Communicating Uncertainty in Explanations of the Outcomes of Machine Learning Models

Ingrid Zukerman
Dept. of Data Science and AI
Faculty of Information Technology
Monash University, Australia
ingrid.zukerman@monash.edu

Sameen Maruf*
Oracle
Melbourne, Australia
sameen.maruf@gmail.com

Abstract

We consider two types of numeric representations for conveying the uncertainty of predictions made by Machine Learning (ML) models: confidence-based (e.g., "the AI is 90% confident") and frequency-based (e.g., "the AI was correct in 180 (90%) out of 200 cases"). We conducted a user study to determine which factors influence users' acceptance of predictions made by ML models, and how the two types of uncertainty representations affect users' views about explanations. Our results show that users' acceptance of ML model predictions depends mainly on the models' confidence, and that explanations that include uncertainty information are deemed better in several respects than explanations that omit it, with frequency-based representations being deemed better than confidence-based representations.

1 Introduction

There is a large body of research on how to communicate the uncertainty associated with predicted outcomes, in particular in healthcare (Freeman, 2019; Simpkin and Armstrong, 2019; Spiegelhalter, 2017; Zipkin et al., 2014). In that research, the uncertainty is derived from simple historical population averages, e.g., iPrevent provides such information to enable patients to assess their risk of breast cancer. However, in the age of personalised medicine, the uncertainty is obtained from the predictions of Machine Learning (ML) models, which are tailored to individuals by learning complex relationships between a prediction (e.g., a disease) and a large number of variables. Understanding this uncertainty is essential to improve medical decision making (Begoli et al., 2019). However, there is relatively little research on conveying the uncertainty of predictions made by ML models.

In this paper, we consider two types of numeric representations for conveying the uncertainty of ML predictions: *Confidence* and *Confidence+Frequency* (denoted *%Frequency*). The Confidence representation was proposed in (Cau et al., 2023) to convey how certain an AI is of its prediction (e.g., "The AI is 80% confident of the predicted outcome"); and the %Frequency representation, which is best practice for conveying population-based statistics in healthcare (Freeman, 2019; Trevena et al., 2013), gives a frequency out of a *reference class* (a base population), and the corresponding percentage. The reference class may be *generic* (e.g.,"*Out of 200 people*, 160 (80%) will develop this side effect") or *tailored* (e.g., "*Out of 200 people like you*, …"). We chose the latter, as recommended in (Trevena et al., 2013).

We describe a user study that examines (1) the influence of these two representations of uncertainty and other factors on users' acceptance of the predictions of an ML model; and (2) users' views about explanations featuring these representations of uncertainty. Our study was conducted in a healthcare scenario, sourced from the Busselton dataset (Knuiman et al., 1998), where an AI uses demographic, medical and lifestyle information to predict whether a person is at risk of *Coronary Heart Disease* (*CHD*). *Concessive-contrastive* explanations for these predictions, without uncertainty information, were used as a baseline. We chose these explanations owing to their support in the literature (Biran and McKeown, 2017; Maruf et al., 2023; Miller, 2019).

Table 1 shows a sample scenario, a concessive-contrastive explanation for an at-risk prediction, and a Confidence and a %Frequency representation of uncertainty. The baseline explanation follows the general template used in (Maruf et al., 2024) for the concessive-contrastive component of conservative explanations. It starts with a preamble which mentions feature values that support an outcome that differs from the predicted one ("even though" part), and ends with a resolution which mentions feature values that overcome the values

*Work done while the author was at Monash University.

Proceedings of the 17th International Natural Language Generation Conference, pages 30–46
September 23–27, 2024. ©2024 Association for Computational Linguistics

Table 1: Instance from the Busselton dataset (top part), a concessive-contrastive explanation of the AI's prediction, and a Confidence and %Frequency representation of the uncertainty of this prediction.

At-risk Scenario – ResidentID 83:
You are a 76 year old female whose weight is optimal, does not drink, but smokes 10 cigarettes a day. You also have optimal blood pressure, borderline total and HDL cholesterol, and high triglycerides. But on the upside, you are not diabetic.
Concessive-contrastive explanation (baseline)
Even though you have optimal blood pressure, the AI predicts that you are at risk of a coronary event because you are between 72 and 79 years old and have a high level of triglycerides.
Confidence representation of uncertainty
Based on its past performance, the AI is 90% confident that you are at risk of a coronary event.
%Frequency representation of uncertainty (tailored)
The AI is 90% confident that you are at risk of a coronary event. This confidence is based on the AI's past performance, where out of 200 residents like you (same age, blood pressure and level of triglycerides), it correctly predicted that 180 (90%) were at risk of a coronary event.

in the preamble to yield the predicted outcome.[1]

Our user study considers four research questions:

RQ1: How does the type of uncertainty information (Confidence or %Frequency) affect the likelihood of accepting a prediction, compared to a baseline explanation that omits this information?

RQ2: Which factors affect the likelihood of accepting a prediction when uncertainty information is added to a baseline explanation?

RQ3: How do percentages in Confidence and %Frequency representations and the size of the reference class in %Frequency representations affect the acceptance of a prediction when uncertainty information is added to a baseline explanation?

RQ4: How does uncertainty information affect users' views about four explanatory attributes: completeness, presence of extraneous information, helpfulness to understand the AI's reasoning, and support for decision making? (Hoffman et al., 2018).

This paper is organised as follows. Section 2 presents related work on conveying uncertainty. Section 3 describes our experimental design, followed by our results in Section 4. Section 5 summarises key findings and discusses future work.

2 Related Work

There has been substantial research in communicating the uncertainty associated with predicted outcomes, in particular in healthcare (Freeman, 2019; Simpkin and Armstrong, 2019; Spiegelhalter, 2017; Zipkin et al., 2014). Most of that research considers how to convey probabilities derived from historical population-based statistics, focusing on modality selection (i.e., words, numbers or graphs), and within each modality, on selecting a specific format (e.g., probabilities, percentages or natural frequencies for numeric representations).

Gigerenzer (2003) demonstrated that natural frequencies are more understandable than probabilities, and that it is essential to provide a reference class. But in later review articles, Freeman (2019) and Spiegelhalter (2017) argued that both percentages and frequencies are required. These insights have informed best practice in uncertainty representations shown to patients (e.g., iPrevent).

Research on communicating uncertainty also considered the effect of other factors on users' perceptions of risk, such as communicative intent (Spiegelhalter, 2017), risk type (absolute or relative) (Gigerenzer, 2003), framing of an outcome (positive or negative) (Peters et al., 2011), context (e.g., information about a population at a lower risk) (Lipkus et al., 2001), and users' numeracy (Vromans et al., 2020).

Our work is inspired mainly by the research of Vromans et al. (2020) and Cau et al. (2023). Vromans et al. (2020) studied the interaction between the specificity of the reference class in frequency representations (generic versus tailored) and presentation format (words only versus words and numbers) when communicating population-based statistics. They found that patients deemed tailored risks to be *less accurate and higher* than generic risks when the risks were presented in words only, but not when words were combined with numbers.

Cau et al. (2023) examined the interaction between the correctness of an ML model, the explanation style and the model's confidence in its prediction (expressed as a percentage), e.g., "the AI is 45% confident that the price will increase".

The research described in this paper advances the state-of-the-art in that (1) it compares the influence of Confidence and %Frequency representations of uncertainty on users' acceptance of ML predictions (which differ from population-based historical predictions); (2) it considers the influence of three new factors, viz *predicted outcome*, *size of the reference class* and *level of concern about a coronary event*, on users' acceptance of a prediction, in addition to factors from the literature, viz *confidence percentage* (Cau et al., 2023), *(dis)agreement between AI and user predictions* (similar to (Maruf et al., 2023)) and *users' numeracy* (Vromans et al., 2020);

[1] We eschew varying the generated text, e.g., by using Large Language Models, as this may vitiate the experiments.

