ILLNESS COGNITION: PRIOR EXPECTATIONS FOR HEALTH

By

TALIA ROBBINS

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Psychology

Written under the direction of

Pernille Hemmer

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2019

ABSTRACT OF THE DISSERTATION

Illness Cognition: Prior Expectations for Health

by TALIA ROBBINS

Dissertation Director:

Pernille Hemmer

In this dissertation, I will investigate how people make judgments and decisions in the domain of health. My overarching goal is to bridge the gap between behavioral health research and cognitive science. While both fields have made important strides in health decision making, insufficient communication between fields prevents health researchers from benefiting from important findings in cognition, and vice versa. For example, previous models in behavioral health are based mainly on patient health data, and have not been evaluated in terms of their implications for cognition, or computationally. Of particular interest in my research program is the importance of prior expectations, and I will focus on three inter-related questions bearing on the importance of prior expectations. Chapters 2 and 3 will evaluate people's prior expectations for illness statistics; chapter 4 will address how people use prior expectations for prediction of illness durations; and sections 5 and 6 will investigate how these prior expectations are integrated with new evidence (e.g., a diagnosis from a doctor). Furthermore, chapter 5 will propose a rational model to describe how people's health judgments change as they encounter new information.

Acknowledgments

I would like to express my deepest appreciation to my advisor, Dr. Pernille Hemmer, for her invaluable support and guidance throughout my graduate career. She has always encouraged me to ask my own questions and has provided me with invaluable experience in both in research and mentoring undergraduate students. I would also like to thank my committee members, Dr. Jacob Feldman, Dr. Julien Musolino, and Dr. Gretchen Chapman for their insightful comments and feedback. Special thanks to the current and former members of the Priors and Memory (Prime) lab, without whom this work would not be possible. Specifically, I wish to thank Kimele Persaud, Daniel Wall, Joseph Sommer, Chrystal Spencer, and Kevin Pei.

I would also like to thank the staff in the psychology department and the Center for Cognitive Science here are Rutgers for their help over the years, namely Anne Sokolowski, Jo'Ann Meli, and Tamela Wilcox.

Chapter 2, in partial, is a reprint of the material as it appears in: Robbins, T., & Hemmer, P. (2017). Explicit Predictions for Illness Statsitics. In Gunzelmann, G., Howes, A., Tenbrink, T., & Davelaar, E. (Eds.), Proceedings of the 39th Annual Meeting of the Cognitive Science Society. London, UK: Cognitive Science Society. The dissertation author was the primary investigator and author of this paper. Chapter 3, in partial, is a reprint of the material as it appears in: Robbins, T., & Hemmer, P. (2017). Lay Understanding of Illness Probability Distributions. In Kalish, C, Rau, M, Zhu, J, & Rogers, T.T. (Eds.), Proceedings of the 38th Annual Meeting of the Cognitive Science Society. Madison, WI: Cognitive Science Society. The dissertation author was the primary

investigator and author of this work. Completion of this work was supported by the National Science Foundation CAREER Grant (1453276).

Dedication

This dissertation is dedicated to my grandfather, Dr. Murray Robbins. My grandfather was an esteemed chemist at Bell Labs, with patents and impressive discoveries to his name, who spent his free time teaching elementary schoolers science. There was nothing he loved more than to share his love of science with others, especially his grandchildren. Whenever I struggle with my research, I am reminded that as my grandfather was working on his PhD, his advisor passed away and he had to start entirely over. He was a force of nature. He will be an extremely tough act to follow, and I sincerely hope I can live up to his legacy.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Judgments and decisions are assumed to originate with a person's experience with the world. People combine their prior expectations with the available evidence in order to arrive at a decision. This means that when someone makes a suboptimal decision, one of two things is at play: the person is using a *flawed process* to arrive at the answer, or the person is working with *faulty information*. A particularly important question for decision-making research concerns how expectations influence judgments and decisions. While much of the well-known work of Tversky and Kahneman suggests that decision processes are flawed (e.g., 1974, 1992), there is also evidence that people use their expectations optimally (e.g., Griffiths & Tenenbaum, 2006). For example, people's predictions for life spans and movie run times are quite accurate in the aggregate.

This suggests not only that the judgment process is optimal, but that people's expectations are consistent with real-world statistics. When making a suboptimal decision, it is possible that rather than using a biased process, people are using a normative model, but with flawed information. For instance, if the decision process is rational (Bayesian), decisions are based on a combination of observed noisy data and an accurate probabilistic model of the environment (i.e., expectations). However, if those expectations are incorrect, it can lead to flawed judgments and decisions. This can account for flawed judgments under an optimal framework by assuming differences in prior expectations, or mapping expectations from a known domain to an unknown domain. Each time a person experiences a new event, they should update their prior probability for that event by integrating the new information. This should result in events that are experienced more often having prior expectations that more closely reflect environmental statistics. For those that are less

commonly experienced, people might use their prior expectations using events for which they have more knowledge, when making inferences.

Expectations are particularly important within the domain of health, where patient expectations directly impact health outcomes (Peters, 2006). For instance, when a person is diagnosed with a new illness, they are tasked with integrating this new information about their health in order to adapt their expectations and respond appropriately—which in some cases may mean daily medication or lifestyle changes. If a person does not have strong expectations (based on previous experience) for a new illness they are experiencing, such as duration, severity, and symptoms of that illness, they are unlikely to manage it correctly. As a person's health is constantly changing, good decision makers must continually adjust their expectations to track these changes.

While cognitive research on how people adjust to environmental changes can be applied to understanding health behavior, health decisions have been suggested to operate differently from other decision processes. Health numeracy, for instance, is significantly worse than numeracy in other areas (Levy et al., 2014). In order to understand how people update expectations about their health, a prominent theoretical model in behavioral health, the Common Sense Model of illness cognition (CSM, Leventhal, 1990), has been used to describe patient behavior. The CSM argues that people have a model of illness that they can use to understand new illnesses and symptoms (Leventhal, 1992). Prior history of illness episodes is often based largely on acute illnesses because they are more commonly experienced, while for chronic illnesses expectations are based largely on an abstraction about the way illnesses function in general (Leventhal et al., 1992). Furthermore, the CSM argues that people apply expectations for illnesses they are familiar with (generally acute

illnesses) to illnesses they know less about (generally chronic illnesses). This model has been validated empirically mainly on the basis of patient health data, but has not been evaluated computationally, or in terms of its implications for cognition. However, it is important to examine both cognitive research from laboratory studies and from behavioral health research that focuses on patient outcomes and decisions in the real-world.

The overarching goal of this work is to bridge the gap between behavioral health research and cognitive science. This dissertation seeks to evaluate people's prior expectations for health, and how these expectations influence judgments and decisions. My research focuses on three important and inter-related questions bearing on the importance of prior expectations; (1) What are peoples' prior expectations for illness statistics? (2) How do people use prior expectations for prediction? And (3) How are prior expectations for illness statistics integrated with evidence to influence judgments? What follows is a brief overview of what will be covered in each of the chapters of this dissertation.

Chapters 2 and 3 evaluate people's prior expectations for illness statistics. This work appears in the 2017 and 2018 *Proceedings of the Annual Meeting of the Cognitive Science Society and* was presented at the 2017 *Annual Meeting of the Society for Mathematical Psychology,* and the 2017 *Annual Meeting of the Psychonomic Society.* This work was also presented at the 2017 *Annual Meeting for the Association of Psychological Science* and received the NIDCR Building Bridges Travel Award. These chapters evaluate participant expectations for the mean and shape of illness duration distributions, as well as the distribution as a whole. Results illustrate that participants' estimations for the mean duration and distributional form of illnesses they are familiar with (e.g., the common cold) closely reflect average illness duration data; while estimations for illnesses they have less

experience with (e.g., appendicitis) show a pattern of systematic overestimation. I also found that participants estimates for the form of chronic illness distributions were similar to those for acute illnesses—that is, they may have transferred their understanding of the shape of illness distributions from acute illnesses they are familiar with, to chronic illnesses they do not have experience with.

Chapter 4 will address how people use prior expectations for prediction. This work appears in the 2016 *Proceedings of the Annual Meeting of the Cognitive Science Society,* and was presented at the 2016 *Annual Meeting of the Psychonomic Society* and the 2016 *Annual Meeting of the Society for Mathematical Psychology.* The results from this work suggest that participants predictions for illness durations for acute illnesses more closely match average illness duration data than do chronic illnesses. Additionally, results from predictions in this chapter are consistent with participant estimates from Chapters 2 and 3.

Chapter 5 will investigate how new information (e.g., a diagnosis from a doctor) is integrated into existing expectations when people are asked to make repeated judgments under uncertainty. This work was presented at the 2018 *Annual Meeting of the Psychonomic Society* and the 2018 *International Convention of Psychological Science.* The results in this chapter illustrate that judgments made in the domain of health are different than those made in other domains, and that the source of the information influences judgment change, such that authority figures are trusted more than online resources or experience judgments.

Chapter 5 will also propose a rational model to describe how people's health judgments change as they encounter new information. Past work has suggested that people discount advice egocentrically—meaning that they underweight advice from others

relative to their own judgments (e.g., Yaniv & Kleinberger, 2000). However, those experiments have not measured people's confidence both before and after receiving advice. For this reason, they cannot determine whether participant's weighting of their own advice is appropriate given their level of confidence in their own judgment. I implement and compare several models, including a simple weighted confidence model, an egocentric discounting model, and a model which assumes that people simply take the average of source and personal judgments. I choose these models because the averaging model has been described as a normative model for how people *should* integrate advice, and the egocentric model as a descriptive model of how people *do* integrate advice (e.g., Yaniv & Kleinberger, 2000). However, it may be the case that what appears to be egocentric discounting is more accurately described as greater confidence in personal judgments than source information. The weighted confidence model assumes that initial judgments and source information are combined using a weighting structure based on a person's confidence in their own judgment, as well as their confidence in the source.

Chapter 5 will also ask whether the order in which health information is provided is important for judgment change, as the order of information has been found to influence decision-making with the last piece of information being weighted more strongly (Bergus, Chapman, Levy, Ely, & Opplinger, 1998). This experiment expands on those investigations by asking how differing levels of confidence in the source might interact with order effects. The combination of research in this dissertation provides a broad understanding of the cognitive mechanisms that influence prior expectations and judgments in the domain of health.

Chapter 2: Explicit Estimations for Illness Statistics

This work appears in the 2017 *Proceedings of the Annual Meeting of the Cognitive Science Society and* was presented at the 2017 *Annual Meeting of the Society for Mathematical Psychology,* and the 2017 *Annual Meeting of the Psychonomic Society.* This work was also presented at the 2017 *Annual Meeting for the Association of Psychological Science* and received the NIDCR Building Bridges Travel Award.

Abstract

People's predictions for real-world events have been shown to closely match environmental statistics (e.g., Griffiths and Tenenbaum 2006). However, health judgments have been shown to differ from judgments in other domains (Levy et al., 2014). With this in mind, we focus on assessing participant expectations for illness durations. Specifically, we assess expectations for both the mean and form of illness duration distributions. We assess understanding for both acute illnesses for which people might have experience, as well as chronic conditions for which people are less likely to have knowledge. Our data illustrates important differences in how people make estimations for the duration of acute and chronic illnesses.

Introduction

Imagine that you have the flu and need to decide whether you will be better in time to travel to a conference this weekend. You are now faced with predicting how long you will be sick. For this inference, you will need to use your knowledge of real-world statistics, including both the mean duration and most likely form of the duration distribution.

People have been shown to make optimal predictions for the duration of many real-world events (Griffiths & Tenenbaum, 2006). In these domains, people's beliefs about the underlying distribution of quantities (e.g., cake baking times are captured by a bimodal distribution) have been shown to be accurate in the aggregate. In order to extend these findings to the domain of health, in the current study, we assessed people's expectations for illness durations by asking them directly what they thought the mean and form of illness duration distributions were. This allowed us to evaluate whether people have an internal model for real-world statistics that they can consciously access and use to make estimations.

Understanding illness duration is critical for illness identification. For instance, imagine you have a cough and high fever, and thinking you have the flu you try to estimate how long you will be sick. One thing you will draw on is your understanding of the real-world distribution of durations for different illnesses. If your symptoms begin to fade after three days, this may confirm your suspicion that you have the flu, since this is within the normal distribution for the flu. However, if you are still sick after 10 days, you might begin to believe you have a different illness such as the common cold, because you know that 10 days is reasonable within the distribution of duration for the common cold. This estimation requires an understanding of the entire distribution of illness duration, rather than just the

mean or some conditional duration. With only the mean of the distribution, you would not know how much variation in duration is normal, or at which point a particular illness is unlikely given the duration of your symptoms.

Illness further provides an interesting example for prediction because people have different levels of experience for different illnesses—e.g., common illnesses such as the cold, or less common illnesses such as appendicitis. Experience may also differ between acute (e.g., cold) and chronic (e.g., asthma) illnesses. An acute illness is defined as one which can be cured with treatment, while a chronic illness is defined as one that can be managed but not cured. Differing levels of experience between chronic and acute illnesses may influence the accuracy of a person's expectations, and different expectations might be appropriate for different illnesses, given personal experience.

The observer's prior experiences play an important role, as optimal predictions are assumed to follow Bayesian principles. Bayes rule gives a principled account of how people should update their prior beliefs given evidence from the world. Each time a person experiences an illness, they should update their prior probability distributions for the duration of that illness. This would result in illnesses that are experienced more often having very accurate prior distributions. For illnesses that are less commonly experienced, people might adjust their prior beliefs to those of illnesses for which they have more knowledge of the correct form of the distribution, when making inferences. While people might use evidence from other sources when updating their priors, evidence that is personally experienced is better integrated than information acquired in other ways (Sallnas, Rassmus-Grohn, & Sjostrom, 2000).

In Experiment 1, we simply asked participants to predict the mean duration of each

of six illnesses. In Experiment 2, we sought to assess whether people could make estimations of the correct form of illness distributions. To do this, we gave participants four distribution options—each fit to the clinical data for that illness—and asked them to select the distribution form that best described that illness. Because each of the distribution options was fit to the clinical data, consistent selection of the correct distribution would clearly illustrate that there is a correspondence between people's internal model and the true statistics of the environment.

Experiment 1

Participants

Ninety-Nine Mechanical-Turk workers from the United States participated in exchange for $1.

Materials

We selected six illnesses—five acute and four chronic (see Table 1.1)—intended to span a range of durations and familiarity. Familiarity was determined based on prevalence statistics for the number of people diagnosed with that illness each year (see Table 1.1). Table 1.1 also includes the source of the clinical data used for the illness duration distributions. We first needed to determine the mean and correct form of the six illness distributions. Illness durations have been found to be well modeled by a type of distribution known as a survival function, which includes Gamma, Exponential, and Weibull. The Erlang distribution is a special case of the Gamma distribution, where α must be an integer, which is often used to model illness duration and illness stages in

Table 1.1: *Sources for Clinical Data (from least to most prevalent)*

Illness	Source of Clinical Data	Prevalence (per 10,000)
Acute (in order of prevalence)		
Appendicitis	Atema et al. (2015)	9
Seasonal Flu	Kohno et al. (2010)	1250
Common Cold	Gwaltney (1967)	2360
Chronic (in order of prevalence)		
COPD*	Oswald-Mammosser et al. (1995)	4.5
Type II Diabetes	http://www.cdc.gov/diabetes /statistics/duration/fig1.htm	860
Chronic Heart Disease	Proudfit et al. (1983)	1130

*COPD refers to chronic obstructive pulmonary disease

transmission models of infectious disease, and to infer parameters from clinical data (Krylova & Earn, 2013). For this reason, we assume Erlang is the correct distribution for the six illnesses. The clinical data provides a context to understand the mean and form of distributions and compare to participant responses (see Table 1.1 for clinical data sources).

Procedure

Participants were asked to make an estimation about the total duration of each of the six illnesses. The question read: *"Given that you meet someone with illness X, what do you think will be the total duration of their illness?"* Participants responded by typing in a number and selecting a unit of time from a dropdown menu presented on the computer screen. The experiment was performed using the Qualtrics interface. The order of presentation was randomized.

For the illnesses used in this experiment, duration can mean different things. For instance, for acute conditions, illness duration lasts from the time at onset, to the time at cure; whereas for chronic conditions, illness duration lasts from the time at onset to the time at death. In order to assess whether participants understood these important distinctions, they were also asked to categorize each illness using one of five labels: "Lasts a short time, will go away completely even without treatment", "Can vary in length,

Table 1.2: *True and estimated illness durations*

Illness	Mean Duration	Participant Response Duration	% using unit of time (correct unit is **bolded**)				
			Hrs	Days	Wks	Months	Yrs
Acute							
Appendicitis	42 hrs	471.6(SD=969.5) hrs	**8.4**	32.6	39.0	12.6	7.4
Seasonal Flu	3.3 days	8.9(SD=4.5) days	2.1	**37.9**	56.8	3.2	0
Common Cold	4.1 days	6.3(SD=3.2) days	1	**65.3**	33.6	0	0
Chronic							
COPD*	7.5 yrs	36.6(SD=22.0) years	0	1	0	5.3	**93.7**
Type II Diabetes	10.1 yrs	36.0(SD=22.5) years	0	1	0	5.3	**93.7**
Chronic Heart Disease	3.8 yrs	26.4(SD=20.0) years	0	1	2.1	2.1	**94.7**

* COPD stands for Chronic Obstructive Pulmonary Disease

requires immediate treatment, but can be cured", "Is long term, requires treatment, but can eventually be cured", "Lasts the rest of a person's lifetime, treatment can only manage symptoms, it cannot be cured, but does not necessarily cause death", "Varies in length, treatment can only manage symptoms, cannot be cured, eventually causes death". Participants were also asked several basic demographic questions (e.g., age and experience with the six illnesses) which are not analyzed here.

