CONSULTING MEDICAL TEXTS

A cancer diagnosis can spur one to research various treatment modalities, especially by investigating medical studies. The reliability of the studies, the worthiness of the results, and deciphering the data are not always straight forward. This chapter gives an overview of the nature of medical studies and how they are analyzed. Understanding this information will greatly help in conducting one's own research.

STUDY DESIGN

Within a study, there are many variables to consider, such as:

The design of the study

Studies are complex, requiring knowledge, skill, time, and effort. Defects in the quality of the study can call into question the reliability of the results.

The determination of what is being measured (the question or hypothesis)

What is the underlying purpose of the study? Which variables are to be measured (or not measured)? Is there a hypothesis that is being tested?

The choice of participants and how they are selected

That is the size of the study? Small cohorts (groups) are not as representative as large ones. What is the target population? Do the samples accurately reflect the entire population?

Ethical considerations

Is anyone put in danger by the study? Are all risks revealed to the participants? Are there any conflicts of interest, such as being paid or sponsored by the makers of a device, drug or product under study?

The possibility of being affected by other confounding factors

Are there other factors not being studied that could affect the outcome of the study? Have such factors been taken into account when quantifying the results?

The quality of the data

How and when are the data collected and analyzed, and by whom? What statistical tools are being used to verify accuracy? By its very nature, all random sampling has some degree of error. The question is, how much?

How the study is financed and run

Who is financing the study and what are the criteria? Who is taking on the burden of collecting the results and putting them into writing? What is the duration of the study? Is the data blind to those conducting the study?

The influence of biases, whether intentional or incidental

Is the purpose of the study to prove a predetermined result? Will those conducting the study benefit more from one result than another? Do publishing requirements of medical journals favor one type of study over another?

How the results are being presented

Is there a control or placebo group against which the results are compared? Is the study favorably reviewed by peer groups qualified to make critical observations? Are the conclusions verified by the data?

These considerations show that there is much more to doing research than just going to the internet, pulling up some studies, and jumping directly to the conclusions. One must understand the total design of the study and be cognizant of its strengths and weaknesses.

TYPES OF STUDIES

Informational reports

Virtually every proton therapy center keeps records of their results, often following the subjects for years. Documenting a specific group of patients with regard to some quantifiable outcome is more like a report than a carefully crafted study. That does not mean, however, that the data are not useful.

Prospective and longitudinal studies

These types of observational studies follow the subjects over a period of time to discover the factors for a specific outcome. They are watched to see what risk factors may enhance the likelihood of getting the studied condition. One example would be to follow smokers and non-smokers to see who gets lung cancer. For someone already diagnosed with cancer, an experimental study or clinical trial would be more suitable.

Cohort studies are best conducted over long periods of time, which makes them very expensive. Five years is a common milepost for surviving cancer, although some studies continue for decades. Prospective studies are affected by changes in the cohort, such as death from a different cause, dropping out of the study, or a change in lifestyle affecting the outcome. It is also common for the nature of the intervention itself to change over time. Cohort studies produce true incidence rates and relative risks.

Longitudinal studies involve two groups, one with the studied condition and one (the control group) without. At the end of a specific length of time, the groups are compared with regard to the specific event.

Confounding factors are those which pertain to both the object of the study and the end event. In studying overall survival of cancer, for example, age, smoking habits, gender or other factors can be confounders. These must be taken into consideration to modify the raw data.

Retrospective studies are performed a posteriori, looking back to the past for data. They are often criticized for being susceptible to confounding factors.

Experimental studies

Experimental studies are different from observational studies in that they engage in an active intervention with the subjects in an attempt to change the outcome of an existing condition. Such studies are commonly used to quantify results for proton therapy. The existing condition is the cancer; the intervention is the radiation. Participants are chosen or recruited in some way. Finding enough participants for studies is often a problem, sometimes forcing the cancelation of the study.

Clinical trials are the most recognizable experimental studies, in which two or more clinical treatments are compared, such as proton therapy and x-rays. Subjects are randomly assigned to each cohort. When neither the researcher nor the subject know which group the subject is in, it is called a double blind randomized clinical trial, considered the gold standard of medical studies.

Non-randomized studies

Without randomization or blinding, the quality of the data comes into question. That being said, this is the most frequent type of trial. When enough data is collected, the sheer volume becomes statistically relevant.

A single institution collecting data only from its own patients is liable to bias. For a sample to be generalizable to a larger population, it must properly represent that population. This is accomplished through diversity. Homogeneity is a problem. For example, proton therapy is less likely to be available to the uninsured and lower economic levels, and to under-represent ethnic diversity. For those who fit within that homogeneous cohort, however, the reported results may be relevant.

Observational studies can make associations, but cannot of themselves assess cause and effect. It is difficult to draw wider conclusions, such as causality, from a single non-randomized study. One must use caution if using such information to make clinical decisions.

