Info Vis

Original Article

Information Visualization
1–11

DOI: 10.1177/1473871618807121
journals.sagepub.com/home/ivi
SAGE

Graphical uncertainty representations for ensemble predictions

Alexander Toet[1], Jan BF van Erp[1,2], Erik M Boertjes[3] and Stef van Buuren[4,5]

Abstract

We investigated how different graphical representations convey the underlying uncertainty distribution in ensemble predictions. In ensemble predictions, a set of forecasts is produced, indicating the range of possible future states. Adopting a use case from life sciences, we asked non-expert participants to compare ensemble predictions of the growth distribution of individual children to that of the normal population. For each individual child, the historical growth data of a set of 20 of its best matching peers was adopted as the ensemble prediction of the child's growth curve. The ensemble growth predictions were plotted in seven different graphical formats (an ensemble plot, depicting all 20 forecasts and six summary representations, depicting the peer group mean and standard deviation). These graphs were plotted on a population chart with a given mean and variance. For comparison, we included a representation showing only the initial part of the growth curve without any future predictions. For 3 months old children that were measured at four occasions since birth, participants predicted their length at the age of 2 years. They compared their prediction to either (1) the population mean or to (2) a "normal" population range (the mean ± 2(standard deviation)). Our results show that the interpretation of a given uncertainty visualization depends on its visual characteristics, on the type of estimate required and on the user's numeracy. Numeracy correlates negatively with bias (mean response error) magnitude (i.e. people with lower numeracy show larger response bias). Compared to the summary plots that yield a substantial overestimation of probabilities, and the No-prediction representation that results in quite variable predictions, the Ensemble representation consistently shows a lower probability estimation, resulting in the smallest overall response bias. The current results suggest that an Ensemble or "spaghetti plot" representation may be the best choice for communicating the uncertainty in ensemble predictions to non-expert users.

Keywords

Uncertainty visualization, growth charts, user evaluation study

Introduction

Multivariate time-series forecasting has important applications in many different domains such as economy, health care, meteorology, seismology, and astronomy. Simulation models are a primary tool in the generation of predictions. When the parameter space for a simulation is too large or too complex to be fully explored, an ensemble of runs can give an impression of the potential range of model outcomes. A visualization of such an ensemble of predictions should properly convey the variability between its members. It is

[1]Department of Perceptual and Cognitive Systems, TNO, Soesterberg, The Netherlands
[2]Research Group Human Media Interaction, University of Twente, Enschede, The Netherlands
[3]Department of Data Science, TNO, The Hague, The Netherlands
[4]Department of Statistics, TNO, Leiden, The Netherlands
[5]Department of Methodology & Statistics, Faculty of Social and Behavioural Sciences, University of Utrecht, Utrecht, The Netherlands

Corresponding author:
Alexander Toet, Department of Perceptual and Cognitive Systems, TNO, Kampweg 55, 3769 DE Soesterberg, The Netherlands.
Email: lex.toet@tno.nl

