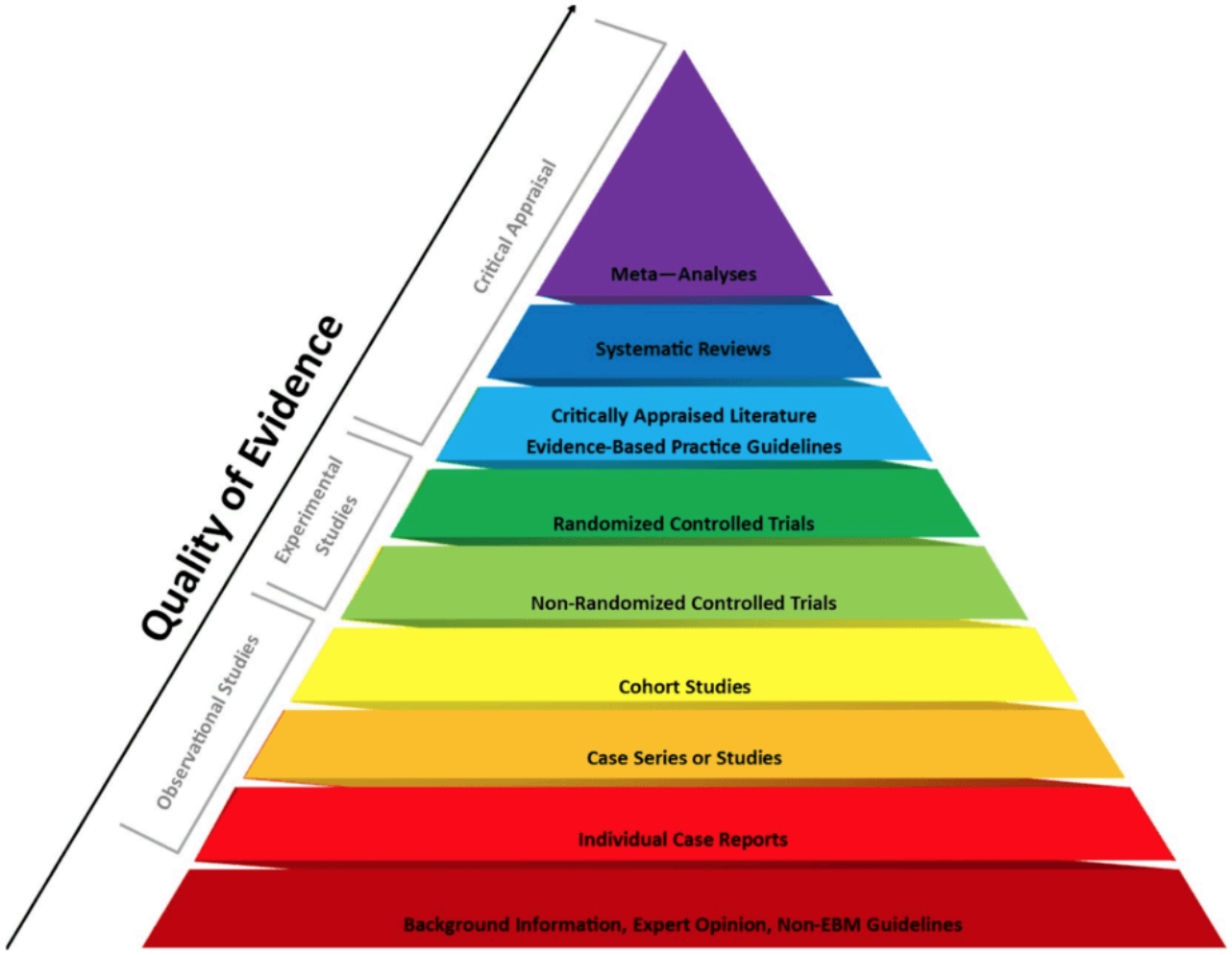
#188 - AMA #30: How to Read and Understand Scientific Studies

PA peterattiamd.com/ama30

Peter Attia December 20, 2021



In this “Ask Me Anything” (AMA) episode, Peter and Bob dive deep into all things related to studying studies to help one sift through all the noise to find the signal. They define the various types of studies, how a study progresses from idea to execution, and how to identify study strengths and limitations. They explain how clinical trials work, as well as the potential for bias and common pitfalls to watch out for. They dig into key factors that contribute to the rigor (or lack thereof) of an experiment, and they discuss how to measure effect size, differentiate relative risk from absolute risk, and what it really means when a study is statistically significant. Finally, Peter lays out his personal process when reading through scientific papers.

Below is a sneak peek of the episode. If you’re a member, you can listen to this full episode on your private RSS feed or on our website . If you are not a member, subscribe now to gain immediate access to this full episode, our entire back catalog of AMA episodes, plus more great benefits!

