Casey's Problem: Interpreting and Evaluating a New Test

JAMES E. SMITH
Fuqua School of Business
Duke University
Durham, North Carolina 27708

ROBERT L. WINKLER
Fuqua School of Business
Duke University

Casey, the newborn daughter of one of the authors of this paper, received a positive result on an experimental medical screening test, indicating that she may lack an enzyme required to digest certain fats. The interpretation of this test result was complicated by uncertainty about the false-positive rate for the test—this was the first positive reading—and the prevalence of the medical condition. We used a simple Bayesian model to help assess the probability that Casey actually had the enzyme deficiency and to help better understand the role and value of this screening test. The model we used and, more generally, our style of analysis could also be used with other new diagnostic tests, such as tests used in manufacturing and environmental contexts as well as other medical situations.

Casey Katharine Carswell Smith, the second daughter of Lori Carswell and one of the authors of this paper, was born January 4, 1995. Casey was an apparently healthy baby, born after a more-or-less routine pregnancy and delivery. As with most newborns, she had a small sample of blood drawn from her heel shortly after her birth, which was sent off to a lab for routine tests. Unlike most newborns, however, Casey was also part of an experimental screening program in which a new procedure was being used to test blood for certain genetic metabolic disorders. This experimental test had been applied to approximately 13,000 newborns, and Casey

0092-2102/99/2903/0063/$5.00
This paper was refereed.

DECISION ANALYSIS—INFERENCE, APPLICATION
HEALTH CARE—DIAGNOSIS

was the first to test positive.

The test results came back when Casey was approximately six weeks old. The screening test indicated that Casey had an elevated level of a particular molecule (the long-chain acylcarnitine species called C14:1) in her blood, an expression of the lack of an enzyme required to digest a particular form of long-chain fats. This is a very rare condition that, if not treated, often presents itself as a sudden and mysterious death (included in the broad category of sudden infant death syndrome or SIDS) or severe illness after a few months of life, often accompanied by permanent damage to vital organs. Casey exhibited no obvious symptoms of this problem—although none would necessarily be expected at her age—and her older sister and parents did not have any known metabolic disorders. There is no clearly defined treatment regimen for this condition, but the doctor suggested that, given that the condition had been identified before any damage was done, certain dietary practices and other precautions could save Casey's life and give her a good chance of enjoying a high quality of life.

The experimental test was developed at Duke University Medical Center and had been applied to every child born at Duke Hospital and Durham Regional Hospital in the three years prior to Casey's birth. The test involves extracting materials from the newborn's blood and evaluating these materials in a tandem mass spectrometer to look for telltale signs of undigested fats in the blood. The test is noninvasive because it makes use of blood samples that are routinely collected from newborns to screen for other metabolic disorders (specifically, PKU deficiencies). In addition to testing for the specific long-chain defect indicated in Casey's case, this new test screens for a variety of different metabolic disorders including more common medium-chain defects as well as other long-chain defects. Casey's result was the first positive for any of the defects in the history of the screening program.

When we speak of a "positive test result," it is important to remember that the test result does not, in itself, constitute a definitive diagnosis. Instead, doctors typically interpret a positive result on a screening test as an indication to perform more specific follow-up tests. Nevertheless, the doctor spoke frequently of "confirming the diagnosis" in his conversations with the parents and, when pressed by the parents for a probability, indicated that he thought the probability that Casey had the deficiency was in the 80 to 90 percent range.

The question we struggled with was how to think about the results of this test. In a sense, this is a classic Bayes' rule exercise, like those encountered in statistics and decision analysis courses. We need to consider the base rate (or, in epidemiological terms, the prevalence) of the condition and the false-positive and false-negative rates for the test and then calculate the posterior probability of Casey having the enzyme deficiency given the result of the test. But in this case there is a twist: because the test has no history of positive results—true or false—we need to think especially carefully about the false-positive rate. Clearly the false-positive rate for this test cannot be too high; otherwise we would have observed some false positives in the 13,000 tests that had been performed. But does this mean it is zero? Moreover, because the condition for which Casey tested positive is very rare and identified fairly recently, there is a question about the prevalence of the condition.

Casey's parents were deeply concerned by the doctor's diagnosis and the apparently dire nature of the condition. But, being aware of the need to think carefully when interpreting screening tests, they were suspicious of the doctor's assessment and pressed him for details concerning the likelihood of the condition and the false-positive rate. They enlisted the help of their friend and "consultant" (Winkler), who is an expert in Bayesian statistics and could provide more objective support and advice. Though we now know more about Casey's condition and this new test, the analysis we describe was based on our state of information while waiting for results from follow-up tests.