Results

Given that participants could respond with any unit of time, we first normalized participant responses to the unit of time for the clinical distributions. Responses were then filtered for outliers. Data was excluded in the following way: unreasonably large responses (defined as those 3 standard deviations greater than the mean response for a given illness) and participants who had more than two data points excluded based on the above criteria. The responses analyzed were 85 for appendicitis, 90 for the seasonal flu, 90 for the common cold, 90 for COPD, 90 for chronic heart disease, and 90 for type II diabetes.

First, we examined people's ability to characterize the durations of acute and

chronic illnesses. Chronic illnesses are lifelong, which is a critical difference from acute illnesses which are curable. To determine whether participants had basic knowledge of the illnesses they were making estimations about, we examined their responses to questions asking to characterize each illness. For the common acute illnesses—common cold and seasonal flu—92% of participants correctly responded that the illnesses were short term and curable. For the less common acute illness—appendicitis—81% and of participants respectively labeled these illnesses as short term. For the four chronic illnesses 74%-84% of participants correctly responded that these illnesses were lifelong. This clearly shows that people understand the chronicity of the chronic and common acute illnesses.

We then evaluated the accuracy of participant's mean responses (see Table 1.2). A qualitative evaluation of the data illustrates that participant responses were close to the mean of the clinical data for more prevalent acute illnesses (i.e., common cold and seasonal flu), and that participants overestimated the duration of chronic illnesses.

In order to assess the similarity between participant responses and the clinical data, we used a *two* one-sided t-test approach (e.g., Limentani et al., 2005). We used this approach as it allows us to test for practical equivalence (e.g., Rogers, Howard, & Vessey, 1993). A one-sample t-test might find a significant difference between a population mean of seven days and a participant response mean of eight days. While this difference is significant, it places too rigid a standard for our purposes, leading to an inaccurate conclusion that participants do not understand the mean of that illness. For this reason, we set a criterion considering accuracy to be within one standard deviation of the mean of the empirical illness distributions (standard deviations for each illness are displayed in Figure 1.1). We then conducted a t-test on either end of this threshold to determine if participant

responses were significantly greater than the lower threshold, and significantly less than the upper threshold.

We found that for the common cold, responses were within the one standard deviation of the true mean—meaning the estimates were practically equivalent to the true mean (upper threshold: Common cold: $t(89)=-6.9$. $p<.0$; lower threshold: Common cold: $t(89)=13.4$. $p<.01$). For the other five illnesses, responses were found to be greater than the lower end of the threshold, but not less than the higher end of the threshold, suggesting a pattern of overestimation, (Appendicitis: $t(84)=4.3$. $p<.01$, Seasonal flu: $t(89)=13.0$, $p<.01$, COPD: $t(89)=14.9$, $p<.01$, Type II diabetes: $t(89)=20.4$, $p<.01$, Chronic heart disease: $t(89)=20.0$. $p<.01$).

Given that participants were not within the one standard deviation threshold for five

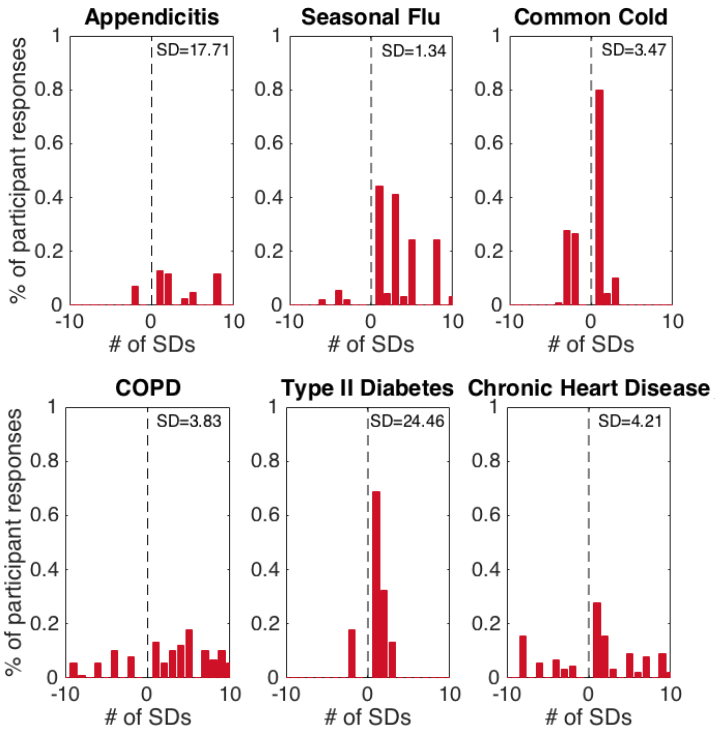

Figure 1.1: Red bars show the percentage of participants that were X number of standard deviations from the mean. Positive numbers indicate estimations above the mean, and negative numbers indicate estimations below the mean.

illnesses, we wanted to further examine participant mean responses. Therefore, we calculated the percentage of participants at each standard deviation from the mean (see Figure 1.1). For the common cold, the majority of participants (Approx. 80%) were within one standard deviation, as illustrated in the TOST. For the seasonal flu more than 70% of participants were within four standard deviations of the mean, which may seem like a large deviation from the correct response, however it is also important to note that the standard deviations varied greatly between illnesses. For the seasonal flu, the standard deviation was only 1.73 days, meaning that more than 70% of participants responded within 6.8 days of the true mean. Conversely, for the least prevalent acute illness, appendicitis, only 34% of participants were within four standard deviations of the true mean, with some participants being up to 80 standard deviations away (corresponding to 1416 hours or 59 days). This illustrates that participants had lower agreement, and mean estimations that were further from the average illness duration from the clinical data.

For the chronic conditions, fewer participants were within four standard deviations of the mean, with 31% for COPD, 100% for type II diabetes, and 61% for chronic heart disease. Participant responses were all within four standard deviations of the mean for type II diabetes because the standard deviation is 24 years.

We then examined the unit of time participants used to respond (see Table 1.2). For the acute illnesses, multiple units of time can be used to express the same value; i.e., a one week long illness can be characterized as seven days or one week. For seasonal flu and common cold, more than 80% of participants responded with either the clinical (days) or the adjacent and reasonable (weeks) unit of time. For the least prevalent acute illness—appendicitis—participants used the clinical or adjacent unit of time only 40% of the time.

For the three chronic illnesses, 92% to 95% of participants chose the clinical unit of time. The results suggest that participants could reliably use the clinical unit of time when estimating durations of prevalent acute illnesses and chronic illnesses.

Experiment 2

Participants

Forty Mechanical-Turk workers from the United States participated in exchange for $2. The participants had not participated in Experiment 1.

Materials

The same six illnesses from Experiment 1 were used. We selected four distributions as response options in the distributional form task: Erlang, Gaussian (a.k.a. Normal),

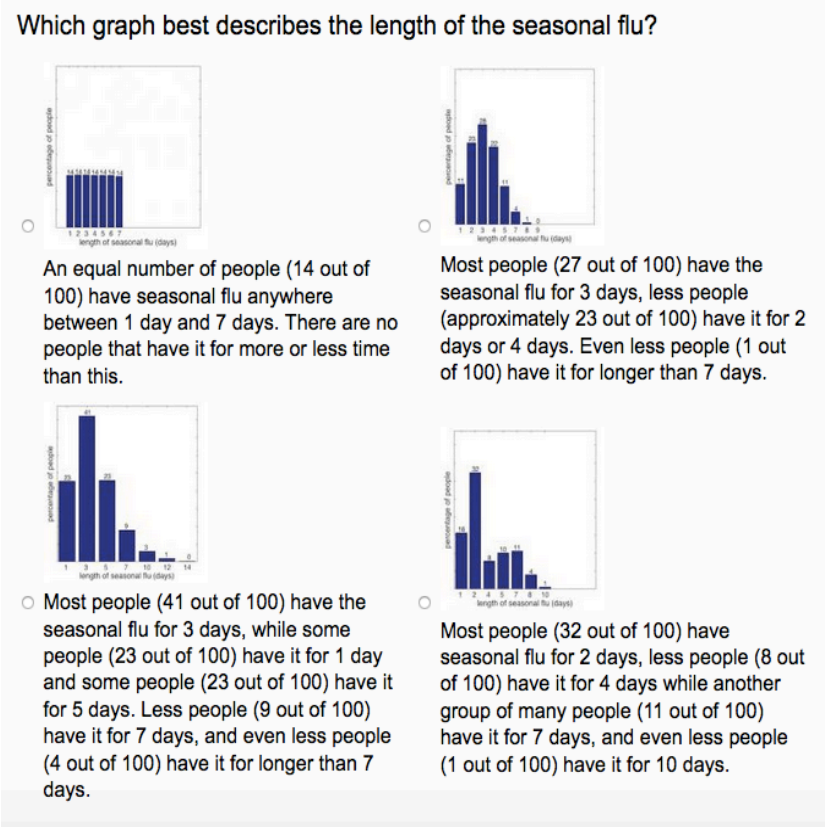

Figure 1.2: Screenshot of experimental interface for sample question (seasonal flu). Distribution types, top left to bottom right, are: Uniform, Normal, Erlang, and Bimodal.

Uniform, and Bimodal. These distributions were chosen as they can reasonably describe illness durations. The Erlang, which was always the correct answer, was chosen because illness distributions have been found to be well modeled by this distribution and provide a good fit for all the clinical distributions. Normal was chosen because the bell-curve is ubiquitous, and in some cases is very close to the Erlang distribution. This allows us to evaluate how well participants can discriminate very similar distributions. Bimodal was chosen because for chronic illnesses it might be reasonable to assume that there is one group of people who die immediately, and another group that lives with the illness for a longer time. Lastly, uniform was chosen because simple Bayesian prediction models assume a single uninformative (or uniform) prior (e.g., Gott, 1993). Selecting the uniform form of the distribution might suggest observers using a heuristic insensitive to prior beliefs.

Distributions were presented to participants as histograms of the average total duration of an illness. For each illness, the presented histograms were created by producing the best fit to the clinical data for that illness for each of the four distributions. In this way, participants' choice of distribution would be based solely on distribution form. The histograms were presented with descriptive captions. The captions for each distribution form were consistent for all illnesses. Captions described several critical points on the graph using frequencies out of 100 (see Figure 1.2). The descriptions for each distribution form were matched to illustrate the same number of points on the histogram. Four naïve raters evaluated the relationship between the descriptors and the histograms and in all cases found them to be well-matched and easily understood. The experiment was presented using the Qualtrics interface.

Procedure

Participants were first shown instructions on how to read graphs in our task. They then completed a training task, with two training sessions of four trials each. For each trial, participants were shown one histogram (illustrating one of the four distributions types used throughout this experiment) and asked to match it to one of four captions. The training trials were designed to illustrate duration without referencing illnesses. One set depicted the amount of time it takes for a person to turn into a zombie after being bitten, and the second set depicted the number of licks it takes to get to the center of a tootsie pop.

After the training task, participants were asked to choose the appropriate histogram from the four distribution options for each of the six illnesses (presented one at a time) by selecting it with a radio button. Both question and choice order were randomized.

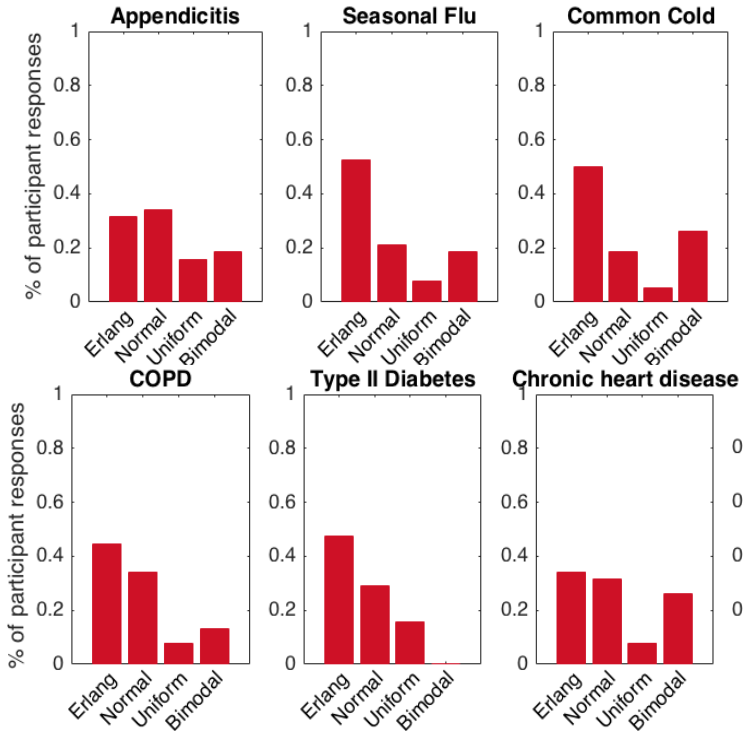

Figure 1.3: Red bars show the percentage of participants that chose a distribution choice.

Results

Data were excluded if participants answered two or more questions incorrectly in each of the two four trial training-sets. This removed two participants' data from analysis.

First, we assessed the proportion of trials for which participants chose the clinical distribution (Erlang). Participants chose Erlang 42% of the time, which was significantly greater than chance (25%), based on a one-sided Binomial test ($p<.01$). It was also chosen significantly more often than any of the other distributions: Normal $X^2(1,N=342)=11.8$, $p<.01$, Uniform $X^2(1,N=342)=93.9$, $p<.01$, and Bimodal $X^2(1,N=342)=48.0$, $p<.01$.

While participants selected Erlang with the greatest frequency overall, we were further interested in how frequently they chose the Erlang distribution for each individual illness. We performed a one-sided Binomial test and found that for four of six illnesses, participants chose the Erlang distribution at a level higher than chance (i.e., significantly more than 25% of participants): seasonal flu (53%, $p<.01$), common cold (50%, $p<.01$), COPD (45%, $p<.01$), and type II diabetes (47%, $p<.01$). Participants did not select any of the other distributions at a level higher than chance. See Figure 1.3 for the proportion of participants that chose each distribution option for the six illnesses.

Lastly, we performed a chi squared test to determine whether participants selected the Erlang distribution significantly more often than the other distribution choices. Participants chose Erlang more often than Uniform for five out of six illnesses: seasonal flu $X^2(1,N=38)=18.0$, $p<.01$, common cold $X^2(1,N=38)=19.0$, $p<.01$), COPD $X^2(1,N=38)=13.3$, $p<.01$, chronic heart disease $X^2(1,N=38)=7.9$, $p<.01$, and type II diabetes $X^2(1,N=38)=8.8$, $p<.01$.

Erlang was chosen significantly more than Bimodal for five of six illnesses: seasonal flu $X^2(1,N=38)=9.7$, $p<.01$, common cold $X^2(1,N=38)=4.5$, $p=.03$, COPD $X^2(1,N=38)=9.2$, $p<.01$, and type II diabetes $X^2(1,N=38)=23.6$, $p<.01$.

Participants chose Erlang significantly more than Normal for two out of six illnesses: seasonal flu $X^2(1,N=38)=8.1$, $p<.01$, and common cold $X^2(1,N=38)=8.4$, $p<.01$. As shown above, Erlang was chosen significantly more often than any other distribution for both seasonal flu and common cold.

General Discussion

We evaluated people's estimations for the mean and form of duration distributions within the domain of health. Examining people's representations of illness duration statistics is important, because it allows us to understand the correspondence between people's beliefs and the statistics of the environment—in this case—illness statistics. In addition, these experiments shed light on people's internal representations of real-world statistics.

When examining participants' estimates for the mean, we found that for more prevalent acute illnesses (i.e., common cold and seasonal flu), participant estimations more closely reflected clinical data for average illness durations. We also found a pattern of overestimation for chronic illnesses and less-prevalent acute illnesses.

The pattern of overestimation for chronic illnesses might be explained by people applying a probabilistic model of life expectancy to their understanding of the distribution form for illness durations. Because they have little experience with chronic illnesses, and they understand that chronic illnesses are life-long, their overestimation might be due to a strategy of applying parameters from the true distribution of lifespans (adjusted slightly to account for decreased life-expectancy with a chronic illness) to their knowledge that

illnesses follow the form of an Erlang distribution. Their ability to select the appropriate distribution form for these illnesses suggests that they can use knowledge of the form of other illness distributions even if they do not have enough experience to set the parameters. This overestimation might also be adaptive in terms of planning for the future. For chronic illnesses, it may be safer to assume a longer duration to plan sufficiently for the future, i.e., retirement savings.

When evaluating participant understanding of the form of illness duration distributions, participants show knowledge of the form of the underlying illness distributions, choosing the assumed clinical distribution (Erlang) more frequently than any other distribution. When broken down by illness, they chose the clinical distribution more frequently for the most prevalent acute illnesses: seasonal flu and common cold.

While participants often inferred the form to be the normal distribution, this may be explained by the similarity of many of the normal fits to the Erlang fits. This occurred because the normal distributions were truncated by a lower duration bound of zero. We deliberately included the Normal distribution because of the potential confusability with the clinical distribution. As such, the fact that participants still chose the clinical distribution as the correct form overall, suggests they have strong beliefs about the form of illness duration distributions and that these correspond to the environmental statistics.

A logical next step for this work would be to ask participants to independently generate distributions, rather than asking them to select from a limited number of options. Goldstein & Rothschild (2014) have shown that participants can generate these distributions when presented with data, which suggests that this method could be used to evaluate peoples' internal representations of real-world statistics.

Our results shed light on people's representations for both the form and mean of illness duration distributions. Significantly, the most prevalent acute illnesses—the common cold and seasonal flu—are also the ones for which participants consistently demonstrate knowledge of the distribution mean form that closely resembles average clinical data. Taken together, the data suggests that people may have an internal representation of illness statistics that they can consciously access.

