Helping Doctors and Patients Make Sense of Health Statistics

Gerd Gigerenzer,[1,2] Wolfgang Gaissmaier,[1,2] Elke Kurz-Milcke,[1,2] Lisa M. Schwartz,[3] and Steven Woloshin[3]

[1]Max Planck Institute for Human Development, Berlin; [2]Harding Center for Risk Literacy, Berlin; [3]The Dartmouth Institute for Health Policy and Clinical Practice's Center for Medicine and the Media, Dartmouth Medical School

SUMMARY *Many doctors, patients, journalists, and politicians alike do not understand what health statistics mean or draw wrong conclusions without noticing. Collective statistical illiteracy refers to the widespread inability to understand the meaning of numbers. For instance, many citizens are unaware that higher survival rates with cancer screening do not imply longer life, or that the statement that mammography screening reduces the risk of dying from breast cancer by 25% in fact means that 1 less woman out of 1,000 will die of the disease. We provide evidence that statistical illiteracy (a) is common to patients, journalists, and physicians; (b) is created by nontransparent framing of information that is sometimes an unintentional result of lack of understanding but can also be a result of intentional efforts to manipulate or persuade people; and (c) can have serious consequences for health.*

The causes of statistical illiteracy should not be attributed to cognitive biases alone, but to the emotional nature of the doctor–patient relationship and conflicts of interest in the healthcare system. The classic doctor–patient relation is based on (the physician's) paternalism and (the patient's) trust in authority, which make statistical literacy seem unnecessary; so does the traditional combination of determinism (physicians who seek causes, not chances) and the illusion of certainty (patients who seek certainty when there is none). We show that information pamphlets, Web sites, leaflets distributed to doctors by the pharmaceutical industry, and even medical journals often report evidence in nontransparent forms that suggest big benefits of featured interventions and small harms. Without understanding the numbers involved, the public is susceptible to political and commercial manipulation of their anxieties and hopes, which undermines the goals of informed consent and shared decision making.

What can be done? We discuss the importance of teaching statistical thinking and transparent representations in primary and secondary education as well as in medical school. Yet this requires familiarizing children early on with the concept of probability and teaching statistical literacy as the art of solving real-world problems rather than applying formulas to toy problems about coins and dice. A major precondition for statistical literacy is transparent risk communication. We recommend using frequency statements instead of single-event probabilities, absolute risks instead of relative risks, mortality rates instead of survival rates, and natural frequencies instead of conditional probabilities. Psychological research on transparent visual and numerical forms of risk communication, as well as training of physicians in their use, is called for.

Statistical literacy is a necessary precondition for an educated citizenship in a technological democracy. Understanding risks and asking critical questions can also shape the emotional climate in a society so that hopes and anxieties are no longer as easily manipulated from outside and citizens can develop a better-informed and more relaxed attitude toward their health.

Address correspondence to Gerd Gigerenzer, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin; e-mail: gigerenzer@mpib-berlin.mpg.de.

INTRODUCTION

In a 2007 campaign advertisement, former New York City mayor Rudy Giuliani said, "I had prostate cancer, 5, 6 years ago. My chance of surviving prostate cancer—and thank God, I was cured of it—in the United States? Eighty-two percent. My chance of surviving prostate cancer in England? Only 44 percent under socialized medicine" (Dobbs, 2007). For Giuliani, these health statistics meant that he was lucky to be living in New York and not in York, since his chances of surviving prostate cancer appeared to be twice as high. This was big news. As we will explain, it was also a big mistake. High-profile politicians are not the only ones who do not understand health statistics or misuse them.

In this monograph, we—a team of psychologists and physicians—describe a societal problem that we call *collective statistical illiteracy*. In *World Brain* (1938/1994), H.G. Wells predicted that for an educated citizenship in a modern democracy, statistical thinking would be as indispensable as reading and writing. At the beginning of the 21st century, nearly everyone living in an industrial society has been taught reading and writing but not statistical thinking—how to understand information about risks and uncertainties in our technological world. The qualifier *collective* signals that lack of understanding is not limited to patients with little education; many physicians do not understand health statistics either. Journalists and politicians further contribute to the problem. One might ask why collective statistical illiteracy is not a top priority of ethics committees, medical curricula, and psychological research. One reason is that its very nature generally ensures that it goes undetected. Many of our readers might not have sensed that anything was wrong with Giuliani's conclusion, had we not highlighted it. Humans are facing a concealed societal problem.

In this monograph, we define statistical illiteracy in health care and analyze its prevalence, the damage it does to health and emotion, its potential causes, and its prevention. We argue that its causes are not simply inside the minds of patients and physicians—such as the lack of a math gene or a tendency to make hard-wired cognitive biases. Rather, we show that statistical literacy is largely a function of the outside world and that it can be fostered by education and, even more simply, by representing numbers in ways that are transparent for the human mind. To give the reader a sense of the problem, we begin with three examples.

I. STATISTICAL ILLITERACY IN PATIENTS, PHYSICIANS, AND POLITICIANS

The three cases that follow illustrate the three main points in this monograph: Statistical illiteracy (a) is common to patients, physicians, and politicians; (b) is created by nontransparent framing of information that may be unintentional (i.e., a result of lack of understanding) or intentional (i.e., an effort to manipulate or persuade people); and (c) can have serious consequences for health.

The Contraceptive Pill Scare
In October 1995, the U.K. Committee on Safety of Medicines issued a warning that third-generation oral contraceptive pills increased the risk of potentially life-threatening blood clots in the legs or lungs twofold—that is, by 100%. This information was passed on in "Dear Doctor" letters to 190,000 general practitioners, pharmacists, and directors of public health and was presented in an emergency announcement to the media. The news caused great anxiety, and distressed women stopped taking the pill, which led to unwanted pregnancies and abortions (Furedi, 1999).
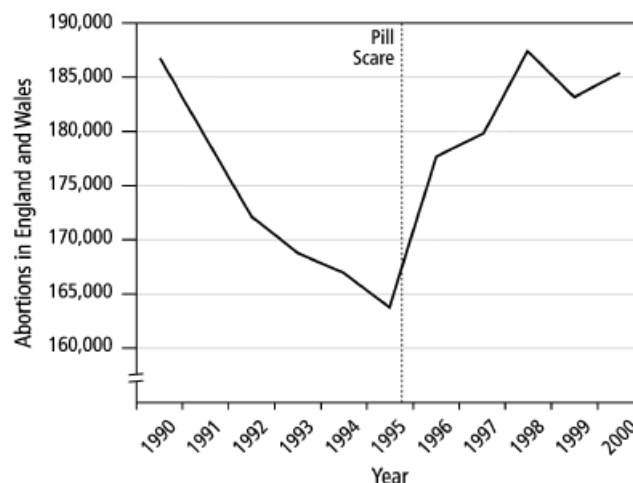

Fig. 1. Reversal of downward trend in number of abortions in England and Wales following the 1995 pill scare.

How big is 100%? The studies on which the warning was based had shown that of every 7,000 women who took the earlier, second-generation oral contraceptive pills, about 1 had a thrombosis; this number increased to 2 among women who took third-generation pills. That is, the *absolute risk* increase was only 1 in 7,000, whereas the *relative* increase was indeed 100%. Absolute risks are typically small numbers while the corresponding relative changes tend to look big—particularly when the base rate is low. Had the committee and the media reported the absolute risks, few women would have panicked and stopped taking the pill.

The pill scare led to an estimated 13,000 additional abortions (!) in the following year in England and Wales. Figure 1 shows that, before the alert, abortion rates had been on the decline since 1990, but afterwards, this trend was reversed (Furedi, 1999). Women's confidence in oral contraceptives was undermined, and pill sales fell sharply. For every additional abortion, there was also one extra birth, and the increase in both was particularly pronounced in teenagers, with some 800 additional conceptions among girls under 16. The resulting cost increase to the National Health Service for abortion provision has been estimated at about £46 million ($70 million at that time). Ironically, abortions and pregnancies are associated with an increased risk of thrombosis that exceeds that of the third-generation pill. The pill scare hurt women, hurt the National Health Service, and even hurt the pharmaceutical industry. Among the few to profit were the journalists who got the story on the front page.

The 1995 pill scare was not the first one. Similar scares had occurred in 1970 and 1977, and after each one, the abortion rate rose (Murphy, 1993). And most likely, the 1995 scare will not be the last. Few citizens know the simple distinction between a relative increase ("100% higher") and an absolute increase ("1 in 7,000"). Medical journals, information brochures, and the

media continue to inform the public in terms of relative changes, if only because big numbers make better headlines and generate more attention. But big numbers can also raise unnecessary anxieties and unrealistic hopes. When the next scare arrives, teenagers and adults will be as unprepared as ever to understand health statistics, creating another wave of abortions.

Few Gynecologists Understand Positive Mammograms

Since a large proportion of women participate in mammography screening, a key health statistic each gynecologist needs to know is the chances that a woman who tests positive actually has breast cancer. Mammography generates many false alarms. To avoid unnecessary anxiety or panic, women have a right to be informed what a positive test result means. Think of a woman who just received a positive screening mammogram and asks her doctor: Do I have breast cancer for certain, or what are the chances? Ninety-nine percent, 90%, 50%, or perhaps less? One would assume that every physician knows the answer. Is that so?

One of us (GG) trained about 1,000 gynecologists in risk communication as part of their continuing education in 2006 and 2007. At the beginning of one continuing-education session in 2007, 160 gynecologists were provided with the relevant health statistics needed for calculating the chances that a woman with a positive test actually has the disease:

> Assume you conduct breast cancer screening using mammography in a certain region. You know the following information about the women in this region:
>
> - The probability that a woman has breast cancer is 1% (prevalence)
> - If a woman has breast cancer, the probability that she tests positive is 90% (sensitivity)
> - If a woman does not have breast cancer, the probability that she nevertheless tests positive is 9% (false-positive rate)

A woman tests positive. She wants to know from you whether that means that she has breast cancer for sure, or what the chances are. What is the best answer?

> A. The probability that she has breast cancer is about 81%.
> B. Out of 10 women with a positive mammogram, about 9 have breast cancer.
> C. Out of 10 women with a positive mammogram, about 1 has breast cancer.
> D. The probability that she has breast cancer is about 1%.

Gynecologists could derive the answer from the health statistics provided, or they could simply recall what they should have known anyhow. In either case, the best answer is C—that is, that only about 1 out of every 10 women who test positive in screening actually has breast cancer. The other 9 are falsely alarmed (Kerlikowske, Grady, Barclay, Sickles, & Ernster, 1996a, 1996b). Note that the incorrect answers were spaced about an order of magnitude away from the best answer, in order
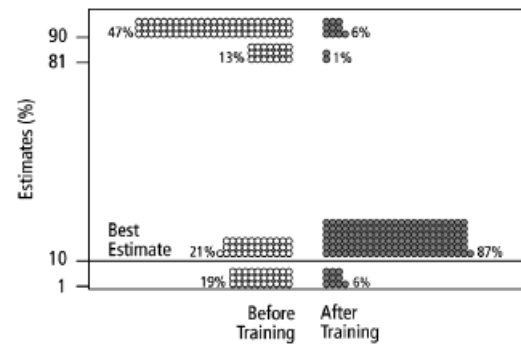

Fig. 2. Estimates by 160 gynecologists of the probability that a woman has breast cancer given a positive mammogram, before and after receiving training in how to translate conditional probabilities into natural frequencies.

to make it easier for the doctors. Figure 2 (left side) shows the 160 gynecologists' answers prior to training. Disconcertingly, the majority of them grossly overestimated the probability of cancer, answering "90%" or "81%." Another troubling result was the high variability in physicians' estimates, ranging between a 1% and 90% chance of cancer. The number of physicians who found the best answer, as documented in medical studies, was slightly less than chance (21%).

Do these physicians lack a gene for understanding health statistics? No. Once again, health statistics are commonly framed in a way that tends to cloud physicians' minds. The information is presented in terms of *conditional probabilities*—which include the sensitivity and the false-positive rate (or 1 − specificity). Just as absolute risks foster greater insight than relative risks do, there is a transparent representation that can achieve the same in comparison to conditional probabilities: what we call *natural frequencies*. Here is the same information from the above problem translated into natural frequencies:

> Assume you conduct breast cancer screening using mammography in a certain region. You know the following information about the women in this region:
>
> - Ten out of every 1,000 women have breast cancer
> - Of these 10 women with breast cancer, 9 test positive
> - Of the 990 women without cancer, about 89 nevertheless test positive

After learning during the training session how to translate conditional probabilities into natural frequencies, the gynecologists' confusion disappeared; 87% of them now understood that 1 in 10 is the best answer (Fig. 2, right). How can this simple change in representation turn their innumeracy into insight? The reason is that natural frequencies facilitate computation, as explained in Figure 3. Natural frequencies represent the way humans encoded information before mathematical probabilities were invented in the mid-17th century and are easy to "digest" by our brains. Unlike relative frequencies and conditional probabilities, they are simple counts that are not normalized
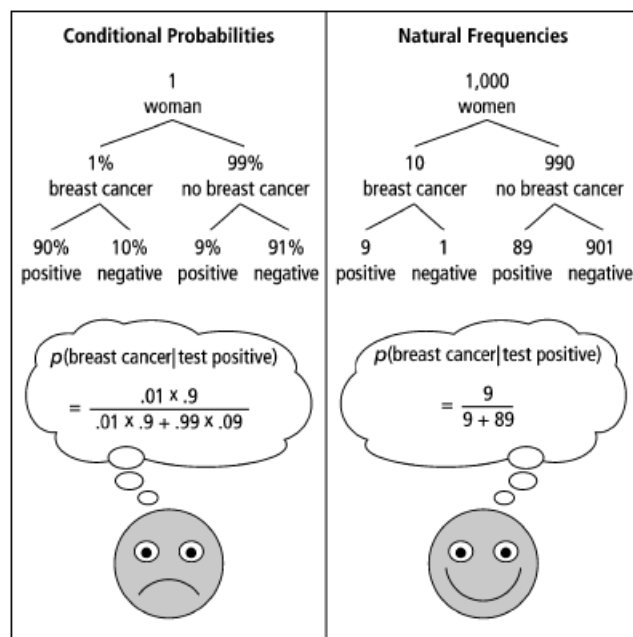
Conditional Probabilities

1
woman

1%
breast cancer

99%
no breast cancer

90%
positive

10%
negative

9%
positive

91%
negative

$p(\text{breast cancer}|\text{test positive})$

$= \dfrac{.01 \times .9}{.01 \times .9 + .99 \times .09}$

Natural Frequencies

1,000
women

10
breast cancer

990
no breast cancer

9
positive

1
negative

89
positive

901
negative

$p(\text{breast cancer}|\text{test positive})$

$= \dfrac{9}{9 + 89}$

Fig. 3. Two ways of calculating the probability that a woman who tests positive in mammography screening actually has breast cancer (positive predictive value). The left side illustrates the calculation with conditional probabilities, and the right side with natural frequencies. The four probabilities at the bottom of the left tree are conditional probabilities, each normalized on base 100. The four frequencies at the bottom of the right tree are natural frequencies. The calculation using natural frequencies is simpler (smiling face) because natural frequencies are not normalized relative to base rates of breast cancer, whereas conditional probabilities (or relative frequencies) are, and need to be multiplied by the base rates. (The formula to calculate the positive predictive value is known as *Bayes's rule*.)

with respect to base rates (Gigerenzer & Hoffrage, 1995, 1999). That is, the four natural frequencies in Figure 3 (right side: 9; 1; 89; and 901) add up to the total number of 1,000 women, whereas the four conditional probabilities (left side) do not add up to 100%—instead each pair is normalized with respect to the base rates of cancer or no cancer, respectively.

This study illustrates a fundamental problem in health care: Many physicians do not know the probabilities that a person has a disease given a positive screening test—that is, the *positive predictive value*. Nor are they able to estimate it from the relevant health statistics when those are framed in terms of conditional probabilities, even when this test is in their own area of specialty (Hoffrage & Gigerenzer, 1998). If you want to find out yourself if this is the case, ask your doctor. The result also shows that there is a fast and efficient cure. Yet doctors' and patients' collective innumeracy is a largely unknown problem in health care that continues to cause undue fear in the public. Months after receiving a false-positive mammogram, 1 in 2 women reported considerable anxiety about mammograms and breast cancer, and 1 in 4 reported that this anxiety affected their daily mood and functioning (Lerman et al., 1991). Everyone who participates in screening should be informed that the majority of sus-

picious results are false alarms. We face a large-scale ethical problem for which an efficient solution exists yet which ethics committees, focusing their attention instead on stem cells, abortion, and other issues that invite endless debates, have not yet noticed.

Higher Survival Does Not Mean Longer Life

Back to Rudy Giuliani. While running for president, Giuliani claimed that health care in the United States was superior to health care in Britain. Giuliani apparently used data from the year 2000, when 49 British men per 100,000 were diagnosed with prostate cancer, of which 28 died within 5 years—about 44%. Using a similar approach, he cited a corresponding 82% 5-year survival rate in the United States, suggesting that Americans with prostate cancer were twice as likely to survive as their British counterparts. Giuliani's numbers, however, are meaningless for making comparisons across groups of people that differ dramatically in how the diagnosis is made. In the United States, most prostate cancer is detected by screening for prostate-specific antigens (PSA), while in the United Kingdom, most is diagnosed by symptoms. The bottom line is that to learn which country is doing better, you need to compare mortality rates. To understand why, it is helpful to look at how "5-year survival" and mortality statistics are calculated. We'll start with survival.

Five-year survival is the most common survival statistic, but there is nothing special about 5 years. The statistic can be calculated for any time frame. Imagine a group of patients all diagnosed with cancer on the same day. The proportion of these patients who are still alive 5 years later is the 5-year survival rate. Here is the formula for the statistic:

$$5\text{-year survival rate} = \frac{\substack{\text{number of patients diagnosed with cancer} \\ \text{still alive 5 years after diagnosis}}}{\text{number of patients diagnosed with cancer}}$$

To calculate a mortality rate, imagine another group of people. The group is *not* defined by a cancer diagnosis. The proportion of people in the group who are dead after 1 year (the typical time frame for mortality statistics) is the "mortality rate." Here is the formula:

$$\text{Annual mortality rate} = \frac{\text{number of people who die from cancer over 1 year}}{\text{number of people in the group}}$$

The key difference to notice between these two kinds of statistics is the word *diagnosed*, which appears in the numerator and denominator of survival statistics but nowhere in the definition of mortality. Screening profoundly biases survival in two ways: (a) It affects the timing of diagnosis and (b) it affects the nature of diagnosis by including people with nonprogressive cancer. The first is called the *lead-time bias*, illustrated in Figure 4. Imagine a group of prostate cancer patients currently diagnosed at age 67, all of whom die at age 70. Each survived only 3 years, so the 5-year survival of this group is 0%. Now imagine that the same group is diagnosed with prostate cancer by PSA tests earlier, at age 60, but they all still die at age 70. All have now survived 10
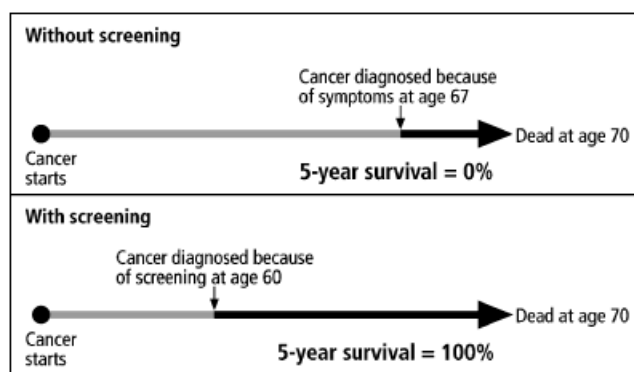
Fig. 4. Lead-time bias. Even if the time of death in not changed by screening—and thus no life is saved or prolonged—advancing the time of diagnosis in this way can result in increased 5-year survival rates, causing such statistics to be misleading.

years and thus their 5-year survival rate is 100%. Even though the survival rate has changed dramatically, nothing has changed about the time of death: Whether diagnosed at age 67 or at age 60, all patients die at age 70. This simple example demonstrates how survival rates can be increased by setting the time of diagnosis earlier, even if no life is prolonged or saved.

The second phenomenon that leads to spuriously high survival rates is the *overdiagnosis bias*, illustrated in Figure 5. Overdiagnosis is the detection of pseudodisease—screening-detected abnormalities that meet the pathologic definition of cancer but will never progress to cause symptoms in the patient's lifetime. These are also called nonprogressive cancers. Figure 5 (top) shows 1,000 men with progressive cancer who do not undergo screening. After 5 years, 440 are still alive, which results in a survival rate of 44%. Figure 5 (bottom) shows a population of men who participate in PSA screening and have cancer. The test detects both people with progressive and those with nonprogressive cancer. Imagine that screening detects 2,000 people with nonprogressive cancers—who by definition will not die of cancer in the following 5 years. These are now added to the 440 who survived progressive cancer, which inflates the survival rate
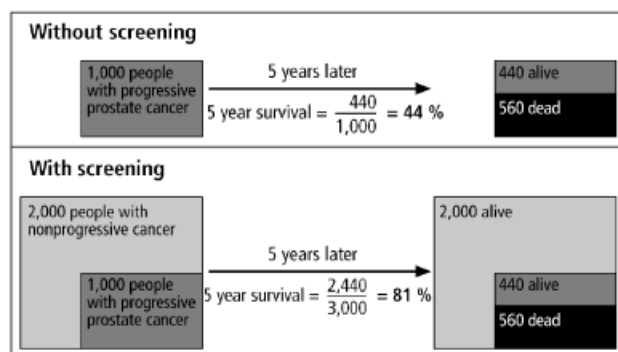

Fig. 5. Overdiagnosis bias. Even if the number of people who die is not changed by screening—and thus no life is saved or prolonged—screening-detected nonprogressive cancers can inflate the 5-year survival rates, causing such statistics to be misleading.

to 81%. Note that even though the survival rate has changed dramatically, the number of people who die has not changed at all.

While the concept of nonprogressive cancer may seem implausible to many people including clinicians, basic scientists have begun to uncover biological mechanisms that halt the progression of cancer (Folkman & Kalluri, 2004; Mooi & Peeper, 2006; Serrano, 2007). These mechanisms apply to many cancers—including one of the most dreaded, lung cancer. Amazingly, with computed tomography (CT) screening, almost as many nonsmokers were found to have lung cancer as smokers (Sone et al., 2001). Given that smokers are 15 times as likely to die from lung cancer, the computed tomography (CT) scans had to be finding abnormalities in nonsmokers that were technically cancer (based on their microscopic appearance) but that did not behave in the way lung cancer is expected to behave—as a progressive disease that ultimately kills (see also Welch, Woloshin, et al., 2007).

Due to overdiagnosis and lead-time bias, changes in 5-year survival rates have no reliable relationship to changes in mortality. For example, consider the 20 most common solid tumors in the United States over the last 50 years. Changes in 5-year survival were completely uncorrelated with changes in mortality (correlation coefficient = 0.0). That means that knowing about changes in survival tells you nothing about changes in mortality (Welch, Schwartz, & Woloshin, 2000)! In the context of screening, survival is always a biased metric. In the United States, screening for prostate cancer using the PSA test began in the late 1980s and spread rapidly, despite the lack of evidence that it saves lives. As a result, the number of new prostate cancer diagnoses soared. In Britain, PSA testing was introduced later and is still not routinely used. Consequently, new prostate cancer diagnoses (i.e., incidence) in Britain have risen only slightly. This largely explains why 5-year survival for prostate cancer is so much higher in the United States. The most recent figures (which differ from those cited by Giuliani) are 98% 5-year survival in the United States versus 71% in Britain.

But the real story is about mortality: Are American men half as likely to die from prostate cancer as British men are? The answer is no; the risk is about the same: About 26 prostate cancer deaths per 100,000 American men versus 27 per 100,000 in Britain (Shibata & Whittemore, 2001). If we use Giuliani's concern with prostate cancer for judging a health-care system, the "socialist" English system appears to win since there are fewer diagnoses— that is, less overdiagnosis—but about the same mortality rate. Looking at the incidence and mortality data together suggests that many American men have been unnecessarily diagnosed (i.e., overdiagnosed) with prostate cancer during the PSA era and have undergone unnecessary surgery and radiation treatment, which often leads to impotence and/or incontinence.

Giuliani is not the only politician who has failed to appreciate the difference between survival rates and mortality rates. A recent report by the U.K. Office for National Statistics on cancer-

survival trends noted that 5-year survival for colon cancer was 60% in the United States compared to 35% in Britain. Experts dubbed this finding "disgraceful" and called for government spending on cancer treatment to be doubled. In response, then-Prime Minister Tony Blair set a target to increase survival rates by 20% over the next 10 years, saying, "We don't match other countries in its prevention, diagnosis and treatment" (Steimle, 1999, p. 1189). In fact, despite these large differences in 5-year survival, the mortality rate for colon cancer in Britain is about the same as the rate in the United States.

Conclusion

These three examples illustrate the theme of this monograph: the collective statistical illiteracy of patients, physicians, and politicians, as well as the considerable costs health systems pay as a consequence. The more widespread this illiteracy, the easier it is to manipulate the opinions of both doctors and patients, such as through campaigns promoting screening based on improved 5-year survival (see Part IV). We have also shown that there is a cure to this phenomenon that would be easy to implement: using transparent health statistics instead of the prevalent confusing ones, such as absolute risks instead of relative risks, natural frequencies instead of conditional probabilities, and mortality rates instead of 5-year survival rates when judging the value of screening (see Part VI). Framing information in a way that is most readily understood by the human mind is a first step toward educating doctors and patients in risk literacy.