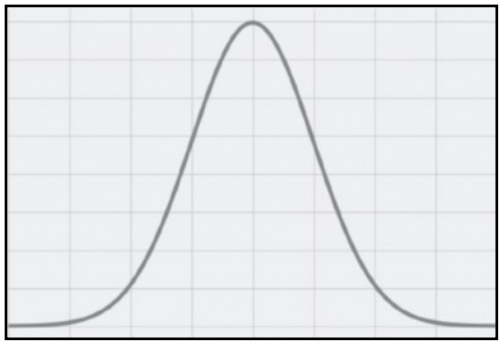
Additional study

For an in depth comprehensive and up-to-date online course pertaining to medical studies, see Cochrane Training: https://training.cochrane.org/handbook/current.

BASIC STATISTICAL ANALYSIS

In this section, you will find a brief overview of statistical concepts commonly used to quantify and analyze medical studies.

Bell curve



Most data are distributed with the greatest number of events in the middle and decreasing numbers in both outward directions. Any type of analysis should specify whether or not it assumes a normal distribution of data.

Parameters

Every test or study must designate the nature of what is being measured, the context, and the type of measurement, thereby setting the limits for the study. All parametric tests assume a bell curve or normal distribution of data. The narrower or more restrictive the limits, the more specific will be the result. A parameter can be seen as a characteristic element or factor defining that which is being studied. It is useful in evaluating the performance, status or condition of something.

Comparisons

When comparing two values, four results are possible. The symbols for these comparisons are shown below.

Less than: < (pointing to the left)
Equal to or less than: ≤
Equal to or greater than: ≥ (pointing to the right)
Greater than: >

A group of people divided into two cohorts, one of which is less than 65 years of age and the other 65 or older, could be summarized as follows: <65 vs ≥65.

Mean

An average expresses in a single number the predominant value for a set of data. In most cases, the average means the mean. The mean is the sum of all the values divided by the number of values. If the values are 2, 4, 6, 8, 10, dividing the total (30) by the number of values (5) gives a mean of 6.

Median

The median is the middle value in a series of values arranged in ascending order. In the example of 2, 4, 6, 8, 10 the median is the same as the mean (6, the middle number). Half of the values are higher and half lower than the median. Whereas the mean takes into account the *value* of all the numbers, the median takes into account the *position* of the numbers.

Suppose a study wishes to quantify the ages of a group of 11 people. Those ages are 2, 5, 7, 10, 14, 25, 29, 42, 55, 68, and 70. The sum is 327. The mean would be 327/11 = 29.73. The median would be 25. (If there is an even number of values the median is the average of the middle two.) Since this is a normal distribution, the mean and median are fairly close. Now, let us add one more age to the list: 98. The mean is now 425/12 = 35.41 whereas the median is 27 (the average of 25 and 29).

In the above case, the median better represents the whole population. Another example would be representing a group of salaries. Perhaps the mean is $50,000 a year. But if you add a billionaire to the mix, the mean could be in the millions of dollars whereas the median would be relatively close to the $50,000. Again, the median better represents the whole population.

Percentiles

Percentages are part of our daily life. A real estate sales commission might be 6%. Our tax bracket might be 20%. As a statistical value, a *percentile* is a point along a continuum of data divided into 100 units (100%). The data must be arranged in ascending order. As with the median, the percentile is a location marker. The 10th percentile would indicate 9 of the 100 values are less than that point and 90 are higher. The 30th percentile would indicate 29 of the 100 values are less and 70 are greater. And so forth. Percentiles are the same as percentage.

Per-*cent*-ile signifies 100 units. There are also *deciles* in which the data is divided into ten divisions. Similarly, *quintiles* sub-divide the data into five equal groups.

n

A lower case *n* stands for any positive integer indicating the number of values. If you are counting eggs in a dozen, n = 12. In medical studies, *n* typically denotes the number of people in a particular cohort. Suppose a study sought to compare the amount of exercise between men and women over the age of 60. The cohort might be described as follows: Men (n = 50) and women (n = 47) all > 60 years of age. The two different values for *n* indicate there were 50 men and 47 women being studied.

The value for *n* is important. The more participants in a study, the more reliable the results. A study with *n = 5* would be much less significant than one in which *n = 275*.

Mode

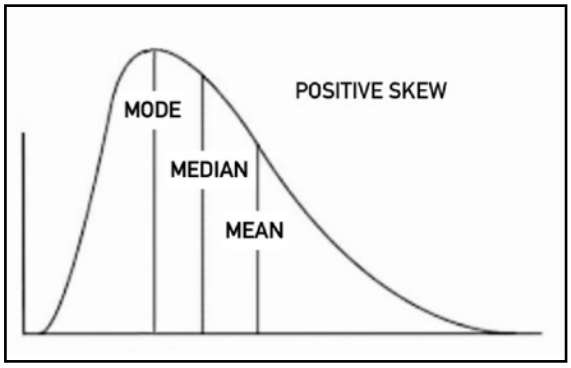
The mode is simply the highest value in a group of data. A common example is a bar graph illustrating the eye color of a certain population. The bar for brown eyes is typically

much higher than the rest. So it is the mode. It is possible to have two modes, or even more, depending on the interpretation.

This term is a bit confusing because many electronic devices have a mode button that changes the method of operation. One might speak of a camera set in automatic mode. This is a completely different use and meaning for the word *mode*. In statistics, the mode is simply the highest data point (see below).