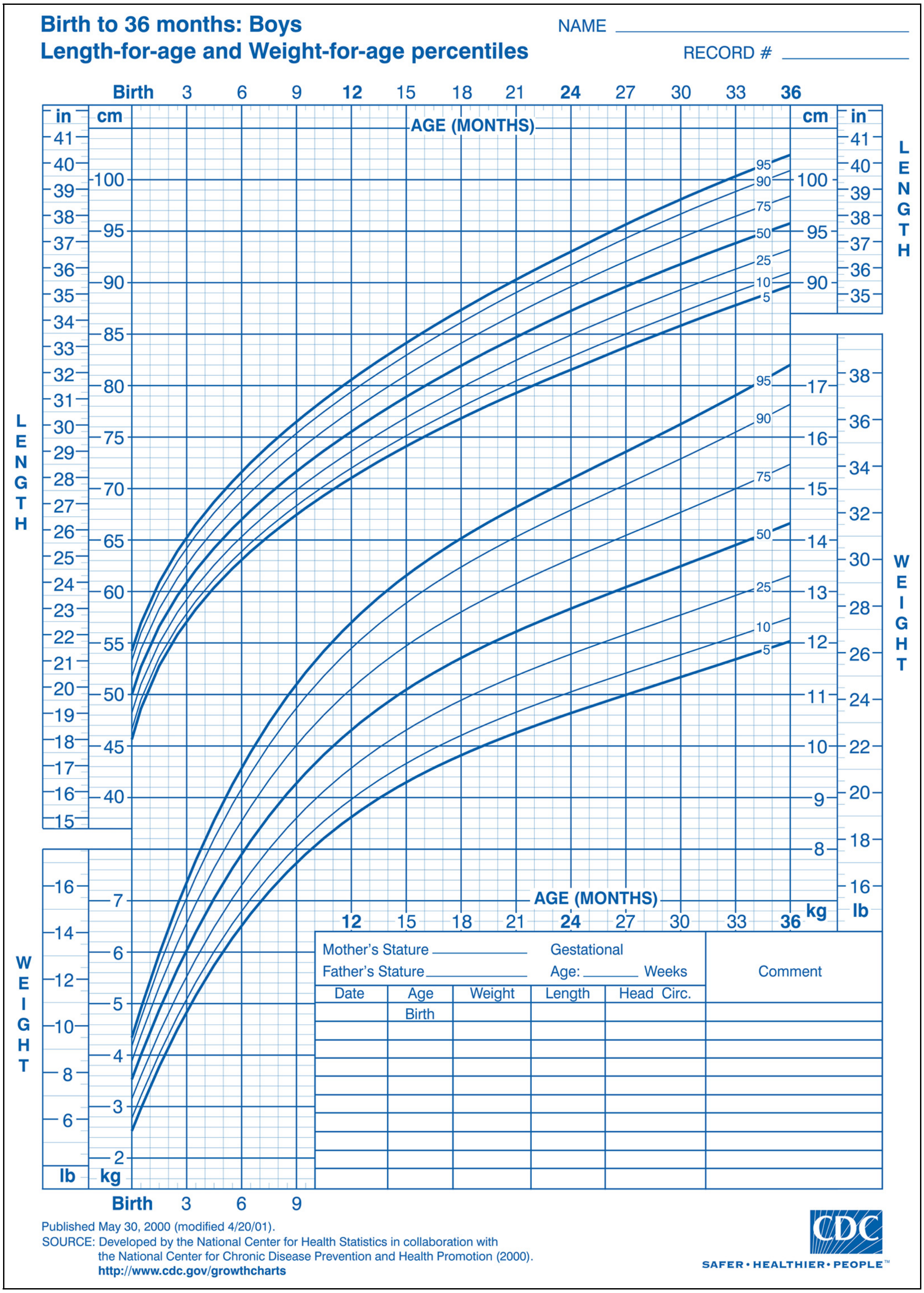

Figure 1. Example of a growth chart depicting length and weight development of boys from birth to 36 months (obtained from http://www.cdc.gov).

still an open research question how an ensemble prediction should be presented graphically so that users correctly understand and interpret the underlying uncertainty distribution.[1]

In two previous studies on visual uncertainty representations, we found that the perceived uncertainty was affected by its graphical representation,[2] particularly for probability estimates that fall outside the

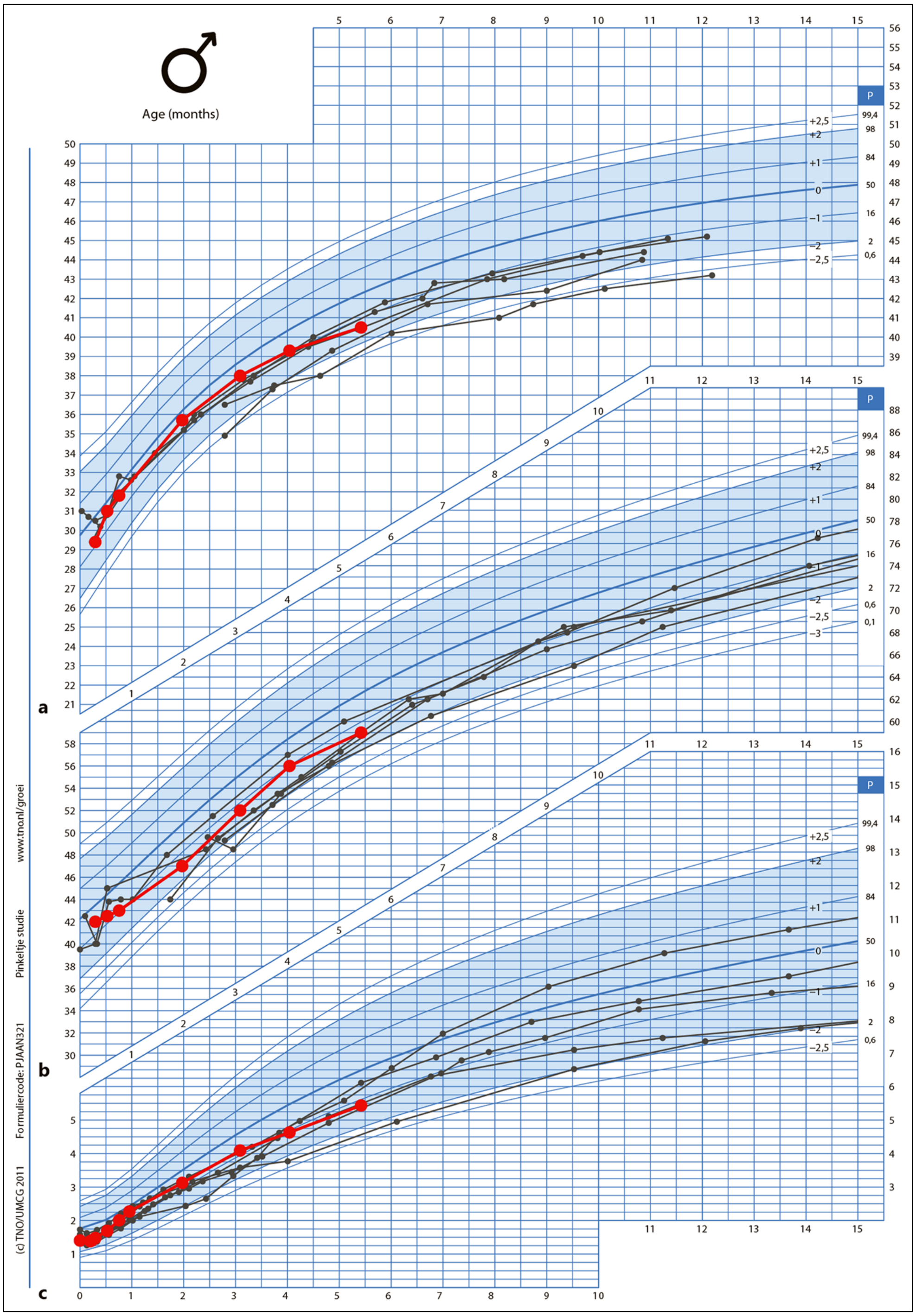

Figure 2. Growth charts for preterm boys born at 32 weeks of gestation. The data are plotted as the growth curves of one index child (red) and five matched children (black). Shown are the (a) head circumference (cm), (b) length (cm), and (c) weight (kg) as a function of age (months).

depicted uncertainty range.[3] It appears that people apply a model of the uncertainty distribution that closely resembles a Bell-shaped (normal) distribution when a graphical representation of the uncertainty range provides no explanation what the depiction exactly indicates.[2,3] Of all graphical uncertainty representations investigated, ensemble plots resulted in an internal model that most closely resembled a normal distribution.[2] The perceived probability of "extreme values" (i.e. values far outside the uncertainty range)