II. WHAT IS STATISTICAL LITERACY?

Statistical literacy in health does not require a degree in statistics. Rather, it means that citizens have basic competencies in understanding health statistics. For instance, statistical literacy implies that a person would recognize that comparing survival rates across countries where screening practices differ dramatically is nonsense and that the statistics cited by Giuliani do not mean that men in the United States are better off than in the United Kingdom.

It is desirable to define statistical literacy in concrete terms. We are aware that one could come up with a long textbook-like list, but a curriculum in statistics is precisely not our intention. What we are instead looking for are insights that can be taught in a short time and whose efficacy has been proven by psychological studies. To this end, we propose a list of insights that all patients and physicians should understand and questions that everyone should know to ask. We call this *minimal statistical literacy in health*.

Minimal Statistical Literacy in Health

Minimum statistical literacy applies to every medical decision, from whether a child's tonsils should be removed to whether an adult should take cholesterol-lowering medication. Minimal literacy focuses on the main concepts (like absolute risks) rather than the more advanced topics of variability (e.g., confidence intervals). Tables 1 and 2 serve as an illustration.

Learning to Live With Uncertainty
Understand that there is no certainty and no zero-risk, but only risks that are more or less acceptable.

For instance, the risk chart in Table 1 shows that women who never smoked have a much smaller risk of lung cancer than do smokers, but that risk still is not zero. Similarly, women with breast cancer genes BRCA-1 or BRCA-2, who face a high risk of breast cancer, do not necessarily develop breast cancer. And women who undergo radical bilateral mastectomy—despite lowering their breast cancer risk—can still develop it (Hartmann et al., 1999).

Questions to Ask About All Risks
Risk of what? Understand the outcome to which the risk refers. For instance, the numbers in Table 1 refer to dying from disease, not getting the disease or developing a symptom.

Time frame? Understand the time the risk refers to. The frequencies of dying in Table 1 refer to a period of 10 years for all age groups. Time frames such as the "next 10 years" are easier to imagine than the widely used "lifetime" risks, are more informative because risks change over time, and are long enough to enable action being taken.

How big? Since there are no zero risks, size is what matters. Size should be expressed in absolute terms (e.g., 13 out of 1,000 women smokers age 50 die of heart disease within 10 years; see Table 1) or in comparative terms, relating the risk to a more familiar one. For example, for a 55-year-old American woman who is a smoker, the risk of dying from lung cancer in the next 10 years is about 10 times as high as dying from a car accident during the same time.

Does it apply to me? Check to see whether the risk information is based on studies of people like you—people of your age or sex, or people with health problems similar to yours. Table 1 shows that age matters for all causes of death, whereas whether one is a smoker or not is relevant for lung cancer but not colon cancer.

Screening Tests
Understand that screening tests may have benefits and harms. Benefits include the possibility of finding disease earlier, when treatment may be less invasive and/or more effective. Harms include costs, inconvenience, and false alarms—and in our view, the most important harm of overdiagnosis. Overdiagnosis can be defined as the detection of pseudodisease or abnormalities that would never progress to cause symptoms in the patient's lifetime. For instance, it has been estimated that about 25% of breast cancers detected by mammography are overdiagnoses (Schwartz & Woloshin, 2007). The best evidence for overdiagnosis in lung cancer comes from studies of CT scans,

TABLE 1

Risk Chart for U.S. Women and Smoking (from Woloshin, Schwartz, & Welch, 2008)

Find the line closest to your age and smoking status[†]. The numbers tell you **how many of 1,000 women will die in the next 10 years from....**

Age	Smoking	Vascular Disease		Cancer					Infection			Lung Disease	Accidents	All Causes Combined
		Heart Disease	Stroke	Lung Cancer	Breast Cancer	Colon Cancer	Ovarian Cancer	Cervical Cancer	Pneumonia	Flu	AIDS	COPD		
35	Never smoker	1			1						1		2	14
	Smoker	**1**	**1**	**1**	**1**						**1**		**2**	**14**
40	Never smoker	1			2	1	Fewer than 1 death				1		2	19
	Smoker	**4**	**2**	**4**	**2**						**1**	**1**	**2**	**27**
45	Never smoker	2	1	1	3	1	1				1		2	25
	Smoker	**9**	**3**	**7**	**3**	**1**	**1**		**1**		**1**	**2**	**2**	**45**
50	Never smoker	4	1	1	4	1	1						2	37
	Smoker	**13**	**5**	**14**	**4**	**1**	**1**		**1**			**4**	**2**	**69**
55	Never smoker	8	2	2	6	2	2	1	1		1		2	55
	Smoker	**20**	**6**	**26**	**5**	**2**	**2**	**1**	**1**		**9**		**2**	**110**
60	Never smoker	14	4	3	7	3	3	1	1		2		2	84
	Smoker	**31**	**8**	**41**	**6**	**3**	**3**	**1**	**2**		**18**		**2**	**167**
65	Never smoker	25	7	5	8	5	4	1	2		3		3	131
	Smoker	**45**	**15**	**55**	**7**	**5**	**3**	**1**	**4**		**31**		**3**	**241**
70	Never smoker	46	14	7	9	7	4	1	4		5		4	207
	Smoker	**66**	**25**	**61**	**8**	**6**	**4**	**1**	**7**		**44**		**4**	**335**
75	Never smoker	86	30	7	10	10	5	1	8		6		7	335
	Smoker	**99**	**34**	**58**	**10**	**9**	**4**		**14**		**61**		**7**	**463**

Note: Grey shading means fewer than 1 death per 1000 women.

† A never smoker has smoked less than 100 cigarettes in her life and a current smoker has smoked at least 100 cigarettes or more in her life and smokes (any amount) now.

which detected almost 10 times the amount of lung cancer than X-rays and, as mentioned before, diagnosed almost as many nonsmokers as smokers as having lung cancer (Sone et al., 2001).

Overdiagnosis leads to harm through overtreatment. The treatment of nonprogressive cancers results in unnecessary surgery and other invasive treatments—treatments that can only harm patients since they are being treated for a "disease" that would never have harmed them if left untreated.

Understand that screening tests can make two errors: false positives and false negatives. A false positive (false alarm) occurs when a test is positive (for example, a test for Down syndrome) in people who do not have the disease (no Down syndrome present). The false-positive rate is the proportion of positive tests among clients without the condition (Table 2). A false negative (miss) occurs when a test is negative in someone who does have the disease. The false-negative rate (miss rate) is the proportion of negative tests among clients with the condition.

Understand how to translate specificities, sensitivities, and other conditional probabilities into natural frequencies. Specificities and sensitivities continue to confuse physicians and patients alike. The specificity is the proportion of negative tests among clients without the condition; the sensitivity is the proportion of positive tests among clients with the condition (Table 2). Figure 3 illustrates how these can be translated into natural frequencies in order to facilitate deriving the positive predictive value.

TABLE 2

Four Possible Test Outcomes

Test result	Down syndrome	
	Yes	No
Positive	82% Sensitivity	8% False-positive rate
Negative	18% False-negative rate	92% Specificity

Note. Testing for a disease (here: Down syndrome by measuring fetal nuchal-translucency thickness) can have four possible outcomes: a positive result given disease, a positive result given no disease, a negative result given disease, and a negative result given no disease. The rates with which these four results occur are called sensitivity (or true positive rate), false positive rate, false negative rate, and specificity (true negative rate). The two shaded areas indicate the two possible errors, false positives and false negatives (data adopted from Snijders, Noble, Sebire, Souka, & Nicolaides, 1998).

Understand that the goal of screening is not simply the early detection of disease; it is mortality reduction or improvement of quality of life. Screening is testing for hidden disease in people without symptoms. It is only useful if early detection results in earlier treatment that is more effective or safer than later treatment. For instance, many smokers, current and past, wonder whether to get a CT scan to screen for lung cancer. While CT scans can clearly find more early-stage cancers, there is no evidence for reduced mortality rates. That is why no professional group currently recommends the test (in fact the American College of Chest Physicians now recommends against routine CT screening).

Treatment
Understand that treatments typically have benefits and harms. Benefits include risk reduction—the lower probability of experiencing a feared outcome, such as getting or dying from disease. Treatment harms include bothersome or potentially even life-threatening side effects that result from medications or surgery. The value of treatment is determined by comparing the benefits (i.e., how much risk there is to reduce) and the harms.

Understand the size of the benefit and harm. Always ask for absolute risks (not relative risks) of outcomes with and without treatment.

Questions About the Science Behind the Numbers
Quality of evidence? A basic distinction is between evidence from a properly randomized controlled trial (Grade I evidence), well-designed cohort or case-control studies without randomization (Grade II), and opinions from respected authorities based on clinical experience (Grade III).

What conflicts of interest exist? Conflicts of interest can be inferred from the source that funded the study or from the goals of the institution that advertised the health statistics (see Part V).

III. HOW WIDESPREAD IS STATISTICAL ILLITERACY?

In health care, statistical illiteracy is typically presented as a problem faced by patients, sometimes by the media, and almost never by physicians. In this section, we analyze the collective statistical illiteracy of all three groups.

Do Patients Understand Health Statistics?

A citizen in a modern technological society faces a bewildering array of medical decisions. Should a pregnant woman undergo prenatal screening for chromosomal anomalies at age 35? Should parents send their teenage daughters for cervical cancer vaccination using Gardasil, despite reports that the vaccine could lead to paralysis? Whom should one trust? If citizens want to make informed decisions, they need more than trust: They need to understand health statistics. The evidence in this section documents, however, that most citizens (a) are not aware of basic health information, (b) do not understand the numbers if they encounter the information, and (c) tend to cherish the illusion of certainty about diagnostic results and treatments or follow the heuristic "trust your doctor"—both of which make risk literacy appear of little relevance. What follows is not an exhaustive overview but an analysis of the main issues. We begin with an elementary skill, called *basic numeracy.*

Basic Numeracy
To analyze the prevalence of low numeracy and gauge the extent to which it impairs communication about health risks, Schwartz, Woloshin, Black, and Welch (1997) developed a simple three-question scale. The first question tests the respondent's ability to convert a percentage to a concrete number of people (out of 1,000), the second tests the ability to translate in the other direction, and the third tests basic familiarity with chance outcomes (Table 3). The test was applied to a random sample of female veterans in New England, 96% of whom were high-school graduates, and whose average age was 68. Forty-six percent were unable to convert 1% to 10 in 1,000, 80% were unable to convert 1 in 1,000 to 0.1%, and 46% were unable to correctly estimate how many times a coin would likely come up

TABLE 3
The Basic Numeracy Assessment Scale

Task	Question
Convert a percent to a proportion	1. A person taking Drug A has a 1% chance of having an allergic reaction. If 1,000 people take Drug A, how many would you expect to have an allergic reaction? —person(s) out of 1,000
Convert a proportion to a percent	2. A person taking Drug B has a 1 in 1,000 chance of an allergic reaction. What percent of people taking Drug B will have an allergic reaction? —%
Basic probability	3. Imagine that I flip a coin 1,000 times. What is your best guess about how many times the coin would come up heads in 1,000 flips? —times out of 1,000

TABLE 4

Percentage of U.S. Adults Aged 35 to 70 Giving Correct Answers to Basic Numeracy Questions (See Table 3), Overall and by Education Level

Question	Overall $n = 450$	Educational attainment			
		High school diploma or less $n = 131$	Some college $n = 151$	College degree $n = 103$	Post-graduate degree $n = 62$
Convert 1% to 10 in 1,000	70	60	68	79	82
Convert 1 in 1,000 to 0.1%	25	23	21	30	27
How many heads in 1,000 coin flips?	76	62	76	87	86

Note. Schwartz & Woloshin (2000). Based on a nationally representative sample of U.S. citizens.

heads in 1,000 flips, with the most common incorrect answers being 25, 50, and 250. The women's scores on this test strongly correlated with their ability to accurately interpret the benefit of mammography after being presented with standard risk-reduction information: Only 6% of women answering just one basic numeracy question correctly could accurately interpret the data, compared to 40% of those answering all three questions correctly. Thus, basic numeracy seems to be a necessary precondition for minimal statistical literacy.

Table 4 shows the prevalence of low numeracy skills among U.S. adults—overall and stratified by educational attainment. The skills of the general adult public with high-school education correspond roughly to those of the female veterans, whereas the skills of people with higher education are better on average. Note again the great difficulty large parts of the public, like the female veterans, have with translating small frequencies into percentages. Only 25% of the population could correctly convert 1 in 1,000 to 0.1%. Even among the highest education groups, at most 30% could solve this translation task. Lipkus, Samsa, and Rimer (2001) even found that only 21% of well-educated adults could answer this question correctly.

Medical Data Interpretation Test
To test beyond basic numeracy, Schwartz, Woloshin, and Welch (2005) developed the medical data interpretation test (which includes some of the minimum statistical literacy introduced above). Its goal is to test the ability to make comparisons, such as between treatments—a fundamental requirement for informed decision making. Table 5 shows the answers of 178 participants with a broad range of educational attainment and backgrounds (recruited from advertisements in local newspapers, an outpatient clinic, and a hospital open house; the individual multiple-choice questions can be found in Schwartz et al., 2005). Item nonresponses ("left blank") were low, suggesting that respondents understood the questions. Item difficulty varied widely, from 20% to 87% correct answers. The item that proved most difficult for the participants was number 5 in the section "knowledge basis for comparisons." The multiple-choice question was: "Which piece of information would be the best evidence that Gritagrel [a new drug against strokes] helped

people?" Seventy percent of participants chose the answer "Fewer people died from strokes in the Gritagrel group than in the placebo group" and only 20% correctly chose "Fewer people died for any reason in the Gritagrel group than in the placebo group." The distinction is important. Few medications have been shown to reduce the chance of death overall, and such a finding would reassuringly mean that (at least in this study) Gritagrel had no life-threatening side effects that substituted death from stroke with death from another cause. The medical data interpretation test appears to have reasonable reliability and validity (Schwartz et al., 2005).

There is no single study that tests all aspects of minimal statistical literacy, and in what follows we review studies that address selected issues.

The Illusion of Certainty
The first item in minimal statistical literacy is learning to live with uncertainty. To appreciate the importance of health statistics, patients need to understand that there is no certainty in the first place. As Benjamin Franklin (1789/1987) once said: "In this world, there is nothing certain but death and taxes." The term *illusion of certainty* refers to an emotional need for certainty when none exists. This feeling can be attached to test results that are taken to be absolutely certain and to treatments that appear to guarantee a cure.

Even very good tests make errors. For instance, a 36-year-old American construction worker tested negative on ELISA tests 35 times before it was established that he was infected with HIV (Reimer et al., 1997). A series of what appears to be 35 misses in a row is an extreme case. Yet in one-time applications of tests, both false positives and misses are typical. In a nationwide survey in 2006, 1,000 German citizens over 18 were asked: "Which of the following tests are absolutely certain?" (Fig. 6). While only 4% believed an expert horoscope to give absolutely accurate results, a majority of citizens believed that HIV tests, fingerprints, and DNA tests were absolutely certain, even though none of these are (Gigerenzer, 2002, 2008). In contrast to these tests, which tend to make relatively few errors, the much less reliable result of a mammography (positive or negative mammogram) was rated as "absolutely certain" by 46% of the women

TABLE 5

Proportion of Correct, Incorrect, and Missing Answers to the 18 Items on the Medical Data Interpretation Test for 178 Participants

Knowledge basis for comparisons	Answered correctly (%)	Answered incorrectly (%)	Left blank (%)
Know that a denominator is needed to calculate risk	75	24	1
Know that denominators are needed to compare risks in 2 groups	45	54	1
Know that the base rate is needed in addition to relative risk to determine the magnitude of benefit	63	36	1
Know that a comparison group is needed to decide whether benefit exists	81	18	1
Know that lowering all-cause mortality provides better evidence of benefit than lowering a single cause of death	20	79	1
Comparison tasks			
Select "1 in 296" as a larger risk than "1 in 407"	85	14	1
Inferred items[a]			
Rate the riskiness of a 9 in 1,000 chance of death as the same as a 991 in 1,000 chance of surviving	61	37	2
Select a larger risk estimate for deaths from all causes than deaths from a specific disease	30	69	1
Select a larger risk estimate for a 20-year risk than for a 10-year risk	39	60	1
Calculations related to comparisons			
Calculate risk in intervention group by applying relative risk reduction to a baseline risk	87	11	2
Calculate 2 absolute risk reductions from relative risk reductions and baseline risks and select the larger	80	19	1
Calculate relative risk reduction from 2 absolute risks	52	46	2
Calculate absolute risk reduction from 2 absolute risks	77	19	4
Calculate the number of events by applying absolute risk to number in group	72	22	6
Context for comparisons			
Know that age and sex of individuals in the source data are needed	47	51	2
Know that age of individuals in the source data is needed	60	39	1
Know that risk of other diseases is needed for context	62	35	3
Know that, for male smokers, the risk of lung cancer death is greater than prostate cancer death	60	37	3

Note. [a]These items were based on a total of 5 separate questions.

and by 42% of the men. Yet its miss rate is about 10%, and the false-positive rate is almost as high. A university education is only a slight safeguard against the illusion of certainty: One out
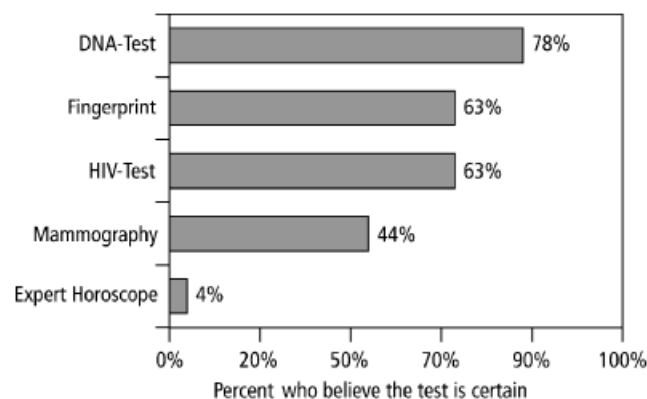

Fig. 6. The illusion of certainty. Shown are results from face-to-face interviews conducted in 2006, in which a representative sample of 1,016 German citizens was asked: ''Which of the following tests are absolutely certain?'' (Gigerenzer, 2008).

of three women with a university degree also believed that mammograms are absolutely certain.

When women participate in a 10-year program of annual mammography, the chances of a false alarm multiply: Every other woman without cancer can expect one or more false-positive test results (Elmore et al., 1998). Schwartz, Woloshin, Sox, Fischhoff, & Welch (2000) asked a stratified sample of 479 American women without breast cancer to estimate the chance of a false-positive result during a 10-year program. The median answer was 20% (an underestimate, but in the right ballpark), with 99% of the women believing that false positives occur. The fact that so many German women say that a singular test result is absolutely certain, whereas almost all the American women respond that false positives can occur in a series of 10 tests, may be related to the different perception of the singular as opposed to the repeated test. At the same time, given that the German women were asked for certainty of result and most mammography results are negative, their response may largely reflect the belief that if the test result is negative, one can be sure of not having cancer. In fact, many women say that they participate in

screening to be sure that they do not have cancer. Similarly, genetic testing is often perceived as infallible: In a survey in the Netherlands, one third of respondents failed to understand that a prenatal test such as amniocentesis is not absolutely certain, as well as that if a person has a genetic predisposition for a disease, this person will not necessarily get the disease (Henneman, Timmermans, & van der Wal, 2004, pp. 11–12).

The illusion of certainty may also result from confusion between early detection and prevention. Pro-screening campaigns in various countries have used the term "cancer prevention," wrongly suggesting that early detection could prevent the risk of getting cancer. In a cross-cultural study, over 4,000 randomly sampled women aged 15 and above were asked whether it is correct that "regular mammography every 2 years in women who are well prevents the risk of contracting breast cancer" or that mammography "reduces the risk" or "does not have any influence on the risk" (the correct answer). Noteworthy proportions of women in Switzerland (10%), the United Kingdom (17%), the United States (26%), and Italy (33%) shared the illusion of certainty that screening would prevent cancer (Domenighetti et al., 2003).

Screening is intended to detect existing cancers at an early stage. So it does not reduce the risk of getting breast cancer; it increases the number of positive diagnoses. Nevertheless, 57%, 65%, 69%, and 81% of the same random sample of women in the United States, Switzerland, the United Kingdom, and Italy, respectively, believed that screening reduces or prevents the risk of getting breast cancer (Domenighetti et al., 2003). An equally astounding 75% of a representative sample of German women who participated in mammography screening wrongly believed that screening reduces the risk of developing breast cancer (*Apotheken Umschau*, 2006).

Understanding Basic Risks

Patients at Auckland Hospital, New Zealand, were asked: "What do you feel is the likelihood of you having a heart attack over the next 12 months?" This likelihood depends on individual risk factors, such as age, sex, a previous cardiac event, a family history of coronary heart disease, diabetes, smoking, and other known factors. Yet patients' risk estimates showed no correlation with any of these factors (Broadbent et al., 2006). The authors reported that there was also no optimistic bias, in which individuals tend to systematically underestimate threats to their health; perceived risks were simply unrelated to the actual risk. In a study in Switzerland, people were shown to lack even minimum medical knowledge on the risk factors for stroke, heart attack, chronic obstructive pulmonary disease, and HIV/AIDS. No participant was able to answer all questions correctly—on average, they only got one third right. The number correct was only moderately higher for people with personal illness experience (Bachmann et al., 2007).

Why do patients in these studies know so little about their risk factors? One possibility is that clinicians may be ineffective in communicating risks and do not notice how inaccurate their patients' perceptions of future risks are. Other studies indicate that patients may still have a good qualitative sense of their risk, whereas their quantitative judgments are strongly influenced by the framing of the questions asked (Woloshin, Schwartz, Black, & Welch, 1999).

Another potential reason why patients lack understanding of basic risks is that they rarely ask questions. Audiotapes of 160 adult patients' visits to doctors in North Carolina revealed that in only one out of four visits did the patient and doctor actually discuss risks or benefits (Kalet, Roberts, & Fletcher, 1994). Only few (about one in six) of these discussions were initiated by the patient, and in the majority of the discussions, the physician stated the risk with certainty (e.g., "You will have a heart attack if you don't lose weight"). Moreover, of the 42 patients who said that they actually had discussed risks with their doctors, only 3 could recall immediately after the discussion what was said. Yet almost all (90%) felt that they had their questions answered, had understood all that was said, and had enough information. Similarly, Beisecker and Beisecker (1990) reported that only few patients actively engage in information-seeking behavior in their consultations with physicians, and Sleath, Roter, Chewning, and Svarstad (1999) concluded that patients often do not ask questions about medications. In a review of 20 interventions directed at increasing patient participation, 11 assessed patient asking behavior. Congruent with the results reported above, question-asking behavior was generally low, and it was not easy to increase it: Out of the 11 interventions, only 5 resulted in significant increases in question asking (Harrington, Noble, & Newman, 2004). In contrast, patients who more actively engage during their encounters with physicians are more likely to understand treatment rationales and recommendations, are more satisfied with their health care, and even have better clinical outcomes (e.g., Roter & Hall, 1993; Street, 2001). In sum, the few studies available suggest that many patients are reluctant to ask questions, which is at odds with the goal of shared decision making.

Understanding That Screening Tests May Have Benefits and Harms

Sir Muir Gray, knighted by the British Queen for his contribution to health-care issues, is known for saying that "All screening programmes do harm; some do good as well, and, of these, some do more good than harm at reasonable cost" (Gray, Patnick, & Blanks, 2008, p. 480). What does the public know about the benefits? Consider mammography screening, where the absolute risk reduction of dying from breast cancer is in the order of 1 in 1,000 women. Let us take any estimate between 0 and 5 in 1,000 as correct. Only 6% of the women in random samples in four countries had the correct information. In contrast, 60%, 44%, 37%, and 37% of the women in the United States, Italy, the United Kingdom, and Switzerland, respectively, believed that out of 1,000 women the absolute risk reduction is 80 women or
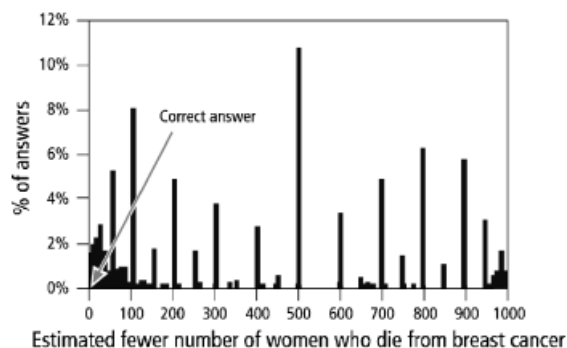
Fig. 7. What does a 25% relative risk reduction mean? A representative sample of 1,000 German citizens was asked: ''Early detection with mammography reduces the risk of dying from breast cancer by 25%. Assume that 1,000 women aged 40 and older participate regularly in screening. How many fewer would die of breast cancer?'' The best estimate is about 1 in 1,000, but most people grossly overestimated.

more (Domenighetti et al., 2003). A similar overestimation of benefits has been reported for PSA screening (Gigerenzer, Mata, & Frank, 2008). Whereas in these studies no information about relative risk reduction was given, Gigerenzer (2008) posed the following problem to a representative sample of 1,000 German citizens: "Early detection with mammography reduces the risk of dying from breast cancer by 25%. Assume that 1,000 women aged 40 and older participate regularly in screening. How many fewer would die of breast cancer?" Figure 7 shows the large variability in the understanding of this health statistic and the small proportion of citizens who understand that it means around 1 in 1,000. The most frequent estimate was 500 out of 1,000—that is, an overestimation by orders of magnitudes.

