Lay Understanding of Illness Probability Distributions

Talia Robbins (talia.robbins@rutgers.edu)
Pernille Hemmer (pernille.hemmer@rutgers.edu)
Department of Psychology, 152 Frelinghuysen Road
Piscataway NJ, 08901 USA

Abstract

Our central question is: how accurate 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 accuracy 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 accurately estimate the mean and form of distributions for acute illnesses.

Keywords: Decision-making; Probability; Health; Cognition

Introduction

How accurate 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, how accurate 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; Robbins & Hemmer, 2017). 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 whether people hold accurate expectations 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 accurate 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 accurate understanding of this distribution 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 (Robbins & Hemmer, 2017).

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 (2003), 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) can people accurately represent the form of illness distributions? (2) can people accurately represent the mean of illness distributions? (3) are there differences in accuracy between acute and chronic illnesses? (4) are there individual differences in the strategies people use to generate these distributions?

Methods

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

Illnesses We selected six illnesses, including both acute (appendicitis, seasonal flu, and the common cold) and chronic (COPD, type-II diabetes, and chronic heart disease) 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. The illnesses were also intended to span a range of duration and familiarity. Familiarity was determined based on prevalence statistics for people diagnosed with that illness each year (see Table 1). Lastly, the ground truth for illness durations, against which participant accuracy was measured, was determined from clinical data (see Table 1 and Figures 3 and 4).

Distribution Builder We use a variation of the Distribution Builder of Goldstein et al. (2003). See Figure 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

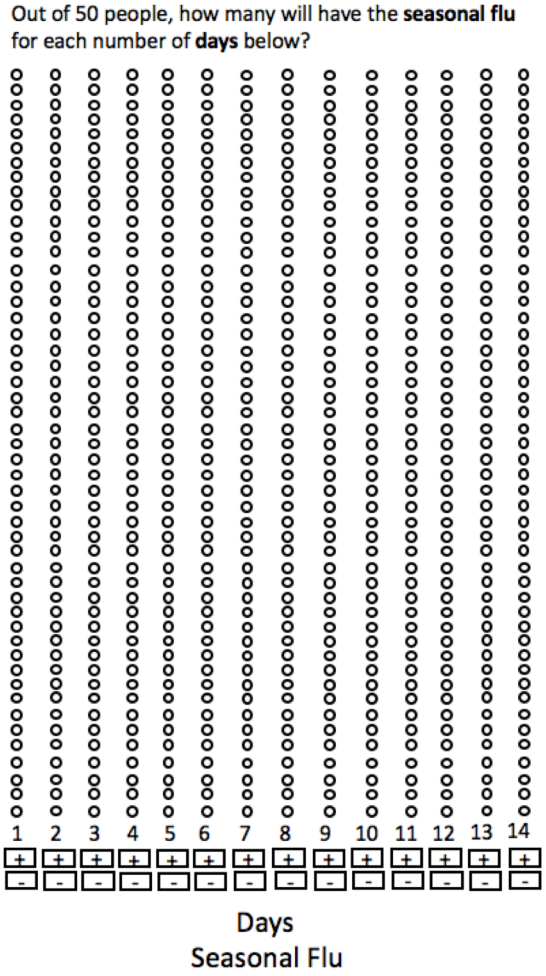

Figure 1: Sample distribution builder as seen by the participants. Participants could add or remove 'virtual people' from each bin (which represented an amount of time with an illness) using the plus and minus signs below that bin. Here, the circles are white because they have not been filled with 'virtual people', if the plus button is selected the empty bin is filled with a red circle.

Table 1: *Sources for Clinical Data (in order of prevalence)*

Illness (Prevalence/10,000)	**Source of Clinical Data**
Acute (in order of prevalence)	
Appendicitis (9)	Atema et al. (2015)
Seasonal Flu (1250)	Kohno et al. (2010)
Common Cold (2360)	Gwaltney (1967)
Chronic (in order of prevalence)	
COPD (4.5)	Oswald-Mammosser et al. (1995)
Type II Diabetes (860)	https://taliarobbinsrutgers.wordpress.com/empirical-data
Chronic Heart Disease (1130)	Proudfit et al. (1983)

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 (see results for information on the clinical distributions). 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

Ground Truth

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 3 for the clinical duration distributions for the six illnesses in this experiment, with corresponding Erlang distribution fits. The clinical data provides a ground truth for the distributions of durations (see Table 1 for clinical data sources).