We discuss:

- The ever changing landscape of scientific literature [2:15];
- The process for a study to progress from idea to design to execution [4:15];
- The various types of studies and how they differ [7:30];
- The different phases of a clinical trial [19:15];
- Observational studies and the potential for bias [26:30];
- Experimental studies: Randomization, blinding, and other factors that make or break a study [44:00];
- Power, p-values, and statistical significance [56:15];
- Measuring effect size: Relative risk vs. absolute risk, hazard ratios, and “Number Needed to Treat” [1:07:45];
- How to interpret confidence intervals [1:17:30];
- Why a study might be stopped before its completion [1:23:45];
- Why only a fraction of studies are ever published and how to combat publication bias [1:31:30];
- Why certain journals are more respected than others [1:40:30];
- Peter’s process when reading a scientific paper [1:43:45]; and
- More.

§
How to Read and Understand Scientific Studies

§

Show Notes

§

The ever changing landscape of scientific literature [2:15]

- Peter was on The Tim Ferriss Show in June where the topic of studies came up
- Peter has also written a five part series on Studying Studies
- This podcast is designed to:
    - i) help people make sense of the “*ever changing landscape of scientific literature and how to distinguish between the signal and the noise of the research news cycle*”; and
    - ii) be a primer for people to really understand the process of scientific experiments and everything from how studies are published and obviously what some of the limitations are

The process for a study to progress from idea to design to execution [4:15]

Broad steps:

- Hypothesis: The default position in an experiment is that there is no relationship between two phenomena. This is called the null (i.e., zero) hypothesis.
- Experimental design
- power analysis
- IRB
- primary and secondary outcomes, protocol, stats plan, and preregistration

Overview:

Step 1: Null hypothesis

- In theory, it should start with a hypothesis—"*Good science is generally hypothesis-driven*."
- Null hypothesis: Take the position that there is no relationship between two phenomena
- For instance, the hypothesis might be that drinking coffee makes your eyes turn darker
- Then it must be framed in a way that says the null hypothesis is that when you drink coffee, your eyes do not change in color in any way
- That would imply that the alternative of hypothesis is that when you drink coffee, your eyes do change color
- But there's nuance to this… because am I specifying what color it changes to? Does it get darker? Does it get lighter? Does it change to blue, green? Does it just get the darker shade
- To be able to formulate that cleanly is the first step here

Step 2: Conduct an experimental design

- How are you going to test that hypothesis?
  A really elegant way to test this is using a randomized controlled experiment and even better if it's possible to blind it
- Other design question would be:
  - How long should we make people drink coffee?
  - How frequently should they drink coffee?
  - How are we going to measure eye color?

Step 3: Power analysis

- A very important variable is how many subjects will you have
- That will depend on a number of things, including how many arms you will have in this study
- It comes down to doing something that's called a power analysis

Step 4: IRB approval

If this study involves human subjects or animal subjects, you will have to get something called an Institutional Review Board to approve the ethics of the study

Step 5: Determine your primary and secondary outcomes

In short…

- Get the protocol approved, develop a plan for statistics, and then pre-register the study
- And in parallel to this, you have to have funding

*Note: there are some studies that are not experimental where some of these steps are obviously skipped

The various types of studies and how they differ [7:30]

Three main categories of studies/papers:

- 1) observational studies
- 2) experimental studies
- 3) papers that analyze/review those studies

