldsymbol{M_1}$	0.940	(0.914-0.965)	385.150	0.944	(0.921-0.966)	369.481
$\boldsymbol{M_2}$	0.972	(0.960-0.985)	319.738	0.970	(0.957-0.982)	293.051
$\boldsymbol{M_2^*}$	0.974	(0.960-0.987)	323.602	0.973	(0.961-0.985)	287.291

Abbreviations: AUC, area under the curve for the receiver operating characteristic curve; BIC, bayesian information criterion

Decision accuracy performance

The proportion tests of specificity for the combined models compared to actual antibiotic treatment are reported in **Table 10**. As expected, the combined mechanical models did not significantly improve the actual specificity rate, in the low bias group. Critically, in the high bias group, the proportion test between $\boldsymbol{M_1}$ and actual antibiotic treatment revealed that a combined mechanical model with human intuition did not succeed to improve decision accuracy. Data revealed that to improve actual specificity it was necessary to include physicians' deviation from expected decision ($\boldsymbol{M_2}$). Finally, the best specificity rate was reached with $\boldsymbol{M_2^*}$ once physicians' metacognitive reports were taken into account.

Table 10: Proportion test (2 tailed) of specificity for combined models (bootstrapped (B_i) and mechanical (M_i) models) compared to actual antibiotic treatment (D)

	Low bias				High bias			
	Mean	Δ	z	p value	Mean	Δ	z	p value
D	0.779				0.638			
Models								
$\boldsymbol{B_0}$	0.141	-0.638	-25.297	<.001	0.226	-0.411	-15.913	<.001
$\boldsymbol{B_1}$	0.274	-0.505	-19.980	<.001	0.368	-0.270	-10.336	<.001
$\boldsymbol{B_2}$	0.779	0.000	0.000	1	0.672	0.034	1.372	0.170
$\boldsymbol{B_2^*}$	0.790	0.010	0.493	0.622	0.703	0.065	2.665	0.008
$\boldsymbol{M_0}$	0.110	-0.669	-26.595	<.001	0.372	-0.266	-10.179	<.001
$\boldsymbol{M_1}$	0.168	-0.612	-24.187	<.001	0.395	-0.243	-9.297	<.001
$\boldsymbol{M_2}$	0.774	-0.005	-0.243	0.809	0.714	0.076	3.122	0.002
$\boldsymbol{M_2^*}$	0.749	-0.031	-1.431	0.152	0.770	0.132	5.544	<.001

Abbreviations: Δ, proportion difference

Figure 2 summarizes the statistical aids performance for the low and high bias group. As can be observed, the ROC curves for the bootstrapped and mechanical models similarly outperformed the stated subjective probability, for the low bias group. On the other hand, for the high bias group, the ROC curves of the mechanical models were greater than the ROC curves of the bootstrapped models, which were greater than the stated subjective probability. Furthermore, the combined models $\boldsymbol{B_2^*}$-***D*** or $\boldsymbol{M_2^*}$-***D*** (red line) outperformed the specificity of the actual antibiotic treatment decision (pink line) only in the high bias group.

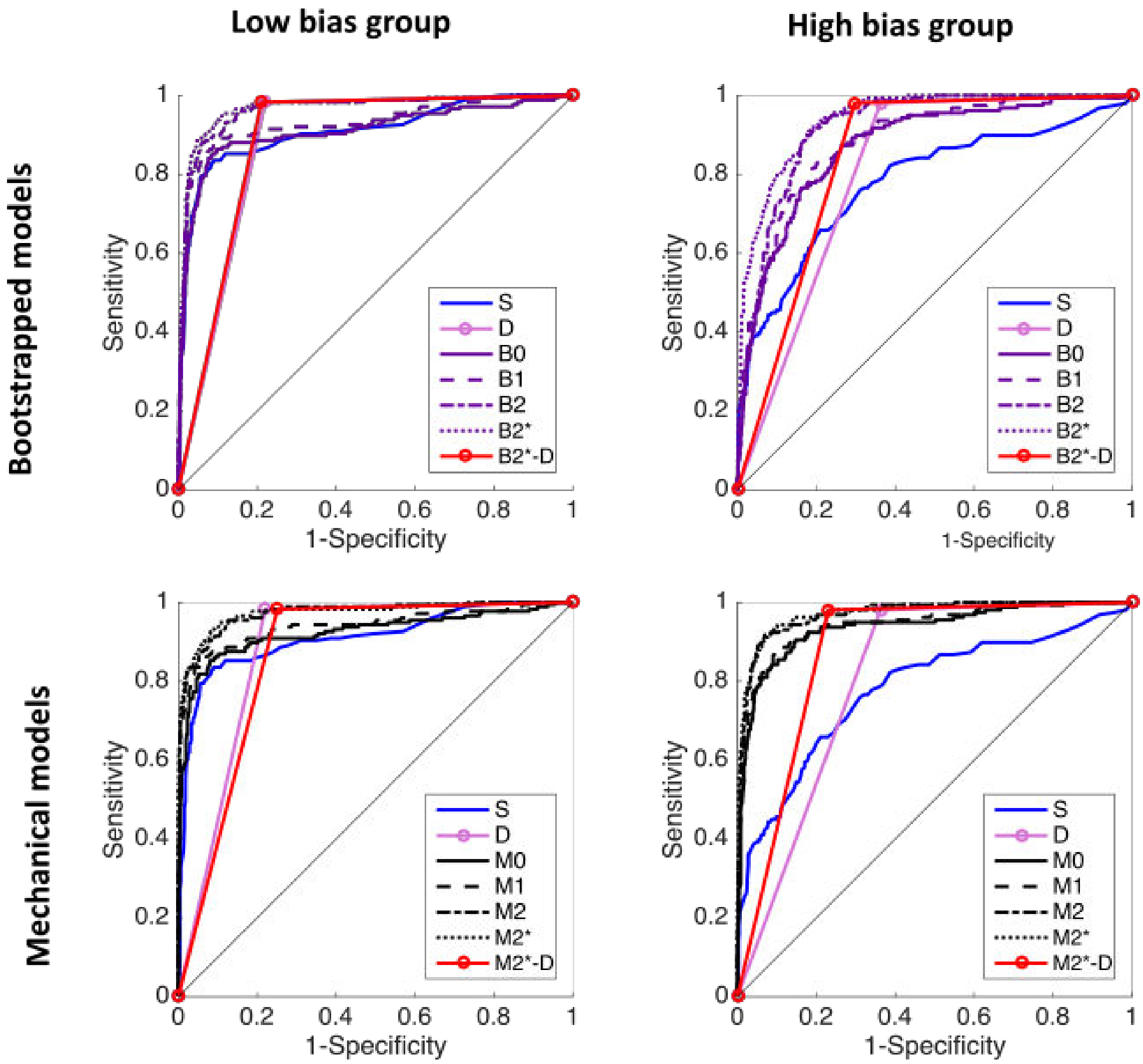


Figure 2: *Receiver operating characteristic curves to detect BI, for the low and high bias group, for stated subjective probability (**S**), antibiotic treatment (**D**), combined models (bootstrapped ($\boldsymbol{B_i}$) and mechanical ($\boldsymbol{M_i}$) models) and antibiotic treatment decision reached with the best combined model ($\boldsymbol{B_2^*}$**-D** or $\boldsymbol{M_2^*}$**-D**)*

Discussion

We developed statistical aids for physicians suffering from probability distortion in their clinical judgment to improve the accuracy of their antibiotic treatment decisions for febrile infants at risk of bacterial infection. We observed that a mechanical aid combined with human intuition could improve diagnostic accuracy but not decision accuracy. To improve physicians' actual decision, it was necessary to include in the combined model physicians' observed deviation from expected decision ("residual decision").

Residual decision

Several explanations might explain why residual decision was relevant information.

First, physicians were asked to report their belief that infants had BI on a probability scale. Critically, such report may not capture the richness of the clinical judgment. In particular, physicians could be more or less confident in their judgment and certainty is likely to affect their decisions (Lutfey et al, 2009). Second, to make final decision physicians may not only rely on their clinical judgment but also share it with others (Charles et al, 1997). For example, Bergman et al (2006) observed that practice type, such as solo or two-person practice and group practice, explained variability in the treatment of febrile infants. Critically, it has been documented that shared decision making could outperform individual decision (Bahrami et al, 2010). Finally, characteristics of the clinical practice may influence decision. In particular, Bergman et al (2006) found that practice-site fixed effects explained a substantial part of the variability in treatment, and they suggested that one potential explanation for this effect might be that practice-specific guidelines are established among physicians of a site. In our data, we cannot explicitly test whether practice-specific guidelines explained the observed deviation from expected decision, but we can explore whether the cohesion between clinical judgment and medical decision varied across practice sites. By pediatric emergency departments (EDs) (N=10), we computed the AUC ROC curve to detect antibiotic treatment for stated subjective probabilities (see **Table S4**). Importantly, we found heterogeneity in the ability of stated subjective probabilities to predict antibiotic treatment, which provided support for the existence of practice-specific guidelines in EDs.

Limitations and further development

Our study has several limitations. First, our statistical aids were developed and applied to the same dataset. To validate our results, a leave-one-out cross- validation is in progress. Second, one may be concerned about the feasibility of implementing our statistical aids in clinical settings. The development of our combined mechanical models required the statistical analysis of physicians' judgment and decision. Alternatively, a more straightforward method would be to combine directly the mechanical model with physicians' reports (for example, see Blattberg & Hoch, 1990). Our approach should only be implemented if it provides substantial benefit, compared to the straightforward method. We assessed whether the straightforward method would have yielded the same results in our data. We estimated combined mechanical models with human reports as follows: (1.1) a combined mechanical model with physicians' judgment ($\boldsymbol{M_1}$***- reports***), (2.1) the combined model ($\boldsymbol{M_1}$***- reports***) with physicians' decision ($\boldsymbol{M_2}$***- reports***), and (2.2) the model ($\boldsymbol{M_2}$***- reports***) augmented with physicians' error detection ($\boldsymbol{M_2^*}$***- reports***). **Table S5** compares the diagnostic and decision accuracies between combined models with human residuals and combined models with human reports. The AUC of the combined models with residuals was slightly greater. Importantly, the improvement in specificity could only be reached with our approach. Further investigation would be needed to evaluate the conditions under which the straightforward method and our approach would be best suited.

Conclusion

A mechanical model combined with physicians' intuition and deviation from expected decision substantially improved the specificity of antibiotic treatment, for physicians suffering from probability distortion. External validation studies of our statistical aids would be necessary. If further validated, future studies could evaluate the effect of providing physicians with such decision support system on decision accuracy.

References

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081-1085.

Beckstead, J. W. (2017). The Bifocal Lens Model and Equation: Examining the Linkage between Clinical Judgments and Decisions. *Medical Decision Making*, *37*(1), 35-45.

Bergman, D. A., Mayer, M. L., Pantell, R. H., Finch, S. A., & Wasserman, R. C. (2006). Does clinical presentation explain practice variability in the treatment of febrile infants?. *Pediatrics*, *117*(3), 787-795.

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model+ 50% manager. *Management Science*, *36*(8), 887-899.

Brunswik, E. (1952). The conceptual framework of psychology. *Psychological Bulletin*, *49*(6), 654-656.

Charles, C., Gafni, A., & Whelan, T. (1997). Shared decision-making in the medical encounter: what does it mean?(or it takes at least two to tango). *Social science & medicine*, *44*(5), 681-692.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668-1674.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological bulletin*, *73*(6), 422.

Hainguerlot, M., Gajdos, V., Vergnaud, J-C. (2017). Impact of probability distortion on medical decision: a field study. Manuscript in preparation.

Hajjaj, F. M., Salek, M. S., Basra, M. K., & Finlay, A. Y. (2010). Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice. *Journal of the Royal Society of Medicine*, *103*(5), 178-187.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of Lens model studies. *Psychological Bulletin, 134*(3), 404-426.

Lutfey, K. E., Link, C. L., Marceau, L. D., Grant, R. W., Adams, A., Arber, S., ... & McKinlay, J. B. (2009). Diagnostic certainty as a source of medical practice variation in coronary heart disease: results from a cross-national experiment of clinical decision making. *Medical Decision Making*, *29*(5), 606-618.

McKinlay, J. B., Potter, D. A., & Feldman, H. A. (1996). Non-medical influences on medical decision-making. *Social science & medicine*, *42*(5), 769-776.

Milcent, K., Faesch, S., Gras-Le Guen, C., Dubos, F., Poulalhon, C., Badier, I., ... & Nissack, G. (2016). Use of procalcitonin assays to predict serious bacterial infection in young febrile infants. *JAMA pediatrics*, 170(1), 62-69.

Pauker, S. G., & Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine*, *302*(20), 1109-1117.

Van den Bruel, A., Thompson, M., Buntinx, F., & Mant, D. (2012). Clinicians' gut feeling about serious infections in children: observational study. *British Medical Journal*, *345*, e6144.

Van den Bruel, A., Haj-Hassan, T., Thompson, M., Buntinx, F., Mant, D., & European Research Network on Recognising Serious Infection investigators. (2010). Diagnostic value of clinical features at presentation to identify serious infection in children in developed countries: a systematic review. *The Lancet*, 375(9717), 834-845.

Woolley, A., & Kostopoulou, O. (2013). Clinical intuition in family medicine: more than first impressions. *The annals of family medicine*, *11*(1), 60-66.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B*, *367*(1594), 1310-1321

Supplementary materials

Supplementary results

Descriptive statistics

In our data (N=1848), the prevalence of BI was 18.1% (334/1848). Among the infants with definite BI (284/334), 257 had urinary tract infection. Among the possible BI (50/334), possible pneumonia was the most frequently diagnosed (35/50). Detailed data on diagnoses are reported in **Table S1**.

Table S1: Bacterial Infections
Value are numbers (percentage)

	N	No	(%)
BI	1848	334	(0.18)
Definite BI	1848	284	(0.15)
Bacteremia	284	10	(0.04)
Bacterial meningitis	284	7	(0.02)
Urinary tract infection	284	257	(0.90)
Pneumonia	284	0	(0.00)
Otitis	284	4	(0.01)
Gastroenteritis	284	3	(0.01)
Soft tissue infection	284	1	(0.00)
Bone and joint infection	284	0	(0.00)
Other definite BI	284	2	(0.01)
Possible BI	1848	50	(0.03)
Possible pneumonia	50	35	(0.70)
Possible otitis	50	11	(0.22)
Other possible BI	50	4	(0.08)

N: Total number of available data
Abbreviation: BI, bacterial infection

The statistics concerning clinical judgments (posttest probability and probability interval) as well as medical decision (antibiotic treatment) are reported in **Table S2**. Overall, physicians' probability estimates that infants had BI were pessimistic compared to the prevalence of BI observed in our data. The mean of the posttest probability was 25.43% whereas we observed a rate of BI of 18.1%. Antibiotic treatment was administered to 41% of the infants.

Table S2: Judgments and medical decisions
Values are numbers (percentage)
unless stated otherwise by*: mean (sd)

	N	No	(%)
Posttest probability*	1848	25.43	(27.93)
Lower bound posttest probability*	1848	18.30	(26.21)
Upper bound posttest probability*	1848	34.17	(29.55)
Antibiotic treatment	1848	766	(0.41)
Reason for antibiotic, BI for sure	766	217	(0.28)
Reason for antibiotic, BI likely	766	200	(0.26)
Reason for antibiotic, PP	766	182	(0.24)
Reason for antibiotic, None	766	167	(0.22)

N: Total number of available data
Abbreviation: PP, precautionary principle

Probability distortion classification

Estimated subjective probabilities were plotted versus objective probabilities in **Figure S1,** with probabilities expressed on log odds scale. The slope of the linear fit was closer to one for the low bias group (b=0.562) than for the high bias group (b=0.285). Moreover, the variance was better explained by a linear fit for the low bias group compared to the high bias group (R^2=0.854 vs R^2=0.590).

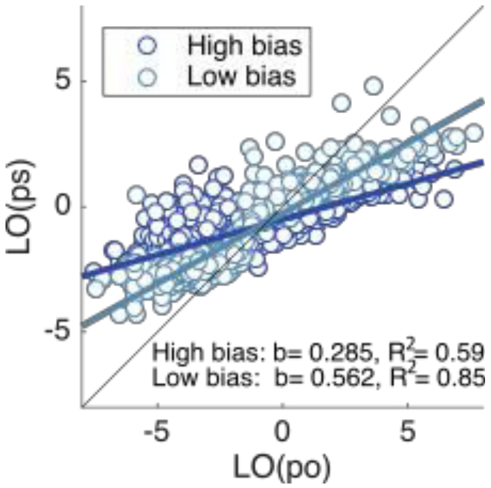


Figure S1*: Predicted subjective probabilities versus objective probabilities on log odds scale by high and low probability distortion groups. The dark and light blue lines are the best linear fits for the high and low bias probability distortion groups respectively.*

Table S3: Predictor variables
Values are numbers (percentage)
Unless stated otherwise by*: mean (sd)

	N	No	(%)
Demographic and neonatal data			
>30 days	1848	1483	(0.80)
Male sex	1848	1100	(0.60)
GBS screening at 8 months of pregnancy	1199	973	(0.81)
Detection of GBS	864	149	(0.17)
Premature rupture of membranes >12 hours	1729	181	(0.10)
Vaginal delivey	1813	1300	(0.72)
Maternal fever >38°C during labor &/ delivery	1788	77	(0.04)
NPS	1335	385	(0.29)
Detection of bacteria in NPS	347	68	(0.20)
Length of pregnancy (weeks of amenorrhea)*	1768	38.98	(1.53)
Neonatal fever	1811	37	(0.02)
Neonatal parenteral antibiotic treatment	1802	55	(0.03)
Neonatal oral antibiotic treatment	1801	15	(0.01)
Medical history			
Admission to the ED in August	1848	81	(0.04)
Duration of fever (hours)*	1795	13.96	(20.45)
Maximum temperature*	1835	38.65	(0.50)
Fever during the first 48 hours after vaccination	1837	65	(0.04)
Breast feeding at the admission	1827	843	(0.46)
Decrease in food consumption	1842	889	(0.48)
Diminished alertness	1843	292	(0.16)
Deterioration of general conditions	1842	567	(0.31)
Hypotonia, hyporesponsiveness	1841	400	(0.22)
Chills	1831	91	(0.05)
Another family member has fever	1812	513	(0.28)
Cough	1843	702	(0.38)
Breathing difficulty	1845	253	(0.14)
Rhinitis	1844	877	(0.48)
Vomiting	1843	299	(0.16)
Diarrhea	1842	259	(0.14)
Clinical examination			
Rectale temperature (°C)*	1843	37.98	(0.71)
Heart rate (/min)*	1802	158.59	(21.31)
Respiratory rate (/min)*	1351	43.68	(12.12)
Weight (g)*	1837	4833.03	(929.69)
Oxygen saturation measurement (%)*	1187	98.99	(1.55)
Systolic blood pressure (mmHg)*	751	94.34	(15.32)
Diastolic blood pressure (mmHg)*	747	55.40	(12.72)
Abnormal respiration	1839	256	(0.14)
Poor peripheral perfusion	1844	316	(0.17)

Irritability	1842	873	(0.47)
Weak or absent cry	1842	134	(0.07)
Poor or absent response to family members	1846	225	(0.12)
Moaning	1841	216	(0.12)
Poor or absent tone	1846	172	(0.09)
Poor or absent spontaneous motor skills	1846	59	(0.03)
Poor or absent eye contact	1842	101	(0.05)
Moderate or weak general appearance	1844	366	(0.20)
Abnormal lung auscultation	1846	207	(0.11)
Abnormal capillary refill time	1838	49	(0.03)
Abnormal anterior fontanelle	1844	35	(0.02)
Clinical signs of dehydration	1845	17	(0.01)
Rash	1843	112	(0.06)
Joint anomaly	1846	2	(0.00)
Erythemathous throat	1831	139	(0.08)
Rhinitis	1843	804	(0.44)
Acute otitis media	1840	284	(0.15)
Lymphadenopathy	1846	3	(0.00)
Diarrhea	1845	214	(0.12)
Vomiting	1844	167	(0.09)
Hepatomegaly and/ or splenomegaly	1846	14	(0.01)
Phimosis (boys)	1014	140	(0.14)
Circumcision (boys)	1022	31	(0.03)
Laboratory tests			
WBC (/mmm3)*	1833	10917.78	(5200.71)
Lymphocyte (/mmm3)*	1799	4662.20	(2558.21)
Myelaemia	1786	73	(0.04)
C- reactive protein (mg/L)*	731	43.86	(46.42)
Capillary blood glucose (mmlo/L)*	548	5.18	(1.32)
Granulocyte neutrophile (/mmm3)*	1803	4604.60	(3371.77)
Monocyte (/mmm3)*	1772	1382.41	(943.26)
Platelets (/mmm3)*	1812	406096.64	(123983.97)
Fibrinogen (g/L)*	113	3.97	(1.46)
Blood lactate (mmlo/L)*	48	2.91	(1.46)
Blood culture	1848	1068	(0.58)
CSF analysis	1637	1030	(0.63)
WBC count in CSF (/mmm3)*	1148	73.08	(423.91)
RBC count in CSF (/mmm3)*	1101	2523.97	(24754.70)
CSF Gram stain	1219	7	(0.01)
UDT n°1	1827	1499	(0.82)
CBEU n°1	1845	1189	(0.64)
Nitrites in UDT n°1	1476	141	(0.10)
WBC in UDT n°1	1489	380	(0.26)
WBC count in CBEU n°1 (/ml)*	1147	61830409.26	(7.79e+08)
RBC count in CBEU n°1 (/ml)*	1106	1.82e+08	(6.01e+09)
CBEU Gram stain n°1	999	322	(0.32)

Urine analysis positive n°1	1741	335	(0.19)
UDT n°2	1480	58	(0.04)
CBEU n°2	1486	205	(0.14)
Nitrites in UDT n°2	54	6	(0.11)
WBC in UDT n°2	57	24	(0.42)
WBC count in CBEU n°2 (/ml)*	177	730378.64	(3866341.69)
RBC count in CBEU n°2 (/ml)*	165	6192698.24	(77846569.44)
CBEU Gram stain n°2	157	60	(0.38)
Urine analysis positive n°2	168	62	(0.37)
NBS	1794	21	(0.01)
Cells in the NBS	12	6	(0.50)
Granulocyte in the NBS	13	6	(0.46)
NBS Gram stain	13	5	(0.38)
Nasopharyngeal search for viral infection	1835	556	(0.30)
Detection of a virus in nasopharyngeal search	286	79	(0.28)
Stool test	1820	125	(0.07)
Virological analysis of stool sample	1806	167	(0.09)
Detection of a virus in stool sample	52	11	(0.21)
Additional sampling n°1	1848	102	(0.06)
Additional sampling n°2	1848	14	(0.01)
Additional sampling n°3	1848	7	(0.00)
Chest radiography	1832	1194	(0.65)
Pulmonary hyperinflation	1194	66	(0.06)
Alveolar consolidation	1194	56	(0.05)
Pleural effusion	1194	0	(0.00)
Bronchial syndrome	1194	214	(0.18)
Interstitial syndrome	1194	6	(0.01)
Atelectasis	1194	12	(0.01)
Additional radiography n°1	1848	35	(0.02)
Additional radiography n°2	1848	2	(0.00)
Additional radiography n°3	1848	0	(0.00)

N: Total number of available data

Abbreviations: CBEU, cyto-bacteriological examination of the urines; CSF, cerebrospinal fluid; GBS, group B streptococcus; NBS, nasopharyngeal bacterial sampling; NPS, neonatal peripheral samplings; RBC, red blood cell; UDT, urine dipstick test; WBC, white blood cell.

Table S4: AUC of ROC curve to detect antibiotic treatment for stated subjective probabilities by centers

	N	AUC	(95%IC)
All centers	1848	0.819	0.798- 0.839
Center 1	748	0.778	0.739- 0.816
Center 2	261	0.832	0.784- 0.879
Center 3	150	0.774	0.668- 0.880
Center 4	111	0.810	0.728- 0.892
Center 5	110	0.825	0.750- 0.900
Center 6	97	0.876	0.805- 0.947
Center 7	77	0.819	0.707- 0.931
Center 8	76	0.952	0.889- 1.000
Center 9	64	0.900	0.820- 0.980
Center 10	62	0858	0.7650- 0.951

Abbreviations: AUC, area under the curve for the receiver operating characteristic curve
Note: We restricted our analysis to centers with at least 50 observations per center, 5 centers were excluded.

Table S5: Summary of AUC and proportion test (2 tailed) of specificity for mechanical models (M_i) with residuals and with reports compared to observed reports

	Judgment		Decision	
	AUC	(95%IC)	Specificity	Δ
Observed reports	0.855	(0.829-0.881)	0.711	
Models with residuals				
M_1- residuals	0.942	(0.925-0.958)	0.278	-.433**
M_2- residuals	0.971	(0.962-0.980)	0.746	+.035*
M_2^*- residuals	0.973	(0.964-0.982)	0.775	+.064**
Models with reports				
M_1- reports	0.940	(0.923-0.957)	0.340	-.371**
M_2- reports	0.967	(0.958-0.977)	0.696	-.015
M_2^*- reports	0.971	(0.962-0.980)	0.732	+.021

Abbreviations: AUC, area under the curve for the receiver operating characteristic curve; Δ, proportion difference
Note: Models with residuals were estimated separately per group. Models with reports were estimated by pooling observations from the two groups.
*P*values: *<5%; **<1%

Part 2: Sources of probability distortion:

laboratory experiments

Chapter 4: Metacognitive ability predicts learning cue-stimulus association in the absence of external feedback

Marine Hainguerlot[1], Jean-Christophe Vergnaud[1,*] , Vincent de Gardelle[2,*]

[1]Centre d'Economie de la Sorbonne, CNRS UMR 8174, Paris, France

[2]CNRS and Paris School of Economics, Paris, France.

[*]denotes equal contribution

Abstract

Learning how certain cues in our environment predict specific states of nature is an essential ability for survival. However learning typically requires external feedback, which is not always available in everyday life. One potential substitute for external feedback could be to use internal feedback, such as the confidence we have in our decisions. Under this hypothesis, if no external feedback is available, then the agents' ability to learn about predictive cues should increase with the quality of their confidence judgments (i.e. metacognitive efficiency). We tested and confirmed this novel prediction in an experimental study using a perceptual decision task. We evaluated in separate sessions the metacognitive abilities of participants (N=65) and their abilities to learn about predictive cues. As predicted, participants with greater metacognitive abilities learned more about the cues. Knowledge of the cues improved accuracy in the perceptual task. Our results provide strong evidence that confidence can act as an implicit feedback signal, improving learning and performance.

Introduction

Developing knowledge about one's own environment, e.g. learning how environmental cues predict the occurrence of future stimuli or rewards, is essential to optimize our behavior. Past research has shown that such (probabilistic) associations can be learned, for instance using reinforcement mechanisms [1,2] or Bayesian inference [3,4]. Typically, these learning mechanisms require external feedback. How agents learn when external feedback is unavailable remains, however, unclear. This is highly problematic given that one may easily consider situations in which agents make repeated decisions but receive feedback only after a long delay, if at all. For instance, a radiologist inspecting mammograms for potential tumors can only obtain decisive feedback on a case via surgery. In such situations, can the decision makers learn about potential predictive cues (e.g. environmental risk factors such as chemical exposure) that might help improve their decisions?

One intriguing hypothesis is that decision confidence could be useful in such situations. Specifically, even though external feedback is unavailable, the radiologist might still rely on her internal feedback, that is, her sense of confidence in her diagnostic accuracy. For example, let's consider that when examining a group of patients who have all been exposed to a particular chemical product, the radiologist's confidence that they have a malignant tumor is higher than usual. If she uses her confidence in detecting a tumor to learn about the toxicity of the chemical product, after a series of examinations she would infer that being exposed to this chemical product increases the risk of having breast cancer. In support of this hypothesis, several recent studies offer evidence that reinforcement learning can be driven by confidence when external feedback is absent, in particular via midbrain dopamine signals representing the confidence “prediction error” involved in the learning process [5,6].

Here, we develop and test another empirical prediction on the basis of this hypothesis. Specifically, we reasoned that if confidence plays the role of feedback to guide learning about cue-stimulus associations, then successful learning (that is, the ability of the agent to identify the predictive value of a cue) should reflect the quality of the agent’s confidence judgments. To understand why, let’s consider the most extreme cases. If the radiologist can produce perfect confidence judgments, she would assign high confidence to her correct diagnostics and low confidence to her incorrect diagnostics. In a way, the radiologist now has perfect information about the presence or absence of the tumor, which she can link to the presence or absence of chemical exposure, so as to learn whether exposure increases the risk of breast cancer or not. On the other hand, if the radiologist’s confidence judgments poorly discriminate between correct and incorrect diagnostics, i.e. if confidence does not bear any information about diagnostic accuracy, then the radiologist would not be able to learn the association. In other words, we predict that successful identification of an environmental cue should depend on the agent’s ability to evaluate her own decisions, which is a form of metacognition (for a mathematical argument, see the Supplementary Materials).

We set out to test this prediction in a laboratory experiment. To do so, we engaged participants in a difficult perceptual task, which served as a basis for us to evaluate their abilities to form confidence judgments and their abilities to identify predictive environmental cues, in two distinct sessions. As explained above, if confidence in decision is used as a substitute for external feedback, then being able to distinguish between one’s correct and incorrect decisions should be key in identifying the cues. In other words, if our hypothesis is true, the assessment of metacognitive abilities in one experimental session should predict whether or not participants will learn about the cues in the other experimental session. Furthermore, our experiment also allowed us to explore whether the identification of cue-stimulus associations eventually influenced participants' behavior in the perceptual task.