While our analysis was specific to Casey's situation, the general structure of Casey's problem arises more frequently, whenever new diagnostic tests are evaluated or deployed. For example, similar situations arise when considering new screening tests for other medical conditions, new environmental assays, or new screening procedures for credit checks or quality control. We believe that the general framework and model we used could also be useful in these other situations. However, our primary goal is to demonstrate by example how some rough quantitative judgments and simple analysis can improve understanding, communication, and decision making in a highly charged and emotional situation, fraught with uncertainty and potentially grave outcomes.

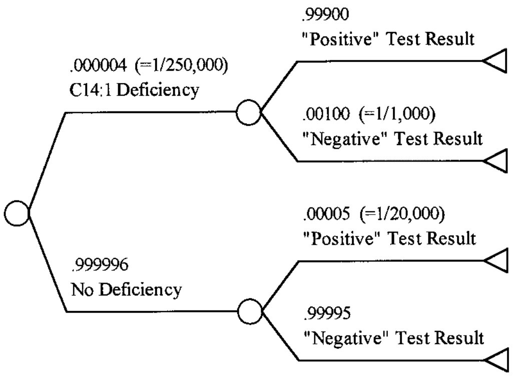

Figure 1: To determine the likelihood that Casey had the enzyme deficiency, we constructed a simple probability tree showing the probabilities of the various possible outcomes.

What to Think?

We began our analysis of Casey's problem by drawing a probability tree (Figure 1) and trying to determine appropriate numbers to place in it. The first question we considered was the prevalence of the specific enzyme deficiency indicated in Casey's case; that is, the probability that a randomly selected newborn would have the C14:1 enzyme deficiency. The doctor reported that there were very few known cases of people living with this condition; but, given that most people suffering from this condition had died as infants, this alone does not tell us much about the likelihood of the condition. We also knew, however, that every SIDS death in North Carolina in the previous five years had been autopsied and that this condition had not been implicated in any of them. We also believed that our doctor and his colleagues, being leading authorities in this field, would be aware of cases in other states where this enzyme deficiency was implicated, certainly if there were many of them. The doctor indicated that he estimated that, in total, long-chain enzyme deficiencies occur in about one in 40,000 newborns, but that Casey's specific C14:1 deficiency was among the rarer of the eight long-chain deficiencies that had been identified. Based on all of this, we estimated the prevalence to be one in 250,000. Recognizing our uncertainty about the prevalence, we considered a range from one in 100,000 to one in a million when performing sensitivity analysis.

Next we considered the false-positive rate; that is, the probability of obtaining a positive test result if Casey did not actually have the deficiency. Given the rarity of the condition that Casey tested positive for, we knew that this was a critical assumption, and we didn't have much guidance. Other screening tests that we knew more about—such as HIV tests, drug tests, amnio fetal protein tests—had false-positive rates in the one-in-100 to one-in-1,000 range. (Sox et al. [1988] lists false-positive and false-negative rates for a variety of medical tests.) But we knew this test couldn't have a false-positive rate in that range because, if the rate were this high, we would almost certainly have seen some false positives in the 13,000 newborns tested before Casey. We took a figure of one in 20,000 to be a rough estimate for our initial analysis and considered a range from one in 5,000 to one in 1,000,000 for sensitivity analysis. It was clear to us that we would have to think about and model this uncertainty more carefully.

Finally, we considered the false-negative rate; that is, the probability of obtaining a negative test result if Casey actually had the deficiency. Here again we had little hard evidence, but we suspected that the rate was not high. Based on what we knew about the design of the test, we figured that if the deficiency were present, it would probably be detected. We also knew that there had been no known cases in which a child had the deficiency and passed the test. However, given the rarity of this condition and the small number of newborns tested, this observation is hardly surprising. We assumed a false-negative rate of one in 1,000 and considered a range from one in 100 to one in 1,000,000.

Given these base-case numbers, we applied Bayes' rule to calculate the probability that Casey had the C14:1 enzyme deficiency given the test result:

$$
\begin{aligned}
&P(\text{Has Def.} \mid + \text{ Result}) \\
&\quad = P(\text{Positive Result} \mid \text{Has Def.})\, P(\text{Has Def.})/ \\
&\qquad \{P(\text{Positive Result} \mid \text{Has Def.})\, P(\text{Has Def.}) \\
&\qquad + P(\text{Positive Result} \mid \text{No Def.})\, P(\text{No Def.})\} \\
&\quad = (.999)(.000004)/\{(.999)(.000004) \\
&\qquad + (.00005)(.999996)\} \\
&\quad = .00000396/\{(.00000396) + (.00005000)\} \\
&\quad = .0740.
\end{aligned}
$$

Thus, given the positive test result, we estimated the probability of Casey actually having the deficiency to be approximately 7.4 percent or one in 13.5. This probability suggests a legitimate cause for concern for the parents, but it is much lower than the 80- or 90-percent probability suggested by the doctor.