What does the public know about the harms? Schwartz et al. (2000) asked a stratified sample of 479 American women and found them to be quite knowledgeable about false positives, tending to view them as an acceptable consequence of screening. Yet very few had ever heard of other potential harms. Ninety-two percent believed that mammography could not harm a woman without breast cancer. Only 7% agreed that some breast cancers grow so slowly that these would never affect a women's health, and only 6% had ever heard of ductal carcinoma in situ, even after the researchers explained what that means: a breast abnormality that can be picked up by mammograms but that does not always become invasive. Nevertheless, almost everyone with ductal carcinoma in situ is treated by surgery. This problem—the detection of "pseudodisease"—is arguably the most important harm of screening, as it results in unnecessary surgery and radiation (Welch, 2004).

This unbalanced view of screening may have important consequences for new screening tests. A random sample of 500 Americans was asked whether they would rather receive $1,000 in cash or a free total-body CT scan. Seventy-three percent said they would prefer the CT scan (Schwartz, Woloshin, Fowler, & Welch, 2004). Yet total-body CT scans are not endorsed by any professional medical organization and are even discouraged

by several because screening tests like this can result in important harm.

Understanding Test Results

Patients in a clinic in Colorado and in a clinic in Oklahoma were asked about standard tests for diseases such as strep throat infection, HIV, and acute myocardial infarction (Hamm & Smith, 1998). Each patient judged (a) the probability that a person has the disease before being tested (base rate), (b) the probability that a person tests positive if the disease is present (sensitivity), (c) the probability that a person tests negative if the disease is absent (specificity), and (d) the probability that a person has the disease if test results are positive (positive predictive value). Most patients estimated the four probabilities to be essentially the same—independent of whether the base rate was high or low or the test accurate or not. This result held independently of whether the patients had been tested or treated for the disease or had accompanied a family member or friend who had been tested or treated for it at a doctor's office. The fact that even experienced patients did not understand health statistics suggests that their doctors either never explained the risks or failed to communicate them properly. Studies with university students show that they too have difficulties drawing conclusions from sensitivities and specificities (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995).

Understanding Treatment Outcomes

More treatment is not always better. From the 1890s until about 1975, in the footsteps of surgeon William Halsted, the standard treatment for breast cancer was mastectomy, which involves complete removal of the breast, surrounding tissues, and lymph nodes. Systematic studies, however, indicated that lumpectomy, a less invasive procedure, is as effective as mastectomy but with less harm to the patient (National Institutes of Health Consensus Conference, 1991). Despite this "good news," many physicians and women nevertheless stick with mastectomy. Even after being reminded of the equivalent beneficial effects, half of the surgeons surveyed said they would choose mastectomy over breast-conserving surgery for themselves (Collins, Kerrigan, & Anglade, 1999). This may have been an informed decision on their part (perhaps because of their desire to reduce their chance of recurrence) but also could have been based on the illusion that more invasive treatment is more effective.

A prominent example is the former First Lady Barbara Bush, who underwent a mastectomy in 1987 despite her physician's recommendation for a lumpectomy. Many American women copied her decision, which led to a significant drop in breast-conserving surgery that had been on the increase beforehand (Wong & King, 2008). Interviews with these women indicate that most believe mastectomy to provide certainty that the cancer cannot recur, and feel personally responsible to do everything possible to ensure this. Family members who share the belief that more aggressive treatment is always better tend to support or

even demand it. A 53-year-old communications director with a graduate degree, for instance, reported the reaction of her three daughters to her diagnosis: "'Mom, just have them both off. Just please, we want you around, just please have it taken care of.' By that, they meant mastectomy" (Wong & King, 2008, p. 586).

Understanding the Difference Between Relative and Absolute Risk Reduction

Is perceived treatment efficacy influenced by framing information in terms of relative and absolute risk reduction? In a telephone survey in New Zealand, respondents were given information on three different screening tests for unspecified cancers (Sarfati, Howden-Chapman, Woodward, & Salmond, 1998). In fact, the benefits were identical, except that they were expressed either as a *relative risk reduction*, as an *absolute risk reduction*, or as the *number of people needed to be treated* (screened) to prevent one death from cancer (which is 1/absolute risk reduction):

- Relative risk reduction: If you have this test every 2 years, it will reduce your chance of dying from this cancer by around one third over the next 10 years
- Absolute risk reduction: If you have this test every 2 years, it will reduce your chance of dying from this cancer from around 3 in 1,000 to around 2 in 1,000 over the next 10 years
- Number needed to treat: If around 1,000 people have this test every 2 years, 1 person will be saved from dying from this cancer every 10 years

When the benefit of the test was presented in the form of relative risk reduction, 80% of 306 people said they would likely accept the test. When the same information was presented in the form of absolute risk reduction and number needed to treat, only 53% and 43% responded identically. Medical students also fall prey to this influence (Naylor, Chen, & Strauss, 1992), as do patients (Malenka, Baron, Johansen, Wahrenberger, & Ross, 1993), and ordinary people are found to make more "rational" decisions about medication when given absolute risks (Hembroff, Holmes-Rovner, & Wills, 2004). In contrast, Sheridan, Pignone, & Lewis (2003) reported that relative risk reduction would lead to more correct answers by patients, but this is apparently a consequence of improper phrasing of the absolute risks, which was "treatment A reduces the chance that you will develop disease Y by 10 per 1,000 persons" (p. 886). This awkward statement is a hybrid between a single-event probability (it is about "you") and a frequency statement yet is not an absolute risk reduction (Gigerenzer, 2003).

A review of experimental studies showed that many patients do not understand the difference between relative and absolute risk reduction and that they evaluate a treatment alternative more favorably if benefits are expressed in terms of relative risk reduction (Covey, 2007).

In summary, the available studies indicate that very few patients have skills that correspond to minimum statistical literacy in health (cf. Reyna & Brainerd, 2007). Many seek certainty in tests or treatments; benefits of screening are wildly overestimated and harms comparatively unknown; early detection is confused with prevention; and basic health statistics such as the differences between sensitivity and specificity and between absolute and relative risks are not understood. This lack of basic health literacy prevents patients from giving informed consent.

Do Journalists Help the Public to Understand Health Statistics?

The press has a powerful influence on public perceptions of health and health care; much of what people—including many physicians—know and believe about medicine comes from the print and broadcast media. Yet journalism schools tend to teach everything except understanding numbers. Journalists generally receive no training in how to interpret or present medical research (Kees, 2002). A survey of health reporters at daily newspapers in five Midwestern states (70% response rate) found that over 80% had no training in covering health news or interpreting health statistics (Voss, 2002). Not surprisingly, few (15%) found it easy to interpret statistical data, and under a third found it easy to put health news in context. This finding is similar to that of a survey by the Freedom Forum, in which nearly half of the science writers agreed that "reporters have no idea how to interpret scientific results" (Hartz & Chappell, 1997).

The American Association for the Advancement of Science (AAAS) asked more than 1,000 reporters and public information officers what science news stories are most interesting to reporters, their supervisors, or news consumers (AAAS, 2006). The top science topic in the U.S. media is medicine and health, followed by stem cells and cloning, and psychology and neuroscience. In Europe, where national and local newspapers devote many more pages to covering science, topic number one is also medicine and health, followed by environment and climate change. Thus, a minimum statistical literacy in health would do journalists and their readers an excellent service.

Problems with the quality of press coverage, particularly in the reporting of health statistics about medical research, have been documented (Moynihan et al., 2000; Ransohoff & Harris, 1997; Rowe, Frewer, & Sjoberg, 2000; Schwartz, Woloshin, & Welch, 1999a). The most fundamental of these include failing to report any numbers, framing numbers in a nontransparent way to attract readers' attention, and failing to report important cautions about study limitations.

No Numbers

As shown in Table 6, one disturbing problem with how the media report on new medications is the failure to provide quantitative data on how well the medications work. In the United States, Norway, and Canada, benefits were quantified in only 7%, 21%, and 20% of news stories about newly approved prescription medications, respectively. In place of data, many such news

TABLE 6

Percentage of Media Reports Presenting Benefits and Harms of Medications and Other Interventions

Media	Medications/setting	Quantitative information provided	Relative risk reduction only*	Harm Mentioned
			Benefit	
Newly approved medications				
U.S. newspaper[a] (n = 15)	Ropinirole (Requip)	7	0	29
Major Norwegian newspapers[b] (n = 357)	18 newly released medications	21	89	39
Canadian newspaper[c] (n = 193)	Atorvastatin,Celexicob Donepezil, Oseltamivir, Raloxifene	20	39	32
Other medications & interventions				
U.S. newspaper/television[d] (n = 200)	Pravastatin, Alendronate Aspirin	60	83	47
Australian newspaper[e] (n = 50)	All medical interventions	40	N/A	44
Major international newspapers and U.S. national radio/TV[f] (n = 187)	Research results from 5 major scientific meetings	60	35	29

Note. *Percentage among the subset where benefit was quantified; [a]Woloshin & Schwartz, 2006a; [b]Høye, 2002; [c]Cassels et al., 2003; [d]Moynihan et al., 2000; [e]Smith, Wilson, & Henry, 2005; [f]Woloshin & Schwartz, 2006b.

stories present anecdotes, often in the form of patients describing miraculous responses to a new drug. The situation is similar when it comes to the harms of medications: Typically less than half of stories name a specific side effect and even fewer actually quantify it.

Nontransparent Numbers
Table 6 also demonstrates that when the benefits of a medication are quantified, they are commonly reported using only a relative risk reduction format without providing a base rate. Reporting relative risk reductions without clearly specifying the base rates is bad practice because it leads readers to overestimate the magnitude of the benefit. Consider one medication that lowers risk of disease from 20% to 10% and another that lowers it from 0.0002% to 0.0001%. Both yield a 50% relative risk reduction, yet they differ dramatically in clinical importance.

Sometimes there is another level of confusion: It is not clear whether a "percent lower" expression (e.g., "Drug X lowers the risk of heart attack by 10%") refers to a relative or an absolute risk reduction. To avoid this confusion, some writers express absolute risk reductions as "percentage points" (e.g., "Drug X reduced the risk of heart attack by 10 percentage points"). This approach may be too subtle for many readers. The frequency format may make this distinction clearer (e.g., "For every 100 people who take drug X, 10 fewer will have a heart attack over 10 years"). But the most important way to clarify risk reductions is to present the fundamental information about the absolute risks

in each group (e.g., "Drug X lowered the risk of heart attack by 10 in 100: from 20 in 100 to 10 in 100 over 10 years").

Harms are mentioned in only about one third of reports on newly approved medications, and they are rarely if ever quantified. While benefits are often presented in a nontransparent format, harms are often stated in a way that minimizes their salience. This is most dramatic in direct-to-consumer advertisements, which often display the relative risk reduction from the medication in prominent, large letters (without the base rate), but present harms in long lists in very fine print. TV ads typically give consumers more time to absorb information about benefits (typically qualitative claims about the drug, like "It worked for me") than about side effects, resulting in better recall of purported benefits (Kaphingst, DeJong, Rudd, & Daltroy, 2004; Kaphingst, Rudd, DeJong, & Daltroy, 2005). A second technique is to report benefits in relative risks (big numbers) and harms in absolute risks (small numbers). This asymmetry magnifies benefits and minimizes harm. A simple solution (again) is to present both benefits and harms in the same format—in absolute risks.

No Cautions
All studies have limitations. If the press is to help the public understand the inherent uncertainties in medical research, they should state the major limitations and important caveats. Unfortunately, this happens only rarely. In a content analysis of the high-profile media coverage of research presented at five scientific meetings (Woloshin & Schwartz, 2006b), few stories included

cautions about studies with inherent limitations. For example, only 10% of stories about uncontrolled studies noted that it was impossible to know if the outcome really related to the exposure.

These problems are a result not only of journalists' lack of proper training but also of press releases themselves, including those from medical schools. Press releases are the most direct way that medical journals communicate with the media, and ideally they provide journalists with an opportunity to get their facts right. Unfortunately, however, press releases suffer from many of the same problems noted above with media coverage of medical news (Woloshin & Schwartz, 2002). They often fail to quantify the main effect (35% of releases), present relative risks without base rates (45% of those reporting on differences between study groups), and make no note of study limitations (77%). Although medical journals work hard to ensure that articles represent study findings fairly and acknowledge important limitations, their hard work is hence partially undone by the time research findings reach the news media. Better press releases could change this, helping journalists write better stories.

A few newspapers have begun to promote correct and transparent reporting in place of confusion and sensationalism. And there are a number of efforts to teach journalists how to understand what the numbers mean. In Germany, for example, one of us (GG) has trained some 100 German science writers, and in the United States there are MIT's Medical Evidence Boot Camp and the Medicine in the Media program sponsored by the National Institutes of Health and the Dartmouth Institute for Health Policy and Clinical Practice's Center for Medicine and the Media (where two of us, LS and SW, teach journalists from around the world).

Do Physicians Understand Health Statistics?

It is commonly assumed that only patients have problems with health statistics, not their physicians. Most psychological, legal, and medical articles on patient–doctor communication assume that the problem lies in the patient's mind. Doctors may be said to pay insufficient attention to their patients' feelings or not listen carefully to their complaints, consult with them only 5 minutes on average, or withhold information—but rarely is it considered that many doctors might be statistically illiterate (e.g., Berwick, Fineberg, & Weinstein, 1981; Rao, 2008).

Why do doctors need minimum statistical literacy? One important skill that doctors should have is to be able to critically assess the findings of a study in the relevant literature, as is expected from every psychologist or economist. If unable to do so, doctors are more dependent on hearsay or leaflets provided by the pharmaceutical industry to update their knowledge. In entering this largely unknown territory, we begin with a test of basic numeracy.

Basic Numeracy

Schwartz and Woloshin (2000) tested physicians at Dartmouth Hitchcock Medical Center on basic numeracy. Compared to the

TABLE 7

Percentage of Physicians Answering Basic Numeracy Questions Correctly (From Schwartz & Woloshin, 2000)

Question	Physicians at Grand Rounds $n = 85$
Convert 1% to 10 in 1,000	91
Convert 1 in 1,000 to 0.1%	75
How many heads in 1,000 coin flips?	100

general public (Table 4), physicians were better in basic numeracy (Table 7). Nevertheless, only 72% of the physicians could answer all three questions correctly. Just as for laypeople, the most difficult operation for the physicians was to convert 1 in 1,000 into a percentage: One out of four physicians got it wrong. Similar results have been obtained by Estrada, Barnes, Collins, and Byrd (1999), who reported that only 60% of medical staff got all three questions correct.

The Illusion of Certainty

Physicians need to inform patients that even the best tests are not perfect and that every test result therefore needs to be interpreted with care or the test needs to be repeated. Some test results are more threatening than others and need to be handled particularly carefully. One terrifying example is a positive HIV test result. At a conference on AIDS held in 1987, former Senator Lawton Chiles of Florida reported that of 22 blood donors in Florida who had been notified that they had tested positive with the ELISA test, 7 committed suicide. A medical text that documented this tragedy years later informed the reader that "even if the results of both AIDS tests, the ELISA and WB [Western blot], are positive, the chances are only 50-50 that the individual is infected" (Stine, 1999, p. 367). This holds for people with low-risk behavior, such as blood donors. Indeed, the test (consisting of one or two ELISA tests and a Western Blot test, performed on a single blood sample) has an extremely high sensitivity of about 99.9% and specificity of about 99.99% (numbers vary, because various criteria have been used that maximize specificity at the expense of sensitivity, or vice versa). Nonetheless, due to a very low base rate in the order of 1 in 10,000 among heterosexual men with low-risk behavior, the chance of infection can be as low as 50% when a man tests positive in screening. This striking result becomes clearer after these percentages are translated into natural frequencies: Out of every 10,000 men, it is expected that one will be infected and will test positive with high probability; out of the other, noninfected men, it is expected that one will also test positive (the complement to the specificity of 99.99%). Thus, two test positive, and one of these is infected (Fig. 8). AIDS counselors need to properly inform everyone who takes the test.

To investigate the quality of counseling of heterosexual men with low-risk behavior, an undercover client visited 20 public health centers in Germany to take 20 HIV tests (Gigerenzer,
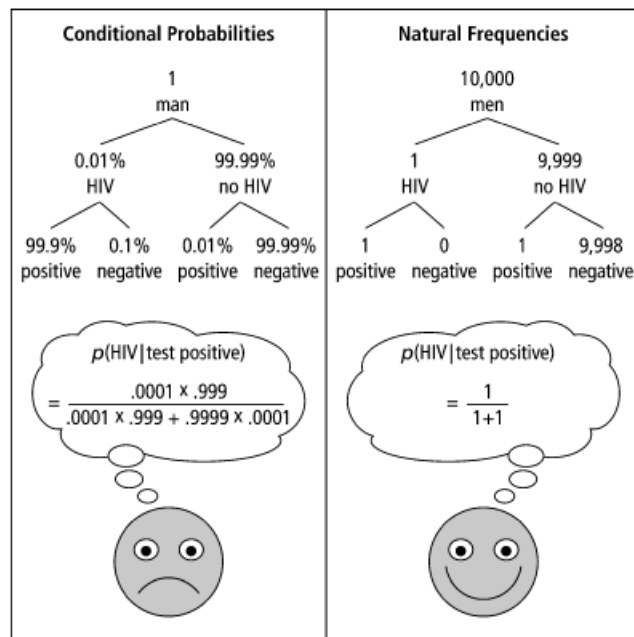
Conditional Probabilities

1
man

0.01%
HIV

99.99%
no HIV

99.9%
positive

0.1%
negative

0.01%
positive

99.99%
negative

$$p(\text{HIV} \mid \text{test positive})$$
$$= \frac{.0001 \times .999}{.0001 \times .999 + .9999 \times .0001}$$

Natural Frequencies

10,000
men

1
HIV

9,999
no HIV

1
positive

0
negative

1
positive

9,998
negative

$$p(\text{HIV} \mid \text{test positive})$$
$$= \frac{1}{1+1}$$

Fig. 8. What does a positive HIV test mean? Shown here are two ways of calculating the chances that a heterosexual man with low-risk behavior who gets a positive HIV test result (positive ELISA test and positive Western blot test) is actually infected with HIV. The information on the left is presented in terms of conditional probabilities. The information on the right is presented in terms of natural frequencies, which simplify the computations and foster insight.

TABLE 8

Answers by 20 AIDS Counselors to the Client's Question: "If One Is Not Infected With HIV, Is It Possible to Have a Positive Test Result?"

1	"No, certainly not"	11	"False positives never happen"
2	"Absolutely impossible"	12	"With absolute certainty, no"
3	"With absolute certainty, no"	13	"With absolute certainty, no"
4	"No, absolutely not"	14	"Definitely not" . . . "extremely rare"
5	"Never"	15	"Absolutely not" . . . "99.7% specificity"
6	"Absolutely impossible"	16	"Absolutely not" . . . "99.9% specificity"
7	"Absolutely impossible"	17	"More than 99% specificity"
8	"With absolute certainty, no"	18	"More than 99.9% specificity"
9	"The test is absolutely certain"	19	"99.9% specificity"
10	"No, only in France, not here"	20	"Don't worry, trust me"

Hoffrage, & Ebert, 1998). The client was explicit about the fact that he belongs to no risk group, like the majority of people who take HIV tests. In the mandatory pretest counseling session, the client asked: "Could I possibly test positive if I do not have the virus? And if so, how often does this happen? Could I test negative even if I have the virus?" Table 8 shows the answers of 20 professional counselors, mostly physicians, to the first question. The first 13 Counselors exhibited the illusion of certainty—although Counselor 10 had a more differentiated view. Counselors 14 to 16 also initially claimed that no false-positive test results ever happened, but when the client asked again whether this was absolutely true, they changed their minds (in contrast to the others, who insisted on their standpoint). Only three counselors (17–19) immediately told the client that false positives can occur since the specificity is not perfect although very high. Counselor 20 provided no concrete information but insisted on blind trust. Note that if no false positives occur, a positive test would imply an HIV infection with certainty. After we sent copies of our article reporting this state of affairs to hundreds of counseling centers, some have begun to train their counselors how to understand HIV test statistics.

PSA Counseling
In 2004, *Stiftung Warentest*, the German equivalent of the U.S. *Consumer Reports*, went beyond testing computer screens and cell phones and began to test the quality of doctors. In the first study, a 60-year-old man (a physician) paid undercover visits to 20 urologists in Berlin, drawn randomly from a total of 135 urologists, and asked for advice on PSA screening. Medical society guidelines call for thorough and systematic counseling before the first PSA test: For instance, counseling should explain that the PSA test can miss cancers or cause false alarms. It should also inform the patient that even in the event of a true positive, not every cancer needs to be treated (i.e., that over-diagnosis exists); there is instead a danger of overtreatment, whereby the treatment does not help the patient but may lead to harms such as incontinence and impotence. The patient should also know that there is no proof that early detection of prostate cancer prolongs life ("Urologen im Test," 2004). Only 2 of the 20 urologists knew the relevant information and were able to answer the patient's questions (and were graded A), and 4 others knew some of the information (grade C). The majority, 14 urologists (half of these graded D and F), could not answer most of the patient's questions, wrongly argued that it was scientifically proven that PSA screening prolongs life, and were not aware of any disadvantages. As one explained to the client, "There is nothing to ponder; at your age you must take the test" (p. 86).

Physicians Are Confused by Sensitivities and Specificities
Hoffrage and Gigerenzer (1998) tested 48 physicians with an average professional experience of 14 years, including radiologists, internists, surgeons, urologists, and gynecologists. The sample had physicians from teaching hospitals slightly over-represented and included heads of medical departments. They were given four problems; one of these was screening for colorectal cancer with the fecal occult blood test (FOBT). Half of the physicians were given the relevant information in conditional probabilities (a sensitivity of 50%, a false-positive rate of 3%,
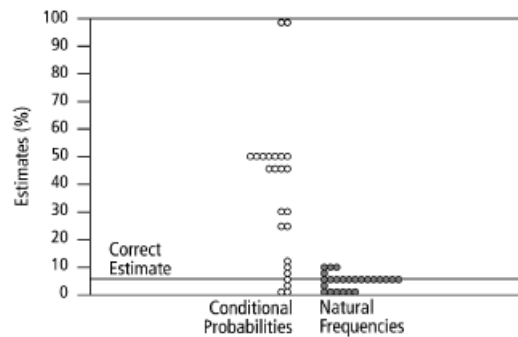
Fig. 9. How to reduce the variability in physicians' judgments. Shown are individual estimates by physicians that a person has colorectal cancer given a positive fecal occult blood test when information was given in conditional probabilities (left) versus natural frequencies (right). Variability decreased dramatically and the correct answer was given more often when numerical information was in natural frequencies (Hoffrage & Gigerenzer, 1998).

Fig. 10. The percentage of physicians' correct estimates of positive predictive values for a range of tests/diseases when information was given in conditional probabilities versus natural frequencies. Natural frequencies fostered diagnostic insight in across all four diagnostic tasks (Hoffrage & Gigerenzer, 1998).

and a prevalence of 0.3%), which is the form in which medical studies tend to report health statistics. The physicians were then asked to estimate the probability of colorectal cancer given a positive test result. Each point in Figure 9 (left) represents one physician. Note that their estimates ranged between a 1% and a 99% chance of cancer! If patients knew this striking variability, they would be rightly concerned. Note that the physicians' answers were not random. The modal answer was 50% (the sensitivity), and four physicians deducted the false-positive rate from the sensitivity (arriving at 47%). When interviewed about how they arrived at their answers, several physicians claimed to be innumerate and in their embarrassment felt compelled to hide this fact from patients by avoiding any mention of numbers.

Yet when the information was provided in natural frequencies rather than conditional probabilities, those who believed themselves to be innumerate could reason just as well as the others. The information was presented as follows: 30 out of every 10,000 people have colorectal cancer. Of these 30, 15 will have a positive FOBT result. Of the remaining people without cancer, 300 will nonetheless test positive. As Figure 9 (right) shows, most physicians estimated the positive predictive value precisely, and the rest were close. Similar results were found for the three other problems (Fig. 10). Thus, the problem is not so much in physicians' minds but in an inadequate external representation of information, which is commonly used in medicine.

Only 18% of physicians and medical staff could infer the positive predictive value from probability information in a study by Casscells, Schoenberger, and Grayboys (1978). Eddy (1982) reported that 95 out of 100 physicians overestimated the probability of cancer after a positive screening mammogram by an order of magnitude. Similarly, Bramwell, West, and Salmon (2006) found only 1 out of 21 obstetricians being able to estimate the probability of an unborn actually having Down syndrome given a positive test, with those giving incorrect responses being fairly confident in their estimates. When the same information was given in natural frequencies, 13 out of 20 obstetricians
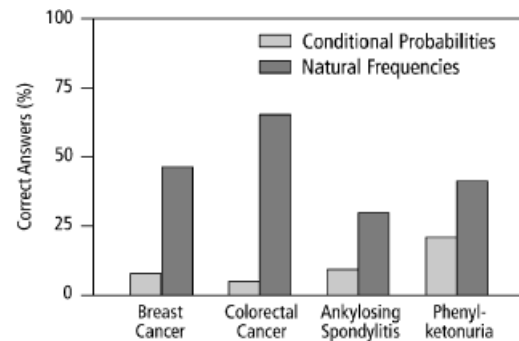
arrived at the correct answer. In one Australian study, 13 of 50 physicians claimed they could describe the positive predictive value, but when directly interviewed, only 1 could do so (Young, Glasziou, & Ward, 2002). Similar effects were reported for members of the U.S. National Academy of Neuropsychology (Labarge, McCaffrey, & Brown, 2003). Ghosh and Ghosh (2005) reviewed further studies that showed that few physicians were able to estimate the positive predictive value from the relevant health statistics.

Studies of legal professionals who evaluated criminal court files involving rape and murder showed similar results. When judges and professors of law had to estimate the probability that the defendant was the source of a DNA trace found on a victim, given the sensitivity and false-positive rate of DNA fingerprinting and base-rate information, only 13% could reason correctly. When the DNA statistics were presented in natural frequencies, 68% of the professionals were successful (Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; Koehler, 1996; Lindsey, Hertwig, & Gigerenzer, 2003).