Our perceptual task was as follows: on each trial, participants indicated which of two circles, the left or right, contained more dots (Figure 1). They received no feedback. Two experimental sessions were taken, in a counterbalanced order across participants. In the “confidence session” (512 trials), participants gave a confidence rating after each decision. Confidence quality was measured as the metacognitive efficiency [7], that is, how good one is at distinguishing between correct and incorrect decisions, after controlling for perceptual performance. In the “learning session” (600 trials), a geometric shape (circle, square or diamond) preceded the stimulus. One shape predicted the left category, one predicted the right category (both with probability p=0.75) and one provided no information about the forthcoming category. We informed participants that there were a 'left', a 'right' and a 'neutral' cue but without specifying the probability of the predictive cue. We instructed participants to learn and use the associations between cues and categories, so as to maximize their performance during the task. At the beginning of the cue learning task, we told participants that they would have to identify these cue-stimuli associations at the end of the session. In our result section, we shall define “learners” and “non-learners” on the basis of whether participants indeed successfully identified the associations between the cues and the stimuli. Both sessions started with a working memory test and an initial calibration phase for the perceptual task (for details see the Methods section).

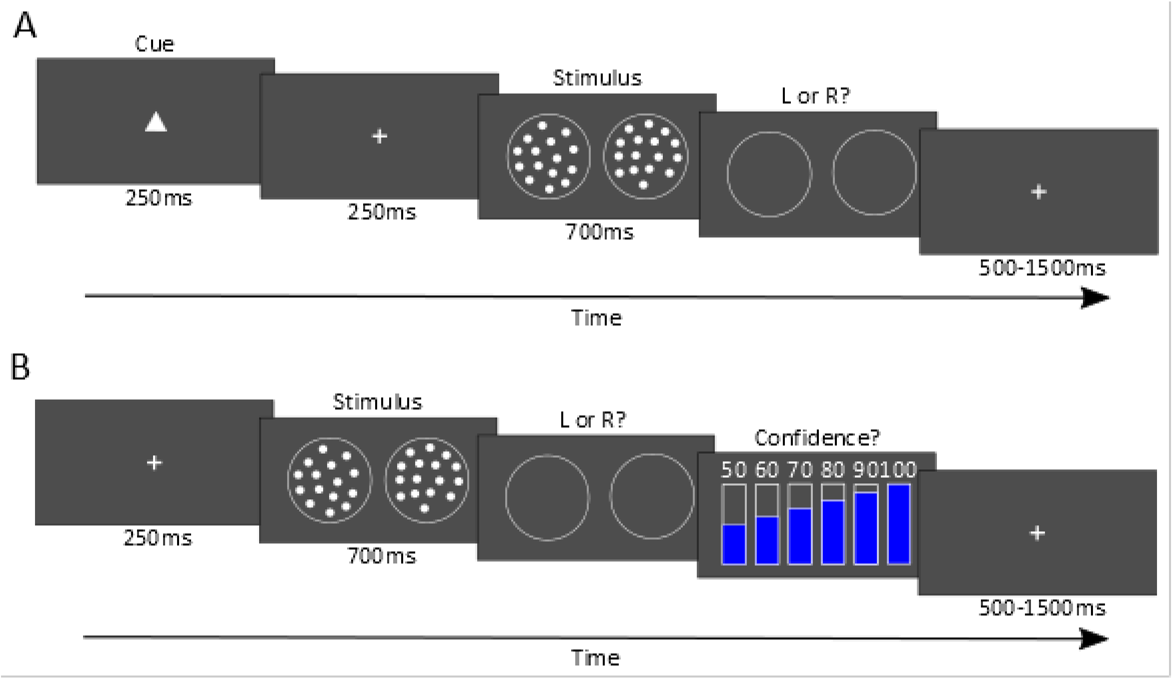


Figure 1: *Experimental Design. In both sessions, participants performed the same perceptual task. On each trial, they had to decide which circle (left or right) contained more dots. (A) Learning session. A cue in the form of a geometric shape (a square, circle or triangle) was presented before the stimulus. Two cues were respectively predictive of the left and right circle whereas one cue was non- predictive. Participants were not informed about the associations between these cues and the stimulus categories. At the end of the session, they had to identify the cue-stimuli associations. (B) Confidence Session. Participants gave a confidence rating after each decision.*

Methods

Participants. 65 individuals (28 females; mean ± SD age, 22.31 ± 3.12 years) were recruited through the Laboratory of Experimental Economics research pool in Paris (LEEP) and gave informed consent to participate. The study was conducted in line with the principles of the Declaration of Helsinki. Participants came to two sessions, which took place 4 days apart.

The experiment was conducted with groups of 15-20 participants. Participants received 13 Euros for participating plus an incentivized bonus as described below.

Ethics statement. Written informed consent was obtained from all participants before the experiment. The research was non-invasive; it involved negligible risks and no collection of nominative/identifying information or health information. Thus, ethics approval was not required under French regulations, and no IRB was consulted before conducting the study.

Summary of the design. Our study involved a simple perceptual categorization task (see Figure 1). On each trial, participants indicated which of two circles, the left or right, contained more dots. They received no feedback. Two experimental sessions were taken, in a counterbalanced order across participants. In the “confidence session” (512 trials), participants gave a confidence rating after each decision. Confidence quality was measured as the metacognitive efficiency [9], that is, the information contained in confidence ratings about the stimuli, after controlling for perceptual performance. In the “learning session” (600 trials), a geometric shape (circle, square or diamond) preceded the stimulus. One shape predicted the left category, one predicted the right category (both with probability p=0.75) and one provided no information about the forthcoming category. We instructed participants to learn and use the associations between cues and categories, so as to maximize their performance during the task. At the end of the session, participants also had to identify these cue-stimuli associations. Both sessions started with a working memory test and an initial calibration phase for the perceptual task.

Perceptual Task. On each trial, after a 250ms fixation cross, two sets of about 100 dots were simultaneously presented for 700ms, one on the left side and one on the right side of the computer screen. Participants had to indicate which set contained more dots, by pressing the corresponding arrow on the keyboard. After the response, the inter-trial interval was jittered between 0.5s and 1.5s. Participants received no feedback about the accuracy of their decision. Response times shorter than 200ms or longer than 2200ms (from stimulus onset) were discouraged by presenting a “too fast” or “too slow” message on the screen. The experiment was run using MATLAB (MathWorks) and Psychotoolbox [13], on screens (resolution 1024 x 768) viewed at normal distance (about 60 cm).

Calibration. Stimulus difficulty was calibrated for each participant at the beginning of each session. Specifically, one circle contained 100 dots while the other circle (the stimulus) contained 100 + *x* dots. The other circle was computed separately for the right (x_r) and left (x_l) stimulus by adjusting with a 2-down 1-up rule [14], to obtain 71% of ”left” or ”right” responses, in two interleaved staircases of 150 trials each. The step size of the staircases was reduced from 20 to 16, 8, 4 and 2 dots on trials 12, 24, 60 and 80 respectively. After the calibration phase, the dots difference x_r and x_l conditional on the right and left stimulus respectively was constant across the session.

Confidence session. Each response was followed by a confidence rating, in which participants indicated their subjective belief that their response just given was correct, on a 6 steps scale ranging from 50% confident (i.e. guess) to 100% confident, in 6 steps of 10%. Participants responded using the numerical keys on the top-left of the keyboard. This confidence rating was incentivized using a probability matching rule [15]. The participant is offered an exchange between his response and a lottery ticket with a probability P of success. The number P is generated randomly on each trial, and compared to the confidence response. If P (the success probability) is greater than the confidence, then the participant’s reward is determined by the lottery. If not, it is determined by the accuracy of the response.

The mechanism was presented to participants as a way to maximize their earnings by providing accurate confidence ratings. Instructions, examples, and a training phase with feedback (40 trials) were included to make sure that participants understood the mechanism. Participants then completed 512 trials in the session.

Metacognitive efficiency measure. We estimated metacognitive sensitivity using the meta-d' method [7] . In a Signal Detection framework, Meta-d' corresponds to the level of type 1 SDT sensitivity (d') that a metacognitively ideal observer would have needed to produce the observed type 2 data. Metacognitive efficiency is defined by the relative measure meta-d'/ d'. If the ratio is equal to one, then the observer is metacognitively ideal. If meta-d'/d' is lower than one, then the observer is metacognitively inefficient. It may also occur that the ratio takes values above one, for instance if additional information is used after the initial choice. Here, we applied the code as provided at http://www.columbia.edu/bsm2015/type2sdt/ with the default settings. In particular, we used the default assumption of "equal variance" between the two stimuli, and the default "cell padding" strategy to avoid empty cells in the confidence x response design, noting that 45 participants out of 65 had at least one empty cell out of 24 (6 ratings x 2 stimuli x 2 responses). We report in the supplementary material the values of meta-d' along with the distributions of confidence ratings for each individual participant.

Learning session. Each trial started with a central cue presented for 250ms, before the fixation cross. The cue was a square, a circle or a triangle. One shape predicted the left category, one predicted the right category (both with probability p=0.75) and one provided no information about the forthcoming category. Participants were not informed about the associations between the cues and the prior probabilities of occurrence of a stimulus but they were informed that there were a 'left', a 'right' and a 'neutral' cue. At the beginning of the cue learning task, they were required to learn about the cue-stimulus associations, in order to optimize their decisions and they were told that at the end of the session, they had to report these associations. Participants completed 600 trials of interleaved sequences of blocks of 8 trials per cue. For each sequence of 8 trials (“block”), the same cue was displayed prior seeing the stimuli. For each block, the predictive cue indicated, in a random order, the forthcoming stimulus correctly on 6 trials (75% valid cue) and the forthcoming stimulus incorrectly on 2 trials (25% invalid cue). Response accuracy was incentivized: participants won 1 point if correct and lost 1 point if incorrect. The value of 1 point was 0.02 €. In addition, participants were informed, at the beginning of the session, that they would be rewarded 2 Euros, 1 or 0 Euro for correctly reporting 3, 1 or none of the associations.

Memory span task. The task consisted of twelve trials. On each trial, participants received a sequence of *m+n* letters and were required to report in the forward order the last *n* letters. The (m,n) combinations were randomly drawn without replacement with *m* = 0, 1, 2 and *n* = 3, 4, 5, 6. Letters were sampled with replacement from a pool (F, H, J, K, L, N, P, Q, R, S, T, Y). After an initial fixation cross of 300ms, letters were presented successively for 300ms on the center of a gray background screen and followed by a 2200ms interval. At the end of the sequence, participants were asked to report the last *n* letters by clicking on the cells of a 4x3 grid displaying the 12 letters of the pool. One point was earned for each item reported in the correct serial position. For example, if participants were instructed to report ”N P T”, they would gain 3 points for responding ”N P T” but 0 point for responding ”L N P”. The maximum score possible was 54 points. The task was programmed in JavaScript and administered to participants through the internet based software REGATE version 9.33. For each participant, we computed the average value of the working memory scores obtained in each session.

Results

Our analysis proceeds in 2 steps. First, we tested our hypothesis that metacognitive efficiency predicted learning, by which we mean the successful identification of the cues by participants. Second, we assessed whether this identification influenced participants' cue usage.

From metacognitive efficiency to successful cue identification

Our main hypothesis related metacognitive efficiency to successful cue identification, which was coded as 1 for participants who correctly classified the 3 cue-stimuli associations (N=35) and 0 (N=30) for participants who identified either 1 (N=27) or 0 association (N=3). Metacognitive efficiency was quantified as the ratio of meta-d' over d' for each individual (M=0.87, SD=0.42). Critically, we found that our main hypothesis was confirmed: metacognitive efficiency was higher in participants who successfully identified the cue-stimuli associations ("learners": M=0.972 SD=0.405) than for participants who did not ("non-learners": M=0.758 SD=0.424), as illustrated in Figure 2A. The difference in metacognitive efficiency between "learners" and "non-learners" was significant (T-test: t(63)=-2.078, p=0.0418).

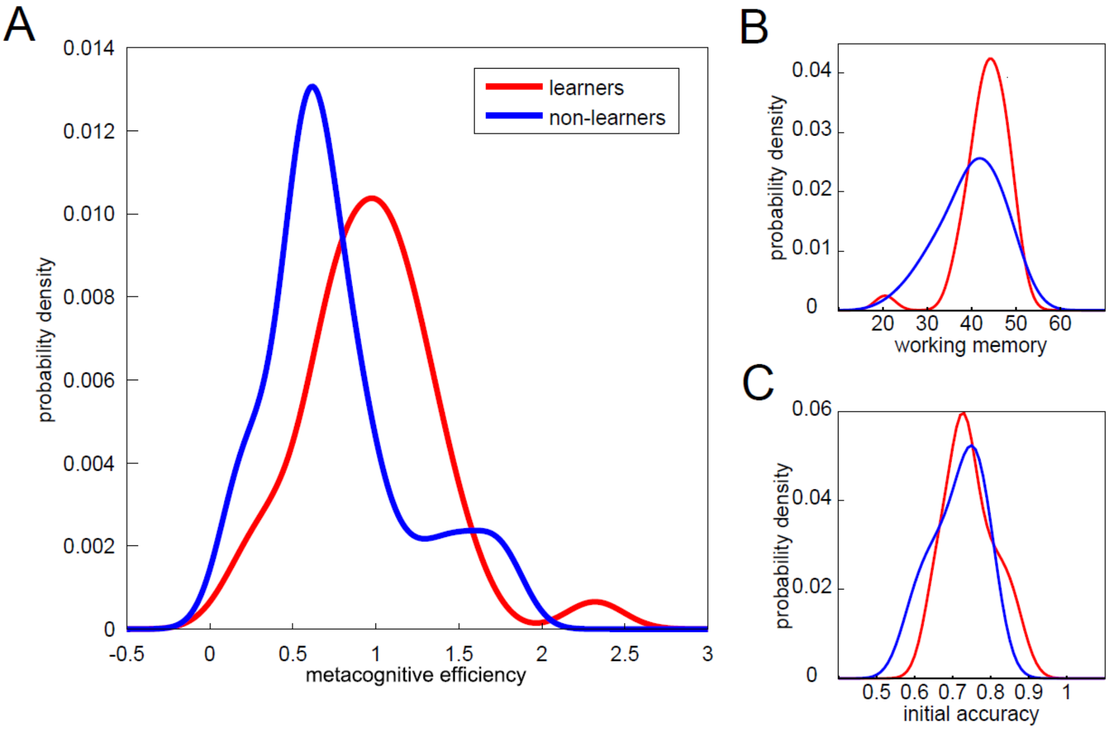


Figure 2*. Predicting cue identification. (A) Distributions of the metacognitive efficiency (ratio of meta-d' over d') across participants who successfully identified the 3 cue-stimuli associations ("learners"), and participants who did not ("non-learners"). (B) Distributions of working memory scores for "learners" and "non-learners". (C) Distributions of the initial accuracy in the cueing session, for "learners" and "non-learners".*

We conducted several robustness checks to ensure that this difference between "learners" and "non-learners" was not due to the some specific features of the metacognitive measure we chose. Firstly, we also found this difference to be significant when we used the log ratio of meta d' over d' instead of the ratio (T-test: t(63)=-2.261, p=0.0272), or when we used a rank test instead of a classic t-test (Rank sum test: z=-2.533, p=0.0113). Secondly, given that the computation of metacognitive efficiency as the ratio of meta-d' over d' can be problematic when participants do not display a lot of variance in their confidence judgments, we verified that the difference between "learners" and "non-learners" was still significant after excluding 7 participants who used the same confidence level in more than 80% of the trials (T-test: t(56)=2.744, p=0.008; Rank sum test: z=3.072, p=0.002). Moreover, we replicated this difference between “learners” and “non-learners” with other measures of metacognitive abilities. The resolution score from the Brier index was higher in participants who successfully learned the cue-stimuli associations compared to participants who did not (M=0.016 SD=0.010; M=0.010 SD=0.009; T-test: t(63)=-2.413, p=0.0188; Rank sum test: z=-2.559, p=0.0105). The difference between the average confidence in correct responses and average confidence in incorrect responses was also greater for “learners” than for “non-learners” (M=0.089, SD=0.063; M=0.058, SD=0.049; T-test: t(63)=-2.197, p=0.0317; Rank sum test: z=-2.323, p=0.0202). These analyses thus confirmed the hypothesized relation between metacognitive efficiency and cue identification.

To ensure that this result was not simply due to inter-individual differences in motivation, perceptual abilities or memory abilities, we conducted a multivariate logit regression analysis in which successful cue identification was predicted from metacognitive efficiency along with additional predictors coding for these factors. In short, although memory and perceptual abilities did facilitate cue identification, metacognitive efficiency still predicted cue identification when we simultaneously controlled for these confounds (β=1.805, se=0.777, p=0.020). We report below the full results of this multivariate regression (see also Table S1 of the Supplementary Materials).

Memory abilities of participants were evaluated using a separate memory test (see Methods) and exhibited a large variability across participants (M=41.7, SD=6.2). We found that greater memory scores were associated with successful cue identification in our multivariate regression (β=0.133, se=0.059, p=0.023). Figure 2B also illustrates how memory scores were greater for “learners” than for “non-learners” (T-test: t(63)=-2.295, p=0.0251). This is expected given that working memory is needed to learn: in order to update her estimation of a cue-stimulus association, the participant needs to remember both the cue presented at the beginning of the trial and her previous estimate of the cue-stimulus associations for this cue.

To control for perceptual abilities in the task, we also included in our multivariate regression a predictor coding for the accuracy in the perceptual task at the beginning of the cueing session (proportion of correct responses over trials 1-96). Although we attempted to calibrate performance at 71% before the experiment, this initial accuracy still varied across participants (M=73%, SD=6%), and we expected this initial accuracy to be predictive of successful learning of the cues. Indeed, any advantage in the perceptual task would increase the amount of information on which learning can be based (for an illustration see the Supplementary Materials). Our multivariate regression confirmed that initial accuracy significantly increased the probability of learning the cues ($\beta=11.812$, se=5.645, $p=0.036$). Figure 2C illustrates how, although the modes of the two distributions appear to be roughly the same, the range of initial accuracy is shifted for “learners” compared to “non-learners”, resulting in a difference between the two groups (T-test: $t(63)=2.029$, $p=0.0467$).

Finally, to control for motivation to do well in the task, we introduced three additional predictors in our regression: one coding for calibrated difference in the number of dots (which would be lower for more motivated participants), one coding for the order in which the sessions were performed (confidence or learning session first) and one coding for the interaction between metacognitive efficiency and this session order. None of these predictors had an effect on successful learning (all $p>0.49$).

From cue identification to cue usage

Our next series of analysis focused on the impact of cue identification on performance in the task. Overall, “learners” exhibited better performance in the task compared to “non-learners” (M=0.754, SD=0.055 vs. M=0.704, SD=0.045; t(63)=4.001, p=0.0002). However, the direction of the influence remains unclear in this analysis: cue identification might lead to an increase in performance, but performance should also help for cue identification, as we explained in the previous section.

To better isolate the effect of cue identification on performance, we thus focused on how participants’ responses followed the information provided by the right and left predictive cues. We found that responses of “learners” were more congruent with the cues than responses of “non-learners” (M=0.684, SD=0.059 vs. M=0.602, SD=0.059; t(63)=5.604, p<0.001). Furthermore, comparing response accuracy when the cue was valid (i.e. in 75% of the trials) to when the cue was invalid (in 25% of the trials), we found that “learners” were indeed more likely to be correct when the cue was valid than when the cue was invalid (valid: M=0.798, SD=0.065; invalid: M=0.657, SD=0.114; t(34)=6.109, p<0.001). Performance of “non-learners”, by contrast, was not modulated by cue validity (valid: M=0.705, SD=0.070; invalid: M=0.702, SD=0.080; t(29)=0.136, p=0.89), and the effect of cue validity was significantly different between the two groups (t(63)=4.418, p<0.001). In other words, not only have “learners” identified that the cue provided valuable information about the stimulus, but they also used this information to increase their performance (see also Table S2). Figure 3 further illustrates how perceptual decisions evolved over time for “learners” and “non-learners”, as a function of the stimulus presented and the cue. For each stimulus there are thus 3 curves, corresponding to the 3 cues, which get separated as time passes, for “learners” but not for “non-learners”.

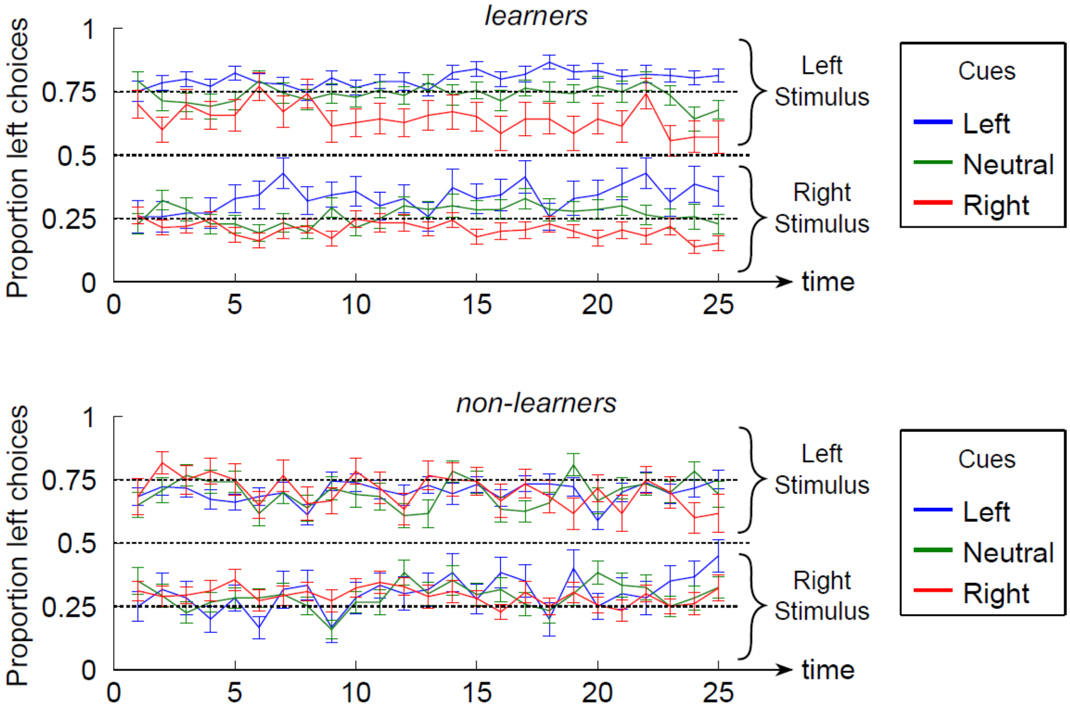


***Figure 3**. Timecourse of the cue influence. The proportion of left choices is plotted for each stimulus and cue, as a function of time, separately for "learners" and "non-learners" participants. Each point in time represents a block of 8 trials. Error bars represent the mean and the standard error of the mean across participants, in each group.*

To further investigate how "learners" and "non-learners" differed in their use of the cue, we conducted an analysis based on Signal Detection Theory (Green & Swets, 1966) to separate perceptual sensitivity and decision criterion. In particular, we evaluated the difference between the decision criteria adopted for the cues that predicted left and right responses. Critically, "learners" adopted different criteria for the right and left predictive cues (t(34)=5.991, p<0.001) whereas "non-learners" did not seem to have changed their criteria in a consistent manner (t(29)=0.061, p=0.9520), resulting in a significant difference between the two groups (t(63)=4.479, p<0.001). For completeness, we also looked at perceptual sensitivity (d', averaged across all 3 cues), and found that it was also higher for "learners" compared to “non-learners” (M=1.334, SD=0.436 vs. M=1.113, SD=0.302; t(63)=2.332, p=0.0229). This may reflect the fact that perceptual sensitivity provides information that can be used to learn about the cues, such that participants with greater sensitivity are more likely to end up as “learners” compared to participants with lower sensitivity.

As a final step, we investigated whether metacognitive efficiency directly predicted cue usage across participants. The correlations between metacognitive efficiency and average performance (r=-0.020, p=0.877) or congruency with the cue (r=-0.027, p=0.828) were not significant. However, we could find an effect of metacognitive efficiency on performance when we controlled for inter-individual differences in memory, perceptual abilities and motivation. To do so, we conducted a multivariate regression analysis (similar to the one in the previous section) in which final performance (here defined as the mean accuracy after trial 96 in the cueing session) was predicted from metacognitive efficiency, along with memory scores, the calibrated difference in the number of dots, the session order, the interaction between metacognitive efficiency and session order, as well as initial performance in the cueing session (mean accuracy before trial 96) that may capture inter-individual differences in perceptual abilities (possibly due to an imperfect calibration of the stimuli). We found, indeed, that final accuracy was significantly predicted by initial accuracy (β=0.643, se=0.076, p<0.001). More critically, final accuracy in the cueing session was also positively affected by metacognitive efficiency (β=0.027, se=0.011, p=0.014). None of the other predictors were significant in this analysis. The full results of this regression are reported in Table S1 of Supplementary Materials.

Discussion

Identifying cues that can help performing a difficult task is a useful ability, in a world where agents have to make decisions under uncertainty. Here, we show that inter-individual heterogeneity in the ability of participants to evaluate their own performance, measured in an independent session, could predict the successful classification of predictive cues. Importantly, this successful learning of predictive cues also translated into actual benefits in terms of task performance.

Our study was based on the hypothesis that confidence could be used as an “internal feedback” signal when external feedback is not available for learning about cues, and our result seem to confirm this hypothesis. Before going further, however, we would like to address one alternative account of the relation we found between cue identification and metacognitive efficiency. Specifically, metacognitive efficiency may have no impact on learning during the task and may only help participants to introspect and evaluate their use of the cues at the end of the task, when asked to report the cue-stimulus associations. If so, participants with higher metacognitive efficiency would be better able to report the cue-stimulus associations that correspond to their cue usage. We tested this prediction in our data by looking at whether the ordering of the cues in their post-hoc identifications corresponded to the ordering of the cues in terms of cue usage (using the decision criteria calculated with Signal Detection Theory). We found 14 participants for whom the two orderings were indeed equal, and they did not have a better metacognitive efficiency than the other participants (T-test: t(63)=0.0129, p=0.99). In sum, metacognitive efficiency did not seem to increase the correspondence between cue usage and cue identification. Thus, our working hypothesis is that metacognitive efficiency helps participants to learn about the cues while they are doing the task.

Our study relates in particular to two recent studies investigating the role of confidence in learning, although with different learning problems. One study used a category learning task: participants had to learn how perceptual stimuli defined in a two-dimensional space were mapped onto two category labels [5]. In this study, participants could learn the mappings from some ‘association’ trials in which exemplars were presented in conjunction with the correct category label, while on the ‘test’ trials they had to categorize the stimuli and give their confidence. The other study investigated a different form of learning, namely perceptual learning [6]. In this paradigm, participants were engaged in a perceptual categorization task, and they knew the mappings between responses and categories from the start (e.g. ‘your task is to press left if the stimulus is tilted clockwise relative to the reference’). Performance nevertheless improved over the course of the experiment, as if participants could learn to better extract the information from the stimuli. Our study implements yet another form of learning, as participants here received two types of information, a perceptual stimulus and a symbolic cue, and had to learn about the symbolic cue on the basis of their estimation of perceptual performance. Together, this set of studies thus provides support for the involvement of confidence in learning, in an increasing range of situations.

One other noticeable difference between the present study and the two aforementioned studies is that in these studies confidence judgments were collected during the learning sessions, while in our paradigm confidence was only evaluated from a distinct session taking place on a different day. This is interesting because it shows that confidence does not need to be explicitly evaluated and verbalized to be used for learning. These results echo previous research showing that confidence is evaluated automatically during decision-making [8]. We would also argue that since confidence was used to guide the evaluation of a symbolic cue, it should be readily available at an abstract level of representation [9,10].

The present work was motivated by the hypothesis that confidence can be used to guide learning, when no external feedback is available. Our study provides support for this hypothesis, and extends past research along these lines, as discussed above. We believe nonetheless that many questions remain open regarding the ability of humans to learn from their own confidence. How fast can observers learn in this situation? Are observers able to evaluate the precision of what they learn, and to use this precision in optimal manner? These questions have been addressed in situations in which observers learn about event probabilities from the past history of events [11,12]. They remain unanswered in situations in which learning needs to be based on confidence.

The specific learning mechanism that might be at play also deserves more scrutiny. In our paradigm, participants might have updated their estimate of the value of the cue from their confidence in the current perceptual decision. Past studies [5,6] have specifically put forward a particular form of confidence-based reinforcement learning, in which the prediction error on confidence would be used to update the synaptic weights between sensory detectors and decision units, so as to increase perceptual performance over time. Yet, alternative mechanisms could be considered as well. For instance, a reinforcement learning mechanism can achieve optimal inference if the learning rate, rather than being a constant, changes appropriately as a function of the estimated volatility of the environment. Humans seem able to do so when they learn from event history [4]. Whether this holds true in confidence-based learning without external feedback is an open empirical question. Do we know when to trust our own confidence? We believe that this issue opens exciting perspectives for further research.