To understand the impact of changes in the assumptions on this result, we varied each of the numbers over the ranges indicated earlier and summarized the effects using a tornado chart (Figure 2). This analysis showed that reasonable variation in the false-negative rate has little impact on

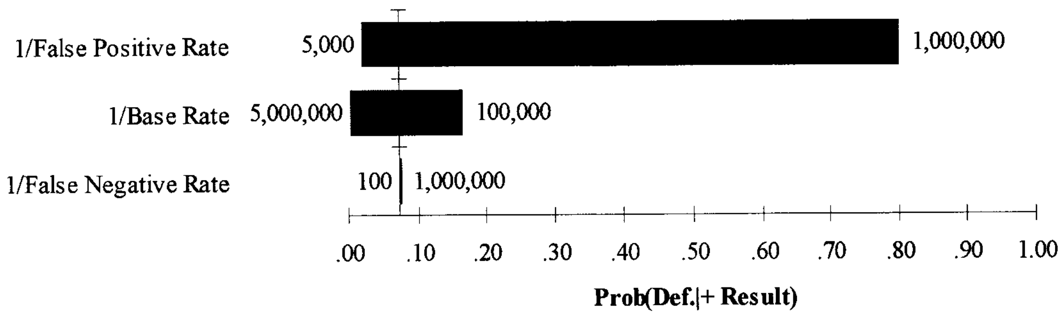

Figure 2: To understand the impact of changes in our assumptions on the probability that Casey has the deficiency, we constructed this tornado chart. The bars of the chart show the posterior probabilities given by varying the assumptions, one at a time, from their base-case values. The vertical line shows the probability (.074) with the base-case assumptions.

the posterior probability. Increasing the prevalence from one in 250,000 to one in 100,000 increases the posterior probability to about .17, still well below the probability the doctor suggested. Decreasing the false-positive rate to one in a million, however, increases the posterior probability to .80. The magnitude of this effect confirmed our earlier intuition that we needed to think more carefully about the false-positive rate. This is what we did next.

How Accurate Is the Test?

Our sensitivity analysis showed the key impact of the false-positive rate. To improve our analysis, we needed to acknowledge that little was known about the false-positive rate and to think carefully about how to use the information we did have. Given that there were no false positives in 13,000 previous tests, it was clear that the false-positive rate could not be very high. But given the low prevalence of Casey's hypothesized enzyme deficiency, it would take extreme confidence in the test to get posterior probabilities in the 80- to 90-percent range. How should we think about this false-positive rate? What do 13,000 negative test results tell us about the false-positive rate?

We chose to approach the questions about the false-positive rate as a Bayes' rule problem as well. First, we needed to ask what our beliefs about the true false-positive rate would have been before we saw 13,000 negative test results by specifying a prior probability distribution on this quantity. Then we needed to consider the likelihood of seeing 13,000 negative test results given the different possible true false-positive rates. We could then update our distribution on the false-positive rate using Bayes' rule.

To simplify our analysis, we assumed that the uncertainty surrounding the false-positive rate could be described using a beta-binomial model. In using this model, we assumed (1) that successive test results could be modeled as independent draws from a Bernoulli process with a false-positive probability p for each test and (2) that our prior distribution on the false-positive probability could be described by a beta distribution. Assumption (1) seemed natural and appropriate in this setting and, given the ability of the beta distribution to capture a variety of different shapes and forms of distributions, assumption (2) seemed reasonable given the nature of the test and the circumstances of the analysis. (Clemen [1996] and Winkler [1972] give more details on the beta distribution and the beta-binomial model.)

We worked together to develop a prior distribution that described our beliefs about the likelihood of false positives before we saw the results of the experimental screening tests. With the beta distribution, one way to specify a distribution is to specify an expected false-positive rate (the mean of the distribution) and an equivalent sample size. In assigning an expected false-positive rate, we thought about the accuracy of other screening tests and thought that before we had seen any experimental screening results it might have been reasonable to assess an expected false-positive rate of about 0.001. This was consistent with the performance of other tests we were familiar with [Sox et al. 1988]. Given how the test works, it seemed that it could be quite accurate, but unlike the tests reviewed by Sox et al., before the screening experiment that Casey was part of, it had not been used on large populations and could still prove disappointing.

With the beta distribution, the equivalent sample size describes how spread out this prior distribution is; it can be interpreted, intuitively, as if our prior state of information is equivalent to having seen this many observations. We chose an equivalent sample size of 1,000. Using the beta distribution, this implies that, before seeing any results from the screening test, we felt that there would be a 10-percent chance of having a false-positive rate less than one in 9,482, a 50-percent chance of having a false-positive rate less than one in 1,442, and a 90-percent chance of having a false-positive rate less than one in 434 (see Figure 3). In the good cases (for example, the 10th percentile case), this would be one of the most accurate medical screening tests available; in the medium and bad cases, its performance would be more routine. We thought that this prior appropriately described our uncertainty before the screening test, but we also recognized the need to do more sensitivity analysis and, if necessary, to review the assumptions with the doctor.