Table 2: ***Classes***, *features* and values, Busselton dataset.

Predicted classes:	***Not at risk of CHD, At risk of CHD***			
age (in years):	61	...	...	95
gender:	female			male
weight status:	optimal	underweight	overweight	obese
daily std. drinks:	0	...	...	44
daily cigarettes:	0	...	...	75
blood pressure:	optimal	normal-to-high		high
total cholesterol:	low	normal	borderline	high
HDL cholesterol:	optimal	borderline		low
triglycerides:	low	normal	borderline	high
diabetes:	no			yes

and (3) it examines how uncertainty information in general and our two types of uncertainty representations influence users' views about explanations that convey the predictions of ML models.

3 Experimental Setup

We describe our dataset, the design of our user study,[2] our experiments and our participant cohorts.

3.1 Dataset

Owing to the prevalence and importance of uncertainty information in healthcare, we chose a dataset from the medical domain, specifically, the Busselton dataset (Knuiman et al., 1998). This dataset contains demographic, medical and lifestyle information for a group of people, and information about whether they developed coronary heart disease (CHD) within ten years of the initial data collection, which is encoded as *predicted class* (Table 2). The dataset was pre-processed as described in Appendix A, and we trained a decision tree that predicts whether a person is at risk of CHD (Figure 1, Appendix A).

The explanations we showed in this study were based on the feature values in the path between the root of the decision tree and a prediction (Guidotti et al., 2019; Stepin et al., 2020). However, we manually added feature values, so that all the baseline explanations are of similar length, thereby obviating this experimental variable (according to Lombrozo (2016), explanation length influences users' perceptions).

3.2 User study design

The research questions were addressed by means of two experiments: (1) between subjects – one group of participants saw only Confidence representations, and another group saw only %Frequency representations; and (2) within subject – each participant saw a Confidence representation followed by a %Frequency representation. We conducted both experiments for the following reasons. On one hand, within-subject experiments generally yield stronger results than between-subjects experiments, especially for relatively low numbers of participants. However, the presentation of %Frequency representations after Confidence representations in the within-subject experiment may influence users' opinions about these representations.

Specificity of the %Frequency representation. As mentioned in Section 2, Vromans et al. (2020) found no difference in the effect of generic and tailored frequency representations when words are combined with numbers (they did not investigate numbers alone). Nonetheless, we chose tailored representations, as they are in line with medical practice (e.g., iPrevent).[3]

Independent variables. Our experiment has three intrinsic independent variables, viz *predicted outcome* (at-risk, not-at-risk), *confidence of the AI in its prediction* and *reference class size* (only for %Frequency representations); and three extrinsic independent variables, viz *(dis)agreement between AI and user predictions* ('agree', 'disagree'), and two participant features – *level of concern about CHD* and *numeracy*. The reference class for a tailored %Frequency representation is the number of people in the dataset who share the features of the current patient that were mentioned in the baseline explanation, e.g., blood pressure, age and level of triglycerides for the example in Table 1. The level of concern about CHD was provided by participants ('Not at all concerned': 1 to 'Extremely concerned': 5). Following Vromans et al. (2020), participants' numeracy was assessed using Fagerlin et al.'s (2007) *Subjective Numeracy Scale* (*SNS*), which correlates well with mathematical test measures of objective numeracy. The SNS consists of eight self-assessment numeracy questions (on a 6-point Likert scale; Table 9, Appendix B), and participants' *Subjective Numeracy Score* (*SNSc*) is the average of their answers' scores in the SNS.

We chose two values for confidence {high (90%), low (65%)}, and two values for reference class size {large (200 patients), small (20 patients)} out of 1000 people. For example, a low-confidence prediction for a large reference class talks about "130

[2]We have addressed the recommendations for human evaluation in (Howcroft et al., 2020). The experiment and data are available here.

[3]Our wording for %Frequency representations resembles that used in (Vromans et al., 2020). However, they used frequencies to clarify medical terms, which do not always match lay-people's understanding, e.g., "common (occurs in 10 out of 100 people)".

(65%) out of 200 patients", while a high-confidence prediction for a small reference class talks about "18 (90%) out of 20 patients". It is worth noting that the confidence values and reference class sizes are not based on the dataset; rather, they were chosen to represent distinct categories, and numbers that are easy to process. Specifically, their values were selected so that they are significantly different, but at the same time, we wanted a low confidence to be substantially higher than random chance (in contrast with (Cau et al., 2023), where low-confidence values were between 12-55%). These choices are somewhat arbitrary, and additional research is required to ascertain the effect of other options.

Scenarios. Eight scenarios are required to cover all the combinations of the three intrinsic variables. However, to avoid participant fatigue, our scenarios comprise only four combinations of *predicted outcome*, *confidence percentage* and *reference class size*: {at-risk, high, large}, {at-risk, low, small}, {not-at-risk, low, large} and {not-at-risk, high, small}.

3.3 The experiments

After signing a consent form, participants filled a demographic questionnaire, followed by the body of the survey and a numeracy test.

The body of the survey consists of the following components: an immersive narrative about a retirement village that has purchased an AI to predict whether the residents are at risk of CHD; a brief account of how an AI makes predictions, plus the features and values that were input to the AI to predict CHD (Figure 2, Appendix C); a sample scenario to prepare participants for the survey; and four scenarios presented in random order.

Scenario description. Each scenario began with a narrative like that at the top of Table 1, which includes feature values for a particular patient. Participants were then asked to make an educated guess about the outcome for this patient, and to indicate how sure they were about this guess on a 7-point Likert scale ('Very unsure': 1 to 'Very sure': 7). A 7-point scale is used throughout our experiment, in line with recent best practice recommendations in (van der Lee et al., 2021). After participants entered how sure they were about their guess of the outcome, they were shown the AI's prediction and a concessive-contrastive explanation similar to the explanation in the second segment of Table 1, and they were asked again how likely they were to accept the AI's prediction on a 7-point Likert scale ('Extremely unlikely': 1 to 'Extremely likely': 7).

At this point, the between-subjects and within-subject arms of the experiment diverge, but each arm displays the same four scenarios (in random order). To detect unreliable responses, at the end of each scenario, we asked an attention question about the background information or the explanation.

Between-subjects experiment (*Confidence* and *%Frequency* cohorts). There were two groups in this experiment: one group saw a Confidence uncertainty representation (third segment in Table 1), and the other saw a %Frequency representation (bottom segment in Table 1). After seeing the uncertainty representation, participants in both groups were asked again how likely they were to accept the AI's prediction. Participants in the %Frequency group were also asked what prompted their decision — response options were "number of people similar to me" (reference class), "percentage of correct predictions" (confidence) or both.

Participants in both groups were then asked to rate the initial (baseline) explanation with respect to four explanatory attributes: completeness, presence of irrelevant/misleading/contradictory information, helpfulness for understanding the AI's reasoning, and support in deciding whether to accept the AI's prediction (Hoffman et al., 2018). Next, they were asked whether adding the uncertainty representation (which is different for each group) would yield improvements with respect to each of the explanatory attributes, compared to the initial explanation.

Within-subject experiment (*Combined* cohort). Participants saw a Confidence representation followed by a %Frequency representation — this order was chosen because %Frequency representations subsume Confidence representations. After each representation, participants were asked how likely they were to accept the AI's prediction, which yields two likelihoods of acceptance for the same confidence percentage. Also, like the above %Frequency cohort, participants were asked what prompted their decision (Figure 3, Appendix C).

As for the between-subjects experiment, participants rated the initial explanation with respect to the four explanatory attributes (top panel of Figure 4, Appendix C). But here, they were asked which uncertainty representation they would add to improve the explanation in terms of each attribute — options were Confidence, %Frequency, 'Either' or 'None' (middle panel of Figure 4, Appendix C).

Table 3: Descriptive statistics for the Confidence, %Frequency and Combined groups (number of participants) – two options with the most participants; and Subjective Numeracy Score (on a 6-point Likert scale).