References

1. Sutton, R. S. & Barto, A. G. *Reinforcement Learning*. (MIT Press, 1998).
2. Rescorla, R. & Wagner, A. in *Classical Conditioning II: Current Research and Theory* (eds. Black, A. & Prokasy, W.) 64–99 (Appleton-Century-Crofts, 1972).
3. Courville, A. C., Daw, N. D. & Touretzky, D. S. Bayesian theories of conditioning in a changing world. *Trends Cogn. Sci.* **10,** 294–300 (2006).
4. Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. Learning the value of information in an uncertain world. *Nat Neurosci* **10,** 1214–21 (2007).
5. Daniel, R. & Pollmann, S. Striatal activations signal prediction errors on confidence in the absence of external feedback. *NeuroImage* **59,** 3457–3467 (2012).
6. Guggenmos, M., Wilbertz, G., Hebart, M. N. & Sterzer, P. Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife* **5,** (2016).
7. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* **21,** 422–430 (2012).
8. Lebreton, M., Abitbol, R., Daunizeau, J. & Pessiglione, M. Automatic integration of confidence in the brain valuation signal. *Nat Neurosci* **18,** 1159–1167 (2015).
9. de Gardelle, V. & Mamassian, P. Does Confidence Use a Common Currency Across Two Visual Tasks? *Psychol. Sci.* **25,** 1286–1288 (2014).
10. de Gardelle, V., Le Corre, F. & Mamassian, P. Confidence as a Common Currency between Vision and Audition. *PLOS ONE* **11,** e0147901 (2016).
11. Meyniel, F., Schlunegger, D. & Dehaene, S. The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLOS Comput. Biol.* **11,** e1004305 (2015).
12. Meyniel, F. & Dehaene, S. Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proc. Natl. Acad. Sci.* **114,** E3859–E3868 (2017).
13. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10,** 433–436 (1997).
14. Levitt, H. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.* **49,** Suppl 2:467+ (1971).

15. Massoni, S., Gajdos, T. & Vergnaud, J.-C. Confidence measurement in the light of signal detection theory. *Front. Psychol.* **5,** (2014).

Supplementary Material

SUPPLEMENTARY TABLES

	(1) Logit Cue identification			(2) OLS Final performance		
	b	*se*	*p*	*b*	*se*	*p*
Metacog	1.805	0.777	0.020	0.027	0.011	0.014
W. Memory	0.133	0.059	0.023	0.001	0.001	0.172
Initial accuracy	11.812	5.645	0.036	0.643	0.076	<0.001
Calibrated dots	0.074	0.107	0.490	0.003	0.002	0.077
Order	-0.034	0.590	0.954	0.013	0.009	0.152
Metacog*Order	-0.002	0.305	0.996	0.006	0.004	0.176

Table S1: *Multivariate regression results. (1) Logit regression with successful cue identification as a dependent variable. (2) OLS regression with final performance as a dependent variable. Explanatory variable: metacognitive efficiency. Control variables: working memory, initial accuracy, calibrated difference in the number of dots, order in which the sessions where performed (=1 if confidence session first, 0 otherwise) and interaction term between metacognitive efficiency and session order.*

		Response rate left	Accuracy, All trials	Accuracy, valid cue trials	Accuracy, invalid cue trials	SDT estimates	
						c	d’
« Learners »	Neutral cue	0.502 (0.071)	0.737 (0.068)	-	-	0.004 (0.230)	1.340 (0.479)
	Left cue	0.685 (0.076)	0.767 (0.062)	0.801 (0.081)	0.665 (0.138)	0.215 (0.287)	1.354 (0.488)
	Right cue	0.316 (0.093)	0.758 (0.063)	0.795 (0.096)	0.648 (0.132)	-0.240 (0.361)	1.308 (0.431)
« Non-learners »	Neutal cue	0.492 (0.107)	0.703 (0.050)	-	-	-0.022 (0.333)	1.126 (0.325)
	Left cue	0.598 (0.095)	0.701 (0.063)	0.701 (0.100)	0.700 (0.117)	-0.010 (0.291)	1.095 (0.324)
	Right cue	0.393 (0.081)	0.707 (0.069)	0.708 (0.095)	0.704 (0.110)	-0.014 (0.246)	1.117 (0.388)

Table S2: *Descriptive statistics on cue usage for “learners” (N=35) and “non- learners” (N=30). Response rate left, Average accuracy rate, accuracy rate conditional on valid and invalid cues, and SDT estimates decision criterion (c) and sensitivity (d’); with standard deviation reported between parentheses.*

	Mean (SD)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(1) Cue learn	+0.54 (0.50)	1	+0.25 *	+0.27 *	+0.29 *	+0.28 *	+0.11	-0.05	+0.25 *	+0.45 ***	+0.48 ***	+0.58 ***	+0.58 ***	-0.22	+0.49 ***	+0.28 *	+0.13
(2) Metaratio	+0.87 (0.42)		1	+0.93 ***	+0.65 ***	+0.12	-0.07	-0.10	-0.21	-0.02	+0.02	-0.03	-0.03	+0.01	-0.03	-0.02	-0.10
(3) log(metaratio)	-0.27 (0.57)			1	+0.63 ***	+0.10	-0.04	-0.06	-0.15	+0.05	+0.09	-0.02	+0.01	+0.07	-0.04	+0.05	-0.03
(4) Brier Resol.	+0.01 (0.01)				1	+0.17	-0.09	-0.09	-0.03	+0.13	+0.16	+0.02	+0.05	+0.07	-0.01	+0.14	+0.22
(5) W. memory	+41.72 (6.21)					1	-0.33 **	+0.06	-0.08	+0.01	+0.03	+0.06	+0.02	-0.09	+0.05	-0.01	+0.09
(6) Cal. Dots	+9.95 (3.36)						1	-0.24	+0.47 ***	+0.47 ***	+0.45 ***	+0.26 *	+0.40 ***	+0.23	+0.14	+0.46 ***	+0.19
(7) Order	+0.54 (0.50)							1	-0.06	+0.00	+0.02	+0.06	+0.02	-0.12	+0.09	-0.04	-0.08
(8) Initial perf.	+0.73 (0.07)								1	+0.84 ***	+0.77 ***	+0.46 ***	+0.67 ***	+0.32 **	+0.21	+0.77 ***	+0.11
(9) Overall perf.	+0.73 (0.06)									1	+0.99 ***	+0.59 ***	+0.82 ***	+0.33 **	+0.30 *	+0.92 ***	+0.27 *
(10) Final perf.	+0.73 (0.06)										1	+0.60 ***	+0.82 ***	+0.31 *	+0.31 *	+0.92 ***	+0.29 *
(11) Cong. Rate	+0.65 (0.07)											1	+0.94 ***	-0.54 ***	+0.93 ***	+0.30 *	+0.10
(12) Acc. : valid	+0.75 (0.08)												1	-0.21	+0.75 ***	+0.59 ***	+0.16
(13) Acc. : invalid	+0.68 (0.10)													1	-0.79 ***	+0.58 ***	+0.11
(14) Criterion bias	+0.25 (0.46)														1	-0.00	+0.06
(15) d' : cue sess	+1.23 (0.39)															1	+0.36 **
(16) d' : conf. sess	+1.21 (0.43)																1

Table S3: *Descriptive statistics and Pearson correlations for all the variables. N= 65 participants. p-values: *p<0.05; **p<0.01; ***p<0.001*

The variables are: (1) Successful identification of the cues after the cueing session. (2) Ratio of Meta-d' over d'. (3) Logarithm of this ratio. (4) resolution index from the Brier score. (5) average score in the working memory tests. (6) Calibrated difference in number of dots in the cueing session. (7) Order of the cueing and confidence session. (8) Initial performance in the cueing session. (9) Overall performance in the cueing session. (10) Final performance in the cueing session. (11) Rate of responses congruent with the predictive cue. (12) Accuracy in the valid cue trials. (13) Accuracy in the invalid cue trials. (14) Difference in the decision criteria for the left and right predictive cues. (15) d' in the cueing session. (16) d' in the confidence session.

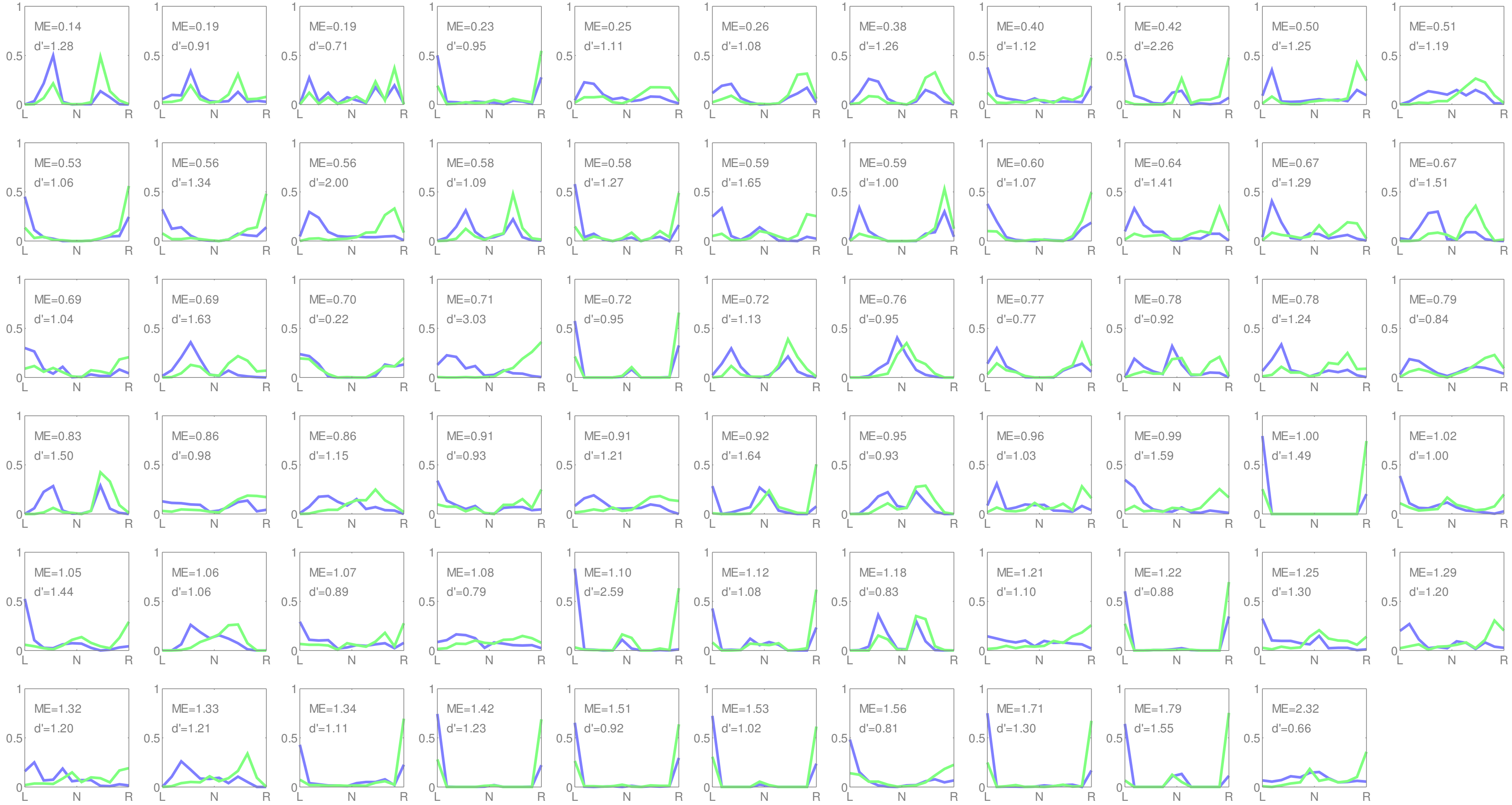


Figure S1. *Individual data for the confidence session**.** For each individual, the figure indicates the metacognitive efficiency (ME), the perceptual d' and plots the distributions of response x confidence ratings (from high confidence left to high confidence right response) for the left stimulus (in blue) and the right stimulus (in green).*

SUPPLEMENTARY ANALYSES

Mathematical argument

We propose here a formalization of our hypothesis linking metacognition to cue identification. We claim that the metacognitive ability of a participant, that is, her ability to assign a lower confidence to her errors, is positively related to her ability to identify the predictive values of the cues. This claim rests on the assumption that both metacognition and learning rely on how well the participant can process the information provided by the stimuli.

On a given trial, after receiving the stimulus ($S = R\ or\ L$) the observer would form a belief about the stimulus being *R*, noted $P(S = R)$, or *L*, noted $P(S = L)$. This belief will be the basis for the observer's decision and confidence judgment: when $P(S = R) > P(S = L)$, the observer selects the response R and her confidence is equal to $P(S = R)$, and when $P(S = R) < P(S = L)$, she reports L with confidence $P(S = L)$.

In a cue learning situation, this belief would be the basis for identifying the predictive value of the cue, that is, the probability π of the stimulus R when the cue is present. The observer can form an estimate $\hat{\pi}$ of this value by considering her average belief about the presence of the stimulus R in these trials.

To formally relate $\hat{\pi}$ to the metacognitive ability of the observer, let us examine how it depends on the observer's average confidence when she is correct ($\overline{C_C}$) and her average confidence when she is incorrect ($\overline{C_I}$). As there are two types of stimuli and two types of responses, there will be 4 types of trials, described in the table below. Here, we also assume that the subject's accuracy rate A and confidence when correct or incorrect do not depend on the stimulus being R or L, for sake of simplicity.

Stimulus presented (with probability p)	Response made (with probability p)	Confidence in the response	Belief about R being present
R, with p= π	R, with p= A	$\overline{C_C}$	$\overline{C_C}$
	L, with p= $1-A$	$\overline{C_I}$	$1-\overline{C_I}$
L, with p= $1-\pi$	L, with p= A	$\overline{C_C}$	$1-\overline{C_C}$
	R, with p= $1-A$	$\overline{C_I}$	$\overline{C_I}$

Across these 4 cases, we can define the net belief that the stimulus R was presented, that is, the estimate $\hat{\pi}$ made by the observer.

$$\hat{\pi} = \pi A\overline{C_C} + \pi(1-A)(1-\overline{C_I}) + (1-\pi)A(1-\overline{C_C}) + (1-\pi)(1-A)\overline{C_I}$$

Furthermore, if we assume that the observer's mean confidence is equal to her accuracy rate A, we can produce a relation between $\hat{\pi}$, A and the difference between $\overline{C_C}$ and $\overline{C_I}$, which we will denote by r for the resolution of confidence with respect to errors. This difference quantifies metacognitive ability in a simplistic manner but it is appropriate for the present illustration.

The assumption that the average confidence is A across all trials gives us: $A\overline{C_C} + (1-A)\overline{C_I} = A$

By definition we also have $r = \overline{C_C} - \overline{C_I}$

Combining these two equations, we find that $\overline{C_C} = A - Ar + r$ and that $\overline{C_I} = A - Ar$

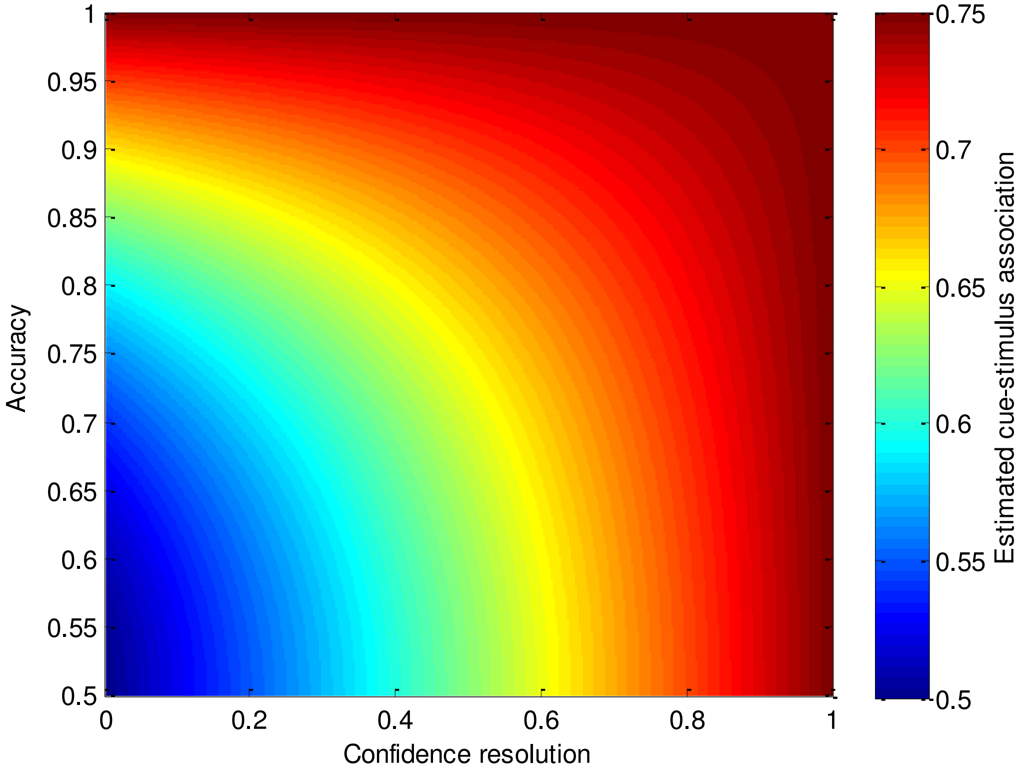
Note that the extreme case of perfect error detection would correspond to $r = 1, \overline{C_C} = 1, \overline{C_I} = 0$ and that the extreme case of null resolution would correspond to $r = 0, \overline{C_C} = A, \overline{C_I} = A$

Introducing these expressions in our calculation of $\hat{\pi}$ yields, after simplification:

$$\hat{\pi} = \pi - 2A(2\pi - 1)(1 - A)(1 - r)$$

Thus, for any given π greater than 0.5, and assuming that both A and r are between 0 and 1, this expression tells us that $\hat{\pi}$ should always fall below π, and will get closer to π when accuracy gets closer to 1 or when resolution gets closer to 1.

The figure below illustrates how $\hat{\pi}$ depends on *A* and *r*, for a target value of $\pi = 0.75$. We clearly see that the estimated probability of the cue-stimulus association approaches the target value of .75 (in red) when accuracy and confidence resolution increases. In other words, both perceptual ability and metacognitive ability improve cue identification.



Chapter 5: Perceptual overconfidence and suboptimal use of symbolic cues: A theoretical and empirical analysis

Marine Hainguerlot[1], Thibault Gajdos[2], Jean-Christophe Vergnaud[1¶], Vincent de Gardelle[3¶*]

[1]Centre d'Economie de la Sorbonne CNRS UMR 8174, Paris, France.

[2]Aix Marseille University, CNRS, LPC, Marseille, France

[3]CNRS and Paris School of Economics, Paris, France.

¶ denotes equal contribution

Abstract

Whereas our perceptual systems appear able to combine sensory cues optimally, humans are often suboptimal when they have to integrate non-sensory information: observers typically exhibit a conservatism bias by which they under-react to cues that are informative and relevant for the task. We describe here a computational model in which this conservatism is due to the observers' overconfidence in their sensory abilities. Within this model, we derive theoretical measures of perceptual overconfidence, conservatism, and the associated suboptimal performance in a task that requires combining sensory and non-sensory cues. We tested this model in a psychophysical experiment on a large group of participants (N=69), in which we measured in separate sessions the perceptual overconfidence of observers and their performance during a cue combination task. Critically, our data reveals that perceptual overconfidence and conservatism in the presence of cues were both present, but that they were uncorrelated across participants, calling for a reassessment of the candidate model. Further analyses indicated that participants' overconfidence correlated with a decrease in decision sensitivity when decision cues were offered, which mediated the relation between overconfidence and suboptimal performance in the task. Our findings offer new perspectives on how human observers combine sensory and non-sensory information towards decision making.

Introduction

Optimal combination of sensory cues in perceptual systems

Many studies have now documented the fascinating ability of our perceptual systems to achieve Bayes-optimal inference when combining information from different sensory cues. Our perceptual systems appear to give weight each cue with its reliability, as an optimal observer would do. This optimal cue combination has been demonstrated within the visual system, which ideally combines low-level cues for localizing texture edges [1] or the slant of a surface [2,3], but also across modalities, as observers optimally combine haptic and visual information about the size of an object [4,5], or visual and auditory information about the location of a sound source [6]. Together these studies emphasize the remarkable ability of our perceptual systems to correctly evaluate the reliability of, and to use appropriately, each source of information. This ability contrasts strikingly with the behavior of agents in tasks that involve non-sensory information.

Suboptimal "conservatism" in the face of symbolic cues

Before a pedestrian decides to cross the road, she considers the speed and acceleration of the car that is approaching, and the color of the pedestrian light which also gives some information about whether this car should stop or not. Here, the evaluation of the car's behavior is based on both sensory inputs that can be used to estimate whether the car is stopping or not, and a symbolic cue that indicates that the car has a high probability to stop or not. Yet, when a sensory stimulus is accompanied by a symbolic cue providing statistical information about the nature of the stimulus, observers do not seem able to perform optimal combination. Although they are able to combine the two pieces of information together, observers typically under-react to the information provided by the symbolic cue. This phenomenon is sometimes referred to as “sluggish beta” or “conservatism”. It is a well-established bias in human decision-making, demonstrated in psychophysical experiments using basic visual stimuli [7–10], in memory tasks [11,12] but also in experiments emulating real-world decisions such as, for instance, engineering quality-control decisions [13–15], or military decisions about a combat target being a friend or a foe [16].

Overconfidence in human decisions.

In the next section, we outline a model that connects conservatism to overconfidence, that is, the over-estimation of one’s own ability. Overconfidence is also a well-established bias in human decisions [17]. It can be observed in a wide variety of tasks, including not only perceptual tasks [18–20] and sensory-motor tasks [21], but also knowledge tasks [22] and tasks that mimic real-life decisions in the medical, economic or military domain [23–26].

A decision model relating overconfidence to conservatism.

Our theoretical model for decision-making is formulated in the framework of statistical decision and signal detection theory [8]. Here, in each trial the agent has to identify a hidden state of nature as either *A* or *B*, on the basis of two elements: a sensory cue (*x*) and an abstract/symbolic cue (*y*). Both the sensory and the abstract cues are generated by the hidden state of nature, in an independent fashion. Both cues thus provide statistical information about the hidden state of nature, and it is assumed that the agent knows the generative models linking the states of nature to the observed values of the cues. For instance, the agent knows that state *A* generates *y=1* with a 75% probability and *y=0* with 25% probability, and vice-versa for state *B*. Similarly she would know how the value *x* is distributed when state *A* is presented and when state *B* is presented. In addition, the agent may have a prior belief about the occurrence of *A* or *B* in the current trial. With all this in mind, and using Bayes' rule, the agent can thus evaluate which state of nature was present, by computing the ratio of the posterior probabilities of *A* or *B* (eq. 1a).

$$\frac{P(A|x,y)}{P(B|x,y)} = \frac{P(A)}{P(B)} \times \frac{P(x|A)}{P(x|B)} \times \frac{P(y|A)}{P(y|B)} \qquad \text{(eq. 1a)}$$

Formally, it is equivalent to consider a decision variable *DV* that is defined as the logarithm of this posterior probability ratio. The agent that would respond according to the sign of *DV* would maximize expected accuracy. In the present situation, one can conveniently express this decision variable as the sum of 3 quantities (eq. 1b): the log-odds of the prior probability (*LP*), the log-odds of the probabilistic information provided by the sensory information (*LS*), and the log-odds of the probability provided by the abstract cue (*LC*).

$$DV = \log\left(\frac{P(A|x,y)}{P(B|x,y)}\right) = \log\left(\frac{P(A)}{P(B)}\right) + \log\left(\frac{P(x|A)}{P(x|B)}\right) + \log\left(\frac{P(y|A)}{P(y|B)}\right) = LP + LS(x) + LC(y)$$

(eq. 1b)

How would overconfidence affect this ideal decision making process? Typically, overconfidence is defined as the overestimation of one's own abilities, in this case the ability to process the sensory signal. Overconfidence thus corresponds to the overestimation of the precision of the generative model of sensory signals, which would result in the overweighing of these signals in the formation of the decision variable. Note that overconfidence does not apply to the symbolic cues, since their reliability is already explicitly established and known to the agent, and does not depend on the ability of the agent. As a consequence, the relative influence of the sensory cue and the symbolic cue will be affected, with the symbolic cue having less impact on the decision variable than what it should have under the optimal conditions. In other words, when the agent is overconfident about a sensory signal, it is more difficult for the symbolic cue to provide enough evidence so as to counteract the sensory signal. In sum, the overconfident agent would under-react to informative and relevant decision cues, which is exactly what is described as "conservatism" or "sluggish beta" in the empirical studies mentioned in the previous section. We note that the link described here between overconfidence and conservatism is similar to propositions made earlier in the context of perceptual tasks [27] and financial decisions [28,29]. Importantly, overconfidence and conservatism cause a loss in performance, because the agent is not following the decision rule that maximizes expected accuracy.

Our empirical approach

In the present study, our goal was to evaluate the computational scheme summarized above by measuring in the same decision-makers both 1) overconfidence about sensory signals and 2) suboptimal performance when sensory signals are accompanied by symbolic cues. As detailed below, the model predicts a quantitative link between these two measures, which should be mediated by conservatism, that is, the insufficient reaction to the information provided by the symbolic cue. We tested this quantitative prediction in a laboratory experiment, using a psychophysical task that allowed us to consider precisely defined measures of task performance, overconfidence and cue combination.

In a nutshell, we engaged participants (N=69) in a simple perceptual decision task: they had to identify which of two sets presented on the computer screen contained more dots (see Fig 1). Each participant completed two experimental sessions, 4 days apart. In the “confidence” session, after each decision participants indicated their subjective probability of success on a quantitative scale. In the “cueing” session, each trial included a cue indicating with 75% validity the nature of the forthcoming stimulus. Before the task, the meaning and validity of these cues was fully explained to participants, who were instructed to optimally integrate the cue with the stimulus information on each trial. We then tested whether overconfidence measured in the “confidence session” would relate to the suboptimal use of the symbolic cue information, as predicted by our computational model (see Methods).

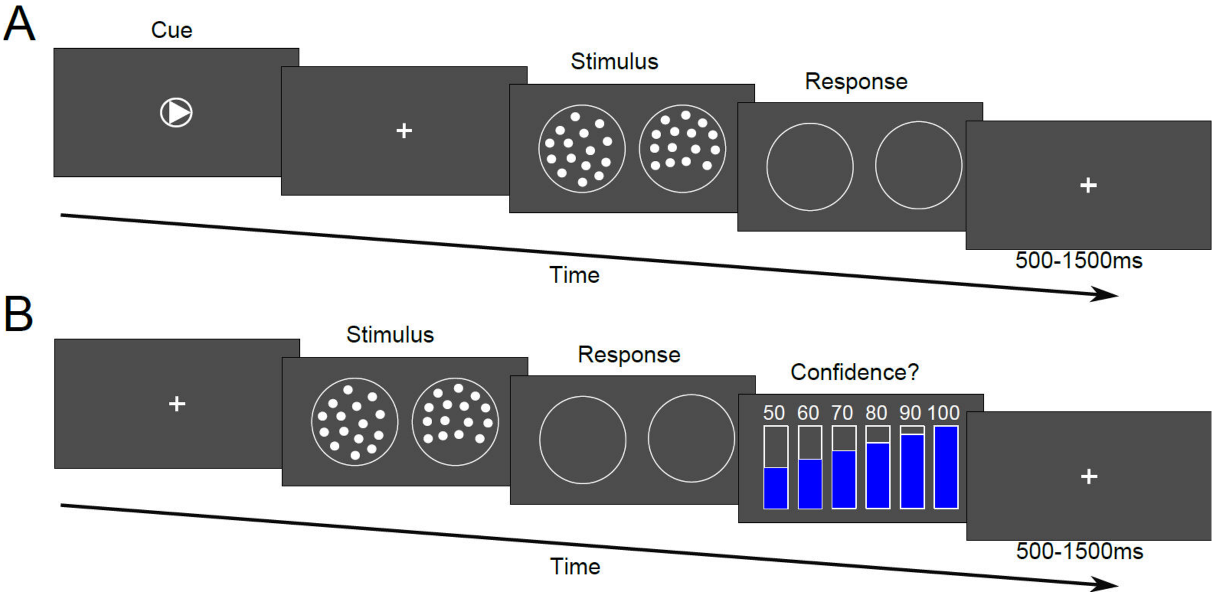


Fig 1. Experimental paradigm. *Participants had to indicate which circle (left or right) contained more dots.* ***(A)*** *In the cueing session, stimuli were presented after a neutral or 75% valid cue that participants had to optimally use to make their decisions.* ***(B)*** *In the confidence session, decisions were followed by confidence judgments on an incentivized probability rating scale.*

Methods

Participants.

69 individuals (39 females; mean age = 23 years, SD = 2.5 years) were recruited through the Laboratory of Experimental Economics of Paris. This sample size resulted from the choice to have a large sample and to conduct the experiment in groups of 10-15 participants. Participants received 13 Euros for participating plus an incentivized bonus described below.

Ethic statement.

The study was conducted in line with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants before the experiment. No nominative/identifying information was collected. No health information was collected from participants other than gender and age. The research involved negligible risks. In this situation, as per current French regulations, ethics approval was not required, therefore no IRB was consulted before conducting the study.

Stimuli and Task.

There were two sessions, a confidence session and a cueing session, which took place 4 days apart. The order of the two sessions was counterbalanced across participants. The experiment was run using MATLAB (MathWorks) and Psychotoolbox [50], on screens (resolution 1024 x 768) viewed at normal distance (about 60 cm). On each trial, after a 250 fixation cross, two sets of about 100 dots were simultaneously presented for 700ms, one on the left side and one on the right side of the computer screen. Participants had to indicate which set contained more dots, by pressing the corresponding arrow on the keyboard. After the response, the inter-trial interval was jittered between 0.5s and 1.5s. Participants received no feedback about the accuracy of their decision. Response times shorter than 200ms or longer than 2200ms (from stimulus onset) were discouraged by presenting a "too fast" or "too slow" message on the screen.