Accuracy and Range of Responses

We first assess participant accuracy 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 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.

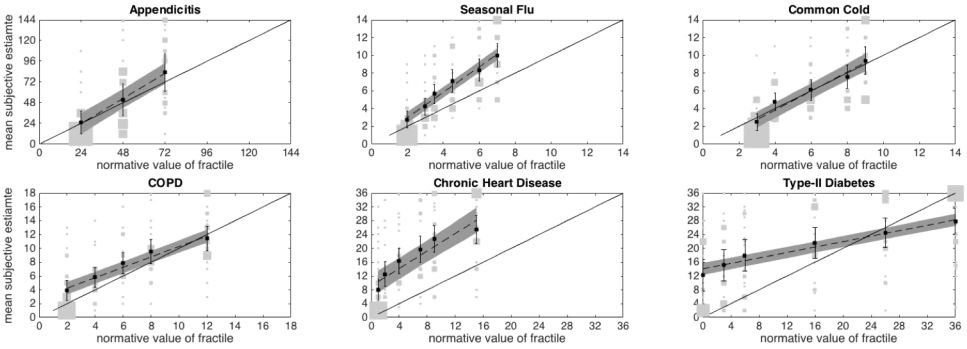

Figure 2: Accuracy for the 1st, 11th, 26th, 50th, 75th, 90th, and 100th fractiles. Light grey squares are individual responses, sized proportionately to number of responses. Black squares and error bars represent the mean of individual responses and standard errors for a given normative value. Dashed 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.

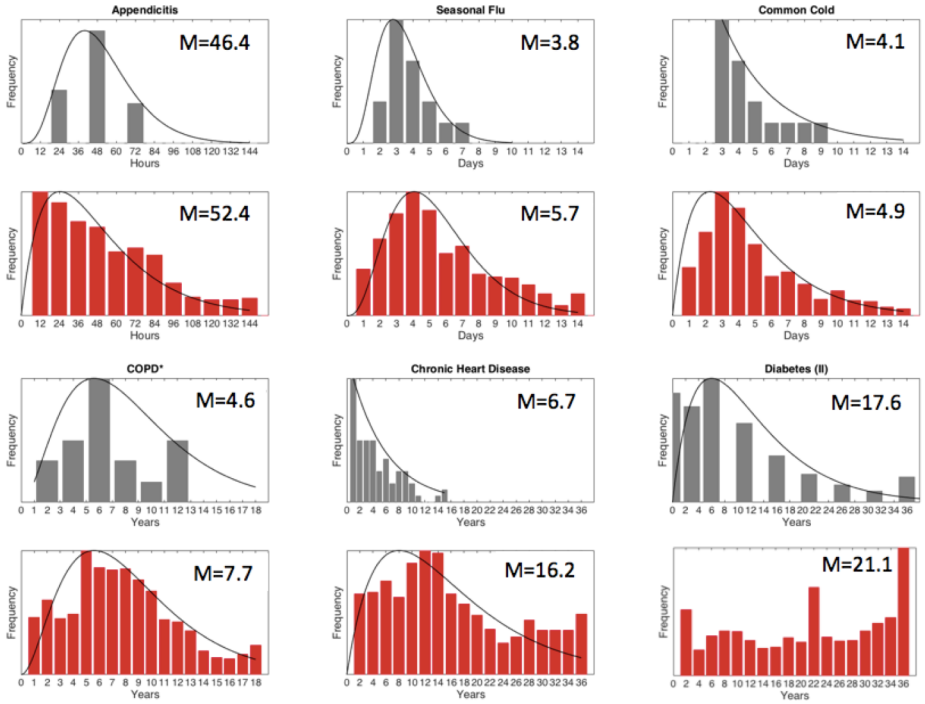

Figure 3: The first and third rows show histograms of clinical data for six illnesses with best fitting Erlang distributions (excluding diabetes, which could not be fit by the Erlang distribution). Grey bars show the frequency of each illness duration, black lines show the Erlang fit to clinical data. 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.