Skewness

Nonparametric tests are used to describe data that are skewed or not normally distributed. Just as a bell curve gives an image of normal distribution, *skewness* portrays an irregular non-symmetrical distribution of values. Skewed with more data to the right is "positive" and to the left, "negative." The mode, median, and mean are no longer all in the center of the graph.



Measuring spread

Mean, median, and mode are simple and easy to understand, but do not reflect all there is to know about the data. Two very disparate sets of data could have the same mean. In one case, the data could congregate close to the mean, whereas in the other, it could be a mile wide. Measurements of spread tell us about the dispersion of the data, certainty, and variance.

Standard deviation

The standard deviation indicates how far the data are from the mean. There are three main standard deviations, each representing a specific percentage of the total data. The greater the standard deviation (SD), the further away from the mean the data are located. Being further away is relevant because the mean is less representative of the whole.

As shown below, one SD (represented by the lowercase Greek letter sigma) includes 68.27% of the bell curve (often rounded to 68%). Two SD include 95.45% (often rounded to 95.5%) and three SD include 99.73% (often rounded to 99.7%). Only three values in a thousand lie outside of three SD.

It is possible to calculate the variance of each individual data point from the mean to calculate a specific number value for one SD. While the percentage of data represented by a SD remains the same, the actual value represented by the SD can vary greatly.

For example, studying age within a given population, suppose one SD has a calculated value of plus or minus 7 years,. If the mean age is 43, then the range of ages would be as follows:

1 SD = (43-7) - (43+7) = 36 - 50. The spread is the mean of 43 plus or minus 7.

2SD = (43-14) - (43+14) = 29 - 57. Here, the spread is 43 plus or minus 14.

3SD = (43-21) - (43+21) = 22 - 64. Here, the spread is 43 plus or minus 21.

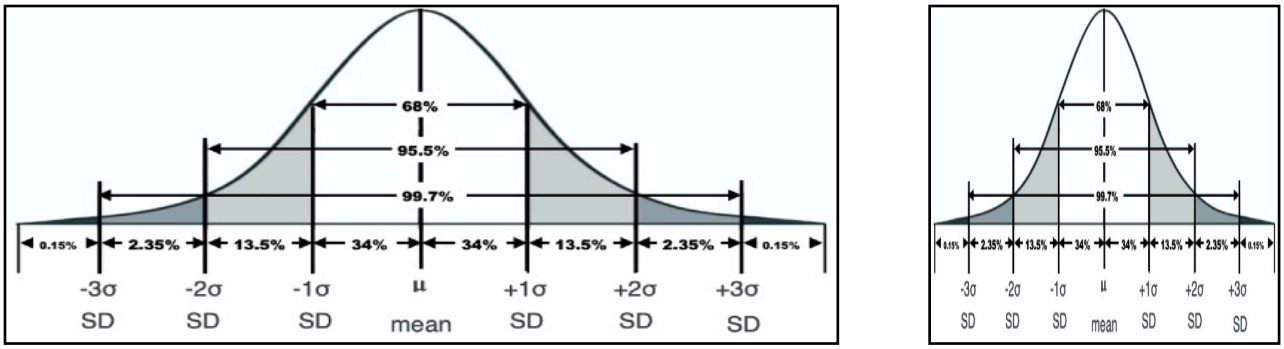
On the other hand, if the SD equalled only 1 year rather than seven, the range would be quite different:

1 SD = (43-1) - (43+1) = 42 - 44. The spread is the mean of 43 plus or minus 1.

2SD = (43-2) - (43+2) = 41 - 45. Here, the spread is 43 plus or minus 2.

3SD = (43-3) - (43+3) = 40 - 46. Here, the spread is 43 plus or minus 3.

The second case, with a spread of only 6 years within 3 SD, represents a very tight set of data, very close to the mean. A small value for SD gives a greater degree of confidence in the statistical conclusions. Below, the same graph can be used to illustrate spread-out data (left, SD = 7 years) or more dense data (right, SD = 1 year). .



Probability

Probability, abbreviated with the lowercase letter *p*, indicates how likely a hypothesis is to be true. This is one of the most used qualifiers for medical studies and is the first thing to look for. Somewhat confusing is the fact that it gives the probability for what is known as the *null hypothesis*, namely, that there is *no* difference between the results from two cohorts (the intervention group and the control group, for example) or an event and random chance. Such a state of affairs (no difference) is not the result researchers are looking for. They want a large and significant difference between the groups. Therefore, they want the probability for the null hypothesis to be as low as possible.

The *p* value is really a percentage. A value of $p = 0.50$ means there is a 50/50 probability (50%). If $p = 0.05$, then the null hypothesis is true in only 5% of the cases. This is considered the threshold value for statistical significance. It means the study result is 95% likely to be true (wrong only 5%). A value of $p = 0.01$ is considered highly significant and $p = 0.001$ very highly significant (happening by chance only once in 1,000 events or one tenth of one percent).

A value of $p = 0.20$, while not considered significant, might still be suggestive of a meaningful relationship, thereby leading to further study.

Statistical significance does not automatically indicate clinical importance or viability. Statistics are a useful tool for analyzing data and outcomes, whereas clinical significance is rooted in the impact on clinical practice. Statistics are mathematical constructs whereas clinical practices affect real people.