Calibration.

Stimulus difficulty was calibrated for each participant at the beginning of each session. Specifically, one circle contained 100 dots, and the other was adjusted with a 2-down 1-up rule (Levitt, 1971), to obtain 71% of “left” or “right” responses, in two interleaved staircases of 150 trials each. The step size of the staircases was reduced from 20 to 16, 8, 4 and 2 dots on trials 12, 24, 60 and 80 respectively.

Symbolic cueing session.

In this session, each trial started with a central cue presented for 250ms, before the fixation cross. The cue was either a triangle pointing to the left or the right side of the screen, indicating the correct response with 75% validity (cue condition), or a diamond providing no information (no-cue condition). These two conditions were administered in interleaved mini-blocks of 8 trials each, for a total of 256 trials per condition. Participants were fully informed of the meaning of these cues, and instructed to use both the stimulus and the cue to make the best possible decisions. Response accuracy was incentivized: participants won 1 point if correct and lost 1 point if incorrect, and points were converted to a bonus payment at the end of the experiment, with 1 point= 0.02 Euros. A training phase with feedback (96 trials) was included.

Confidence session.

In the confidence session, each response was followed by a confidence rating, in which participants indicated their subjective belief that their response just given was correct, on a 6 steps scale ranging from 50% confident (i.e. guess) to 100% confident, in 6 steps of 10%. Participants responded using the numerical keys on the top-left of the keyboard. This confidence rating was incentivized using a probability matching rule [20], which is a variant of the Becker-DeGroot-Marschak rule classically used in experimental economics [51]. The participant is offered an exchange between his response and a lottery ticket with a probability P of success. The number P is randomly determined on each trial (with a uniform distribution between 0 and 1), and compared to the confidence response. If P is greater than the confidence, then the participant's reward is determined by the lottery. If not, it is determined by the accuracy of the response. The mechanism was presented to participants as a way to maximize their earnings by providing accurate confidence ratings. Instructions, examples, and a training phase with feedback (40 trials) were included to make sure that participants understood the mechanism. Participants then completed 512 trials in the session.

Results

An empirical relation between overconfidence & suboptimal cue combination

We first established our basic measures of participants' overconfidence. From the "confidence session", we evaluated raw overconfidence as the average confidence minus the average accuracy for each observer (Fig 2A). We found that although there was a large heterogeneity across participants (M=0.08, SD=0.11), overconfidence was highly significant at the group level (t(68)=6.42, p=1.6e-08). With a more model-based approach (see methods), we also developed a novel measure of overconfidence, by considering the ratio of the subjective sensitivity (d'_{subj}) evaluated from confidence ratings over the actual sensitivity (d') evaluated from performance in the perceptual task. This model-based measure was largely correlated with our initial raw measure across participants (r=0.85, p=9.6e-21). Subjective estimations of sensitivity were twice as large as actual values of sensitivity (ratio d'_{subj}/d': M=2.18, SD=1.31), indicating significant overconfidence over the group (T-test vs. 1: t(68)=13.89, p=1.5e-21).

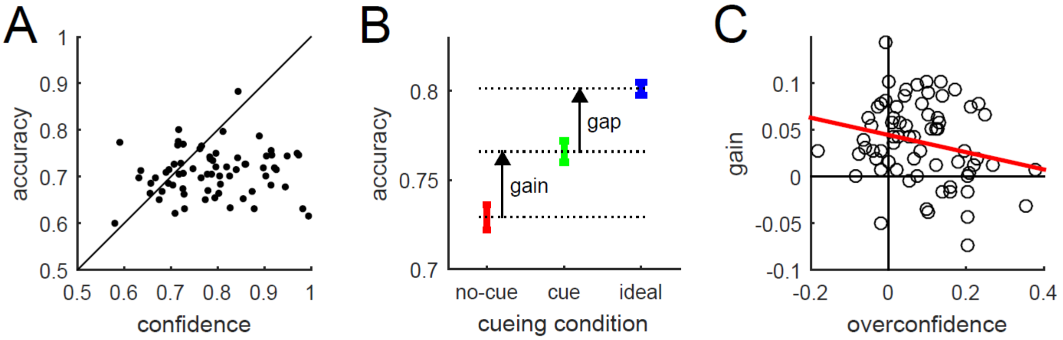


Fig 2.Overconfidence and cue combination performance. ***(A)*** *Average accuracy and confidence for each observer in the confidence session.* ***(B)*** *Average performance across observers in the cueing session, in the presence of a non-informative cue (no-cue condition), of an informative cue (cue condition), and for virtual observers perfectly integrating the cue (ideal condition). Error bars represent SEM.* ***(C)*** *The relation between raw overconfidence (i.e. average confidence minus average accuracy) and the gain in*

performance in the presence of the cue (relative to the no-cue condition). Each dot is an individual observer (N=69). The line represents the best-fitting regression.

We then turned to the “cueing session” to quantify how well observers can combine the symbolic cue information with their sensory information. We did so in two ways (Fig 2B). First, we found a significant gain in performance (t(68)=-7.12, p=8.6e-10) in the cue condition (M=0.77, SD=0.05) relative to the no-cue condition (M=0.73, SD=0.06), indicating that the cue information was used, at least to some extent. Nonetheless, this performance still fell well below the performance that would have been achieved, had participants optimally integrated the cue and maintained the same sensitivity in all cue conditions (M=0.80, SD=0.03, see methods for the computation of this ideal performance). Indeed, a significant gap was found between the actual and ideal performance (t(68)=-9.07, p=2.5e-13). These two measures of cue combination performance (the raw performance gain, and the model-based performance gap) were strongly negatively correlated (r=-0.67, p=2.0e-10), as expected from their mathematical definitions.

Crucially, we then confirmed the expected relation between our measures of overconfidence and our measures of cue combination performance (Fig 2C). Indeed, our raw measure of overconfidence was negatively correlated with the raw performance gain across participants (r=-0.232, p=0.055), and our model-based measure of overconfidence was positively correlated with the model-based performance gap (r=-0.251, p=0.038). Both correlations indicated that more overconfident participants benefited less from the cue information.

Discarding conservatism as a mediating variable

The crucial test of the model described above involves conservatism, that is, the extent to which participants might fail to fully adjust their decision criteria in the presence of the cue. Ideally, participants should have adjusted their decision criteria in log-odds to incorporate the log-odds of the cue (that is *LC=log(0.75/0.25)≈1.1*). Our SDT analyses (see methods) however indicated that participants set their criteria only half-way through this ideal value (criterion adjustment in log-odds: M=0.58, SD=0.48), resulting in a significant conservatism (ratio actual criteria over ideal criteria: M=0.53, SD=0.42, T-test vs. 1: t(68)=-8.99, p=3.5e-13). At the group level, this overall amount of conservatism appeared in line with the overall amount of overconfidence as estimated from our model-based measure: the criterion adjustment that could be predicted on the basis of our model-based measure of overconfidence indeed largely overlapped with the criterion adjustment observed in the empirical data (Fig 3A).

To our surprise, however, the relation between overconfidence and conservatism did not hold when examining their covariation across participants. Conservatism was uncorrelated with either raw overconfidence (r=-0.049, p=0.691) or model-based overconfidence (r=-0.092, p=0.450). Conservatism thus cannot be a mediator between overconfidence and suboptimal cue combination performance. A Sobel test for mediation analyses indeed provided no support for the mediation (with raw measures: p=0.35; with model-based measures: p=0.23). Furthermore, conservatism and overconfidence appeared to simultaneously predict suboptimal cue combination in an independent manner, in a multivariate regression analysis conducted on our raw measures (conservatism: b=0.024, p=0.044; overconfidence: b=-0.088, p=0.063) or on our model-based measures (conservatism: b=0.042, p<0.001; overconfidence: b=-0.005, p=0.045).

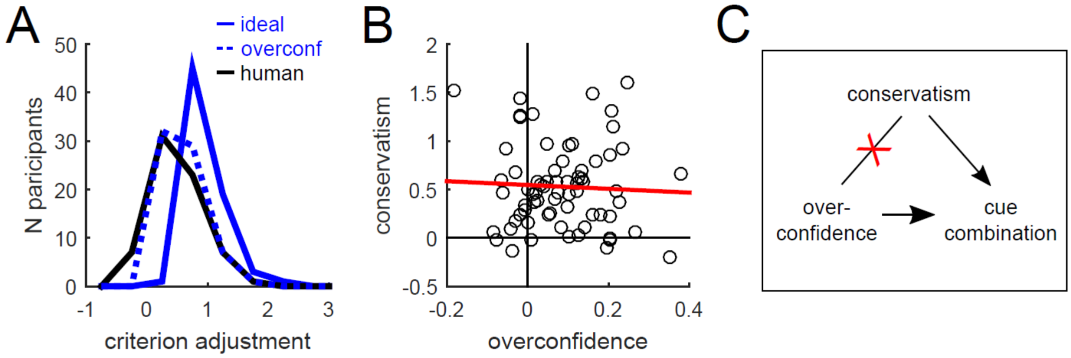


Fig 3. ***Overconfidence and conservatism. (A)*** *Distribution of the adjustment of decision criteria in the presence of a symbolic cue, as measured empirically and predicted theoretically for ideal observer and overconfident observers.* ***(B)*** *The relation between conservatism (criterion under-adjustment) and raw overconfidence. Each dot is an individual observer (N=69). The line represents the best-fitting regression.* ***(C)*** *Summary of the mediation analysis for conservatism.*

An exploratory analysis of motivation as a mediating variable

Since conservatism did not mediate the empirical relation between overconfidence and suboptimal cue combination performance, we carried out an exploratory analysis to investigate another possible mediator for this relation. In particular, we focused on the possibility that variations in sensitivity, rather than in the adjustment of decision criteria, might mediate the relation between overconfidence and suboptimal cue combination. We shall offer a possible intuition for such mediation in the discussion section.

To evaluate this mediation, we relaxed our assumption that sensitivity was constant in the “cueing session”, and we evaluated separate values of sensitivity (*d’*) and decision criterion (*c*) for each cue condition. The change in sensitivity between the cue condition (M=1.21, SD=0.39) and no-cue condition (M=1.26, SD=0.36) was not systematic at the group level (M=-0.05, SD=0.28, t(68)=-1.51, p=0.13, Fig 4A). Nonetheless the large heterogeneity of this sensitivity change across observers was a potential leverage to understand the relation between overconfidence and cue combination performance.

We thus applied the same mediation analysis as used in the previous section for the conservatism, but now assessing whether sensitivity loss (i.e. sensitivity in the no-cue condition minus sensitivity in the cue condition) could be the mediator between raw overconfidence (i.e. average confidence minus average accuracy) and our raw measure of cue combination performance (i.e. performance gain in the cue condition relative to the no-cue condition). We focused on raw measures here since the rational for the analysis was not based on the SDT model presented in the introduction. First, we evaluated the correlation between raw overconfidence and sensitivity loss and found a significant positive correlation between the two measures (r=-0.265, p=0.028). In other words, more overconfident participants tended to exhibit lower sensitivity in the cue condition relative to the no-cue condition, and less overconfident observers tended to exhibit higher sensitivity in the cue condition than in the no-cue condition (Fig 4B). Second, we entered sensitivity loss together with raw overconfidence to predict cue combination performance. In this regression analysis, we found that sensitivity loss had a negative and significant impact on performance gain (b=-0.099, p=4.8e-9). This was expected since a lower sensitivity in the presence of the cue, all other things being equal, would necessarily deteriorate performance in the cue condition, and thus lower performance gain. Critically, introducing sensitivity loss in the regression annihilated the effect of overconfidence on performance gain (from b=-0.093, p=0.055 to b=-0.024, p=0.52). Sobel’s test confirmed that sensitivity change was a mediator of the relation between overconfidence and performance gain (p=0.016).

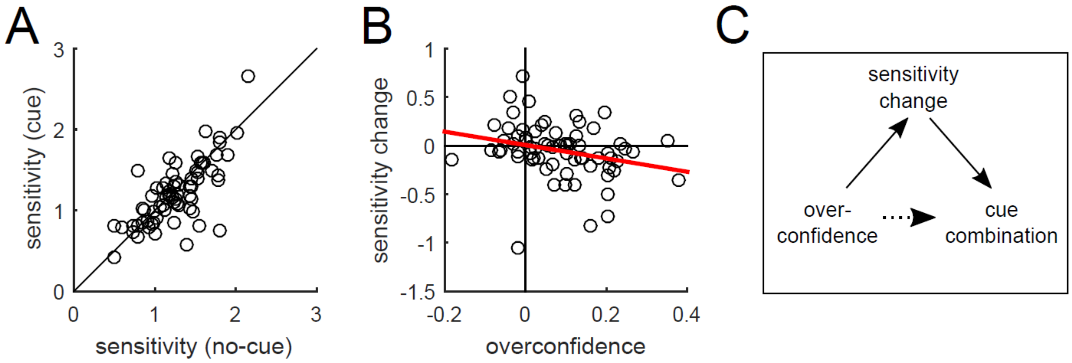


***Fig 4. Overconfidence and sensitivity change. (A)** Sensitivity in the presence of a valid symbolic cue and in the no-cue condition. **(B)** The relation between sensitivity change and raw overconfidence. The line represents the best-fitting regression. In both panels, each dot is an individual observer (N=69). **(C)** Summary of the mediation analysis for sensitivity change.*

Discussion

In our introduction, we described a theoretical model according to which overconfidence would induce conservatism and result in suboptimal combination of the information provided by sensory stimuli and symbolic cues. Within this model, we provided quantitative and direct operational measures of overconfidence (the overestimation of one's own choice accuracy), conservatism (the tendency to under-adjust decision criteria when receiving a symbolic cue), and suboptimal cue combination (the inability to improve choice accuracy when a cue is offered). We then tested the proposed model using a psychophysical task, which allowed precise control over the task parameters and unambiguous definitions of choice accuracy in each trial.

Contrary to what was prescribed by the model, overconfidence and conservatism appeared uncorrelated across participants. In other words, our initial hypothesis that overconfidence would scale with the overweighting of private signal relative to the symbolic cue was not supported by our data. It should be emphasized that this absence of correlation is not likely to be due to poor experimental measures of overconfidence and conservatism, as both were clearly manifest in our data, in line with previous studies reporting overconfidence [18,21] and conservatism [7,8,30,31] using perceptual tasks. In addition, we did find that both measures separately predicted cue combination performance, despite the fact that cue combination performance and overconfidence were measured in separate experimental sessions taken 4 days apart by participants.

That overconfidence and conservatism were uncorrelated came as an unexpected result, but a recent study [7] has reported the same pattern in a visual detection task, although with a different experimental strategy. The authors of this study concluded that the way participants adjusted their decision criteria was unrelated to the mis-estimation of the noise affecting the internal signal during the decision process, which one might relate to overconfidence. Our study provides further evidence along these lines, although with a different estimation of overconfidence.

Our data prompted us to consider that a different mechanism may underlie the empirical relation between overconfidence and cue combination performance. Indeed, our final analysis suggests that the change in perceptual sensitivity in the presence of a cue is a mediator for the relation between overconfidence and cue combination performance. Specifically, we found that overconfidence was positively correlated with this sensitivity loss, which in turn impacted negatively on performance. A potential explanation for this phenomenon might be that when a symbolic cue is offered overconfident observers reduce their efforts more than well-calibrated participants, which would eventually result in a sensitivity loss. Further investigation is needed to study the effect of overconfidence on the strategic allocations of resources.

Before we conclude, we would like to examine how the question of the link between overconfidence and the integration of cues towards task performance is also relevant in other domains, beyond psychophysical investigations. In the behavioral finance literature, for example, overconfidence is an important issue, and some models have been proposed on similar grounds as our initial SDT model. These models show how overconfidence should, theoretically, induce overreaction to private information and under-reaction to public signals, which results in excess volatility on markets [28], abnormally high trading volumes [29], and lower profits [29]. However, the empirical data on the issue is actually mixed, and whether the theoretical link holds in practice is still debated [24,32–36]. Our empirical results also call for a better understanding of the mechanisms underlying the relationship between overconfidence and loss in performance when private information is provided.

More generally, how humans use (or fail to use) external advices in their decisions is a central issue in the abundant literature on decision aid and decision support systems (for reviews see e.g. [37–39]). Critically, whereas such systems may outperform human judgments in some situations [40–44], it has also been recurrently shown that human decision-makers fail to use them, and hence make errors that could have been avoided. This failure has been documented in various decision-making domains, including financial decisions [45], sport evaluations [46], and medical judgments [47,48]. We note that previous research has also blamed overconfidence for these failures [45,49], as "One of the dangers of overconfidence is that one feels that no assistance is needed" [46]. Our explanatory analysis also suggests that future research could investigate whether, in addition, one of the dangers of overconfidence is that one feels that no effort is needed.

References

1. Landy MS, Kojima H. Ideal cue combination for localizing texture-defined edges. J Opt Soc Am A Opt Image Sci Vis. 2001;18: 2307–2320.

2. Hillis JM, Watt SJ, Landy MS, Banks MS. Slant from texture and disparity cues: optimal cue combination. J Vis. 2004;4: 967–992. doi:10.1167/4.12.1

3. Knill DC, Saunders JA. Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Res. 2003;43: 2539–2558.

4. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002;415: 429–33. doi:10.1038/415429a

5. Gepshtein S, Banks MS. Viewing geometry determines how vision and haptics combine in size perception. Curr Biol CB. 2003;13: 483–488.

6. Alais D, Burr D. The ventriloquist effect results from near-optimal bimodal integration. Curr Biol CB. 2004;14: 257–262. doi:10.1016/j.cub.2004.01.029

7. Ackermann JF, Landy MS. Suboptimal decision criteria are predicted by subjectively weighted probabilities and rewards. Atten Percept Psychophys. 2015;77: 638–658. doi:10.3758/s13414-014-0779-z

8. Green DM, Swets JA. Signal Detection Theory and Psychophysics. John Wiley and Sons; 1966.

9. Murrell GA. Combination of evidence in a probabilistic visual search and detection task. Organ Behav Hum Perform. 1977;18: 3–18. doi:10.1016/0030-5073(77)90015-0

10. Ulehla ZJ. Optimality of perceptual decision criteria. J Exp Psychol. 1966;71: 564–569. doi:10.1037/h0023007

11. Healy AF, Kubovy M. A comparison of recognition memory to numerical decision: How prior probabilities affect cutoff location. Mem Cognit. 1977;5: 3–9.

12. Healy AF, Kubovy M. The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. Mem Cognit. 1978;6: 544–553.

13. Botzer A, Meyer J, Bak P, Parmet Y. User settings of cue thresholds for binary categorization decisions. J Exp Psychol Appl. 2010;16: 1–15. doi:10.1037/a0018758

14. Botzer A, Meyer J, Parmet Y. Mental effort in binary categorization aided by binary cues. J Exp Psychol Appl. 2013;19: 39–54. doi:10.1037/a0031625

15. Chi C-F, Drury CG. Do people choose an optimal response criterion in an inspection task? IIE Trans. 1998;30: 257–266.

16. Wang L, Jamieson GA, Hollands JG. Trust and Reliance on an Automated Combat Identification System. Hum Factors. 2009;51: 281–291. doi:10.1177/0018720809338842

17. Hoffrage, U. (2016). 13 Overconfidence. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, Second Edition,291

18. Baranski JV, Petrusic WM. The calibration and resolution of confidence in perceptual judgments. Percept Psychophys. 1994;55: 412–428.

19. Kvidera S, Koutstaal W. Confidence and decision type under matched stimulus conditions: overconfidence in perceptual but not conceptual decisions. J Behav Decis Mak. 2008;21: 253–281. doi:10.1002/bdm.587

20. Massoni S, Gajdos T, Vergnaud J-C. Confidence measurement in the light of signal detection theory. Front Psychol. 2014;5. doi:10.3389/fpsyg.2014.01455

21. Mamassian P. Overconfidence in an Objective Anticipatory Motor Task. Psychol Sci. 2008;19: 601–606. doi:10.1111/j.1467-9280.2008.02129.x

22. West RF, Stanovich KE. The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. Psychon Bull Rev. 1997;4: 387–392.

23. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. Am J Med. 2008;121: S2-23. doi:10.1016/j.amjmed.2008.01.001

24. Biais B, Hilton D, Mazurier K, Pouget S. Judgemental Overconfidence, Self-Monitoring, and Trading Performance in an Experimental Financial Market. Rev Econ Stud. 2005;72: 287–312. doi:10.1111/j.1467-937X.2005.00333.x

25. Camerer C, Lovallo D. Overconfidence and excess entry: An experimental approach. Am Econ Rev. 1999; 306–318.

26. Johnson DD., McDermott R, Barrett ES, Cowden J, Wrangham R, McIntyre MH, et al. Overconfidence in wargames: experimental evidence on expectations, aggression, gender and testosterone. Proc R Soc B Biol Sci. 2006;273: 2513–2520. doi:10.1098/rspb.2006.3606

27. Kubovy M. A possible basis for conservatism in signal detection and probabilistic categorization tasks. Percept Psychophys. 1977;22: 277–281.

28. Daniel K, Hirshleifer D, Subrahmanyam A. Investor Psychology and Security Market under- and Overreactions. J Finance. 1998;53: 1839–1885.

29. Odean T. Volume, Volatility, Price, and Profit When All Traders Are Above Average. J Finance. 1998;53: 1887–1934. doi:10.1111/0022-1082.00078

30. Gorea A, Sagi D. Failure to handle more than one internal representation in visual detection tasks. Proc Natl Acad Sci U S A. 2000;97: 12380–12384. doi:10.1073/pnas.97.22.12380

31. Morales J, Solovey G, Maniscalco B, Rahnev D, de Lange FP, Lau H. Low attention impairs optimal incorporation of prior knowledge in perceptual decisions. Atten Percept Psychophys. 2015;77: 2021–2036. doi:10.3758/s13414-015-0897-2

32. Barber BM, Odean T. Boys will be Boys: Gender, Overconfidence, and Common Stock Investment. Q J Econ. 2001;116: 261–292. doi:10.1162/003355301556400

33. Deaves R, Lüders E, Luo GY. An Experimental Test of the Impact of Overconfidence and Gender on Trading Activity. Rev Finance. 2009;13: 555–575. doi:10.1093/rof/rfn023

34. Fellner-Röhling G, Krügel S. Judgmental overconfidence and trading activity. J Econ Behav Organ. 2014;107: 827–842. doi:10.1016/j.jebo.2014.04.016

35. Olsson H. Measuring overconfidence: Methodological problems and statistical artifacts. J Bus Res. 2014;67: 1766–1770. doi:10.1016/j.jbusres.2014.03.002

36. Glaser M, Weber M. Overconfidence and trading volume. Geneva Risk Insur Rev. 2007;32: 1–36. doi:10.1007/s10713-007-0003-3

37. Swets JA, Dawes RM, Monahan J. Psychological science can improve diagnostic decisions. Psychol Sci Public Interest. 2000;1: 1–26.

38. Berner ES, La Lande TJ. Overview of clinical decision support systems. Clinical decision support systems. Springer; 2007. pp. 3–22. Available: http://link.springer.com/chapter/10.1007/978-0-387-38319-4_1

39. Maserejian NN, Lutfey KE, McKinlay JB. Do Physicians Attend to Base Rates? Prevalence Data and Statistical Discrimination in the Diagnosis of Coronary Heart Disease. Health Serv Res. 2009;44: 1933–1949. doi:10.1111/j.1475-6773.2009.01022.x

40. Meehl PE. Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence. Northvale, N.J.: Echo Point Books & Media; 2013.

41. Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. Science. 1989;243: 1668–1674. doi:10.1126/science.2648573

42. Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: A meta-analysis. Psychol Assess. 2000;12: 19–30. doi:10.1037//1040-3590.12.1.19

43. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. Jama. 2005;293: 1223–1238.

44. Bishop MA, Trout JD. Epistemology and the Psychology of Human Judgment [Internet]. Oxford University Press; 2005. Available: http://www.oxfordscholarship.com/view/10.1093/0195162293.001.0001/acprof-9780195162295

45. Whitecotton SM. The effects of experience and confidence on decision aid reliance: A causal model. Behav Res Account. 1996;8: 194–216.

46. Arkes HR, Dawes RM, Christensen C. Factors influencing the use of a decision rule in a probabilistic task. Organ Behav Hum Decis Process. 1986;37: 93–110. doi:10.1016/0749-5978(86)90046-4

47. Short D, Frischer M, Bashford J. Barriers to the adoption of computerised decision support systems in general practice consultations: a qualitative study of GPs' perspectives. Int J Med Inf. 2004;73: 357–362. doi:10.1016/j.ijmedinf.2004.02.001

48. Böckenholt U, Weber EU. Use of formal Methods in Medical Decision Making A Survey and Analysis. Med Decis Making. 1992;12: 298–306.

49. Sieck WR, Arkes HR. The recalcitrance of overconfidence and its contribution to decision aid neglect. J Behav Decis Mak. 2005;18: 29–53. doi:10.1002/bdm.486

50. Brainard DH. The Psychophysics Toolbox. Spat Vis. 1997;10: 433–436.

51. Becker GM, DeGroot MH, Marschak J. Measuring utility by a single-response sequential method. Behav Sci. 1964;9: 226–232.

Supplementary materials

Model-based measure of overconfidence.

Following Signal Detection Theory, let's assume that a given state of nature (i.e. *A* = right has more dots vs. *B* = left has more dots) generates an internal response or private signal (noted *x*) that the agent compares with a decision criterion (noted *c*). We assume also that across trials, both states generate normally distributed values of *x*, with equal variance σ^2 and with means μ_A and μ_B. Without loss of generality and for simplicity, we further assume that μ_A and μ_B are symmetric around 0, and that $\sigma=1$. The common value of the generative means is then equal to *d'/2*, where *d'* is the standard SDT measure of the sensitivity of the participant.

$$P(x|A) = N(x, +d'/2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-d'/2)^2} \qquad \text{(eq. 2a)}$$

$$P(x|B) = N(x, -d'/2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+d'/2)^2} \qquad \text{(eq. 2b)}$$

We can now calculate the logarithm of the likelihood ratio that the ideal agent would consider, which we noted *LS* in our introduction (eq. 1b). This log-likelihood ratio takes a simple form and scales with the actual sensory input *x* (eq. 3). The scaling factor is the sensitivity value, *d'*, for an ideal agent who correctly estimates the generative model.

$$LS(x) = \log\left(\frac{P(x|A)}{P(x|B)}\right) = -\frac{1}{2}\left(x - \frac{d'}{2}\right)^2 + \frac{1}{2}\left(x + \frac{d'}{2}\right)^2 = d'x \qquad \text{(eq. 3)}$$

A non-ideal agent might underestimate or overestimate her own abilities, which would correspond to an underestimation or overestimation of *d'*, and use a subjective value d'_{subj} instead of *d'* to amplify the sensory signal. In particular, an overconfident agent would use a value *d'ₛᵤbⱼ* greater than *d'* such that sensory signals would be systematically have more influence on choices than what they should.

Empirically, we used the confidence ratings to evaluate the overconfidence of participants. To explain how we did this, let us consider the information conveyed by *x*, separately for the ideal agent who uses the correct value of *d'* (eq. 4a), and for the non-ideal agent who uses the subjective estimate *d'ₛᵤbⱼ* (eq. 4b). In the confidence session, there are no cues and we consider that the prior is balanced, such that the posterior probability ratio amounts to the likelihood ratio.

$$\log\left(\frac{P_{obj}(A|x)}{P_{obj}(B|x)}\right) = LS(x) = \log\left(\frac{P_{obj}(x|A)}{P_{obj}(x|B)}\right) = d'x \qquad \text{(eq. 4a)}$$

$$\log\left(\frac{P_{subj}(A|x)}{P_{subj}(B|x)}\right) = LS_{subj}(x) = \log\left(\frac{P_{subj}(x|A)}{P_{subj}(x|B)}\right) = d'_{subj}x \qquad \text{(eq. 4b)}$$

From these two equations, we can evaluate the ratio of *d'ₛᵤbⱼ* over *d'* that is the relative bias in confidence for the agent. If this ratio is greater than 1, then the agent is overconfident. For a given internal signal *x*, this ratio relates the subjective probabilities expressed about the states of nature to the objective probabilities of these states (eq. 4c).

$$\log\left(\frac{P_{obj}(A|x)}{P_{obj}(B|x)}\right) = \frac{d'}{d'_{subj}} \log\left(\frac{P_{subj}(A|x)}{P_{subj}(B|x)}\right) \qquad \text{(eq. 4c)}$$

As we cannot observe the sensory signal *x* directly, we used the conjunction of the response and the confidence rating in each trial as a proxy for this variable. We thus formed subsets of trials according to the conjunction of response and confidence ratings, and we evaluated on these subsets the subjective and objective probabilities of a given state. We then conducted a linear regression to predict the objective probabilities from the subjective probabilities, across these subsets of trials. The inverse of the slope in this regression corresponds to our estimate of the ratio d'_{subj}/d'.

Model-based measure of conservatism and suboptimal cue combination.