Relative Risk Reductions Can Cause Exaggerated Perceptions of Treatment Effects

In one of the earliest studies published on this topic, Naylor et al. (1992) found that physicians rated the effectiveness of a treatment higher when the benefits were described in terms of a relative risk reduction ("A medical intervention results in a 34% relative decrease in the incidence of fatal and nonfatal myocardial infarction") rather than as an absolute risk reduction ("A medical intervention results in a 1.4% decrease in the incidence of fatal and nonfatal myocardial infarction—2.5% vs. 3.9%"; p. 920) or a number needed to treat ("77 persons must be treated for an average of just over 5 years to prevent 1 fatal or nonfatal myocardial infarction"; p. 920). Yet one cannot blame this misunderstanding on the physicians alone, since the authors of the study themselves incorrectly specified the absolute risk reduction as "a 1.4% decrease" (p. 920) instead of a decrease by

1.4 percentage points (see above). More recently, Mühlhauser, Kasper, and Meyer (2006) presented results from three diabetes prevention studies to participants in European diabetes conferences (160 nurse educators, 112 physicians, 27 other professionals). When results were presented as relative risk reduction, 87% of the health professionals evaluated the effect of the preventive intervention as important or very important. However, when the same results were presented by giving the corresponding fasting plasma glucose values, only 39% of the health professionals evaluated the effect similarly.

After interviewing one of us (GG) on the confusion caused by relative risks, an editor of a medical journal who also heads a teaching hospital in Switzerland asked all 15 gynecologists in his department what the widely known 25% risk reduction by mammography really means. How many fewer women die of breast cancer? One physician thought that 25% means 2.5 out of 1,000, another, 25 out of 1,000; the total range of the answers was between 1 and 750 in 1,000 women (Schüssler, 2005). A group of 150 gynecologists who took a course in risk communication by GG as part of their continuing education were also asked what the 25% risk figure meant. Using an interactive voting system, the physicians could choose between four alternatives:

> Mammography screening reduces mortality from breast cancer by about 25%. Assume that 1,000 women age 40 and over participate in mammography screening. How many fewer women are likely to die of breast cancer?
>
> - 1 [66%]
> - 25 [16%]
> - 100 [3%]
> - 250 [15%]

The numbers in the brackets show the percentage of gynecologists who gave the respective answer. Two thirds understood that the best answer was 1 in 1,000. Yet 16% believed that the figure meant 25 in 1,000, and 15% responded that 250 fewer women in 1,000 die of breast cancer. The overestimation of the benefit was most pronounced among physicians in their 50s and 60s, with 21% and 27%, respectively, estimating "250 out of 1,000." After the training session in risk communication, all physicians understood the correct estimate—except one, who still insisted that the answer had to be 250 out of 1,000.

Do physicians understand the number needed to treat, which is defined as the number of patients that must be treated in order to save the life of one patient? It is also called "number needed to harm," since treatments typically have side effects. Few studies have been conducted on this question (Covey, 2007). In a survey of 50 Australian physicians, only 8 could understand and explain number needed to treat to others (Young et al., 2002). Studies in the US and Europe have consistently shown that physicians and medical students prefer relative-risk reductions to number needed to treat (see Ghosh & Ghosh, 2005). British

researchers submitted four identical proposals for funding a cardiac rehabilitation and a breast cancer screening program, except that the benefit was presented either in relative risk reduction, absolute risk reduction, the absolute values from which the absolute risk reduction is computed, or number needed to treat (Fahey, Griffiths, & Peters, 1995). Only 3 out of the 140 reviewers (members of the Anglia and Oxford health authorities) noticed that the four proposals were equivalent, and when the benefits were described in relative risk reductions, the authorities saw the program as having the greatest merit and were most willing to fund it.

In her meta-analysis on the effect of presenting information in terms of absolute risks versus relative risks, Covey (2007) analyzed 13 experiments that investigated physicians and 3 experiments that investigated other health professionals, which show how physicians and health professionals can be consistently manipulated by framing the treatment effect differently. The results reviewed in this section demonstrate that even professionals are likely to evaluate effects as more beneficial when they are presented as relative risk reduction.

Geography Is Destiny

If medical practice were always founded on the best scientific evidence, then practices involving similar patients would not differ largely between hospitals and regions, with every patient receiving the most appropriate treatment known. Reality is different, however. Medical practice is often based not on scientific evidence but rather on local habits. The *Dartmouth Atlas of Health Care* documents the striking variability in the use of surgical treatments across all regions in the United States. For instance, the proportion of women in Maine who have undergone a hysterectomy ranges from less than 20% to more than 70% between regions. Similarly, 8% of the children in one community in Vermont had their tonsils removed, whereas this figure was as high as 70% in others. In Iowa, the proportion of men who have had prostate surgery varies between 15% and more than 60% (Center for the Evaluative Clinical Sciences Staff, 1996).

These numbers indicate that surgical treatments are often not based on evidence. Population differences that would necessitate disparities in treatments as large as those reported within the same state are unlikely. Instead, the tendency to follow local custom is the single most important explanation for regional differences in medical practice (Eddy, 1996). These local customs may be the result of the uncertainty about the outcome of many medical treatments. Unlike new medications, which the U.S. Food and Drug Administration (FDA) ensures are tested, surgical procedures and medical devices are not systematically subjected to evaluation (although even with FDA approval, use of medication is still extremely variable).

Collective statistical illiteracy may be one major reason why regional customs outweigh evidence. If evidence is neither understood nor communicated properly, few will be able to recognize that something might be wrong with what their local peers

are usually doing. Improved statistical skills might provide doctors and patients with the momentum to reduce this unwanted geographical variation and to practice shared decision making based on the best scientific evidence, a huge and necessary step toward evidence-based medicine (Barry, Fowler, Mulley, Henderson, & Wennberg, 1995).

Specialty Is Destiny
Similarly, if treatments are based on the scientific evidence, it should barely matter which specialist one happens to consult. However, aside from geography, the physician's specialization all too frequently determines treatment. The treatment of localized prostate cancer in the United States, for instance, generally depends on whom the patient visits. A study found that some 80% of urologists recommended radical surgery, whereas some 90% of radiation oncologists recommended radiation treatment (Center for the Evaluative Clinical Sciences Staff, 1996, p. 135). This pattern of variation suggests that doctors treat patients according to their specialty and that patients are not generally advised about their options in a way that encourages them to participate in decision making.

Collective Statistical Illiteracy
In this section, we showed that statistical illiteracy exists among patients, physicians, and journalists. The high degree of this form of innumeracy is often striking. We called this phenomenon collective illiteracy, and it is collective in two senses. First, it exists among all three groups simultaneously, and second, the groups influence each other. Doctors influence patients' understanding of health issues, and the media influence both. In this way, shared statistical illiteracy becomes a stable phenomenon whose existence is rarely noticed.

IV. CONSEQUENCES OF STATISTICAL ILLITERACY

Consumers are bombarded with messages promoting the latest new test, drug, or treatment. Many of these messages employ techniques that deliberately and insidiously exploit limited statistical literacy in order to convince the audience that they are at high risk of illness (and do not know it) and would be foolish or irresponsible not to buy the advertised service or product. We discuss two consequences of misleading advertising in this section: emotional manipulation and impediments to informed consent and shared decision making.

Susceptibility to Manipulation of Anxieties and Hopes
The advertisements in Figure 11 are an illustrative sample of those that try to raise anxieties or hopes. In the first example, one of the most prestigious cancer centers in the United States informs the reader that "as national mortality rates for prostate cancer fluctuated between 1960 and 1990, five year survival rates for prostate cancer among MD Anderson patients contin-

ued to improve." The implication is that higher 5-year survival rates would mean that more lives are saved, as Giuliani implied. Yet as we have shown, there is no relationship between the survival rate and the mortality rate. The ad compares the survival rates at MD Anderson with the mortality rates in the United States. The statistically illiterate reader, who may not notice the difference and has never heard of lead-time bias and overdiagnosis bias, is led to conclude that the center has made considerable progress in treating patients.

In each of the advertisements, the message explicitly or implicitly overstates a risk, a benefit, or both. Such ads contribute to a climate of anxiety and concern, even when the event is as rare as brain cancer. Whereas readers with adequate statistical literacy would know which questions to ask (e.g., how large is the risk, how large is the benefit, what is the state of the evidence), readers without these skills are likely to accept the messages at face value and undergo testing or treatment that is not in their best interest. Some may think that it is better to play it safe, even when an illness is rare. But these additional tests trigger a cascade of unnecessary medical intervention, overdiagnosis, and overtreatment that may result in harm, which means there is nothing "safe" about this strategy. For the severely ill, these harms generally pale in comparison to the potential benefits. But for those experiencing mild symptoms (or who have mild forms of disease), the harms become much more relevant. And for the many labeled as having predisease, or for those who are "at risk" but destined to remain healthy, or for those who have pseudodisease, treatment can only cause harm. An epidemic of diagnoses can be as dangerous to our health as disease is (Welch, Schwartz, & Woloshin, 2007).

Informed Consent and Shared Decision Making Undermined
In April 2007, the American College of Physicians—the largest medical specialty society in the United States—issued new guidelines on screening mammography for women aged 40 to 49. Rather than calling for universal screening, the guidelines recommend that women make an informed decision after learning about the benefits and harms of mammography (Schwartz & Woloshin, 2007). Yet many doctors do not understand the potential benefits and harms of mammography, including what a positive mammogram means. Collective statistical illiteracy makes informed consent science fiction.

The term *informed consent* refers to an ideal of how doctors and patients interact. Patients should be informed about the pros and cons of a treatment and its alternatives, and should decide on this basis whether they want to undergo treatment. To emphasize that the goal of informed consent is not simply obtaining patients' consent to doctors' decisions, the term *shared decision making* is often used instead (Moumjid, Gafni, Bremond, & Carrere, 2007). Yet studies indicate that clinicians rarely communicate the uncertainties about risks and benefits of
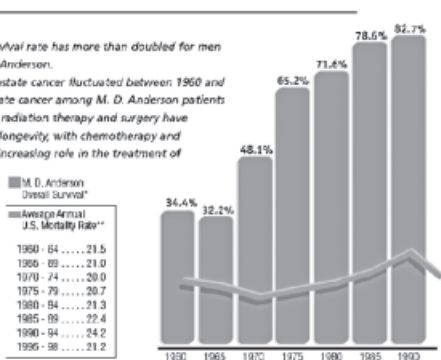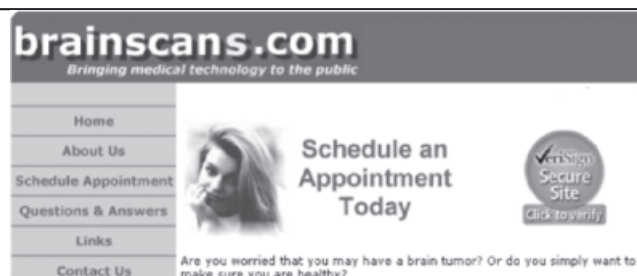
Message	Tactic	Consequence
	Implies that higher 5-year survival for prostate cancer means lives are being saved.\n\nThe problem is that there is no relationship between 5-year survival and mortality.	Confusion about progress against prostate cancer.\n\nUndue enthusiasm for the medical center.
	Implies that people need a brain scan to be sure they are healthy.\n\nImplicit message is that brain cancer is common (it is quite rare) and that screening is beneficial (there is no evidence).	Anxiety about brain cancer.\n\nUndue enthusiasm for testing.
	Implies that MRI is better than mammography because it finds more cancers. Confuses goal of screening (reducing death from breast cancer) with early detection (finding small cancers).\n\nThis is problematic because many of the "extra" cancers found represent overdiagnosis or for which treatment can only cause harm.	Anxiety about breast cancer.\n\nUndue enthusiasm for MRI screening for breast cancer.

Fig. 11. Tactics used in a selection of health messages to manipulate consumers' anxieties and hopes, and the consequences of such manipulation.
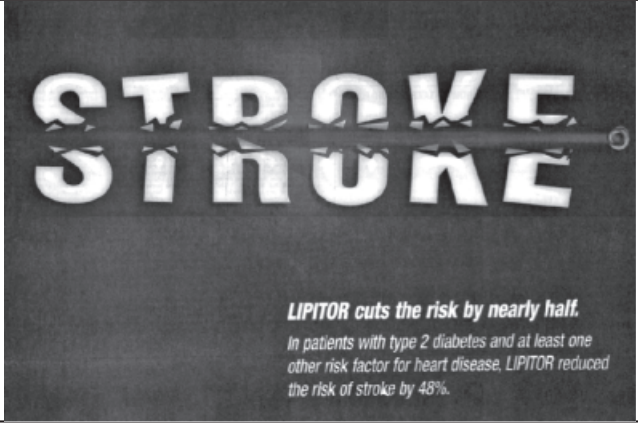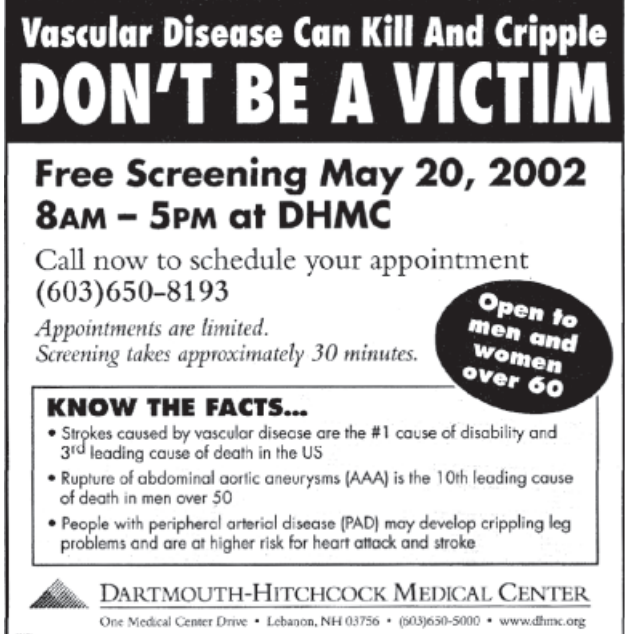
Message	Tactic	Consequence
	Implies that Lipitor substantially reduces stroke risk (by 50%). Benefit in absolute terms is small: At 4 years, 2.8% of patients taking sugar pill had a stroke compared to 1.5% taking Lipitor.	Undue enthusiasm for treatment.
	Implies that the chance of being crippled or killed by vascular disease is high and that the various screenings offered will keep you from being "a victim." The U.S. Preventive Services Task Force recommends against routine screening for carotid artery disease, peripheral artery disease, and abdominal aortic aneurysm (except for older males who have smoked).	Undue anxiety about vascular disease. Undue enthusiasm for testing.

Fig. 11. (Continued)

treatments to patients (Braddock, Edwards, Hasenberg, Laidley, & Levinson, 1999). Shared decision making can be seen as a middle ground between "doctor knows best" paternalism and rampant consumerism. Although there is no unanimous definition, key aspects are the exchange of information between the physician and the patient and the involvement of both patient and physician in making the decision (Towle & Godolphin, 1999). Informed shared decision making thus requires that patients and doctors understand the benefits and harms of different treatment options. The classical view is that the technical knowledge about risks and benefits is held by the physician and is shared with the patients to enable them to decide according to their preferences (Charles, Gafni, & Whelan, 1997).

As we have reviewed in this article, statistical illiteracy not only is typical for patients but also exists among physicians. Thus, even with good will, some doctors would not be able to inform their patients adequately without two essential skills:

understanding health statistics and communicating these in a transparent form. If both patients and physicians do not have minimal literacy in health statistics, an effective risk communication cannot take place and informed shared decision making is impossible.

This fundamental obstacle for the ideal of shared decision making has been rarely noticed, and is not a major topic at conferences on shared decision making and patient information. Their focus instead tends to be on patients as the problem, due to either their lack of knowledge or their emotional distress when forced to deal with uncertainty. Moreover, many physicians are concerned that their patients would no longer trust them if they disclosed their own uncertainty (Politi, Han, & Col, 2007). Similarly, the legal doctrine of informed consent deals with voluntary consent to biomedical research and medical treatment, the question of how much information suffices (an issue in malpractice trials), the patient's competence, and the right to

refuse treatment. In contrast, doctor's statistical literacy has not yet been recognized as an issue, but is simply taken for granted. Physicians protect themselves against patients who might turn into plaintiffs by having them give their written consent. But informed consent involves more than just signing a form.

V. CAUSES OF STATISTICAL ILLITERACY

Why does collective statistical illiteracy persist? And why is it not more of an issue at medical conferences, including those on informed consent and shared decision making? One obvious reason is the lack of training in statistical thinking in primary education and medical training, which we discuss in Section VI. In the present section we analyze factors specific to the patient–physician relationship and the health care environment.

Today, health statistics and randomized trials are an indispensable part of clinical practice. Yet medicine in fact has held a long-standing antagonism toward statistics. For centuries, treatment was based on "medical tact" in relation to the individual patient and on an ethic of personal trust rather than quantitative facts, which were dismissed as impersonal or irrelevant to the individual. The numerical method was alien to European therapeutic ethos, and equally so to 19th-century American medical practice, which presumed that disease was specific to the "natural" constitution of the individual (Warner, 1986). Some of the rare and mostly neglected early advocates for statistical thinking in medicine are described in Coleman (1987). When averages became accepted much later, in 20th-century medicine, statistics redefined health as the "normal" rather than the "natural" state, with normality characterized by averages. Even in the 1940s and 1950s, Sir Austin Bradford Hill (1897–1991), who introduced the first large-scale clinical trials, spoke of medical opposition to statistics in his lectures at medical schools (Porter, 1995).

In 1937, an editorial in *The Lancet* stressed the importance of statistics for both laboratory and clinical medicine, and criticized physicians' "educational blind spot" (Fig. 12). In 1948, the British Medical Association (BMA) Curriculum Committee recommended the inclusion of statistics in medical education. They proposed 10 lectures with additional time for exercises, ranging from teaching core concepts such as chance and probability to interpreting correlations (Altman & Bland, 1991). Yet two decades passed before the General Medical Council (GMC), in 1967, echoed the BMA recommendation (Morris, 2002). Not until 1975 did statistics become a mandatory subject in medical schools within the University of London, and it took 10 more years in Austria, Hungary, and Italy (Altman & Bland, 1991, p. 230). By comparison, in psychology and other social sciences, statistics were already institutionalized as part of university curricula in the 1950s (Gigerenzer & Murray, 1987). Doctors working on higher degrees such as an MD were thereafter encouraged to do their own research. Yet the quality of this research has been criticized by statisticians as being the product
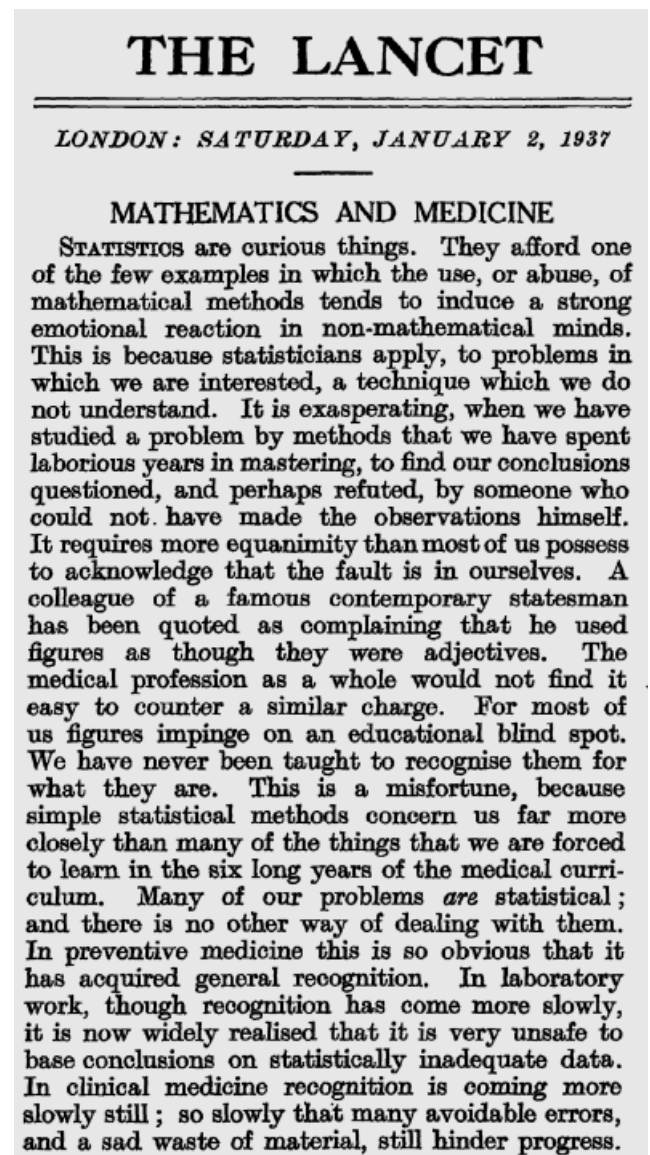

THE LANCET

LONDON: SATURDAY, JANUARY 2, 1937

MATHEMATICS AND MEDICINE

STATISTICS are curious things. They afford one of the few examples in which the use, or abuse, of mathematical methods tends to induce a strong emotional reaction in non-mathematical minds. This is because statisticians apply, to problems in which we are interested, a technique which we do not understand. It is exasperating, when we have studied a problem by methods that we have spent laborious years in mastering, to find our conclusions questioned, and perhaps refuted, by someone who could not have made the observations himself. It requires more equanimity than most of us possess to acknowledge that the fault is in ourselves. A colleague of a famous contemporary statesman has been quoted as complaining that he used figures as though they were adjectives. The medical profession as a whole would not find it easy to counter a similar charge. For most of us figures impinge on an educational blind spot. We have never been taught to recognise them for what they are. This is a misfortune, because simple statistical methods concern us far more closely than many of the things that we are forced to learn in the six long years of the medical curriculum. Many of our problems *are* statistical; and there is no other way of dealing with them. In preventive medicine this is so obvious that it has acquired general recognition. In laboratory work, though recognition has come more slowly, it is now widely realised that it is very unsafe to base conclusions on statistically inadequate data. In clinical medicine recognition is coming more slowly still; so slowly that many avoidable errors, and a sad waste of material, still hinder progress.

Fig. 12. Excerpt from a *Lancet* 1937 editorial (''Mathematics and Medicine,'' 1937) documenting the emerging realization that statistics is actually relevant for both laboratory and clinical medicine.

of inexperienced researchers in a hurry or of "Mickey Mouse trials" published solely to decorate curricula vitae (Altman & Bland, 1991, p. 224). The problem is less the physicians themselves than the organization of medicine and the academic structure of biostatistics. Young biostatisticians are rewarded for theoretical work, less so for applications to medicine. The new emerging relation between patient, physician, and biostatistician is depicted in a cartoon from 1978 (Fig. 13).

The long and enduring opposition to health statistics can be traced back to the struggle between three 19th-century visions of the physician: artist, determinist, or statistician (Gigerenzer et al., 1989, chaps. 2 & 4). We argue that these professional ideals go hand in hand with patients' corresponding ideals, which even today fuel the mixture of feelings about health

Fig. 13. The new Holy Trinity in medicine (Rimm & Bortin, 1978). This cartoon was a reaction to the statistical revolution in medicine in the 1970s. The physician continues to play God for the patient, but no longer for himself. For him, God's voice is in the verdict of the biostatistician, ''significant'' (i.e., ''$p < .05$'') or ''not significant.'' The biostatistician, finally, sees God in the mirror.

statistics: The artist embodies paternalism and requests blind trust from the patient, the determinist strives for perfect knowledge of causes and invites the illusion of certainty in patients, and the statistician relies on facts rather than medical charisma, paving the way for shared decision making. The first two ideals effectively deter interest in health statistics.

Paternalism and Trust

Physicians who think of themselves as artists place their trust in charisma, personal experience, and skill. They rely on personal intuition rather than impersonal numbers and exhibit characteristic faith in their own judgment. Risueño d'Amador (1836, pp. 634–635) argued before the Royal Academy of Medicine in Paris that the use of statistics was antimedical, for it aimed ''not to cure this or that disease, but to cure the most possible out of a certain number.'' Following the law of the majority would condemn individual patients to death. Therefore, the physician must rely on intuition, not on the mechanical collection and use of health statistics. In this view, the use of statistics was anti-scientific—it presupposed a level of homogeneity among patients that might be appropriate for physics but was utterly unrealistic in medicine.

For the physician-as-artist, the patient resembles a unique sculpture that is molded and shaped and therefore essentially passive. The artist assumes responsibility for the patient's body, and the patient engages in a paternalistic relationship of trust and obedience. Paternalism is defined as a hierarchical relationship in which a figurehead (the father, *pater* in Latin) makes decisions on behalf of others (the children) for their own good. Today, paternalism remains widespread, but it would be wrong to simply attribute it to physicians with an antiquated sense of their profession. Involved are *two* players who react to each other's expectations in a game of trust. Discussions among physicians indicate that many are ambivalent about being regarded as omniscient and omnipotent godlike figures, and would instead prefer being able to admit when they are uncertain about the best treatment (Gigerenzer, 2002). Yet they also tend to believe that patients want a father figure and might switch to another doctor who is willing to play this role. As mentioned above, medical organizations—including the American College of Physicians, the U.S. Preventive Services Task Force, and the Academy of Family Physicians—explicitly recommend that every man should weigh the pros and cons of PSA screening because the benefits (mortality reduction) are unclear, while severe harms (incontinence and impotence) occur in one third to two third of surgeries following a positive test. Yet among patients who participated in PSA screening, 68% said that it was because their doctor told them to, and 16% reported that their wife or girlfriend influenced their decision (Federman, Goyal, Kamina, Peduzzi, & Concato, 1999). The paternalist heuristic "If you see a white coat, trust it" is decidedly not a decision strategy of the uneducated public only. Consider neoclassical economists, whose doctrine includes weighing all pros and cons of alternatives, emphasizing rational choice rather than trust. Yet two thirds of more than 100 American economists surveyed said that they had not weighed any pros and cons of PSA screening but only followed their doctor's recommendation. Another 7% said that their wives or relatives had exerted an influence on the decision (Berg, Biele, & Gigerenzer, 2008).