Confidence interval

The purpose of statistical analysis is to determine from a sample or a study the true value as it would apply to the whole population. A study of 100 subjects is meant to generalize to the public as a whole. But samples by their nature contain a degree of uncertainty. Rather than depending on a single number, such as a mean or median, the confidence interval (CI) gives a range within which there is a 95% probability that the true value for the entire population lies. (Specifically, the CI is plus or minus two standard errors from the mean, a computation too technical to include here.) The narrower the CI, the greater the confidence in the mean.

How are SD and CI different? Standard deviation serves as a description of the location (spread) of the values. The percentages given for SD are for the distribution of the values in relation to the mean. A narrower SD shows a distribution of values closer to the mean. It is about the data. The CI is considered an estimate that the true value for the whole population lies within that range. It pertains to the generalization of the study findings. It shows how accurately the sample is likely to represent the larger population.

The CI is related to the size of the study. The greater the value for *n* the narrower the CI, as it can more accurately estimate the true value from a larger sample.

The null hypothesis purports that there is no significant difference in the results being studied. Any CI that has a range including zero, such as -0.86 to +0.98, would indicate zero confidence in value for the true population mean. That would automatically conclude that the result is not significant. Significance is a refutation of the null hypothesis. Zero confidence is the same as the null hypothesis. On the other hand, a 95% CI that *does not include zero* in its range is the same as $p = 0.05$ and therefore statistically significant.

Risk, odds, and ratios

Risk and odds are not the same, although they are sometimes mistakenly considered synonymous. They measure different things.

Risk includes a time element, estimating whether or not, over a period of study, a certain risk event will happen. For example, a study of the risk for smokers getting lung cancer during the term of the study would be calculated by dividing the number of cases of lung cancer (say, 54) by the total number of subjects in the study, say, 100 ($n = 100$). Hence, risk = 54/100 or 0.54. Risk is essentially the same as the probability of occurrence of an event or outcome happening in the future. Risk will be a number between 0 and 1. It can be read as a percentage (54%).

Odds represent the chance of a single event happening as opposed to not happening. Rather than a measurement of how many events happen within an entire group over time, odds measure the likelihood of a single event at a specific moment of time. Consider an adverse effect that happens 40 times out of 100 (happens 40 times, doesn't happen 60 times) the odds would be how many times it happened (40) divided by how many times it didn't (60), hence

40/60 = 0.667, sometimes expressed as 667:1000 or 66.7%. (The risk would be 40 divided by *100* such that 40/100 = 0.40.) Odds can range from 0 to infinity.

The risks and odds from different cohorts can be compared by using a ratio. The risk ratio is also called the relative risk. The ratio is calculated by dividing the risk/odds of the subject group by the risk/odds of a second group (say, a control group). When the risk/odds values are equivalent, the ratio equals one (any number divided by the same number equals one). When one or other of the risk/odds is different, the ratio reflects the difference as follows.

>1.0 Risk/odds is increased in the subject group compared to the control group

<1.0 Risk/odds is decreased in the subject group compared to the subject group

For example, consider a study measuring whether obesity leads to high cholesterol. In the subject group, 62 events of high cholesterol happen out of 200 participants. The risk would be 62/200 = 31.0. This risk of higher cholesterol would be 31%. In the control group, 22 events happen in a group of 180. The risk would be 22/180 = 12.22. To compare the risk of the two groups, divide the event ratio for the subject group by that of the control group, such that 31.00/12.22 = 2.54. Being greater than 1.0, the risk ratio shows higher risk in the subject group. If there were more events in the control group (say, 42 out of 100) than in the subject group (say, 32 out of 100), then theratio would be less than 1.0 (32/42 = 0.76).

Suppose there is a study comparing an old cancer drug and a new cancer drug. The group receiving the old drug would be the control group, and the new drug the subject group. If they both had the same number of events (death), the drugs would be considered equal. If the new drug had fewer events, then the risk ratio would be lower (dividing a lower number by a larger number always resulting in a value under 1.0). If you are the manufacturer of the new drug (or the person taking it), you want the RR to be as low as possible.

The risk ratio is commonly used in meta-analyses that gather data from multiple studies.

The same procedure is used for comparing odds. You divide the odds of the subject group by the odds of the control group. As with the RR, results greater than 1.0 signify the odds in the event group are higher whereas results less than 1.0 denote lower odds compared to the control group.

To elaborate on the risk/odds ratios, the 95% confidence interval (CI) often gives the range in which the ratio is found. Here is a typical result for a study.

Women who had received RT (radiation therapy) for left-sided breast cancer had a higher risk of CV (cardiovascular) death than those who received RT for a right-sided breast cancer (RR: 1.12, 95% CI: 1.07-1.18, P < 0.001).

Interpreting the above statistics: The ratio is greater than 1.0, meaning the risk was greater for those who received radiotherapy for left-sided breast cancer. The CI shows the RR to be right in the middle of the range and does not include zero. Therefore we know the result is statistically significant. In fact, the probability of less than 0.001 means the result is very highly significant.

The odds ratio compares two odds at a point in time. The risk ratio (relative risk), however, is a ratio of two probabilities, representing risk over time.