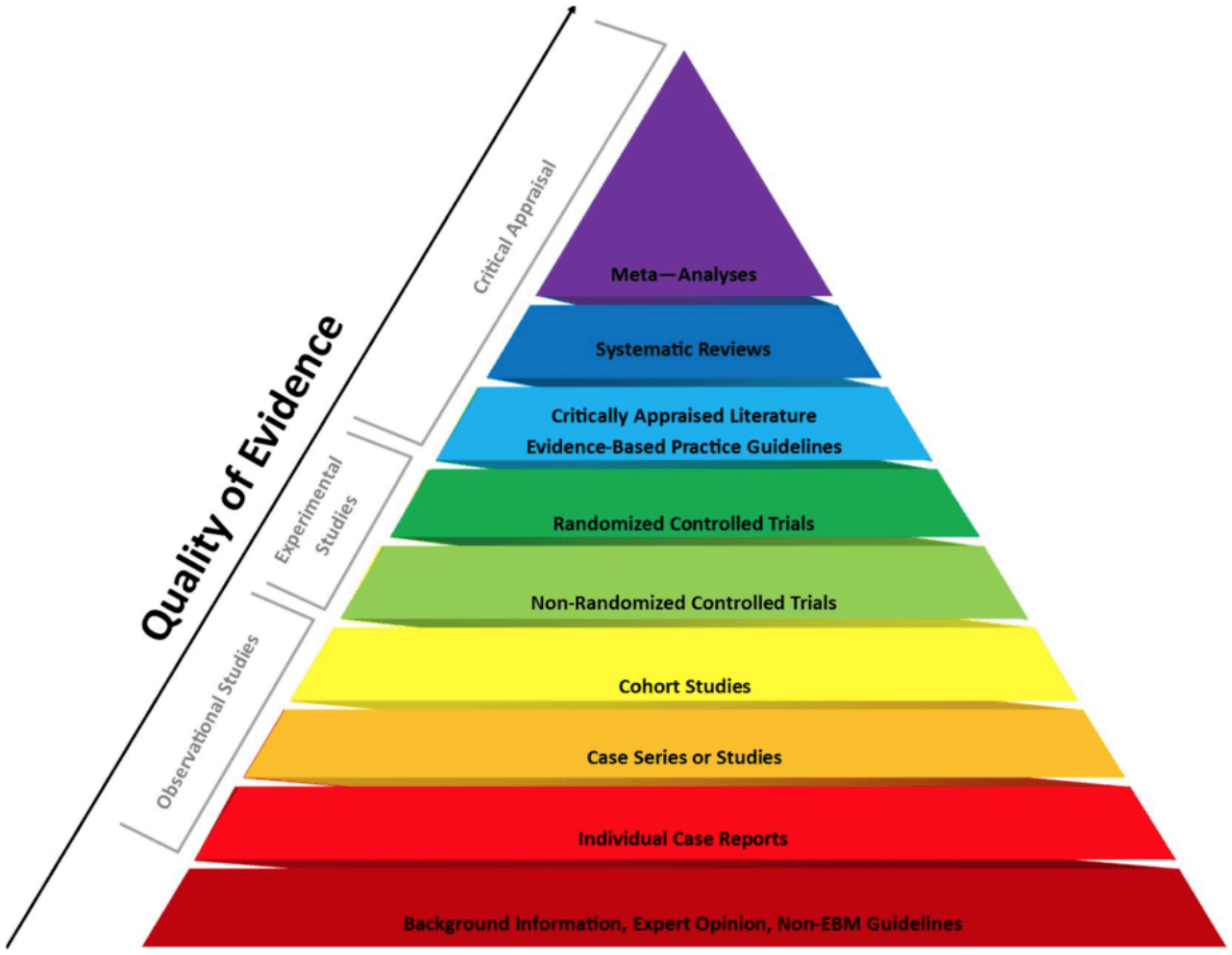


Figure 1. The evidence-based medicine pyramid. Credit: Purdue University

NOTE: One thing Peter doesn't like about the pyramid above is that it puts a hierarchy in place that suggests a meta analysis better than a randomized control trial, which is not necessarily true

Individual case report

- Example: Peter wrote one while he was at the NIH —
    - It was about a patient who had come into the clinic with metastatic melanoma and their calcium was dangerously high
    - The first assumption was that this patient had metastatic disease to their bone and that they were lysing bone and calcium was leaching into their bloodstream
    - That tuned out to not be the case — instead, they had something that had not been previously reported in patients with melanoma, which was, they had developed this parathyroid hormone related like hormone in response to their melanoma
    - This is a hormone that exists normally, but it doesn't exist in this format
    - So their cancer was causing them to have more of this hormone that was causing them to raise their calcium level
    - It was interesting because it had never been reported before in the literature, so Peter wrote up an individual case report
    - *What's the value in that?* Well, the next time a patient with melanoma shows up to clinic and their calcium is sky-high and someone goes to the literature to search for it, they'll see that report and it will hopefully save them time in getting to the diagnosis
- On the podcast with Steve Rosenberg, he emphasizes the importance of these types of observations
    - To start the process for a study—from idea, to design, to execution—you must have a hypothesis that begins with an observation
    - In other words, an interesting observation is hypothesis generating and it might kickstart a larger trial
    - That said, with one observation/case report, you can't make any statement about the frequency of this in the broader subset of patients or make any comment about any intervention that may or may not change the outcome of this

Case series or set of studies [11:15]

- Here you're basically doing the same thing, but you would look at more than one patient
- So for example, looking back at one's clinical practice and noticing that 27 patients over the last 40 years have demonstrated a very unusual finding
- For instance, one could write a paper that looks at all spontaneous regressions of cancer —this is incredibly rare, but there are certainly enough of them that one could write a case series.