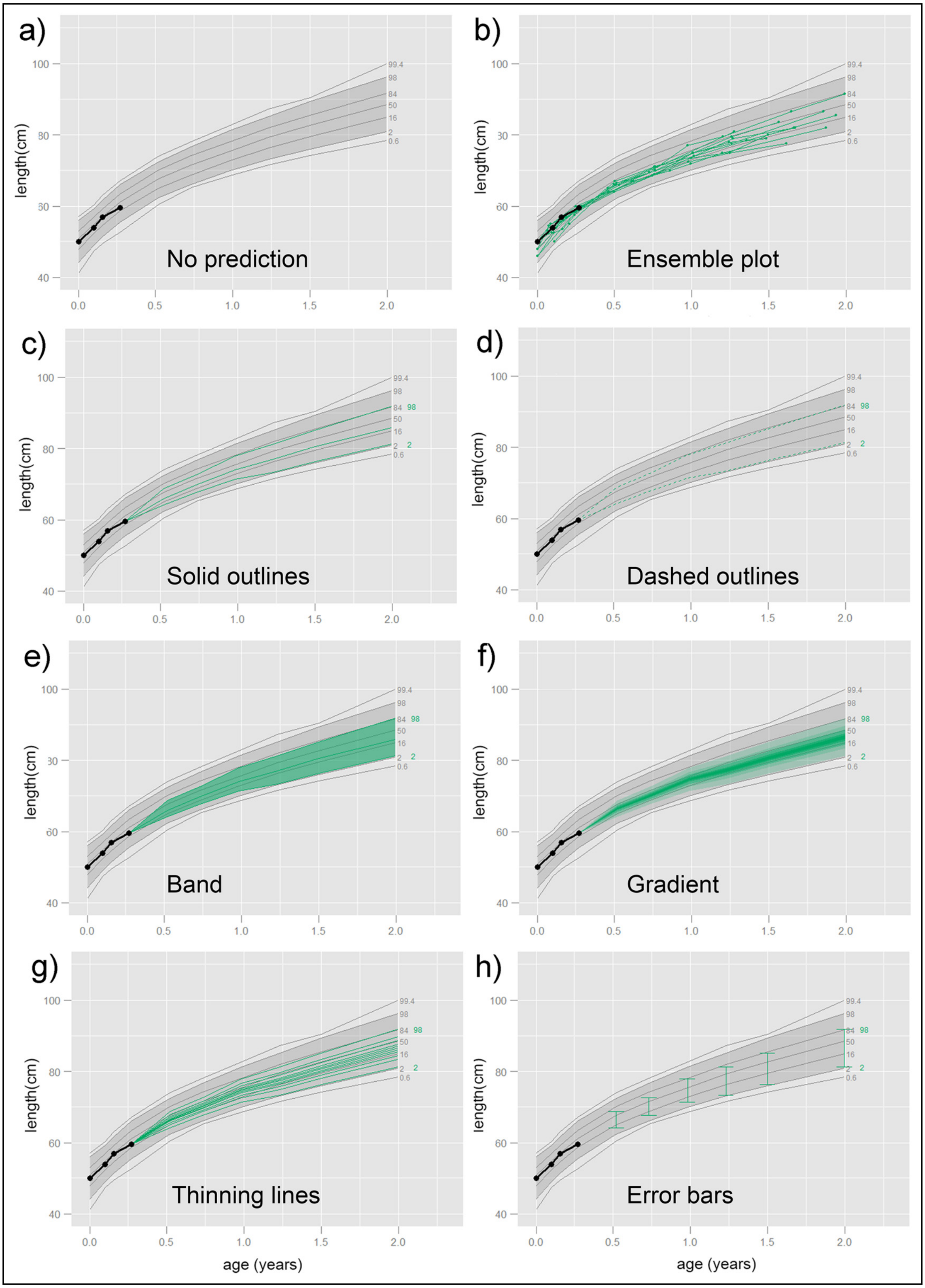

Figure 3. Example of the graphical growth prediction types used in this study. The represented uncertainty area corresponds to the range between 2 and 98 percentile peer curves. (a) No growth predictions. (b) Ensemble of matched peers plots. (c) Solid and (d) dashed outlines of the uncertainty area. Uncertainty area filled with, respectively, (e) uniform (Band) intensity, (f) intensity Gradient, (g) quasi parallel lines with an increasing spacing (decreasing density) from the center outward (Thinning Lines), and (h) Error Bars visualization using "traditional" error bars.

was affected by the visualization type, with denser fills leading to higher perceived probability of values within that area.[3] In addition, we found that the assumed (Bell-shaped) model of the uncertainty distribution depended on a participant's numeracy: people with low numeracy tended to adopt a "flatter" interpretation (i.e. all values are judged more or less equally probable) than those with high numeracy.[2,3] In practice, this means that people with a low numeracy overestimate the probability of extreme values while people with a high numeracy have a more realistic interpretation of uncertainty visualizations when these are presented without any further explanation. The cause of this effect (which has also been observed in several previous studies) is still unknown.[4–7] Thus, it remains a challenge to design uncertainty visualizations that will be correctly interpreted by all users, independent of their numeracy.