Chapter 3: Independent Generation of Illness Statistics

This work appears in the 2018 *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Abstract

Our central question is: what are laypeople's statistical intuitions about probability distributions within the domain of health? Specifically, can participants produce entire probability distributions for the duration of illnesses? While a large body of decision making research has suggested that people use a flawed process to arrive at decisions, we posit that participants may be using an optimal process, but with flawed information. To this end, we assess expectations in terms of both the mean and form of distributions for both acute illnesses for which people might have experience, and chronic conditions for which people are less likely to have experience. We find that participants can estimate the mean and form of distributions for acute illnesses.

Introduction

What are laypeople's statistical intuitions about probability distributions within the domain of health? Decision processes are assumed to originate with a person's experience with the world, meaning that when someone makes a suboptimal decision, one of two things is at play: the person is using a *flawed process* to arrive at the answer, or the person is working with *faulty information*. In this paper, we focus on the latter: that is, what are people's prior expectations?

Biased vs. Optimal use of Expectations

Decision making research often focuses on people's apparent inability to make rational choices. People have discounted future outcomes (Koopsman, 1960) and anchored their judgments to irrelevant starting points (Tversky & Kahneman, 1974). While it has been assumed that this is due to a flawed decision process, it is also conceivable that people are working with flawed information.

While much of the Tversky and Kahneman work suggests that decision processes are flawed (e.g., 1974, 1992), there is also evidence that people use their expectations optimally (Griffiths & Tenenbaum, 2006). For example, people's predictions for life spans and movie run times are quite accurate in the aggregate. This suggests not only that judgments are optimal, but that people's expectations are consistent with real-world statistics. However, it is not clear what expectations people hold for the full probability distributions for events.

Normative Model

An alternative explanation for biased decision making is that people are using a normative model, but with flawed information. Assuming that the decision process is

rational (Bayesian), decisions are based on a combination of observed noisy data and an accurate probabilistic model of the environment (i.e., expectations). However, if those expectations are incorrect, it can lead to flawed decisions. This framework can account for flawed decisions under an optimal framework by assuming differences in prior expectations, or mapping expectations from a known domain to an unknown domain. Each time a person experiences a new event, they should update their prior probability for that event by integrating the new information. This should result in events that are experienced more often having very fine-tuned prior expectations. For those that are less commonly experienced, people might adjust their prior expectations using events for which they have more knowledge, when making inferences.

Probability Distributions Underlying Health Decisions

In this paper, we specifically investigate people's ability to produce the entire probability distribution for illness durations. There are many situations where understanding only the descriptive statistics (e.g., the mean) of a probability distribution is inadequate, and knowledge of the full probability distribution is required. Imagine you have a cough and high fever, and think you have the flu. The mean duration of the flu is 3 days, and the range is between 1 and 7 days. Additionally, there is a diminishing likelihood of the flu after 3 days. If you are applying the wrong probability distribution, you might misestimate the rate of improvement you should be expecting, i.e., the decrease after the mean. Conversely, if you have an understanding of this distribution that closely reflects the environmental statistics, and find yourself still sick after 7-10 days, you might begin to believe you have a different illness. Not only are you outside the range, but also, you have reached a point in the distribution where the likelihood of having the flu is very small. This

estimation can be critical, as illness durations outside the true distribution of durations might signal an urgent need to seek care.

Furthermore, this investigation is important in the domain of health for three reasons: (1) health decisions have been assumed to be irrational, for example, people fail to adhere to medication regimens with up to 50% non-adherence (Baroletti & Dell'Orfano, 2010), neglect preventative care (Peters, McCaul, Stefanek, & Nelson, 2006), and fail to seek care when necessary (Finnegan et al., 2000). However, it is unclear whether this is due to a flawed process or a flawed understanding of illness statistics. (2) Little work has been done to assess people's expectations for illness durations. (3) Illnesses provide a simple way to assess the normative model, as different illnesses have different degrees of experience (e.g., between acute and chronic illnesses). For instance, while you have probably personally experienced the cold many times, you may not have experienced heart disease, and therefore you would need to use a different approach when making inferences about heart disease. People may have different representations of the underlying probability distributions in cases where they do or do not have personal experience. We use this to motivate our experimental task, in which we ask participants to construct illness distributions for both acute and chronic illnesses, to evaluate how their prior expectations might differ between the two. While participants are being asked a different question about chronic illnesses (as they are evaluating time until death) previous work in this area has illustrated that people do, in fact, understand that these chronic illnesses terminate in death (See Chapter 2).

In addition to an influence of experience, there might also be individual differences in the representation of probability distributions. To measure both individual differences, and

differences between acute and chronic illnesses, we adapt this Distribution Builder of Goldstein, Johnson, & Sharpe (2008), to measure people's prior expectations for illness duration probability distributions. This paradigm has previously been used to measure people's ability to reproduce data they have recently experienced (e.g., numbers on balls in a bag), finding that people can accurately represent the mean and form of probability distributions.

In this experiment, we sought to answer the following questions: (1) how do people

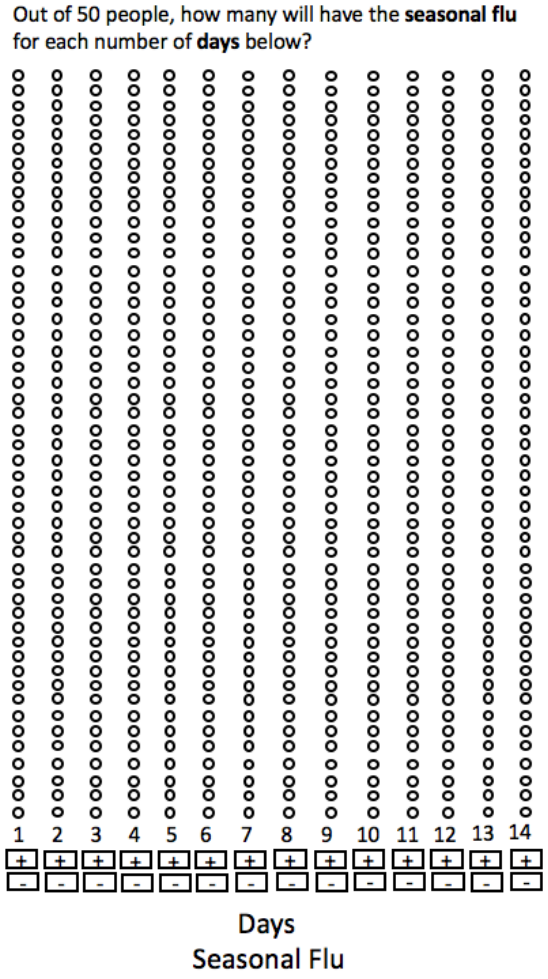

Figure 2.1: Sample distribution builder as seen by the participants. Participants could add or remove 'virtual people' from each bin (which represented an amount of time with an illness) using the plus and minus signs below that bin. Here, the circles are white because they have not been filled with 'virtual people', if the plus button is selected the empty bin is filled with a red circle.

represent the form of illness distributions? (2) how do people represent the mean of illness distributions? (3) are there differences in estimations between acute and chronic illnesses? (4) are there individual differences in the strategies people use to generate these distributions?

Experiment

Participants

Twenty Mechanical-Turk workers participated in exchange for $1 (based on the number of participants used by Goldstein et al. (2014) in the same task). The task lasted 8.75 minutes on average.

Materials

We use a variation of the Distribution Builder of Goldstein et al. (2008). See Figure 2.1. Participants were asked to indicate how many people out of fifty would have an illness for a given period of time. They were given fifty 'virtual people' to build their distribution. The number of bins in each column corresponded to the number of 'virtual people' (represented as red circles) the participants needed to place (i.e., the question was to indicate how many people out of 50 would have an illness for a particular period of time). These 50 bins allowed participants to assign all 'virtual people' to one column if they chose to.

Below each column were plus and minus buttons that could be used to add or remove 'virtual people' from each bin. Below the plus and minus signs was the unit of time, in either hours, days, or years. The columns of the distribution builder correspond to the periods of time that participants could use to respond. For each illness, we used the most common reporting unit of time and the range of available durations from the clinical

data. We chose the amount of time and number of columns to be equivalent within the chronic and acute illness categories. Each column corresponded to 12 hrs. for appendicitis (12 col.), 1 day for seasonal flu and common cold (14 col.), 1 year for COPD (18 col.), and 2 years for chronic heart disease and type-II diabetes (18 col.).

Procedure

Participants were first given instructions on how to read and understand the distribution builder (e.g., what the number of circles above the durations mean), as well as how to read a sample graph with a distribution of movie grosses. They were then randomly shown one of two check questions, to evaluate whether they understood the probability distributions. For example, they were shown a distribution of cake baking times and asked: "*The graph below shows how many of 50 cakes will bake for each amount of time (in minutes). According to this graph, how many cakes out of 50 will bake for 40 minutes?*" If they answered the first question incorrectly, they were corrected and given a second check question. If they first received the cake question, they received a question about movie run times. After these questions, participants saw task-specific instructions, explaining how they would use the distribution builder to create illness duration distributions (e.g., how to add and subtract 'virtual people' by using the plus and minus buttons). They were then given two questions to evaluate whether they read the instructions (i.e., "do you need to use all 50 people when answering a question?", and "do the units of time change between questions?").

Lastly, participants were directed to the task. For each of six illnesses, presented in random order, participants were asked "*how many people out of 50 have illness x for each*

period of time?" Participants could not continue to the next trial until all 50 'virtual people' had been assigned to bins.

Results

For each of the six illnesses we assumed a functional form of Erlang. Illness durations have been found to be well modeled by a type of distribution known as a survival function, which includes Gamma, Exponential, and Weibull. The Erlang distribution is a special case of the Gamma distribution, where α must be an integer, which is often used to model illness duration and illness stages in transmission models of infectious disease, and to infer parameters from clinical data (Krylova & Earn, 2013). See Figure 2.2 for the clinical duration distributions for the six illnesses in this experiment, with corresponding Erlang distribution fits.

We first assess participants expectations as compared to the clinical data as a whole. We calculated the fractiles for the distributions of all 6 illnesses. A fractile is defined as the value of a distribution for which some fraction of the sample lies below (e.g., the 90th fractile is the value 90% of the sample lies below). We performed a quantitative analysis of the accuracy for each of the six illnesses, for the seven key fractiles: 1st, 11th, 26th, 50th, 75th, 90th, and 100th in the same way as Goldstein et al. (2014). Figure 2.2 shows the subjective estimates as a function of normative values of the fractile, where correct answers fall on the solid black line. The figure shows that participants are more accurate for the acute illnesses, i.e., their responses lie closer to the black line than for the chronic illnesses, which show a systematic pattern of overestimation. The figure further shows that participants, on average, did not use all the available units of time for any of the illnesses, as evidenced by the fact that the 100th percentile is not the maximum available unit of time.

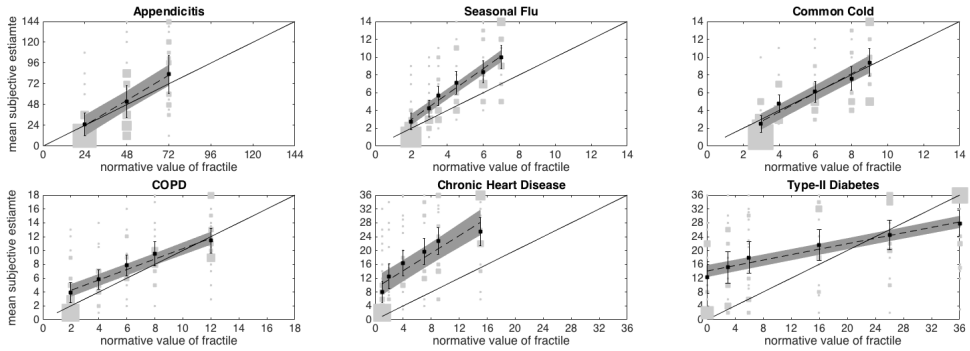

Figure 2.2: Accuracy for the 1st, 11th, 26th, 50th, 75th, 90th, and 100th fractiles. Light grey squares are individual responses, sized proportionately to number of responses. Black squares and error bars represent the mean of individual responses and standard errors for a given normative value. Dashed lines are linear trends of individual responses with standard error in dark grey. Axes are scaled for the y axis to include all responses in light grey squares. Normative 100th fractile can be read off the x axis.

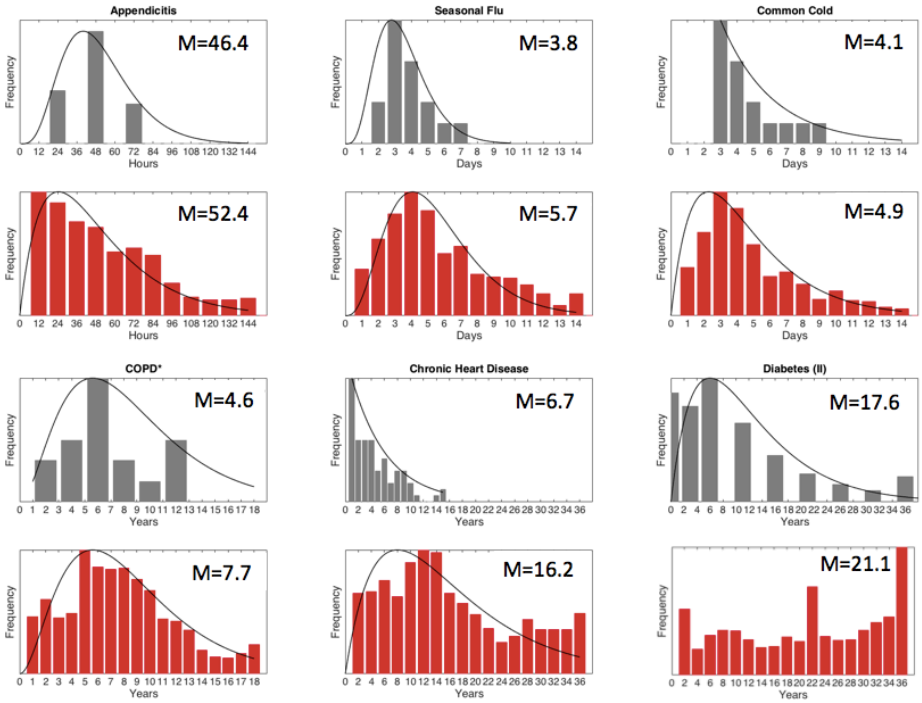

Figure 2.3: The first and third rows show histograms of clinical data for six illnesses with best fitting Erlang distributions (excluding diabetes, which could not be fit by the Erlang distribution). Grey bars show the frequency of each illness duration, black lines show the Erlang fit to clinical data. M gives the distribution mean. The second and fourth rows (red bars) show histograms of participant data displayed in the same manner as the clinical data.

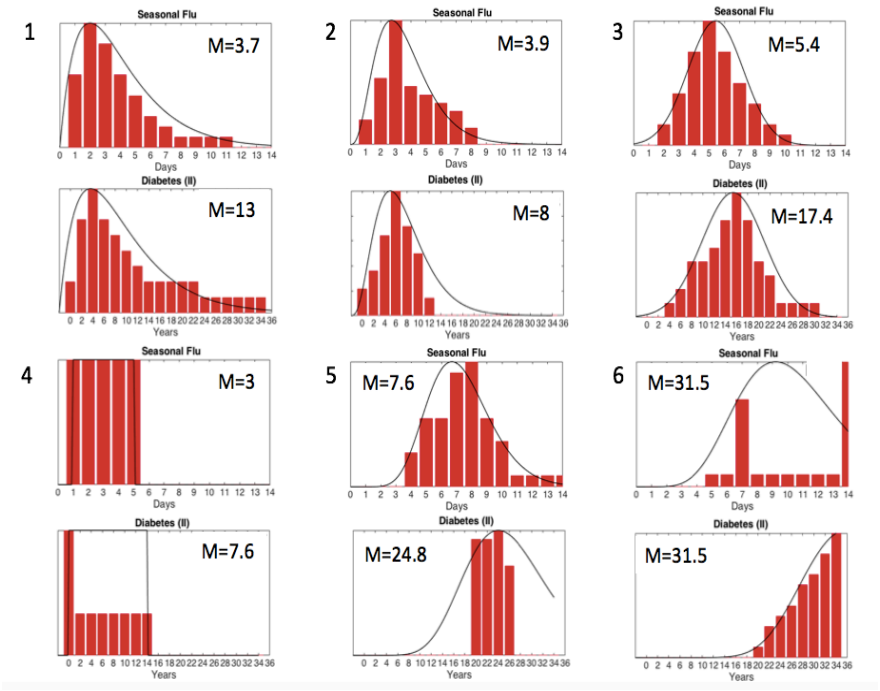

Figure 2.4: Samples of strategies used by participants in our task. Each pair of panels shows two samples, one from an acute (seasonal flu) and one from a chronic illness (type-II diabetes). See figure 2.2 for clinical data (ground truth). From top left to bottom right: 1. correctly estimate the distribution for all illnesses (2 pps.) 2. correctly estimate the distribution for acute but not chronic (6 pps.) 3. consistently use the normal distribution (3 pps.) 4. consistently use the uniform distribution (3 pps.) 5. consistently overestimate (5 pps.) 6. show no consistent pattern (1 pps.).

We then evaluated participants' ability to represent the form of the illness distributions. To compare participant responses to the true clinical data, we simply aggregated participant responses to reveal the aggregate probability distributions for each of the six illnesses (see Figure 2.3). We first performed a qualitative evaluation of whether participant responses reflected the distributional form, specifically the Erlang. For five of the six illnesses (excluding type-II diabetes) participant responses appear to be well fit by an Erlang distribution (see Figure 2.3).