Paternalism is practiced in many forms. *Concealed paternalism* is an extreme form in which physicians do not even inform patients about tests or treatments performed on them. It is not infrequent in the United States, where doctors routinely do PSA screening tests

on men without obtaining their consent. For instance, about one third of men without prostate cancer were unaware that their physician had ordered a PSA test (Federman et al., 1999). Concealed paternalism is in part a reaction to the unpredictabilities of the U.S. legal system that encourage physicians to practice defensive medicine—to protect themselves against potential lawsuits—rather than do what they consider best for the patient (there are no legal consequences for overdiagnosis, only for underdiagnosis). For instance, in 2003, Daniel Merenstein, a young family physician in Virginia, was sued because he did not automatically order a PSA test for a patient. Merenstein had followed the recommendations of the medical organizations and informed the man about the pros and cons, who then declined to take the test. The patient unfortunately developed a horrible, incurable form of prostate cancer. The plaintiff's attorney claimed that the PSA test was standard in the Commonwealth of Virginia and that Virginia physicians routinely do the test without informing their patients. The jury exonerated Merenstein, but his residency was found liable for $1 million. After this experience, he feels he has no choice but to overdiagnose and overtreat patients, even at the risk of causing unnecessary harm: "I order more tests now, am more nervous around patients; I am not the doctor I should be" (Gigerenzer, 2007, p. 161).

A glamorous version of paternalism is found in public health messages that replace physicians with celebrities as trustworthy authorities. Once again, the goal is to persuade people to do the "right" thing rather than encourage them to make informed decisions. For example, celebrity endorsements of cancer screening typically consist of messages asserting that the celebrity's life was saved by screening or that the life of a loved one was lost due to failure to be screened. In the United States, these celebrity messages are widely heard and have increased the number of people undergoing screening (Larson, Woloshin, Schwartz, & Welch, 2005).

Paternalism and its counterpart, trust in authority, make patients' grasp of health statistics superfluous. Moreover, patients who desire a paternalistic relationship want care, not cure by numbers—so they would be unable to detect whether or not their physician understands health statistics. Paternalism is one potential cause of collective statistical illiteracy.

Determinism and the Illusion of Certainty

The second vision of the physician is that of a determinist who relies on experimentation to find the true causes of disease and eventually will be able to treat these with certainty. This view, like that of the physician-as-artist, has been hostile to health statistics. To understand why, it is important to realize that, before the early 20th century, experiment and statistics were opposed practices. For experimenters, collecting numbers was evaluated as unscientific. Science was about causes, not chances. The determinist believed that through careful experiments, science could teach the physician to control every detail, so that averages and medical intuition alike would be rendered otiose. In Paris, the famous physiologist Claude Bernard vehemently opposed the "medical tact" promoted by Risueño d'Amador as charlatanism, but also rejected statistics as proposed by P.C.A. Louis (1787–1872). Bernard argued that being content with an average means failing to deal with the variation that is of supreme importance when curing patients. There exists, he insisted, no average pulse, but only a resting, working, or eating pulse. Nor is there average urine, for urine during fasting is different from urine during digestion. How could a physician interested in curing each patient, and not just some proportion, remain content with averages? In Bernard's (1865/1957, pp. 137–138) own words:

> A great surgeon performs operations for [a kidney] stone by a single method; later he makes a statistical summary of deaths and recoveries, and he concludes from these statistics that the mortality law for this operation is two out of five. Well, I say that this ratio means literally nothing scientifically and gives us no certainty in performing the next operation; for we do not know whether the next case will be among the recoveries or the deaths. What really should be done, instead of gathering facts empirically, is to study them more accurately, each in its special determinism.

Determinism prevailed, although some medical researchers, such as Louis in Paris and Ignaz Semmelweis (1818–1865) in Vienna, collected numbers. Louis, known as the father of modern medical statistics, showed that bloodletting in pneumonia had no effect on outcome. Semmelweis discovered that the incidence of fatal puerperal fever could be drastically cut from about 20% to 1% by requiring physicians to wash their hands between examinations. Semmelweis's discovery of a general cause, cleanliness, was largely ignored at a time when each patient and thus each cause of death were believed to be unique. Outraged by the indifference or outright hostility of the medical profession, Semmelweis eventually had a mental breakdown and was confined to an institution where he died shortly after—ironically, by what appears to have been a wound infection. Louis and Semmelweis are today considered to be forerunners of "evidence-based medicine."

It is to the credit of Sir Ronald Fisher (1890–1962) that the opposition between the experimenters and the statisticians was finally dissolved in the 1920s. Fisher joined experimentation with statistics, and after they had become two sides of the same coin, experimentation radically changed, now being defined by randomization, repetition, and other statistical concepts (Gigerenzer et al., 1989). Based on Fisher's work, Sir Austin Bradford Hill (1897–1991) promoted the new union between experimentation and statistics as an indispensable part of medicine.

Although statistics suppressed determinism, its traces have not been entirely wiped out. Specifically, determinism has survived in the form of the *illusion of certainty* in patients' minds, fostered by information brochures and advertisements. An illusion of certainty is defined as the belief that some event is absolutely certain even when such certainty does not exist. It is a major emotional obstacle toward learning to live with uncertainty.

Figure 6 showed that large proportions of the general public have illusory certainty about the perfection of tests, including HIV testing and mammography. This illusion is not simply a product of the individual mind but, as we have seen, has its historical origins in deterministic medical science. Today, it is fueled by health messages that claim or suggest certainty. For instance, the philanthropic Burda Foundation has established a network against colorectal cancer; according to its Web site: "It has been proven that with early detection, almost 100% of colorectal cancer cases could be prevented or cured" (Felix Burda Stiftung, 2008). When we inquired about where to find this evidence, the head of the foundation's marketing and communication department responded that he could not recall the precise study, but that researchers—mostly in U.S. studies—found that 60% to 90% of colorectal cancers can be prevented. Since physicians always overlook something, he explained, it follows that colorectal cancer is theoretically 100% curable. An example of a more suggestive illusion of certainty is the brain-scan advertisement in Figure 11, where the reader is asked: "Do you simply want to make sure you are healthy?"

A subtle way to induce the illusion of certainty is by analogies, such as combat metaphors that liken "war" on cancer to recent military triumph (Wong & King, 2008). In this militarized narrative, cancer is the enigmatic enemy, described as "lawless," "savage," and "relentless." This suggests that one can "slash," "burn," or "poison" the cancer cells with surgery, radiation therapy, and chemotherapy, respectively. Once the cancer is killed, the enemy is beaten, and the war is won. And the earlier the enemy is detected and the more slashing and burning that take place, the faster and more decisive the victory will be.

To summarize, determinism and its psychological counterpart, the illusion of certainty, make health statistics appear to be a wasted enterprise. The goal is certainty, rather than learning how to live with uncertainty. Like paternalism and trust, this ideal is incompatible with the quest for health statistics. Yet these factors are not the only ones. Conflicts of interest ensure that physicians and patients learn about only part of the relevant health statistics, which are framed in a way to serve particular purposes rather than to create an informed citizenship.

Conflicts of Interest

There are various players in public health with goals that can conflict with transparent risk communication—goals such as pushing a political agenda, attracting media attention, selling a new drug, increasing compliance with screening, or trying to impress physicians. Conflicts of interest lead to omission of relevant information and the use of nontransparent framing.

At issue is the distinction between content and form. All information can be communicated in several forms. The degree of transparency is empirically defined by the proportion of people in a population who can correctly understand it. Transparency is relative to expertise. For instance, when information necessary to estimate the chances that a baby has Down syndrome was presented in terms of conditional probabilities, obstetricians, midwives, and patients alike found it to be nontransparent. When the information was instead given in the form of natural frequencies, it proved to be much more transparent to the obstetricians than to the other groups (Bramwell et al., 2006). When we speak of transparent versus nontransparent forms in this article, we thus oversimplify what is a gradual matter and dependent on population. Transparent forms include absolute risks, natural frequencies, mortality rates, and, in general, statements about frequencies or depictions of frequencies in pictures. Nontransparent forms include relative risks, conditional probabilities such as sensitivities and specificities, survival rates, and statements about single events that do not specify the reference class. As the case of Giuliani illustrates, misunderstandings by nontransparent information go largely unnoticed since the issue has not yet been subject to public awareness.

Do Medical Journals Provide Transparent Information?

Where do nontransparent statistics come from? One hypothesis is that they originate from innumerate physicians, patients, and journalists, who are both manufacturers and victims of statistical confusion. Yet surprisingly, nontransparent health statistics such as relative risks without the base rate often appear in leading medical journals, and it is often from these sources that the numbers spread to physicians, the media, and the public. Nuovo, Melnikow, and Chang (2002) analyzed 359 articles that reported randomized trials in the years 1989, 1992, 1995, and 1998 that were published in *Annals of Internal Medicine*, *British Medical Journal (BMJ)*, *Journal of the American Medical Association (JAMA)*, *The Lancet*, and *The New England Journal of Medicine*. Only 25 articles reported absolute risk reduction, and 14 of these 25 also included the number needed to treat, which is simply the inverse of the absolute risk reduction. That is, only about 7% of the articles reported the results in a transparent way. The same journals, along with the *Journal of the National Cancer Institute*, were analyzed again in 2003/2004 (Schwartz, Woloshin, Dvorin, & Welch, 2006). Sixty-eight percent of 222 articles failed to report the absolute risks for the first ratio measure (such as relative risks) in the abstract; about half of these did report the underlying absolute risks elsewhere in the article but the other half did not. An analysis of *BMJ*, *JAMA*, and *The Lancet* from 2004 to 2006 found that in about half of the articles, absolute risks or other transparent frequency data were not reported (Sedrakyan & Shih, 2007). These analyses indicate that one reason why physicians, patients, and journalists talk about relative risk reductions in isolation is because the original studies regularly provide the information in this nontransparent form. Fortunately, the major medical journals, through initiatives like CONSORT (http://www.consort-statement.org/) and the international peer review congresses (http://jama.ama-assn.org/cgi/content/full/298/20/2420), are paying increasing attention to these issues.

Yet readers can be misled more directly than just via non-transparent framing. In some cases, benefits and harms of treatments are reported in different currencies: benefits in big numbers (relative risk reduction), but harms in small numbers (absolute risk increases). We call this technique *mismatched framing*. For instance, the *Guide to Clinical Preventive Services* of the U.S. Preventive Services Task Force (2002) states the relative risk reduction (not the absolute risk reduction) when describing the benefits of screening—"sigmoidoscopy screening reduced the risk of death by 59% for cancers within reach of the sigmoidoscope" (p. 93); but when the harms associated with the procedure are described, these are reported in absolute risks—"Perforations are reported to occur in approximately 1 of 1,000–10,000 rigid sigmoidoscopic examinations" (p. 94). An analysis of three major medical journals, *BMJ*, *JAMA*, and *The Lancet* from 2004 to 2006 revealed that when both benefits and harms of therapeutic interventions were reported, 1 in 3 studies used mismatched framing and did not report the benefits in the same metric as the harms. In most cases, relative risks were reported for benefits, and absolute frequencies were reported for harms (Sedrakyan & Shih, 2007).

The prevalent use of relative risks (and odds ratios) is sometimes defended on the basis that these ratio measures are transportable to different populations with different baseline risks, or that they summarize two numbers in one. But these features are also their main weakness, since they conceal the underlying absolute risks. Relative risk estimates are meaningless for understanding the chances of experiencing either a benefit or a harm. Even when readers understand relative risks, they cannot judge the clinical significance of the effect unless the underlying absolute risks are reported. As mentioned before, a relative risk reduction of 50% is compatible with both a substantial mortality reduction from 200 to 100 in 10,000 patients and a much smaller reduction from 2 to 1 in 10,000 patients. If the absolute risks are reported, the relative risks can be derived from these, but not vice versa. Randomized trials provide some of the best information in medicine, but unless the results are reported adequately, assessing and comprehending the information is difficult.

Why do medical journals not make transparency a requirement for submissions? One answer is competing interests. One third of the trials published in the *BMJ* and between two thirds and three quarters published in the major North American journals were funded by the pharmaceutical industry (Egger, Bartlett, & Juni, 2001). Richard Smith (2005), former editor of the *BMJ* and former chief executive of the *BMJ Publishing Group*, explained the dependency between journals and the pharmaceutical industry:

> The most conspicuous example of medical journals' dependence on the pharmaceutical industry is the substantial income from advertising, but this is, I suggest, the least corrupting form of dependence. . . . For a drug company, a favourable trial is worth thousands of pages of advertising . . . Publishers know that phar-

maceutical companies will often purchase thousands of dollars' worth of reprints, and the profit margin on reprints is likely to be 70%. Editors, too, know that publishing such studies is highly profitable, and editors are increasingly responsible for the budgets of their journals and for producing a profit for the owners. . . . An editor may thus face a frighteningly stark conflict of interest: publish a trial that will bring US$100,000 of profit or meet the end-of-year budget by firing an editor.

It is in the very interest of pharmaceutical companies to present the results in a way that is most likely to impress the readers and, particularly, the doctors who receive the reprints. And relative risk reductions for the benefits of one's drug are an efficient means toward this end. "Journals have devolved into information laundering operations for the pharmaceutical industry," wrote Richard Horton (2004, p. 9), editor of *The Lancet*.

Are Patients Likely to Find Transparent Information in Medical Pamphlets and Web Sites?

Pamphlets. Information on breast cancer screening should provide information about the potential benefits and harms, so that a woman can make an informed decision whether she wants to participate or not. If she participates, she also needs information about the positive predictive value. An investigation of 58 pamphlets informing women about breast cancer screening in Australia (Slaytor & Ward, 1998) found that a majority of pamphlets (35, or 60%) included information about the lifetime incidence rate, but only 1 pamphlet included the risk of actually dying of breast cancer (Table 9). Naturally, the incidence rates loom larger than the mortality rates and thus contribute to raising anxiety, and campaigns selectively reporting incidence rates have been criticized for this reason (Baines, 1992). Most important, the mortality rate, not the incidence rate, is relevant for screening, since the goal of screening is to reduce mortality, whereas it cannot reduce incidence. The information about benefits and harm that women would need to make an informed decision, in contrast, was scarce in these pamphlets (consistent with patients' lack of knowledge; see Part II). Only 22% of the Australian pamphlets reported the benefit in quantitative terms, always in relative risk reductions, and never in a transparent form, such as in absolute risk reductions. No information about potential harms was available. The most important information about the test quality, that about 9 out of 10 women who test positive do not have cancer, was never mentioned. An analysis of German brochures (Kurzenhäuser, 2003) revealed a similar picture, apart from the specific attention given to the dangers of X-rays. A few German pamphlets did, however, provide information about benefits and harms in a transparent way. In Austrian pamphlets, in contrast, there was a striking absence of relevant information (Rásky & Groth, 2004), except for constant assurances that the potential harms of X-rays are negligible and that mammography can save lives. Like in Australia, information about the positive predictive value was never provided. All

TABLE 9

Percentage of Informational Materials That Provide Specific Pieces of Information About Breast Cancer Screening to Patients in Various Countries

	Pamphlets (Australia)[a] $n = 58$	Pamphlets (Germany)[b] $n = 27$	Pamphlets (Austria)[c] $n = 7$	Web sites (8 countries)[d] $n = 27$	Invitations (7 countries)[e] $n = 31$
Baseline risk					
Lifetime risk of developing breast cancer	60	37	43	44	32
Lifetime risk of dying from breast cancer	2	4	0	15	n/a
Benefits from screening					
Relative risk reduction of death from breast cancer	22	7	0	56	23
Absolute risk reduction of death from breast cancer	0	7	0	19	0
Number needed to screen to avoid one death from breast cancer	0	4	0	7	0
Harms					
Overdiagnosis and overtreatment (e.g., carcinoma in situ)	n/a	11	n/a	26	0
Harms from X-rays	n/a	44	100	15	n/a
Psychological distress related to false positive results	n/a	11	n/a	37	n/a
Test properties					
Proportion of women who are recalled (positive tests)	14	11	14	44	19
Proportion of breast cancers detected by mammography (sensitivity)	26	19	0	26	23
Proportion of women who test negative among those without breast cancer (specificity)	0	4	0	0	0
Proportion of women with breast cancer among those who test positive (positive predictive value)	0	15	0	15	0

Note. The table lists all mentions of the respective piece of information, independent of whether the piece of information was given correctly. It is based on different studies, and not all studies assessed all pieces of information (n/a).
[a]Slaytor & Ward (1998); [b]Kurzenhäuser (2003); [c]Rásky & Groth (2004); [d]Jorgensen & Gøtzsche (2004); [e]Jorgensen & Gøtzsche (2006).

7 of the Austrian pamphlets mentioned that early detection increases the chance for complete recovery, but all were mute on the size of this increase. It is telling that when a recent German pamphlet (from the women's health network Nationales Netzwerk Frauen und Gesundheit; not included in Table 9) informed women about screening in a more comprehensive and transparent way, the Austrian Association of Physicians asked their members to remove it from their shelves because they feared it would lead to lower compliance (Noormofidi, 2006). This is the same association that, when *The Lancet* published a meta-analysis finding homeopathy to have no effect (Shang et al., 2005), responded that meta-analyses are an interesting instrument for theoretical science but of little relevance to clinical practice (Österreichische Ärztekammer, 2005).

Mismatched framing also occurs in pamphlets and leaflets. Yet as Table 9 shows, it can only occur in the few that actually provide information about both benefits and harms. For instance, one leaflet explained that hormone replacement therapy "has been proven to protect women against colorectal cancer (by up to more than 50 percent)" whereas the risk of breast cancer "may possibly increase by 0.6 percent (6 in 1,000)" (see Gigerenzer, 2002, p. 206). Looking up the absolute risk reduction, which was not reported, one finds that the 50% benefit corresponds to an absolute number that is less than 6 in 1,000. In a study, this leaflet was given to 80 women between age 41 and 69; 75% of these incorrectly understood the numbers to mean that hormone replacement therapy prevents more cases of cancer than it produces, whereas only 4% correctly understood that the opposite was the case (Hoffrage, 2003).

Invitations for screening. In countries with publicly funded screening, eligible citizens are often made aware of these programs by letters of invitation. Thus, by sheer numbers of citizens reached, such letters are—alongside physicians—potentially the most important source of information about screening. Invitation letters would be the ideal opportunity to provide the patients with balanced, transparent information about screening, so that they can make informed decisions. Yet there is a conflict of interest built into the system: Those who are responsible for the screening program are also responsible for

designing the invitations, which puts their goal of increasing compliance at odds with increasing transparency. For example, German health authorities, addressing women between 50 and 69, say that it is important that as many women as possible participate and this is best reached by personal invitations (Bundesministerium für Gesundheit, 2002b). The official leaflet sent to all women in Germany in this age group contains much useful information, including that 5% will be recalled (i.e., test positive) and that 80% of these do not have cancer, but includes no information about the size of the potential benefit (Kassenärztliche Bundesvereinigung, 2004). If women were told that it is indeed unclear whether the benefits of mammography screening outweigh its harms, some might decide against it; thus, transparent health statistics are likely to decrease compliance in this case.

Jorgensen & Gøtzsche (2006) investigated letters of invitations to breast cancer screening in seven countries with publicly funded screening: Australia, Canada, Denmark, New Zealand, Norway, Sweden, and the United Kingdom (Table 9). Most of the invitations (97%) stated the major benefit of screening, the reduction in breast cancer mortality. However, the very few (23%) that also mentioned the size of the benefit always did so in terms of relative risk reductions rather than absolute risk reductions. None of the invitations included information about potential harms or the positive predictive value. Instead, most invitations used persuasive wording and prespecified appointments. Thus, the invitation letters clearly aim at compliance rather than at informing the public.

If citizens look for additional information on the Internet, does this provide a more balanced perspective?

Web sites. A study of 27 Scandinavian and English speaking Web sites demonstrated that all those of advocacy groups and governmental institutions (24 Web sites in total) recommended screening and favored information that shed positive light on it (Jorgensen & Gøtzsche, 2004). Only few mentioned the major potential harms of screening: overdiagnosis and overtreatment. Three Web sites of consumer organizations had a more balanced perspective on breast cancer screening and included information on both the potential benefits and harms. In total, very few sites met the standards of informed consent, as specified by the General Medical Council's (1998) guidelines for patient information.

Mismatched framing was also used in the National Cancer Institute's Risk Disk, intended to help women make informed decisions about whether to use tamoxifen for the primary prevention of breast cancer (Schwartz, Woloshin, & Welch, 1999b). The benefit of tamoxifen is stated with the following relative risk reduction: "Women [taking tamoxifen] had about 49% fewer diagnoses of invasive breast cancer." In contrast, the harm of more uterine cancer was presented as "the annual rate of uterine cancer in the tamoxifen arm was 30 per 10 000 compared to 8 per 10 000 in the placebo arm" (National Cancer Institute, 1998). And in fact, the current Breast Cancer Prevention Study Fact Sheet (National Cancer Institute, 2005) presents the 49% statistic and no numbers for the increased risk of uterine cancer.

This problem is not limited to information about cancer. For example, advice on the World Wide Web about how to manage fever in children at home was similar: Complete and accurate information was rare, and some Web sites contained advice that should in fact be discouraged (Impiccatore, Pandolfini, Casella, & Bonati, 1997). Rigby, Försstrom, Roberts, Wyatt, for the TEAC-Health Partners (2001) estimated that one quarter of the messages disseminated by Internet health information services are false. These results are alarming, given that many people use the Internet to acquire information about health issues—in the European Union, this number is as high as 23% (see Jorgensen & Gøtzsche, 2004).

How Accurate Are Leaflets Distributed to Doctors? For the busy physician with limited time to keep abreast of medical research, advertisement leaflets by the pharmaceutical industry are a major source of further education. These are directly sent to doctors or personally handed to them by well-dressed representatives. A leaflet typically summarizes the results of a published study for the physician in a convenient form. Do doctors get accurate summaries? Researchers from the German Institute for Quality and Efficiency in Health Care searched for the original studies and compared these with the summaries in 175 leaflets (Kaiser et al., 2004). The summaries could be verified in only 8% of the cases (!). In the remaining 92% of cases, key results of the original study were often systematically distorted or important details omitted. For instance, one pamphlet from Bayer stated that their potency drug Levitra (Vardenafil) works up to 5 hours—without mentioning that this statistic was based on studies with numbed hares. Should doctors have wanted to check the original studies, the cited sources were often either not provided or impossible to find. In general, leaflets exaggerated baseline risks and risk reduction, enlarged the period through which medication could safely be taken, or did not reveal severe side effects of medication pointed out in the original publications.

The spread of advertising for medical products reflects the increase in the commercialization of medicine—and profits from the statistically illiterate, who are unlikely to ask the tough questions. Even for advertisements placed in medical journals, selective reporting of results has been documented (Villanueva, Peiró, Librero, & Pereiró, 2003). In the United States, direct-to-consumer advertising constitutes the single largest effort to inform the public about prescription drugs—on which pharmaceutical companies spent more than $5 billion in 2007. These ads typically assert the benefit of the drug with personal statements (e.g., "It works for me") or with data on popularity of the drug ("Over a million people have begun to take this drug to manage their diabetes"). But the ads fail to provide the most fundamental information consumers need to make informed decisions: How well does the drug work, and what are the side effects? (Woloshin, Schwartz, Tremmel, & Welch, 2001). The

education of patients and physicians alike is too important to be left to the pharmaceutical industry and pseudoeducational campaigns that promote sales.

Do Political Institutions Promote Informed Citizens? In 2001, the German government proposed mammography screening for all women between ages 50 and 69: "Mammography screening could reduce mortality from breast cancer by 30%, that means, every year about 3500 deaths could be prevented, ca. 10/day" (cited in Mühlhauser & Höldke, 2002, p. 299). Note the use of a relative risk reduction, suggesting a big benefit, instead of the absolute risk reduction, which is in the order of 1 in 1,000. Furthermore, the public is not informed that there is no evidence that the total mortality is reduced by screening—that is, that no lives are saved. The estimated 3,500 women are the decreased number of women who die of breast cancer within 10 to 15 years, whereas the total number of deaths remains the same in this period for women who participate in screening or not (Gøtzsche & Nielsen, 2006). The Berlin Chamber of Physicians (Ärztekammer Berlin, 2002, March 21) protested in a 2002 press release against a general screening program on the grounds that there is no scientific evidence that the potential benefits of screening are higher than its harms, and that the parliament's health committee overstated benefits and downplayed harms. Two days later, the German Minister of Health, Ulla Schmidt, responded in a press release that there is sufficient evidence in favor of screening because "there is an up to 35% reduction in breast cancer mortality" (Bundesministerium, 2002a). Note once again the use of relative risk reduction. When one of the authors (GG) clarified what this number means in an interview in the German weekly *Die Zeit*, the advisor of the Secretary of Health, Professor Karl Lauterbach defended the use of relative risk reduction by responding that "In justifying the programs, the Secretary of Health does not inform individual women, but the public. If an individual doctor advises patients, he should, as Mr. Gigerenzer, state the absolute risk and its reduction" (Lauterbach, 2002, p. 16). According to this logic, transparency is for individual women, not for the public. It is a pity that a democratic government confuses taxpayers about the benefits of a program that they ultimately finance. But political interests reign over transparency in health in other countries, too.