In this study, we adopted a pediatric growth chart use case to investigate how different graphical representations affect the user's interpretation of the uncertainty distribution underlying a given ensemble (growth) prediction. Accurate assessment and monitoring of growth in children is of critical importance for early identification of defects associated with treatable conditions versus growth variations associated with normal conditions. Pediatricians and other healthcare providers typically use growth charts to monitor a child's physical development over time. Conventional growth diagrams represent the distribution of continuous measures (e.g. height, weight, and head circumference) against age for a large population of children as a series of percentile curves (e.g. Figure 1). By plotting a child's measures in this population growth chart, they can be compared to the parameters of children of the same age and sex. This enables both health professional and parents to determine whether the child is growing normally (defined as the population mean ± 2(standard deviation (SD))) or not, which may lead to nutritional or therapeutic interventions. Since children generally maintain a fairly smooth growth curve, growth charts can also be used to predict a child's future growth (through extrapolation of the current curve). However, extrapolation based on large peer populations typically contains considerable variance. To improve the prediction for an individual child, it has been proposed to base the extrapolation on a limited set of well-chosen, matched peers instead of the whole population.[8] This new technique to generate ensemble growth predictions is known as curve matching[9] or similarity-based forecasting.[10] The key idea of curve matching is to find peers (references) in databases that are similar to the current child in their initial growth curves (measured at multiple moments in time, for example, at birth and at 1, 2,

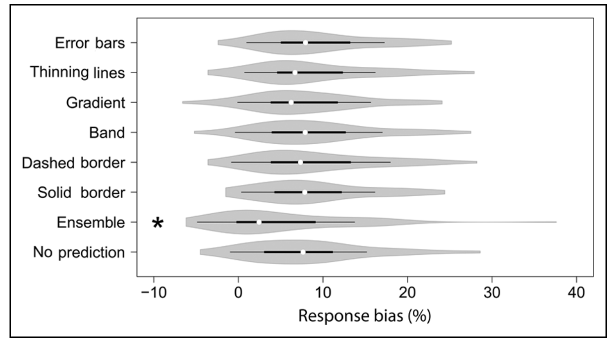

Figure 4. The response bias (mean response error in %, with positive values indicating an overestimation) for each of the eight different uncertainty visualizations. The violin plot represents the interquartile range (thick black bar in the center), the 95% confidence interval (thin black line extending from the black bar), and the median (white dot) of the response error, together with the kernel density plot (shaded area) computed over all responses (i.e. for judgments relative to both the population mean and the normal range) and over all 16 children (N = 64). Significant differences are indicated with *.

and 3 months of age) and in a set of covariates that are known to influence growth (see van Buuren[9] for details of the actual matching procedure). The historical data of the real growth of the matched (existing) children provide an ensemble prediction for the growth pattern of the current child. By providing more accurate and realistic growth predictions, curve matching may enhance the interpretation and prediction of individual growth curves. However, it is currently unknown how the ensemble result of curve matching should be presented to users (medical professionals) and, in particular, to non-experts (parents). A correct perception of the prediction and its variance is of utmost importance to reach an accurate decision whether or not to intervene in a child's growth.

In our previous studies on graphical uncertainty representations,[2,3] users had to compare a predicted range to a single value (e.g. in a weather forecast case: the probability that the temperature will exceed 20°C). However, in many domains (like engineering, quality control, social sciences, and life sciences), predictions (with a given uncertainty) must be compared to a population (with a given variance) and not to a fixed value. This complicates the task in the sense that the user should be able to build an internal model of—and compare—two distributions (the predicted ensemble distribution to the given population) instead of only one. In this study, we therefore ask users to compare the predicted growth of an individual child (with a given uncertainty) to that of the population of children (with a given variance) to answer relevant

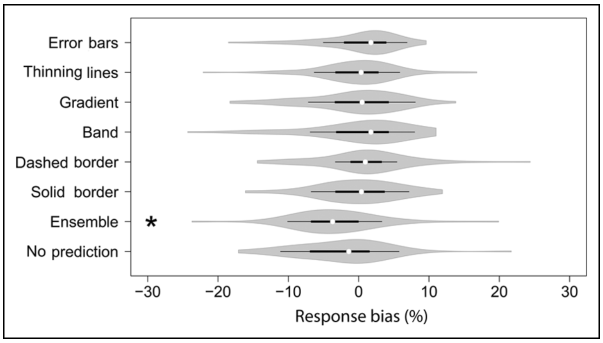

Figure 5. As Figure 4, for judgments relative to the population mean.

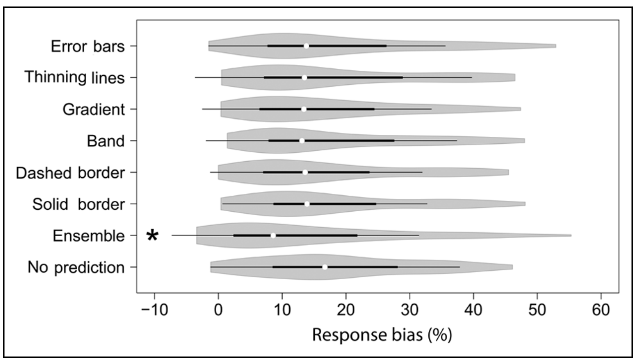

Figure 6. As Figure 4, for judgments relative to the normal range.

questions like "What is the probability that this child will become taller/shorter than the average child?" and "What is the probability that this child will become taller/shorter than what may be considered a normal length (mean ± 2SD)?" We investigate the extent to which eight different graphical representations (see Figure 3) convey the underlying uncertainty in an ensemble of children's growth predictions. We are interested in how people interpret these graphical uncertainty representations, whether their interpretation depends on their numeracy, whether there are biases (systematic errors) in their interpretation, and, if so, whether different types of visualization induce different biases. The key findings of this study are as follows. Numeracy correlates negatively with bias magnitude (i.e. people with lower numeracy show larger response bias). Compared to the summary plots that yield a substantial overestimation of probabilities, and the No-prediction representation that results in quite variable predictions, the Ensemble representation consistently shows a lower probability estimation, resulting in the smallest overall response bias. These results are relevant for other application domains that require a comparison between two distributions.