Cohort studies [12:00]

- Cohort studies are larger studies and they can be **retrospective** or they can be **prospective**

- A *retrospective* observational cohort study would be looking backwards in time
    - For instance, go back and look at all the people who have used saunas for the last 10 years and look at how they're doing today relative to people who didn't use saunas over the last 10 years
    - It's observational, no intervention, and the hope when you do this is that you're going to see some sort of pattern
    - Undoubtedly, you will see a pattern, but the real question is, *will you be able to establish causality in that pattern?*
- A *prospective* cohort study would be looking forward in time
    - For instance, one would follow people over the next 5-10 years who use saunas and compare them to a similar number of people who don't
    - In a forward looking fashion, we're going to be examining the other behaviors of these people and ultimately what their outcomes are—*Do they have different rates of death, heart disease, cancer, Alzheimer's disease, other metrics of health that we might be interested in?*
    - Again, not intervening and not an experiment per se, just observing

Experimental studies [13:30]

- A **non-randomized** trial sometimes gets referred to as an open label trial
    - This is where you take two groups of people and you give one of them a treatment and you give the other one, either a placebo or a different treatment, but *you don't randomize them* (i.e., There's a reason that they're in that group)
    - For example, you may want to study the effect of a certain antibiotic on a person that comes in the ER
        - So you take all the people that come in who look a certain way (fever, white blood cell count, etc.) and you're going to give them the antibiotic
        - And the people who come in but they don't have those exact signs or symptoms, you do NOT give an antibiotic and you're just going to follow them
    - In summary, there are enormous limitation to non-randomized trials because presumably there's a reason you're making that decision. And that reason will undoubtedly introduce bias
- A **randomized** controlled trial is referred to as the "gold standard"
    This is where whatever question you want to study, you study it, but you attempt to take all bias out of it by randomly assigning people into the treatment groups
        "Blinding" is another important aspect that Peter will discuss later in the podcast

Meta-analysis [16:15]

- This is a statistical technique where you can combine data from multiple studies that are attempting to look at the same question
- Each study gets a relative weighting and the weighting of a study is a function of its precision (sample size, other events in the study)
- For instance, larger studies, which have smaller standard errors are given more weight than smaller studies with larger standard errors, for example.

- You'll know you're looking at a meta analysis because usually there will be a figure somewhere in there that will show across rows all of the studies.

- Let's say there's 10 studies included in the meta-analysis, and then they'll have the hazard ratios for each of the studies
- They'll represent them usually as little triangles, the triangle will represent the 95% confidence interval of what the hazard ratio is (basically a marker of the risk)
- You'll see all 10 studies and then they'll show you the final summation of them at the bottom, which of course you wouldn't be able to deduce looking at the figure, but it takes into account that mathematical waiting

The truth about meta-analyses

- On the surface, meta-analyses seem really, really great because if one trial, one randomized trial is good, 10 must be better
- But as James Yang once once said during a journal club about a meta analysis that was being presented, he said something to the effect of, "*1000 sows ears makes not a pearl necklace*" — just an eloquent way to say that garbage in garbage out.
- If you do a meta-analysis of a bunch of garbage studies, you get a garbage meta-analysis
- So a meta-analysis of great randomized control trials will produce a great meta-analysis.

The different phases of a clinical trial [19:15]

When talking about human clinical trials, the different "phases" is the phraseology is used by the FDA here in the United States

Origins of this process…

- Say you have an interesting idea—e.g.., a color cancer drug or molecule that you think will have some benefit
- Say you've done some interesting experiments in animals, maybe started with some mice and you went up to some rats and maybe even you've done something in primates
- Now you're really committed to this as the success of this and it's both safe and efficacious in animals
- You now decide you want to foray into the human space
- Step 1 is you have to file for something called an IND, an Investigational New Drug application that basically sets your intention of testing this as a drug in humans
- If you can the IND, you go to phase 1