Within SDT, and assuming a balanced prior, the ideal decision criterion corresponds to the value *c* of the sensory evidence for which the decision variable switches sign (eq. 5). It can be defined as a function of the value of *d'* and the cue information *LC* (eq. 5a and Fig 3, blue line). For the overconfident agent, the predicted criterion (Fig 3, blue dotted line) would take into account the subjective value d'_{subj} instead of the correct value (eq. 5b). In our analyses, we assumed that overconfidence was identical between the "confidence session" and the "cueing session", and we extrapolated the subjective sensitivity of participants in the cueing session from their objective sensitivity and their overconfidence.

$$DV = 0 \Leftrightarrow LS(x) + LC = 0 \Leftrightarrow d'x + LC = 0$$

$$c_{ideal} = -\frac{1}{d'}\,LC \qquad \text{(eq. 5a)}$$

$$c_{subj} = -\frac{1}{d'_{subj}}\,LC \qquad \text{(eq. 5b)}$$

$$\frac{c_{subj}}{c_{ideal}} = \frac{d'}{d'_{subj}} \qquad \text{(eq. 5c)}$$

Our measure of conservatism is the ratio of the actual criterion adjustment over the ideal criterion adjustment. According to our model, this quantity should be inversely related to our overconfidence estimate (see eq. 4c and eq. 5c). To evaluate the actual criterion adjustment, we fitted with maximum likelihood a SDT model to the "cueing session" data, separately for each participant. This fitted model had 4 parameters: a constant sensitivity and a decision criterion that was free to vary between the 3 cueing conditions (left, right, neutral). We used the semi-distance between the estimated criteria for the left cue and right cue trials as a measure of the actual criterion adjustment, for each participant (Fig 3, black line). The ideal criterion adjustment was derived (using eq. 5a) from the correct value of *LC=log(3)* and the estimated sensitivity of the participant.

We also derived a model-based measure of suboptimal cue combination, by considering the gap between the actual accuracy of the participant and the expected accuracy of this participant could have achieved (*EA*, eq. 6, where Φ is the cumulative standard normal distribution) if the sensory information and the cue information had been combined optimally. This expected accuracy corresponds to the accuracy of an SDT agent with sensitivity equal to the *d'* of the observer and with c_{ideal} as a decision criterion.

$$EA = P(B)\Phi(c_{ideal} + d'/2) + P(A)\big(1 - \Phi(c_{ideal} - d'/2)\big) \qquad \text{(eq. 6)}$$

Chapter 6: Optimal probabilistic cue integration with visual evidence is limited by working memory capacity[11]

Marine Hainguerlot[1], Jean-Christophe Vergnaud[1,*], Vincent de Gardelle[2,*]

[1]Centre d'Economie de la Sorbonne, CNRS UMR 8174, Paris, France

[2]CNRS and Paris School of Economics, Paris, France.

[*]denotes equal contribution

[11] Please note that the experimental data used in this chapter come from the same experiment reported in Chapter 5 (for more details on the experimental design see the General introduction)

Abstract

In detection problems, it is typically observed that humans fail to combine sensory information with probabilistic cues in an optimal fashion. When the probability of the target occurring is unequal, observers should bias their response accordingly to maximize their accuracy. It is typically found, though, that observers do not bias their response as much as they should, a phenomenon called “sluggish beta” or “conservatism”. Previous work has suggested that appropriately biasing the response is cognitively demanding. We investigated whether working memory was a source of conservatism in visual detection. Participants performed a simple perceptual categorization task in which symbolic cues indicating with 75% validity the nature of the forthcoming stimulus. We measured participants’ working memory capacity in a running memory span task. Critically, our data revealed that working memory predicted the extent to which participants biased their responses. Furthermore, a mediation analysis showed that working memory predicted gain in accuracy via the adaptation of response bias. Our results suggest that working memory capacity contributes to individual differences in the ability to integrate visual evidence with probabilistic cues.

Introduction

In breast cancer screening, the radiologist should be more willing to categorize an abnormal spot on the mammogram as a malignant tumor if the patient has mother, sister, or daughter with breast cancer, as this factor is known to double the risk for the patient (see e.g. McPherson, Steel, Dixon, 2000). However, research also shows that humans typically fail to incorporate such probabilistic cues optimally. Even when the informative values of the cues are fully explained to them, observers insufficiently adjust their decisions, which may result in diagnostic errors. This suboptimal behavior, sometimes referred to as “conservatism” or “sluggish beta”, has been demonstrated both in and outside the laboratory, with simple auditory (Green & Swets, 1966; Tanner, 1956) and visual detection tasks (Ackerman & Landy, 2015; Ulehla, 1966), memory tasks (Healy & Kubovy, 1978), and clinical (Lusted, 1976), air traffic (Bissenet, 1981) or quality control (Harris & Chaney, 1969) decision tasks.

Various explanations have been proposed for conservatism, including probability matching (Healy & Kubovy, 1981; Lee & Janke, 1965; Thomas & Leffe, 1970), probability distortion (Ackermann & Landy, 2015; Kubovy, 1977) or incorrect modeling assumptions made by researchers (Maloney & Thomas, 1991). Here, we investigate another potential source of conservatism that has received only little attention so far, namely working memory. Indeed, a recent study on memory (Konkel et al., 2015) suggested that because combining the stimulus information and the cue information is cognitively demanding, it may also reflect the working memory abilities of participants. In other words, decision-makers with better working memory should be better able to adjust their behavior to the cue offered to them, and they should exhibit less conservatism.

We tested this hypothesis in the domain of perceptual decision making. After performing a standard running memory span task (Pollack et al., 1959; Broadway et al., 2010), participants (N=69) were engaged in a visual task in which they had to combine a simple stimulus with symbolic probabilistic cue in order to make a decision (see Figure 1). We anticipated that participants would exhibit conservatism in the perceptual decision task, in the sense that they would fail to fully use the information provided by the cue, as typically reported in the literature. Furthermore, and most critically, we predicted a negative relationship between this conservatism and working memory scores: conservatism would be more pronounced for agents exhibiting lower working memory scores.

Methods

Participants

69 individuals (39 females; mean age = 23 years, SD = 2.5 years) were recruited through the Laboratory of Experimental Economics of Paris and gave informed consent to participate. The study was conducted in line with the principles of the Declaration of Helsinki. Participants first completed a running memory span task and then performed the cue combination task. They received 5 Euros for participating plus an incentivized bonus described below.

Running memory span task

On each trial, participants encoded a sequence of *m+n* letters and were required to report in the forward order the last *n* letters. Letters were sampled with replacement from a pool (F, H, J, K, L, N, P, Q, R, S, T, Y). The task consisted of twelve trials, covering all possible combinations with *m*=0, 1, 2 and *n*=3, 4, 5, 6, in a random order. Each trial began with a fixation cross for 300ms. Letters were then presented successively for 300ms on the center of a gray background screen followed by a 2200ms interval. At the end of the sequence, participants were shown a 4x3 grid displaying the 12 letters of the pool and they had to report the last *n* letters by clicking on the corresponding cells in the grid, in the correct order. One point was earned for each item reported in the correct serial position. For example, if participants were instructed to report ”N P T”, they would gain 3 points for responding ”N P T” but 0 point for responding ”L N P”. The maximum score possible was 54 points; points were converted to payment with 1 point= 0.05 Euros. The task was programmed in JavaScript and administered to participants through the internet based software REGATE version 9.33 (Zeiliger).

Cue combination task

Each trial involved a central cue presented for 250ms, a 250ms fixation cross, and then the stimulus presented for 700ms. The stimulus was composed of two sets of about 100 dots, one on the left side and one on the right side of the computer screen, and participants’ task was to indicate which set contained more dots, by pressing the corresponding arrow on the keyboard. The cue was either a triangle pointing to the left or the right side of the screen, indicating the correct response with 75% validity (predictive cue condition), or a diamond providing no information (neutral cue condition). Participants were fully informed of the meaning of these cues, and they were explicitly instructed to use both the stimulus and the cue to make the best possible decisions. To ensure that they would, response accuracy was incentivized: participants won 1 point if correct and lost 1 point if incorrect, and points were converted to a bonus payment at the end of the experiment, with 1 point= 0.02 Euros. The predictive and neutral cue conditions were administered in interleaved mini-blocks of 8 trials each, for a total of 256 trials per condition. Participants received no feedback about the accuracy of their decision. Response times shorter than 200ms or longer than 2200ms (from stimulus onset) were discouraged by presenting a "too fast" or "too slow" message on the screen. The inter-trial interval was jittered between 0.5s and 1.5s.

After the instructions, participants received 300 trials of a calibration phase without the probabilistic cues, in which we adjusted the stimulus difficulty for each participant. Specifically, one circle contained 100 dots, and the other was adjusted with a 2-down 1-up rule (Levitt, 1971), to obtain 71% of “left” or “right” responses, in two interleaved staircases of 150 trials each. The step size of the staircases was reduced from 20 to 16, 8, 4 and 2 dots on trials 12, 24, 60 and 80 respectively. After this calibration phase, but before the main task, participants received 96 training trials in the same conditions as in the main task, except for the fact that they received feedback on their response accuracy during this training. The whole task was run using MATLAB (MathWorks) and Psychotoolbox [52], on screens (resolution 1024 x 768) viewed at normal distance (about 60 cm).

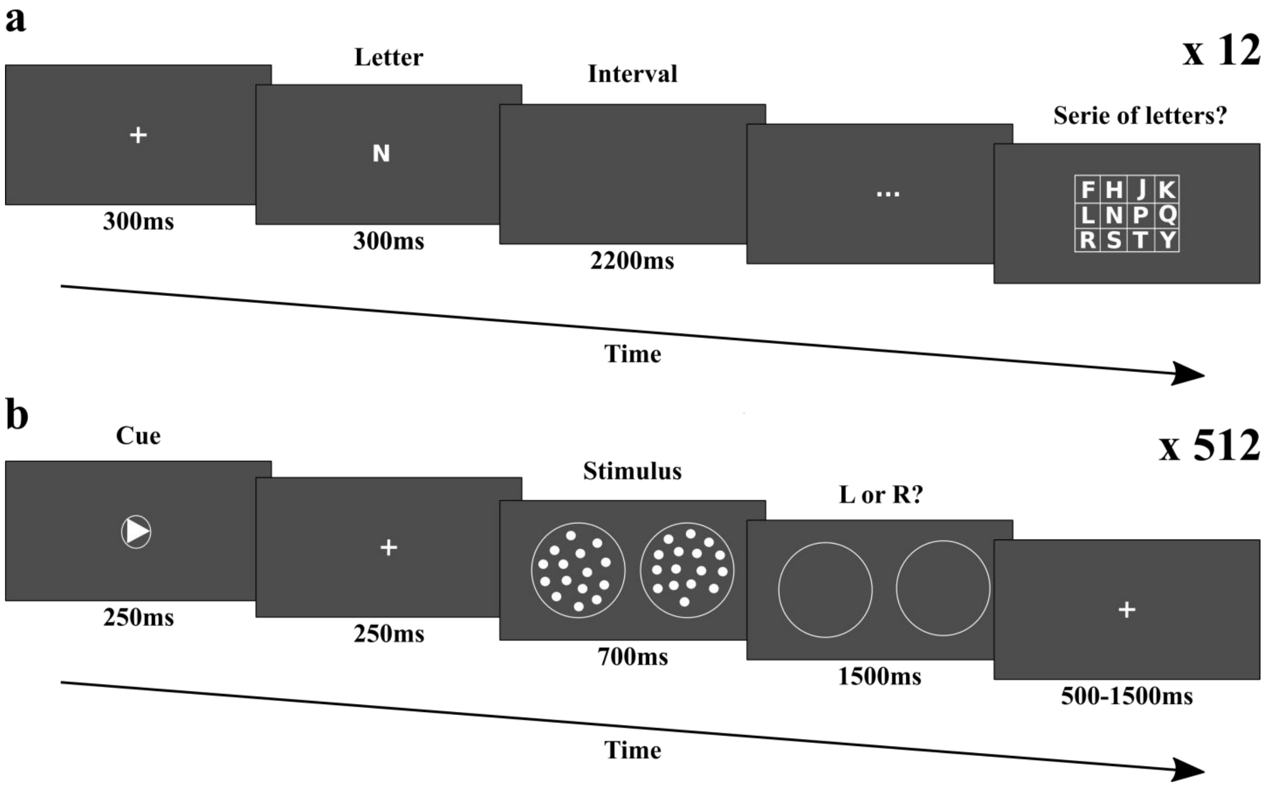


Figure 1: *Experimental Design.* ***(a)*** *Running memory span task. On each trial, a sequence of letters was displayed. Participants had to report in the forward order the last 3, 4, 5 or 6 letters.* ***(b)*** *Cue combination task. On each trial, participants had to decide which circle (left or right) contained more dots. Before the stimulus, a cue in the form of a geometric shape was presented to indicate the nature of the forthcoming stimulus. A triangle pointing to the left or to the right indicated with 75 % validity the left or the right circle respectively. A diamond indicated that the forthcoming stimulus was equally likely to be the left or the right circle.*

Results

As expected, participants' responses were indeed influenced by the cues, as shown by an increase in response accuracy for valid compared to neutral cues (84% vs. 73%, t(68)= 11.139, p<0.001) and a decrease in accuracy for invalid compared to neutral cues (54% vs. 73%, t(68)= -9.569, p<0.001). We then took the difference between accuracy in valid and invalid trials as a crude measure of the influence from the cues on perceptual decision making, and we assessed whether this influence from the cues would correlate with working memory scores, across participants. Here, the working memory score is the total number of correctly reported items in the memory task (M= 39.362, SD= 8.331). Critically, we found a significant correlation (r= 0.375, p= 0.002), which provided a first confirmation of our hypothesis: participants with higher working memory scores were also more affected by the cues in the perceptual decision-making task (**Fig2A**). It is worth noting that working memory scores were not correlated with accuracy in the neutral cue condition (r=-0.049, p=0.692), ruling out a possible interpretation in terms of a general willingness of participants to engage in the task.

We next applied Signal Detection Theory (Green & Swets, 1966), to distinguish between two parameters determining the observer's decisions: the decision criterion, that is the propensity to respond right or left, and the sensitivity, that is the ability to discriminate between the right and the left circle. Conservatism is typically associated with the insufficient adjustment of the decision criterion within this framework. For each participant, we estimated separate values of sensitivity and criterion for each cue (left, right, neutral), and we used the semi-distance between the estimated criteria (in log-odds units) for the left cue and right cue trials as a measure of how observers adjusted their decision criterion given the cue.

Participants did adjust their criteria (criterion adjustment: M=0.53, SD=0.465) but only halfway through the optimal adjustment (i.e. log(0.75/0.25)≈1.1). The ratio of actual criterion adjustment over ideal adjustment was significantly below one (M= 0.482, SD= 0.423; t(68)= -10.16, p<0.001) indicating conservatism across the cohort. Critically, confirming our hypothesis, we found that working memory was significantly correlated with criterion adjustment (r= 0.296, p= 0.013) (**Fig2B**). Of note, working memory was not correlated (r= -0.167, p= 0.171) with the change in sensitivity between the predictive and neutral cue conditions, which we defined as the ratio of the average sensitivity in the predictive cues over the sensitivity in the neutral cue (M= 0.982, SD= 0.232). It was also not correlated with the average sensitivity of the observer (r=-0.075, p=0.539).

Finally, we conducted a mediation analysis to evaluate the determinants of accuracy gains from the cues (i.e. the difference in accuracy between predictive and neutral cue trials). We found that accuracy gains could be predicted from working memory scores (b=0.001, p=0.039) and sensitivity change (b=0.123, p<0.001), but that entering criterion adjustment as an additional predictor of accuracy gains (b=0.022, p=0.01) suppressed the effect of working memory on accuracy gains (b=0.0004, p=0.392) and not that of sensitivity change (b=0.1349, p=0.01). A Sobel test confirmed that criterion adjustment fully mediated the relation between working memory scores and accuracy gains (t= 2.1626, p=0.0153, one-tailed).

Discussion

In this study, we showed that in a perceptual decision task, participants with higher working memory scores exhibited less conservatism when receiving explicit probabilistic cues that would help them in their decisions. These findings might have practical applications in applied decision-making situations, as conservatism-related errors are more likely to arise and should be monitored more closely if the situation imposes constraints on working memory.

Conservatism was previously associated with several explanations: a probability matching strategy observers (Healy & Kubovy, 1981; Lee & Janke, 1965; Thomas & Leffe, 1970), distortions in the subjective representation of probabilities (Ackermann & Landy, 2015; Kubovy, 1977) or incorrect assumptions made by researchers about the shape of the distribution of sensory evidence in observers (Maloney & Thomas, 1991). We suggest that working memory is a factor distinct from the above, noting that these earlier accounts indeed made no mention of working memory. Nevertheless, we acknowledge that whether and how this influence of working memory relates to earlier accounts of conservatism remains an open empirical question that requires further research. In any case, we insist that these different accounts are not mutually exclusive, and could take place at the same time.

Our results also suggest that combining sensory information with explicit probabilistic information is a cognitively demanding operation. The nature of the limits imposed by working memory remains unclear however. Does working memory simply protect observers from forgetting the probabilistic value of the cue on some trials? Alternatively, is our working memory score a proxy for general computational resources needed to quantitatively evaluate and combine the probabilistic information from the cue and the stimulus? Another possibility is suggested by the recent study of Ackerman and Landy (2015), who concluded that “subjects move their criterion away from the neutral criterion by estimating how much they stand to gain by such a change based on the slope of subjective gain as a function of criterion”. Although this explanation does not invoke working memory as a limiting factor, it might share computational resources involved in estimating the expected subjective gain or in facing the burden of changing one’s own decision criterion from one trial to the next.

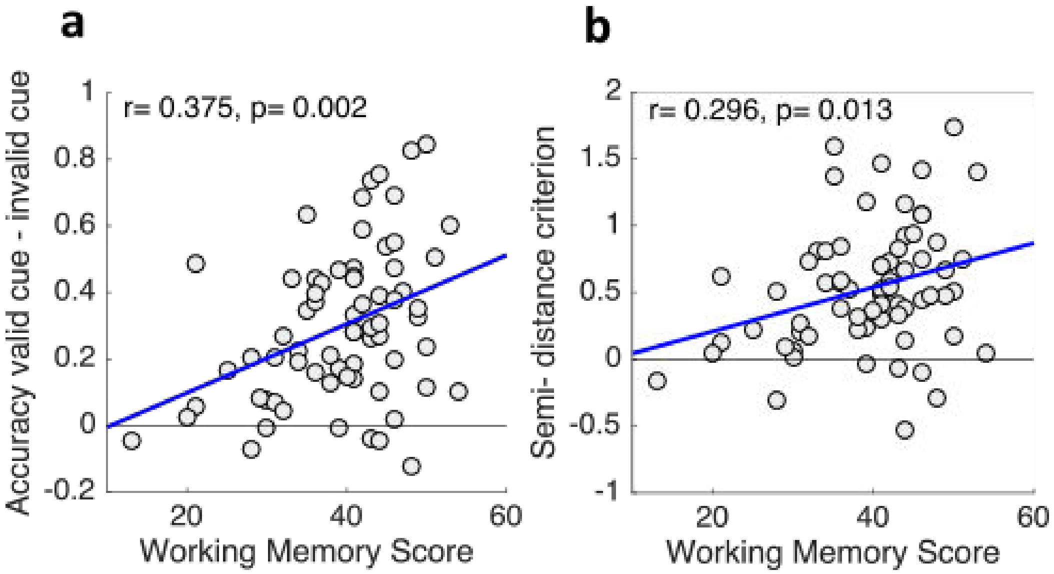


Figure 2: *The graphs show the relationship between working memory capacity and response bias measured as* ***(a)*** *the difference between the accuracy rate in valid and invalid cue trials;* ***(b)*** *the semi-distance criterion.*

References

Ackermann, J. F., & Landy, M. S. (2015). Suboptimal decision criteria are predicted by subjectively weighted probabilities and rewards. *Attention, Perception, & Psychophysics*, *77*(2), 638-658.

Bisseret, A. (1981). Application of signal detection theory to decision making in supervisory control The effect of the operator's experience. *Ergonomics*, *24*(2), 81-94.

Broadway, J. M., & Engle, R. W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods*, *42*(2), 563-570.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review*, *12*(5), 769-786.

Maloney, L.T., Thomas E.A.C (1991). Distributional assumptions and observed conservatism in the theory of signal detectability. *Journal of Mathematical Psychology*. 35(4):443–70.

Green, D.M, Swets J.A. (1966). Signal Detection Theory and Psychophysics. John Wiley and Sons.
Harris, D. H., & Chaney, F. D., (1969). Human factors in quality assurance. New York: Wiley.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of learning and motivation*, *22*, 193-225.

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & cognition*, *6*(5), 544-553.

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 344.

Konkel, A., Selmeczy, D., & Dobbins, I. G. (2015). They can take a hint: Older adults effectively integrate memory cues during recognition. *Psychology and aging*, *30*(4), 781.

Lee, W., & Janke, M. (1965). Categorizing externally distributed stimulus samples for unequal molar probabilities. Psychological Reports, 17(1), 79-90.

Lusted, L. B. (1976). Clinical decision making. In D. Dombal & J. Grevy (Eds.), Decision making and medical care. Amsterdam: North Holland.

Pollack, I., Johnson, L. B., & Knaff, P. R. (1959). Running memory span. *Journal of experimental Psychology*, *57*(3), 137-146.

Quak, M., London, R. E., & Talsma, D. (2015). A multisensory perspective of working memory. *Frontiers in human neuroscience*, *9*.

Thomas, E. A., & Legge, D. (1970). Probability matching as a basis for detection and recognition decisions. *Psychological Review*, *77*(1), 65.

Ulehla, Z. J. (1966). Optimality of perceptual decision criteria. *Journal of Experimental Psychology*, *71*(4), 564.

Supplementary materials

Supplementary results

	Response rate right	Accuracy, All trials	Accuracy, valid cue trials	Accuracy, invalid cue trials	SDT estimates	
					c	d’
Neutral cue	0.508 (0.066)	0.729 (0.058)	-	-	-0.029 (0.204)	1.264 (0.364)
Left predictive cue	0.253 (0.118)	0.764 (0.057)	0.840 (0.103)	0.535 (0.207)	0.497 (0.496)	1.213 (0.460)
Right predictive cue	0.743 (0.066)	0.768 (0.060)	0.841 (0.098)	0.549 (0.190)	-0.479 (0.440)	1.213 (0.426)

Table S1: *Response rate right, Average accuracy rate, accuracy rate conditional on valid and invalid cues, and SDT estimates decision criterion (c) and sensitivity (d’); with standard deviation reported between parentheses.*

	Mean (SD)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Working memory score	39.362 (8.331)	1	0.305*	-0.375**	0.375**	0.088	0.296*	-0.167	-0.264*	0.171	0.064
(2) Gain accuracy valid cue trials	0.112 (0.083)		1	-0.728***	0.875***	0.763***	0.766***	0.060	-0.171	0.159	0.150
(3) Gain accuracy invalid cue trials	-0.187 (0.162)			1	-0.969***	-0.112	-0.790***	0.574***	0.327**	-0.034	-0.186
(4) Accuracy valid – invalid cue	0.299 (0.230)				1	0.355**	0.834***	-0.383***	-0.293*	0.093	0.185
(5) Average gain in accuracy	0.037 (0.043)					1	0.366**	0.628***	0.061	0.214^{+}	0.041
(6)Semi- distance criterion	0.530 (0.465)						1	-0.181	-0.234^{+}	0.075	0.087
(7) Sensitivity change ratio	0.982 (0.232)							1	0.244*	0.170	-0.154
(8) Age	23.638 (3.125)								1	0.350**	-0.011
(9) Level of Studies[1]	3.420 (0.976)									1	-0.061
(10) Male Sex	0.435										1

Table S2: *Descriptive statistics and Pearson correlations for all the variables ([1] except Spearman correlations for the variable Level of Studies coded from 0 to 5 with: 0, vocational certificate; 1, high school diploma; 2, 2 year technical college degree; 3, undergraduate degree; 4, postgraduate degree; 5, PhD degree. N=69 participants. p- values: $^{+}$ <10%; *<5%; **<1%; ***<0,1%*

	(1) OLS Gain in accuracy		(2) OLS Semi-distance criterion		(3) OLS Gain in accuracy	
	b	t	b	t	b	t
D1 : Working Memory	0.001*	2.111	0.015*	2.317	0.0004	0.862
D2 : Sensitivity change	0.123***	7.028	-0.272	-1.149	0.135***	9.470
ME : Semi- distance criterion					0.044***	6.033

Table S3: *Regression results for mediation analysis with working memory and sensitivity change as independent variables, semi-distance criterion as mediator. D: Independent variables; ME: Mediator variable. [p- values: *p<0.05; **p<0.01;***p<0.001]*

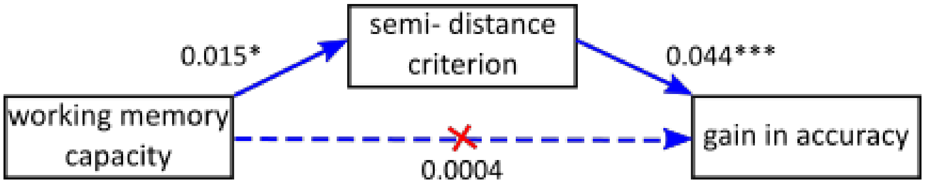


Figure S2: *Mediation analysis. Blue arrows represent the mediation. Semi- distance criterion fully mediates the impact of working memory capacity on the gain in accuracy, controlling for sensitivity change. [p- values: *p<0.05; **p<0.01;***p<0.001]*

General Conclusion

This thesis studies probability distortion in clinical judgment to compare physicians' judgment with statistical models. To address this "physician versus model" debate, we developed a theoretical framework of how physicians may process the information. It allowed us to derive the cognitive mechanisms through which physicians may depart from models, and, as a result, how their subjective probabilities may deviate from the objective probabilities. Within this framework, we considered that physicians form their clinical judgment by integrating an analytical component and an intuitive component. We documented that physicians may suffer from several biases in the way they process and integrate the information. They may misevaluate both the analytical and the intuitive components. They may inaccurately integrate the analytical and/ or intuitive component(s).

The dissertation gathers findings from the field and the lab that we believe can shed light on the debate. With actual medical data practice, we found that physicians' analytical component was biased. They were not as good as the statistical models at integrating consistently medical evidence. They underweighted medical evidence but also attributed values to non-relevant medical evidence, compared to our statistical model. As a result, physicians in our dataset over-estimated small probabilities that the patient had the disease and under- estimated large probabilities. Importantly, our analysis revealed that their biased probability judgment might cause unnecessary health care treatment. How then can we improve physician judgment? First, we explored whether replacing physician judgment by the probability generated from our statistical model, could improve actual medical decision accuracy. Our data analysis revealed that our statistical score, which combined the analytical model with the intuitive component of the physician, was not enough. It was necessary to include physicians' observed deviation from their expected decision in our statistical score to actually improve decision. Secondly, we tested in the lab factors that may affect human information processing. We found that participants' ability to learn about the value of the analytical component, in the absence of external feedback, depends on the quality of their intuitive component and their working memory capacity. In a second experiment, we found that participants' ability to integrate the analytical and intuitive components together depends on their working memory capacity. On the other hand, we found no support in favor of the hypothesis that a misevaluation of the intuitive component affects integration.

This thesis has several methodological limitations which generate further research questions. First, in chapters 1 and 3, the statistical model and the "analytical man" are developed on the very simple assumption that medical evidence is integrated linearly. There is no doubt that machine learning can do more than a simple linear integration, it can capture complex and nonlinear relationships in the data. Moreover, it is very simplistic to assume that physicians only integrate the evidence linearly (Hoffrage & Marewski, 2015). As a result, the intuitive component contained in physicians' judgment that we operationally measure as the part of the judgment that is not explained by a linear integration (i.e. "residual judgment") is very likely to also capture physicians' ability to interpret omitted evidence or to take evidence in a nonlinear way. Further work should compare the statistical model with the "analytical man" with more sophisticated assumptions on the integration of evidence. We may wonder whether we still find a comparative advantage of the physician over the model by specifying more complex integration of the evidence. In other words, what is made of “residual judgment”? To what extent is “residual judgment” made of intuition per se? A second limitation is that the questions we addressed in chapters 1 and 3 in the medical field are tested on a set of data that does not contain information per physician. In particular, we did not control for characteristics of the physician that may affect his clinical judgment and decision such as her experience, personality traits and sensitivity to risk. It would be important to address these empirical questions with a richer dataset.

Third, in chapter 3, we derived and tested our combined statistical scores on the same dataset. It would be necessary to cross validate and externally validate these scores. Finally, the generalizability of our findings in chapters 4, 5, and 6 from the lab may be questionable. To test the factors that may affect human information processing, we used a simple perceptual task. Moreover, most of the participants were students not physicians. The next step would be to investigate these factors with more ecological tasks and with physicians as participants.

Our findings lead to important implications. Overall, we found in chapter 1 that the biased analytical physician makes poorer decision. Thus, it would be necessary to consider solutions that may “debiase” the analytical physician. In the thesis, we started to explore two potential solutions. In chapter 3, we considered “replacing” the analytical physician by the statistical model. In chapters 4, 5, and 6, we considered “educating” and “supporting” the analytical physician by first identifying factors that may affect the way physicians process the information.

To “replace” the analytical physician, we would need to further investigate how and when physicians are willing to integrate combined statistical aids in their decision. Critically, medical decision ultimately lies in the hands of the physician who may be unwilling to use the combined statistical score from the model (Davis, 1993). We believe that it may be important to explore physicians’ acceptance of combined aids. Indeed, several studies (Sieck & Arkes, 2005; Whitecotton, 1996) have shown that experts tend to under use predictions from the machines. In particular, their (perceived) level of expertise has been documented as a potential source of neglect (Kaplan, 2001) as experts have doubts about the quality of the forecast. Thus, it would be interesting to investigate whether emphasizing that the statistical aid results from the combination of the physician himself and the model can enhance physician’s use of the aid. Importantly, it may also be crucial to facilitate the integration of statistical scores by providing visual aids (Garcia-Retamero & Hoffrage, 2013).