Casey was the first to test positive.

With the beta-binomial model, it is easy to update the prior distribution based on the results of the screening tests using Bayes' rule. Given a prior with a mean of one in 1,000 and equivalent sample size of 1,000, if (hypothetically) we had a new sample of 1,000 screening tests that included five false positives, we would take our new equivalent sample size to be 2,000 (= 1,000 + 1,000) and revise our expected false-positive rate up to six (= 1 + 5) in 2,000. In the actual case, having seen no false positives in 13,000 samples, we had a new equivalent sample size of 14,000 (= 1,000 + 13,000), and we revised our expected false-positive rate down to one in 14,000. Observing no false positives thus shifted the distribution for the false-positive rate to the left (Figure 3).

Thus, given that we had seen no false positives in 13,000 trials, we believed that, if Casey did not have the enzyme deficiency, there was a one-in-14,000 chance that she would have a positive test result. To calculate the probability that she truly had the enzyme deficiency, we used this false-positive rate in the Bayes'-rule calculation described earlier. This gives a probability that Casey has the C14:1 deficiency of 5.3 percent.

Like our earlier calculation, this one is also based on some fairly rough assumptions, and it is important to understand the sensitivity of our results to these assumptions. This posterior probability (the probability that Casey has the C14:1 deficiency given the test result) varies depending on the expected false-positive rate and the equivalent sample size assumed for the prior distribution on the false positive rate. As we decrease the prior expected false-positive rate (Figure 4), holding constant the equivalent sample size, Casey's probability of having the deficiency increases, but it does not reach the 80- to 90-percent range until the prior expected false-positive rate decreases to about one in 100,000. As we decrease the equivalent sample size, holding constant the expected false-positive rate, Casey's probability of having the deficiency increases, but reaches a probability of only 34 percent with an equivalent sample size of 100. Considering changes in both parameters simultaneously, we found that to reach probabilities in the 80- to 90-percent range we would need an unlikely combination of expecting an extremely low false-positive rate and yet having a small equivalent sample size. This analysis gave us confidence that the doctor's probability was too high and that a more appropriate probability would be in the one to 10 percent range with five percent being our base-case probability.

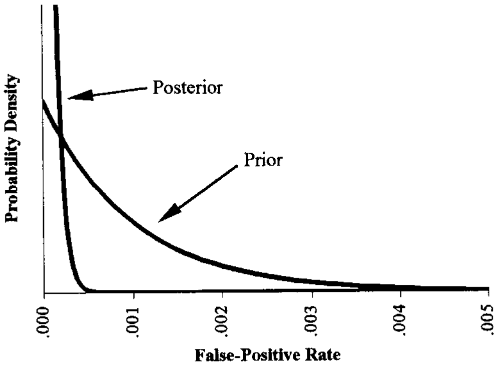

Figure 3: The prior probability density function describes our uncertainty about the false-positive rate before seeing any results from the screening tests. The posterior density function describes our uncertainty after observing no false positive in 13,000 trials.

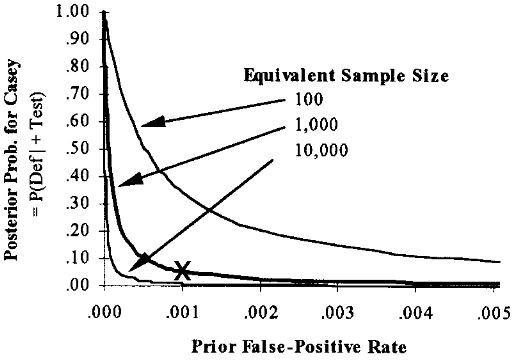

Figure 4: This chart shows how the probability that Casey has the deficiency varies with our assumptions about the prior distribution for the false-positive rate. Our base-case assumptions yield a probability of 5.3 percent and are marked with an X in the figure. The middle curve shows probabilities based on an equivalent sample size of 1,000 and varying prior expected false-positive rates. The other curves show the same results with different equivalent sample sizes.

How Could the Doctor Be So "Wrong"?

As you may have guessed, follow-up tests performed shortly after learning of Casey's positive test result revealed that Casey's enzyme system was functioning normally. Repeating the test on another spot blood sample from the original newborn screening card gave results identical to the first, but a repeat of the same test with a new sample and some other tests indicated no problems. While the doctor believes that these results suggest a transient impairment to Casey's enzyme system present at birth (perhaps due to a delayed maturation of her enzyme system), Casey's test result was certainly a false positive in that she showed no evidence of an enzyme deficiency in these follow-up tests.