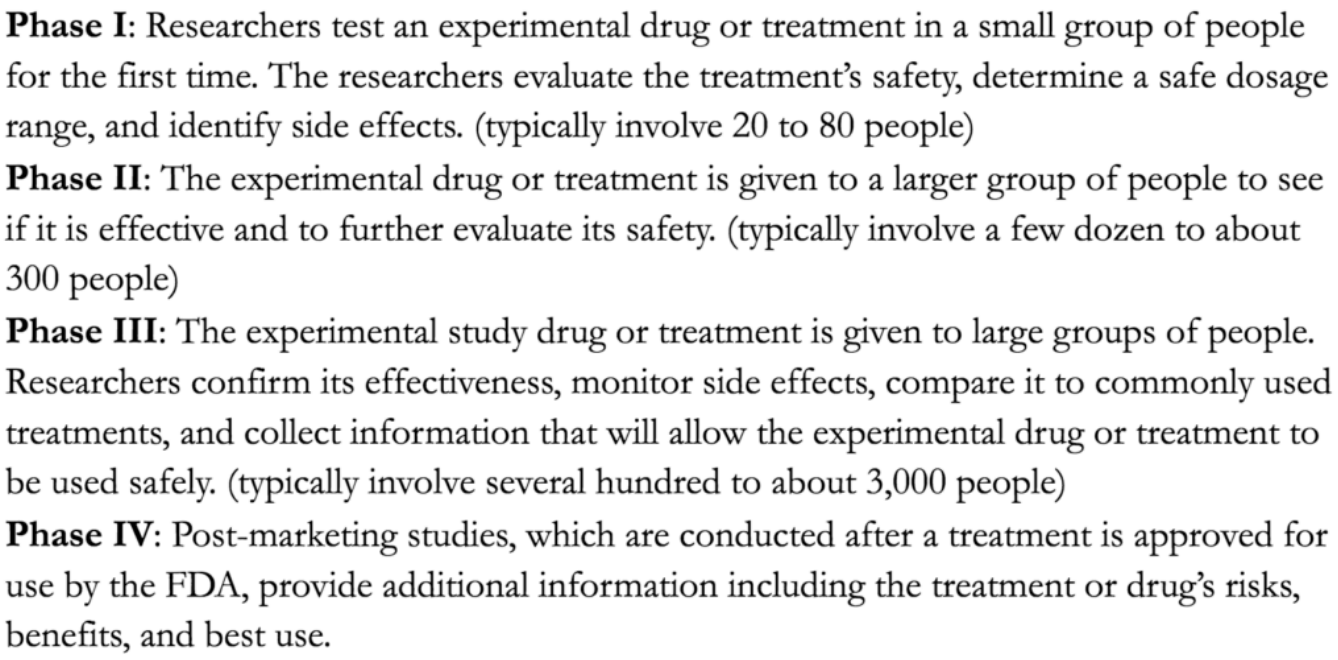


Figure 2. Different phases of a trial.

Phase 1

- Phase I is geared specifically to dose escalate this drug from a very, very low level to determine what the toxicity is across a range of doses that will hopefully have efficacy
- Very small studies, usually less than 100 people typically done in cohorts
- E.g., the first 12 people are going to be at 0.1 mg/kg, and assuming we see no adverse effects there, we'll go up to 0.15 mg/kg per kilogram for the next 12 people. If we have no issues there, we'll escalate it to 0.25 and so on
- Notice there is nothing in there about whether the drug works or not
- The people in the study are going to be patients that all have colon cancer—often metastatic colon cancer
    - They have progressed through all other standard treatments without much luck
- If the drug gets through phase I safely, then it goes to phase II

Phase 2

- The goal of phase II is to continue to evaluate for safety, but also to start to look for efficacy
- This is done in an open label fashion—they're not randomizing patients to one drug versus the other typically
- Usually it's at point where the investigators now we think that one or two doses are going to produce efficacy
- The dose levels were deemed safe in the phase I, so now we're now going to take patients and give them this drug and look for an effect
- A lot of times, if there's no control arm in this study, you're going to compare the drug to the natural history
- For instance, let's assume that we know that patients with metastatic colon cancer on standard of care, have a medium survival of X months

- Now we're going to give these patients this drug and see if that extends it anymore
- Typically these are very small studies — Can be in the 20-50 range, maybe up to a few hundred people
- There are lots of things that can introduce bias to a phase II if it does not have randomization
- The goal would be to still randomize in phase II, because you really do want to tease out efficacy
- So if a compound succeeds in phase II, which means it continues to show no significant adverse safety effects — note, no adverse events doesn't mean no side effects…it's just that it doesn't have side effects that are deemed unacceptable for the risk profile of the patient
- So if it shows efficacy without adverse events, you then proceed to phase III

Phase 3

- Phase III is a really rigorous trial
- It's typically a log step up in the number of patients to thousands of patients
- This is absolutely either a placebo-controlled trial or it can be standard of care versus standard of care plus this new agent
- It will be randomized and whenever possible, it is blinded
- And these are typically longer studies
- And because you have so much more sample size, you're going to potentially pick up side effects that weren't there in the first place
- Now you really have that gold standard for measuring efficacy
- And it's on the basis of the phase I, phase II and mostly phase III data that a drug will get approved or not approved for broad use, which leads to a fourth phase

Phase 4

- This is a post marketing study
- Phase IV studies take place after the drug has been approved, and they're used to basically get additional information because once a drug is approved, you now have more people taking it
- And they may also be using this to look at other indications for the drug
- In AMA #29, a phase IV trial with semaglutide was discussed
    - It was a trial that used semaglutide to look at obesity versus its original phase III trials, which we're looking at diabetes
    - Semaglutide is a drug that's already been approved — the trial is basically expanding the indication for semaglutide, in this case so that insurance companies would actually pay for it for a new indication
- In phase 4, you're also looking for whether there is another side effect that was missed in the phase 3 trial

Observational studies and the potential for bias [26:30]

Are there things you look for in an observational study that increase or decrease your confidence in it?

- **Selection bias** is a big one, says Peter
- The selection of subjects into a study, or their likelihood of staying in the study, leads to a result that is not representative of the target population. This can threaten RCTs as well.