Related work

It is still an open research question how ensemble predictions and their inherent uncertainty should be presented so that users correctly understand and interpret the results.[1,11–13] Several graphical formats have been proposed to plot a summary representation of an ensemble of predictions, such as line graphs with error bars, glyphs, contour box plots, bar charts, ribbon plots or fan charts (line graphs with an uncertainty "band"), threshold maps (using colorscale or grayscale), and summary tables.[14–19] Summary representations provide users an explicit representation of the spread of the ensemble predictions. However, previous research has shown that such summary representations can lead to misinterpretations of the underlying data distribution.[11,12] It appears that the visual characteristics of these representations differentially affect the user's interpretation of the underlying uncertainty distribution.[2,3,12,20] A more straightforward approach than using summary representations is to visualize the ensemble prediction by plotting a small but representative subset (e.g. Figure 2) of its members (a "spaghetti plot" or "ensemble plot"). Such a representation provides users an implicit representation of the spread of the ensemble members. It appears that users prefer representations showing individual data sources to aggregated data representations,[21] and they seem to create an internal (mental) model based on weighted averaging of the individual data sources,[22] leading to improved inferences.[21] In the context of hurricane trajectory prediction visualizations, it has, for instance, been observed that ensemble plots lead to a more accurate interpretation of the underlying distribution.[12,13] However, users may also misinterpret the variability in this representation as a sign that the data are not very trustworthy ("anything goes"[23]). In this study, we compare the capability of six different graphical summary representations and ensemble plots to convey the underlying uncertainty distribution in ensemble predictions. The six selected summary representations are all commonly used in uncertainty visualizations, and our previous studies showed that no single one of them systematically outperforms all the others.[2,3]

Materials and methods

Participants

In total, 64 people, naïve about the goal of the experiment, participated voluntarily (32 male, 32 female, aged 19–62 years with a mean age of 36 years). The participants were recruited randomly from the TNO database of volunteers and received €10 for their participation. There were no inclusion or exclusion criteria. The experimental protocol was reviewed and approved by the TNO Internal Review Board on experiments with human participants (Ethical Application Ref: TNO-IRB-2015-10-3) and was in accordance with the Helsinki Declaration of 1975, as revised in 2013[24] and the ethical guidelines of the American Psychological Association. All participants gave their written consent prior to the experiment. Participants were asked to report their age, gender, and their highest level of education completed (seven categories: 1 = "primary education/no education," 2 = "lower vocational education," 3 = "lower secondary education," 4 = "higher secondary education," 5 = "BSc," 6 = "MSc," and 7 = "PhD").

Data

The Social Medical Survey of Children Attending Child Health Clinics (SMOCC) cohort is a nationally representative cohort of 2151 children born in the Netherlands in 1988–1989.[25,26] Data were obtained at nine occasions between birth and 30 months of age. Each record in the data corresponds to a visit. Records without valid ages or with missing scores on all developmental indicators were eliminated. The total number of available records was 16,538, pertaining to 2038 infants.

We selected 16 different children for this experiment (their identifiers in the SMOCC database were 10032, 10089, 13010, 14066, 15069, 15094, 31027, 31029, 34021, 34024, 35033, 52021, 52073, 52091, 70031, and 72010). The selection resulted in an equal number of children with a high or low probability of growing taller or smaller than the population mean (±2SD). For each child, we computed a set of 20 best matching peers based on their sex and initial growth curve. The data of the 20 matched peers were used to calculate the predicted mean ± 2SD at several ages.

Measures

Probability estimates. Participants were asked to estimate the probability that the length of a small child (about 3 months old, that had only been measured at four previous occasions) would either reach a length above or below the population mean (questions Q1 and Q2: see Supplementary Material) or the "normal" range (the mean ± 2SD; questions Q3 and Q4) at the age of 2 years. On each stimulus presentation, participants reported their probability estimate by placing a slider along a horizontal bar with endpoints labeled, respectively, "Very unlikely" and "Very likely." The slider position was mapped to values between 0 and 100.

Numeracy. Previous research has shown that people's understanding of visual uncertainty representations depends on their level of numeracy.[2,3,7] We, therefore, assessed the participants' numeracy using seven items from the Rasch-based numeracy scale listed in Weller et al.[27] This scale measures the ability to understand, manipulate, and use numerical information, including probabilities. It correlates with previous (more extensive) objective and subjective numeracy scales, but is a better linear predictor of risk judgments than prior measures.[27] In this study, item 5 from the original scale was omitted because it is equivalent to item 4. Each correct answer was rewarded one point so that the total score on this test ranges between 0 and 7.

Stimuli

In this section, we describe the eight different uncertainty visualizations that were investigated in this study: a plot without any predictions, an ensemble plot showing a small representative set of predictions, and six summary plots (for an overview see Figure 3). These visualizations were all produced by plotting the (predicted) data of an individual child in the population growth chart based on the population mean ± 2SD (mimicking the growth charts that are commonly used). In this type of chart, lengths within a range of 2SD from the mean are considered as normal, while lengths that are more than 2SD above or below the mean are considered as long and short, respectively, compared to their peers.

A "No-prediction" representation (Figure 3(a)), showing only the first four length measurements of a given child, served as a baseline. This representation only shows the currently available information without any growth prediction.

The *Ensemble representation* (Figure 3(b)) shows the actual growth curves of the 10 best matching peers resulting from the curve matching procedure, including the marks representing the actual measurements.

The six remaining uncertainty visualizations are summary plots (uncertainty area outlined by solid or dashed borders or filled with uniform or gradient intensity or with quasi parallel lines or a standard error bars: see Figure 3(c)–(h)), similar to the ones we previously investigated in our previous studies on the perceived uncertainty in, respectively, the spatial position of earth layers[2] and in temperature forecasts.[3] The mean and SD of the set of 20 best matching peers served to construct the predicted uncertainty region for these summary plots, such that this region had a width of 4SD around its mean (similar to the population graph). We included these summary representations since they are all commonly used in uncertainty visualizations and our previous studies showed that no single one of them systematically outperformed all the others. In the rest of this section, we will discuss the construction of the six summary representations in more detail.