To evaluate whether the Erlang distribution provided a good fit to participant data, a chi square goodness of fit test was calculated comparing the observed data to the Erlang

distributional fits. For the five illnesses for which we could calculate an Erlang fit (excluding type-II diabetes) there was no significant deviation from the Erlang distribution fits, meaning that the Erlang provided a good fit to the data. To evaluate whether another distribution might also provide a good fit, we checked whether people were using the normal distribution, as it is a common distribution in the environment, and one for which there is a standardized test. We use the Kolmogorov-Smirnov test of normality, and all distributions were found to significantly deviate from normality: *appendicitis:* D(359)=.86, $p<.001$, *seasonal flu:* D(359)=.85, $p<.001$, *common cold:* D(359)= .72, $p<.001$, *COPD:* D(359)=.76, $p<.001$, *chronic heart disease*: D(359)=.89, $p<.001$, *type-II diabetes:* D(359)=.93, $p<.001$.

Next, we sought to evaluate participant expectations for the mean of illness duration distributions. A qualitative comparison illustrates that the means calculated from participant data closely aligned with the clinical means for all the acute illnesses, while overestimating for the chronic illnesses. See Figure 2.2 for means.

To perform a quantitative evaluation of whether mean responses were accurate relative to the clinical mean, we used a *two*-one-sided t-test approach (TOST; e.g., Limentani et al., 2005). This approach allows us to test for practical equivalence (e.g., Lakens et al., 1993). A one-sample t-test might find a significant difference between a population mean of seven days and a participant response mean of eight days. This places too rigid a standard for our purposes, leading to an inaccurate conclusion that participants responses do not reflect the clinical illness distribution. Another advantage of the TOST approach is its utility for large data sets like ours (20 participants x 50 estimates) so that

the null hypothesis can be supported in situations where a one sample t-test might indicate a significant difference (Lakens, 2017).

For this reason, we set a criterion for accuracy to be two bins from the true illness duration distributions (see procedure and Figure 2.2 for bin sizes). We then conducted a t-test on either end of this threshold to determine if participant responses were significantly greater than the lower threshold, and less than the upper threshold.

Given that we showed our data is not normally distributed, to perform a t-test (which assumes normality), we log transform our data. We found that for appendicitis, seasonal flu, the common cold, and type-II diabetes, responses were within threshold of the true mean, i.e., practically equivalent to the true mean (upper threshold: appendicitis: $t(999)=25.5$, $p<.001$; seasonal flu: $t(999)=46.5$, $p<.001$; Common cold: $t(999)=19.9$, $p<.001$; type II diabetes: $t(999)=7.3$, $p<.01$; lower threshold: Appendicitis: $t(999)= -23.1$, $p<.001$; seasonal flu: $t(999)= -10.1$, $p<.01$; common cold: $t(999)= -24.5$, $p<.001$; Type II diabetes: $t(999)= -10.2$, $p<.01$). For the other two illnesses, responses were found to be greater than the lower end of the threshold, but not less than the higher end of the threshold, suggesting a pattern of overestimation, (COPD: $t(999)=41.5$. $p<.001$, chronic heart disease: $t(999)=63.5$, $p<.001$).

To examine how participants approached this task on an individual level, we examined each participant's distributions, and divided them into 6 strategies: participants that 1. correctly estimate the distribution for all illnesses (2 participants (pps.)) 2. correctly estimate the distribution for acute but not chronic illnesses (6 pps.) 3. consistently use the normal distribution (3 pps.) 4. consistently use the uniform distribution (3 pps.) 5. consistently overestimate (5 pps.) 6. show no consistent pattern (1 pp.). Figure 2.4 provides

examples of these strategies. It is important to note that for those who used a strategy of overestimation 2 out of 5 still used an approximation of the Erlang distribution.

Discussion

The primary question we sought to answer was: what are people's statistical intuitions for probability distributions in the domain of health? We found that, on average, people have mental representations of probability distributions for illness duration that closely reflect clinical data, and can produce the full probability distribution.

Recall that this investigation had four central questions, the first of which was: can people accurately reproduce the form of illness distributions? We found that for five out of the six illnesses participant data in the aggregate reflected the correct form of the distribution (see Figure 6).

Our second question was: can people accurately reproduce the mean of illness distributions? We found that for acute illnesses, participants produced a mean that closely reflected the clinical distributions, while overestimating for 2 of the 3 chronic illnesses. Importantly, we limited the range of responses for each illness, meaning participants could not overestimate as significantly as they might have, had a wider range of values been available. However, as illustrated by Figure 2.2, they appear to understand that these illnesses have a limited range, as their mean subjective estimate at the 100^{th} fractile was less than the maximum available value for all illnesses.

Our third question was, are there differences between acute and chronic illnesses? It is clear that differences exist, such that participants could produce the approximate mean and form of the clinical distribution for all three acute illnesses, but only produced the mean of one and form of two chronic illnesses. High accuracy for the distributional form of

chronic illnesses illustrates that participants used their understanding of how illness durations are generally distributed, and apply this to their understanding of illnesses they had less experience with.

Our fourth question was, are there individual differences in the strategies people use to generate these distributions? While participants used the appropriate Erlang distribution in the aggregate, we identified six strategies that participants used on an individual level. Importantly, 8 out of 20 participants used the Erlang distribution as their main strategy, which was the most popular. Some of the participants who used a strategy of overestimation also produced Erlang distributions, meaning a total of 10 participants could produce the Erlang distributional form.

Taken together, these results help to answer a central question of this investigation: when a person makes poor decisions, is the process flawed, or are the prior expectations flawed? Our results indicate that people's prior expectations are, on average, accurate for acute illnesses, but may be flawed for chronic illnesses. This result helps to inform research showing that medication adherence for chronic illnesses is worse than for acute illnesses (Baroletti & Dell'Orfano, 2010). If people are using the right process to make decisions about their health, poor short-term decisions with long term consequences for chronic illnesses may be caused by the expectations participants hold about the duration of those illnesses.

Future work should focus on how those expectations for the durations of chronic illnesses might be reduced. For instance, doctor's expectations for the knowledge of their patients are often misaligned (Street & Haidet, 2011). Doctors could use this method to understand and improve their patient's expectations. This direction is further supported by

work in which eliciting full probability distributions allowed financial planners to gain improved insight into the monetary expectations of people when planning for retirement (Goldstein et al., 2008).

The work presented here illuminates how people internally represent real-world statistics, illustrating that people can produce entire probability distributions. Eliciting these distributions can help us gain important insight into the information people are using when making decisions.

Chapter 4: Implicit Predictions for Illness Statistics

This work appears in the 2016 *Proceedings of the Annual Meeting of the Cognitive Science Society and* was presented at the 2016 *Annual Meeting of the Society for Mathematical Psychology,* and the 2016 *Annual Meeting of the Psychonomic Society.*

Abstract

People have been shown to make predictions for many real world events that closely reflect the environmental statistics. The ability to make accurate predictions might be particularly important in the domain of health, where illness knowledge directly influences patient outcomes. Therefore, we sought to investigate people's ability to make predictions for illness durations. We evaluated predictions for both acute and chronic illnesses, as judgments for chronic illnesses have been shown to be influenced by people's knowledge of acute illnesses. In two experiments, we asked participants to estimate the duration of six illnesses—three acute and three chronic—and we compared their judgments to the Bayesian optimal prediction determined from clinical distributions. For acute illnesses, people were able to estimate both the median duration and the shape of the distribution of illness durations. For chronic illnesses, people estimated the shape of the distribution, but overestimated the median duration. We discuss the possible strategies people may be employing that lead to systematic overestimation.

Introduction

Imagine that you have had a cold for a week, and need to decide if you will feel better in time for a trip beginning in two days. You are now faced with a question: How much longer will I be sick? Likewise, later in life, you could find yourself diagnosed with a chronic illness such as diabetes, and would then be faced with a new question: Given this diagnosis, what is my life expectancy? In both cases, you must make the best calculation possible to prepare for the future, and you will likely use your past experience with illnesses to make this estimation. The internal calculation must be based not only on your understanding of the average illness duration, but also on the shape of the duration distribution. In this way, as you progress further into an illness, you can adapt your duration estimate.

How people make judgments and predictions is dependent on their understanding of the statistical regularities of the world. People are well-calibrated to these regularities, and their predictions are, on average, quite accurate (e.g., Brady & Oliva, 2008; Griffiths & Tenenbaum, 2006; Hemmer & Persaud, 2014; Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Hedges, & Vevea, 2000). For example, when people were asked to make predictions about events including movie grosses and life spans, they could accurately capture both the shape and median of the distributions for these events (Griffiths & Tenenbaum, 2006). For less ubiquitous areas, such as the reign of pharaohs, people were able to capture the shape of the distribution (i.e., approximately Erlang distributed) but overestimated the duration. This is likely because they were able to apply their understanding of lifespan, but did not account for the age of mortality in the era of pharaohs. As such, it appears that people are calibrated for many events, and can apply

their expectations to those events for which they have less knowledge.

Our ability to use our understanding of the regularities of the environment to make accurate judgments and predictions is particularly important within the domain of health. Everyone faces a problem of estimating illness statistics at some point, and these internal calculations can directly impact patient health (Peters, McCaul, Stefanek, & Nelson, 2006). For instance, people's estimates of cancer risk directly influence their likelihood to receive cancer screenings (Peters et al., 2006); people who inaccurately estimate their risk are therefore unlikely to be screened regularly, increasing the chance of having a cancer go undetected.

While little is known about people's understanding of illness statistics, prominent theoretical models of illness cognition make explicit assumptions about peoples' understanding of illness statistics. For example, CSM (Leventhal et al., 1992) makes specific claims about how patients make predictions and decisions concerning their health. The CSM asserts that people construct representations of an illness based on symptoms, and that these representations guide their decisions and behavior. The CSM also argues that patients tend to apply their over-learned model for managing acute illnesses (whereby symptoms are temporary and the illness is cured with some treatment) when attempting to manage chronic illnesses. This feedback loop, in which treatment reduces symptoms (often leading to a cure), does not exist for chronic illnesses, and is suggested as a factor in low adherence for chronic illnesses within the CSM framework. However, the CSM argues that people can still use strategies such as applying their understanding for the acute illness statistics when making predictions for chronic illnesses, and health decision making more generally.

In this experiment, I sought to evaluate people's understanding of illness statistics, namely the median and shape of illness duration distributions. Motivated by the common sense hypothesis that people have a well-learned model for acute illnesses, and less defined models for chronic illnesses, we asked participants to make estimations for both acute and chronic illnesses. As such, in cases where participants have very little knowledge of an illness, the best strategy may be to use their understanding of the distribution of other illnesses for which they have more experience. We used the median of the distribution as a good estimate for illness duration, as it is the point at which the illness duration is equally likely to be longer or shorter.

In Condition 1, we tested participants' ability to capture the median and shape of both acute and chronic illness distributions by asking them to predict total illness duration based on current duration. In Condition 2, we repeated this task with older populations, to evaluate whether age played a role in these estimations, either due to increased experience, or an effect of using themselves as a reference point in estimations (e.g., assuming a later age of onset).

Modeling Approach

We followed the modeling approach of Griffiths and Tenenbaum (2006) to compare subjective performance to the optimal prediction from the clinical distributions using Bayes rule, under the assumption that people's prediction judgments follow optimal statistical principles. Bayes rule gives a principled account of how people should make predictions about the total duration of a particular illness given the duration of the illness thus far. In Equation 1, below, if d_{total} indicates the amount of time the average person experiences an illness, and d indicates the duration of the illness thus far, we can estimate d_{total} given d as

follows:

$$p(d_{\text{total}}d) \propto p(d|d_{\text{total}})p(d_{\text{total}}) \quad (1)$$

The posterior probability $p(d_{total}|d)$ is based on a combination of $p(d_{\text{total}})$, the prior probability of the total illness duration, and $p(d|d_{\text{total}})$, the likelihood of the current duration given the average total duration.

To model prediction for illness duration we obtained the clinical data for the duration of 6 illnesses. Optimal predictions from an Erlang distribution follow an approximately linear function (see Figure 3.1 bottom panels) with a slope of 1 and a non-zero intercept.

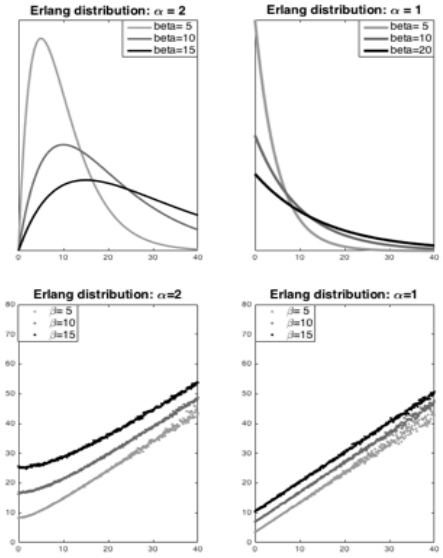

Figure 3.1: Top row shows probability density functions, bottom row shows simulated optimal prediction for Erlang distributions. Column 1 shows the Erlang distribution with α=2, which is a special case of the Gamma distribution. Column 2 shows the Erlang distribution with α=1, this reduces to an exponential distribution

As a concrete example, the common cold is approximately Erlang distributed with median of M=4.5 days. Using Equation 1, the participant's task is to calculate $p(d_{\text{total}}|d)$ for every possible cold duration for someone met on day d of their cold. If someone has had a cold for 3 days, you would expect their total duration to be around 4.5 days. Likewise, if

someone has had a cold for 6 days (which is longer than the median), you might estimate approximately 9.5 days. In this way, prediction would include both an understanding of the median duration and the Erlang distribution of duration.

Condition 1

Participants

One-hundred and eighty-eight Rutgers students participated in exchange for course credit.

*Material***s**

For each illness we sampled five data points from the distribution over duration to be used as probes in the experimental task (see Table 3.1 for illnesses, median durations and duration probes, listed in order of median duration). Following the procedure of Griffiths and Tenenbaum (2006), samples were obtained by fitting Erlang distributions to each of the six distributions. See Figures 3.2 and 3.3, top row, for the Erlang fits to the empirical illness distributions, and Table 3.1 for parameter values.

Procedure

Each participant completed one duration prediction trial for each of the six illnesses, given the duration probe. On each trial participants were asked: "*Given that you meet someone who has had illness X for time period Y, what do you think will be the total duration of their illness?"* They responded by entering a number, and choosing a unit of time from a dropdown menu (options included hours, days, weeks, months, and years). The duration probe was randomly selected from the fixed set of 5 possible probes and illness presentation order was randomized

Participants were instructed that they were being asked to predict *total* duration, not remaining duration (see Appendix A for the experimental instructions) and given a sample

question to evaluate whether they understood the instructions (see Appendix B for experimenter instructions for administering sample questions). If participants answered the test question incorrectly, the task was explained again, followed by a second sample question. If they answered the second question incorrectly, they would be excused from participating in the experiment. No participants failed the second sample question.

Table 3.1: *Illnesses and Durations*

			Presented Durations					Parameters	
Illness	Median Duration	Time Unit	T1	T2	T3	T4	T5	α	β
Acute									
Appendicitis	42	Hours	15	20	28	41	67	5	.9
Seasonal Flu	3.3	Days	1	2	3	5	7	5	.7
Common Cold	4.7	Days	2	4	5	8	20	1	3
Chronic									
COPD	7.5	Years	1	2	4	6	11	3	2.8
Chronic Heart Disease	3.8	Years	1	3	5	10	16	2	4
Type II Diabetes	10.1	Years	1	5	9	14	32	3	.4

*COPD refers to chronic obstructive pulmonary disease

One hundred and thirteen participants received the experimental procedure as described above. Seventy-five participants received three additional questions asking about their personal experience, however this data is not presented here.

Results

The following data was excluded from analysis: data points smaller than the presented duration ($d_{total}<d$), unreasonably large responses (defined as those 3 standard deviations greater than the median response for a given illness duration probe), participants who responded using negative numbers, and participants who had more than two data points excluded based on the above criteria. The responses analyzed were 162 for appendicitis, 171 for the seasonal flu, 170 for the common cold, 170 for COPD, 173 for

chronic heart disease, and 170 for type II diabetes.

To evaluate whether people's predictions for illness durations captured the median and shape of the real-world distribution, we first calculated optimal predictions from the Erlang prior: $d^* = d + \beta \log 2$, where d^* is the predicted value of d_{total} and d is the duration probe (see Griffiths and Tenenbaum, 2006 appendix for the derivation of the prediction equation). Figures 3.2 and 3.3, second row, show participant predictions for total durations given the duration probe with optimal predictions calculated from the Erlang distributions, as well as best-fitting Erlang predictions to participant data.

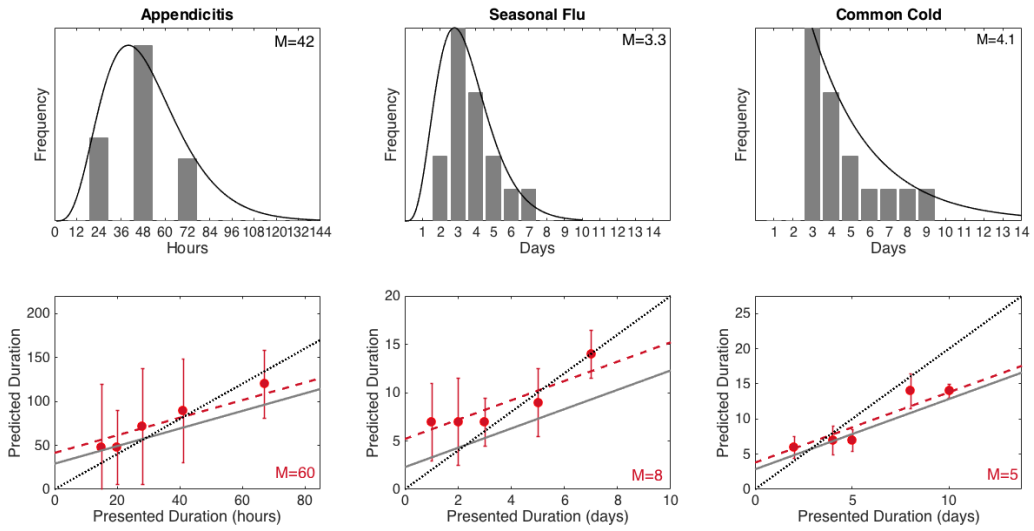

Figure 3.2: The top row shows real world distributions for the durations of the three acute illnesses and corresponding Erlang distribution fits. The second row shows participant predictions for illness duration in Experiment 1. Red circles show the median predicted duration as a function of presented duration, with error bars indicating the 68% confidence interval (estimated by a 1000 sample bootstrap). The red dashed line is the fits from the Erlang prior to participant responses and the gray line shows the Bayesian optimal prediction, and the black dotted line illustrates an uninformative prior.