Attribute	**Option**	**Confidence** (29)	**%Frequency** (28)	**Combined** (29)
Gender	Male / Female	19 / 10	16 / 12	13 / 16
Age	25-34 / 35-44	12 / 7	10 / 8	10 / 12
Ethnicity	Caucasian	23	19	21
English proficiency	High	29	27	29
Education	Bachelor / Some college, no degree	12 / 15	14 / 8	20 / 5
ML expertise	Low / Medium	12 / 14	15 / 10	12 / 15
Concern about CHD	Extremely–Moderately / Somewhat–Slightly	15 / 9	13 / 11	7 / 19
Subjective Numeracy Score (SNSc)	Mean (standard deviation)	4.52 (1.08)	4.64 (0.92)	4.58 (0.89)

3.4 Participants

Our survey was implemented in the Qualtrics survey platform, and conducted on Connect (a Cloud Research platform (Litman and Robinson, 2020)). Participants spent 25 minutes on the experiment on average, and were paid \$8-\$10 USD. Their responses were validated based on their answers to the attention questions and the time they spent on each scenario, yielding 86 valid responses out of 101. Table 3 shows descriptive statistics for the retained participants from the three cohorts: Confidence and %Frequency (between subjects) and Combined (within subject). To determine whether the cohorts are similar, we compared the *Subjective Numeracy Scores* of each pair of groups (Wilcoxon rank-sum test). We did not find any statistically significant differences between the scores of the three groups.

4 Results

We report the results for research questions RQ1-RQ4. Statistical significance was adjusted with Holm-Bonferroni correction for multiple comparisons (Holm, 1979), where applicable; results with $0.05 < p\text{-}value < 0.1$ are designated as *trends*.

4.1 RQ1 and RQ2

RQ1 considers the effect of the *type of uncertainty representation* (Confidence or %Frequency) on the likelihood of accepting a prediction, compared to a baseline explanation that omits uncertainty information. We define this dependent variable as
$DiffLikely{=}AcceptLikely_{uncertain}{-}AcceptLikely_{init}$

We use difference in likelihoods, rather than absolute likelihoods, because we observed a high variability between participants' absolute likelihoods of prediction acceptance. A similar observation was made in (van der Bles et al., 2019) with respect to verbal expressions of uncertainty.

RQ2 considers the influence of five of the independent variables described in Section 3.2 on *DiffLikely*: the discrete variables *predicted outcome*, *confidence of the AI in its prediction*, *(dis)agreement between AI and user predictions* and *participants' level of concern about CHD*, and the continuous variable (or covariate) *Subjective Numeracy Score* (*SNSc*). *Reference class size* was excluded from RQ2, because the Confidence group did not receive this information.

We employed ANCOVA to analyse the data for RQ1 and RQ2, as it adjusts for the effects of covariates. However, inspection of the assumptions for ANCOVA revealed that *(dis)agreement between AI and user predictions* and *level of concern about CHD* are not independent of the covariate *SNSc* in the within-subject experiment. Hence, we excluded these two variables from our initial analysis — the results appear in Table 10, Appendix D. Our results show that *SNSc* has no statistically significant impact on *DiffLikely*. We therefore removed this covariate, and reintroduced the excluded variables. ANOVA was employed to re-analyse the data for RQ1 and RQ2, as all the variables are now discrete — the results appear in Table 11, Appendix D.

Table 4 displays the mean (standard deviation) of the likelihood of accepting a prediction after seeing the baseline explanation, and the mean (standard deviation) of the difference after viewing the uncertainty information (*DiffLikely*), broken down according to *type of uncertainty* and the variables that had a statistically significant effect in either experiment: *predicted outcome*, *confidence of the AI in its prediction* and *(dis)agreement between AI and user predictions*. Statistically significant differences are boldfaced, and trends are italicised. The analysis of the effect of the independent variables on the likelihood of accepting predictions after seeing baseline explanations appears in Appendix D.

Type of uncertainty. The leftmost *DiffLikely* column in the top segment of Table 4 shows no statistically significant effect of *type of uncertainty* in the between-subjects experiment ($F(1, 223) = 0.136$, $p\text{-}value = 0.713$), while the rightmost *DiffLikely*

Table 4: Likelihood of accepting predictions after a baseline explanation, and difference after adding uncertainty information (*DiffLikely*), for the between-subjects cohorts (left-hand side) and the within-subject cohort (right-hand side), broken down by *type of uncertainty*, *predicted outcome*, *confidence percentage* and *(dis)agreement between AI and user predictions*: mean (std. dev.); statistically significant differences in means (*p-value* < 0.01) are **boldfaced**, and trends ($0.05 <$ *p-value* < 0.1) are *italicised*.

		Between subjects				Within subject			
		Baseline explanation		***DiffLikely***		**Baseline explanation**		***DiffLikely***	
		Mean	(std. dev.)	Mean	(std. dev.)	Mean	(std. dev.)	Mean	(std. dev.)
Type of uncertainty	Confidence	4.56	(1.75)	0.147	(1.02)	5.21	(1.50)	$-$*0.138*	(1.27)
	Frequency	5.01	(1.66)	0.098	(1.10)	5.21	(1.50)	*0.155*	(1.35)
Predicted outcome	at-risk	**5.57**	(1.25)	*0.009*	(1.01)	**5.90**	(0.93)	$-$**0.207**	(1.25)
	not-at-risk	**3.99**	(1.75)	*0.237*	(1.10)	**4.52**	(1.63)	**0.224**	(1.35)
Confidence percentage	high	4.76	(1.76)	**0.500**	(0.96)	5.05	(1.55)	**0.526**	(1.11)
	low	4.80	(1.68)	$-$**0.254**	(1.02)	5.36	(1.42)	$-$**0.509**	(1.30)
AI predict vs User predict	agree	**5.85**	(0.95)	0.052	(0.94)	**6.05**	(0.85)	$-$**0.266**	(1.20)
	disagree	**4.00**	(1.73)	0.174	(1.14)	**4.24**	(1.48)	**0.324**	(1.38)

column shows a trend in the within-subject experiment ($F(1, 227) = 3.544$, *p-value* $= 0.061$). According to this trend, %Frequency representations increased the likelihood of acceptance, while Confidence representations reduced it.[4]

Predicted outcome. Even though *predicted outcome* is domain specific, we consider this variable, as the notions of good and bad outcomes are general. According to the second segment of Table 4, in both experiments, there is a statistically significant difference between the likelihood of accepting a prediction for the two values of *predicted outcome* {at-risk, not-at-risk}, after seeing the baseline explanations (*p-value* $\ll 0.001$): at-risk predictions have a higher likelihood of acceptance than not-at-risk predictions. The uncertainty information has a statistically significant effect on *DiffLikely* in the within-subject experiment ($F(1, 227) = 7.664$, *p-value* $= 0.006$), but shows only a trend in the between-subjects experiment ($F(1, 223) = 3.023$, *p-value* $= 0.084$), where *DiffLikely* changes mainly for the not-at-risk prediction. After viewing the uncertainty information, the acceptance likelihood of not-at-risk predictions increased in both experiments, and the acceptance likelihood of at-risk predictions decreased in the within-subject experiment.

Confidence percentage. The third segment of Table 4 indicates that *confidence percentage* has a statistically significant influence on *DiffLikely* (between subjects $F(1, 223) = 33.074$, within subject $F(1, 227) = 62.07$, *p-value* $\ll 0.001$ for both). In both experiments, a low prediction confidence led to a reduction in the acceptance likelihood of a prediction, and a high prediction confidence led to an increase. However, recall that a low prediction confidence is 65%, which is substantially higher than random chance. This suggests that people may require a high level of confidence in order to increase their likelihood of accepting an ML prediction.