The fact that Casey's result was a false positive does not imply that our five-percent probability was right and that the doctor's 80- to 90-percent probability was wrong; we could simply have been among the lucky 10 or 20 percent to have a false positive. Nevertheless, this experience and our modeling do suggest the need to be careful in interpreting test results. Because many people find the five-percent probability surprising or counterintuitive, we briefly review some of the psychological heuristics and biases affecting people's intuitive probability judgments in these kinds of situations so that you might be aware of these traps if you find yourself in an analogous situation.

We believe that the doctor was employing what Kahneman and Tversky refer to as the representativeness heuristic when forming his intuitive evaluation of Casey's situation. Kahneman and Tversky [1974] suggest that when judging the probability that some object belongs to a particular category, people typically use a heuristic in which the judgments are based on the degree to which the object exhibits characteristics that are representative or stereotypical of the category. The more the object resembles the stereotype, the higher the judged probability. While this kind of similarity is an important consideration in forming probability judgments, it does not consider the prevalence of objects in the category and consequently leads to predictable biases in which subjects ignore or underestimate the effect of the base rate or prior probability.

In this setting, we believe that the representativeness heuristic was operating at two different levels. First and most obviously, given that Casey had tested positive for the C14:1 enzyme deficiency, she appeared to fit the stereotype of somebody with this deficiency and it was natural for the doctor to assume that she had the condition regardless of how rare it is. This is a common problem with the interpretation of test results, exhibited by lay people and physicians. For example, Eddy [1982] did a study in which he asked physicians to judge the probability that a woman had a malignant breast tumor based on an X-ray that correctly classifies 80 percent of the malignant tumors and 90 percent of benign tumors. This woman was judged to have a one-percent probability of having a malignant tumor prior to having an X-ray come back positive for a malignancy. In this study, Eddy found that 95 out of 100 doctors estimated the probability to be about 75 percent. Applying Bayes' rule, the probability is only 7.5 percent.

In Casey's situation, the doctor, with a background in genetic testing and a role as a research physician, is familiar with Bayes' rule and is likely to be aware of this phenomenon. But given the extremely low false-positive rates, it is easy to see how someone—even an expert—reasoning intuitively could exhibit this bias. In a note debriefing the parents after the follow-up tests had been completed, the doctor emphasized this similarity in test results: "All the patients with the [long-chain enzyme deficiency] had stood out quite flagrantly, and Casey's sample looked like theirs." This similarity and the fact that the false-positive rate for the test is, in his words, "essentially zero," seemed to be dominating his thinking. But the prevalence is also essentially zero, and to make sense of the comparison, most of us would have to do some calculations along the lines of those outlined earlier.

We also suspect that the representativeness heuristic played a role in the doctor's arriving at an estimate of essentially zero for the false-positive rate. Kahneman and Tversky [1974] and others have found that, when forming probability estimates in this kind of setting, intuitive judgments tend to be dominated by the sample proportion and are insufficiently sensitive to the size of the sample and prior probability. Having no false positives in 13,000 tries is certainly stereotypical performance for a test that never generates false positives and, following the representativeness heuristic, one might naturally assume the false-positive rate to be zero. In the beta-binomial model, given our prior distribution (equivalent to having seen one false positive in 1,000 tries), one could not arrive at an expected false-positive rate of one in 100,000 or less without having performed the test more than 99,000 times with no false positives. Here again, though Casey's physicians were trained in statistics and likely to be aware of the importance of sample size, Casey's situation is extreme and likely to lead people making intuitive judgments into this familiar trap.

The anchoring-and-adjustment heuristic may have also played a role in the doctor's intuitive evaluation. When using this heuristic to estimate probabilities, people anchor on a probability from some apparently analogous situation and adjust the probability for use in the present situation. The anchors are sometimes inappropriate and the adjustments are often insufficient [Kahneman and Tversky 1974]. For example, a doctor might anchor on a high posterior probability, thinking only about how accurate the test is, and adjust that number insufficiently to reflect the specific characteristics of the test and the base rate of the condition being tested. Alternatively, the doctor might focus on a posterior probability for some other test—for example, one that might give a 90-percent posterior probability—and then adjust insufficiently for the situation at hand. The use of such a heuristic seems particularly likely when interpreting a new test since physicians don't have much experience with positive test results and have not had the opportunity to observe, for example, that only a small proportion of those who test positive actually have the condition.

How Much Is the Test Worth?

