

Micronumerosity

Eric Blair

14 June 2007

Today's subject is two studies of money and its relation to medical efficacy. Both found no relation, but this non-finding may be suspect.

The first, found via this blog¹, explains that doctors aren't nearly as effective as one would hope. His key citation is this early '80s study by RAND Corp² (PDF). [I actually posted much of the below as a comment on the blog, but the guy has on one of those 'I will post only comments that agree with me' filters on the blog.]

The second, found via the New York Times³, is this study of PA cardiac surgeons⁴ (PDF).

The RAND study followed several thousand people over almost a decade. It paid for the health care of some of them, and those people naturally visited the doctor (statistically) significantly more than the control group who had to pay for their own health care.

However, the study did not find a statistically significant difference in many standard measures of health, such as death rate, between those who received paid health care and those who purchased their own. We conclude, therefore, that public subsidies for healthcare is stupid and people should fend for themselves.

The NY Times interpretation of the PA study went along the same lines: it looked at what insurance companies were paying for health care at each of two dozen hospitals, and the success rate of those hospitals. Again, it found little difference in the death rates from one hospital to the next.

Catastrophic events Having spent a reasonable amount of time with both studies, I encourage the reader interested in the running gag that is US health care finance to look at both reports. That said, there's a problem with the statistics: The studies aren't very powerful.

Most major medical events, death included, are catastrophic events, in the formal sense that they happen very infrequently but are a big deal when they happen. Colloquial catastrophes, like earthquakes, are also catastrophic in the statistical sense. Doing

¹<http://www.blog.sethroberts.net/2007/06/10/the-twilight-of-expertise-part-2/>

²<http://www.rand.org/pubs/reports/R3055/>

³<http://www.nytimes.com/2007/06/14/health/14insure.html?ex=1339473600&en=3b83aad4ce8bdb79&ei=5090&partner=rssuserland&emc=rss>

⁴http://graphics8.nytimes.com/packages/pdf/business/20070614_INSURE.pdf

stats on catastrophic events is difficult, due to the difficulty of gathering enough observations. Our guest blogger from a few episodes ago⁵ has often pointed out to me that the ocean is very clumpy, and is therefore impossible to sample. Either you'll get the 99.9% dead spots where nothing is going on, or the 0.1% where a huge menagerie of critters are collectively following the currents and feeding off of each other.

[When I first arrived in DC, I was lounging in a coffee shop attempting to keep myself occupied, when a TV crew came in, looking for man-on-the-street interviews. The interviewer showed me a paper with a news release about terrorism futures, for about, oh, three seconds tops, which was long enough to read the single highlighted line about TERRORISM FUTURES. The videotape rolled, the mic was shoved in my face, and I was supposed to say, 'I am apalled!!!'. But they got an unlucky draw; instead I said something like, 'I just got my PhD from a major research institution—the department that had a hand in the development of markets like these, even—and these markets do a wonderful job of aggregating information. However, terrorist events themselves—not to be confused with ancillary events around terrorism—are catastrophic events, and so it's very difficult to aggregate information about them.' What else could I say? I didn't bother to see if I showed up on TV that night.]

Table four in the RAND study states that the range in deaths over the various treatments was from 0.9% to 1.1%, for a total of forty deaths in the entire study.

So bear with me here (or if you're wimpy, skip to the boldface difference below): assume a binomial distribution, and a true rate of death for those paying for health care of 1%. In order for us to reject the null hypothesis at a $p=0.01$ level, the 2000-member test group would need to see under ten deaths, or a death rate of under 0.5%. A halving of the death rate is what most of us would call 'miraculous'.

Now let's say we had 2 million people in the test group instead of two thousand. Then we could reject the null hypothesis at the 1% level if, instead of observing the 20,000 deaths expected, we observe only about 19,680—a death rate of 0.98%. A 0.02% drop in death rate is beginning to look like something that could actually happen, and it would be both socially and statistically significant at this scale.

The difference is what statisticians call *power*. How much true difference does there need to be between one group and another before the test is able to actually detect that difference? To give a physical metaphor, some people have crappy vision and can't distinguish letters from a distance; others have powerful vision and can easily tell the difference. With catastrophic events, we're trying to read very faint lines, and so we need to gather lots of data to reasonably detect them and say that the control's line is definitely different from the case's line.

The RAND test is low-power because one rate has to be fully half the other before it can state a difference with confidence. The PA data is also low-power for the same reason. I leave the stats as an exercise to the reader, but the number of surgeries is around 20,000, and the rate of various measures of success for various procedures range from 1.9% for death rate to 19% for readmission (reading from page one of the study). This means that the power will be better than RAND's, but still not incredibly good. The PA study includes no regressions or hypothesis tests, which one could argue that they were right to do.

⁵<http://fluff.info/blog/arch/00000214.htm>

Interpreting a low-power study So what are we to make of a study that gathered data, but did not gather enough data to state anything with statistical significance? Well, we're back to eyeballing it and using our intuition. For more money you get a longer hospital stay; people who get free health care see the doctor more; people like to overcharge insurance companies. These things make sense, so when we see that there is no evidence *against* these intuitive statements, then I suppose we can bolster our belief a bit. But then there are things that are counter to intuition, like how health care is irrelevant to human health, or additional services are completely useless, which seem a bit counterintuitive. [Got cancer? Walk it off.]

There seems to be a popular belief that studies are standalone events that conclusively prove or disprove; newspapers are happy to push this perception because 'a recent study proved a sexy fact' sounds a lot better than 'a recent study marginally raised our subjective belief regarding a certain sexy fact'. But that's what a single study does: it marginally raises or lowers our confidence in the truth of a statement. I've already discussed this extensively in the context of creationism⁶. A study that has low power certainly contributes less to the debate than a powerful study would, but it is still data that we can scrutinize, in conjunction with all the other data that we have about the issue.

⁶<http://fluff.info/blog/arch/00000159.htm>