(Dis)agreement between AI and user predictions. Maruf et al. (2023) studied the influence of (dis)agreement between AI predictions and users' estimates of these predictions on users' views about explanations. Here, we determine whether *(dis)agreement between AI and user predictions* affects prediction acceptance, in particular *DiffLikely*. According to the bottom segment of Table 4, the likelihood of accepting a prediction after seeing the baseline explanations is statistically significantly higher when the predictions of the AI and the user agree than when they disagree (*p-value* $\ll 0.0001$ for both experiments). *(Dis)agreement between AI and user predictions* has no statistically significant effect on *DiffLikely* in the between-subjects experiment ($F(1, 219) = 1.167$, *p-value* $= 0.281$), but has a statistically significant effect in the within-subject experiment ($F(1, 223) = 6.072$, *p-value* $= 0.015$). After seeing the uncertainty information, the acceptance likelihood of AI predictions that agreed/disagreed with the user's decreased/increased. This suggests that uncertainty information moderates users' initial inclination to accept AI predictions on the basis of agreement with their own predictions or lack thereof.

Subjective Numeracy Score (SNSc). People's numeracy has been found to affect their perceptions of risk, especially when uncertainty is presented in different modalities, e.g., numbers versus

[4]The cohorts in the between-subjects experiment correspond to the types of uncertainty, which explains the different mean ratings for accepting a prediction after seeing the baseline explanations (leftmost 'Baseline explanation' column). In contrast, the cohort in the within-subject experiment saw the same baseline explanations independently of *type of uncertainty*, hence the invariant rating (mean 5.21 and standard deviation 1.50, rightmost 'Baseline explanation' column).

Table 5: Likelihood of accepting predictions for the Confidence representation (top segment) – high and low confidence (between-subjects Confidence cohort – left-hand side, and within-subject experiment – right-hand side); and for the %Frequency representation (bottom segment) – high and low confidence and large and small reference class (between-subjects %Frequency cohort – left-hand side, and within-subject experiment – right-hand side): mean (std. dev.); statistically significant differences in means (*p-value* < 0.01) are **boldfaced**.

Confidence representation	**Between subjects**				**Within subject**			
	High Confidence		*Low Confidence*		*High Confidence*		*Low Confidence*	
	Mean	(std. dev.)	Mean	(std. dev.)	Mean	(std. dev.)	Mean	(std. dev.)
Baseline explanation	4.57	(1.92)	4.55	(1.57)	5.05	(1.56)	5.36	(1.42)
DiffLikely	**0.431**	(0.99)	**−0.138**	(0.98)	**0.414**	(1.08)	**−0.690**	(1.22)
%Frequency representation	**Between subjects**				**Within subject**			
	High Confidence		*Low Confidence*		*High Confidence*		*Low Confidence*	
	Mean	(std. dev.)	Mean	(std. dev.)	Mean	(std. dev.)	Mean	(std. dev.)
Baseline explanation	4.96	(1.56)	5.05	(1.76)	5.05	(1.56)	5.36	(1.42)
DiffLikely	**0.571**	(0.93)	**−0.375**	(1.05)	**0.638**	(1.15)	**−0.328**	(1.37)
	Large reference class		*Small reference class*		*Large reference class*		*Small reference class*	
	Mean	(std. dev.)	Mean	(std. dev.)	Mean	(std. dev.)	Mean	(std. dev.)
Baseline explanation	4.86	(1.64)	5.16	(1.67)	5.22	(1.43)	5.19	(1.57)
DiffLikely	**0.429**	(1.06)	**−0.232**	(1.04)	0.259	(1.21)	0.052	(1.48)

words (Spiegelhalter, 2017; Vromans et al., 2020). However, *SNSc* has no statistically significant impact on *DiffLikely* in our experiments (between-subjects $F(1, 223) = 0.316$, *p-value* $= 0.574$; within-subject $F(1, 227) = 2.137$, *p-value* $= 0.145$). This indicates that users' numeracy, at the levels exhibited by our participants, is not relevant when comparing simple numeric representations.

Participants' concern about CHD. This variable was considered because people who are concerned about CHD may be biased towards a particular outcome. However, participants' *concern about CHD* has no statistically significant impact on the likelihood of accepting a prediction or on *DiffLikely* in both experiments (between-subjects $F(4, 219) = 0.243$, *p-value* $= 0.913$; within-subject $F(4, 223) = 1.743$, *p-value* $= 0.142$).

Finding 1 *The* confidence percentage *in an uncertainty representation has the strongest influence on* DiffLikely*– high values increase acceptance likelihood and low values decrease it. The* predicted outcome *and* (dis)agreement between AI and user predictions *have some influence on* DiffLikely.

4.2 RQ3

RQ3 examines the influence of *confidence percentage* (Confidence and %Frequency representations) and *reference class size* (%Frequency representation) on the likelihood of accepting a prediction, compared to a baseline explanation that omits uncertainty information (*DiffLikely*).

We employed ANOVA to analyse the data for RQ3 — the results appear in Table 14, Appendix D. Table 5 displays the mean (standard deviation) of the likelihood of accepting a prediction and the mean (standard deviation) of the difference after viewing the uncertainty information (*DiffLikely*) for the Confidence and %Frequency representations, for both cohorts of the between-subjects experiment (left-hand side) and for the within-subject experiment (right-hand side). The results for *confidence percentage* are consistent with the results in Table 4 — a high percentage (90%) increases acceptance likelihood, and a low percentage (65%) decreases it (statistically significant, *p-value* < 0.01 for both experiments). Looking at *reference class size*, a large class (200) led to an increase in acceptance likelihood, and a small class (20) led to a decrease, for the %Frequency cohort in the between-subjects experiment (statistically significant, *p-value* < 0.001). However, this effect was not observed in the within-subject experiment, where the %Frequency representation followed the Confidence representation. Rather, an interaction effect was observed (trend; Table 14, Appendix D); Tukey's HSD test for the interaction indicates that a low *confidence percentage* for a small reference class led to a lower *DiffLikely* (mean ≤ 0) than a high *confidence percentage* regardless of *reference class size* (mean > 0.5) (statistically significant, *p-value* < 0.01).

Finding 2 *Finding 1 with respect to* confidence percentage *was corroborated for both types of uncertainty representation.* Reference class size *also influences* DiffLikely*, but the effects differ for the two experimental conditions.*

4.3 RQ4

RQ4 considers the effect of adding uncertainty information to a baseline explanation on users' opinions about four explanatory attributes: complete-

Table 6: Participant views about adding uncertainty information in terms of four explanatory attributes – one-proportion Z-test applied to Confidence and %Frequency cohorts of the between-subjects experiment together: number of 'Yes' replies (total number of replies), χ^2 statistic, *p-value* after Holm-Bonferroni correction; statistically significant results are **boldfaced**.

Attribute	Uncertainty (228)	χ^2 statistic	adjusted *p-value*
+Complete	188	94.776	**1.76E-15**
+Relevant, −Misleading, ...	161	37.934	**3.66E-09**
+Helpful for understanding	181	77.583	**1.76E-15**
+Enable better decisions	192	105.37	**1.76E-15**

ness, presence of irrelevant/misleading/contradictory information, helpfulness for understanding the AI's reasoning, and support in making a decision (Hoffman et al., 2018).

First, we examine overall effects, in terms of improving a baseline explanation, as reflected by the total number of 'Yes' replies to whether the uncertainty information would make the explanation (1) more complete, (2) more relevant, less misleading or less contradictory, (3) more helpful for understanding the AI's reasoning, and whether this information would (4) enable participants to make a better decision about accepting the AI's prediction (Section 3.3). Table 6 displays the results of a one-proportion Z-test applied to the Confidence and %Frequency cohorts together (between-subjects experiment)[5] — the second column shows the number of 'Yes' replies (out of 228 responses). As seen in Table 6, most participants thought that uncertainty information improves baseline explanations in terms of the four explanatory attributes (statistically significant, *p-value* $\ll 0.001$).