The *Solid Border* (Figure 3(c)) and *Dashed Border* (Figure 3(d)) visualizations outline an uncertainty range with a width of 4SD (i.e. the area between 2 and 98 percentiles) by, respectively, solid or dashed lines. The *Band* (Figure 3(e)) and *Gradient* (Figure 3(f)) visualizations fill this uncertainty range with, respectively, a uniform or a gradient grayscale distribution.

The *Thinning Lines* (Figure 3(g)) visualization fills the uncertainty area with lines that are quasi parallel and with a decreasing density (increasing spacing) from the center outward. To construct the Thinning Lines representation, we first determined 11 different points (lengths) within the 2SD interval at each age at which the child had been measured, such that (1) these points were normally distributed over the 2SD interval, (2) the distance between these points monotonously increased from the center of the 2SD interval outwards, and (3) their midpoint coincided with the median of the length values thus generated. The Thinning Lines representation was then obtained by connecting the midpoints and by connecting points that were an equal number of steps above or below the midpoints.

The *Gradient representation* was constructed from the Thinning Lines representation. Starting at the borders of the 2SD interval and going toward the center, the area between each pair of lines was filled with a transparent fill so that the gradient started light at the edges and became darker toward the center. Finally, the *Error Bars* (Figure 3(h)) visualization uses conventional error bars to delineate the interval between the between the 2 and 98 percentiles.

The four questions pertaining to the probability judgments (Q1–Q4: see Supplementary Material) were distributed over the stimuli such that each question was asked for four children (Q1 for 10032, 10089, 13010, and 14066; Q2 for 15069, 15094, 31027, and 31029; Q3 for 34021, 34024, 35033, and 52021; and Q4 for 52073, 52091, 70031, and 72010). In addition, the selection was such that each question was asked for two children with a high probability of growing taller than the population mean (10032 and 10089 for Q1; 15069 and 31029 for Q2; 34021 and 35033 for Q3; 70031 and 72010 for Q4) and for two children with a high probability of becoming smaller than the population mean (13010 and 14066 for Q1; 15094, 31027 for Q2; 34024 and 52021 for Q3; and 52073 and 52091 for Q4). The predictions for each child were shown in all eight graphical formats, resulting in a total of 128 (=16 children × 8 representations) stimuli. The stimuli were presented in random order to each participant. The images had a size of 800 × 600 pixels.

Procedure

The experiment was performed self-paced and online. The participants were presented with 128 different growth diagrams of small children (about 3 months old) that had only been measured at four points in time, both with and without predictions of their future growth. Their task was to interpret these diagrams and judge the probability that a given child will reach a certain length at the age of 2 years. The type of uncertainty estimate (i.e. the probability that the child's length at the age of 2 years will either be below or above the population mean or normal range), the identity of the child, and the type of uncertainty representation varied randomly between stimulus presentations. On each presentation participants gave their probability estimates by placing a slider along horizontal bar with endpoints labeled, respectively, "Very unlikely" and "Very likely." After, respectively, 48 and 96 stimuli had been presented (i.e. at 1/3 and at 2/3 of the stimulus presentations), a growth chart was shown in combination with a control question to test whether participants were actually paying attention or whether they were merely responding randomly. These control questions had the same layout as the four regular questions, but asked the participants to place the response slider at one of its endpoints. After judging all 128 growth curves, the seven-item Rasch-based numerical test was presented, and the participants were asked for their demographical data.

Data analysis

The statistical data analysis was done with IBM SPSS 22.0 for Windows (www.ibm.com). For all analyses, a probability level of $p < 0.05$ was considered to be statistically significant. Using SPSS 22.0, we computed the correct answer (i.e. the actual probability that a child would be above/below the normal mean/region at an age of 2 years) for each child from the distributions of both the normal and the peer populations. The difference between the participant's response and the correct answer (i.e. the response error) was used for further analysis. The data with this article is publicly available at figshare (growth estimates data) with doi: 10.6084/m9.figshare.4052373.

Results

A Shapiro–Wilk test of normality showed that the data were not normally distributed. Therefore, we used non-parametric tests to further analyze the data.

Overall performance

Figure 4 shows the response bias (mean response error) for each of the eight different uncertainty visualizations, combined over all estimates (i.e. both relative to the mean and normal range of the population), over all 16 stimuli, and over all 64 observers. A

Friedman test showed that there was a significant difference between the eight different visualizations: $\chi^2(7) = 51.828$; $p < 0.001$. A post hoc analysis showed that the Ensemble visualization differed significantly from all other visualizations: it had the lowest median overall bias (an overestimation of probabilities by 2.5%; $p < 0.001$, with a medium effect size: Cohen's d = 0.52). There were no significant differences between the remaining seven visualizations. Hence, overall the bias resulting from all summary representations was similar to the bias in the absence of an uncertainty representation.

Effect of comparison to the mean or to the range

Next, we investigated whether the response bias differed for the different types of comparisons the respondents were asked to make (i.e. a judgment relative to the normal mean or relative to the normal range of the population).

Figure 5 shows the response bias when respondents were asked to compare the prediction to the population mean (Q1 and Q2). Performance is generally acceptable with a median value close to 0. A Friedman test showed that the Ensemble representation yielded significantly more negative bias (an underestimation of probabilities by 3.7%; $p < 0.001$) for judgments relative to the mean than the other methods: $\chi^2(7) = 56.424$; $p < 0.001$, with a medium effect size: Cohen's d = 0.56. This means that users underestimated the probability that a child will either become taller or shorter than the population mean when basing their judgments on the Ensemble representation. All other methods yielded biases closer to zero (ranging between −1.40% and 1.75%; $p < 0.001$), meaning these representations induce only small systematic biases for this type of judgment.