One common reaction to hearing of Casey's false positive is to think that the test is a waste of time, money, and emotion. But we cannot draw this conclusion based on Casey's outcome; we must compare the possible outcomes with the test to those without the test. To get a sense of the value of the test, we constructed a de-

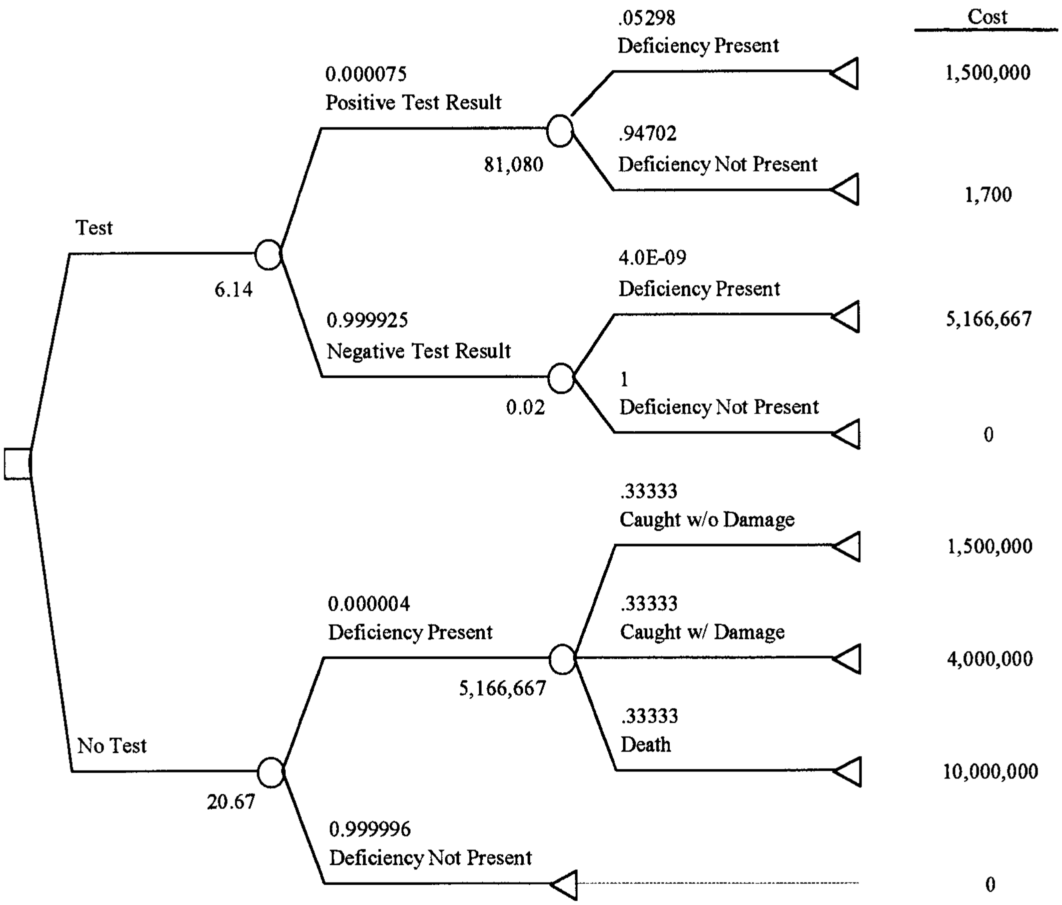

Figure 5: To understand the value and role of the test, we constructed a decision tree showing each of the possible outcomes and the costs in each scenario. The numbers beneath each node in the tree represent the expected costs given that you are in the state corresponding to that node.

cision tree showing the possible outcomes with and without the test (Figure 5). In constructing this tree, we relied on our own judgment. Because this was intended as a quick, rough analysis, it should be interpreted more as a demonstration of the kind of analysis required to address this question than as an accurate evaluation of this particular test. Our goal was to better understand the purpose of the test and its medical benefits.

Starting at the bottom of the tree, we first considered the possible outcomes in the case where no test was performed. In this case, we assumed Casey's probability of having the enzyme deficiency to be one in 250,000, our expected prevalence for the condition. If she does not have the condition, she would live a normal life, and we assigned a cost of zero dollars to this situation. If she does have the enzyme deficiency, the question is whether the doctors would catch the condition before it killed Casey or caused permanent damage; we assumed, for the sake of argument, that each of these outcomes was equally likely.

To capture the value of the test, we had to place some monetary value on the various possible health states. We assessed a value of life of $10 million that would be lost in the event of Casey's death. This number is to be interpreted as a small-risk value of life as in Howard's [1980] model for valuing life risks; it is appropriate only for valuing small risks of death—one in 250,000 qualifies as a small risk—and it reflects the parents' trade-off between quality and length of life and risk aversion. While economists' estimates of the average small-risk value of life are typically in the $2 to $6 million range, the higher number for Casey reflects her parents' preferences and economic situation.