In 1997, the National Institutes of Health Consensus Development Conference on Breast Cancer Screening for Women Ages 40 to 49 was convened at the request of the director of the National Cancer Institute (NCI). The expert panel reviewed the medical studies and concluded with a 10-to-2 vote that there is insufficient evidence to recommend screening for this age group and that "a woman should have access to the best possible relevant information regarding both benefits and risks, presented in an understandable and usable form" (National Institutes of Health Consensus Development Panel, 1997, p. 1015). At the news conference, Richard Klausner, Director of the NCI, said he was "shocked" by this evidence, and that night a national

television program began its news coverage with an apology to American women for the panel's report. Eventually, the Senate voted 98 to 0 for a nonbinding solution in favor of mammography for women in their 40s. The director of the NCI asked the advisory board to review the panel's report, a request that they first declined, but in March 1997, the board voted 17 to 1 that the NCI should recommend mammography screening every one or two years for women in this age group—against the conclusion of its own expert panel (Fletcher, 1997). The voting members of the NCI advisory board are appointed by the U.S. president, not by the medical experts in the field, and are under great pressure to recommend cancer screening.

In 2002, new studies became available that again indicated that the benefits of mammograms may not outweigh the risks, and Donald Berry, chairman of the department of biostatistics at M.D. Anderson Cancer Center explained this result to the Senate, but to no avail. The Bush administration restated the recommendation and Andrew von Eschenbach, the director of the NCI at that time, announced that women in their 40s should get mammograms (Stolberg, 2002).

The mesh between medicine and politics is visually captured in two stamps (Figure 14). The U.S. Postal Service has used commemorative stamps depicting matters of historical, social, and cultural importance to the nation. The mechanisms for choosing stamps were designed to insulate the Postal Service from special interest groups. But in 1996, a California surgeon and founder of a nonprofit advocacy organization for breast cancer research approached Representative Vic Fazio (D-Calif.) with the idea of issuing a fund-raising stamp (Woloshin & Schwartz, 1999). In August 1997, the Breast Cancer Research Stamp Act was signed into U.S. law, against the objections of the Postal Service. The denomination was 40 cents, of which 8 cents went to federal research on breast cancer. The nation's first-ever fund-raising stamp was issued in 1998 at a White House ceremony hosted by First Lady Hillary Rodham Clinton and Postmaster General William Henderson. The idea for a prostate cancer stamp emerged in Congress in reaction to the breast cancer stamp. The Postal Service once more opposed the bill calling for a new semipostal stamp, and eventually a regular stamp that promoted "annual checkups and tests" was released.

Fig. 14. U.S. Postal Service stamps promoting breast and prostate cancer screening—an illustration of the intersection between medicine and politics.

Evidence did not seem to matter. Just 2 years before the stamp's release, in 1996, the U.S. Preventive Service Task Force had concluded that "routine screening for prostate cancer with digital rectal examinations, serum tumor markers (e.g., prostate-specific antigen), or transrectal ultrasound is not recommended" (p. 119). Against the scientific evidence, the Postal Service became a vehicle for special interest groups.

Summary

In this section we argued that there is a network of causes for collective statistical illiteracy. Statistical thinking is a latecomer in medical practice and research, which had been dominated by two conflicting models of physicians: the godlike artist and the scientific determinist, both of whom rejected statistics. These ideals go hand in hand with unconditional trust and illusions of certainty in patients, for whom statistical information appears of little relevance. Now that these two visions of the patient–physician relationship are beginning to crumble in the age of information, organizations with other interests spend much subtle energy in preventing citizens from receiving the relevant information about potential benefits and harms of medical treatments in a transparent form. The sad part of this story is that, to a considerable degree, democratic governments and medical organizations that disseminate information pamphlets play their part in this game.

VI. THERAPY

The network of factors we have described—competing interests, trust, paternalism, and illusion of certainty—provides a challenge for change. Yet if we can change one fundamental factor, some of the other obstacles might fall like a row of dominos. In our opinion, this would be education of the public in statistical thinking combined with training in transparent framing. An educated citizenship will know what questions to ask, what information is missing, and how to translate nontransparent statistics into transparent ones. But that necessitates rethinking how statistical thinking is taught.

Medical doctors tend to think of psychologists as therapists, useful for the emotionally disturbed but not for members of their own trade. Research and training in transparent risk communication, however, is a field in which cognitive psychologists can actually help doctors. In this last section, we define the task that psychological and medical researchers should address: the efficient training of pupils, medical students, and doctors in understanding risks and uncertainties. We also discuss sources of resistance.

Teach Statistical Literacy in School

Statistical thinking is the most useful part of mathematics for life after school. Today, however, almost all of the available time is spent on the mathematics of certainty—from algebra to geometry to trigonometry. If children learned to deal with an uncertain world in a playful way, much of collective statistical illiteracy would be history. But for the teacher, like for the doctor, sta-

tistical thinking is a late arrival: Elementary and high schools have been "probability free" even longer than medical schools have been. In 1992, when Michael Shaughnessy reviewed the situation in the United States, he reported that only 2% of college-bound high-school students had taken a course in probability and statistics, whereas 90% of these students had taken a course in algebra (Shaughnessy, 1992). The Quantitative Literacy Project (Gnanadesikan, Scheaffer, & Swift, 1987) and the Middle Grades Mathematics Project (Phillips, Lappan, Winter, & Fitzgerald, 1986) were among the pioneering programs to make some inroads into the teaching of probability and statistics in the middle grades.

National school systems differ profoundly in the time allotted to different areas within mathematics. Germany's educational system, for instance, traditionally paid very little attention to teaching data analysis and probability. In recent years this has changed, and competencies in data analysis and probability are now a mandatory part of national curricula from elementary school to grade 12. Yet that alone does not solve the problem. Many teachers are simply not prepared to teach statistics. Performance of German students in statistics and probability as measured by the 2003 Programme for International Student Assessment (PISA) continued to be relatively weak. PISA documented a relatively stronger performance for American 15-year-olds in the area of "uncertainty" as compared to "quantity" and "shape and space." However, this result has to be seen against the low overall performance of the U.S. students, putting their competence in dealing with "uncertainty" at a similar unsatisfactory level as that of the German students. The U.S. National Council of Teachers of Mathematics (NCTM) has announced its commitment to teaching data analysis and probability in grades pre-kindergarten to 12, as described in its *Principles and Standards for School Mathematics* (NCTM, 2000), and declared data analysis and probability its "Professional Development Focus of the Year," providing additional resources and continuing education. The NCTM prefaced its *Principles* with a simple truth: "Young children will not develop statistical reasoning if it is not included in the curriculum."

Today, the mathematics curriculum in many countries includes probability and statistics. Yet research on the effect of teaching has shown that while students can learn how to compute formal measures of averages and variability, they rarely understand what these statistics represent or their importance and connection to other concepts (Garfield & Ben-Zvi, 2007). Few pupils learn to see a connection between statistics in school and what is going on in their world. Why do schools contribute so little to statistical literacy? We believe that there are four factors. Statistical thinking is taught

(a) too late in school
(b) with representations that confuse young minds
(c) with boring examples that kill motivation, and
(d) by teachers who are unversed in statistical thinking

Statistical Literacy Should Be Taught as Early as Reading and Writing

An essential requirement for starting early is a discrete (not continuous) concept of probability. Children can easily understand natural numbers, whereas proportions and continuous quantities are more difficult (Butterworth, 1999; Gelman & Gallistel, 1978). Yet many mathematics educators insist that probability needs to be introduced as a continuous variable, along with continuous distributions. This theoretical vision is a major obstacle to a successful head start with statistical thinking. For instance, at a conference on teaching statistics in school, where we showed that children can easily understand statistics with discrete representations (such as the absolute number of cases, as in Figs. 3 & 8), a mathematics professor asked why the frequentistic, discrete concept of probability was being emphasized, as opposed to the subjective, continuous concept (according to which a continuous probability distribution describes a person's degree of belief in a proposition, such as that the next president of the United States will be Republican; see Savage, 1972). He seems to have been thinking about philosophical schools of probability, not about children.

In recent years, a consensus has emerged from the recommendations of professional associations (e.g., the NCTM and the German Gesellschaft für Didaktik der Mathematik) that instruction in statistics and probability should begin in primary school. This understanding is new and revolutionary, given that generations of students in the 20th century have learned statistics and probability only in their later secondary and tertiary education.

Start With Transparent Representations

Teaching statistics early is not sufficient. It is also essential to represent probabilistic information in forms that the human mind can grasp. To this end, visual and hands-on material can enable a playful development of statistical thinking. For instance, tinker-cubes are lego-like units that first graders can use to represent simple events, to combine to represent joint events, and to count to determine conditional frequencies (Kurz-Milcke & Martignon, 2007; Kurz-Milcke, Gigerenzer & Martignon, 2008). At a later age, visualization software such as Fathom (Finzer & Erickson, 2006; www.keypress.com/x5656.xml) and TinkerPlots (Konold & Miller, 2005; www.keypress.com/x5715.xml; Biehler, Hofmann, Maxara, & Prömmel, 2006) are available for exploring and manipulating data sets (Garfield & Ben-Zvi, 2007). By starting with concrete representations of risks, children can build up confidence in understanding the basic concepts, and will less likely develop a math phobia when continuous concepts are introduced at a later point.

Consider a particularly challenging task: Bayesian inference, which is needed in medicine to derive the positive predictive value from a prior probability (e.g., the base rate of a disease) and from the sensitivity and the false-positive rate of a test (see Fig. 3). For decades, psychologists had concluded that even adults are doomed to fail—"man is apparently not a conserva-tive Bayesian: he is not a Bayesian at all" (Kahneman & Tversky, 1972, p. 450), and "our minds are not built (for whatever reason) to work by the rules of probability" (Gould, 1992, p. 469). Yet when the information is presented in natural frequencies rather than conditional probabilities, even fourth to sixth graders can reliably solve these tasks (Zhu & Gigerenzer, 2006). Computer-programmed tutorials showed that people can learn how to translate conditional probabilities into natural frequencies in less than 2 hours (Sedlmeier & Gigerenzer, 2001). Most important, learning was not only fast but also remained stable after weeks of subsequent tests, whereas students who were taught how to insert probabilities into Bayes's rule (see Fig. 3, left side) forgot fairly quickly what they had learned (see also Ruscio, 2003). Statistical literacy is more than learning the laws of statistics; it is about representations that the human mind can understand and remember.

Teach Real-World Problem Solving, Not Applying Formulas to Toy Problems

People love baseball statistics, are interested in graphs about stock indices, have heard of probabilities of rain, worry about the chance of a major earthquake, and are concerned about cholesterol and blood pressure. How safe is the contraceptive pill? What is the error margin for polls and surveys? Is there a probability that extraterrestrial life exists? Personal relevance is what makes statistics so interesting.

To build up motivation, curricula should start with relevant everyday problems and teach statistics as a problem-solving method. However, in most curricula, statistics is taught as a formal mathematical discipline, where problems are purely decorative. One begins with a law of probability and then presents problems that can be safely answered by this law—which is why the use of randomizing devices such as coins, dice, and urns abound. Even when a textbook gives itself an applied feel, the content is more often than not only secondary. This approach leads to a continuous stream of more or less boring examples that do their best to kill young people's curiosity and motivation.

Is lack of motivation the reason students learn so little about statistics? The sparse evidence available suggests that the answer is no (Martignon & Wassner, 2005). Forty mathematics teachers who taught at German Gymnasien (grades 5–13) were asked to rate their students' interest, attentiveness, motivation, and comprehension when being taught probability and statistics compared to the rest of mathematics education. Many teachers reported that their students were more interested, attentive, and motivated when being taught probability and statistics than they were when being taught other types of mathematics (Fig. 15). Yet, strikingly, this did not lead to better comprehension. We believe that this dissociation can largely be overcome by beginning with real-world problems and transparent representations, and recently textbooks have incorporated these principles from psychological research (Gigerenzer, 2002). For instance, one
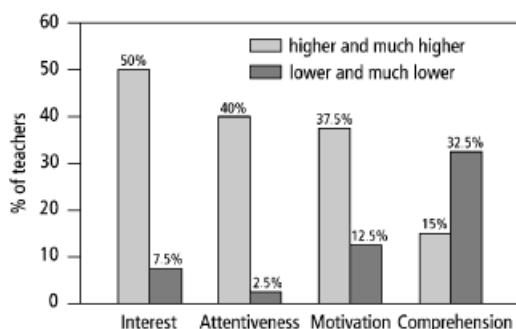
Fig. 15. Mathematics teachers' judgments about students' attitudes to and comprehension of probability and statistics compared to the rest of mathematics. For instance, 50% of the teachers said that students' interest was higher for probability and statistics, 7.5% estimated it as lower, and the others as equal. Note the discrepancy between interest, attentiveness, and motivation on the one hand, and comprehension on the other (Martignon & Wassner, 2005).

secondary school textbook (Jahnke & Wuttke, 2005) introduces Bayes's rule with the real story of a 26-year-old single mother who tested positive in a routine HIV test at a Virginia hospital, lost her job, moved into a halfway house with other HIV-positive residents, had unprotected sex with one of them, eventually developed bronchitis, and was asked by her new doctor to take the HIV test again. She did, and the result was negative, as was her original blood sample when it was retested. The poor woman had lived through a nightmare because her physicians did not understand that there are false alarms even when both the ELISA and the Western blot test are positive. After hearing this example, the students are given the relevant information in natural frequencies and can compute that the positive predictive value of the two tests combined is only about 50%, not 100% as the original physicians had assumed. Here, students are taken from a real and gripping problem to statistical thinking. As a next step, they can learn where to find the relevant information themselves and how to ask questions about the assumptions for applying statistical principles to the real world—questions which do not arise when only toy problems (e.g., involving cards or dice) are used.

Statistical literacy demands rethinking the teaching of statistics. Is mathematical statistics an end in itself or a useful tool for solving problems? In our opinion, to be helpful for patients, physicians, and journalists alike, it should be taught as a disciplined problem-solving technique. One great disappointment of motivated students is when they find out that school statistics has little to do with their own world.

Teach Teachers First
Studies on pre-service and in-service K–12 teachers suggest that both groups have troubles in understanding and teaching statistics (for an overview, see Garfield & Ben-Zvi, 2007). For instance, elementary-school teachers have difficulties in finding out the median of data sets presented graphically (Bright & Friel, 1998). As long as teachers themselves do not understand,

they are likely to resist. Similar to what occurs in medical training, resistance to statistics education is rarely articulated openly in print but is indirectly present through the lack of support for its actual attainment. Unexpressed concerns can be detrimental to an undertaking. For this reason, we are making an effort to explicate four major concerns of teachers, beginning in the early grades and continuing on through the middle ones:

- Concern # 1: *There are simply more important things in the elementary math curriculum; in other words, something else would suffer from it*
- Concern # 2: *Statistics is about games of chance and touches upon content that is simply not appropriate for children in the elementary grades*
- Concern # 3: *We experience difficulties teaching probability and statistics to high-school and even college students, let alone to students in elementary school*
- Concern # 4: *In spite of my education as a math teacher, I know very little about data analysis, probability, and teaching in this area of mathematics*

The first concern is one of mathematics educators who do not seem to realize that statistical thinking is indispensable. It also reflects the hierarchy within the mathematics profession, with abstract mathematics at the top and empirical statistics at the bottom. In our view, the traditional emphasis on Latin as a foreign language in schools provides an apt comparison. After 4 years of Latin, although students showed improved skill in grammar-related activities, such as letter-exact reading and forming complex sentences, they did not learn a modern Romance language (i.e., Spanish) more easily than did a group lacking proficiency in Latin (Haag & Stern, 2003). Learning the mathematics of certainty cannot be assumed to simply transfer to readily learning statistics and probability, nor can it be assumed to be more important.

The second concern is very peculiar to statistics education. Historically, games of chance were an early topic of probability theory, but not the only one, the others being the analysis of mortality tables for insurance and the evaluation of the reliability of testimony in court (Daston, 1988). Yet the connection with games of chance can evoke moral protest. In the 1980s, Israeli psychologist Ruma Falk devised a hands-on probability game in which young children could develop their intuitions. Children had to choose one of two disks (like two roulette wheels) to spin before making a move on a game board. Each disk was divided into sectors of two colors, one color favorable and one unfavorable. The challenge was to identify and spin the disk with the higher probability of a favorable outcome.

The game was sharply criticized by parents and educators as being "uneducational." They objected to the notion of a game in which one might make a correct choice (of the disc with a higher probability of success) and yet obtain an unfavorable outcome,

while on the other hand, an incorrect decision may be rewarded. Obviously, they wished for a consistent, "just" system. Implied in their criticism was the expectation that good decisions would always be reinforced, while bad ones would never be. (Falk & Konold, 1992)

This concern involves a double misunderstanding. Statistics is not only about games of chance but about health and other everyday issues as well. And real life is not always fair in every instance, even if it hopefully is in the long run.

The third and fourth concerns need to be addressed in teacher training. A radical solution would be to take teaching of statistical thinking out of the hands of mathematics teachers and turn it into a problem-solving field. Such a new field could be called "statistical reasoning" and might help young people make better decisions about health, drugs, alcohol use, driving, biotechnology, and other relevant issues. This teaching revolution is related to Moore's (1997) "new pedagogy" designed to overcome the "professional fallacy" that introductory courses are a step in the training of formal statisticians.

How Can Primary and Secondary School Contribute to Statistical Literacy?

We recommend that primary and secondary schools begin teaching statistical thinking as a problem-solving discipline in its own right, not as an appendage to math education. In this way, a majority of citizens could reach minimal or even higher levels of statistical literacy. With this basic knowledge, patients, physicians, and journalists would no longer be as easily confused by numbers, which could directly impact on some of the other causes mentioned in Part V. Statistical thinking as a problem-solving discipline puts the solution of individual and social problems first, using statistical tools as a means toward that end. The goals of this discipline include the following:

- To learn that societal problems can be solved by critical thinking instead of mere belief, trust in authority, or violence
- To develop empirical thinking by formulating competing hypotheses and collecting and analyzing data to test them
- To develop critical thinking skills in evaluating the applicability of various statistical models to real-world problems
- To learn to use transparent representations and computer-based visualization techniques

Teaching statistical thinking as problem solving can be directly connected to teaching health in schools. Steckelberg, Hülfenhaus, Kasper, Rost, and Mühlhauser (2007, 2008) developed a curriculum and a test of critical health literacy for grade 11 secondary-school students, both as a 1-week project and over a longer period. The curriculum contains six modules, ranging from recognizing fallacies and misinterpretations of data representations to designing experiments to understanding systematic reviews to appraising patient information. The curriculum was well accepted by students, who perceived it as personally beneficial, and increased their competence in health literacy.

Teach Statistical Literacy in Medical Training

As described in the previous section, not until in the late 20th century did medical schools begin to teach statistics, and there are still medical organizations, physicians, and students who tend to see statistics as inherently mathematical and clinically irrelevant for the individual patient (Altman & Bland 1991; Gigerenzer, 2002). This attitude is reinforced by curricula focusing on analysis of variance and multiple regression techniques; transparent risk communication is rarely recognized as an essential part of medical training and is not part of the general medical curriculum in Germany and the United States. To check whether there have been any recent changes, we contacted the Association of American Medical Colleges (AAMC), the national association that accredits U.S. medical schools, and asked if there "are any ongoing AAMC initiatives addressing numeracy (sometimes called 'statistical literacy') in medical school education?" The answer was "There are currently no AAMC initiatives in this area."

Statisticians have long criticized the fact that many introductory statistics texts in medicine are not written by experts on statistics and, furthermore, that this lack of expertise is even sold as a strength, as the renowned British statistician Michael J.R. Healy noticed:

> I do not know a single discipline other than statistics in which it is a positive recommendation for a new text book, worthy of being quoted on the dust cover, that it is not written by a specialist in the appropriate field. Would any medical reader read, would any medical publisher publish, my new introduction to brain surgery—so much simpler and more clearly written than those by professional brain surgeons, with their confusing mass of detail? I trust not. (Healy, 1979, p. 143)

As a result, some textbooks contain gross errors (see Altman & Bland, 1991; Eddy, 1982; Schönemann, 1969). Errors in textbooks and journals include confusion of conditional probabilities, as when equating the positive predictive value with the sensitivity, or the p-value with the probability that the null hypothesis is correct. These errors, however, also have a long history in psychology (Gigerenzer, 2004).

Yet it is important to go beyond this common critique. A curriculum with standard statistical techniques does not guarantee understanding health statistics, as we demonstrated in Part III. In contrast, teaching medical students transparent representations does foster understanding (Hoffrage, Gigerenzer, Krauss, & Martignon, 2002; Kurzenhäuser & Hoffrage, 2002). We believe that statistical literacy is more important for clinical practice than specific statistical techniques are (Appleton, 1990). In the end, medical schools need to ensure that

every graduate has minimal statistical literacy, if not a more advanced understanding.

Transparency

With the spread of democracies in the last century, transparency has become as highly valued as free speech and free press, for instance when fighting against corruption or for public access to disclosed information. The Vienna philosopher and political economist Otto Neurath (1882–1945) is one of the fathers of this social movement, who in the 1920s and '30s developed a strikingly beautiful symbolic way to represent economic facts to the largely uneducated Viennese public. This method allowed everyone to understand statistics in a "blink of an eye" by using pictorial representations called "isotypes" that conform to the psychology of vision (e.g., Neurath, 1946). Neurath's isotypes have not yet been adapted to health statistics, but various graphic representations are in use (Elmore & Gigerenzer, 2005; Galesic, Garcia-Retamero, & Gigerenzer, in press; Paling, 2003; Kurz-Milcke et al., 2008; Lipkus, 2007; Schapira, Nattinger, & McHorney, 2001). Here we focus on transparent tables and numbers (see also Fagerlin, Ubel, Smith, & Zikmund-Fisher, 2007; Peters, Hibbard, Slovic, & Dieckmann, 2007).

Numbers, Not Only Words

An important response to statistical illiteracy is to give the public more numbers. Patients have a right to learn how big benefits and harms of a treatment are. Qualitative risk terms are notoriously unclear. There are attempts to standardize verbal expressions, such as the EU guideline for drug labels and package leaflets, where specific terms are defined for frequency intervals. However, people seem to overestimate the frequencies of side effects based on those labels (Steckelberg, Berger, Köpke, Heesen, & Mühlhauser, 2005). Moreover, terms such as "unlikely" are interpreted differently from context to context. For example, more severe side effects are estimated to occur less frequently than less severe side effects described by the same qualitative term (Fischer & Jungermann, 1996). Patients tend to overestimate risks when disclosed verbally, and are less likely to comply if information is given numerically (Young & Oppenheimer, 2006). For both written and verbal information, patients had a more accurate perception of risk when it was numerical as opposed to verbal (see the review by Trevena, Davey, Barratt, Butow, & Caldwell, 2006). Therefore, risk should always be specified numerically.

Contrary to popular belief, studies report that a majority of patients do prefer numerical information to care only (Hallowell, Statham, Murton, Green, & Richards, 1997; Wallsten, Budescu, Zwick, & Kemp, 1993). Some studies have addressed differences between patients who do and do not prefer to see numbers. For instance, men who prefer to communicate with their physicians in words only ("no numbers, please") more often also prefer early aggressive surgery for prostate cancer over watchful waiting (Mazur, Hickam, & Mazur, 1999).

Data Tables: Drug Facts Boxes

While tables are routinely used to communicate data in scientific articles, there seems to be a hesitancy to use them in communicating with the general public. But tables are a practical way to look at and compare a series of numbers. To be efficient, such a table should be simple—that is, focus on the relevant information. We have developed a one-page summary of drug information at the heart of which is a study-findings table summarizing the benefit and side-effect data from trials used in the Food and Drug Administration's (FDA's) drug-approval process (Schwartz, Woloshin, & Welch, 2007). Compare the drug box on tamoxifen (Table 10) with the original advertisement (Fig. 16).

The table format provides a structure for readers to help them think about drug performance. By being given data outcomes side by side, readers are reminded that understanding an effect entails comparing what would happen with and without the drug. Similarly, presented with information about benefit and harm on the same page, readers are reminded that judging whether a drug is "worth it" means comparing good and harmful effects. Benefit needs to be judged in the context of harm, and vice versa. A small benefit may not be seen as sufficient if there are significant harms. Alternatively, significant harms may be tolerable in the context of substantial benefit. Another positive effect of presenting data symmetrically (i.e., providing absolute event rates for outcomes with and without the drug) is that information about benefit and harm is given equal weight: The numerical information is given in both percentages and frequencies. We have tested the drug box in two studies and both have demonstrated that people (even those with lower educational attainment) like it, think the data are valuable, and, most importantly, can understand information presented (Woloshin, Schwartz, & Welch, 2004; Schwartz, Woloshin, & Welch, 2007). We hope that such tables can become a routine element in communicating data to the public.

Transparent Numbers

In our final section, we summarize transparent and nontransparent ways to communicate health statistics (Table 11). They are arranged in pairs, with definitions and examples provided. In the literature, one sometimes finds a general distinction between probability format and frequency format. Yet there are different kinds of probabilities and frequencies, and some are less confusing than others (Brase, 2002, 2008; Gigerenzer & Hoffrage, 1995). For instance, an unconditional probability statement that specifies a reference class is clear ("The probability that a 50-year-old American woman will die of colon cancer in the next 10 years is 2 in 1,000"), whereas conditional probabilities tend to confuse ("the probability of colon cancer given a positive screening test" is often mistaken for "the probability of a positive screening test given colon cancer"). Table 11 distinguishes various kinds of probability and frequency representations.