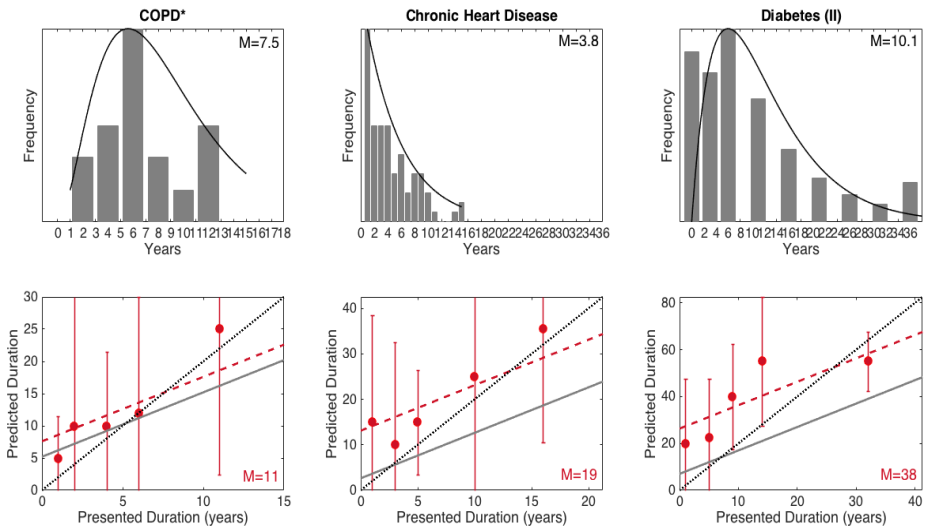

Figure 3.3: The top row shows real world distributions for the durations of the three chronic illnesses and corresponding Erlang distribution fits. The second row shows participant predictions for illness duration in Experiment 1. Red circles show the median predicted duration as a function of presented duration, with error bars indicating the 68% confidence interval

A qualitative comparison for the three acute illnesses suggested that the best fitting predictions to the participant data was relatively close to the Bayesian optimal prediction for the clinical distributions. This can be seen by the closeness of the red line to the grey line, as well as the closeness of the respective medians (see Figure 3.2 row 2). In this way, participants' predictions for acute conditions were consistent both with the median and shape (following a linear trend with a slope of 1) of the assumed Erlang distribution of the empirical data. For chronic conditions (see Figure 3.3 row 2) participants overestimated the median, while they still captured the form of the distributions. It is also important to note that for all illnesses where participants did not accurately estimate the median, the pattern was to systematically overestimate the duration, a result elaborated on in the conclusions.

In order to quantitatively evaluate the difference between participant predictions

and the Bayesian optimal prediction, we performed a bootstrap analysis. For each illness, we drew a random sample with replacement from participant responses for each duration probe. We then used these samples to fit the optimal Erlang prediction (as we did for the red dashed line in Figures 9 and 10), recording the median. This procedure was repeated 1000 times, resulting in 1000 medians for each illness. Finally, we computed the bootstrap 95 percentile confidence interval. If the accurate median fell within that confidence interval, the bootstrapped samples were all practically equivalent to the true median. For confidence intervals see Table 3.2.

The bootstrap analysis found that the clinical median fell within the bootstrap confidence interval for 2 out of 3 acute illnesses (i.e., appendicitis and the common cold), and 1 of the 3 chronic illnesses. For the chronic illnesses, it is important to note that the confidence intervals were very large, suggesting low agreement within subjects. This pattern was also the case for appendicitis, which is the least prevalent acute illness. This illustrates that participant predictions closely reflected the clinical data for the common acute illnesses overall, but did not when making predictions for the chronic illnesses. While

Table 3.2: *Bootstrap 95 percentile confidence intervals for means fit to participant data as compared to means for the real world distribution*

		Experiment 1 Confidence Interval		**Experiment 2** Confidence Interval	
Illness	True Median	Lower	Upper	Lower	Upper
Acute					
Appendicitis	42	29.7	654.7	16.2	324.0
Seasonal Flu	3.3	4.0	13.9	3.8	14.4
Common Cold	4.1	2.6	8.9	2.9	9.8
Chronic					
COPD	7.5	4.6	47.9	7.8	49.3
Chronic Heart Disease	8.9	9.5	57.1	11.3	54.5
Type II Diabetes	10.1	18.5	73.0	21.4	67.5

*COPD refers to chronic obstructive pulmonary disease

the clinical median for the seasonal flu fell outside the confidence interval for participant responses, participants may have been using a prior expectation for the duration of the common cold when making these estimations, causing them to overestimate the duration.

Condition 2

In condition 1 we found that participants were generally able to capture the shape of illness distributions, and more closely captured the median for acute than chronic conditions. Very few participants had personal experience or familiarity with the chronic illnesses which may be a result of the sample being drawn from college students ranging in age from 18-24. An older population might have more experience, and thus, their predictions might be closer to the optimal prediction for chronic conditions. It is also possible that an older population would assume a later age of onset than younger participants, who may be using themselves as a reference point. Estimating a later age of onset might lead to lower estimations of total duration, making them closer to the clinical duration. Therefore, in condition 2 we sought to examine prediction from an older participant sample. In this experiment we used the same experimental paradigm with participants on Mechanical Turk who were aged 40 or older.

Participants

One hundred and thirty-five Mechanical Turk workers aged 40 or older from the United States were paid $1 for their participation.

Procedure

Both the materials and procedure were identical to that of condition 1.

Results

Data in this experiment was analyzed using the same exclusion criteria from

condition 1. The responses analyzed were 112 for appendicitis, 120 for the seasonal flu, 114 for the common cold, 119 for COPD, 117 for chronic heart disease, and 116 for type II diabetes.

In the same manner as condition 1, we calculated optimal predictions as well as the best fitting Erlang prediction to the observed participant data (see Figures 3.4 and 3.5). The most striking result is illustrated by comparing results to those in condition 1. A qualitative comparison of the best-fitting predictions to the data relative to the Bayesian optimal prediction revealed that participant performance in this task closely paralleled that of condition 1, as reflected in the similarity of the median estimations, and the shape of the predictions fitting the Erlang prediction function.

For the quantitative assessment, we replicated the bootstrap procedure from condition 1, and calculated the bootstrap 95 percentile confidence intervals for the responses of the older adults (see Table 3.2). When comparing the results to the true distributions, we found that the true median fell within the confidence interval for 2 out of 3 acute illnesses (i.e., appendicitis, and the common cold) and none of the 3 chronic illnesses.

Using the bootstrap samples, we also compared the medians between experiments. The confidence intervals for both conditions overlapped for all 6 illnesses, illustrating that the two groups responses were practically equivalent to one another. These results indicate that overall, older participants did not perform differently than college aged participants.

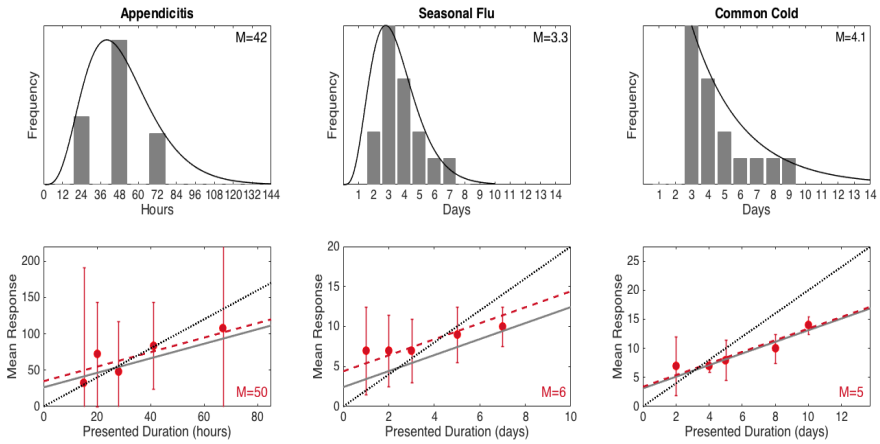

Figure 3.4: The top row shows real world distributions for the durations of the three acute illnesses and corresponding Erlang distribution fits. The second row shows participant predictions for illness duration in Experiment 1. Red circles show the median predicted duration as a function of presented duration, with error bars indicating the 68% confidence interval (estimated by a 1000 sample bootstrap). The red dashed line is the fits from the Erlang prior to participant responses, the gray line shows the Bayesian optimal prediction, and the black dotted line illustrates an uninformative prior.

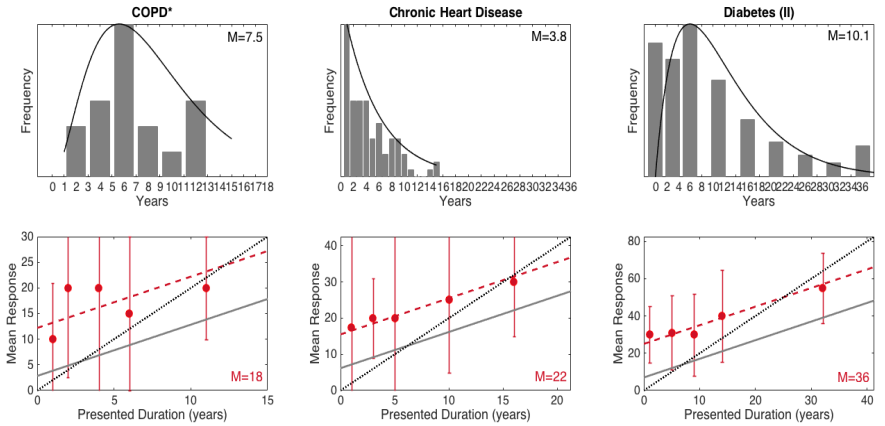

Figure 3.5: The top row shows real world distributions for the durations of the three chronic illnesses and corresponding Erlang distribution fits. The second row shows participant predictions for illness duration in Experiment 1. Red circles show the median predicted duration as a function of presented duration, with error bars indicating the 68% confidence interval (estimated by a 1000 sample bootstrap). The red dashed line is the fits from the Erlang prior to participant responses, the gray line shows the Bayesian optimal prediction, and the black dotted line illustrates an uninformative prior.

Discussion

In this paper, we applied the paradigm of Griffiths and Tenenbaum (2006) assessing optimal prediction for everyday events to the domain of health. We measured how people make predictions about illness durations and compared performance for acute and chronic conditions. The data show that participant responses closely matched the optimal predictions for both the form and median of the illness distributions for acute conditions, with near perfect performance for the common cold. While previous research has suggested that health decisions operate differently from other decision processes (Levy et al., 2014), we show that for acute illnesses for which people have experience, participants follow optimal statistical principles and have understanding of the regularities of illness distributions.

Furthermore, for chronic conditions the data show that while responses follow the form of the clinical distribution, the median durations are systematically overestimated. This is in line with the common sense model (Leventhal et al., 1992) which suggests that people should be able to apply their understanding of acute illnesses to judgments about chronic illnesses, but that a lack of experience with chronic illnesses might also lead to misalignment when applying the acute model (e.g., overestimating duration).

A strategy of overestimation might be adaptive in terms of planning for the future (whether that be short or long term). Recall the opening scenario where you were asked to imagine that you had a cold for a week, and needed to predict if you would be feeling better in time for a trip beginning in two days. For other illnesses, where you might be unsure of the duration, how would you make an estimation of when you are likely to recover? You might take an illness you understood better, such as a cold, and adjust upward to ensure

yourself an adequate recovery time. The same may be true for chronic illnesses. When planning for the future (e.g., retirement savings), it may be safer to assume a longer duration. Indeed, when planning for the future, it may be safer to overestimate the duration of an illness rather than risk underestimating the duration.

In addition, optimism about lifespan for chronic illnesses may be important for positive health behaviors. People who report higher levels of optimism about their condition report being less bothered by symptoms (Scheier & Carver, 1985) and show faster recovery from surgery (Scheier, Owens, Magovern, Leferebve, Abbot, & Carver, 1989). For this reason, it might be advantageous to overestimate the duration of chronic illnesses. This could signal optimism, which might, in turn, help patients to engage in behaviors that are good for their health, and remain healthier longer.

A critical feature of chronic illnesses that might make prediction for total duration more complex is that by definition persons with chronic conditions have not yet experienced the duration of that illness in its entirety, and therefore do not have knowledge of the total duration. There is evidence that successful predictions require not just some experience in a domain, but a relevant amount of experience. For instance, when asked to estimate the duration of bus routes, participants systematically underestimated the durations (Stephens, Dunn, Rao, & Li, 2015). The authors posited that this was because bus riders rarely complete a journey through an entire bus route, and rather only know the length of their typical journey. This may explain why participant performance seemed to improve for younger participants who had personally experienced the seasonal flu, but not for those who had experienced chronic heart disease. Participants who have experienced

the flu have experienced it in its entirety, and therefore have some firsthand knowledge of its duration.

Understanding illness duration information has important implications for health decision making (McAndrew et al., 2008). People's understanding of illness duration is directly linked to their health decisions, and ultimately to their health care seeking behavior. For example, if you attribute your symptoms to the common cold, but still find yourself sick after three weeks, you may re-evaluate your illness assignation. Furthermore, accurate understanding of illness statistics impacts patient doctor communication. Doctors often have misaligned expectations of their patients' illness knowledge (Street & Haidet, 2011), incorrectly believing that their patients have knowledge more closely matching their own. This causes poor communication about illnesses and treatment, and ultimately affects patient health decision making leading to low adherence to treatment regimens.

The significance of the work presented here is both in its novelty—to our knowledge this is the first investigation assessing people's judgments for illness statistics—and in its importance in understanding people's ability to make optimal statistical judgments. The findings extend our knowledge of how people make judgment about everyday events to health-based decisions. As such, provides an important step in understanding how people reason about illnesses and illness outcomes, and it provides a foundation for future investigations into patient judgments and decisions.

Chapter 5: Information Integration and Judgment Change for Health

This work was presented at the 2018 *Annual Meeting of the Psychonomic Society,* and the 2019 *International Convention of Psychological Science.*

Abstract

How is new information integrated into existing expectations? This question has important implications for how people make judgments and decisions as they gather new evidence from various sources. The current investigation is focused on understanding the influence that both the domain the problem is presented in and the source of incoming evidence have on repeated judgments. Previous work has suggested that the best strategy is to simply take an average of all the available evidence, while people tend to be egocentric, weighting their own judgments more strongly than those of others (e.g., Yaniv & Kleinberger, 2000). However, it may be the case that judgments can be approximated by a weighted confidence model which assumes that judgments are a weighted combination of prior expectations and third-party evidence. In the following experiments, participants are asked to judge the likelihoods of different problems given a set of symptoms both before and after receiving evidence from either 1 or 2 outside sources. Results show a main effect of source and an interaction between source and domain. Participant responses in the two judgment tasks are also found to be well approximated by the weighted confidence model.

Introduction

An important area in both cognition and health research deals with how new information is integrated into existing expectations when people are asked to make repeated judgments under uncertainty. There are many factors that have been shown to be important when making a judgment given information from a third party. For instance, people have been shown to form opinions about the quality of advice for different advisors (Yaniv & Kleinberger, 2000) and are sensitive to the number of cues that are available to each advisor (Budescu, Rantilla, Yu, & Karelitz, 2003).

In the domain of health, this issue often arises when making a judgment about what illness you are likely to have given particular symptoms. In this case, your initial judgment is based on your prior expectations for what different illnesses tend to look like. Each time you receive a new piece of information (e.g., information from WebMD), you are likely to update this initial judgment to reflect the new information. In this way, the source of information can be an important predictor of judgment change. More specifically, sources that are trusted more may prompt a greater change in judgment. Indeed, the credibility of a source has been shown to positively influence judgments such that people rely more on sources they believe are credible (e.g., Birnbaum & Stegner, 1979).

The following experiments focus on three sources of information that people often rely on when seeking healthcare information: doctors, WebMD, and past illness experience of the person in question. More than 90% of people express at least some level of trust in their doctors (Hall, Dugan, Zheng, & Mishra, 2001) and this trust is a strong predictor of medication adherence (Piette, Heisler, Krein, & Kerr, 2005). Online resources also appear to have a significant influence, with 70% of people saying they were prepared to act upon

information obtained on the web (Silence, Briggs, Harris, & Fishwick, 2007); however, this may be problematic, as symptom checkers have low accuracy, with the correct diagnosis being listed within the top 20 diagnoses only 58% of the time (Semigram, Linder, Gidengil, & Mehrotra, 2015). Lastly, past experience with illness is an important component of how people treat and manage their illnesses (e.g., McAndrew et al., 2008; Leventhal et al., 1992).

One factor that might interact with the source of new information is the domain the problem is presented in. For instance, health judgments have been shown to differ from judgments in other domains. Specifically, health numeracy is worse than numeracy in pure math or finance (Levy et al., 2014). More broadly, decision processes often differ based on the domain the problem is presented in. For instance, in the case of the Wason Card Selection task (Wason, 1960), participants performed significantly better when the problem was presented in a real word domain (Cosmides, 1989). With this in mind, I will be comparing judgments across three domains: health, laptops, and cars. I chose these domains because the task in each is identical: to judge the likelihood of a problem given particular symptoms. Additionally, participants in each domain are given information from an authority figure, an online resource, and an evaluation based on the past experience of the person experiencing the problem.