SUMMARY

Cochrane, a British organization, has as its mission organizing medical research findings to enable the best evidence-based choices. It publishes studies that meet their high standards online in the Cochrane Library, holding them to be the most reliable and up-to-date evidence. The following quotation from Cochrane Reviews helps put statistics into perspective.

"A leap of faith is always required when applying any study findings to the population at large" or to a specific person. "In making that jump, one must always strike a balance between making justifiable broad generalizations and being too conservative in one's conclusions." In addition to issues about risk of bias and other domains determining the certainty of evidence, this leap of faith is related to how well the identified body of evidence matches the posed PICO (Population, Intervention, Comparator(s) and Outcome) question. As to the population, no individual can be entirely matched to the population included in research studies. At the time of decision, there will always be differences between the study population and the person or population to whom the evidence is applied; sometimes these differences are slight, sometimes large.
In particular, the following issues can help people make better informed decisions….

- *information on all important outcomes, including adverse outcomes;*
- *the certainty of the evidence for each of these outcomes, as it applies to specific populations and specific interventions; and*
- *clarification of the manner in which particular values and preferences may bear on the desirable and undesirable consequences of the intervention.*

This chapter has touched only the surface of how to read and interpret studies. Below is a quick summary of this chapter.

<u>Designs</u>

Informational: Surveys, gathering of information

Prospective and Longitudinal studies: Monitoring over time to determine risk factors

Experimental studies: Measuring the effect(s) of specific interventions. Includes clinical trials and randomized double-blind studies.

Non-randomized studies: Accumulation of data within a specific setting or population.

<u>Statistical concepts</u>

Bell curve (normal distribution) vs. skewed data

Comparisons:

< (less than)

> (greater than)

Mean: Average of all the data (sum divided by number of data points)

Median: The center value in a set of sequential data

Mode: The largest value

Percentiles: Location along a scale of 100 units

n: The number of members in a cohort (group)

Standard deviation: Division of data on a bell curve quantifying spread from the mean

1SD contains 68% of the data

2SD contains 95.5% of the data

3SD contains 99.7% of data

p: Probability. One of the first parameters to look for.

p = 0.05 or lower is significant (95% certainty)

p = 0.01 is highly significant (99% certainty)

p - 0:001 is very highly significant (99.9% certainty)

Confidence interval (CI): A range of values 95% certain to contain the mean for the true population (beyond the studied cohort). If the range does *not* contain zero, it is statistically significant (the same as *p = 0.05)*.

Risk*:* An observation over time quantifying number of risk events divided by sum of all subjects.

Odds: Chances of a specific event calculated by dividing favorable outcomes by unfavorable outcomes.

Risk ratios and odds ratios: Comparisons such that

1	Equivalency
>1.0	Risk/odds higher in the subject group than the control group
<1.0	Risk/odds less in the subject group than the control group

Suggested reading:

Medical Statistics Made Easy, M. Harris and G. Taylor (2004)

Medical Statistics from Scratch, David Bowers (2008)

CASE STUDY

There is a competition between photon technology (x-rays) and proton therapy. The following essay explains how a widely touted trial supposedly detrimental to proton therapy actually had just the opposite result. This is a tale of poor design and manipulated statistics.

The study comes from MD Anderson and Massachusetts General Hospital, both of which have proton centers. A MedPage Today headline screams: "NO REDUCED TOXICITY FOR PROTON TX vs. IMRT in NSCLC." (See MedPage Today.) Proton Tx means proton treatment. The talking points in the article are given as follows:

◆ The use of proton therapy to treat non-small cell lung cancer (NSCLC) does not result in decreased toxicity compared with intensity-modulated (photon) radiotherapy (IMRT).

◆ These findings challenge what is considered to be the major advantage of proton therapy — i.e., superior dose distribution that avoids healthy tissues and reduces toxicity.

Since both photons and protons kill cancer, the attention goes to how they do it. Protons stop when they reach the target and x-rays do not. That cannot be changed. It gives protons an advantage that is hard to challenge. This (single) study believes it has refuted the proton advantage. Busy doctors who just scan the headline while quickly perusing the literature are left with a conclusion which is not proven by the data. It is a case in which the study design and analysis had a major role in the misleading outcome.

Because the lungs are a moving target, this is one of the most challenging areas for radiology in general. It would seem like a perfect opportunity to compare these two modalities. The purpose of the study was not to determine the effectiveness in killing the cancer. Rather, it measured levels of toxicity, which is to say, damage to tissues. The study concedes that proton therapy exposes less heart tissue to radiation than x-rays, but that wasn't their focus.

Yet, what if it were? Suppose they were measuring heart exposure. Then the headline would have been "Proton Therapy Exposes Heart to Less Damage than IMRT." I think that would be extremely relevant to the patient.

Instead they were measuring radiation pneumonitis (RP), which is inflammation of the lung, and local failure (LF). After one year, the combined rates of toxicity were 17.4% for for IMRT and 21.1% for proton therapy. X-rays win, or so it would seem. This is all the critics needed to hear. Have any of them looked at the study? If they did, they would find the results are not only unconvincing, they are exactly the opposite. After appropriate adjustments were made to the study design at the midpoint of the study, protons heartily outperformed X-rays.