Next, we examine users' views about adding a Confidence or a %Frequency representation to baseline explanations. For the between-subjects experiment, we counted the 'Yes' replies to the above questions; and for the within-subject experiment, we counted the number of times the Confidence representation or the %Frequency representation was selected when asked which of these representations would improve the four explanatory attributes listed above (middle panel of Figure 4, Appendix C) — users chose very few 'Either' and 'None' options, which we excluded from our analysis. The results of the two-proportions Z-test applied to the cohorts of the between-subjects experiment appear on the left-hand side of Table 7, and the results of the one-proportion Z-test applied to the cohort of the within-subject experiment appear on the right-hand side. The Confidence and %Frequency columns show the number of 'Yes' replies for the corresponding representations.

As seen in Table 7 (left-hand side), no statistically significant differences were found when comparing the representations seen by the Confidence cohort with those seen by the %Frequency cohort — there was only a trend whereby %Frequency representations were deemed more complete than Confidence representations. These results are not surprising, as each cohort saw only one uncertainty representation, which was deemed to be a valuable addition to a baseline explanation (Table 6). However, when participants in the within-subject experiment directly compared the two types of uncertainty representation, the %Frequency representation was deemed better than the Confidence representation with respect to all explanatory attributes (statistically significant, *p-value* $\ll 0.001$).

Finding 3 *Both types of uncertainty representations are deemed to add value to baseline explanations in terms of the four explanatory attributes, with %Frequency representations being considered better than Confidence representations.*

5 Conclusion

This research focuses on the influence of uncertainty information on the acceptance of predictions made by ML models. Our main contributions are: (1) determining factors that influence users' acceptance of these predictions; and (2) comparing the influence of Confidence and %Frequency uncertainty representations on users' views about explanations.

Our results show that when uncertainty information is incorporated in an explanation of the prediction of an ML model, users' likelihood of accepting the prediction is influenced by the model's *confidence percentage* — high percentages (90%) increase the likelihood of acceptance (compared to a baseline explanation without uncertainty information), while low percentages (65%) decrease this likelihood. This finding suggests that people may require a high level of confidence in order to increase their likelihood of accepting an ML prediction. *Reference class size* influenced the likelihood of prediction acceptance, with a large class (200 out of 1000) increasing this likelihood and a small class (20 out of 1000) decreasing it (for the %Frequency cohort).

Predicted outcome and *(dis)agreement between*

[5] The within-subject experiment was exluded, as its questions differ from those in the between-subjects experiment.

Table 7: Participant views about adding a Confidence versus a %Frequency representation in terms of four explanatory attributes – two-proportions Z-test for the between-subjects experiment, and one-proportion Z-test for the within-subject experiment: number of Confidence and %Frequency replies (total number of replies), χ^2 statistic, *p-value* after Holm-Bonferroni correction; statistically significant results are **boldfaced**, and trends ($0.05 <$ *p-value* < 0.1) are *italicised*.

	Between subjects				Within subject			
Attribute	Confidence (116)	%Frequency (112)	χ^2 statistic	adjusted *p-value*	Confidence	%Frequency (116)	χ^2 statistic	adjusted *p-value*
+Complete	88	100	6.200	*0.0511*	14	90	54.087	**3.84E-13**
+Relevant, −Misleading, ...	87	74	1.780	0.3642	8	83	60.176	**2.60E-14**
+Helpful for understanding	89	92	0.718	0.3968	17	85	44.010	**3.27E-11**
+Enable better decisions	92	100	3.547	0.1789	10	90	62.410	**1.12E-14**

AI and user predictions influenced prediction acceptance for baseline explanations (without uncertainty information), with participants being more likely to accept at-risk predictions than not-at-risk predictions, and ML model predictions that agreed with their own predictions than ML model predictions that disagreed. However, uncertainty information moderated these effects, increasing the likelihood of accepting the less-acceptable predictions and decreasing the likelihood of accepting the more-acceptable ones.

Users deemed explanations that include uncertainty information to be better, in terms of the four explanatory attributes, than baseline explanations that omit uncertainty information. When the two types of uncertainty representations were seen separately, users deemed them to be similar in terms of their effect on the four explanatory attributes. However, when seen together, %Frequency representations were deemed to be better than Confidence representations by the vast majority of users.

Limitations and future work

User study. We could not recruit real users who were personally engaged with the CHD scenario, and employed crowd-workers instead. This is a common limitation when evaluating NLG systems, which we tried to mitigate by having a narrative immersion at the start of our experiment.

Uncertainty representation. Our study considers two numerical methods for representing uncertainty, viz Confidence and %Frequency. In the future, it is worth investigating additional modalities, such as words and graphs, e.g., charts and icon arrays (Spiegelhalter, 2017; Zipkin et al., 2014), as well as combinations of modalities.

Confidence percentage and reference class size. As mentioned in Section 3.2, our choices for *confidence percentage* and *reference class size* are somewhat arbitrary. Additional levels of confidence and reference class sizes should be investigated, as well as the interaction between these two variables.

Additional factors and interactions between them. Our experiment considers the effect of six independent variables on prediction acceptance, viz *type of uncertainty*, *predicted outcome*, *confidence of the AI*, *(dis)agreement between AI and user predictions*, *concern about CHD* and *Subjective Numeracy Score*. However, as seen in Section 2, there are many more factors examined in the literature, e.g., communicative intent (Spiegelhalter, 2017), risk type (Gigerenzer, 2003), framing of an outcome (Peters et al., 2011) and context (Lipkus et al., 2001). Combinations of these factors should be investigated in the future.

In addition, according to Lombrozo (2016), explanation length influences users' perceptions. To obviate the potential effect of the length difference between %Frequency and Confidence representations on their relative ratings, content would have to be added to the latter. However, this would influence other explanatory attributes of this representation, e.g., completeness and relevance.

Aleatoric and epistemic uncertainty. The uncertainty of ML predictions comes from two main sources (Hüllermeier and Waegeman, 2021): *aleatoric* (due to chance) and *epistemic* (due to insufficient information in the prediction models themselves) — a distinction that is critical in decision making (Senge et al., 2014). In the future, we will derive these types of uncertainty for the predictions made by ML models, and investigate how to communicate them.

Acknowledgments

This research was supported in part by grant DP190100006 from the Australian Research Council. Ethics approval for the user study was obtained from Monash University Human Research Ethics Committee (ID-24208).

References

E. Begoli, T. Bhattacharya, and D. Kusnezov. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23.

O. Biran and K. McKeown. 2017. Human-centric justification of Machine Learning predictions. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI 2017, pages 1461–1467, Melbourne, Australia.

F.M. Cau, H. Hauptmann, L.D. Spano, and N. Tintarev. 2023. Supporting high-uncertainty decisions through AI and logic-style explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 251–263, Sydney, Australia.

A. Fagerlin, B.J. Zikmund-Fisher, P.A. Ubel, A. Jankovic, H.A. Derry, and D.M. Smith. 2007. Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making*, pages 672–680.

E. Frank, M.A. Hall, and I.H. Witten. 2016. *The WEKA Workbench – Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4 edition. Morgan Kaufmann Publishers, San Francisco, California.

A.L.J. Freeman. 2019. How to communicate evidence to patients. *Drug and Therapeutics Bulletin*, 57(8):119–124.

G. Gigerenzer. 2003. *Reckoning with risk: Learning to live with uncertainty*. Penguin Books Ltd.

R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23.

R.R. Hoffman, S.T. Mueller, G. Klein, and J. Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

S. Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

D.M. Howcroft, A. Belz, M.A. Clinciu, D. Gkatzia, S.A. Hasan, S. Mahamood, S. Mille, E. Van Miltenburg, S. Santhanam, and V. Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, INLG 2020, pages 169–182, Dublin, Ireland.

E. Hüllermeier and W. Waegeman. 2021. Aleatoric and epistemic uncertainty in Machine Learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

M.W. Knuiman, H.T. Vu, and H.C. Bartholomew. 1998. Multivariate risk estimation for coronary heart disease: the Busselton health study. *Australian & New Zealand Journal of Public Health*, 22:747–753.