Furthermore, identifying the factors that affect human information processing may help to design cognitive interventions to “teach” or develop technologies to “support” physicians. In particular, teaching physicians about their biases and specific skills has been recommended as a potential debiasing intervention (Croskerry et al.; 2013a, Croskerry et al., 2013b; Croskerry, 2003). In our case, we could consider teaching physicians about their integration process of the analytical and intuitive component. We could also teach them to better evaluate the quality of their intuitive component. Another potential educational strategy may be to train physicians’ working memory, but the literature suggests that there may be only small room for improvement (Melby- Lervag & Hulme, 2014). Alternatively, technologies could be developed to enable physicians to encode less information in working memory.

What’s next? This dissertation opens several exciting avenues for research. First, it would be interesting to understand the gap between judgment and decision that we observed in chapter 3. The literature has already documented that non-clinical factors may alter physicians’ decision (Hajjaj et al, 2010; McKinlay et al, 1996). But, and to the best of our knowledge, it remains an open question to study the factors that may explain a valid deviation from expected decision. Several lines of research could be considered. In particular, physicians’ confidence may affect their decision (Lutfey et al, 2009). People are capable of robust evaluation of their decisions, thus they may correct and adapt their decision if they detect that they have made an error (Yeung & Summerfield, 2012). Also, to make a final decision, physicians may not only rely on their clinical judgment but share it with others (Charles et al, 1997). Shared decision making has proven that two heads can be better than one (Bahrami et al, 2010).

Another exciting topic would be to investigate what is made of intuition. In the thesis, we chose to adopt the empirical identification of intuition through the “residual judgment”. More precisely, intuition was defined as the part of the physician judgment not explained by the integration of the evidence. On the other hand, another line of research directly investigates intuition by asking physicians to report their feeling of intuition. We believe that one promising area of research would be to bridge the gap between these two approaches. In particular, to what extent is the “residual judgment” related to subjective reports of intuition?

Bibliography

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341-382.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081-1085.

Beckstead, J. W. (2017). The Bifocal Lens Model and Equation: Examining the Linkage between Clinical Judgments and Decisions. *Medical Decision Making*, *37*(1), 35-45.

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model+ 50% manager. *Management Science*, *36*(8), 887-899.

Brunswick, E., "The Conceptual Framework of Psychology," in International Encyclopedia of Unified Science, Vol. 1, No. 10, University of Chicago Press, Chicago, 1952.

Charles, C., Gafni, A., & Whelan, T. (1997). Shared decision-making in the medical encounter: what does it mean?(or it takes at least two to tango). *Social science & medicine*, *44*(5), 681-692.

Chen, J. H., & Asch, S. M. (2017). Machine learning and prediction in medicine− beyond the peak of inflated expectations. *The New England journal of medicine*, *376*(26), 2507-2509.

Croskerry et al. (2013a). Cognitive debiasing 1: origins of bias and theory of debiasing. *British Medical Journal Quality and Safety* 2013; 22:ii58–64.

Croskerry et al (2013b). Cognitive debiasing 2: impediments to and strategies for change. *British Medical Journal Quality and Safety,* 22:ii65–72.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine*, *78*(8), 775-780.

Davis (1993). User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. *International journal of man-machine studies, 38*(3), 475-487.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668-1674.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. Psy- chological Bulletin, 81, 95-106.

Donnelly, L. (2017). Forget your GP, robots will "soon be able to diagnose more accurately than almost any doctor". *The Telegraph*, March 7, available on: <http://www.telegraph.co.uk/technology/2017/03/07/robots-will-soon-able-diagnose-accurately-almost-doctor/>

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, *13*(2), 171-192.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*.

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine*, *83*, 27-33.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology*, *56*(11), 1129-1135.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological bulletin*, *73*(6), 422.

Green, D.M, Swets J.A. (1966). Signal Detection Theory and Psychophysics. John Wiley and Sons.

Harris, D. H., & Chaney, F. D., (1969). Human factors in quality assurance. New York: Wiley.

Greenhalgh, T. (2002). Intuition and evidence--uneasy bedfellows?. *Br J Gen Pract*, *52*(478), 395-400.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, *12*(1), 19.

Hajjaj, F. M., Salek, M. S., Basra, M. K., & Finlay, A. Y. (2010). Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice. *Journal of the Royal Society of Medicine*, *103*(5), 178-187.

Hoffrage, U., Marewski, J.N. (2015). Unveiling the lady in black: Modeling and aiding intuition *Journal of Applied Research in Memory and Cognition*, 4 (3) pp. 145-163.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American psychologist*, *64*(6), 515.

Kaplan et al (2001). The effects of predictive ability information, locus of control, and decision maker involvement on decision aid reliance. *Journal of Behavioral Decision Making,* 14, pp. 35-50.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of Linear Judgment: A Meta-Analysis of Lens Model Studies. *Psychological Bulletin*, *134*(3), 404-426.

Klein, G. A. (2004). *The power of intuition: How to use your gut feelings to make better decisions at work*. Crown Business.

Lutfey, K. E., Link, C. L., Marceau, L. D., Grant, R. W., Adams, A., Arber, S., ... & McKinlay, J. B. (2009). Diagnostic certainty as a source of medical practice variation in coronary heart disease: results from a cross-national experiment of clinical decision making. Medical Decision *Making*, *29*(5), 606-618.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, *21*(1), 422-430.

Massoni, S., Gajdos, T., & Vergnaud, J. C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in psychology*, *5*.

McClish, D. K., & Powell, S. H. (1989). How well can physicians estimate mortality in a medical intensive care unit?. *Medical Decision Making*, *9*(2), 125-132.

McKinlay, J. B., Potter, D. A., & Feldman, H. A. (1996). Non-medical influences on medical decision-making. *Social science & medicine*, *42*(5), 769-776.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental psychology*, *49*(2), 270-291.

Milcent, K., Faesch, S., Gras-Le Guen, C., Dubos, F., Poulalhon, C., Badier, I., ... & Nissack, G. (2016). Use of procalcitonin assays to predict serious bacterial infection in young febrile infants. *JAMA pediatrics*, *170*(1), 62-69.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, *375*(13), 1216.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... & Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.

Poses, R. M., Cebul, R. D., & Wigton, R. S. (1995). You can lead a horse to water-improving physicians' knowledge of probabilities may not affect their decisions. *Medical Decision Making*, *15*(1), 65-75.

Sieck & Arkes (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Makin*, *18*(1), 29-53.

Sorum, P. C., Stewart, T. R., Mullet, E., González-Vallejo, C., Shim, J., Chasseigne, G., ... & Grenier, B. (2002). Does choosing a treatment depend on making a diagnosis? US and French physicians’ decision making about acute otitis media. *Medical Decision Making*, *22*(5), 394-402.

Stolper, E., Van de Wiel, M., Van Royen, P., Van Bokhoven, M., Van der Weijden, T., & Dinant, G. J. (2011). Gut feelings as a third track in general practitioners' diagnostic reasoning. *Journal of general internal medicine*, *26*(2), 197-203.

Van den Bruel, A., Thompson, M., Buntinx, F., & Mant, D. (2012). Clinicians' gut feeling about serious infections in children: observational study. *Bmj*, *345*, e6144.

Whitecotton, S. M., Sanders, D. E., & Norris, K. B. (1998). Improving predictive accuracy with a combination of human intuition and mechanical decision aids. *Organizational behavior and human decision processes*, *76*(3), 325-348.

Wickens, C. D., Gordon, S. E., Liu, Y., & Lee, J. (2004). *An introduction to human factors engineering.*Pearson Prentice Hall.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology & human performance*. Psychology Press.

Whitecotton, S. M. (1996). The effects of experience and a decision aid on the slope, scatter, and bias of earnings forecasts. *Organizational Behavior and Human Decision Processes, 66*(1), 111-121.

Woolley, A., & Kostopoulou, O. (2013). Clinical intuition in family medicine: more than first impressions. *The annals of family medicine*, *11*(1), 60-66.

Yaniv, I., & Hogarth, R. M. (1993). Judgmental versus statistical prediction: Information asymmetry and combination rules. *Psychological Science*, *4*(1), 58-62.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B*, *367*(1594), 1310-1321

Résumé substantiel

Distorsion de probabilité dans le jugement clinique:

Etude de terrain et Expériences en laboratoire

Avec le développement du machine learning, la question de savoir si nous devrions remplacer le jugement humain par les prédictions des machines pour améliorer la prise de décision est d'actualité. Cette question n'a pas épargné la prise de décision médicale (Chen et Asch, 2017; Donnelly, 2017; Obermeyer et Emanuel, 2016). Le but de cette thèse est d'examiner si le remplacement du jugement des médecins par des modèles statistiques peut améliorer la qualité des décisions médicales.

Pour comparer le médecin avec le modèle, nous utiliserons la distorsion de probabilité comme mesure centrale. La distorsion de probabilité correspond à la différence entre les probabilités subjectives de l'homme et les probabilités objectives du modèle. Par ailleurs, cette thèse va au-delà de la question de la mesure de la performance du médecin par rapport au modèle. Cette thèse étudie également les processus cognitifs qui peuvent conduire les médecins à s'écarter du modèle. Comprendre les raisons de la distorsion de probabilité peut aider à mieux aborder le débat "homme versus modèle" dans la prise de décision médicale.

Un cadre théorique est proposé pour formaliser la façon dont les médecins traitent l'information. Ce cadre nous permet de dériver les mécanismes cognitifs par lesquels le médecin peut s'écarter du modèle et, par conséquent, comment ses probabilités subjectives peuvent s'écarter des probabilités objectives. Ce cadre servira de base théorique pour motiver les questions abordées dans la thèse. Pour chaque question adressée, un résumé des méthodes et des résultats est présenté. Tout d'abord, nous résumons les principales constatations de la littérature comparant le jugement des médecins aux modèles statistiques.

Littérature médecin versus modèle statistique

Jugement de prédiction et de probabilité

Les comparaisons entre les prédictions des modèles statistiques et celles des médecins remontent aux années 50 avec les travaux pionniers de Meehl (1954). Il est généralement reconnu que les algorithmes linéaires simples peuvent mieux diagnostiquer les conditions ou prédire les résultats que les médecins (Dawes et al, 1989; Grove et al, 2000; Ægisdóttir et al, 2006).

Cependant, les comparaisons entre les estimations de probabilité des médecins et celles générées par les modèles statistiques ont donné des résultats mitigés sur la supériorité des modèles (pour un résumé, voir O'Hagan et al, 2006). D'une part, on constate que les modèles statistiques sont mieux calibrés: les estimations de probabilité générées diffèrent moins des fréquences réelles. D'autre part, les estimations de probabilité des médecins semblent mieux faire la distinction entre les patients qui souffrent ou non de la condition. (p. ex.: McClish & Powell, 1989).

Les forces distinctes des modèles statistiques et des médecins sont au cœur du débat "homme versus modèle": quelle quantité d'information est disponible et comment l'information est-elle traitée? Les médecins peuvent avoir plus d'information, mais les modèles statistiques traitent mieux l'information.

La force des modèles statistiques: le jugement analytique

Les modèles statistiques produisent un jugement analytique fondé sur un ensemble d'hypothèses qui précisent comment intégrer les données. Les modèles sont reconnus pour leur capacité à intégrer les données de façon consistante (Karelaia et Hogarth, 2008). Présentés avec le même ensemble de données médicales, les modèles statistiques peuvent être mieux calibrés que les médecins parce qu'ils pondèrent de façon optimale les données, ils ne sont pas biaisés.

La force des médecins: le jugement intuitif

Le jugement du médecin implique non seulement des processus analytiques, mais aussi des processus intuitifs (Greenhalgh, 2002; Stolper et al, 2011; Wooley et Kostopoulou, 2013) et leur intuition peut être une information valide (Van den Bruel, 2012). Les modèles statistiques ne peuvent générer que des estimations de probabilité basées sur l’évidence médicale qu'on leur apprend à utiliser, alors que les médecins peuvent avoir accès à plus d'informations, en particulier leur jugement intuitif, pour mieux discriminer la présence ou l'absence de la maladie.

Combinaison: le modèle statistique + le médecin intuitif

En raison de leur complémentarité, il a été recommandé de tirer parti de leurs forces respectives en combinant le modèle statistique avec le jugement intuitif du médecin pour améliorer l'exactitude du jugement probabiliste (Blattberg et Hoch, 1990; Yaniv et Hogarth, 1993; Whitecotton et al, 1998).

Un cadre théorique du jugement du médecin

Pour aborder le débat "homme versus modèle" dans la thèse, nous proposons un cadre théorique qui décrit comment les médecins peuvent traiter l'information. Dans ce cadre, nous distinguons les processus analytiques et intuitifs qui peuvent être impliqués dans le jugement du médecin. Nous proposons une formalisation Bayésienne qui définit comment les médecins peuvent former leur jugement clinique en intégrant leurs processus analytiques et intuitifs. Ce cadre servira de base théorique pour motiver les questions abordées dans la thèse.

Un modèle de traitement de l'information du médecin

Pendant l'examen d'un patient, les médecins traitent de grandes quantités d'information provenant de leur environnement pour établir un diagnostic qui, éventuellement, guidera leur décision. Ici, nous adaptons un modèle de traitement de l'information de Wickens et al. (2015) à la situation de l'examen clinique. Nous décrivons ci-dessous un modèle simple en quatre étapes de traitement de l'information du médecin (voir aussi la **Figure 1**).

(5) *Données d'entrées:* Le médecin recueille de l'information sur le patient (indices de diagnostic) de deux façons. Certains indices (x_i) sont explicites en ce sens qu'ils prennent une valeur particulière. Par exemple, le médecin observe que la température du patient est de 37°C. Le médecin peut également recevoir un signal interne (x) qui correspond à une impression sur la présence ou l'absence de la maladie et qui ne peut pas être facilement reliées aux indices de diagnostic. Par exemple, le médecin peut avoir l'impression que quelque chose ne va pas chez le patient sans pouvoir expliquer pourquoi.

(6) *Traitement central:* Forte de ses connaissances et de ses expériences antérieures, elle récupère dans sa mémoire de long-terme une interprétation significative de l'association entre les informations recueillies dans les échantillons et l'apparition de la maladie. Par exemple, elle considère que l'observation d'une température de 37°C est généralement rassurante. D'autre part, elle se souvient que son sentiment que quelque chose ne va pas avec le patient a déjà été un signal d'alarme avec d'autres patients. En tout, elle conserve en mémoire de travail deux composantes informatives: une composante analytique (c. -à-d. l'évaluation des indices explicites) et une composante intuitive (c. -à-d. l'évaluation de son signal interne). Elle intègre ensuite les composantes analytique et intuitive pour juger si le patient est atteint ou non de la maladie.

(7) *Décision:* Le médecin décide de traiter ou non la patiente en fonction de son jugement quant à savoir si la patiente peut avoir la maladie et les conséquences associées à sa décision

(8) *Apprentissage:* Enfin, après avoir observé le résultat de sa décision, le médecin peut mettre à jour l'interprétation qu'elle attribue aux informations recueillies. Elle garde ensuite la valeur actualisée dans sa mémoire de long terme.

Notez que ces étapes dépendent de plusieurs capacités cognitives: la mémoire de travail, la mémoire de long terme, les ressources d'attention et effort. Ci-dessous, nous décrivons brièvement leurs rôles potentiels dans le modèle de traitement de l'information.

Mémoire de travail: La mémoire de travail est essentielle pour conserver et manipuler l'information à court-terme: "C'est la mémoire temporaire qui garde l'information active pendant que nous l'utilisons ou jusqu' à ce que nous l'utilisions" (Wickens et al, 2004). La mémoire de travail est nécessaire pour garder les composantes analytique et intuitif actifs afin de les intégrer ensemble. Elle est également nécessaire pour mettre à jour la valeur des indices de diagnostic en fonction de l'information obtenue à partir du résultat (c. -à-d. le feedback).

Mémoire de long-terme: La mémoire de long-terme est responsable du stockage et de la récupération d'informations sur la valeur des indices de diagnostic à long-terme. Il correspond au processus d'apprentissage.

Ressources d'attention et effort: En particulier, une attention sélective est nécessaire pour sélectionner les indices de diagnostic à traiter (ceux qui ont la plus grande valeur informative perçue) et les indices de diagnostic à ignorer. La capacité d'effectuer plusieurs tâches à la fois, d'affecter les ressources d'attention et d'effort à différentes tâches est également nécessaire.

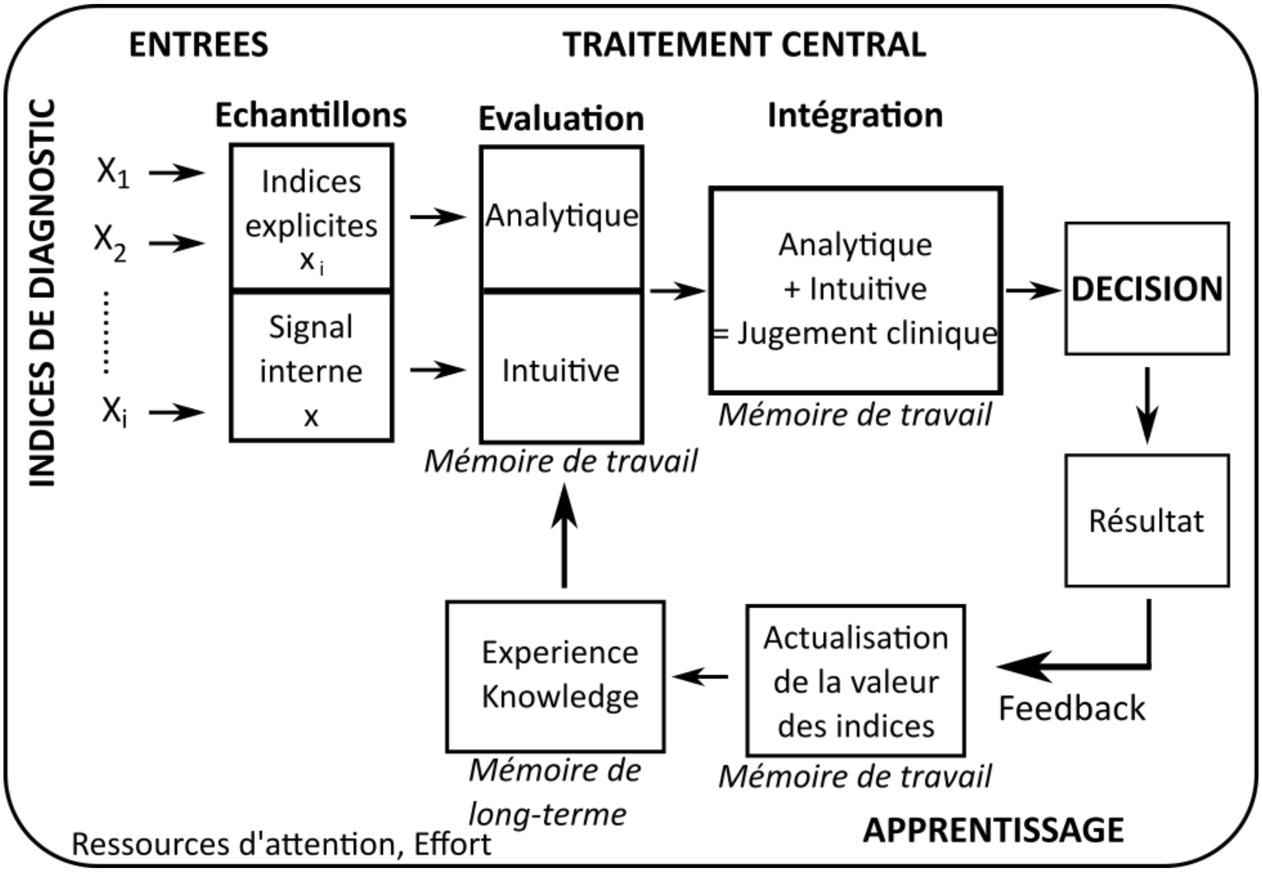


Figure 1: *Un modèle simple de traitement de l'information du médecin*

Modèle bayésien de traitement central de l'information

Nous modélisons la façon dont un médecin forme son jugement clinique sur la maladie du patient: $Y = 1$ ou $Y = 0$. Nous définissons d'abord l'évaluation des indices explicites et du signal interne. Deuxièmement, nous modélisons comment les composantes analytique et intuitive sont intégrées dans le jugement clinique. Le modèle est développé en log odds. Par souci de simplicité, nous supposons que les indices explicites x_i et le signal interne x sont indépendants.

Echantillons d'indices de diagnostic X_i

Le médecin observe les valeurs prises par les indices explicites (x_i) et le signal interne (x). Par souci de simplicité, nous considérons que les indices explicites peuvent prendre 2 valeurs: 1 ou 0.

Connaissance de x_i et x

Nous considérons que le médecin a en mémoire de long-terme des connaissances sur la sensibilité $P^s(x_i = 1|Y = 1)$[12] et la spécificité $P^s(x_i = 0|Y = 0)$ des repères explicites x_i. De même, le médecin stocke en mémoire de long-terme la distribution de probabilité du signal interne conditionnelle à la présence $P^s(x|Y = 1)$ ou absence de maladie $P^s(x|Y = 0)$.

Évaluation des indices explicites

Selon ses connaissances, elle évalue chaque indice x_i en calculant le poids d'évidence (log odds) comme suit: $W^s_{x_i=a} = ln\frac{P^s(x_i=a|Y=1)}{P^s(x_i=a|Y=0)}$ où $a = 0{,}1.$ La composante analytique correspond à la somme du poids d'évidence de x_i : $\sum_{i=1}^{k} W^s_{x_i}$.

Evaluation du signal interne

Nous postulons que le médecin évalue également son signal interne par le poids d'évidence comme suit:

$$W^s_{x=a} = ln\left(\frac{P^s(x=a|Y=1)}{P^s(x=a|Y=0)}\right), \text{où } a \in \mathbb{R}$$

W^s_x correspond à la composante intuitive.

[12] Où P^s correspond à la probabilité subjective. Nous notons P^o la probabilité objective.

Intégration des composantes analytique et intuitive

Enfin, elle forme sa probabilité subjective en révisant sa croyance antérieure sur la maladie $P^{s}_{prior}(Y=1)$ par l'équation d'intégration suivante:

$$ln\frac{P^{s}(Y=1|X,x)}{P^{s}(Y=0|X,x)} = ln\frac{P^{s}_{prior}(Y=1)}{P^{s}_{prior}(Y=0)} + \alpha\sum_{i=1}^{k} W^{s}_{x_i} + \beta W^{s}_{x} \qquad \text{(eq1)}$$

où les paramètres α et β capturent comment le médecin peut distordre les composantes analytique et intuitive respectivement.

Médecin versus jugement clinique idéal

Dans l'ensemble, le médecin peut souffrir de plusieurs biais dans la façon dont elle traite et intègre l'information. Premièrement, elle peut mal évaluer les composantes analytique et intuitive par rapport à leurs valeurs objectives $W^{o}_{x_i}$ et W^{o}_{x}. Deuxièmement, elle peut intégrer de façon inexacte les composantes analytique et/ou intuitive.

Terminologie

Dans le reste de la thèse, nous utilisons les termes suivants:

- « Modèle statistique »: jugement généré par un ensemble d'hypothèses précisant comment intégrer au mieux les données.
- "Homme analytique": partie du jugement du médecin expliquée par une intégration analytique de l’évidence.
- « Homme intuitif »: partie du jugement du médecin expliquée par l'intuition.

- Jugement du médecin: probabilité subjective du médecin que le patient soit atteint de la maladie.
- Décision du médecin: décision du médecin de traiter ou non le patient.

Questions abordées dans la thèse

Dans ce cadre, la thèse aborde les trois questions suivantes:

- Est-ce qu'un homme analytique biaisé prend de moins bonnes décisions?
- La combinaison du modèle statistique avec l'homme intuitif améliore-t-elle la qualité de la décision?
- Quels sont les facteurs qui influencent le traitement de l'information de l’homme?

Est-ce qu'un homme analytique biaisé prend de moins bonnes décisions?

Tel que décrit précédemment, le jugement du médecin comporte non seulement une composante analytique, mais aussi une composante intuitive. La mesure dans laquelle les médecins utilisent des processus analytiques et intuitifs peut varier. Ainsi, même s'il est bien documenté que les médecins sont biaisés dans leur intégration analytique des données, cela peut ne pas avoir d'incidence sur la qualité de leur décision. Par exemple, les médecins pourraient se fier à leur intuition pour prendre une décision. Par ailleurs, leur composante intuitive pourrait compenser leur composante analytique biaisée.

Dans le chapitre 1, notre but est d'évaluer, à l'aide de données médicales sur le terrain, si un médecin analytique biaisé prend de moins bonnes décisions. Notre ensemble de données médicales contient pour chaque patient: des informations sur la présence ou l'absence de la maladie, l' évidence médicale disponible, le jugement probabiliste du médecin que le patient ait la maladie et sa décision de traitement.

Selon notre cadre théorique, le jugement du médecin se compose de deux composantes: une partie analytique et une partie intuitive. Nous devons séparer ces deux composantes. Sur le plan opérationnel, nous proposons de séparer le jugement du médecin en deux composantes: le jugement linéaire et le jugement résiduel. Nous définissons la composante analytique comme un jugement linéaire qui contient la partie du jugement du médecin qui s'explique par une intégration linéaire de l'évidence médicale. La composante intuitive correspond au jugement résiduel qui saisit la partie du jugement qui n'est pas expliquée par une intégration linéaire. Il peut capter l'intuition des médecins, mais aussi leur capacité à intégrer les données de façon non linéaire. Pour évaluer si le médecin est biaisé dans son jugement linéaire, nous comparons la probabilité du médecin prédite par le jugement linéaire à la probabilité de maladie prédite par le modèle linéaire. Nous quantifions le biais dans la partie analytique comme la distorsion entre la probabilité du médecin prédite par le jugement linéaire et la probabilité de la maladie prédite par le modèle linéaire. Enfin, nous testons si la distorsion de probabilité nuit à la qualité des décisions médicales.

Méthode

La **Figure 2** illustre notre méthode.

Approche modèle de Lens

Quelle est la capacité du médecin à intégrer les données médicales disponibles par rapport à un modèle statistique? Pour répondre à cette question, nous utilisons l'approche du modèle de Lens (Brunswick, 1952; Goldberg, 1970). Nous considérons que le jugement du médecin et la présence ou l'absence de la maladie peuvent être modélisés comme deux fonctions linéaires distinctes des indices disponibles dans l'environnement (Dawes et Corrigan, 1974; Einhorn et Hogarth, 1975).

La présence ou l'absence de la maladie est modélisée comme une fonction linéaire d'un ensemble d'indices $X_i, i = 1,..,k$, comme suit:

$$ln\frac{P^o(Y=1|X)}{P^o(Y=0|X)} = \beta_o^o + \sum_{i=1}^{k} \beta_i^o\, x_i \qquad \text{(eq2a)}$$

De même, le jugement probabiliste du médecin selon lequel un patient est atteint de la maladie est modélisé comme une fonction linéaire d'un ensemble d'indices $X_i, i = 1,..,k$, comme suit:

$$ln\frac{P^s(Y=1)}{P^s(Y=0)} = \beta_o^s + \sum_{i=1}^{k} \beta_i^s\, x_i + \mu \qquad \text{(eq2b)}$$

où le terme d'erreur μ correspond au "jugement résiduel".

Interprétation de β_i^o versus β_i^s

Le coefficient β_i correspond au log odds ratio pour un indice x_i donné.

Les odds définissent la probabilité que la maladie se produise:

$$Odds = \frac{P(Y=1)}{P(Y=0)}$$

Le ratio des odds correspond aux odds que la maladie se produise étant donné que $x_i = 1$ par rapport aux odds que la maladie se produise lorsque $x_i = 0$. Le ratio des odds mesure l'association entre la présence ou l'absence de la maladie et l'indice x_i:

$$Odds\ Ratio = \frac{\frac{P(Y=1|x_i=1)}{P(Y=0|x_i=1)}}{\frac{P(Y=1|x_i=0)}{P(Y=0|x_i=0)}}$$

Nous pouvons observer que la façon dont le médecin pondère les indices en termes de log odds ratio (c-à-d β_i^s) peut différer de deux façons distinctes par rapport à β_i^o. Tout d'abord, elle peut sur- ou sous-pondérer les indices pertinents (*biais de sur- ou sous-pondération*): $\frac{\boldsymbol{\beta_i^s}}{\boldsymbol{\beta_i^o}} < 1$ ou $\frac{\boldsymbol{\beta_i^s}}{\boldsymbol{\beta_i^o}} > 1$. Deuxièmement, elle peut pondérer des indices non pertinents (*biais de pondération erroné*): $\beta_i^s \neq \beta_i^o = 0$.

Analyse de la distorsion de probabilité: Jugement linéaire versus modèle linéaire

En appliquant les poids de régression correspondants aux indices, on peut estimer en log odds ($Lo(p) = \frac{p}{1-p}$) la probabilité du médecin prédite par le jugement linéaire ($Lo(\widehat{P^s})$) et la probabilité de maladie prédite par le modèle linéaire ($Lo(\widehat{P^o})$). Pour comparer le jugement linéaire avec le modèle linéaire, nous traçons $Lo(\widehat{P^s})$ en fonction de $Lo(\widehat{P^o})$.

Nous utilisons le modèle linéaire général pour estimer un paramètre de pente de la distorsion lorsque les probabilités sont transformées en log odds, comme suit: $Lo\left(\widehat{P^s}\right) = \beta_0 + \beta_1 Lo(\widehat{P^o})^{13}$ (Zhang & Maloney, 2012).

Quand β_1=1, il n' y a pas de distorsion dans la pente. Quand $\beta_1 < 1$, les médecins surestiment les petites probabilités et sous-estiment les grandes probabilités. Quand $\beta_1 > 1$, les médecins sous-estiment les petites probabilités et surestiment les grandes probabilités.