The proton technology used was the older PSPT (passively scattered proton therapy) technique. This utilizes a shaped beam. A more recent technology is pencil beam scanning (PBS), in which a series of dots are laid down in a sequence of rows and layers. The improvement in accuracy with PBS increased the number of cancers treatable by proton therapy from 20% to 80%. When this study took place, only PSPT was available.

Breaking down the results a little more, we find that LF slightly favored proton therapy, 10.5% vs. 10.9% for X-rays. So the difference was in the RP inflammation factor. Specifically, at one year protons had an RP rate of 10.5% vs. 6.5% for IMRT. This is where the design of the trial comes into play. IMRT has been around for many years, whereas proton therapy at the time of this study had been at MD Anderson for six years and was still developing. Whether intentionally or inadvertently, the design of the trial put protons at a considerable disadvantage, as will become apparent below.

Jeffrey Bradley, MD, director of the S. Lee Kling Center for Proton Therapy at Washington University School of Medicine in St. Louis, Missouri, responded to this study as follow.

"I was part of a team that reviewed their adverse events ... I went through every case that had an event -- either a failure or a pneumonitis.

"They began treating [contralateral lymph nodes] with PSPT (passively scattered proton therapy), and eventually they learned that protons were better when all of the target was on the same side of the chest. They started having some early events when they were trying to chase down lymph nodes on the opposite side of the chest. Those patients had some pneumonitis."

The IMRT procedure was to treat all of the lymph nodes from one side. Since the x-rays continue through the body, this is standard practice. The design of the trial was to have protons do the same thing, going through the chest to treat the lymph nodes on the other side. Presumably this was meant to treat protons and photons equally. But that is not how protons are used. Protons reach the target from the closest side, not across the chest. The unfavorable results for the protons came from using them inappropriately. Other factors described below elaborate on how detrimental this design was. Dr. Bradley:

"Their proton planning skills got better over time; the incidence of pneumonitis went to zero after trial midpoint, because they stopped trying to treat patients with contralateral lymph nodes where the proton dose was scattering to the opposite side of the chest."

Going beyond the target to treat the lymph nodes on the opposite side did not utilize the greatest advantage protons offer and in so doing was responsible for increased exposure to healthy tissue, causing RP. After they improved their methods for proton therapy by treating the lymph nodes from the closest side, the incidence of RP was reduced to zero. ZERO! That was the true result, not that protons did worse. The RP for IMRT stayed the same throughout the trial because it could not be improved. The second half of the trial, then, properly reflected the actual difference between X-rays and protons, which was an RP of zero for protons and 6.5% for IMRT. That clearly that shows that protons are superior to X-rays and *not* more toxic, as claimed.

An important factor in any study is how it is reported. Despite the fact that the second half of the test had a clear superiority for protons, the report included in the calculation the unfavorable results of the first half of the study, allowing the early statistics to drag down and dilute the results achieved later in the study. Even if obliged to report all of the data, the

significance of the second half could have been emphasized so as to counter the improper headline.

The poor study design and misleading statistics were not the only biases in the trial. Dr. Bradley continues:

"So by design, they tried to give the same radiation dose to both arms, but you've taken away an advantage of proton beam therapy, because allowing a higher dose with protons is one of the advantages of the technology.

"Let's say you had an ice hockey team, and you can't use your best skaters -- everybody on the team has to be able to skate at the same level. You know you're going to get a similar outcome. Almost all experts in proton radiation therapy feel like they handicapped themselves by that study design."

Proton therapy's star players had to stay on the bench. As if that weren't enough, Dr. Bradley reveals yet one more obstacle for proton therapy to overcome, a larger target.

"In addition, the targets were larger on the proton arm than they were on the IMRT arm. It was early in the proton experience and they were worried about missing the target so they enlarged their aperture. They ended up treating, circumferentially, about 8-10 mm wider around the tumor with the proton arm. We've learned not to do that."

An aperture is a device that shapes the radiation to conform to the target. This is used only for PSPT, as it radiates the target all at once. Pencil beam scanning works a different way. In this trial, the aperture for X-rays was smaller than the one for protons. That means protons were guaranteed to expose more tissue around the target, thereby increasing the probability of inflammation. And still, even then, protons out performed X-rays in the second half of the trial.

Since most proton therapy centers now have pencil beam scanning, the study was well obsolete at the time it was published (2018). Yet it continues to be quoted widely.

These were the obstacles to the best performance of proton therapy:

1. The study design had a faulty technique for protons for half of the test.

2. Passive double scattering technology was used, which at the time was the only choice but is now not a typical result.

3. The aperture was enlarged to expose more healthy tissue.

4. The dose rate was minimized to equal that of the photons.

The performance of the proton therapy was acknowledged in the statistics is this way. *Exploratory analysis showed that the RP and LF rates at 12 months for patients enrolled early versus after the trial midpoint were:*

	Before midpoint	After midpoint	
IMRT	*21.1%*	*18.2%*	*(about 10% less) p = 0.47*
PSPT	*31.0%*	*13.1%*	*(about 58% less) p = .027*

This relevant difference was buried in the statistics and not emphasized.