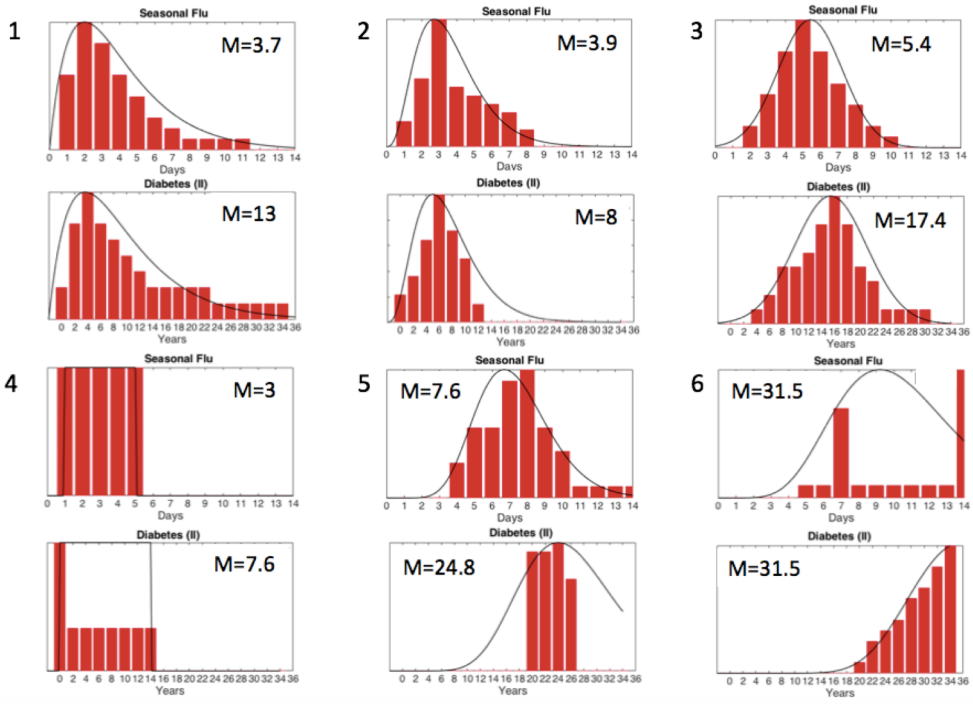

Figure 4: Samples of strategies used by participants in our task. Each pair of panels shows two samples, one from an acute (seasonal flu) and one from a chronic illness (type-II diabetes). See figure 2 for clinical data (ground truth). From top left to bottom right: 1. correctly estimate the distribution 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.).

Understanding of distributional form

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 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 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:* $D(359)=.93$, $p<.001$.

Understanding of the mean

Next, we sought to evaluate participant accuracy 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 3 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 do not understand the mean illness duration. 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 3 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: $t(999)=63.5$, $p<.001$).

Individual differences in strategy

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 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: how accurate are people's statistical intuitions for probability distributions in the domain of health? We found that, on average, people have accurate mental representations of probability distributions for illness duration, 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 accurately reflected the correct form of the distribution (see Figure 3).

Our second question was: can people accurately reproduce the mean of illness distributions? We found that for acute illnesses, participants accurately reproduced the mean, 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 3, 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 in accuracy between acute and chronic illnesses? It is clear that differences exist, such that participants could reproduce the mean and form for all three acute illnesses but could only reproduce 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 correct 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 decisions for chronic illnesses may be caused by flawed information.

Future work should focus on how those expectations might be corrected. 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.

Acknowledgments

This research was supported by the National Science Foundation grant #1453276.

References

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. *Acad Emerg Med,* 1015-1024.

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

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.

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.

Griffiths, T.., & Tenenbaum, J. (2006). Optimal predictions in everyday cognition. *Psychol Sci*, *17*, 767–773.

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

Hemmer, P. & Steyvers, M. (2009). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review*, 16, 80-87.

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.

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

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.

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. *Ann Behav Med*, *31*, 45–52.

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.

Robbins, T., & Hemmer, P. (2017). Explicit Predictions for Illness Statistics. 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.

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.