I.M. Lipkus, M. Biradavolu, K. Fenn, P. Keller, and B.K. Rimer. 2001. Informing women about their breast cancer risks: truth and consequences. *Health communication*, 13(2):205–226.

L. Litman and J. Robinson. 2020. *Conducting online research on Amazon Mechanical Turk and beyond*. Sage Publications.

T. Lombrozo. 2016. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10):748–759.

S. Maruf, I. Zukerman, E. Reiter, and G. Haffari. 2023. Influence of context on users' views about explanations for decision-tree predictions. *Computer Speech & Language*, 81:101483.

S. Maruf, I. Zukerman, X. Situ, C. Paris, and G. Haffari. 2024. Generating simple, conservative and unifying explanations for logistic regression models. In *Proceedings of the 17th International Conference on Natural Language Generation*, INLG 2024, Tokyo, Japan.

T. Miller. 2019. Explanation in Artificial Intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

E. Peters, Hart P.S., and L. Fraenkel. 2011. Informing patients: the influence of numeracy, framing, and format of side effect information on risk perceptions. *Medical Decision Making*, 31(3):432–436.

J.R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, California.

R. Senge, S. Bösner, K. Dembczynski, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, and E. Hüllermeier. 2014. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29.

A.L. Simpkin and K.A. Armstrong. 2019. Communicating uncertainty: a narrative review and framework for future research. *Journal of General Internal Medicine*, 34:2586–2591.

D. Spiegelhalter. 2017. Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, 4(1):31–60.

I. Stepin, J.M. Alonso, A. Catala, and M. Pereira. 2020. Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers. In *Proceedings of the IEEE World Congress on Computational Intelligence*, WCCI, pages 1–8, Glasgow, Scotland.

L.J. Trevena, B.J. Zikmund-Fisher, A. Edwards, W. Gaissmaier, M. Galesic, P.K.J. Han, J. King, M.L. Lawson, S.K. Linder, I. Lipkus, E. Ozanne, E. Peters, D. Timmermans, and S. Woloshin. 2013. Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers. *BMC Medical Informatics and Decision Making*, 13(2).

A.M. van der Bles, S. van der Linden, A.L.J. Freeman, J. Mitchell, A.B. Galvao, L. Zaval, and D.J. Spiegelhalter. 2019. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6:181870.

C. van der Lee, A. Gatt, E. van Miltenburg, and E.J. Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:1–24.

R.D. Vromans, S.C. Pauws, N. Bol, L.V. van de Poll-Franse, and E.J. Krahmer. 2020. Communicating tailored risk information of cancer treatment side effects: Only words or also numbers? *BMC Medical Informatics and Decision Making*, 20:277.

D.A. Zipkin, C.A. Umscheid, N.L. Keating, E. Allen, K. Aung, R. Beyth, S. Kaatz, D.M. Mann, J.B. Sussman, D. Korenstein, C. Schardt, A. Nagi, R. Sloane, and D.A. Feldstein. 2014. Evidence-based risk communication: a systematic review. *Annals of internal medicine*, 161(4):270–280.

A The Busselton dataset

We employed a version of the dataset that was pre-processed by Maruf et al. (2023). This dataset has two classes: whether someone will experience a CHD event or not within ten years of the initial data collection. We recoded these classes as *at risk of a coronary event* and *not at risk of a coronary event* respectively. In addition, in order to fit in with our narrative about a retirement village (Figure 2, Appendix C), we removed patients under the age of 61.

The dataset was split into 80% training and 20% test sets using proportional sampling (we did not cross-validate, as average classifier accuracy is tangential to this research). Table 8 shows the two classes in our evaluation dataset, and the breakdown of the training/test sets. We employed the J48 classifier (Quinlan, 1993) in WEKA (Frank et al., 2016) to learn a decision tree — the resultant decision tree has 24 nodes (Figure 1), and achieved an accuracy of 78.4% and 68.8% on the training and test set respectively.

Table 8: Breakdown of classes for the training and test sets, Busselton dataset (patients over 60 years old).

Partition	*Not at risk*	*At risk*	**Total**
Training	459	166	625
Testing	99	46	145
Total	558	212	770

```
Age <= 69.1: No
Age > 69.1
|   Age <= 78.7
|   |   Triglyce-cat = low: No
|   |   Triglyce-cat = desirable
|   |   |   Smoke_amt <= 11: No
|   |   |   Smoke_amt > 11
|   |   |   |   Age <= 73.1: No
|   |   |   |   Age > 73.1: Yes
|   |   Triglyce-cat = borderline
|   |   |   BP-cat = Optimal: Yes
|   |   |   BP-cat = Normal-to-High
|   |   |   |   Weight-cat = underweight: No
|   |   |   |   Weight-cat = normal: No
|   |   |   |   Weight-cat = overweight
|   |   |   |   |   Sex = F: No
|   |   |   |   |   Sex = M: Yes
|   |   |   |   Weight-cat = obese: Yes
|   |   |   BP-cat = Mild-Mod-Hyp: Yes
|   |   Triglyce-cat = high
|   |   |   Age <= 71.7: No
|   |   |   Age > 71.7: Yes
|   Age > 78.7: Yes

Number of Leaves  :   15
Size of the tree :   24
```

Figure 1: Pruned decision tree, Busselton dataset (patients over 60 years old), recoded classes and features.

B Subjective numeracy test

Table 9 displays the questions in Fagerlin et al.'s (2007) Subjective Numeracy Scale. All the answers are on a 6-point Likert scale, where 1 indicates a low preference for numerical information or a low proficiency in processing it, and 6 indicates a high preference or proficiency.

Table 9: Questions in the Subjective Numeracy Scale – answers are on a 6-point Likert scale.

1. Please indicate how good you are at each of the tasks listed below:
   - Working with fractions
   - Working with percentages
   - Calculating a 15% tip
   - Figuring out the price of a shirt that is 25% off
2. When reading the newspaper, how helpful do you find tables and graphs that are part of a story?
3. When people tell you the chance of something happening, do you prefer that they use words ("it rarely happens") or numbers ("there's a 1% chance")?
4. When you hear a weather forecast, do you prefer predictions using percentages (e.g., "there will be a 20% chance of rain today") or predictions using only words (e.g., "there is a small chance of rain today")?
5. How often do you find numerical information useful?

C Screenshots from the experiment

Background

Artificial Intelligence (AI) systems are used to generate predictions in different domains, such as health, finance and industry. For example, the AI system used in this study predicts whether a person is at risk of a coronary event or not.

We are developing a computer system that automatically generates explanations for the predictions made by this AI system.

The aim of this study is to find out how good are these explanations, and whether presenting the AI's confidence information changes your perceptions about the explanation. We would appreciate your help in making this determination.

The domain

A seniors village has purchased a state-of-the-art AI system that predicts whether a particular resident is at risk of a coronary event or not. To make these predictions, the AI system takes into account different factors in a resident's profile, such as their age and cholesterol level (see the table below).

AI systems make predictions based on trends and patterns they identify in the data. Our AI system built its prediction model from data obtained from **1000 residents** of the seniors village. These data consist of **ten** personal, lifestyle and medical factors of previous residents. The same factors are then obtained from new residents to predict whether they are at risk of a coronary event. These factors and their possible values are listed below in shades of red (more prone to a coronary event) and blue (less prone to a coronary event). These colours will be used in the situations you will see in the survey.

Personal and Lifestyle Factors	**Possible values**			
Age	61			95
Gender	Female			Male
Weight status based on Body Mass Index (BMI)	Optimal	Underweight	Overweight	Obese
Daily alcohol intake (standard drinks)	0			44
Daily cigarette consumption	0			40
Medical Factors	**Possible values**			
Blood pressure	Optimal		Normal-to-High	High
Total cholesterol	Low	Normal	Borderline	High
HDL cholesterol	Optimal		Borderline	Low
Triglycerides	Low	Normal	Borderline	High
Diabetes	No			Yes

Notes:

This dataset comes from the 1970s, and at that time people only had the option to choose from two genders.