The following three experiments endeavor to answer several questions about how health information is integrated. The first experiment will focus on two questions: (1) Is the source of incoming information an important predictor of judgment change? (2) Are judgments made in the domain of health updated differently from judgments made in other domains? These questions drive the experimental paradigm, which first measures people's

expectations by asking them to make a judgment about the likelihood of a problem given symptoms, and then measures the change in their judgment after new evidence is presented. This evidence was presented in the form of (1) an authority figure (e.g., doctor or mechanic); (2) an online resource (e.g., WebMD or AutoMD); or (3) an evaluation made by the person they are being asked about, based on that person's experience.

Experiment 2 replicates the methods from Experiment 1, with the important difference that I asked participants to rate their confidence in each of the sources. These confidence ratings allow me to inform the weighted confidence model which assumes that participants weight advice based on their confidence in the source of that advice. I implement and compare three models, including a simple weighted confidence model, an egocentric discounting model, and a model which assumes that people simply take the average of source and personal judgments. While averaging opinions has been suggested as the normative model of how people should make judgments—and people do indeed use this strategy in some cases (e.g., Anderson, 1981; Fishcer & Harvey, 1999)—this strategy is only optimal when all pieces of information are equally informative. In most cases, however, not every piece of information is equally informative. Past work has suggested that rather than averaging opinions, people often discount advice egocentrically—meaning, they underweight advice from others relative to their own judgments (e.g., Yaniv & Kleinberger, 2000). However, those experiments have not measured people's confidence both before and after receiving advice. For this reason, they cannot determine whether participant's weighting of their own advice is appropriate given their level of confidence in their own judgment. It may be the case that what appears to be egocentric discounting is more accurately described as greater confidence in personal judgments than source

information. Importantly, the weighted confidence model assumes that initial judgments and source information are combined using a weighting structure based on a person's confidence in their own judgment, as well as their confidence in the source.

In the third and final experiment, I ask participants about the likelihood of different problems given particular symptoms, and then provide them with information from two sources, asking participants to make a judgment three times—(1) one initial judgment based on symptoms; (2) one judgment after receiving information from the first source; and (3) a final judgment after receiving information from the second source. This task allows for the evaluation of the model with multiple pieces of information. Previous research has shown that the order of information has been found to influence decision-making with the last piece of information being weighted more strongly (Bergus, Chapman, Levy, Ely, & Opplinger, 1998). This experiment expands on those investigations by asking how differing levels of confidence in the source might interact with order effects.

Experiment 1

*Participant***s**

Fifty-nine Rutgers students participated in exchange for course credit.

Material

For this experiment, we presented problems from three domains: health, car, and laptops. See tables 4.1-4.3 for a list of problems and symptoms. For the health domain, symptom sets were taken from Epocrates.com, an online database of vignettes created by physicians. For each symptom set, an alternate illness (one with similar symptoms) was chosen by entering the symptoms into WebMD. As such, participants were either presented with the true cause of the symptoms or an alternate reasonable explanation of symptoms.

For the car and laptop domains, car problems were chosen to span a range of severity and familiarity.

Procedure

Participants were asked to answer questions about the probability of a health, car, or laptop problem given symptoms. For instance, "*Your friend Jane (30 years old) has been experiencing a fever, cough, headache, and weakness. She asks for your opinion, how likely do you think it is that she has sinusitis?*" They would then judge the likelihood on a scale from 0 to 100, to two decimal places. Additionally, participants were asked to rate their confidence in their estimation on a scale from 1 to 5. They were asked questions about all 6 symptom sets in each of the three domains. The questions were grouped by domain, and participants received the questions and domains in a random order. For each symptom set, they were randomly presented with a possible cause of their symptoms (either the true problem or the alternate problem). After their initial response, they were then provided with input from a third party. For instance: "*Jane went to the doctor. The doctor thinks Jane does have sinusitis. Jane asks you again, how likely do you think it is that she has*

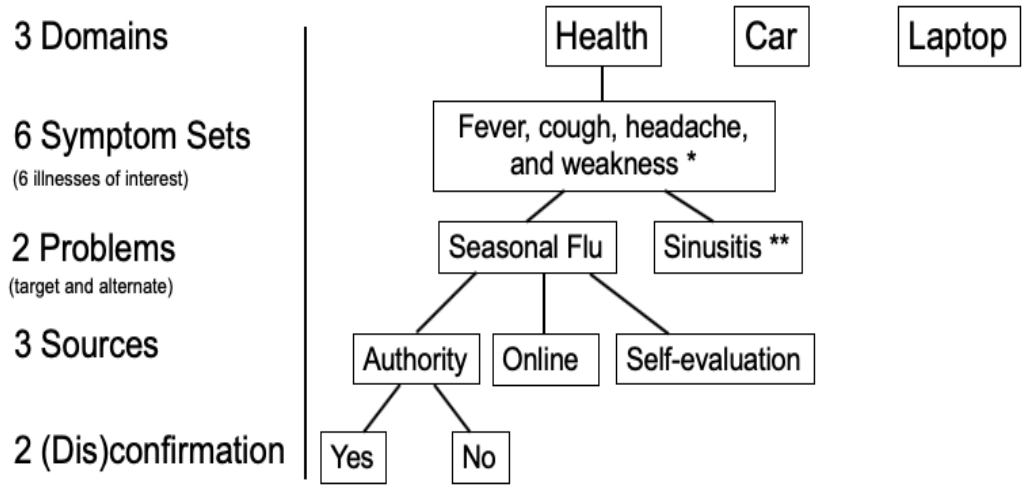

Figure 4.1: A visual representation of the procedure for Experiments 1 and 2.

Table 4.1: *Problems and Associated Symptoms for Health Domain*

True Illness	Alternate illness	Symptoms
Appendicitis	Gastroenteritis	abdominal pain moving from the middle to lower right stomach, nausea, and a low-grade fever
Bacterial Meningitis	Gastroenteritis	a severe headache, fever, light sensitivity, and a stiff neck
Seasonal Flu	Sinusitis	a fever, cough, headache, and weakness
Stroke	Alzheimer's	blurred vision, fatigue, dry skin, frequent urination, and increased thirst
Asthma	Generalized Anxiety Disorder	shortness of breath, wheezing, and waking from sleep from wheezing
Type-II Diabetes	Urinary Tract Infection	blurred vision, fatigue, dry skin, frequent urination, and increased thirst

sinusitis?". They then responded using the same scale for probability and confidence. For each domain there were three available third parties and participants randomly received feedback from either an authority (i.e., a doctor, mechanic, or laptop specialist), an online resource (i.e., WebMD, AutoMD, or LaptopMD), or the persons past experience (i.e., that the person in question had experienced the problem before and had an intuition). Each authority could either confirm or disconfirm the suggested problem. In each domain, participants received each 3rd party source and yes/no combination, exactly once. See figure 4.1 for a visual representation of the procedure.

Table 4.2: *Problems and Associated Symptoms for Car Domain*

True Problem	Alternate Problem	Symptoms
Dead Battery	Bad Starter	an inability to start, slow engine crank, and a lit-up check engine light
Faulty Spark Plugs	Bad Ignition Component	hard starts, trouble starting, engine misfires, high fuel consumption, and a lack of acceleration
Bad Brake Pads/Rotors	Loose Wheel Bearing	loud squeaking or squealing, a vibrating steering wheel and brake pedal and needing to press down hard to brake
Failing Gas Cap	Fuel Filter Leak	a lit-up check engine light, fuel smell, and the gas cap will not tighten all the way
Failing Transmission	Coolant Leak	leaking fluid, a strange smell, and a delay in acceleration
Overheating	Blown Head Gasket	steam pouring out of the hood, high temperature gauge, and a weird smell from engine

Table 4.3: *Problems and Associated Symptoms for Laptop Domain*

True Problem	Alternate Problem	Symptoms
Bad Computer Fan	Damaged Hard Drive	weird noises, overheating, and error messages
Battery Dying	Dirty Air Vents	reduced charge capacity, overheating, and sudden shutdowns
Hard Drive Failure	Loose Screws	slow speed, frequent freezing, blue screen, corrupted data, and weird sounds
Computer Virus	Disk Cache Overload	slowdown, pop-ups, computer crashing, new homepage, and programs starting without warning
Memory/RAM failure	Hard Drive Failure	slowdown, random restarts and freezing, blue screen, corrupted files, and problems installing software
Graphics Card Failure	CPU Failure	slow animation, pictures looking wrong, wavy lines, fuzzy picture, and black screen

Results

The measurement used to evaluate judgment change in this task is the relative proportion of change. To illustrate how this was calculated, imagine the participant made an initial estimate of 30% and then after receiving the source information estimated 60%. Their judgment change is 30, and the total possible change they could have made is 70. The relative proportion of change is calculated as the judgment change over the total possible change.

A three-way ANOVA revealed no significant main effect of domain (see figure 4.2). There was a significant main effect of source, $F(2,1053)=41.64$, $p<.001$ (see figure 4.3).

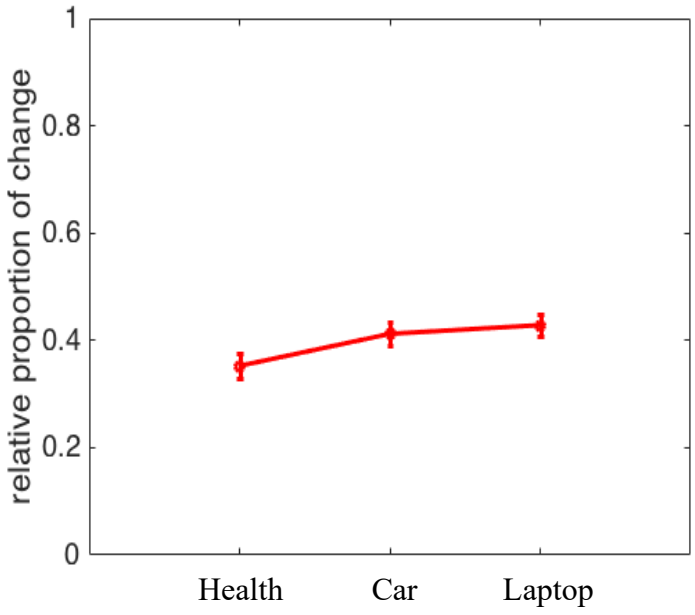

Figure 4.2: The plot shows the relative proportion of change for each of the three domains.

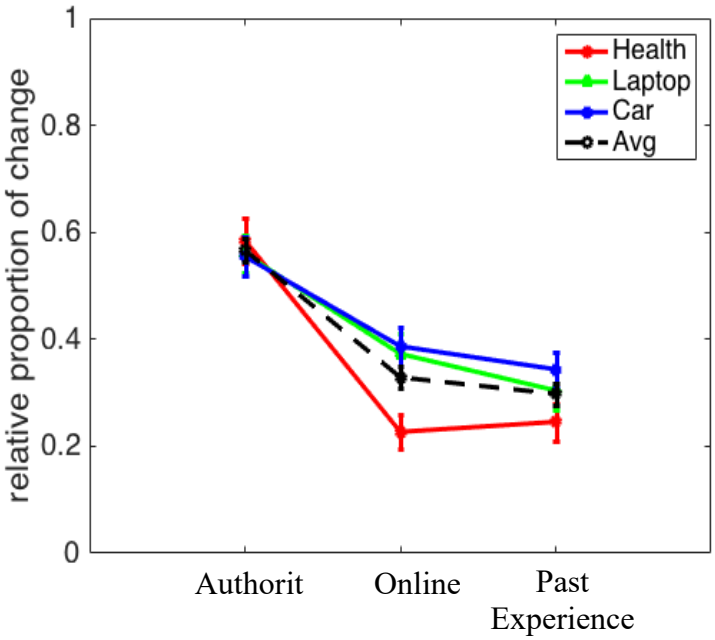

Figure 4.3: The plot shows the relative proportion of change for each of the three sources, separated by domain. The red dots represent health domain, the green dots represent the laptop domain, and the blue dots represent the car domain. Lastly, the black dots represent the relative proportion of change for each source, averaged across all domains.

A post hoc Tukey test showed that the authority and online sources differed significantly at $p < .05$, and the authority and past experience sources differed significantly at $p < .05$. Lastly, there was a significant interaction between source and domain $F(4,1053)=2.3$, $p=.05$.

I also examined the influence of source and domain on confidence in judgements. For each repeated judgment a difference in confidence is measured as confidence 2-confidence 1. A two-way ANOVA revealed no significant main effect of domain. There was a significant main effect of source, $F(2,1053)=20.5$, $p<.001$. A post hoc Tukey test showed that the authority and online sources differed significantly at $p < .05$, and the authority and past experience sources differed significantly at $p < .05$. Lastly, there was a significant interaction between source and domain $F(4,1053)=4.43$, $p<.01$.

Experiment 2

Participants

Fifty-nine Mechanical-Turk workers from the United States participated in exchange for $1 each.

Materials

The materials were identical to those in Experiment 1 with the removal of all questions in the domain of laptops. These questions were removed as Experiment 1 illustrated that the car and laptop domains yielded identical results.

Procedure

The procedure was identical to that of Experiment 1, with the important difference that participants were asked two additional questions to assess their expectations about each of the sources. First, they were asked how confident they were in the advice provided by that source. Second, they were asked two questions about the sources own confidence in their opinion. For instance "When the doctor says that they think a person does have an illness, what do they think the probability of that illness is?" and "When the doctor says that they think a person does NOT have an illness, what do they think the probability of that illness is?". Lastly, the confidence scale was changed to 1-10 (rather than 1-5 as in Experiment 1). This was to provide participants with a wider range of confidence values.

Results

I first replicated the results of Experiment 1. A two-way ANOVA revealed no significant main effect of domain. However, there was a significant main effect of source, $F(2,714)=15.12$, $p<.001$ (see figure 4.4). A post hoc Tukey test showed that the authority and online sources differed significantly at $p < .05$, and the authority and past experience

sources differed significantly at p < .05. Lastly, there was a significant interaction between source and domain $F(4,714)=5.19$, $p<.01$.

Next, I examined the influence of source and domain on confidence in judgements. For each repeated judgment a difference in confidence is measured as confidence 2 - confidence 1. A three-way ANOVA revealed no significant main effect of domain. There was a significant main effect of source, $F(2,708)=14.2$, $p<.001$. A post hoc Tukey test

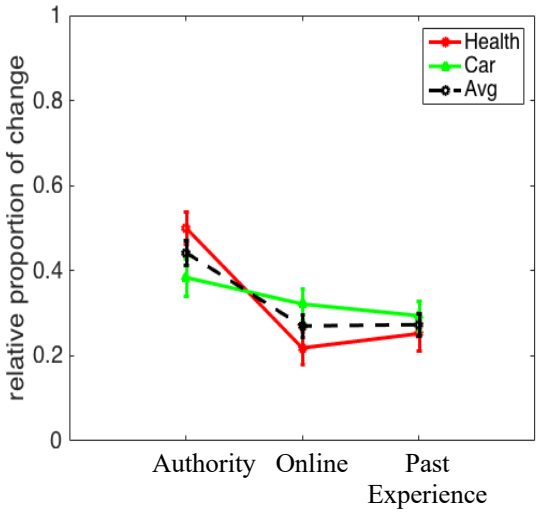

Figure 4.4: The plot shows the relative proportion of change for each of the three sources, separated by domain. The red dots represent health domain and the green dots represent the car domain. Lastly, the black dots represent the relative proportion of change for each source, averaged across all domains.

showed that the authority and online sources differed significantly at p < .05, and the authority and past experience sources differed significantly at p < .05. Lastly, there was a significant interaction between source and domain $F(4,708)=11.04$, $p<.001$.

I also evaluated participant confidence in each of the three sources. As a reminder, each participant was asked to rate their confidence in each of the three sources out of 10. A two-way ANOVA revealed no significant main effect of domain. There was a significant main effect of source, $F(2,354)=66.49$, $p<.001$. A post hoc Tukey test showed that the

authority and online sources differed significantly at $p < .05$, and the authority and past experience sources differed significantly at $p < .05$. There was no significant interaction between domain and source.

Lastly, I evaluated what probability participants thought each source was assigned to its *own* judgment. A two-way ANOVA revealed no significant main effect of domain. There was a significant main effect of source, $F(2,708)=7.6$, $p<.001$. A post hoc Tukey test showed that the authority and online sources differed significantly at $p < .05$, and the authority and past experience sources differed significantly at $p < .05$. This illustrates that participants thought that an authority who confirmed a diagnosis was much more confidence in their judgment than an online resources or a person speaking from experience. Lastly, there was no significant interaction between domain and source. See table 4.4 for means.

Table 4.4: *Means and standard deviations of source probabilities*

	Authority	Online	Past experience
Yes	81.0 (16.0)	66.8 (18.2)	67.9 (16.3)
No	41.4 (32.6)	39.9 (22.2)	42.4 (22.0)

Modeling

To investigate the observed patterns of participant judgments, I compared three generative models that make conflicting assumptions about how people integrate evidence when making judgment: (1) a simple weighted confidence model; (2) an egocentric discounting model; and (3) a model which assumes that people simply take the average of source and personal judgments. I chose these models because the averaging model has been described as a normative model for how people *should* integrate advice, while the egocentric model as a descriptive model of how people *do* integrate advice (e.g., Yaniv & Kleinberger, 2000).