The results of trials should be peer reviewed. In this case, the review was done by Dr. Feng-Ming (Spring) Kong, MD, from the Indiana University School of Medicine (which does not have a proton therapy center). (See: https://ascopubs.org/doi/full/10.1200/

JCO.2017.76.5479) Her review was titled, *"What Happens When Proton Meets Randomization: Is There a Future for Proton Therapy?"*

The title is very suggestive, as if this one flawed trial would somehow imperil the future of proton therapy. *Dr. Kong states,*

"Completion of this study is not trivial because the evaluation of the benefit of a new technology rarely has been done during the century-long history of radiation oncology practice."

The invention of x-rays in the 19th century was a milestone in medical history. For the first time, it made it possible to see beyond the skin, into the body. In less than a year it was being used on the battlefield to locate shrapnel in the injured. Hospitals around the world clamored for x-rays. There were no demands for double blind studies or randomization. The benefits were obvious, just as they are today for proton therapy, for which critics demand more and more studies.

Dr. Kong addressed this subject.

"Some may even argue that conducting a trial to test the significance of such a treatment is unethical, like performing a randomized study to test the value of parachutes, because it will put patients at risk for unneeded radiation complications."

Exactly. Once proton therapy is shown to be superior, is it ethical to expose the opposing cohort to a lesser treatment? Dr. Kong again.

"Such beliefs have been reflected in the history of radiotherapy technology advancement. From the first uses of x-rays and radium for cancer treatment in the early 1900s, to kilovoltage (superficial) x-ray machines and the era of cobalt-60 and megavoltage two-dimensional treatment, to Linac-based three-dimensional conformal technology and the current widespread use of IMRT, technologies have been developed and implemented routinely in the clinic without randomized trials."

Routinely implemented without randomized trials. But now, Dr. Kong feels they are necessary when it comes to proton therapy.

"Comparative clinical outcome data are needed for patients and their families to choose a cancer treatment modality that is not readily available, for physicians to make treatment recommendations, for investors/industry to determine where to spend resources, for insurance companies and government to make reimbursement policies, and for researchers to know how and where to focus their efforts. Thus, a randomized trial is needed to generate unbiased evidence for this extremely costly technology."

Unbiased evidence. Great idea. Dr. Kong tangentially addresses some of the shortcomings of this study.

"Whether a better planning technique such as proton intensity modulation (IMPT) or pencil beam scanning (PBS) would have generated different results is hard to predict."

Actually, it isn't hard to predict at all. Many comparisons have shown the superiority of PBS. In fact, the head of this study, Zhongxing Liao, MD, herself conducted a study to compare double scattering and pencil beam scanning results for this same kind of lung cancer (see

https://mdanderson.elsevierpure.com/en/publications/toxicity-and-survival-after-intensity-modulated-proton-therapy-ve). There she made this conclusion:

IMPT (pencil beam scanning) is associated with lower radiation doses to the lung, heart, and esophagus, and lower rates of grade 3 or higher cardiopulmonary toxicity.

That finding took place *after* this study, but before it was published and reviewed by Dr. Kong. Predictability in statistics is determined in a way that indicates whether or not the results are significant. The value for *p* represents the likelihood that there is *no* significance, and that the result was random or meaningless. Therefore, the lower the value for *p*, the more significant the results. The threshold for significance is considered to be no greater than $p = 0.05$. Here are the results for the study above, comparing IMPT and PSPT. *Gy* represents a measurement of radiation.

IMPT (pencil beam scanning) delivered lower mean radiation doses to the lungs (PSPT 16.0 Gy versus IMPT 13.0 Gy, $p < 0.001$), heart (10.7 Gy versus 6.6 Gy, $p = 0.004$), and esophagus (27.4 Gy versus 21.8 Gy, $p = 0.005$). Consequently, the IMPT cohort had lower rates of grade 3 or higher pulmonary (17% versus 2%, $p = 0.005$) and cardiac (11% versus 0%, $p = 0.01$) toxicities. Six patients (7%) with PSPT and zero patients (0%) with IMPT experienced grade 4 or 5 toxicity. Lower rates of pulmonary (28% versus 3%, $p = 0.006$) and cardiac (14% versus 0%, $p = 0.05$) toxicities were observed in the IMPT cohort even after propensity score matching for baseline imbalances.

Look at the low rates of adverse effects for PBS: 2%, 0%, 0%, 3%, and 0%. These numbers do not make it hard to predict which modality is superior. The probabilities range from significant to very highly significant. It is clear that, in a study in which PSPT already out-performed photons, pencil beam scanning would have done far better yet. Still, despite all of the evidence to the contrary, Dr. Kong takes the traditional approach:

"Personally, as a radiation oncologist, I would not recommend proton therapy for NSCLC outside a clinical trial setting until a clinical benefit is demonstrated in a prospective randomized study."

But read between the lines. She wants "a prospective randomized study." That means this one doesn't fit the bill. She wants a better one. Here's why.

"In contrast to the largest retrospective study of patients from the National Cancer Database, this prospective randomized study failed to prove superiority of proton therapy."

Once the study was properly designed, protons *far* out performed photons. They proved extremely superior. Further, Dr. Kong admits that the above trial contradicts the conclusion of a large retrospective study. Let's take a look. The study was for the same kind of lung cancer.