Est-ce qu'un biais dans la distorsion de probabilité affecte la qualité des décisions médicales?

Empiriquement, nous séparons l'ensemble des données en deux groupes de patients selon le jugement probabiliste du médecin: un groupe haut biais et un groupe bas biais, de sorte que dans le groupe haut biais, la distorsion est plus grande comparativement au groupe bas biais. Nous examinons si la qualité de la décision médicale est altérée dans le groupe haut biais par rapport au groupe bas biais. Pour évaluer la qualité de la décision médicale, nous considérons deux mesures: la sensibilité (c'est-à-dire la proportion de patients atteints de la maladie qui reçoivent un traitement) et la spécificité (c'est-à-dire la proportion de patients non atteints de la maladie qui ne reçoivent pas de traitement).

Données

Nous avons appliqué notre méthode à la détection des infections bactériennes chez les nourrissons fébriles de moins de 3 mois (N=1848) et évalué l'impact du biais de la distorsion de probabilité sur deux dimensions des soins de santé (le traitement antibiotique et l'hospitalisation). Nos données proviennent d'une étude de cohorte prospective du

[13] Notez que les coefficients β_0 et β_1 dans cette équation sont différents des coefficients β_i^o et β_i^s venant de l'approche du Lens modèle.

biomarqueur de la procalcitonine dans la détection des infections bactériennes chez les nourrissons fébriles de moins de 3 mois (Milcent et al., 2016). Les médecins devaient consigner l'information recueillie sur les patients depuis leur admission jusqu' à leur sortie. Ils ont reporté les données démographiques et néonatales, les antécédents médicaux, l'examen physique et les résultats cliniques des analyses en laboratoire demandées. Les médecins devaient déclarer leur estimation de la probabilité que le nourrisson ait eu une infection bactérienne, sur une échelle de 0 à 100 % à deux étapes de la collecte des données: (i) à la fin de l'examen physique (probabilité pré-test); et (ii) après avoir reçu les résultats cliniques des tests en laboratoire (probabilité post-test). Après la deuxième estimation, ils signalaient leur décision d'hospitaliser et de traiter avec des antibiotiques.

-

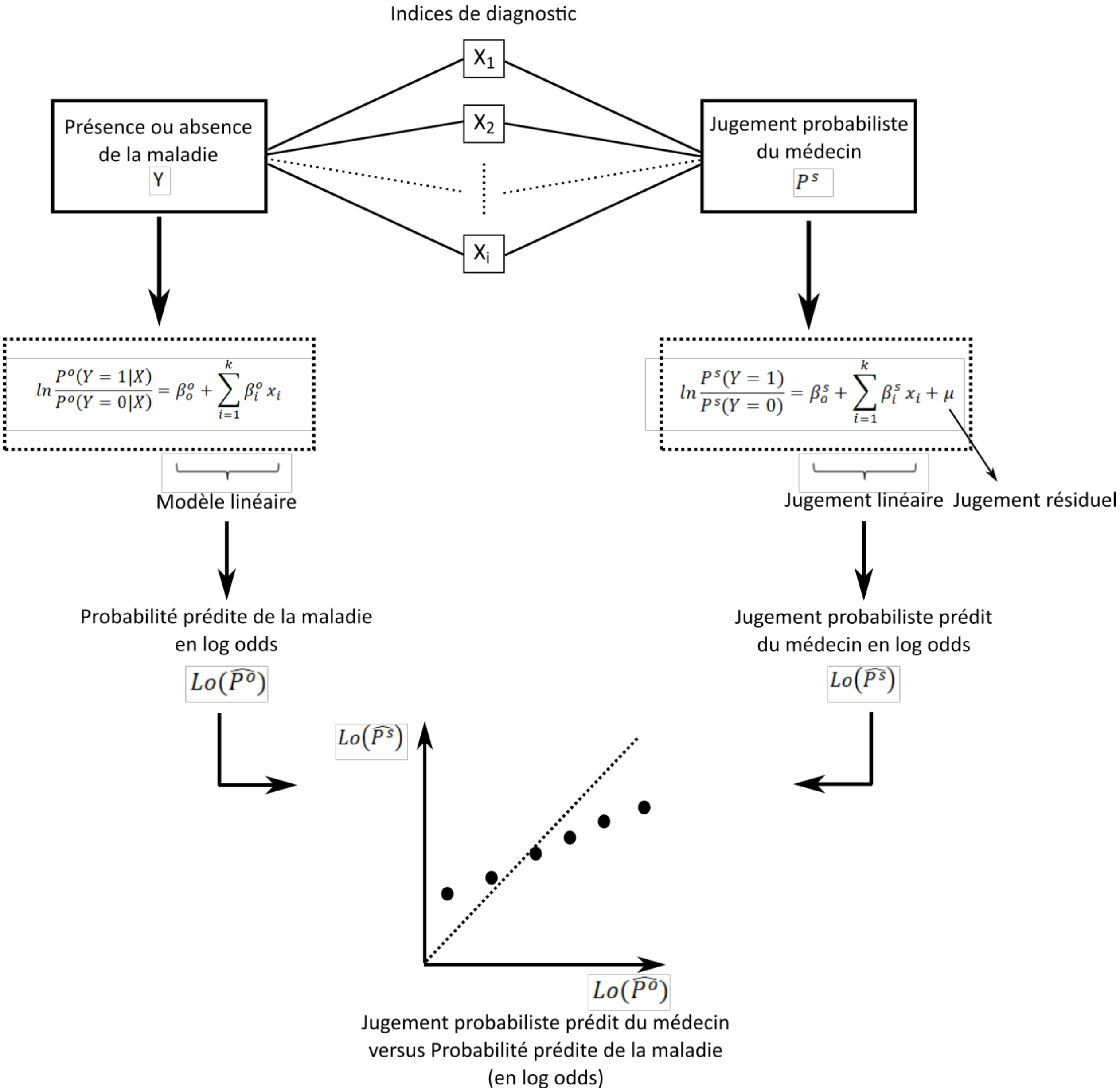

Indices de diagnostic
X_1
X_2
X_i
Présence ou absence de la maladie
Y
Jugement probabiliste du médecin
P^s
$ln\frac{P^o(Y=1|X)}{P^o(Y=0|X)} = \beta_o^o + \sum_{i=1}^{k} \beta_i^o x_i$
$ln\frac{P^s(Y=1)}{P^s(Y=0)} = \beta_o^s + \sum_{i=1}^{k} \beta_i^s x_i + \mu$
Modèle linéaire
Jugement linéaire
Jugement résiduel
Probabilité prédite de la maladie en log odds
$Lo(\widehat{P^o})$
Jugement probabiliste prédit du médecin en log odds
$Lo(\widehat{P^s})$
$Lo(\widehat{P^s})$
$Lo(\widehat{P^o})$
Jugement probabiliste prédit du médecin versus Probabilité prédite de la maladie (en log odds)

Figure 2: *Diagramme de l'approche du modèle de Lens appliqué au jugement*

Résultats

Nous avons trouvé que les médecins surestimaient les petites probabilités de la présence d'infection bactérienne et sous- estimaient les grosses probabilités, comparativement à la probabilité prédite par le modèle linéaire (**Figure 3A**). La **Figure3B** présente la classification des groupes de distorsion haut biais (en bleu foncé) et bas biais (bleu clair). La pente de l'ajustement linéaire était plus proche de 1 pour le groupe à bas biais (b=0,539) que pour le groupe à haut biais (b=0,289), ce qui capture le fait que les estimations des médecins étaient plus erronées dans le groupe à haut biais que dans le groupe à bas biais, par rapport à la probabilité prédite de IB. Finalement, nous avons constaté que la spécificité était significativement plus élevée dans le groupe à bas biais pour les décisions de traitement par antibiotique (0.776 vs 0.642, z=5.745, pvalue<0.001) et hospitalisation (0.406 vs 0.211, z=8.196, pvalue<0.001). Nos résultats suggèrent que la distorsion de probabilité pourrait causer des soins de santé inutiles c'est à dire fournir un traitement antibiotique et admettre à l'hôpital des nourrissons qui ne sont pas atteints de IB.

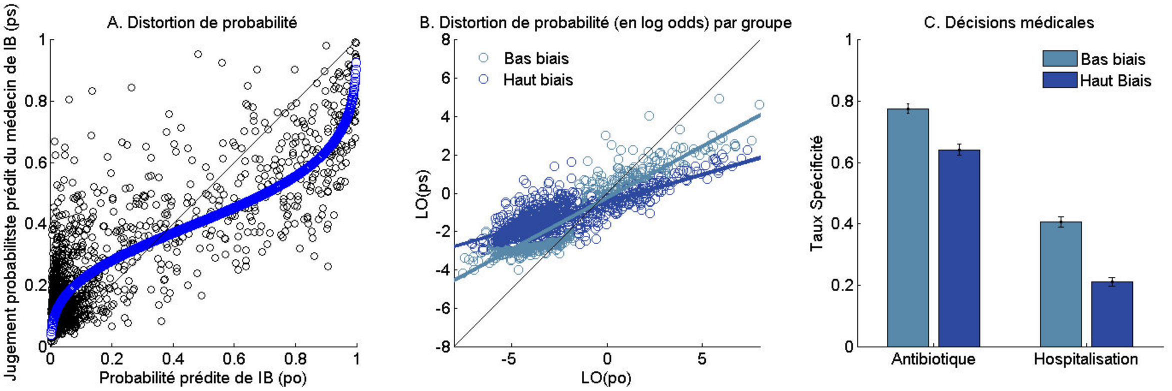


Figure 3. *(A)Jugement probabiliste prédit du médecin de IB (Infection Bactérienne) versus Probabilité prédite de IB (N=1848 patients). (B) Classification des haut et bas biais. (C) Taux de spécificité pour les décisions de traitement par antibiotique et d'hospitalisation par haut et bas biais.*

Combiner le modèle statistique avec l'homme intuitif pour améliorer la prise de décision?

Dans le cadre théorique, nous avons documenté que les médecins peuvent souffrir de plusieurs biais dans la façon dont ils traitent et intègrent l'information. Premièrement, ils peuvent mal évaluer les composantes analytique et intuitive par rapport à leurs valeurs objectives $W_{x_i}^o$ et W_x^o. Deuxièmement, ils peuvent intégrer de façon inexacte les composantes analytique (α) et/ou intuitive (β). Comment pouvons-nous remplacer le jugement des médecins par des modèles statistiques pour améliorer la qualité du jugement? Sur le plan opérationnel, nous pouvons remplacer l'homme analytique $(\alpha W_{x_i}^s)$ par le modèle statistique $(W_{x_i}^o)$. Cependant, nous n'observons pas la valeur optimale de l'intuition (W_x^o). Ainsi, le mieux que nous puissions faire pour améliorer le jugement du médecin, serait de combiner de façon optimale le modèle statistique $(W_{x_i}^o)$ avec l'homme intuitif (W_x^s). L'efficacité de cette approche combinée pour améliorer la qualité du jugement est bien documentée (Blattberg et Hoch, 1990). Toutefois, la question de savoir si ces scores statistiques combinés peuvent ou non améliorer la qualité des décisions réelles reste ouverte. En effet, certaines études dans la littérature suggèrent que les décisions des médecins ne dépendent pas entièrement de leur jugement (Sorum et al., 2002; Beckstead, 2017). La déviation de la décision des médecins par rapport au jugement pourrait également être pertinente.

Dans le chapitre 3, notre objectif est double: (1) évaluer, à l'aide de données médicales de terrain, si la combinaison du modèle statistique avec l'homme intuitif peut améliorer la qualité de la décision; (2) évaluer, sur le même ensemble de données, si la décision réelle des médecins s'écarte de la décision attendue, c'est-à-dire celle qui résulte de leur jugement, et, le cas échéant, si cette déviation constitue une information pertinente à prendre en compte lors de la conception d'un score statistique combiné.

Notre analyse est effectuée sur le même ensemble de données médicales, précédemment décrit pour le chapitre 1. Pour mesurer le modèle statistique et l'homme intuitif, nous utilisons la même approche d'identification que celle décrite précédemment au chapitre 1 (c. -à-d. l'approche du modèle de Lens). Nous définissons la décision attendue du médecin de traiter comme une intégration linéaire des indices de diagnostic et de leur jugement. L'écart par rapport à la décision attendue de traiter est la différence entre la décision réelle et la décision attendue. Nous estimons deux scores statistiques qui combinent: (1) le modèle statistique et l'homme intuitif, (2) le modèle statistique, l'homme intuitif et l'écart observé par rapport à la décision attendue. Enfin, nous testons si ces deux scores statistiques combinés peuvent améliorer la qualité des décisions observées.

Méthodes

Score statistique combinant la probabilité de maladie prédite par le modèle linéaire et le jugement résiduel

Nous estimons le score statistique combiné comme le modèle le mieux adapté pour prédire la présence ou l'absence de la maladie, compte tenu de la probabilité de maladie prédite par le modèle linéaire en log odds ($Lo(\widehat{P^o})$) et le jugement résiduel (μ), comme suit:

$$ln\frac{P\left(Y=1|Lo(\widehat{P^o}),\mu\right)}{P\left(Y=0|Lo(\widehat{P^o}),\mu\right)}=\alpha_0+\ \alpha_1Lo(\widehat{P^o})+\alpha_2\mu$$

Score statistique combinant la probabilité de maladie prédite par le modèle linéaire, le jugement résiduel et la décision résiduelle

- Décision résiduelle

Ici, nous décrivons notre méthode pour identifier la déviation du médecin par rapport à la décision attendue ("décision résiduelle"). La **Figure 4** résume la méthode.

Tout d'abord, en suivant la même approche du modèle de Lens que celle décrite à la section 1, nous modélisons la décision de traiter (T) comme une fonction logistique linéaire des indices x_i et du jugement de probabilité du médecin en log odds ($Lo(P^s)$) comme suit:

$$ln\frac{P(T=1|X,Lo(P^s))\ )}{P(T=0|X,Lo(P^s))}=\beta_o^T+\sum_{i=1}^{k}\beta_i^T\,x_i+\beta_s^T Lo(P^s)$$

En appliquant les pondérations de la régression, nous estimons en log odds la probabilité prédite de traitement ($Lo(\widehat{P^T})$).

Deuxièmement, nous mesurons la décision résiduelle (ω), comme la différence entre la décision réelle de traiter en log odds et $Lo\left(\widehat{P^T}\right)$, comme suit:

$$\begin{cases}\ \ln\left(\dfrac{0.99}{0.01}\right)-Lo\left(\widehat{P^T}\right), & si\ décision\ réelle\ est\ de\ traiter\\ \ln\left(\dfrac{0.01}{0.99}\right)-Lo\left(\widehat{P^T}\right), & si\ décision\ réelle\ est\ de\ ne\ pas\ traiter\end{cases}$$

où nous choisissons d'attribuer la valeur $ln\frac{0.99}{0.01}$ si la décision réelle est de traiter et $ln\frac{0.01}{0.99}$ si la décision réelle est de ne pas traiter, pour résoudre le problème des valeurs infinies.

La décision résiduelle (ω) contient la partie de la décision du médecin qui n'est pas expliquée par le jugement du médecin et l'intégration linéaire des indices.

- Score statistique

Enfin, nous estimons le score statistique combiné comme le modèle le mieux adapté pour prédire la présence ou l'absence de la maladie, compte tenu de la probabilité de maladie prédite par le modèle linéaire en log odds ($Lo(\widehat{P^o})$), du jugement résiduel (μ) et de la décision résiduelle (ω) comme suit:

$$ln\frac{P\big(Y=1|Lo(\widehat{P^o}),\mu,\omega\big)}{P\big(Y=0|Lo(\widehat{P^o}),\mu,\omega\big)}=\gamma_0+\ \gamma_1 Lo(\widehat{P^o})+\gamma_2\mu+\gamma_3\omega$$

Un score statistique combiné peut-il améliorer la qualité de la décision médicale?

Empiriquement, nous déterminons le seuil de décision pour chaque score combiné qui maximise la spécificité (c. -à-d. la proportion de patients sans maladie qui ne reçoivent pas de traitement) sous la contrainte que la sensibilité (c. -à-d. la proportion de patients atteints de maladie qui reçoivent un traitement) soit égale à la sensibilité de la décision de traitement réelle. Nous pouvons ensuite comparer la spécificité obtenue avec le score statistique combiné à la spécificité réelle.

Données

Nous appliquons notre méthode aux données présentées dans la section précédente.

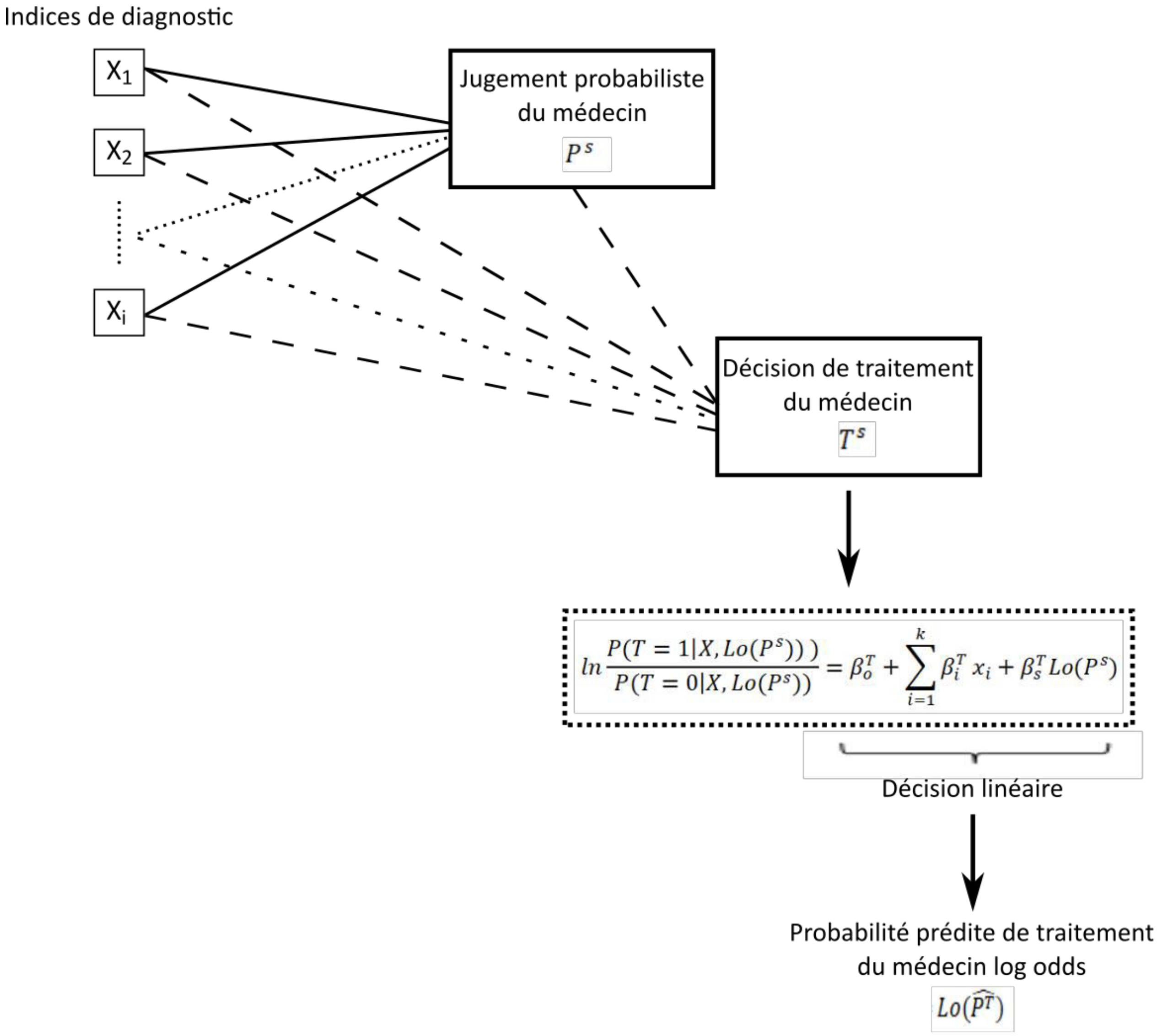

Indices de diagnostic
X_1
X_2
X_i
Jugement probabiliste du médecin
P^s
Décision de traitement du médecin
T^s
$ln\frac{P(T=1|X,Lo(P^s)))}{P(T=0|X,Lo(P^s))} = \beta_o^T + \sum_{i=1}^{k} \beta_i^T x_i + \beta_s^T Lo(P^s)$
Décision linéaire
Probabilité prédite de traitement du médecin log odds
$Lo(\widehat{P^T})$

Figure 4: *Diagramme de l'approche du modèle de Lens appliqué à la décision*

Résultats

Nous avons observé qu'une aide mécanique combinée à l'intuition humaine pouvait améliorer la qualité du jugement dans le groupe des haut biais. Cependant, ce score combiné n’était pas suffisant pour améliorer la spécificité des décisions. Pour améliorer la spécificité des décisions réelles dans le groupe des hauts biais (courbe en rose), il était nécessaire d'inclure dans le modèle combiné l'écart observé des médecins par rapport à la décision attendue (courbe en rouge).

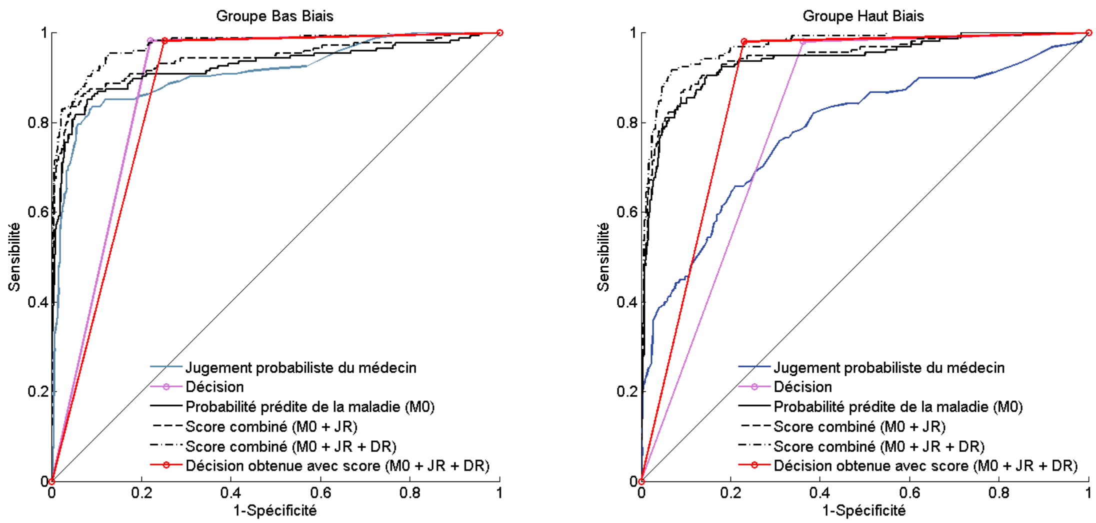


Figure 5: *Courbe ROC ("Receiver Operating Characteristic") détection Infection Bactérienne, pour les groupes bas et haut biais, pour le jugement probabiliste, la décision de traitement par antibiotique réelle, la probabilité prédite par le modèle linéaire (M0), le score combinant la probabilité de maladie prédite par le modèle linéaire et le jugement résiduel (M0 + JR), le score combinant la probabilité de maladie prédite par le modèle linéaire, le jugement résiduel et la décision résiduelle (M0 + JR + DR), et la décision de traitement par antibiotique obtenue avec le score M0 + JR + DR.*

Quels sont les facteurs qui influencent le traitement de l'information de l'homme?

Dans le cadre théorique proposé, nous avons documenté que les médecins peuvent souffrir de plusieurs biais dans la façon dont ils traitent et intègrent l'information. Premièrement, ils peuvent mal évaluer les composantes analytiques et intuitives par rapport à leurs valeurs objectives $W_{x_i}^{o}$ et W_{x}^{o}. Deuxièmement, ils peuvent intégrer de façon inexacte les composantes analytique (α) et/ou intuitive (β). Quels sont les facteurs qui influencent le traitement de l'information de l'homme?

Dans les chapitres 4,5 et 6, nous examinons les sources potentielles de mauvaise évaluation de la composante analytique[14] (chapitre 4) et d'intégration inexacte des composantes analytique et intuitive[15] (chapitres 5 et 6). Le premier facteur que nous étudions est la mémoire de travail. Il est bien documenté que la capacité de conserver l'information en mémoire de travail est limitée dans le temps et en capacité (pour un résumé, voir Wickens et al, 2015). Le deuxième facteur que nous étudions est la mauvaise évaluation de l'élément intuitif [16] car "l'intuition est parfois merveilleuse et parfois imparfaite" (Kahneman & Klein, 2009).

Dans le chapitre 4, nous testons si la capacité des gens à apprendre la valeur de la composante analytique, en l'absence de feedback externe, dépend de la qualité de leur composante intuitive. Nous pensons qu'en l'absence de feedback externe, la seule source d'information qui peut aider les gens à savoir quels indices de diagnostic sont pertinents ou non est leur composante intuitive. De plus, nous testons si la mémoire de travail est également nécessaire pour apprendre dans cette situation.

Dans le chapitre 5, nous examinons si la capacité des personnes à intégrer les composantes analytique et intuitive dépend de la qualité de leur composante intuitive. Nous considérons que les personnes peuvent mal évaluer leur composante intuitive, ce qui affecterait la qualité du processus d'intégration.

[14] c.-à-d. $W_{x_i}^{s} \neq W_{x_i}^{o}$

[15] c.-à-d. $ln\frac{P_{prior}^{s}(Y=1)}{P_{prior}^{s}(Y=0)} + \alpha\sum_{i=1}^{k} W_{x_i}^{s} + \beta W_{x}^{s} \neq ln\frac{P_{prior}^{o}(Y=1)}{P_{prior}^{o}(Y=0)} + \sum_{i=1}^{k} W_{x_i}^{o} + W_{x}^{o}$

[16] c. -à-d. $W_{x}^{s} \neq W_{x}^{o}$

Dans le chapitre 6, nous examinons si la capacité des gens à intégrer les composantes analytique et intuitive dépend de leur capacité de mémoire de travail, en supposant que cette capacité est nécessaire pour manipuler l'information à cette étape.

Pour mesurer la façon dont les gens valorisent et intègrent la composante analytique, nous considérons la situation simple où un seul indice explicite est disponible. L'évaluation de l'indice explicite est mesurée par un report subjectif lorsque la valeur objective n'est pas disponible. L'intégration de l'indice explicite est mesurée en observant comment il est utilisé lorsque la valeur objective est connue.

Pour mesurer la qualité du composant intuitif, nous proposons d'utiliser la confiance dans sa propre décision, lorsqu'il n' y a pas d'indice explicite pour prendre la décision, comme mesure de la valeur qu'on attribue au composant intuitif.

Pour tester nos hypothèses sur l'impact de ces deux facteurs, nous avons effectué deux expériences avec une simple tâche de décision perceptuelle. Dans une expérience, les participants devaient apprendre la valeur d'un indice explicite, en l'absence de feedback externe. Dans une autre expérience, on a demandé aux participants d'intégrer un indice explicite (dont la valeur informative leur a été fournie) avec le stimulus perceptuel. Pour chaque expérience, nous avons mesuré séparément la mémoire de travail et la confiance dans la décision.

Méthode

Dans la section suivante, nous développons la méthode pour mesurer la qualité du composant intuitif et la méthode pour mesurer la capacité d'intégrer le composant analytique et intuitif.

Mesure de la qualité du composant intuitif

Comme décrit précédemment, pour mesurer la qualité du composant intuitif, nous proposons d'utiliser la confiance dans la décision. Nous décrivons ci-après les mesures utilisées pour évaluer la qualité des reports de confiance (Fleming & Lau, 2014).

A partir des reports de confiance, il est important de faire la distinction entre biais et sensibilité. Le biais de confiance, également appelé sous- ou sur-confiance, correspond à la tendance à donner des reports de confiance faible ou élevée. La sensibilité en confiance correspond à la capacité de distinguer entre les bonnes et les mauvaises réponses. La **Figure 6** ci-dessous illustre schématiquement la différence entre sensibilité et biais. Les distributions bleue et rouge représentent respectivement les reports de confiance lorsque l'observateur est correct et incorrect. Par exemple, un observateur peut être bon pour distinguer ses réponses correctes des réponses incorrectes (sensibilité élevée) mais afficher une confiance excessive dans l'ensemble (biais élevé).

Nous quantifions la qualité de la composante intuitive de l'observateur en mesurant les deux dimensions: le biais (surconfiance) et la sensibilité (sensibilité métacognitive).

Biais dans la confiance : Nous avons quantifié le biais dans la confiance comme la différence entre la confiance moyenne et la performance moyenne.

Sensibilité dans la confiance : Nous avons estimé la sensibilité dans la confiance à l'aide de la méthode méta-d' (Maniscalco and Lau, 2012) (http://www.columbia.edu/bsm2015/type2sdt/).

Dans un cadre de détection des signaux, Meta-d'correspond au niveau de sensibilité SDT de type 1 (d') qu'un observateur métacognitivement idéal aurait dû atteindre pour produire les données observées de type 2. L'efficacité métacognitive est définie par la mesure relative meta-d' /d'. Si le rapport est égal à un, alors l'observateur est métacognitivement idéal. Si le méta-d' /d'est inférieur à un, alors l'observateur est métacognitivement inefficace.