- If you hover the mouse over the names of medical factors, you will see a brief description for each of them.
- If you hover the mouse over the values of *weight status*, *blood pressure*, *total cholesterol*, *HDL cholesterol* and *triglycerides*, you will see the range for each value.

Important: AI systems may determine that factors that are relevant to some situations are not relevant to other situations. For example, if a person is more than 70 years old, their weight status may influence the AI system's prediction about their risk of a coronary event. In contrast, the AI system may not need to consider the weight status of people under 70 years of age.

Disclaimer:

The AI system developed for this study is a Machine Learning model that predicts the risk of a coronary event from data pertaining to **a particular population**. Although this system considers relevant factors, it may decide to ignore factors that don't improve the system's prediction accuracy **for this population** --- this decision is based on statistical considerations, **not on medical reasons**.

Figure 2: Background information; narrative immersion for the survey; description of the reasoning of AI systems; features and feature values of a patient; notes and disclaimer.

ResidentID 83:

Assume that you are a *76 year old* ***female*** whose ***weight is optimal***, who ***does not drink***, but *smokes 10 cigarettes a day*. You also have *optimal* blood pressure, *high total cholesterol*, *low HDL cholesterol* and *high triglycerides*. But on the upside, you are ***not diabetic***.

Notes:

- If you hover the mouse over the *underlined values*, you will see their range.
- Click here to look at the glossary of all the factors and their possible values for a patient's profile.

The AI system will predict whether you are at *risk of a coronary event* or *not*.

Before we proceed, please indicate your expectation regarding your risk of a coronary event based on your profile.

At risk of a coronary event	**Not** at risk of a coronary event	Can't decide
○	○	○

How sure are you about your expectation regarding your risk of a coronary event?

Very unsure	Moderately unsure	Slightly unsure	Neither sure nor unsure	Slightly sure	Moderately sure	Very sure
○	○	○	○	○	○	○

Based on your profile, our AI system predicts that you are at **risk of a coronary event**. Recall that the AI built its prediction model from data obtained from **1000 residents**.

Please read the following explanation carefully before you rate it.

Even though you have

- *optimal* blood pressure,

the AI predicts that you are at **risk of a coronary event** because you

- are *between 72 and 79 years old* and
- have a *high level of triglycerides*.

Based on this explanation, how likely are you to accept the AI's prediction?

Extremely unlikely	Moderately unlikely	Slightly unlikely	Neither likely nor unlikely	Slightly likely	Moderately likely	Extremely likely
○	○	○	○	○	○	○

We will now show you **two ways of communicating the AI's confidence in its prediction.** We would like to see how each of them affects your acceptance of the prediction.

Option 1:

Based on its past performance, the AI is 90% confident that you are at **risk of a coronary event**.

In light of the above explanation and this confidence information, how likely are you to accept the AI's prediction?

Extremely unlikely	Moderately unlikely	Slightly unlikely	Neither likely nor unlikely	Slightly likely	Moderately likely	Extremely likely
○	○	○	○	○	○	○

Option 2:

The AI is 90% confident that you are at **risk of a coronary event**. This confidence is based on the AI's past performance, where out of 200 residents like you (same *age*, *blood pressure* and *level of triglycerides*), it correctly predicted that 180 (90%) were at **risk of a coronary event**.

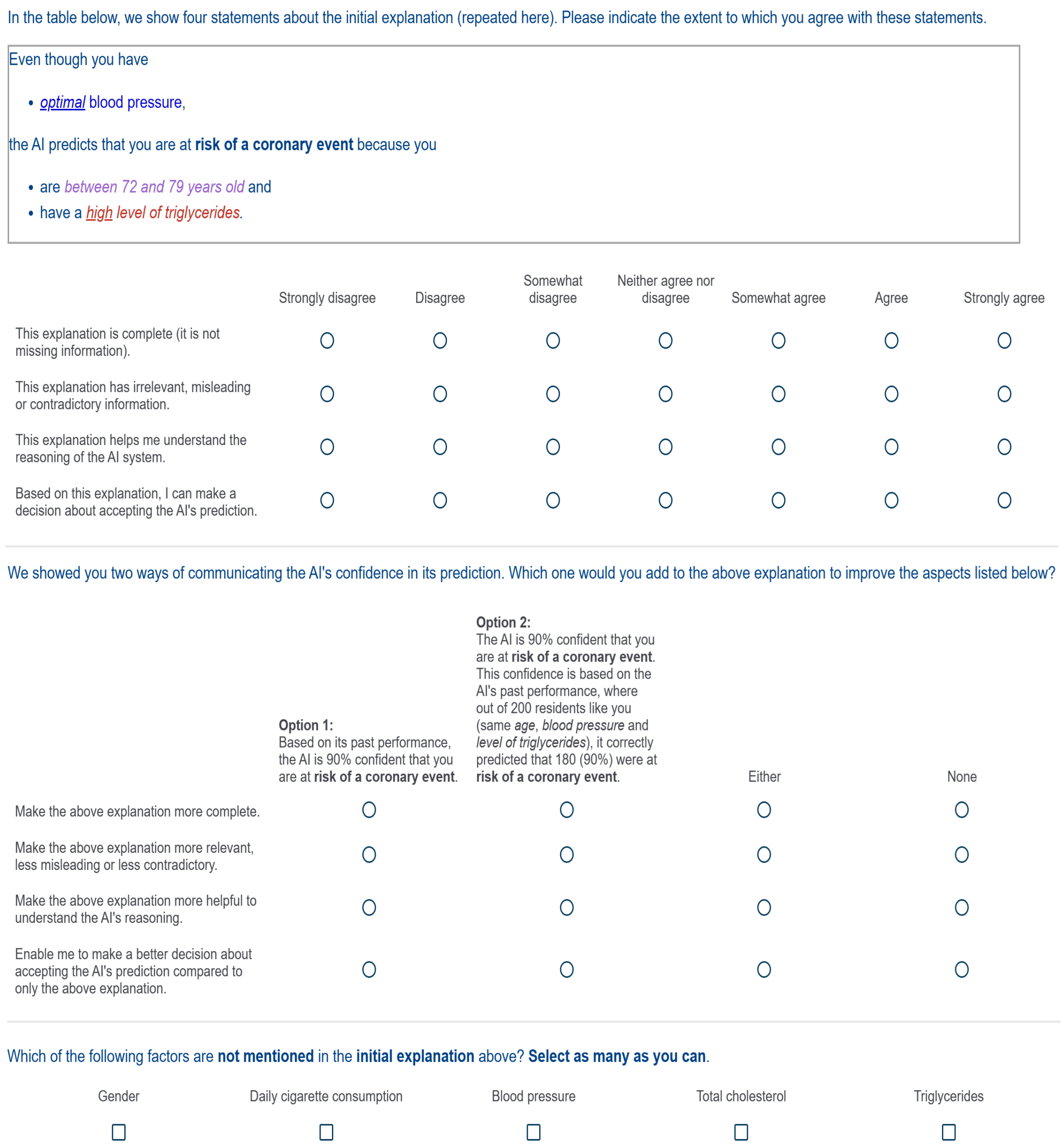
Based on the above explanation and this confidence information, how likely are you to accept the AI's prediction?

Extremely unlikely	Moderately unlikely	Slightly unlikely	Neither likely nor unlikely	Slightly likely	Moderately likely	Extremely likely
○	○	○	○	○	○	○

What prompted your decision?

Number of residents similar to me (200)	Percentage of correct predictions (90%)	Both
○	○	○

Figure 3: First page of the survey for the within-subject group: request for a participant's prediction and their certainty about it; the AI's prediction, associated explanation and request to rate it; two options for communicating uncertainty: Confidence and %Frequency; request for the main factors that prompted the participant's decision.