Figure 6 shows the response bias when respondents were asked to compare the prediction to the population range (Q3 and Q4). Generally, all visualizations show a large overestimation with (except for the ensemble plot) a median bias larger than 10%, while this estimation is critical for this specific use case since it directly relates to the question whether or not to intervene in a child's growth. A Friedman test showed that the Ensemble visualization yielded a significantly smaller bias for judgments relative to the population range (an overestimation of probabilities by 8.6%; $p < 0.001$) than the other methods (overestimation ranging between 13.1% and 16.7%); $p < 0.001$: $\chi^2(7) = 57.074$; $p < 0.001$, with a small effect size: Cohen's d = 0.32. This means that users could better estimate the probability what a child's length will be relative to the population range when basing their judgments on the Ensemble representation.

The general picture that arises from these results is that there is a general overestimation of probabilities and that the ensemble plot generally results in lower estimates than the seven other plots, leading to an underestimation for comparisons to the population mean and a reduced overestimation for comparisons to the population range. An important question is why the ensemble plot leads to lower estimations.

Effect of numeracy and education

Numeracy showed a negative correlation with bias magnitude: the Spearman correlation was r = 0.406 ± 0.001. Numeracy correlated only weakly with education level (r = 0.24) and education level only weekly with bias (r = 0.18). This confirms earlier observations[2,3] that numeracy is strongly related to biases in graphical uncertainty interpretations than education.

Discussion and limitations

In this section, we discuss the results and limitations of this study, present the conclusions of this study, and make some recommendations for future work.

Discussion

In this study we investigated how eight different graphical representations of the uncertainty in an ensemble of children's growth predictions affect the user's interpretation and whether they induce any systematic interpretation biases.

Generally, we find a substantial overestimation of probabilities (similar to previous studies[2,3]), which probably results because people apply a broader internal probability model. However, it appears that non-expert users vary widely in their interpretation of graphical representations of predicted probability distributions. As a result, there is a large variation in the estimated probabilities between different visualizations and types of judgments.

Compared to the seven other visualizations investigated in this study (six summary uncertainty representations and the absence of an uncertainty representation), the Ensemble representation shows a lower probability estimation, resulting in the smallest overall (i.e. over all judgment tasks and cases tested) response bias (a small overestimation of probabilities in all judgment tasks). More specifically, the ensemble plot shows a substantial reduction in the overestimation for comparisons relative to the population range (the most relevant questions in this use case since estimates outside the population range considered normal may lead to interventions) and only a small underestimation for comparisons relative to the mean of the population. Although the ensemble plot yields the (overall) best performance, the importance of different underestimations and overestimations may be use-case dependent. For none of the specific children and questions investigated, we found a significant difference among the six summary plots. In other words, although they may offer benefits compared to No-prediction, they all score similarly and equal or lower than the ensemble plot. The results of the No-prediction visualizations differed significantly from the results obtained with the other prediction visualizations in several conditions. However, there appears to be no systematic pattern in the response bias for the No-prediction plot: the estimation seems to a large degree determined by the pattern seen in the initial data points, which make the predictions variable, resulting in the occurrence of both significant underestimations as well as overestimations.

The results of this study confirm earlier reports that the interpretation of a visualization of an uncertainty distribution depends on its visual characteristics[2,3,12] and that visualization providing access to individual predictions lead to less-biased estimates.[11,22] In addition, it appears that the type of estimate required (e.g. relative to the mean or normal range of a population) also differentially affects the relative observer performance with different visualizations. It has been suggested that this may be a result of the variations in feature saliency between the different representations.[11] Further research is needed to clarify the nature of this relation. It is also not clear why exactly the ensemble plot leads to lower estimates than the summary visualizations. Additional research may show whether this finding reflects the shape of the internal model of the uncertainty distribution that users construct.[2,3]

We found that numeracy correlated negatively with bias magnitude. This means that probability estimates produced by people with lower numeracy differ more from the predicted probability than estimates given by people with higher numeracy. This result agrees with the earlier finding that people with relatively low numeracy have a "flatter" (or more dispersed) interpretation of the underlying uncertainty distribution than those with higher numeracy.[3] Our result that numeracy correlated only weakly with education level agrees with previous findings[28] and with the observation that—even among highly educated samples—the ability to solve basic numeracy problems is (on average) relatively poor.[29]

Given the present finding that the Ensemble representation induced the smallest overall response bias, it appears that this representation may be the best choice for communicating the uncertainty in ensemble predictions to non-expert users. It seems that this representation effectively communicates essential unpredictability through the metaphor of "multiple possible outcomes."

Limitations and future work

We limited our study to eight different uncertainty visualizations which were tested for 16 different cases (children). This resulted in 128 decisions per participant. For an online experiment, this is already quite a large number of trials. The inclusion of more visualizations or a larger number of cases would have resulted in an excessively large number of trials which might have discouraged participants from finishing the online experiment. Note that we only included children that were outliers in the sense that they all had an increased probability of becoming shorter or taller than the average or normal length. Assuming a Gaussian distribution, the number of children who are outliers in length are actually quite small (4%, by construction). Hence, the base probability of abnormality is also quite low. The class bias in our set of stimuli (100% outliers) could induce the adoption of an observer response bias toward outliers, a regression toward the mean, or simply in no response adjustment at all. These responses are mutually exclusive and predict different effects: a regression to the mean will result in an underestimation of being an outlier, while a response bias toward outliers will result in an overestimation of outlier probability. A future study including a (much larger) well-balanced stimulus set may provide more insight into these effects.