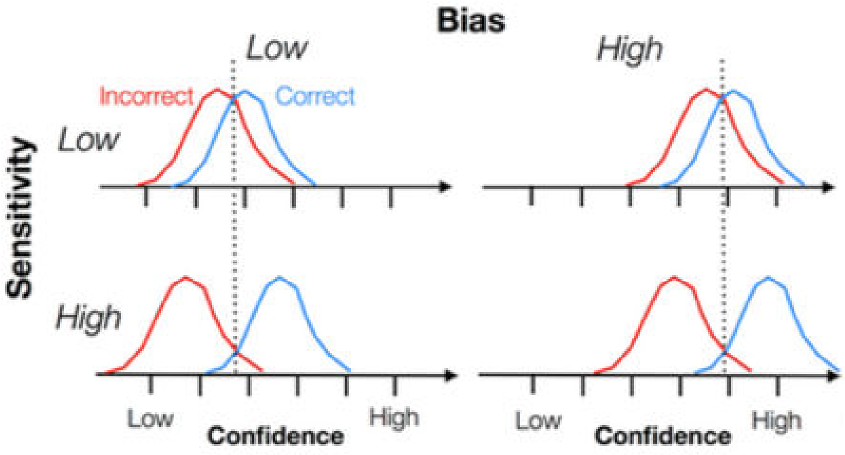


Figure 6: *Représentation schématique montrant la différence entre sensibilité dans la confidence et biais dans la confiance. Extrait de "How to measure metacognition" by Fleming, S. M., & Lau, H. C (2014). Frontiers in human neuroscience, 8. Note traduction des termes en français : Bias, Biais ; Sensitivity, Sensibilité ; Low, Bas ; High, Haut.*

Mesurer la capacité d'intégrer les composantes analytique et intuitive

Pour évaluer dans quelle mesure les participants sont capables d'intégrer les composantes analytique et intuitive, nous avons besoin d'un benchmark optimal. Pour déterminer comment les deux composantes doivent être intégrées de manière idéale, nous devons connaître la valeur objective du composant intuitif (c'est-à-dire le signal interne). Nous utilisons la Théorie de la Détection du Signal (TDS) (Green & Swets, 1966) qui propose une formalisation du signal interne. Nous présentons ci-après le cadre TDS et notre mesure d'intégration.

Théorie de la Détection du Signal

Transcrit dans notre cadre médical, le modèle TDS suppose que le signal interne de l'observatrice suit une distribution gaussienne conditionnelle à la présence de la maladie ($Y = 1$) ($\sim\mathcal{N}({}^{+d'}/_{2}, 1)$) ou à l'absence de la maladie ($Y = 0$) ($\sim\mathcal{N}({}^{-d'}/_{2}, 1)$). La distance entre les deux courbes gaussiennes est égale à d', nommée sensibilité, ce qui correspond à la capacité de l'observatrice de discriminer entre maladie présente ou absente. L'observatrice fixe un critère de décision c sur son axe interne qui détermine au-dessus de quel niveau de signal interne elle répondra “Y=1” (c'est-à-dire "Maladie présente"). La **Figure 7** illustre le modèle TDS.

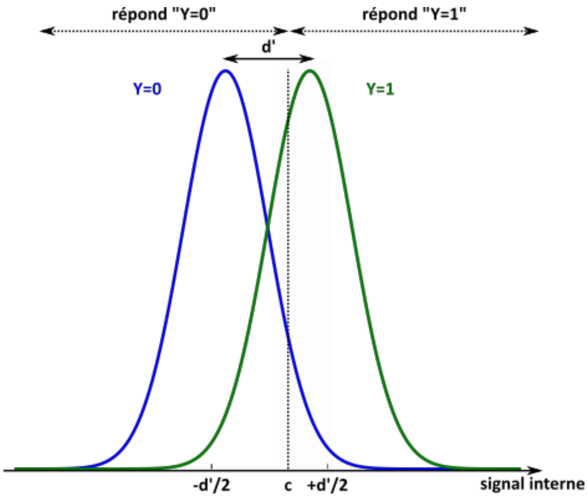


Figure 7: *Diagramme du modèle TDS*

Suivant ce modèle, la probabilité de réponse du médecin dépend du critère c, de la sensibilité d' et de l'état du patient Y tel que décrit dans le tableau ci-dessous.

		Etat du patient	
		Malade	Non malade
Réponse du médecin	Répond “Y=1”	$P(\text{“Y = 1”}\vert Y=1)$ $= 1 - F({}^{-d'}/_{2} + c)$	$P(\text{“Y = 1”}\vert Y=0)$ $= 1 - F({}^{d'}/_{2} + c)$
	Répond “Y=0”	$P(\text{“Y = 0”}\vert Y=1)$ $= F({}^{-d'}/_{2} + c)$	$P(\text{“Y = 0”}\vert Y=0)$ $= F({}^{d'}/_{2} + c)$

où $F(z) = \int_{t=-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$.

Estimation des paramètres TDS critère c et sensibilité d'

Les paramètres c et d' peuvent être inférés à partir des réponses de l’observateur en calculant

$P(\text{“Y = 1”}|Y=1)$ et $P(\text{“Y = 0”}|Y=0)$, comme ci- dessous.

$$P(\text{“Y = 1”}|Y=1) = 1 - F\left({}^{-d'}/_{2} + c\right) = F\left({}^{d'}/_{2} - c\right)$$

Alors ${}^{d'}/_{2} - c = Z(P(\text{“Y = 1”}|Y=1))$ où $Z(f) = F^{-1}(z)$

$$P(\text{“Y = 0”}|Y=0) = F\left({}^{d'}/_{2} + c\right)$$

Alors ${}^{d'}/_{2} + c = Z(P(\text{“Y = 0”}|Y=0))$.

Donc

$$d' = Z\big(P(\text{“Y = 1”}|Y=1)\big) + Z(P(\text{“Y = 0”}|Y=0))$$

$$c = \frac{1}{2}\Big(Z\big(P(\text{“Y = 0”}|Y=0)\big) - Z(P(\text{“Y = 1”}|Y=1)\Big)$$

Application SDT: l'observateur reçoit un indice explicite x_1 et un signal interne x

Ici, nous dérivons l'intégration optimale de l’indice explicite x_1 et du signal interne x.

Etant donné l’indice x_1, le signal interne x et la prior objective $P^o_{prior}(Y = 1)$, l'observateur Bayésien forme sa croyance postérieure selon la règle de Bayes, comme suit:

$$ln\frac{P^o(Y = 1|x_1, x)}{P^o(Y = 0|x_1, x)} = ln\frac{P^o_{prior}(Y = 1)}{P^o_{prior}(Y = 0)} + W^o_{x_1} + W^o_x$$

L’observateur Bayésien décide de répondre "$Y = 1$" si $ln\frac{P^o(Y=1|x_1,x)}{P^o(Y=0|x_1,x)} > 0$

Le critère de décision optimal $c^{opt}(X)$ est la valeur de x tel que $ln\frac{P^o(Y=1|x_1,x)}{P^o(Y=0|x_1,x)} = 0$

Dans le cadre TDS, nous pouvons calculer le poids objectif de l’évidence du signal interne x, comme suit:

$$W^o_x = ln\left(\frac{P^o(x|Y = 1)}{P^o(x|Y = 0)}\right)$$

où $P^o(x|Y = 1) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(x-d'/2\right)^2}$ et $P^o(x|Y = 0) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(x+d'/2\right)^2}$

$$W^o_x = ln\left(\frac{e^{-\frac{1}{2}\left(x-d'/2\right)^2}}{e^{-\frac{1}{2}\left(x+d'/2\right)^2}}\right) = xd'$$

Donc

$$ln\frac{P^o(Y = 1|x_1, x)}{P^o(Y = 0|x_1, x)} = ln\frac{P^o_{prior}(Y = 1)}{P^o_{prior}(Y = 0)} + W^o_{x_1} + xd'$$

Le critère de décision optimal est égal à :

$$x^* = -\frac{1}{d'}\left(ln\frac{P^o_{prior}(Y = 1)}{P^o_{prior}(Y = 0)} + W^o_{x_1}\right) \equiv c^{opt}$$

Ajustement des critères : Nous quantifions la mesure dans laquelle l'observateur est capable d'intégrer l’indice explicite x_1 (c. -à-d. la composante analytique) et le signal interne x (c. -à-d. la composante intuitive) par rapport à l'intégration idéale comme la déviation de l'observateur par rapport au critère de décision optimal: $\frac{c^{obs}}{c^{opt}}$

Données

Protocole expérimental

Nous avons effectué deux expériences (pour plus de détails sur le protocole expérimental, voir la **Figure 8**). Dans chaque expérience, les participants se sont livrés à une tâche simple de décision perceptuelle: ils devaient déterminer lequel des deux ensembles présentés sur l'écran d'ordinateur contenait le plus de points (voir la **Figure 9a**). Chaque participant a suivi deux sessions expérimentales à 4 jours d'intervalle. L'ordre des deux sessions était contrebalancé entre les participants. Dans l'expérience d' « apprentissage de l'indice explicite » (N=65), les participants ont participé à une session d' « apprentissage de l'indice explicite » et à une session de « confiance ». Dans l'expérience d' « intégration de l'indice explicite » (N=69), les participants ont participé à une session d' « intégration de l'indice explicite » et à une session de « confiance ». Chaque session commençait par une tâche de mémoire de travail. Les données de l'expérience d' « apprentissage de l'indice explicite » sont analysées dans le chapitre 4 et les données de l'expérience d' « intégration de l'indice explicite » sont analysées dans les chapitres 5 et 6. Ci-dessous nous décrivons chaque séance.

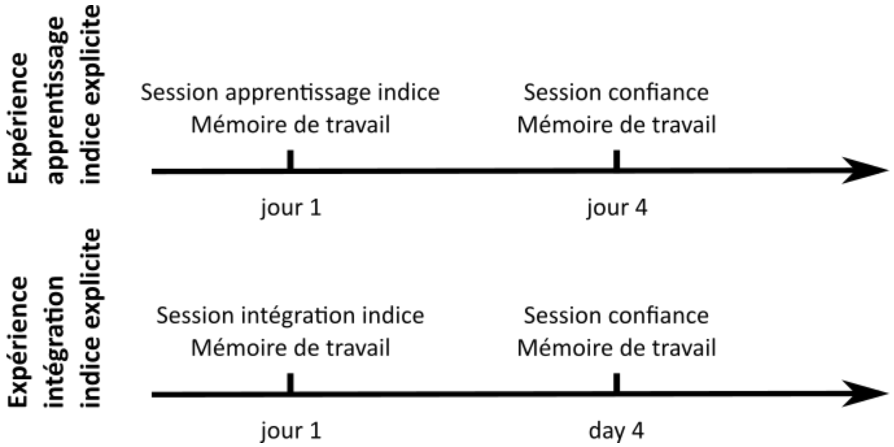

Expérience apprentissage indice explicite
Session apprentissage indice
Mémoire de travail
Session confiance
Mémoire de travail
jour 1
jour 4
Expérience intégration indice explicite
Session intégration indice
Mémoire de travail
Session confiance
Mémoire de travail
jour 1
day 4

Figure 8: *Expérience d'*apprentissage de l'indice explicite *(N=65) et expérience d'intégration de l'indice explicite (N=69). L'ordre des deux sessions était contrebalancé entre les participants.*

Session "Apprentissage de l'indice explicite": Chaque essai débutait par un indice central. L'indice était un carré, un cercle ou un triangle. Une forme prédisait la catégorie de gauche, une autre la catégorie de droite (tous deux avec probabilité p=0,75) et une autre ne donnait aucune information sur la catégorie à venir. Les participants n'ont pas été informés des associations entre les indices et les probabilités de survenue d'un stimulus, mais ils ont été informés qu'il y avait un indice de « gauche », de « droite » et « neutre ». Au début de la session, il leur a été demandé d'apprendre les associations indice-stimulus, afin d'optimiser leurs décisions et on leur a dit qu'à la fin de la session, ils devaient reporter ces associations (par exemple: Quelle forme était prédictive de la catégorie de gauche?)

Session "Intégration de l'indice explicite": Chaque essai débutait par un indice central présenté pendant 250ms, avant la croix de fixation. L'indice était soit un triangle pointant vers la gauche ou la droite de l'écran, indiquant la réponse correcte avec une validité de 75 % (condition avec indice), soit un diamant ne fournissant aucune information (condition sans indice). Les participants ont été pleinement informés de la signification de ces indices et ont reçu pour instruction d'utiliser à la fois le stimulus et l'indice pour prendre les meilleures décisions possibles.

Session "confiance": Après chaque réponse à gauche ou à droite, les participants devaient indiquer leur probabilité subjective de succès sur une échelle quantitative allant de 50 % à 100 % de confiance (voir la **Figure 9b**).

Tâche de mémoire de travail : A chaque essai, les participants ont reçu une séquence de lettres et ont dû reporter les dernières lettres dans l'ordre (voir la **Figure 9c**).

Tests empiriques

Dans le chapitre 4, nous étudions si l'identification correcte par les participants des associations indices-stimulus reportées lors de la session d'apprentissage de l'indice explicite est liée à leur sensibilité en confiance (mesurée à 4 jours d'intervalle) et à leur capacité de mémoire de travail (mesure moyenne sur les deux sessions).

Au chapitre 5, nous évaluons si l'ajustement des critères des participants mesuré lors de la séance d'intégration des indices explicites est lié à leur biais de confiance (mesuré à 4 jours d'intervalle).

Au chapitre 6, nous examinons si l'ajustement des critères des participants mesuré lors de la séance d'intégration des indices est lié à leur capacité de mémoire de travail (mesurée le même jour).

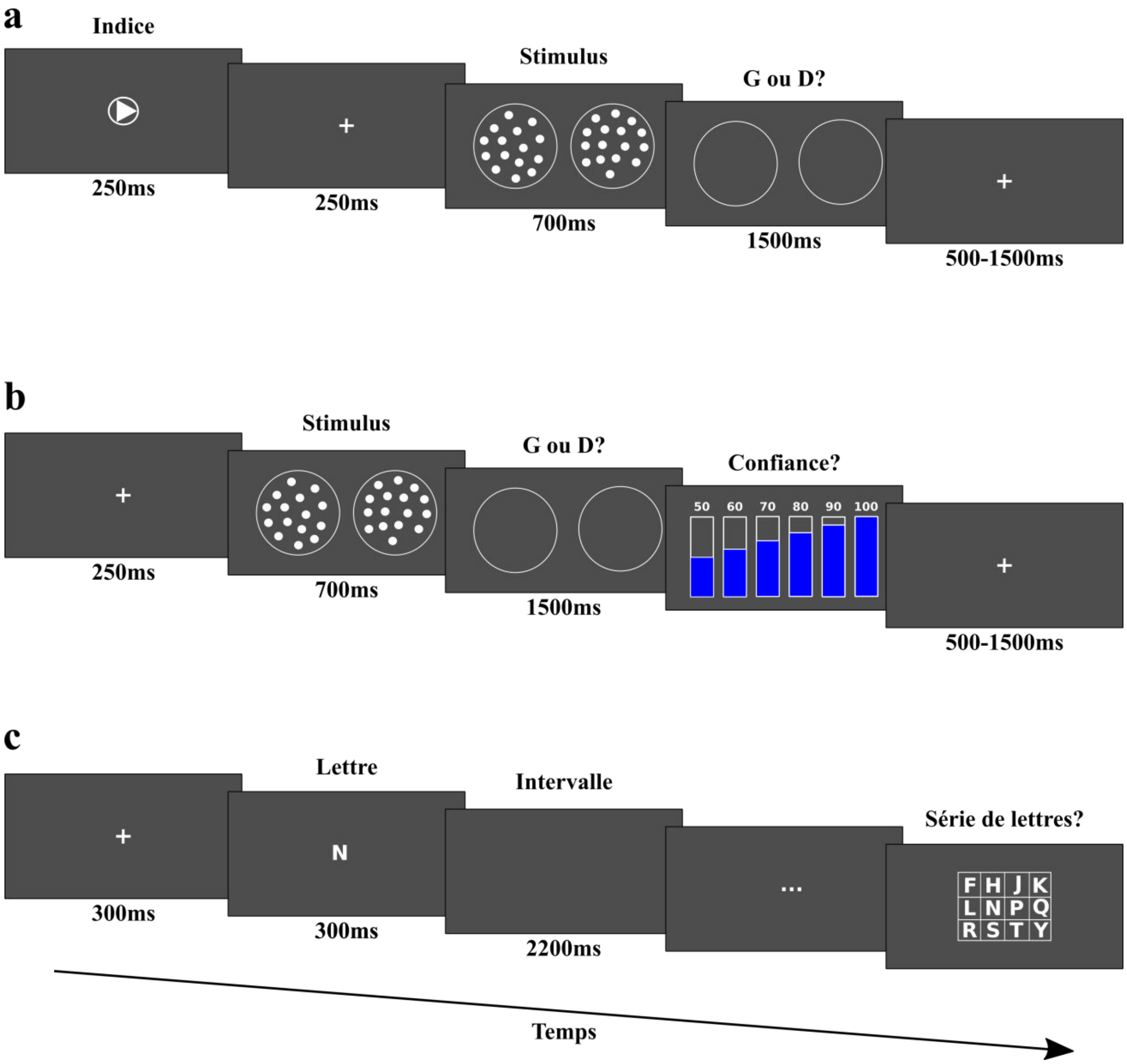


Figure 9: (a) *Session "apprentissage indice explicite" / "Intégration indice explicite".* ***(b)*** *Session "Confiance".* ***(c)*** *Mémoire de travail.*

Résultats

Chapitre 4

Nous avons constaté que notre hypothèse principale a été confirmée: la sensibilité dans la confiance était plus élevée chez les participants qui ont identifié avec succès les associations indices-stimuli ("apprenants": M=0,972 SD=0,405) que chez les participants qui n'en ont pas identifié ("non apprenants": M=0,758 SD=0,424), comme l'illustre la **Figure 10A**. La différence de sensibilité dans la confiance entre les "apprenants" et les "non-apprenants" était significative (test T: t (63)=-2,078, p=0,0418). Par ailleurs, nous avons constaté que des scores de mémoire plus élevés étaient associés à la réussite de l'identification des indices. La **Figure 10B** illustre comment les scores de mémoire étaient plus élevés pour les "apprenants" que pour les "non-apprenants" (test T: t (63)=-2,295, p=0,0251).

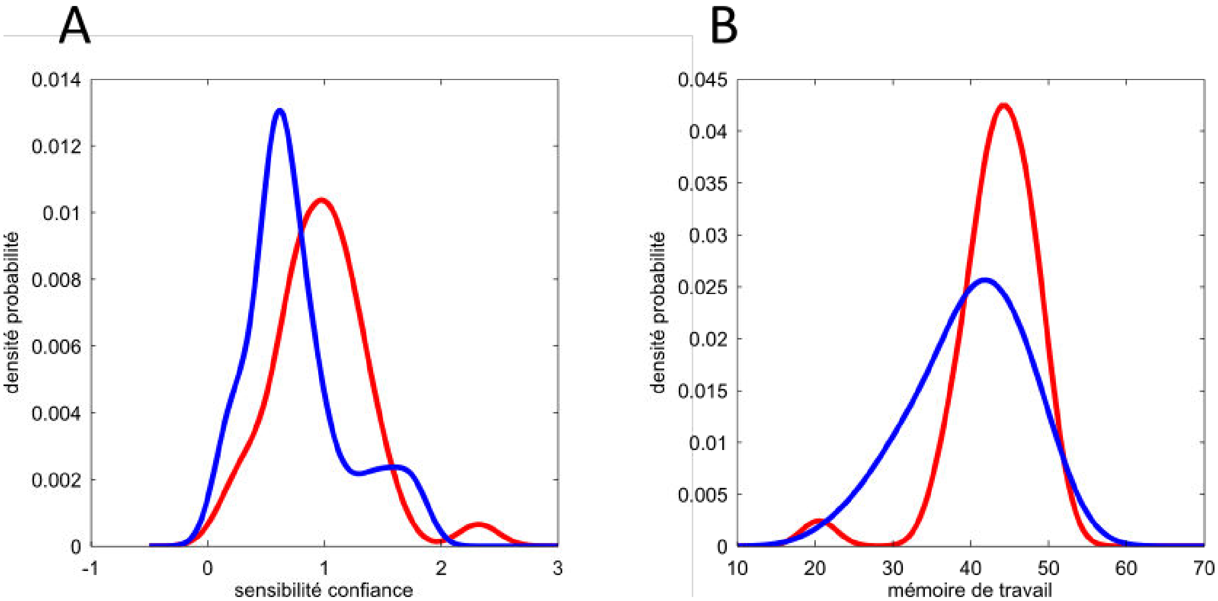


Figure 10. *(A) Distribution de la sensibilité dans la confiance (ratio de méta-d' sur d') parmi les participants qui ont identifié avec succès les 3 associations indices-stimuli ("apprenants") et les participants qui ne l'ont pas fait ("non apprenants"). (B) Distribution des scores de la mémoire de travail pour les "apprenants" et les "non-apprenants".*

Chapitres 5 et 6

À notre grande surprise, nous n’avons pas trouvé de corrélation entre l’ajustement des critères et la sur-confiance (r=-0,049, p=0,691) (Figure 11A). Par ailleurs, confirmant notre hypothèse, nous avons trouvé que la mémoire de travail était significativement corrélée avec l'ajustement des critères (r= 0,296, p= 0,013) (Figure 11B).

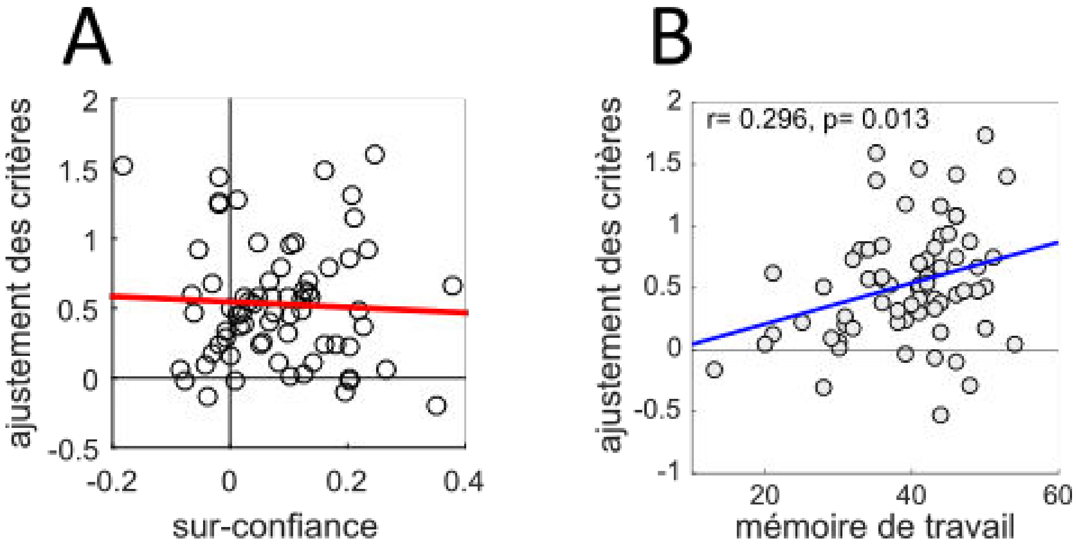


Figure 11: *Relations entre ajustement des critères et (A) sur-confiance ; (B) mémoire de travail. Chaque point est un participant (N=69). Les lignes (rouge et bleue) représentent les meilleures lignes de régression.*

Conclusion

Cette thèse étudie la distorsion de probabilité dans le jugement clinique afin de comparer le jugement des médecins à des modèles statistiques. Pour répondre à ce débat sur l'opposition médecin contre modèle, nous avons élaboré un cadre théorique sur la façon dont les médecins peuvent traiter l'information. Ce cadre nous a permis de considérer les mécanismes cognitifs par lesquels les médecins peuvent s'écarter des modèles et, par conséquent, comment leurs probabilités subjectives peuvent s'écarter des probabilités objectives. Dans ce cadre, nous avons considéré que les médecins forment leur jugement clinique en intégrant une composante analytique et une composante intuitive. Nous avons documenté que les médecins peuvent souffrir de plusieurs biais dans la façon dont ils traitent et intègrent l'information. Ils peuvent mal évaluer les composantes analytique et intuitive. Ils peuvent intégrer de façon inexacte les composantes analytique et/ou intuitive.

La thèse recueille des résultats sur le terrain et dans le laboratoire qui, à notre avis, peuvent éclairer le débat. A partir de données médicales sur le terrain, nous avons constaté que la composante analytique des médecins était biaisée. Ils n'étaient pas aussi bons que les modèles statistiques pour l'intégration des données médicales. Ils sous-pondéraient les évidences médicales, et attribuaient également de la valeur à des évidences médicales non pertinentes, comparativement à notre modèle statistique. Par conséquent, les médecins, dans notre set de données, surestimaient les petites probabilités que le patient soit atteint de la maladie et sous-estimaient les grandes probabilités. Fait important, notre analyse a révélé que leur jugement biaisé sur les probabilités pourrait causer des soins de santé excessifs. Comment alors améliorer le jugement des médecins? Tout d'abord, nous avons cherché à déterminer si le remplacement du jugement des médecins par la probabilité générée par notre modèle statistique pourrait améliorer la qualité des décisions médicales. Notre analyse des données a révélé que notre score statistique, qui combinait le modèle analytique avec la composante intuitive du médecin, n'était pas suffisant. Il était nécessaire d'inclure la déviation observée des médecins par rapport à la décision attendue dans notre score statistique afin d'améliorer la prise de décision. Deuxièmement, nous avons testé en laboratoire des facteurs qui peuvent avoir une incidence sur le traitement de l'information de l’homme. Nous avons constaté que la capacité des participants à apprendre la valeur de la composante analytique, en l'absence de feedback externe, dépend de la qualité de leur composante intuitive et de leur capacité de mémoire de travail. Dans une deuxième expérience, nous avons constaté que la capacité des participants à intégrer les composantes analytique et intuitive dépend de leur capacité de mémoire de travail. D'autre part, nous n'avons trouvé aucune évidence en faveur de l'hypothèse selon laquelle une mauvaise évaluation du composant intuitif affecte l'intégration.

Nos résultats peuvent avoir d'importantes implications. Dans l'ensemble, nous avons constaté au chapitre 1 que le médecin analytique biaisé prend de moins bonnes décisions. Il serait donc nécessaire d'envisager des solutions qui pourraient « débiaiser » le médecin analytique. Dans la thèse, nous avons commencé à explorer deux solutions possibles. Au chapitre 3, nous avons envisagé de « remplacer » le médecin analytique par le modèle statistique. Aux chapitres 4,5 et 6, nous avons identifié des facteurs qui peuvent influencer la façon dont les médecins traitent l'information dans le but d’envisager des interventions qui pourraient être mises en place.

Abstract: Probability distortion in clinical judgment: field study and laboratory experiments

This thesis studies probability distortion in clinical judgment to compare physicians’ judgment with statistical models. We considered that physicians form their clinical judgment by integrating an analytical component and an intuitive component. We documented that physicians may suffer from several biases in the way they evaluate and integrate the two components. This dissertation gathers findings from the field and the lab. With actual medical data practice, we found that physicians were not as good as the statistical models at integrating consistently medical evidence. They over-estimated small probabilities that the patient had the disease and under- estimated large probabilities. We found that their biased probability judgment might cause unnecessary health care treatment. How then can we improve physician judgment? First, we considered to replace physician judgment by the probability generated from our statistical model. To actually improve decision it was necessary to develop a statistical score that combines the analytical model, the intuitive component of the physician and his observed deviation from the expected decision. Second, we tested in the lab factors that may affect information processing. We found that participants’ ability to learn about the value of the analytical component, without external feedback, depends on the quality of their intuitive component and their working memory. We also found that participants’ ability to integrate both components together depends on their working memory but not their evaluation of the intuitive component.

Keywords: Probability distortion, Medical decision-making, Clinical judgment, Intuition, Information processing, Working-memory, Statistical aids

Résumé : Distorsion de probabilité dans le jugement clinique: étude de terrain et expériences en laboratoire

Cette thèse étudie la distorsion de probabilité dans le jugement clinique afin de comparer le jugement des médecins à des modèles statistiques. Nous supposons que les médecins forment leur jugement clinique en intégrant une composante analytique et une composante intuitive. Dans ce cadre, les médecins peuvent souffrir de plusieurs biais dans la façon dont ils évaluent et intègrent les deux composantes. Cette thèse rassemble les résultats obtenus sur le terrain et en laboratoire. A partir de données médicales, nous avons constaté que les médecins n'étaient pas aussi bons que les modèles statistiques à intégrer des évidences médicales. Ils surestimaient les petites probabilités que le patient soit malade et sous-estimaient les probabilités élevées. Nous avons constaté que leur jugement biaisé pourrait entraîner un sur-traitement. Comment améliorer leur jugement? Premièrement, nous avons envisagé de remplacer le jugement du médecin par la probabilité de notre modèle statistique. Pour améliorer la décision, il était nécessaire d'élaborer un score statistique qui combine le modèle analytique, la composante intuitive du médecin et sa déviation observée par rapport à la décision attendue. Deuxièmement, nous avons testé en laboratoire des facteurs qui peuvent influencer le traitement de l'information. Nous avons trouvé que la capacité des participants à apprendre la valeur de la composante analytique, sans feedback externe, dépend de la qualité de leur composante intuitive et de leur mémoire de travail. Nous avons aussi trouvé que la capacité des participants à intégrer les deux composantes dépend de leur mémoire de travail, mais pas de leur évaluation de la composante intuitive.

Mots-clés : Distorsion de probabilité, Décision médical, Jugement clinique, Intuition, Traitement de l’information, Mémoire de travail, Aides statistiques