Use frequency statements, not single-event probabilities. One nontransparent representation we have not discussed so far is a
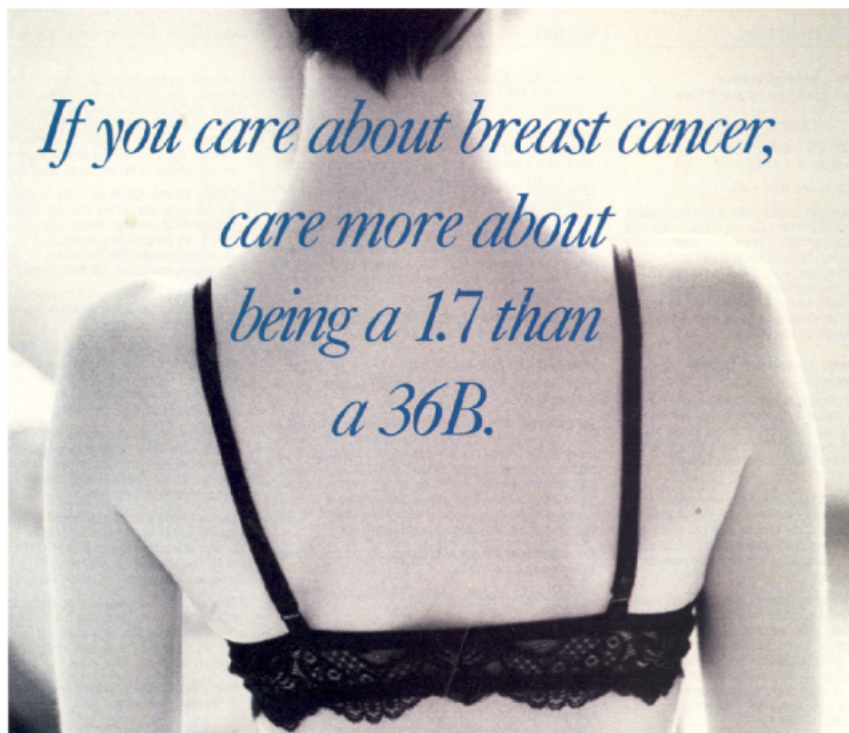
TABLE 10

Risk Chart Summarizing Benefits and Side Effects of a Drug so That Comparison Is Made Easy (From Schwartz, Woloshin, & Welch, 2007)

Prescription drug facts: NOLVADEX (tamoxifen)	
What is this drug for?	Reducing the chance of getting breast cancer
Who might consider taking it?	Women at high risk of getting breast cancer (1.7% or higher risk over 5 years). You can calculate your breast cancer risk at http://bcra.nci.nih.gov/btc.
Who should *not* take it?	Women who are pregnant or breastfeeding
Recommended testing	Have a yearly checkup that includes a gynecological examination and blood tests
Other things to consider doing	No other medicines are approved to reduce breast cancer risk for women who have not had breast cancer

NOLVADEX Study Findings		
13,000 women at high risk of getting breast cancer were given either NOLVADEX or a sugar pill for 5 years. Here's what happened:		
What difference did NOLVADEX make?	Women given a sugar pill	Women given NOLVADEX (20 mg a day)
Did NOLVADEX help?		
Fewer women got invasive breast cancer (16 in 1,000 fewer due to drug)	3.3% 33 in 1,000	1.7% 17 in 1,000
No difference in dying from breast cancer	About 0.09% in both groups or 0.9 in 1,000	
Did NOLVADEX have side effects?		
Life threatening side effects		
More women had a blood clot in their leg or lungs (additional 5 in 1,000 due to drug)	0.5% 5 in 1,000	1.0% 10 in 1,000
More women got invasive uterine cancer (additional 6 in 1,000 due to drug)	0.5% 5 in 1,000	1.1% 11 in 1,000
No difference in having a stroke	About 0.4% in both groups or 4 in 1,000	
Symptom side effects		
More women had hot flashes (additional 120 in 1,000 due to drug)	68% 680 in 1,000	80% 800 in 1,000
More women had vaginal discharge (additional 200 in 1,000 due to drug)	35% 350 in 1,000	55% 550 in 1,000
More women had cataracts needing surgery (additional 8 in 1,000 due to drug)	1.5% 15 in 1,000	2.3% 23 in 1,000
Bottom Line		
No difference in deaths from all causes combined	About 1.2% in both groups or 12 in 1,000	

How long has the drug been in use?
Nolvadex was first approved by the FDA in 1982. Studies show that most serious side effects or recalls of new drugs happen during their first 5 years of approval.

single-event probability statement. It is defined as a statement in which a probability refers to a singular person or event rather than to a class. A good illustration is weather prediction: "There is a 30% probability of rain tomorrow" is a single-event probability. By definition, no reference class is mentioned, but since people tend to think in terms of classes, misunderstanding is inevitable. Some citizens believe the statement to mean that it will rain tomorrow 30% *of the time,* others that it will rain in 30% *of the area*, or that it will rain on 30% *of the days* for which the announcement was made (Gigerenzer, Hertwig, van den Broek, Fasolo, & Katsikopoulos, 2005). The ambiguity of the reference class—time, area, or days—can be avoided by

Fig. 16. The original Nolvadex (tamoxifen) advertisement (compare to Table 10).

making a frequency statement, such as "it will rain in 30% of the days."

Similarly, when in clinical practice a physician tells a patient: "If you take Prozac, you have a 30 to 50% chance of developing a sexual problem, such as impotence or loss of interest," this single-event statement invites misunderstanding. As in the case of probabilities of rain, confusion will mostly go unnoticed. After learning of this problem, one psychiatrist changed the way he communicated the risk to his patients from single-event statements to frequency statements: "Out of every 10 patients who

TABLE 11

Some Confusing and Transparent Representations of Health Statistics

Confusing representation	Transparent representation
Single-event probabilities	*Frequency statements*
Definition: A probability that refers to an individual event or person, as opposed to a class of events or people, is called a single-event probability. In practice, single-event probabilities are often expressed in percentages, and occasionally as "X chances out of 100," rather than as a probability ranging between 0 and 1.	Definition: A frequency states the risk in relation to a specified reference class.
Example: "If you take Prozac, the probability that you will experience sexual problems is 30% to 50% (or: 30 to 50 chances out of 100)."	Example: "Out of every 10 of my patients who take Prozac, 3 to 5 experience a sexual problem."
Relative risks	*Absolute risks*
Definition: A relative risk is a ratio of the probabilities of the event occurring in one group (usually the treatment group) versus another group (usually the control group). The relative risk reduction of the treatment is calculated as 1 minus the relative risk: $$\text{Relative risk reduction} = 1 - \frac{P_{treatment}}{P_{control}}$$	Definition: The absolute risk in both the treatment and the control group is simply the corresponding baseline risk. The absolute risk reduction is calculated by subtracting the absolute risk in the treatment group from the absolute risk in the control group: $$\text{Absolute risk reduction} = P_{control} - P_{treatment}$$
Example: "Mammography screening reduces the risk of dying from breast cancer by about 20%."	Example: "Mammography screening reduces the risk of dying from breast cancer by about 1 in 1,000, from about 5 in 1,000 to about 4 in 1,000."
Survival rates	*Mortality rates*
Definition: The survival rate is the number of patients alive at a specified time *following diagnosis* (such as after 5 years) divided by the number of patients diagnosed.	Definition: The mortality rate is the number of people in a group who die annually from a disease, divided by the total number of people in the group.
Example: "The 5-year survival rate for people diagnosed with prostate cancer is 98% in the USA vs. 71% in Britain."	Example: "There are 26 prostate cancer deaths per 100,000 American men vs. 27 per 100,000 men in Britain."
Conditional probabilities	*Natural frequencies*
Definition: A conditional probability $p(A|B)$ is the probability of an event A given an event B.	Definition: A class of *N* events (persons) is subdivided into groups by two binary variables. The four resulting joint frequencies are called natural frequencies. Note that these are "raw counts" that sum up to *N*, unlike relative frequencies or conditional probabilities that are normalized with respect to the base rates of the event in question. Generalization to more than two variables and variable values are straightforward.
Example: See Figures 3 and 8.	Example: See Figures 3 and 8.

take Prozac, 3 to 5 experience a sexual problem." Psychologically that made a difference: Patients who were informed in terms of frequencies were less anxious about taking Prozac. When the psychiatrist asked his patients how they had understood the single-event statement, it turned out that many had thought that something would go awry in 30 to 50 percent of their sexual encounters (Gigerenzer, 2002). The psychiatrist had been thinking of all his patients who take Prozac, whereas his patients thought of themselves alone. Several studies have shown systematic differences in the interpretation of single-event and frequency statements (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Slovic, Monahan, & MacGregor, 2000; Tan et al., 2005).

Use absolute risks, not relative risks. There exist several reviews of studies comparing relative risks with absolute risks (Covey, 2007; Edwards, Elwyn, Covey, Matthews, & Pill, 2001;

McGettigan, Sly, O'Connell, Hill, & Henry, 1999; Moxey, O'Connell, McGettigan, & Henry, 2003). The common finding is that relative risk reductions lead people to systematically overestimate treatment effects. Why are relative risks confusing for many people? As mentioned before, this statistic is mute about the baseline risks (in Table 11: from 5 to 4 in 1,000) and the absolute effect size (1 in 1,000). Moreover, when patients hear about a 20% risk reduction, they are likely to think that this percentage refers to people like themselves, such as people who participate in cancer screening. Yet it refers to the baseline of people who do not participate in screening and die of cancer.

Use mortality rates, not survival rates. There seem to be no experimental studies about how patients or physicians understand survival rates compared to mortality rates. However, preliminary

evidence suggests that survival rates confuse physicians and make them draw unwarranted conclusions, while mortality rates are clearly understood (Wegwarth & Gaissmaier, 2008).

Use natural frequencies, not conditional probabilities. Estimating the probability of disease given a positive test (or any other posterior probability) is much easier with natural frequencies than with conditional probabilities (sensitivities and specificities). Note that this distinction refers to situations where *two* variables are considered: Natural frequencies are *joint* frequencies, as shown in Figures 3 and 8. Gigerenzer and Hoffrage (1995, 1999) showed that natural frequencies—but not relative frequencies—facilitate judgments. This fact has been repeatedly misrepresented in the literature, where our thesis is often held to be that *all* frequency representations improve judgments (see Hoffrage et al., 2002).

Caution

It should be noted that providing people with accurate, balanced, accessible data on disease risk and treatment benefit could have an untoward side effect. People may be very surprised about how small many of the risks and benefits are. Consequently, they may dismiss as unimportant interventions that physicians see as extremely valuable. For example, in one of our studies (Woloshin, Schwartz, & Welch, 2004), participants were very optimistic about the effectiveness of three different drugs; in each case, these perceptions dropped substantially after seeing the actual data. The effect, however, was similar for all drugs. This is concerning, since one of the drugs, a statin used to treat men with high cholesterol but no prior myocardial infarction, showed a reduction of overall mortality over 5 years from 4 in 100 to 3 in 100 patients. We suspect that many respondents did not appreciate the real magnitude of this effect: Few drugs now being manufactured can match this reduction in all-cause mortality among relatively healthy outpatients. To truly judge how well a drug (or other intervention) works, people need a context—that is, some sense of the magnitude of the benefit of other interventions. Undoubtedly, most people lack such knowledge and overestimate the benefits of drugs. We believe that reactions to benefit data will change as people have more exposure to them; that is, as consumers become better calibrated to effect sizes, they will be better able to discriminate among drugs and interventions. It is important to provide this context to make sure consumers do not discount small but important effects.

Reference Class and Transparency

Much of the mental confusion that defines nontransparency seems to be caused by the reference class to which a health statistic applies (Gigerenzer & Edwards, 2003). Single-event probabilities specify by definition no class of events, and relative risks often refer to a reference class that is different from the class people are thinking of. Sensitivities and specificities are conditional on two different reference classes (patients with disease and patients without disease), whereas natural frequencies all refer to the same reference class (all patients). And survival and mortality rates crucially differ in their denominator—that is, the class of events they refer to. Clarity about the reference class to which a health statistic refers is one of the central tools in attaining health literacy.

VII. THE DREAM OF STATISTICAL LITERACY

Two millennia separated the Athens of Aristotle and the Paris of Claude Bernard, but the two men shared one article of faith: Science is about causes, not chances. Not until 1654, when the French mathematicians Blaise Pascal and Pierre Fermat exchanged letters on gambling problems, did mathematical probability arrive on the scene. This curiously late appearance was christened "the scandal of philosophy" by philosopher Ian Hacking (1975). In the following centuries, the "probabilistic revolution" (Krüger, Gigerenzer, & Morgan, 1987) changed science and everyday life, beginning slowly but resulting in enormous transformations. It turned deterministic physics into statistical mechanics and quantum theory, changed biology by introducing Darwinian variation and random drift, and redefined the nature of scientific experiments by introducing repetition and randomization. Yet this revolution in thought has not yet reached patients and physicians in their understanding of health statistics.

We hope that this monograph stimulates researchers to contribute to solving the problem of collective statistical illiteracy and to develop and implement efficient and transparent representations of health statistics. Nonetheless, the dream of statistical literacy is of a broader scope and is fundamental to a functioning democracy. It embodies the Enlightenment ideal of people's emergence from their self-imposed immaturity. In Kant's (1784) words, "Dare to know!"

REFERENCES

AAAS (American Association for the Advancement of Science). (2006, July 18). *Pinholster presents science communication survey at EuroScience 2006.* Retrieved April 1, 2008, from http://www.aaas.org/news/releases/2006/0718euroscience.shtml

Altman, D.G., & Bland, J.M. (1991). Improving doctors' understanding of statistics. *Journal of the Royal Statistical Society, Series A, 154,* 223–267.

Apotheken Umschau (2006). Mammografie für alle - Ist das sinnvoll? [Mammographies for everyone: A good idea?]. Retrieved April 8, 2008, from http://www.gesundheitpro.de/Onkologie-2-Mammografie-fuer-alle-Brustkrebs-A060425COCHP023606.html

Appleton, D.R. (1990). What statistics should we teach medical undergraduates and graduates? *Statistics in Medicine*, *9*, 1013–1021.

Ärztekammer Berlin. (2002, March 21). Politischer Aktionismus führt zu Über- und Fehlversorgung [Political overreaction leads to overtreatment and mistreatment]. (Press release). Berlin: Author.

Bachmann, L.M., Gutzwiller, F.S., Puhan, M.A., Steurer, J., Steurer-Stey, C., & Gigerenzer, G. (2007). Do citizens have minimum medical knowledge? A survey. *BMC Medicine*, *5*, Article 14. DOI: 10.1186/1741-7015-1185-1114.

Baines, C.J. (1992). Women and breast cancer: Is it really possible for the public to be well informed? *The Canadian Medical Association Journal*, *142*, 2147–2148.

Barry, M.J., Fowler, F.J., Mulley, A.G., Henderson, J.V., & Wennberg, J.E. (1995). Patient reactions to a program designed to facilitate patient participation in treatment decisions for benign prostatic hyperplasia. *Medical Care*, *33*, 771–782.

Beisecker, A.E., & Beisecker, T.D. (1990). Patient information-seeking behaviors when communicating with physicians. *Medical Care*, *28*, 19–28.

Berg, N., Biele, G., & Gigerenzer, G. (2008). *Logical consistency and accuracy of beliefs: Survey evidence on health decision-making among economists*. Unpublished manuscript.

Bernard, C. (1957). *An introduction to the study of experimental medicine* (H.C. Greene, Trans.). New York: Dover (Original work published 1865)

Berwick, D.M., Fineberg, H.V., & Weinstein, M.C. (1981). When doctors meet numbers. *The American Journal of Medicine*, *71*, 991–998.

Biehler, R., Hofmann, T., Maxara, C., & Prömmel, A. (2006). *Fathom 2–Eine Einführung*. Heidelberg: Springer.

Braddock, C.H., Edwards, K.A., Hasenberg, N.M., Laidley, T.L., & Levinson, W. (1999). Informed decision making in outpatient practice: Time to get back to basics. *JAMA: The Journal of the American Medical Association*, *282*, 2313–2320.

Bramwell, R., West, H., & Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: Experimental study. *British Medical Journal*, *333*, 284–286.

Brase, G.L. (2002). Ecological and evolutionary validity: Comments on Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni's (1999) mental model theory of extensional reasoning. *Psychological Review*, *109*, 722–728.

Brase, G.L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, *15*, 284–289.

Bright, G.W., & Friel, S.N. (1998). Graphical representations: Helping students interpret data. In S.P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K–12* (pp. 63–88). Mahwah, NJ: Erlbaum.

Broadbent, E., Petrie, K.J., Ellis, C.J., Anderson, J., Gamble, G., Anderson, D., et al. (2006). Patients with acute myocardial infarction have an inaccurate understanding of their risk of a future cardiac event. *Internal Medicine Journal*, *36*, 643–647.

Bundesministerium für Gesundheit (2002a, March 23). Einführung von Mammographie Screening: Unberechtigte Kritik der Ärztekammer Berlin [Implementation of mammography screening: Unjustified criticism from the Berlin Chamber of Physicians] (Press release).

Bundesministerium für Gesundheit (2002b, September 24). Ulla Schmidt: Neue Schritte zur Qualitätssicherung bei Brustkrebs [New steps towards quality control with breast cancer] (Press release).

Butterworth, B. (1999). *What counts: How every brain is hardwired for math*. New York: Free Press.

Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, *299*, 999–1000.

Cassels, A., Hughes, M.A., Cole, C., Mintzes, B., Lexchin, J., & McCormack, J.P. (2003). Drugs in the news: An analysis of Canadian newspaper coverage of new prescription drugs. *Canadian Medical Association Journal*, *168*, 1133–1137.

Center for the Evaluative Clinical Sciences Staff (Ed.). (1996). *The Dartmouth atlas of health care*. Chicago: American Hospital Association.

Charles, C.A., Gafni, A., & Whelan, T. (1997). Shared decision-making in the medical encounter: What does it mean? (or, It takes at least two to tango). *Social Science and Medicine*, *44*, 681–692.

Coleman, W. (1987). Experimental physiology and statistical inference: The therapeutic trial in nineteenth-century Germany. In L. Krüger, G. Gigerenzer, & M.S. Morgan (Eds.), *The probabilistic revolution, Vol. II: Ideas in the Sciences* (pp. 201–226). Cambridge, MA: MIT Press.

Collins, E.D., Kerrigan, C.L., & Anglade, P. (1999, July/August). Surgical treatment of early breast cancer: What would surgeons choose for themselves? *Effective Clinical Practice* 149–151.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Covey, J. (2007). A meta-analysis of the effects of presenting treatment benefits in different formats. *Medical Decision Making*, *27*, 638–654.

Daston, L. (1988). *Classical probability in the Enlightenment*. Princeton, NJ: Princeton University Press.

Dobbs, M. (2007, October 30). Rudy wrong on cancer survival chances. Retrieved July 21, 2008 from http://blog.washingtonpost.com/fact-checker/2007/10/rudy_miscalculates_cancer_surv.html

Domenighetti, G., D'Avanzo, B., Egger, M., Berrino, F., Perneger, T., Mosconi, P., et al. (2003). Women's perception of the benefits of mammography screening: Population-based survey in four countries. *International Journal of Epidemiology*, *32*, 816–821.

Eddy, D.M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge: Cambridge University Press.

Eddy, D.M. (1996). *Clinical decision making: From theory to practice: A collection of essays from the Journal of the American Medical Association*. Boston: Jones and Bartlett Publishers.

Edwards, A., Elwyn, G.J., Covey, J., Matthews, E., & Pill, R. (2001). Presenting risk information: A review of the effects of "framing" and other manipulations on patient outcomes. *Journal of Health Communication*, *6*, 61–82.

Egger, M., Bartlett, C., & Juni, P. (2001). Are randomised controlled trials in the BMJ different? *British Medical Journal*, *323*, 1253–1254.

Elmore, J.G., Barton, M.B., Moceri, V.M., Polk, S., Arena, P.J., & Fletcher, S.W. (1998). Ten-year risk of false positive screening mammograms and clinical breast examinations. *The New England Journal of Medicine*, *338*, 1089–1096.

Elmore, J.G., & Gigerenzer, G. (2005). Benign breast disease: The risk of communicating risks. *The New England Journal of Medicine*, *353*, 297–299.

Estrada, C., Barnes, V., Collins, C., & Byrd, J.C. (1999). Health literacy and numeracy. *JAMA: The Journal of the American Medical Association*, *282*, 527.

Fagerlin, A., Ubel, P.A., Smith, D.M., & Zikmund-Fisher, B.J. (2007). Making numbers matter: Present and future research in risk communication. *American Journal of Health Behavior*, *31*, 47–56.

Fahey, T., Griffiths, S., & Peters, T.J. (1995). Evidence based purchasing: Understanding results of clinical trials and systematic reviews. *British Medical Journal*, *311*, 1056–1059.

Falk, R., & Konold, C. (1992). The psychology of learning probability. In F.S. Gordon & S.P. Gordon (Eds.), *Statistics for the twenty-first century* (pp. 151–164). Washington, DC: The Mathematical Association of America.

Federman, D.G., Goyal, S., Kamina, A., Peduzzi, P., & Concato, J. (1999). Informed consent for PSA screening: Does it happen? *Effective Clinical Practice*, *2*, 152–157.

Felix Burda Stiftung. (2008, February 19). Felix Burda Stiftung startet neue Medien-Kampagne. [Felix Burda Foundation starts new media campaign] (Press release). Retrieved July 27, 2008, from http://www.felix-burda-stiftung.de/index.php?id=46&tx_cwtpresscenter_pi1[showUid]=236&cHash=988c9aeaf5

Finzer, B., & Erickson, T. (2006). *Fathom. Emeryville*. CA: Key Curriculum Press.

Fischer, K., & Jungermann, H. (1996). Rarely occurring headaches and rarely occurring blindness: Is rarely = rarely? The meaning of verbal frequentistic labels in specific medical contexts. *Journal of Behavioral Decision Making*, *9*, 153–172.

Fletcher, S.W. (1997). Whither scientific deliberation in health policy recommendations? Alice in the Wonderland of breast-cancer screening. *New England Journal of Medicine*, *336*, 1180–1183.

Folkman, J., & Kalluri, R. (2004). Cancer without disease. *Nature*, *427*, 787.

Franklin, B. (1987). *Writings*. New York: The Library of America. (Original work published 1789).

Furedi, A. (1999). The public health implications of the 1995 'pill scare.' *Human Reproduction Update*, *5*, 621–626.

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (in press). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, *75*, 372–396.

Gelman, R., & Gallistel, C.R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.

General Medical Council. (1998). *Seeking patients' consent: The ethical considerations*. London: Author.

Ghosh, A.K., & Ghosh, K. (2005). Translating evidence-based information into effective risk communication: Current challenges and opportunities. *Journal of Laboratory and Clinical Medicine*, *145*, 171–180.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster. (UK version: *Reckoning with risk: Learning to live with uncertainty*, London: Penguin).

Gigerenzer, G. (2003). Why does framing influence judgment? *Journal of General Internal Medicine*, *18*, 960–961.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587–606.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. New York: Viking.

Gigerenzer, G. (2008). *A survey of attitudes about risk and uncertainty*. Unpublished manuscript.

Gigerenzer, G., & Edwards, A.G.K. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal*, *327*, 741–744.

Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B., & Katsikopoulos, K.V. (2005). "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis*, *25*, 623–629.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.

Gigerenzer, G., & Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: A reply to Lewis & Keren and Mellers & McGraw. *Psychological Review*, *106*, 425–430.

Gigerenzer, G., Hoffrage, U., & Ebert, A. (1998). AIDS counselling for low-risk clients. *AIDS Care*, *10*, 197–211.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

Gigerenzer, G., Mata, J., & Frank, R. (2008). *A survey of health knowledge in seven European countries*. Unpublished manuscript.

Gigerenzer, G., & Murray, D.J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge: Cambridge University Press.

Gnanadesikan, M., Scheaffer, R.L., & Swift, J. (1987). *The arts and techniques of simulation*. Palo Alto, CA: Dale Seymour Publications.

Gøtzsche, P.C., & Nielsen, M. (2006). Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews* (4), Article CD001877. DOI: 001810.001002/14651858.CD14001877.pub14651852.

Gould, S.J. (1992). *Bully for brontosaurus: Further reflections in natural history*. New York: Penguin Books.

Gray, J.A.M., Patnick, J., & Blanks, R.G. (2008). Maximising benefit and minimising harm of screening. *British Medical Journal*, *336*, 480–483.

Haag, L., & Stern, E. (2003). In search of the benefits of learning Latin. *Journal of Educational Psychology*, *95*, 174–178.

Hacking, I. (1975). *The emergence of probability*. Cambridge: Cambridge University Press.

Hallowell, N., Statham, H., Murton, F., Green, J., & Richards, M. (1997). 'Talking about chance': The presentation of risk information during genetic counseling for breast and ovarian cancer. *Journal of Genetic Counseling*, *6*, 269–286.

Hamm, R.M., & Smith, S.L. (1998). The accuracy of patients' judgments of disease probability and test sensitivity and specificity. *The Journal of Family Practice*, *47*, 44–52.

Harrington, J., Noble, L.M., & Newman, S.P. (2004). Improving patients' communication with doctors: A systematic review of intervention studies. *Patient Education & Counseling*, *52*, 7–16.

Hartmann, L.C., Schaid, D.J., Woods, J.E., Crotty, T.P., Myers, J.L., Arnold, P.G. et al. (1999). Efficacy of bilateral prophylactic mastectomy in women with a family history of breast cancer. *New England Journal of Medicine*, *340*, 77–84.

Hartz, J., & Chappell, R. (1997). *Worlds apart. How the distance between science and journalism threatens America's future*. Nashville, TN: First Amendment Center.