The National Cancer Database was queried to capture patients with stage I-IV NSCLC treated with thoracic radiation from 2004 to 2012. A total of 243,822 patients (photon radiation therapy: 243,474; proton radiation therapy: 348) were included in the analysis.

On multivariate analysis of all patients, non-proton therapy was associated with significantly worse survival compared with proton therapy (hazard ratio 1.21 [95% CI 1.06-1.39]; P<.01). On propensity matched analysis, proton radiation therapy (n=309) was

associated with better 5-year overall survival compared with non-proton radiation therapy (n=1549), 22% versus 16% (P=.025). For stage II and III patients, non-proton radiation therapy was associated with worse survival compared with proton radiation therapy (hazard ratio 1.35 [95% CI 1.10-1.64], P<.01).

True, the National Cancer Database did not specifically study RP. I think to the patient, survival is a more important parameter. Non-proton radiation had worse survival rates. Well now, why do the critics quote one flawed study about inflammation instead of this much broader perspective about survival? X-rays had almost a quarter of a million chances to outperform the 348 proton patients, but they didn't do it. And again, this was with the less effective passive scattering technology.

Dr. Kong admits that the study may have been skewed against proton therapy. For example:

"Finally, the study design in terms of end point definition, control of confounding factors, and dealing with the lung dosimetric restriction may have confounded the results."

Dosimetric restriction refers to equal doses, whereas normally proton therapy would have a higher dose. This reflects what Dr. Bradley said about eliminating the best players.

"More importantly, the study required patients to meet dosimetric limits for both the PSPT and the IMRT arms, which may have resulted in not being able to enroll patients who would most likely benefit from protons."

Here is still one more disadvantage. They eliminated the best prospects for proton therapy, but not the best ones for IMRT. Disappointingly, after all of these qualifications, admissions, and misgivings, Dr. Kong accepts the skewed results. Then she gets political.

"Although negative results from a phase II study in NSCLC cannot exclude the potential benefit of proton therapy in other clinical situations, such as for pediatric patients, and the cost of proton therapy will be significantly reduced by newer technological changes, this trial should at least cause some pause in hospitals that are building these facilities for competitive reasons and not for cost-effectiveness reasons."

Are we to believe a cancer center planning to spend fifty to a hundred million dollars for a proton therapy center would change their mind because of this single flawed study? Dr. Kong seems to be encouraging such a conclusion. Sadly, she is right. Several planned proton therapy centers (Baton Rouge, New Orleans, Cleveland) have been cancelled or reduced in size based on studies such as this with faulty conclusions and exaggerated headlines. In the end, Dr. Kong gives a nod to protons.

"With the availability of more gantry angles, better imaging guidance, more-accurate dose computation for moving lung cancer targets and low-density of lung tissue, and more-advanced treatment planning technology like pencil beam scanning, we can generate remarkably better plans at every tumor dose level that lead to meaningful benefits for many patients in the clinic and are proven as cost-effective treatment in some specific disease settings."

“Cost effective” is a term used by purveyors of photon technologies to imply that even though protons are superior, the cost is too high to justify. In her review, Dr. Kong seems to be straddling the fence, giving weight to the misleading results but also giving a nod to proton therapy. She admits to some of the faults in the design study while also suggesting ways in which protons can be superior (pencil beam scanning, better dose control, etc.) Why is it that she is not willing to take a stand?

Both Dr. Liao and Dr. Kong were paid for their work by Varian, one of the world’s leading manufacturers of X-ray equipment that also sells proton therapy equipment. She also may have hesitated to openly criticize those who conducted the study, although the shortcomings are implied in this statement.

"The randomized trial should only include patients for whom the use of protons provides a better dosimetric plan. Such a randomized trial will identify patients with proven dosimetric superiority from proton planning to demonstrate whether such a dosimetric advantage can be translated into clinical benefit. Another possible advantage of protons to investigate is that by delivering less dose to much of the body (if achievable), protons may decrease radiation-mediated immune suppression and thereby improve survival in patients with NSCLC."

Note the qualification (“if achievable”) as if there is any doubt when virtually everyone knows protons deliver less dose to healthy tissue. Further investigation shows that Dr. Liao is highly respected and prolific at obtaining multi-million dollar grants for studying proton therapy. Here is an example from a grant proposal.

“This could be a practice-changing trial. Our results may provide level I evidence of the benefits of proton therapy and shape clinical guidelines.” The grant, for $8,734,885, was to show whether the unique dosimetric characteristics of proton beams would reduce dose and volume of the normal lung to radiation, hence further decreasing treatment-related lung toxicity. Designated as RTOG-1308, this level III randomized study sought 330 participants. It shows how costly such studies are. It may also explain why the (questionable) results for this study were so dramatically presented, as if to justify the expense.

In a short video previewing her presentation at the 2017 Multidisciplinary Thoracic Cancer Symposium of the American Society of Clinic Oncologists (ASCO), Dr. Liao mentions the above trial and again repeats the erroneous conclusion that it did not prove proton therapy superior. However, she said secondary findings show that proton therapy is minimally invasive, protective of lung and heart function, protects bone marrow, and allows both increased dose and survival.

That sounds like a win for proton therapy.