Healthy user bias

- Healthy-user bias (a form of selection bias): People who are health-conscious are different from people who are not in myriad ways. [PAMD] This probably can't be controlled without randomized trials.
- Thinking about observational studies—prospective or retrospective—the healthy user bias is one of the more common ones we see in the epidemiology as it pertains to health
- For instance, *Is bacon bad for you?*
    - If you look at observational epidemiology, bacon is almost always bad for you
    - Hazard ratios are around 1.3, meaning it has about a 30% increase in the risk of basically anything you look at, whether it be cancer, heart disease, death, is that directionally
    - The WHO, for example, is looking at over 700 epidemiological studies for red meat consumption
    - When you look at those, you have to ask, *how are they measuring bacon consumption?*
        - They're using these food frequency questionnaires, which introduces **recall bias**
        - But if you look at the bacon data/red meat data on the surface, of course you'd be concerned
        - The problem with these studies is that you can't ever, no matter how much you try to statistically reconcile it, you can't strip out the fact that people make choices not in isolation
        - I.e., Is there any difference between a person who makes a lifelong decision to not eat meat and a person who doesn't? ⇒ Of course there is
        - It's going to come down to many things that go beyond their diet, including things that can't be controlled for
        - You can control for some of things, like smoking—a person who doesn't eat meat is far less likely to smoke than a person who does
        - Also, a person who doesn't eat meat is probably far more likely to exercise or pay attention to their sleep habits, or be more compliant with their medications, etc.
        - People who don't eat meat are basically a proxy for someone who is very, very health conscious
        - This healthy user bias permeates everywhere—it permeates in both directions

- If you look at the epidemiology that started to become very popular about 10 years ago, that was suggesting that diet soda was more fattening than soda—meaning drinking a diet Coke is worse than drinking a Coke
    - On the surface, that doesn't seem to make a lot of sense — diet Coke has no calories in it while coke is full of just liquid sugar
    - Then of course it gets you thinking, "Oh, is it the aspartame or whatever else?"
    - But a far simpler explanation: Look at people who are drinking diet soda versus people who are drinking soda — as a person is becoming more metabolically ill, and they're being informed that they really need to stop drinking soda, they're going to be drinking diet soda
    - So it's very difficult to look at just people drink this, people drink that and they're otherwise identical and simply the only difference between them is what they drink. — "*It just doesn't really hold up.*"

Information or recall bias

- Recall bias is a distortion in the measure of association due to a lack of accurate measurements. For example, errors in accuracy or completeness of the recollections "recalled" (recall bias) regarding events in the past. [PAMD]
- Many people are just shocked to learn how clunky nutritional epidemiology is, says Peter

Food frequency questionnaires

- For example, we have this thing called a food frequency questionnaire where "*you get a call from Billy and he asks you, 'Hey, do you remember how many times a week you ate oatmeal for the past year?*'

    Peter says, "I pay quite a bit of attention to what I eat. I don't know how I'd answer that question."
- The number of foods that are encapsulated in the food frequency questionnaire, vastly different. It only covers a very small portion of it
- And it's actually the foods that I think the epidemiologists often look at, like the red meat consumption and things like that, that people will underestimate when they do these validity studies and actually follow them
- Or they do a food log compared to the food frequency questionnaires—the correlation is so low that it's so underestimated that you're not really getting an accurate picture.

> "*I really feel strongly that we should abandon food frequency questionnaires, and no study should ever be published that includes them*

A more accurate questionnaire would be something like childbirth:

- If we were asking women to recall how many times have you been pregnant? How many times did you either have an abortion or miscarry? And how many times did you deliver it term?
- "Something that profound, yeah, I would feel confident that if you asked a woman that question over the past 10 years of her life, you would get very accurate answers." says Peter

- But, by the way, it still doesn't tell me that I would be able to infer causality if I was trying to look at women who have never had a miscarriage versus women who have had miscarriages
- Just because I look back and ask them to tell me those things doesn't mean that embedded within those differences or other biological or social or economic factors

> "*I think epidemiology has a place, but I think the pendulum has swung a little too far, and it has been asserted as being more valuable than I think it probably is*

Performance bias/Hawthorne effect

- *Performance biases: Hawthorne effect*: This is a type of bias where individuals change aspects of their behavior in response to knowing that they're being observed. Also a threat to RCTs.
- *Simple example*: Anyone who has tried to fastidiously log what they eat every day, just by logging what you eat, you will change what you eat
    - H*ow much more will you change your behavior when you know someone is going to look at it*?
        - Unbelievably so
        - In fact, you could make a case that one of the most efficacious dietary interventions known to man is having somebody watch what you eat every meal
        - Whether you have somebody virtually or literally watching you at every moment eat, especially someone who you're not entirely comfortable with, that's going to have an enormous impact