In the table below, we show four statements about the initial explanation (repeated here). Please indicate the extent to which you agree with these statements.

Even though you have

- *optimal* blood pressure,

the AI predicts that you are at **risk of a coronary event** because you

- are *between 72 and 79 years old* and
- have a *high level of triglycerides*.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
This explanation is complete (it is not missing information).	○	○	○	○	○	○	○
This explanation has irrelevant, misleading or contradictory information.	○	○	○	○	○	○	○
This explanation helps me understand the reasoning of the AI system.	○	○	○	○	○	○	○
Based on this explanation, I can make a decision about accepting the AI's prediction.	○	○	○	○	○	○	○

We showed you two ways of communicating the AI's confidence in its prediction. Which one would you add to the above explanation to improve the aspects listed below?

	Option 1: Based on its past performance, the AI is 90% confident that you are at **risk of a coronary event**.	**Option 2:** The AI is 90% confident that you are at **risk of a coronary event**. This confidence is based on the AI's past performance, where out of 200 residents like you (same *age*, *blood pressure* and *level of triglycerides*), it correctly predicted that 180 (90%) were at **risk of a coronary event**.	Either	None
Make the above explanation more complete.	○	○	○	○
Make the above explanation more relevant, less misleading or less contradictory.	○	○	○	○
Make the above explanation more helpful to understand the AI's reasoning.	○	○	○	○
Enable me to make a better decision about accepting the AI's prediction compared to only the above explanation.	○	○	○	○

Which of the following factors are **not mentioned** in the **initial explanation** above? **Select as many as you can**.

Gender	Daily cigarette consumption	Blood pressure	Total cholesterol	Triglycerides
☐	☐	☐	☐	☐

Figure 4: Second page of the survey for the within-subject group: request to rate the initial explanation on four explanatory attributes; request to rate the influence of the two types of uncertainty representations on these attributes; attention question.

D Experimental results

Table 10 displays the results of the ANCOVA test for research questions RQ1 and RQ2 for the independent variables *uncertainty type*, *predicted outcome*, *confidence percentage* and *SNSc*; Table 11 displays the results of the ANOVA test for research questions RQ1 and RQ2 for the independent variables *uncertainty type*, *predicted outcome*, *confidence percentage*, *(dis)agreement between AI and user predictions* and *level of concern about CHD*.

Table 12 shows the results of the ANCOVA test for accepting a predicted outcome for the independent variables *predicted outcome* and *SNSc* after seeing the baseline explanation; and Table 13 shows the results of the ANOVA test for accepting a predicted outcome for the independent variables *predicted outcome*, *(dis)agreement between AI and user predictions* and *level of concern about CHD* after seeing the baseline explanation. The independent variables *type of uncertainty* and *confidence percentage* were excluded from these analyses, as uncertainty is not part of the baseline explanations.

Table 14 shows the ANOVA results for research question RQ3.

Table 10: ANCOVA results for RQ1 and RQ2 – *uncertainty type*, *predicted outcome*, *confidence percentage* and *SNSc* (between-subjects and within-subject experiments); statistically significant results are **boldfaced**, and trends ($0.05 < \textit{p-value} < 0.1$) are *italicised*.

	Between subjects					**Within subject**				
	DF	Sum of squares	Mean square	F-value	*p-value*	DF	Sum of squares	Mean square	F-value	*p-value*
Uncertainty type	1	0.13	0.13	0.136	0.713	1	4.98	4.98	3.544	*0.061*
Predicted outcome	1	2.96	2.96	3.023	*0.084*	1	10.78	10.78	7.664	**0.006**
Confidence percentage	1	32.44	32.44	33.074	**2.90E-08**	1	62.07	62.07	44.147	**2.23E-10**
SNSc	1	0.31	0.31	0.316	0.574	1	3.00	3.00	2.137	0.145

Table 11: ANOVA results for RQ1 and RQ2 – *uncertainty type*, *predicted outcome*, *confidence percentage*, *(dis)agreement between AI and user predictions*, and *participants' concern about CHD* (between-subjects and within-subject experiments); statistically significant results are **boldfaced**, and trends ($0.05 < \textit{p-value} < 0.1$) are *italicised*.

	Between subjects					**Within subject**				
	DF	Sum of squares	Mean square	F-value	*p-value*	DF	Sum of squares	Mean square	F-value	*p-value*
Uncertainty type	1	0.13	0.13	0.134	0.714	1	4.98	4.98	3.651	*0.057*
Predicted outcome	1	2.96	2.96	2.994	*0.084*	1	10.78	10.78	7.895	**0.005**
Confidence percentage	1	32.44	32.44	32.752	**3.42E-08**	1	62.07	62.07	45.478	**1.31E-10**
AIPredict-vs-UserPredict	1	1.16	1.16	1.167	0.281	1	8.29	8.29	6.072	**0.015**
Concern about CHD	4	0.96	0.24	0.243	0.913	4	9.51	2.38	1.743	0.142
Residuals	219	216.9	0.99			223	304.35	1.36		

Table 12: ANCOVA results for likelihood of prediction acceptance after baseline explanations – *predicted outcome* and *SNSc* (between-subjects and within-subject experiments); statistically significant results are **boldfaced**.

	Between subjects					**Within subject**				
	DF	Sum of squares	Mean square	F-value	*p-value*	DF	Sum of squares	Mean square	F-value	*p-value*
Predicted outcome	1	142.11	142.11	61.46	**1.79E-13**	1	55.17	55.17	31.082	**1.71E-07**
SNSc	1	4.70	4.70	2.032	0.155	1	1.28	1.28	0.721	0.397
Residuals	225	520.20	2.31			113	200.58	1.78		

Table 13: ANOVA results for likelihood of prediction acceptance after baseline explanations – *predicted outcome*, *(dis)agreement between AI and user predictions*, and *participants' concern about CHD* (between-subjects and within-subject experiments); statistically significant results are **boldfaced**.

	Between subjects					**Within subject**				
	DF	Sum of squares	Mean square	F-value	*p-value*	DF	Sum of squares	Mean square	F-value	*p-value*
Predicted outcome	1	142.11	142.11	71.712	**3.49E-15**	1	55.17	55.17	38.72	**9.29E-09**
AIPredict-vs-UserPredict	1	75.61	75.61	38.154	**3.10E-09**	1	45.23	45.23	31.74	**1.39E-07**
Concern about CHD	4	11.40	2.85	1.437	0.223	4	1.31	0.33	0.23	0.921
Residuals	221	437.90	1.98			109	155.32	1.42		

Table 14: ANOVA results for RQ3 – Confidence representation (within-subject experiment and Confidence cohort, between-subjects experiment), and %Frequency representation (within-subject experiment and %Frequency cohort, between-subjects experiment); statistically significant results are **boldfaced**, and trends ($0.05 < p\text{-}value < 0.1$) are *italicised*.

Confidence representation	**Between subjects**					**Within subject**				
	DF	Sum of squares	Mean square	F-value	*p-value*	DF	Sum of squares	Mean square	F-value	*p-value*
Confidence percentage	1	9.39	9.39	9.631	**0.002**	1	35.31	35.31	26.75	**1.00E-06**
Residuals	114	111.12	0.98			114	150.48	1.32		
%Frequency representation	**Between subjects**					**Within subject**				
	DF	Sum of squares	Mean square	F-value	*p-value*	DF	Sum of squares	Mean square	F-value	*p-value*
Confidence percentage	1	25.08	25.08	28.04	**6.31E-07**	1	27.03	27.03	17.29	**6.30E-05**
Reference class size	1	12.22	12.22	13.66	**3.45E-04**	1	1.24	1.24	0.79	0.375
[*Confidence* : *Ref. class size*]	1	0.01	0.009	0.01	0.921	1	5.83	5.83	3.73	*0.056*
Residuals	108	96.61	0.895			112	175.1	1.563		