Healy, M.J.R. (1979). Does medical statistics exist? *Bulletin of Applied Statistics*, *6*, 137–182.

Hembroff, L.A., Holmes-Rovner, M., & Wills, C.E. (2004). Treatment decision-making and the form of risk communication: Results of a factorial survey. *BMC Medical Informatics and Decision Making*, *4*. DOI: 10.1186/1472-6947-1184-1120.

Henneman, L., Timmermans, D.R.M., & van der Wal, G. (2004). Public experiences, knowledge and expectations about medical genetics and the use of genetic information. *Community Genetics*, *7*, 33–43.

Hoffrage, U. (2003). Risikokommunikation bei Brustkrebsfrüherkennung und Hormonersatztherapie. [Risk communication in the early identification of breast cancer and hormone-replacement therapy]. *Zeitschrift für Gesundheitspsychologie*, *11*, 76–86.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*, 538–540.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, *84*, 343–352.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science*, *290*, 2261–2262.

Horton, R. (2004, March 11). The dawn of McScience. *The New York Review of Books*, *51*, 7–9.

Høye, S.P.H. (2002). "New wonder pill"!—What do Norwegian newspapers write about new medications. *Tidsskr Nor Lægeforen*, *122*, 1671–1676.

Impiccatore, P., Pandolfini, C., Casella, N., & Bonati, M. (1997). Reliability of health information for the public on the world wide web: Systemic survey of advice on managing fever in children at home. *British Medical Journal*, *314*, 1875–1879.

Jahnke, T., & Wuttke, H. (Eds.). (2005). *Mathematik: Stochastik*. Berlin: Cornelsen Verlag.

Jorgensen, K.J., & Gøtzsche, P.C. (2004). Presentation on websites of possible benefits and harms from screening for breast cancer: Cross sectional study. *British Medical Journal*, *328*, 148–151.

Jorgensen, K.J., & Gøtzsche, P.C. (2006). Content of invitations for publicly funded screening mammography. *British Medical Journal*, *332*, 538–541.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454. Reprinted in D. Kahneman et al. (1982) (Eds.), Judgment under uncertainty: Heuristics and biases (pp. 1932–1947). Cambridge, UK: Cambridge University Press.

Kaiser, T., Ewers, H., Waltering, A., Beckwermert, D., Jennen, C., & Sawicki, P.T. (2004). Sind die Aussagen medizinischer Werbeprospekte korrekt? *Arznei-Telegramm*, *35*, 21–23.

Kalet, A., Roberts, J.C., & Fletcher, R. (1994). How do physicians talk with their patients about risks? *Journal of General Internal Medicine*, *9*, 402–404.

Kant, I. (1784). Beantwortung der Frage: Was ist Aufklärung? [What is enlightenment?]. *Berlinische Monatsschrift*, *Dezember-Heft* 481–494.

Kaphingst, K.A., DeJong, W., Rudd, R.E., & Daltroy, L.H. (2004). A content analysis of direct-to-consumer television prescription drug advertisements. *Journal of Health Communication*, *9*, 515–528.

Kaphingst, K.A., Rudd, R.E., DeJong, W., & Daltroy, L.H. (2005). Comprehension of information in three direct-to-consumer television prescription drug advertisements among adults with limited literacy. *Journal of Health Communication*, *10*, 609–619.

Kassenärztliche Bundesvereinigung (2004). Einführung eines bundesweiten Mammographie-Screening-Programms [Implementation of a national mammography screening program]. *Beilage zum Deutschen Ärzteblatt*, *4*, 1–44.

Kees, B. (2002, October 18). Newsroom training: Where's the investment! Retrieved March 18, 2008, from the John S. and James L. Knight Foundation Web site: http://www.knightfoundation.org/permalink/178205/224861.

Kerlikowske, K., Grady, D., Barclay, J., Sickles, E.A., & Ernster, V. (1996a). Effect of age, breast density, and family history on the sensitivity of first screening mammography. *Journal of the American Medical Association*, *276*, 33–38.

Kerlikowske, K., Grady, D., Barclay, J., Sickles, E.A., & Ernster, V. (1996b). Likelihood ratios for modern screening mammography: Risk of breast cancer based on age and mammographic interpretation. *Journal of the American Medical Association*, *276*, 39–43.

Koehler, J.J. (1996). On conveying the probative value of DNA evidence: Frequencies, likelihood ratios, and error rates. *University of Colorado Law Review*, *67*, 859–886.

Konold, C., & Miller, C. (2005). *TinkerPlots*. Emeryville, CA: Key Curriculum Press.

Krüger, L., Gigerenzer, G., & Morgan, M.S. (Eds.). (1987). *The probabilistic revolution, Vol. II: Ideas in the sciences*. Cambridge, MA: MIT Press.

Kurzenhäuser, S. (2003). Welche Informationen vermitteln deutsche Gesundheitsbroschüren über die Screening-Mammographie? [What information do German health brochures provide on mammography screening?] *Zeitschrift für ärztliche Fortbildung und Qualitätssicherung*, *97*, 53–57.

Kurzenhäuser, S., & Hoffrage, U. (2002). Teaching Bayesian reasoning: An evaluation of a classroom tutorial for medical students. *Medical Teacher*, *24*, 516–521.

Kurz-Milke, E., Gigerenzer, G., & Martignon, L. (2008). Transparency in risk communication: Graphical and analog tools. *Annals of the New York Academy of Sciences*, *1128*, 18–28.

Kurz-Milke, E., & Martignon, L. (2007). Stochastische Urnen und Modelle in der Grundschule [Stochastic urns and models in elementary school]. In G. Kaiser (Ed.), *Tagungsband der Jahrestagung der Gesellschaft für Didaktik der Mathematik, Berlin*. Hildesheim: Verlag Franzbecker.

Labarge, A.S., McCaffrey, R.J., & Brown, T.A. (2003). Neuropsychologists' ability to determine the predictive value of diagnostic tests. *Clinical Neuropsychology*, *18*, 165–175.

Larson, R.J., Woloshin, S., Schwartz, B., & Welch, H.G. (2005). Celebrity endorsements of cancer screening. *Journal of the National Cancer Institute*, *97*, 693–695.

Lauterbach, K.W. (2002, August 28). 100 000 überflüssige Operationen [Letter to the editor]. *Die Zeit*, p. 16.

Lerman, C., Trock, B., Rimer, B.K., Jepson, C., Brody, D., & Boyce, A. (1991). Psychological side effects of breast cancer screening. *Health Psychology*, *10*, 259–267.

Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics*, *43*, 147–163.

Lipkus, I.M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, *27*, 696–713.

Lipkus, I.M., Samsa, G., & Rimer, B.K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*, 37–44.

Malenka, D.J., Baron, J.A., Johansen, S., Wahrenberger, J.W., & Ross, J.M. (1993). The framing effect of relative versus absolute risk. *Journal of General Internal Medicine*, *8*, 543–548.

Martignon, L., & Wassner, C. (2005). Schulung frühen stochastischen Denkens von Kindern [Teaching children early stochastic thinking]. *Zeitschrift für Erziehungswissenschaften*, *8*, 202–222.

Mathematics and medicine. (1937, January 2). *The Lancet*, i, 31.

Mazur, D.J., Hickam, D.H., & Mazur, M.D. (1999). How patients' preferences for risk information influence treatment choice in a case of

high risk and high therapeutic uncertainty: Asymptomatic localized prostate cancer. *Medical Decision Making*, *194*, 394–398.

McGettigan, P., Sly, K., O'Connell, D., Hill, S., & Henry, D. (1999). The effects of information framing on the practices of physicians. *Journal of General Internal Medicine*, *14*, 633–642.

Mooi, W.J., & Peeper, D.S. (2006). Oncogene-induced cell senescence—Halting on the road to cancer. *New England Journal of Medicine*, *355*, 1037–1046.

Moore, D.S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*, 123–165.

Morris, R.W. (2002). Does EBM offer the best opportunity yet for teaching medical statistics? *Statistics in Medicine*, *21*, 969–977.

Moumjid, N., Gafni, A., Bremond, A., & Carrere, M.-O. (2007). Shared decision making in the medical encounter: Are we all talking about the same thing? *Medical Decision Making*, *27*, 539–546.

Moxey, A., O'Connell, D., McGettigan, P., & Henry, D. (2003). Describing treatment effects to patients: How they are expressed makes a difference. *Journal of General Internal Medicine*, *18*, 948–959.

Moynihan, R., Bero, L., Ross-Degnan, D., Henry, D., Lee, K., Watkins, J. et al. (2000). Coverage by the news media of the benefits and risks of medications. *The New England Journal of Medicine*, *342*, 1645–1650.

Mühlhauser, I., & Höldke, B. (2002). Information zum Mammographiescreening – vom Trugschluss zur Ent-Täuschung [Information on mammography screening: From deception to dis-illusionment]. *Radiologe*, *42*, 299–304.

Mühlhauser, I., Kasper, J., & Meyer, G. (2006). FEND: Understanding of diabetes prevention studies: Questionnaire survey of professionals in diabetes care. *Diabetologia*, *49*, 1742–1746.

Murphy, M. (1993). The contraceptive pill and women's employment as factors in fertility change in Britain 1963–1980: A challenge to the conventional view. *Population Studies*, *47*, 221–243.

National Cancer Institute. (1998). Breast cancer risk tool: An interactive patient education tool [Computer software]. Bethesda, MD: Author.

National Cancer Institute. (2005). Fact sheet: Breast cancer prevention studies. Bethesda, MD: Author. Retrieved July 8, 2008, from http://www.cancer.gov/cancertopics/factsheet/Prevention/breast-cancer

NCTM (National Council of Teachers of Mathematics). (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Institutes of Health Consensus Conference. (1991). Treatment of early-stage breast cancer. *Journal of the American Medical Association*, *265*, 391–395.

National Institutes of Health Consensus Development Panel. (1997). National Institutes of Health Consensus Development Conference statement: Breast cancer screening for women ages 40–49, January 21–23, 1997. *Journal of the National Cancer Institute*, *89*, 1015–1020.

Naylor, C.D., Chen, E., & Strauss, B. (1992). Measured enthusiasm: Does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Annals of Internal Medicine*, *117*, 916–921.

Neurath, O. (1946). *From hieroglyphics to isotypes*. London: Future Books.

Noormofidi, D. (2006, May 3). Zank um Brustkrebs [Strife over breast cancer]. *DieStandard.at*. Retrieved March 28, 2008, from http://diestandard.at/?url=/?id=2431198

Nuovo, J., Melnikow, J., & Chang, D. (2002). Reporting number need to treat and absolute risk reduction in randomized controlled trials. *JAMA: The Journal of the American Medical Association*, *287*, 2813–2814.

Österreichische Ärztekammer (2005). Österreichische Ärztekammer: Homöopathie kein Placebo. Viele internationale Studien ergeben positive Wirkungsweise [Austrian Chamber of Physicians: Homeopathy is not a placebo. Many international studies show positive results]. Retrieved April 5, 2008 from http://www.aerztekammer.at/index.php?id=0000000000020050919115742&aid=xhtml&id=0000000000020050919115742&type=module&noedit=true

Paling, J. (2003). Strategies to help patients understand risks. *British Medical Journal*, *327*, 745–748.

Peters, E., Hibbard, J., Slovic, P., & Dieckmann, N. (2007). Numeracy skill and the communication, comprehension, and use of risk and benefit information. *Health Affairs*, *26*, 741–748.

Phillips, E., Lappan, G., Winter, M.J., & Fitzgerald, G. (1986). *Middle Grades Mathematics Project: Probability*. Menlo Park, CA: Addison-Wesley Publishing Company.

Politi, M.C., Han, P.K.J., & Col, N.F. (2007). Communicating the uncertainty of harms and benefits of medical interventions. *Medical Decision Making*, *27*, 681–695.

Porter, T.M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Ransohoff, D.F., & Harris, R.P. (1997). Lessons from the mammography screening controversy: Can we improve the debate? *Annals of Internal Medicine*, *127*, 1029–1034.

Rao, G. (2008). Physician numeracy: Essential skills for practicing evidence-based medicine. *Family Medicine*, *40*, 354–358.

Rásky, É., & Groth, S. (2004). Informationsmaterialien zum Mammographiescreening in Österreich – Unterstützen sie die informierte Entscheidung von Frauen? [Information on mammography screening in Austria – Does it facilitate women's informed decisions?]. *Sozial und Präventivmedizin*, *49*, 301–397.

Reimer, L., Mottice, S., Schable, C., Sullivan, P., Nakashima, A., Rayfield, M. et al. (1997). Absence of detectable antibody in a patient infected with human immunodeficiency virus. *Clinical Infectious Diseases*, *25*, 98–100.

Reyna, V.F., & Brainerd, C.J. (2007). The importance of mathematics in health and human judgment: Numeracy, risk communication, and medical decision making. *Learning and Individual Differences*, *17*, 147–159.

Rigby, M., Forsström, J., Roberts, R., & Wyatt, J., for the TEAC-Health Partners. (2001). Verifying quality and safety in health informatics services. *British Medical Journal*, *323*, 552–556.

Rimm, A.A., & Bortin, M. (1978). Clinical trials as a religion. *Biomedicine Special Issue*, *28*, 60–63.

Risueño d'Amador, B.J.I. (1836). Mémoire sur le calcul des probabilités appliqué à la médecine [Memoir on the calculation of probabilities applied to medicine]. *Bulletin de l'Academie Royale de Médicine*, *1*, 622–680.

Roter, D.L., & Hall, J.A. (1993). *Doctors talking with patients/patients taking with doctors: Improving communication in medical visits*. London: Auburn House.

Rowe, G., Frewer, L., & Sjoberg, L. (2000). Newspaper reporting of hazards in the UK and Sweden. *Public Understanding of Science*, *9*, 59–78.

Ruscio, J. (2003). Comparing Bayes's Theorem to frequency-based approaches to teaching Bayesian reasoning. *Teaching of Psychology*, *30*, 325–328.

Sarfati, D., Howden-Chapman, P., Woodward, A., & Salmond, C. (1998). Does the frame affect the picture? A study into how attitudes to screening for cancer are affected by the way benefits are expressed. *Journal of Medical Screening*, *5*, 137–140.

Savage, L.J. (1972). *The foundations of statistics*. NY: Dover.

Schapira, M., Nattinger, A., & McHorney, C. (2001). Frequency or probability? A qualitative study of risk communication formats used in health care. *Medical Decision Making*, *21*, 459–467.

Schönemann, P. (1969). Review of *Faktorenanalyse* by K. Uberla and *Dimensionen des Verhaltens* by Kurt Pawlik. *Biometrics*, *25*, 604–607.

Schüssler, B. (2005). Im Dialog: Ist Risiko überhaupt kommunizierbar, Herr Prof. Gigerenzer? [Interview with Gerd Gigerenzer: Can risk be communicated at all?]. *Frauenheilkunde Aktuell*, *14*, 25–31.

Schwartz, L.M., & Woloshin, S. (2000). Physician grand round survey (unpublished data).

Schwartz, L.M., & Woloshin, S. (2007). Participation in mammography screening. *British Medical Journal*, *335*, 731–732.

Schwartz, L.M., Woloshin, S., Black, W.C., & Welch, H.G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127*, 966–972.

Schwartz, L.M., Woloshin, S., Dvorin, E.L., & Welch, H.G. (2006). Ratio measures in leading medical journals: Structured review of accessibility of underlying absolute risks. *British Medical Journal*, *333*, 1248–1252.

Schwartz, L.M., Woloshin, S., Fowler, F.J., Jr, & Welch, H.G. (2004). Enthusiasm for cancer screening in the United States. *Journal of the American Medical Association*, *291*, 71–78.

Schwartz, L.M., Woloshin, S., Sox, H.C., Fischhoff, B., & Welch, H.G. (2000). U.S. women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: Cross sectional survey. *British Medical Journal*, *320*, 1635–1640.

Schwartz, L.M., Woloshin, S., & Welch, H.G. (1999a). Misunderstandings about the effect of race and sex on physicians' referrals for cardiac catheterization. *New England Journal of Medicine*, *341*, 279–283.

Schwartz, L.M., Woloshin, S., & Welch, H.G. (1999b). Risk communication in clinical practice: Putting cancer in context. *Monograph of the National Cancer Institute*, *25*, 124–133.

Schwartz, L.M., Woloshin, S., & Welch, H.G. (2005). Can patients interpret health information? An assessment of the medical data interpretation test. *Medical Decision Making*, *25*, 290–300.

Schwartz, L.M., Woloshin, S., & Welch, H.G. (2007). The drug facts box: Providing consumers with simple tabular data on drug benefit and harm. *Medical Decision Making*, *27*, 655–662.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*, 380–400.

Sedrakyan, A., & Shih, C. (2007). Improving depiction of benefits and harms: Analyses of studies of well-known therapeutics and review of high-impact medical journals. *Medical Care*, *45*, 523–528.

Serrano, M. (2007). Cancer regression by senescence. *New England Journal of Medicine*, *356*, 1996–1997.

Shang, A., Huwiler-Müntener, K., Nartey, L., Jüni, P., Dörig, S., Sterne, J. et al. (2005). Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. *The Lancet*, *366*, 726–732.

Shaughnessy, J.M. (1992). Research on probability and statistics: Reflections and directions. In D.A. Grouws (Ed.), *Handbook of research on mathematical teaching and learning* (pp. 465–499). New York: Macmillan.

Sheridan, S., Pignone, M.P., & Lewis, C.L. (2003). A randomized comparison of patients' understanding of number needed to treat

and other common risk reduction formats. *Journal of General Internal Medicine*, *18*, 884–892.

Shibata, A., & Whittemore, A.S. (2001). Re: Prostate cancer incidence and mortality in the United States and the United Kingdom. *Journal of the National Cancer Institute*, *93*, 1109–1110.

Slaytor, E.K., & Ward, J.E. (1998). How risks of breast cancer and benefits of screening are communicated to women: Analysis of 58 pamphlets. *British Medical Journal*, *317*, 263–264.

Sleath, B., Roter, D.L., Chewning, B., & Svarstad, B. (1999). Question-asking about medications: Physician experiences and perceptions. *Medical Care*, *37*, 1169–1173.

Slovic, P., Monahan, J., & MacGregor, D.G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior*, *24*, 271–296.

Smith, D.E., Wilson, A.J., & Henry, D.A. (2005). Monitoring the quality of medical news reporting: Early experience with media doctor. *The Medical Journal of Australia*, *183*, 190–193.

Smith, R. (2005). Medical journals are an extension of the marketing arm of pharmaceutical companies. *PLoS Medicine*, *2*, e138. DOI: 110.1371/journal.pmed.0020138

Snijders, R.J., Noble, P., Sebire, N., Souka, A., & Nicolaides, K.H. (1998). UK multicentre project assessment of risk of trisomy 21 by maternal age and fetal nuchal-translucency thickness at 10–14 weeks of gestation. *Lancet*, *352*, 343–346.

Sone, S., Li, F., Yang, Z., Honda, T., Maruyama, Y., & Takashima, S. (2001). Results of three-year mass screening programme for lung cancer using mobile lowdose spiral computed tomography scanner. *British Journal of Cancer*, *84*, 25–32.

Steckelberg, A., Berger, B., Köpke, S., Heesen, C., & Mühlhauser, I. (2005). Kriterien für evidenzbasierte Patienteninformationen [Criteria for evidence-based information for patients]. *Zeitschrift für ärztliche Fortbildung und Qualität im Gesundheitswesen*, *99*, 343–351.

Steckelberg, A., Hülfenhaus, C., Kasper, J., Rost, J., & Mühlhauser, I. (2007). How to measure critical health competences: Development and validation of the Critical Health Competence Test (CHC Test). *Advances in Health Sciences Education*. DOI: 10.1007/s10459-10007-19083-10451

Steckelberg, A., Hülfenhaus, C., Kasper, J., Rost, J., & Mühlhauser, I. (2008). *Ebm@school – a curriculum of critical health literacy for secondary school students: Results of a pilot study*. Unpublished manuscript.

Steimle, S. (1999). U.K.'s Tony Blair announces crusade to fight cancer. *Journal of the National Cancer Institute*, *91*, 1184–1185.

Stine, G.J. (1999). *AIDS update 1999: An annual overview of acquired immune deficiency syndrome*. Upper Saddle River, NJ: Prentice-Hall.

Stolberg, S.G. (2002, February 6). Study says clinical guides often hide ties of doctors. *The New York Times*, A17.

Street, R.L.J. (2001). Active patients as powerful communicators. In W.P. Robinson & H. Giles (Eds.), *The new handbook of language and social psychology* (pp. 541–560). New York: Wiley.

Tan, S.B., Goh, C., Thumboo, J., Che, W., Chowbay, B., & Cheung, Y.B. (2005). Risk perception is affected by modes of risk presentation among Singaporeans. *Annals of the Academy of Medicine*, *34*, 184–187.

Towle, A., & Godolphin, W. (1999). Framework for teaching and learning informed shared decision making. *British Medical Journal*, *319*, 766–771.

Trevena, L.J., Davey, H.M., Barratt, A., Butow, P., & Caldwell, P. (2006). A systematic review on communicating with patients

about evidence. *Journal of Evaluation in Clinical Practice*, *12*, 13–23.

Urologen im Test: Welchen Nutzen hat der PSA-Test? [Testing urologists: What are the benefits of a PSA test?]. (2004, February). *Stiftung Warentest*, pp. 86–89.

U.S. Preventive Services Task Force. (2002). *Guide to clinical preventive services: Report of the U.S. preventive services task force* (3rd ed.). Baltimore, MD: Williams & Wilkins.

Villanueva, P., Peiró, S., Librero, J., & Pereiró, I. (2003). Accuracy of pharmaceutical advertisments in medical journals. *The Lancet*, *361*, 27–32.

Voss, M. (2002). Checking the pulse: Midwestern reporters' opinions on their ability to report health care news. *American Journal of Public Health*, *92*, 1158–1160.

Wallsten, T.S., Budescu, D.V., Zwick, R., & Kemp, S.M. (1993). Preference and reasons for communicating probabilistic information in numerical or verbal terms. *Bulletin of the Psychonomic Society*, *31*, 135–138.

Warner, J.H. (1986). *The therapeutic perspective: Medical practice, knowledge, and identity in America, 1820–1885*. Cambridge, MA: Harvard University Press.

Wegwarth, O., & Gaissmaier, W. (2008, July). *The delusive evidence of survival rates*. Paper presented at the XXIX International Congress of Psychology, Berlin, Germany.

Welch, H.G. (2004). *Should I be tested for cancer?* Berkeley: University of California Press.

Welch, H.G., Schwartz, L.M., & Woloshin, S. (2000). Are increasing 5-year survival rates evidence of success against cancer? *JAMA: The Journal of the American Medical Association*, *283*, 2975–2978.

Welch, H.G., Schwartz, L.M., & Woloshin, S. (2007, January 2). What's making us sick is an epidemic of diagnoses. *New York Times*. Retrieved July 27, 2008, from http://query.nytimes.com/gst/fullpage.html?res=9C01EED71630F931A35752C0A9619C8B63&scp=1&sq=What%27s+making+us+sick+&st=nyt

Welch, H.G., Woloshin, S., Schwartz, L.M., Gordis, L., Gøtzsche, P.C., Harris, R. et al. (2007). Overstating the evidence for lung cancer screening: The International Early Lung Cancer Action Program (I-ELCAP) Study. *Archives of Internal Medicine*, *167*, 2289–2295.

Wells, H.G. (1994). *World brain*. London: Cambridge University Press. (Original work published in 1938)

Wolf, A., & Schorling, J.B. (2000). Does informed consent alter elderly patients' preferences for colorectal cancer screening? *Journal of General Internal Medicine*, *15*, 24–30.

Woloshin, S., & Schwartz, L.M. (1999). How can we help people make sense of medical data? *Effective Clinical Practice*, *2*, 176–183.

Woloshin, S., & Schwartz, L.M. (2002). Press releases: Translating research into news. *JAMA: The Journal of the American Medical Association*, *287*, 2856–2858.

Woloshin, S., & Schwartz, L.M. (2006a). Giving legs to restless legs: A case study of how the media helps make people sick. *PLoS Medicine*, *3*, e170 DOI: 110.1371/journal.pmed.0030170

Woloshin, S., & Schwartz, L.M. (2006b). Media reporting of research presented at scientific meetings: More caution needed. *The Medical Journal of Australia*, *184*, 576–580.

Woloshin, S., Schwartz, L.M., Black, W.C., & Welch, H.G. (1999). Women's perceptions of breast cancer risk: How you ask matters. *Medical Decision Making*, *19*, 221–229.

Woloshin, S., Schwartz, L.M., Tremmel, J., & Welch, H. (2001). Direct to consumer drug advertisements for prescription drugs: What are Americans being sold? *The Lancet*, *358*, 1141–1146.

Woloshin, S., Schwartz, L.M., & Welch, H.G. (2004, April 28). The value of benefit data in direct-to-consumer drug ads. *Health Affairs* Web Exclusive, 234–245.

Woloshin, S., Schwartz, L.M., & Welch, H.G. (2008). The risk of death by age, sex, and smoking status in the United States: Putting health risks in context. *Journal of the National Cancer Institute*, *100*, 845–853.

Wong, N., & King, T. (2008). The cultural construction of risk understandings through illness narratives. *Journal of Consumer Research*, *34*, 579–594.

Young, J.M., Glasziou, P., & Ward, J.E. (2002). General practitioners' self rating of skills in evidence based medicine: A validation study. *British Medical Journal*, *324*, 950–951.

Young, S.D., & Oppenheimer, D.M. (2006). Different methods of presenting risk information and their influence on medication compliance intentions: Results of three studies. *Clinical Therapeutics*, *28*, 129–139.

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, *98*, 287–308.