In the scenario in which the enzyme deficiency is discovered before it does any permanent damage, we assumed a 15-percent reduction in the value of Casey's life, reflecting a diminished quality of life and life expectancy due to this deficiency. If the condition is discovered after it has done permanent damage, we assumed a 40-percent reduction in the value of Casey's life for these same reasons. While we did not discuss these potential outcomes in much detail, the doctors had described the case of a woman in North Carolina who was living with this condition and had recently given birth to a healthy baby. In her case, the condition was not identified until after it had done considerable damage, and she lived with severe health problems. This suggested that the condition could be managed through dietary restrictions and other precautions, particularly if the condition were identified before it had done damage. Nevertheless, these management strategies are restrictive and may not be perfect, so we have assumed some decrease in the quality of life and life expectancy, even if the condition is identified before it does damage.

Calculating expected costs in the case in which we don't perform the test, we found an expected cost of $20.67. This figure implies that Casey's parents should be willing to pay up to $20.67 for a hypothetical white pill that would guarantee that Casey would not have the C14:1 enzyme deficiency. The screening test is not such a white pill, but by allowing for the early detection of this deficiency, it makes it more likely that doctors will catch the deficiency before it causes death or damage.

If the test is performed, is negative, and Casey truly does not have the deficiency, she will lead a normal life, and we again assigned a cost of zero dollars. In the event that the test is negative and Casey truly has the deficiency (the test gave a false negative), then the situation is similar to the case in which she has the enzyme deficiency and no test was done. We have assigned a cost for this scenario that is equal to the expected costs of having the deficiency without testing. If the screening test comes back positive and Casey truly has the deficiency, then we face essentially the same situation we would if we had discovered the condition without testing before it did any damage. As in that case, we assigned a cost of $1.5 million, reflecting a diminished quality of life and life expectancy due to this deficiency amounting to 15 percent of the assumed value of Casey's life.

In the event of a false positive, Casey would lead a normal life, but we would have costs associated with follow-up testing and the emotional trauma to the parents in dealing with this false positive. We estimated the costs of follow-up testing to be about $200 and the trauma costs to be about $1,500 (that is, the parents would be willing to pay $1,500 to avoid the ordeal associated with the positive result). These follow-up tests were fairly noninvasive (blood and urine samples) and were performed quickly. If the follow-up tests had been more invasive or required more time, the costs and trauma would have been much higher. There would also be emotional trauma and testing costs in those scenarios in which Casey truly has the enzyme deficiency, but we assume that these costs are included in (and dwarfed by) the loss of value in Casey's life.

Casey is now a thriving four-year-old.

Calculating the expected total costs with the test, we find an expected cost of $6.14: if the test has been done but we do not know the results yet, the value of a hypothetical white pill that would guarantee that Casey would not have the C14:1 enzyme deficiency is $6.14. The value of the test is then the difference in expected costs with and without the test, or $20.67 − $6.14 = $14.53. The test actually costs about $6 per child to perform and thus seems like a good investment.

To understand the sensitivity of these results, we constructed another tornado chart showing how the value of the test changes as we vary these assumptions over their possible ranges (Figure 6). This analysis showed that the critical assumptions affecting the value of the test are the prevalence, the assumed value of life, and the probability of death before detection; if we wanted to accurately determine the value of the test, we would have to think more carefully about these assumptions. Interestingly, the false-positive rate—which was critical in calculating the probability that Casey actually has the deficiency—is not very sensitive in this calculation. This is because the overall probability of a false positive is rather small and the costs associated with such a false positive are very small compared to value of the life potentially saved.

In reviewing these results, you should remember that the value calculated is based on our own assumptions (particularly those concerning the prognosis if Casey were to have the deficiency) and does not include the value of the test in screening for defects other than the C14:1 deficiency or any value of learning about the test for potential application to others. If we were to include the ability of the same test to screen for other deficiencies (the test also screens for the medium-chain enzyme deficiencies, which occur in about one in 23,000 newborns) or the value of learning for others, the overall value of the test would increase substantially, and we would see that this $6 test is a very good investment.