We hypothesized that, in the absence of additional information, the shape of the initial curve may determine the estimated outcome. However, it is not possible to establish a relation based on the limited number of cases investigated here. A future study can investigate this hypothesis by systematically varying the position (relative to the normal range), length, and direction (tendency to go upward or downward) of the initial curve.

Similarly, ensemble representations may establish more trust in a forecast if several members indicate the same magnitude and direction. A future study may investigate this effect by systematically varying the fraction of curves with a similar tendency.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

ORCID iD

Alexander Toet https://orcid.org/0000-0003-1051-5422

References

1. Szafir DA, Haroz S, Gleicher M, et al. Four types of ensemble coding in data visualizations. *J Vision* 2016; 16: 11.
2. Tak S, Toet A, Van Erp J, et al. The perception of visual uncertainty representation by non-experts. *IEEE T Vis Comput Gr* 2014; 20: 935–943.
3. Tak S, Toet A and van Erp J. Public understanding of visual representations of uncertainty in temperature forecasts. *J Cogn Eng Decis Mak* 2015; 9: 241–262.
4. Lipkus IM and Peters E. Understanding the role of numeracy in health: proposed theoretical framework and practical insights. *Health Educ Behav* 2009; 36: 1065–1081.
5. Peters E, Västfjäll D, Slovic P, et al. Numeracy and decision making. *Psychol Sci* 2006; 17: 407–413.
6. Reyna VF, Nelson WL, Han PK, et al. How numeracy influences risk comprehension and medical decision making. *Psychol Bull* 2009; 135: 943–973.
7. Rinne LF and Mazzocco MMM. Inferring uncertainty from interval estimates: effects of alpha level and numeracy. *Judgm Decis Mak* 2013; 8: 330–344.
8. Hermanussen M, Staub K, Assmann C, et al. Dilemmas in choosing and using growth charts. *Pediatr Endocr Rev* 2012; 9: 650–656.
9. Van Buuren S. Curve matching: a data-driven technique to improve individual prediction of childhood growth. *Ann Nutr Metab* 2014; 65: 227–233.
10. Buono P, Plaisant C, Simeone A, et al. Similarity-based forecasting with simultaneous previews: a river plot interface for time series forecasting. In: *Proceedings of the 11th international conference on information visualization (IV '07)*, Zurich, 4–6 July 2007, pp. 191–196. New York: IEEE.
11. Padilla LM, Ruginski IT and Creem-Regehr SH. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cogn Res Princ Implic* 2017; 2: 40.
12. Ruginski IT, Boone AP, Padilla LM, et al. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spat Cogn Comput* 2016; 16: 154–172.
13. Liu L, Boone AP, Ruginski IT, et al. Uncertainty visualization by representative sampling from prediction ensembles. *IEEE T Vis Computer Gr* 2017; 23: 2165–2178.
14. Demeritt D, Cloke H, Pappenberger F, et al. Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environ Hazards* 2007; 7: 115–127.
15. Demeritt D, Nobert S, Cloke HL, et al. The European flood alert system and the communication, perception, and use of ensemble predictions for operational flood risk management. *Hydrol Process* 2013; 27: 147–157.
16. Kootval H. *Guidelines on communicating forecast uncertainty*, 2008. Geneva: World Meteorological Organization, https://www.preventionweb.net/publications/view/26243
17. Pappenberger F, Stephens E, Thielen J, et al. Visualizing probabilistic flood forecast information: expert preferences and perceptions of best practice in uncertainty communication. *Hydrol Process* 2013; 27: 132–146.
18. Sanyal J, Zhang S, Dyer J, et al. Noodles: a tool for visualization of numerical weather model ensemble uncertainty. *IEEE T Vis Comput Gr* 2010; 16: 1421–1430.
19. Whitaker RT, Mirzargar M and Kirby RM. Contour boxplots: a method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE T Vis Comput Gr* 2013; 19: 2713–2722.
20. Miran SM, Ling C, James JJ, et al. User perception and interpretation of tornado probabilistic hazard information: comparison of four graphical designs. *Appl Ergon* 2017; 65: 277–285.
21. Greis M, Avci E, Schmidt A, et al. Increasing users' confidence in uncertain data by aggregating data from multiple sources. In: *Proceedings of the CHI conference on human factors in computing systems*, Denver, CO, 6–11 May 2017, pp. 828–840. New York: ACM.
22. Greis M, Joshi A, Singer K, et al. Uncertainty visualization influences how humans aggregate discrepant information. In: *Proceedings of the CHI conference on human factors in computing systems*, Montreal, QC, Canada, 21–26 April 2018, pp. 1–12. New York: ACM.
23. Lipkus IM and Hollands JG. The visual communication of risk. *J Natl Cancer Inst Monogr* 1999 25; 149–163.
24. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. *J Am Med Assoc* 2013; 310: 2191–2194.
25. Herngreen WP, Reerink JD, van Noord-Zaadstra BM, et al. SMOCC: design of a representative cohort-study of live-born infants in the Netherlands. *Europ J Public Health* 1992; 2: 117–122.
26. Herngreen WP, van Buuren S, van Wieringen JC, et al. Growth in length and weight from birth to 2 years of a representative sample of Netherlands children (born in 1988–89) related to socioeconomic status and other background characteristics. *Annal Human Bio* 1994; 21: 449–463.
27. Weller JA, Dieckmann NF, Tusler M, et al. Development and testing of an abbreviated numeracy scale: a Rasch analysis approach. *J Behav Decis Mak* 2013; 26: 198–212.
28. Ghazal S, Cokely ET and Garcia-Retamero R. Predicting biases in very highly educated samples: numeracy and metacognition. *Judg Decis Mak* 2014; 9: 15–34.
29. Lipkus IM, Samsa G and Rimer BK. General performance on a numeracy scale among highly educated samples. *Med Decis Mak* 2001; 21: 37–44.