For the following section, I will be referring to the source judgment as SJ, the participants initial judgment as IJ, the participants confidence in their initial judgment as IC, and their confidence in the source judgment as SC. Recall from the procedure that for each source, participants were asked how likely that source thought the probability of a problem was when they said yes, and how likely they thought a problem was when they said no. These responses were used as the source judgment in each of the models. In the case of the averaging model, the data was generated by calculating: Final Judgment=(0.5*IJ) + (0.5*SJ). For the egocentric model, we assume that people are weighting their own advice 70% against the source judgment, as previous research has suggested that people tend to weight their own estimates at this rate (Yaniv & Kleinberger, 2000), such that Final Judgment=(0.7*IJ) + (0.3*SJ). The weighted confidence model takes into account their confidence in their initial judgment, and their confidence in the source's judgment, such that Final Judgment=(w*IJ) + ((1-w)*SJ). W was calculated as 1/((SC/IC)+1). In order to evaluate which model provided the best fit to the data, I first calculated difference scores between the participant data and the model predictions (see figure 4.5 for error distributions). I then calculated the mean absolute deviation (MAD). A one-way ANOVA revealed no significant differences in the MAD between the three models.

While each model performed equally well when examining error in the aggregate, it is important to examine which of the models can provide a fit to the *patterns* of participant data. Figure 4.7 shows the simulated data for each of the models. I performed a three-way ANOVA on the simulated model data, in the same way as for the data in experiments 1 and 2. For the averaging model, there was no significant main effect of

domain or source on relative proportion of change. There was an interaction between source and domain, $F(2,714)=3.62$, $p<.05$. For the egocentric model, there was no significant main effect of domain or source. There was a significant interaction between source and domain, $F(2,714)=3.62$, $p<.01$. For the weighted confidence model, there was no significant main effect of domain. There was a significant main effect of source, $F(2,714)=5.88$, $p<.001$, and a significant interaction between source and domain, $F(2,714)=3.54$, $p<.05$.

I computed the log likelihood of the data under each of the models for the relative proportion of change. The summed log likelihood for the averaging model was -1030.9, -3569.2 for the egocentric model, and -816.7 for the weighted confidence model. This illustrates that with the lowest log likelihood, the weighted confidence model provides a better fit to the data.

To examine individual participants, I rank-ordered the models in terms of their fit for each participant. Figure 4.7 shows the distribution of ranks assigned to each model

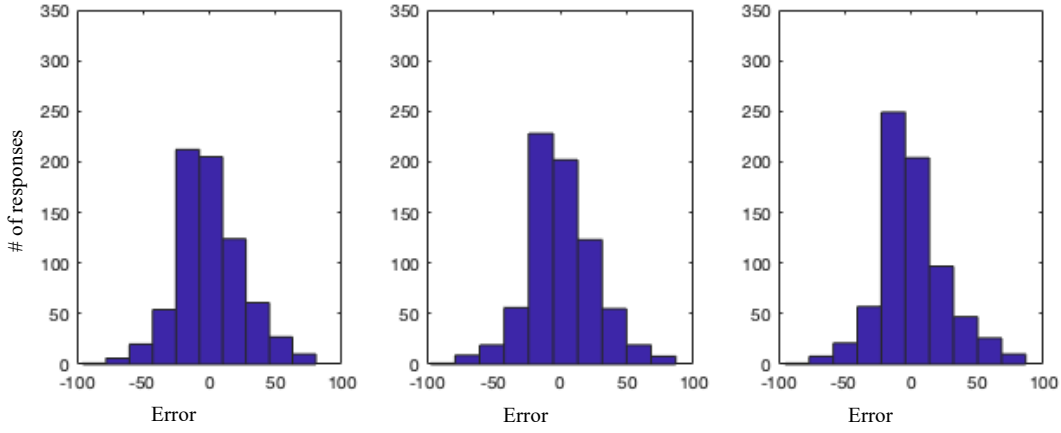

Figure 4.5: histograms show the distribution of errors from the three models (calculated as model prediction-participant response). The first panel shows the averaging model, the second shows the egocentric model, and the third shows the Bayesian model. Values above zero indicate that the model overestimated, while values below zero indicate that the model underestimated.

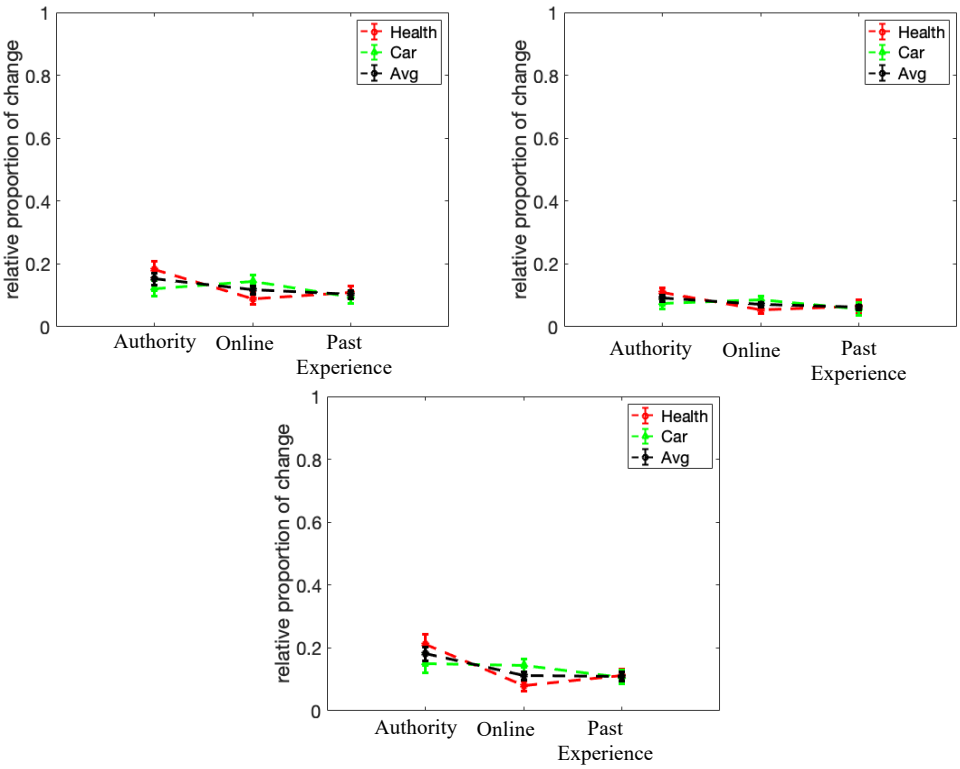

Figure 4.6: The plots show the relative proportion of change for each of the three sources, separated by domain. The red dots represent health domain and the green dots represent the car domain. Lastly, the black dots represent the relative proportion of change for each source, averaged across all domains. Starting on the left, the first plot illustrates the results of the averaging model, the second plot illustrates the results of the egocentric model, and the last illustrates the results of the Bayesian model.

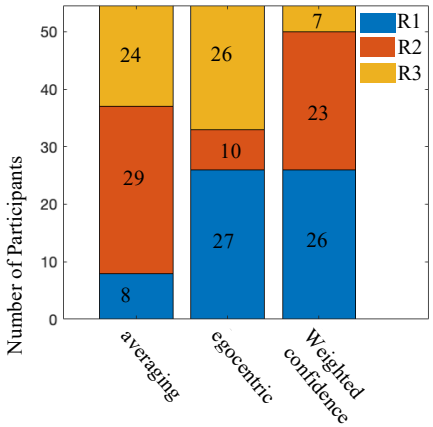

Figure 4.7: Distribution of closeness of fit of three models to participant data.

across the participants in our task; the weighted confidence model fit best for 26 out of 60 participants, and it was ranked either first or second in 50 of 60 participants.

Experiment 3

Participants

One-hundred and eighty Mechanical-Turk workers from the United States participated in exchange for $2.

Materials

The materials were identical to those from Experiment 2, with the addition of two problems in both the health and car domains. This was in the interest of accommodating the procedure (see procedure). See table 4.5 for a list of added problems and associated symptoms.

Table 4.5: *Problems and Associated Symptoms for Laptop Domain*

True Problem	Alternate Problem	Symptoms
Car		
Oil Leak	leaking valve cover	dark puddles under car, smoke from engine, dashboard oil light, engine overheating, smell of burning oil
Bad Alignment	loose suspension component	vibration, wheels pulling to one side, crooked steering wheel
Health		
Mononucleosis	Strep throat	a fever, sore throat, and fatigue
COPD	Bronchitis	shortness of breath, a chronic cough, yellow mucus

Procedure

The procedure in this experiment was identical to that of Experiment 2, with the addition of a second piece of source information. In this experiment, participants were asked to judge the probability of each problem three times: (1) after they were given the

symptom set; (2) after a source initially provided information; and (3) after a second source provided information. Participants were divided into three conditions, each of which only included two of the sources: authority and online resource; authority and past experience; or online resource and past experience. For each symptom set they received one of four combinations: source 1 confirmed the diagnosis and source 2 also confirmed the diagnosis; source 1 confirmed the diagnosis and source 2 refuted the diagnosis; source 1 refuted the diagnosis and source 2 confirmed the diagnosis; or source 1 refuted the diagnosis and source 2 also refuted the diagnosis.

Results

First, I assessed the proportion of trials for which participants chose to side with the first or last piece of evidence. For this assessment, I only included the trials for which the two sources did not agree. Overall, participants chose the last piece of evidence more frequently than the first, $X^2(1,N=1438)=15.1$, $p<.001$. When dividing the data by source, the last piece of evidence was chosen more frequently than the first for the authority

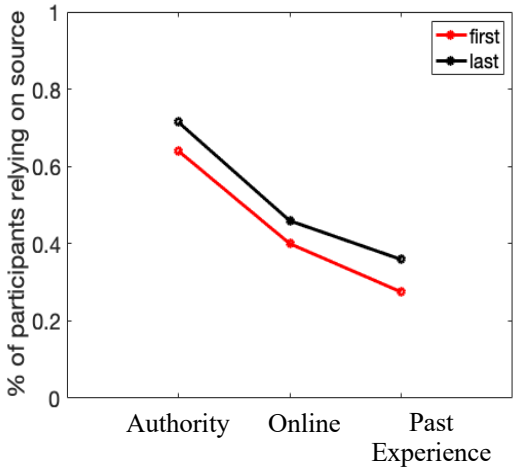

Figure 4.8: Percentage of participants relying on the first or last piece of evidence, separated by source.

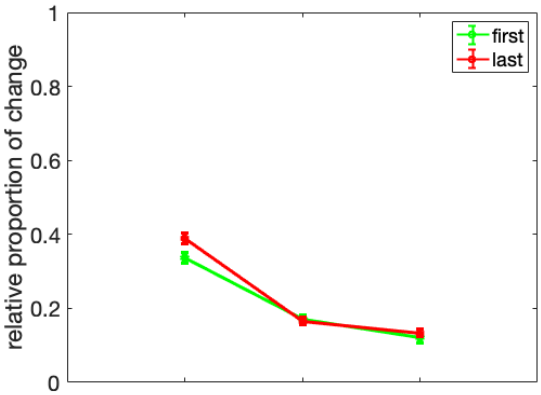

Figure 4.9: The plot shows the relative proportion of change for each of the three sources, separated by first and last judgments. The green dots represent first judgments and the red dots represent the last judgments.

X^2(1,N=1438)=5.7, p<.05, and past experience X^2(1,N=1438)=7.4, p<.01, but not for the online resource (see figure 4.8). This suggests that certain sources may not be more influential when placed last.

Additionally, the source of the information influenced which piece of evidence the participant sided with. Participants sided with the authority more frequently than the online resource X^2(1,N=1438)=83.2, p<.001, and past experience X^2(1,N=1438)=185.7, p<.001. They also sided with the online resource more frequently than the past experience X^2(1,N=1438) =21.3, p<.001.

A three-way ANOVA revealed no significant main effect of domain. However, there was a significant main effect of source, F(2,5748)=231.28, p<.001 (see figure 4.9). A post hoc Tukey test showed that the authority and online sources differed significantly at $p < .05$, and the authority and past experience sources differed significantly at $p < .05$. There was also a significant main effect of whether it was the first or last judgment F(1,5748)=4.36, p<.01. Lastly, there was a significant interaction between the domain and

source $F(1,5748)=9.82$, $p<.001$ and between whether it was the first or last judgment and the source $F(2,5748)=3.13$, $p=.05$.

I also evaluated participant confidence in each of the three sources. As a reminder, each participant was asked to rate their confidence in each of the three sources out of 10. A two-way ANOVA revealed no significant main effect of domain. There was a significant main effect of source, $F(2,714)=104.9$, $p<.001$. A post hoc Tukey test showed that the authority and online sources differed significantly at $p < .05$, and the authority and past experience sources differed significantly at $p < .05$. There was no significant interaction between domain and source.

Lastly, I evaluated the probability judgments that participants thought each source had assigned to its *own* judgment. A two-way ANOVA revealed no significant main effect of domain. There was a significant main effect of source, $F(2,1434)=5.61$, $p<.001$. A post hoc Tukey test showed that the authority and online sources differed significantly at $p < .05$, and the authority and past experience sources differed significantly at $p < .05$. Lastly, there was no significant interaction between domain and source. See table 4.6 for means.

Table 4.6: *Means and standard deviations of source probabilities*

	Authority	Online	Past experience
Yes	79.4 (17.4)	63.2 (20.9)	67.8 (16.8)
No	45.4 (32.1)	39.9 (22.2)	48.0 (22.3)

Modeling

For each of the models, the second judgment was calculated in the same way as in Experiment 2, so I will now discuss how the third and final judgment was calculated. In the case of the averaging model, the data was generated by calculating: Final Judgment=(IJ+SJ1+SJ2)/3. For the egocentric model, we continued to assume that people

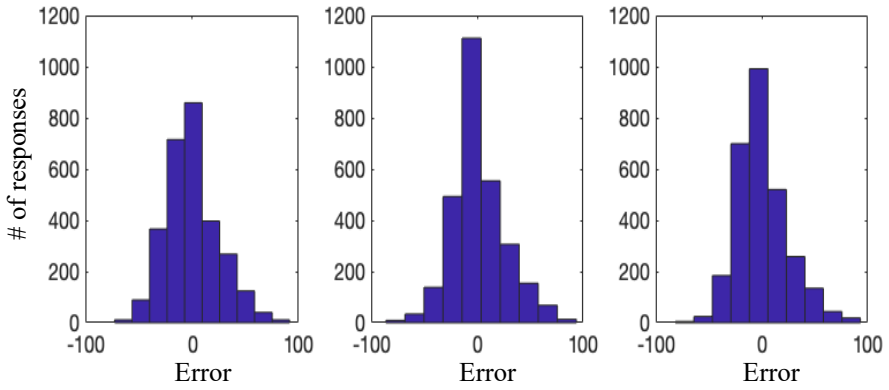

Figure 4.10: histograms show the distribution of errors from the three models (calculated as model prediction-participant response). The first panel shows the averaging model, the second shows the egocentric model, and the third shows the weighted confidence model. Values above zero indicate that the model overestimated, while values below zero indicate that the model underestimated.

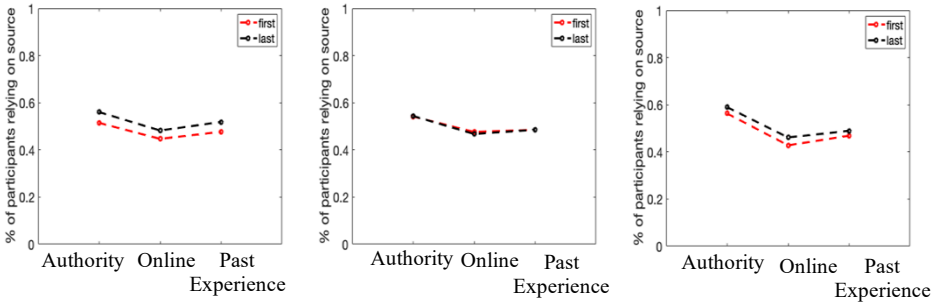

Figure 4.11: Percentage of participants relying on the first or last piece of evidence, separated by source for the three models: averaging, egocentric, and weighted confidence (from left to right).

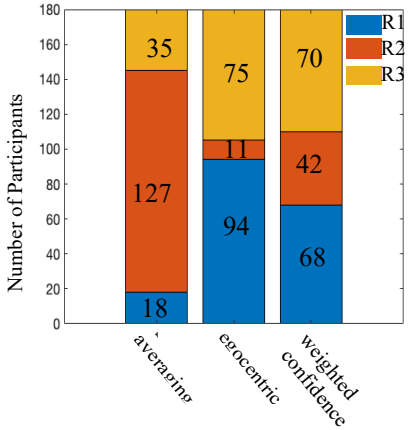

Figure 4.12: Distribution of closeness of fit of three models to participant data.

are weighting their own advice 70% against the source judgments, such that Final Judgment=(0.7*IJ) + (0.3*(SJ1+SJ2/2)). The weighted confidence model takes into into account participants' confidence in their initial judgment, and their confidence in the source's judgment, such that Final Judgment=(w*J2) + ((1-w)*SJ). J2 was calculated in the same way as Experiment 2. W was calculated as 1/((SC/judgment 2 confidence)+1).

In order to evaluate which model provided the best fit to the data, I first calculated difference scores between the participant data and the model predictions for participants final judgment (see figure 4.10 for error distributions). I then calculated the mean absolute deviation (MAD). A one-way ANOVA revealed no significant differences in the MAD between the three models.

While each model performed equally well when examining error in the aggregate, it is important to examine which of the models can provide a fit to the *patterns* of participant data. First, for each of the models I assessed the proportion of trials for which the simulations chose to side with the first or last piece of evidence. For this assessment, I only included the trials for which the two sources did not agree. For the averaging model, the last piece of evidence was chosen more frequently than the first, $X^2(1,N=1438)=5.0$, $p<.05$. When dividing the data by source, I did not find any significant difference between the percentage of participants choosing the first or last judgment (see figure 4.11). For the egocentric model, the first piece of evidence was chosen more frequently than the last, $X^2(1,N=1438)=5.0$, $p<.05$. When dividing the data by source, I did not find any significant difference between the percentage of participants choosing the first or last judgment. For the weighted confidence model, the first and last piece of evidence were chosen equally frequently overall.