- There's a more sinister for of performance bias that creeps up in clinical trials
    - The risk of performance bias may be bigger in observational studies because in an RCT the intervention is generally more standardized and has more control over how the intervention is administered or the level of exposure.
    - But it can occur in RCTs, especially due to differences in treatment groups
        - With a RCT, you might think at the surface, "Wow, this is a really well done study."
        - So you'll take two groups, and let's say it's a weight loss trial
        - We're going to test calorie restriction versus pick your diet, the all potato diet.
        - So the calorie restricted group is given some leaflets, and it tells them how to measure calories, that they need to cut their calories by 25% from baseline. And we'll see you in 12 weeks
        - The potato diet group is given twice weekly counseling sessions on all the different ways you can cook potato so that you don't get fed up and bored of eating potatoes all day on the potato diet group
        - And at the end of the study, the potato diet group lost more weight than the calorie restricted group.
        - It'd be tempting to say, "Well, come on. This is a randomized control trial.
        - but the problem is there's an enormous performance bias in the potato group in that they were given far more attention. They were observed more. They were given more coaching. They had much more of a positive behavioral influence
        - If you're really designing a trial well, you have to flatten the curve on those differences
        - each person in each group should be getting the exact same amount of attention, the exact same amount of touch with the investigators, the exact same type of advice so that you can eliminate difference

> "*I would say that's the number one bias that I see in RCTs that are lifestyle-based, is that very subtle performance bias*

Bias due to primary and secondary outcomes

The idea of differentiating primary from secondary outcomes: Are the findings based on the primary outcome?

- there's some debate about whether you can only have one primary outcome or whether you can have co-primary outcomes, but the primary outcomes are basically the outcomes for which the study is designed around and powered against
- there are lots of secondary outcomes, and they're often exploratory
- It's really important that when people are preregistering studies, they state what the primary outcome is and what the secondary outcomes are
- And typically, a study that fails to meet its primary outcome will be deemed a null study, even fit meets secondary outcomes. So it's just very important to pay attention to the subtlety of that

- a good journal with a preregistered study is going to make that abundantly clear, "*but I can promise you that someone writing about it in the newspaper is virtually never going to make that distinction*"
- It's important to understand that because it gets to this next issue, which is kind of the multiple hypothesis testing problem—Research should be hypothesis seeking or hypothesis testing, but it can also be hypothesis generating
- And so you can use statistical tools to slice and dice data in multiple ways
- you can take many looks at data to see if you actually find something significant there
- You have to be careful because the more you look, the more times you look at something, the more likely you are to find something that is indeed positive
- *Simple example*: Flipping a coin
    - Say you flip a coin once, you've got a 50% chance of Heads
    - And if you get two chances to flip the coin, the probability that you're going to get heads is now 75%. If you get three chances to flip a coin, you're up to 87.5% chance that you're going to get at least one head
    - 10 times, you're basically at 100% likely that you're going to get heads
    - So if you're allowed 10 looks, you have to correct for that.
    - there's something in statistics called a Bonferroni Correction Factor that does force you to do that
    - It forces you to divide your P value by N where N is the number of times you've taken a look at the data, so to speak
    - And therefore, it raises the bar for what is significant

Confounding

How well did the study control for confounding?

- This goes back to the healthy-user bias and selection bias. Confounding occurs when a factor is associated with both the exposure (or treatment) and the outcome (e.g., disease or death), and is not part of the causal pathway from exposure to outcome.
- As a silly example: say that researchers observe that people who consume more hot chocolate are more prone to ski accidents. Does hot chocolate consumption cause ski accidents? Do ski accidents cause hot chocolate consumption? Maybe, but it's more likely that people who live in colder climates consume more hot chocolate, and most skiing occurs in places with colder climates, too (thus increasing the likelihood of ski accidents in these areas). Climate is a confound.
- As John Ioannidis argued in his conversation with Peter on the podcast, **it's really not possible to identify, let alone eliminate, all confounders**.

Experimental studies: Randomization, blinding, and other factors that make or break a study [44:00]

Things to think about:

(Some of these will apply to observational studies)

- Is there a control group? What kind of control (e.g., placebo, crossover)? How appropriate is the control? Crossover groups?
- Is it a randomized experiment?
- Is it blinded? Single? Double?
- How large (or small) is it?
- How long (or short) is it?
- How generalizable are the results? (Are the participants like me?)
- Single site or multi-site? Multisite RCTs are typically more generalizable.
- How big is the association or effect? Is it clinically meaningful?
- Adverse events?
- Who funded the trial?
- Do the authors have any conflicts of interest?
- Is the study adequately powered?