Conclusion

Since the results of the follow-up test were available soon after we did our analysis and before the doctors performed any invasive medical procedures, the main benefit of the analysis was to give Casey's

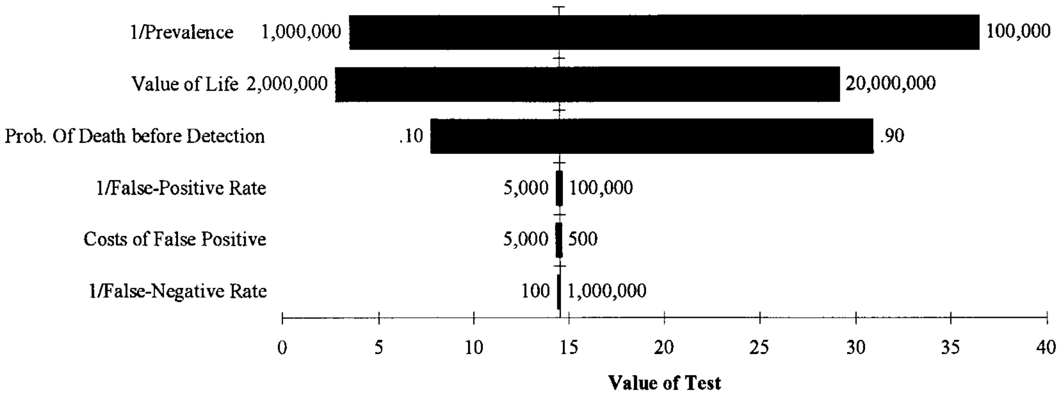

Figure 6: This tornado chart shows how the value of test varies with different assumptions.

parents some peace of mind. While a five-percent chance that one's child has such a serious problem is not an easy thought to bear, it is far less difficult than the 80- or 90-percent chance the doctor initially suggested. In other situations, however, the follow-up tests or procedures might be more invasive or risky (for example, surgery) and the difference between the two probabilities could lead to very different courses of treatment. By constructing the decision tree and calculating the value of the test, we were better able to appreciate the medical benefits of the test and, in the end, the parents felt good about the test even though it generated a false positive for Casey. While Casey's parents were fortunate to have the training to allow them to reason through the implications of the test result, patients and families would lack such training in most cases, and it would be up to the physicians or counselors to help them interpret the test result and its implications appropriately. Our experience demonstrates how some rough quantitative judgments and simple analysis can help patients and families properly understand the nature, magnitude, and severity of the risks they face and avoid the traps associated with inappropriate, but commonly used, heuristics.

While we have focused here on the interpretation and evaluation of the test in terms of its impact on Casey and her parents, the analytic framework could be extended to consider such questions as whether the testing program should be scrapped or adopted for broader use. The analytic structure of Casey's problem—a diagnostic test with unknown performance characteristics—arises in many other settings, including other medical tests, environmental assays, and various other screening procedures, such as those used for credit checks or quality control. In these problems, the model we used and variations on it may prove helpful. More generally, in these other contexts and in many other difficult situations, we believe that some rough quantitative judgments and simple analysis can help improve understanding, communication, and ultimately decision making.

Epilogue

Shortly after Casey's false positive, the screening program identified its first true positive, but for a more common medium-chain defect rather than the long-chain defect suspected in Casey's case. Subsequently the screening program has been adopted more broadly and, starting last year, every child born in the state of North Carolina is now tested using this procedure. With a total of 93,000 newborns tested to date, the program has identified 10 cases of children with medium-chain enzyme deficiencies. None have been found to have the long-chain deficiency suspected in Casey's case. To date, there have been only two false positives, one for a medium-chain deficiency and one (Casey) for the long-chain deficiency. Other states are considering adopting this screening program, but none has yet followed North Carolina's lead.

Casey, we are happy to report, is now a thriving four-year-old with no apparent enzyme deficiencies.

Acknowledgment

We are grateful for helpful comments provided by Bob Clemen, Ward Edwards, Craig Fox, Dennis Fryback, Ralph Keeney, Harold Lehmann, Dennis Lindley, David Millington, Mike Rothkopf, and, particularly, Lori Carswell.

References

Clemen, Robert T. 1996, *Making Hard Decisions: An Introduction to Decision Analysis*, Duxbury Press, Pacific Grove, California.

Eddy, David M. 1982, "Probabilistic reasoning in clinical medicine: Problems and opportunities," in *Judgment Under Uncertainty: Heuristics and Biases*, eds. Daniel Kahneman, Paul Slovic, and Amos Tversky, Cambridge University Press, Cambridge, England, pp. 249–267.

Kahneman, Daniel and Tversky, Amos 1974, "Judgment under uncertainty: Heuristics and biases," *Science*, Vol. 185, pp. 1124–1131.

Howard, Ronald A. 1980, "On making life and death decisions," in *Societal Risk Assessment: How Safe Is Safe Enough*?, eds. Richard C. Schwing and Walter A. Albers, Jr., Plenum Press, New York, pp. 89–113.

Sox, Harold C., Jr.; Blatt, Marshal A.; Higgins, Michael C.; and Marton, Keith I. 1988, *Medical Decision Making*, Butterworth's, Boston, Massachusetts.

Winkler, Robert L. 1972, *Introduction to Bayesian Inference and Decision*, Holt, Rinehart and Winston, New York.