Participant results showed that the source of the information influenced which piece of evidence the participant sided with. I repeated this calculation for each of the three models.

For the averaging model the data sided with the authority more frequently than the online resource $X^2(1,N=1438)=6.8, p<.01$, and past experience $X^2(1,N=1438)=4.2, p<.05$. For the egocentric model the data sided with the authority more frequently than the online resource $X^2(1,N=1438)=6.8, p<.01$, and past experience $X^2(1,N=1438)=4.2, p<.05$. For the weighted confidence model, participants the data with the authority more frequently than the online resource $X^2(1,N=1438)=24.2, p<.001$, and past experience $X^2(1,N=1438)=12.9, p<.001$.

I computed the log likelihood of the data under each of the models for the relative proportion of change. The summed log likelihood for the averaging model was -8850, -19,463 for the egocentric model, and -2814 for the weighted confidence model. This illustrates that the weighted confidence model still provides the best fit to the relative proportion of change in this task.

In order to examine individual participants, I rank-ordered the three models in terms of their fit for each participant. Figure 4.12 shows the distribution of ranks assigned to each model across the participants in my task. While it appears that very few participants had data that most closely matched the averaging model, 68 participants' data ranked the weighted confidence model first, and 94 participants' data ranked the egocentric model first. These results suggest that participants as a whole may not have used a consistent strategy.

General Discussion

There were four central questions that this chapter focused on answering: (1) Is the source of incoming information an important predictor of judgment change? (2) Are judgments made in the domain of health updated differently than judgments made in other domains? (3) Can repeated judgments be approximated by a rational model? (4) How might the order of information interact with the source of information?

First, I found the source of information is an important predictor of judgment change, with the greatest overall judgment change found in relation to the authority. This is not necessarily surprising, given that previous research has shown that people are influenced by the expertise of advisors when making judgments (e.g., Birnbaum & Stegner, 1979). However, it is an important indication that any model which does not explicitly account for confidence in the source of information will not adequately describe the data. This is also reflected not only in participants' confidence in each of the sources, but also in their estimations of what the source thought the probability was. Participants felt that doctors who confirmed a particular illness assigned a fairly high probability (around 80%), while online resources or people with past experience assigned a lower probability (around 66%). This suggests that not only do people have different levels of confidence in different sources, but they assign different probabilities to the words "yes" and "no" depending on who is saying it.

The lack of a main effect for domain (for either judgment change or change in confidence), illustrates that the domain the problem was presented in was not an important predictor of how participants responded in our task. While previous work has found an influence of domain (e.g., Levy et al., 2014; Cosmides, 1989), it is possible that the

important difference in our task is the introduction of uncertainty. While in the tasks described above, participants were shown to perform better in one domain than another, there is no "better" in this task. It would appear from the results above that the mechanism underlying how judgments change with new information remains consistent when making judgments under uncertainty.

While there was no main effect of domain, there was an interaction between domain and source, such that the online resource was trusted significantly less in the domain of health. This is important to note, as previous work demonstrated that people are willing to act upon information obtained from online health resources (Silence, Briggs, Harris, & Fishwick, 2007). It may be the case that while people rely on online resources to make decisions about whether or not to seek care, these resources do not significantly influence their internal calculation of the likelihood of a given illness. For instance, if WebMD tells you that you may be having a stroke, this may not have a large influence on the probability you assign to the likelihood that you have a stroke, but may still lead you to rush to the hospital. In this case, while you are still assigning a low probability to the stroke, the potential cost of not seeking care may be large enough to overcome it. For instance, in emergency departments, physicians often use the rule out worst-case scenario strategy, in which they first evaluate hypotheses for the most serious illnesses, to begin treatment (Croskerry, 2002). This suggests that the cognitive process for arriving at a decision to seek care and the process for judging the likelihood of an illness may be related, but distinct.

The modeling data from Experiment 2 provides some support for the idea that participants use a weighted combination of the source information and their prior expectations to make judgments. While all three models do equally well in the aggregate,

only the weighted confidence model can predict all the patterns seen in the data. Finally, the data from Experiment 3 illustrate that there may be an important interaction between order effects and the source of information. Taken together, the data in this Chapter contribute to our understanding of how judgments change with new evidence.

Chapter 6: Conclusion

This dissertation is centered around expanding our understanding of what expectations for illness statistics look like and how they influence judgment and prediction. These expectations have important consequences. For instance, people often do not seek care for a heart attack because they have misaligned expectations about their symptoms, believing them to be gastrointestinal rather than heart-related, and believing them to be too mild to be consistent with a heart attack (Bunde & Martin, 2006). This misalignment in expectations has been attributed to popular media, as many people expect a heart attack to present the way it does in the movies—i.e., a sharp crushing pain that appears instantaneously—whereas symptoms are often more mild, and present more gradually (Finnegan et al., 2006). This pattern is consistent across multiple chronic illnesses, as 75% of stroke patients could not correctly identify their symptoms as indicating a stroke. Importantly, patients who had previously experienced a stroke were more likely to have accurate expectations about their symptoms (Williams, Bruno, Rouch, & Marriott, 1997), another indication that experience matters. Deciding whether to seek care from a doctor can be broken down into three stages: (1) the perception of symptoms, which leads to prediction about whether those symptoms suggest an illness; (2) a decision about whether that illness requires care from a doctor; and (3) a prediction about whether the benefits of seeking care outweigh the costs (Bunde & Martin, 2006).

This dissertation focuses on stage 1, assessing how people mentally represent illness information. Chapter 2 illustrates that people generally agree on the mean and form of the duration distributions for acute illnesses, but do not have as fine-grained prior expectations for chronic illnesses. This supports the idea that illness experience may

influence prior expectations, causing illnesses that are experienced often to have more fine-grained expectations. This is consistent with patient health data, which illustrates that chronic illnesses are mismanaged at a higher rate than acute illnesses (e.g., Davis, Wagner, & Grovers, 2000; Wagner et al., 2001), with as much as 50% non-adherence (WHO, 2003). For instance, according to the American Heart Association, patient adherence to medication after a heart attack reduces with time (Ho, Bryson, & Rumsfeld, 2009). Diminishing adherence suggests that misaligned expectations for the duration and proper treatment of chronic illnesses and leads to negative consequences, such as increased mortality (Baroletti and Dell'Orfano, 2013).

In Chapter 3, we see that not only do people have expectations for the descriptive statistics of illness duration distributions for acute illnesses, but they can also produce these distributions as a whole. These distributions were more closely aligned with clinical data for acute than chronic illnesses. This illustrates that people do have a representation of illness statistics (whether or not this representation closely aligns with clinical data), and that they can consciously access this information. In Chapter 4, participants make predictions that reflect the environmental statistics for more common illnesses, and importantly, reflect the expectations that they illustrated in Chapters 2-3. This demonstrates that participants are in fact able to use their expectations to make predictions about the future.

In Chapter 5, we see that these expectations change over time as participants are presented with new information. The source of information plays an important role, with information from authorities (such as a doctor or mechanic) prompting greater judgment change. The pattern of changes we observed were best approximated by a weighted

confidence model; weighted confidence models provide a principled account both of how we update our expectations given observed data, and in turn, how these expectations learned from experience influence cognitive processes. When participants combine their prior expectations with new evidence, they assign weights based on their confidence in their own judgment, and their confidence in the source. Chapter 5 also illustrates that repeated judgments can be approximated by a weighted confidence model, and that effects of the order of information may interact with the source of information.

The question posed in Chapter 1 of this dissertation was as follows: When people make bad judgments or decisions, are they using a *flawed process*, or are they working with *faulty information*? The research in this dissertation suggests that the process by which people make judgments may be optimal—people weigh information more strongly when it comes from a source they deem reliable. However, it may also be the case that people are working with flawed information, as we saw people's expectations for chronic illnesses were less consistent than their expectations for acute illnesses, suggesting that flawed information, rather than a flawed process that contributes to poor judgments.

The overarching message of the work discussed in this dissertation is that people form prior expectations for illness statistics, and that these prior expectations are combined optimally with evidence when making judgments in the domain of health. This expanded understanding of what people's expectations for illnesses are and how they change over time may help to improve our understanding of patient health decisions. As doctors' knowledge of their patient's expectations is often poor (Street & Haidet, 2011), further work should focus on how patients' expectations regarding the likelihood of an illness might influence health decisions

References

Anderson, N. H. (1981). *Information integration theory.* Hillsdale, NJ: Lawrence Earlbaum.

Atema, J. J., Gans, S. L., Beene, L. F., Toorenvliet, B. R., Laurell, H., Stoker, J. & Boermeester, M. A. (2015). Accuracy of white blood cell count and c-reactive protein levels related to duration of symptoms in patients suspected of acute appendicitis. *Academic Emergency Medicine,* 1015-1024.

Baroletti, S., & Dell'Orfano, H. (2010). Medication adherence in cardiovascular disease. *Journal of the American Heart Association*, *121*, 1455-1458.

Bergus, G. R., Chapman, G. B., Levy, B. T., Ely, J. W., & Opplinger, R. A. (1998). Clinical diagnosis and order of information. *Medical Decision Making, 18,* 412-417.

Birnbaum, M. H. & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology, 37(1),* 48-74.

Brady, T. F., & Oliva, A. (2008). Statistical Learning Using Real-World Scenes. *Psychological Science*, *19*(7), 678– 685.

Budescu, D. V., Rantilla, A. K., Yu, H., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes, 90,* 178-194.

Bunde, J. & Martin, R. (2006). Depression and prehospital delay in the context of myocardial infarction. *Psychosomatic Medicine, 68,* 51-57.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31,* 187-276.

Croskerry, P. (2002). Achieving quality in clinical decision making: Cognitive strategies and detection of bias. *Quality in Clinical Decision Making, 9(11),* 1184-1204.

Davis, R. M., Wagner, E. G., & Groves, T. (2000). Patients as partners in managing chronic disease. *Evaluation*, *320*, 537-40.

Finnegan, J.R., Meischke, H., Zapka, J. G., Leviton, L., Meshack, A.… Stone, E. (2000). Patient delay in seeking care for heart attack symptoms: Findings from focus groups conducted in five U.S. regions. *Preventative Medicine, 31,* 205-213.

Fischer, I. & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform a simple average? *International Journal of Forecasting, 15,* 227-246.

Goldstein, D. G., Johnson, E. J., & Sharpe, W. F. (2008). Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *Journal of Consumer Research*, *35*, 440–456.

Goldstein, D. G. & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making, 9,* 1-14.

Gott, J.R. (1993). Implications of the Copernican principle for our future prospects. Nature, *363*, 315–319.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767–773.

Gwaltney, J. (1967). Rhinovirus infections in an industrial population: II. Characteristics of illness and antibody response. JAMA, 202, 494-500.

Hall, M. A., Dugan, E., Zheng, B., & Mishra, A. K. (2001). Trust in physicians and medical institutions: What is it, can it be measured, and does it matter? *Milbank Quarterly, 79*, 613-639.

Hemmer, P., & Persaud, K. (2014). Interaction between categorical knowledge and episodic memory across domains. *Frontiers in Psychology*, *5*(June), 584.

Ho, P. M., Bryson, C. L., & Rumsfeld, J. S. (2009). Medication adherence: Its importance in cardiovascular outcomes. *Journal of the American Heart Association*, *119*, 3028-30353.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: prototype effects in estimating spatial location. *Psychological Review*, *98*, 352–76.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*, 220–241.

Kohno, S., Kida, H., Mizuguchi, M., & Shimada, J. (2010). Efficacy and Safety of Intravenous Peramivir for Treatment of Seasonal Influenza Virus Infection. *Antimicrobial Agents and Chemotherapy, 54,* 4568-4574.

Koopsman, T. (1960). Stationary ordinal utility and impatience. *Econometrica, 19,* 287-309.

Krylova, O., & Earn, D.J.D. (2013). Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. *Journal of the Royal Society*, 1-14.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8,* 355-362.

Leventhal, H., Diefenbach, M., & Leventhal, E. A. (1992). Illness cognition: Using common sense to understand treatment adherence and affect cognition interactions. *Cognitive Therapy and Research*, *16*, 143–163.

Levy, H., Ubel, P. A., Dillard, A. J., Weir, D. R., & Fagerlin, A. (2014). Health Numeracy: The Importance of Domain in Assessing Numeracy. *Medical Decision Making*, *34*, 107–115.

Limentani, G.B., Ring, M.C., Ye, F., Bergquist, M.L., McSorely, E. O. (2005) Beyond the t-test: Statistical equivalence testing. *Analytical Chemistry,* 221A-226A.

McAndrew, L. M., Musumeci-Szabo, T. J., Mora, P. A., Vileikyte, L., Burns, E., Halm, E. A., ... Leventhal, H. (2008). Using the common sense model to design interventions for the prevention and management of chronic illness threats: From description to process. *British Journal of Health Psychology*, *13*, 195–204.

Oswald-Mammosser, M., Weitzenblum, E., Quoix, E. (1995). Prognostic factors in COPD patients receiving long-term oxygen therapy. Importance of pulmonary artery pressure. *Chest*, *107*, 1193–1198.

Peters, E., McCaul, K.D., Stefanek, M., & Nelson, W. (2006). A heuristics approach to understanding cancer risk perception: contributions from judgment and decision-making research. *Annals of behavioral medicine: a publication of the Society of Behavioral Medicine*, *31*, 45–52.

Piette, J., Heisler, M., Krein, S., & Kerr, E. A. (2005). The role of patient-physician trust in moderating medication nonadherence due to cost pressures. *Archives of Internal Medicine, 165,* 1749-1755.

Proudfit, W. J., Bruschke, A. V. G., MacMillan, J. P., Williams, G. W. & Sones, M. S. (1983). Fifteen-year survival study of patients with obstructive coronary artery disease. *Circulation, 68,* 986-997.

Rogers, J.L., Howard, K.I., & Vessey, J.T. (1993). Using significance tests to evaluate

equivalence between two experimental groups. *Psychon B Rev*, *113*, 553-565.

Sallnäs, E.L., Rassmus-Grön, K., Sjöström, C. (2000). Supporting presence in collaborative environments by haptic force feedback. *Journal ACM Transactions on Computer-Human Interaction, 7,* 461-476.

Scheier, M. F., Carver, C. S. (1985). Optimism, coping, and health Assessment and implications of generalized outcome expectancies. *Health Psychology*, 4, 219-247.

Scheier, M. F., Matthews, K. A., Owens, J. F., Magovern, G. F., Lefebvre, R. C., Abbot R. A., Carver, C. S. (1989). Dispositional optimism and recovery from coronary artery bypass surgery: The beneficial effects on physical and psychological well-being. *Journal of Personality and Social Psychology,* 57, 1024-1040.

Semigram, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2005). Evaluation of symptom checkers for self diagnosis and triage: Audit study. *BMJ, 351,* 1-9.

Shavelle, R.M., Paculdo, D.R., Kush, S.J., Mannino, D. M., Strauss, D. J. (2009). Life expectancy and years of life lost in chronic obstructive pulmonary disease: Findings from the NHANES III Follow-up Study. *International Journal of Chronic Obstructive Pulmonary Disease,* 137:148.

Silence, E., Briggs, P., Harris, P. R., & Fishwick, L. (2007). How do patients evaluate and make use of online health information? *Social Science & Medicine, 64(9),* 1853-1862.

Stephens, R. G., Dunn, J. C., Rao, L. & Li, S. (2015). Exploring the knowledge behind predictions in everyday cognition: An iterated learning study. *Memory and Cognition, 43,* 1007-1020.

Street, R. L., & Haidet, P. (2011). How Well Do Doctors Know their Patients? Factors

Affecting Physician Understanding of Patients Health Beliefs. *Journal of General Internal Medicine*, *26*, 21–27.

Tversky, & Kahneman (1974). Judgment under uncertainty: Heuristics & Biases. *Science, 185,* 1124-1131.

Tversky & Kahneman (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 297-323.

Wagner, E. H., Austin, B. T., Davis, C., Hindmarsh, M., Schaefer, J., & Bonomi, A. (2001). Improving chronic illness care: Translating evidence into action. *Health affairs*, *20*, 64-78.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.

Williams, L. S., Bruno, A., Rouch, D., & Marriott, D. J. (1997). Stroke patients' knowledge of stroke: Influence on time to presentation. *Stroke, 28,* 912-915.

World Health Organization (2003). *Adherence to long-term therapies: Evidence for action.* Geneva, Switzerland: World Health Organization.

Yaniv, I., Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83(2),* 260-281.

Appendix A

Participant Instructions:

In this experiment, you will be asked to make predictions based on a single piece of information. Please read each question carefully. We are interested in your intuition so please do not make complicated calculations, just tell us what you think.

Specifically, you will be asked to estimate the total duration of different illnesses, based on how long someone has already had the illness. To give you an example of how to think about this question, imagine that you meet a man that is 50 years old and you are asked to estimate the total duration of his life. You might guess that his lifespan is likely to be 79 years of age (because this is the national average).

Importantly, you are NOT being asked how much longer he is likely to live, but rather the total age that he would reach.

PLEASE CALL OVER THE EXPERIMENTER BEFORE CLICKING TO CONTINUE.

Appendix B

Experimenter Instructions:

After participants read the instructions, give them the following test question to ensure that they understand the task:

Given that you meet someone who has had food poisoning for 2 days, what do you expect the total duration of this illness will be?

If they answer with any value LESS than 2 days, explain the task to them again and then ask this follow up question:

Given that you have had a headache for 1 hour, what do you expect the total duration of this illness will be?

If participants answer with any value LESS than 1 hour, they should be excluded from the experiment.