-Randomization

*The most important aspect of an experimental study is **randomization***

- If an experiment isn't randomized, it doesn't mean that it's useless, but it just means it's going to be a lot harder to really make sense of this
- Randomization needs to be a rigorous randomization, you can randomize incorrectly

⇒ *Famous example of an incorrect randomization is with the PREDIMED study*

- When the study was published was kind of a remarkable finding
- It was a very large study, something like 7,500 randomized into three groups, 2,500 per group, given basically two different dietary patterns
  - Mediterranean diet in two versions
  - a low fat diet
- This was a primary prevention study—it was looking at people who are high risk, but who haven't had heart attacks or anything yet, and it was looking at mortality
- The study was actually stopped early because it had such a positive effect— the Mediterranean diet had such a favorable effect relative to the low fat diet that people were dying at a rate far less such that it would've been unethical to continue the study

But then something happened with the PREDIMED study…

- They went back and reanalyzed this PREDIMED group and published that in 2018
- This all started with a guy named John Carlisle who did this analysis where he looked at thousands of studies and he could flag the studies and see, does this truly look like randomization based on some particular statistics?
- And the PREDIMED study was flagged, looking like this doesn't look like proper randomization might be something going on here

- If you did into the story behind the study, you find out more issues:
    - In the New York times, they talked to the lead or the senior investigator.
    - And he said that it turns out that some of the villages or the clinics, the investigators were randomizing the entire clinics to one group
    - Per the NY Times, some of the study participants in the villages complained that some neighbors were receiving free olive oil, while they got only nuts or inexpensive gifts. So the investigator decided to give everyone in each village the same diet. He never told the leaders of the study what he had done.

True randomization vs. cluster randomization

- Example: If you want to study the effects of meditation on attention span of kids…
- It's very different to
    - 1) Just take 100 kids and randomize 50 into one group, 50 into another and separate them; versus
    - 2) Two classes over here, two classes over here. We're going to split those two and two into the effect."
- That's a totally different type of randomization
- One is a true randomization
- One's a cluster randomization
- While you can do the latter (cluster randomization), it requires a different statistical adjustment
- PREDIMED basically had to reanalyze all of their data in light of that
- It turned out in the case of PREDIMED, the results still held, but it will always kind of be a cloud that hangs over it

To put the challenge of doing this PREDIMED study, think of this simple example:

- Imagine randomizing a household
    - "Dad, you're on a Mediterranean diet for the next seven years"
    - "Mom, you're on a low fat diet for the next seven years"
- You can see it starts to get very difficult

-Control groups

- You also want to make sure, is there a control group
- Not all prospective trials have control groups

Crossover

- Sometimes it's a single group where a person serves as their own control, and there's typically a crossover
- So you'll take a group
- You'll randomize them into two
- It's not that one group is getting treatment A and the other group is getting placebo or treatment B
- Both groups get both treatments plus or minus a placebo in different orders

- this is a great statistical tool provided the treatment doesn't interfere with the wash out
- The treatment doesn't interfere with the control session
- The reason this is powerful is you need far fewer subjects when everybody gets to serve as their own control
- So it greatly reduces basically the cost and logistics of a study, but you run into challenges
    - Say 20 people are going to this drug that is supposed to help them exercise better for eight weeks
    - another group is going to take a placebo for eight weeks and exercise
    - and then everybody switches
    - Do you need a gap between the tree treatments because will the effects of that drug linger into the placebo period for one group, which is not what's happening to the other group
    - And even if it is, even if you're only doing it with one group, are you confounding the effect of that treatment?
- The really good ones go A, B, and then B, A. They divide them into two groups and go A, B and B, A
- There's also the Paired T Test
    - The simplicity of the statistic of the Paired T Test is part of its elegance
    - it basically eliminates a lot of variants

-Blinding [51:30]

- So in an ideal world, both the subjects and the investigators should not know who is getting the treatment and who is getting the placebo
- At a minimum, the subjects should not know, this is single blinding
- Double blinding is always preferred if possible because the investigators can be biased towards the drug they are testing and could consciously or subconsciously treat the treatment group different
- This is very important and sometimes very challenging
- *Example*: In the podcast with Rick Doblin, he mentioned that one of the huge challenges of studying psychedelics is it's very difficult to blind anybody, most of all, the user, the subject

    One group is getting Psilocybin and the other group is getting, even if it's niacin which causes some flushing, it's not hard to know which group you're in, and that may affect the results

-Size: Size of study matter

-Duration: the length of the study matters

-Generalizability: *How generalizable are the results?*

- So *is it in a population that replicates or looks like what I'm interested in studying*?
- And there are strengths and weaknesses to mass heterogeneity of study.

- the more heterogeneous a study in terms of its patient population, well, the more generalizable the results are, but the higher the bar for finding it
- Extreme example:
    - for a while it was a relatively unknown kind of dirty little secret of medicine was how many clinical trials involved men only
    - How many drugs were approved for both men and women, but on the basis of only being studied in men
    - the rationale for this was that it was more complicated to study women
    - women, especially premenopausal women, because they have a menstrual cycle, that really changes things hormonally, and therefore, it's more complicated to do studies and look at drug kinetics and all sorts of things in women
    - And so the easier way to do that was to just study it in a homogeneous population of men.
    - Well, of course that poses an enormous problem if you're now trying to extrapolate the utility of that drug in women

-Single site or multi-site?

- Multisite RCTs are typically more generalizable
- The PREDIMED study is a great example of the challenges of a multi-site study
    - So you had a multi-site study, and there were probably significant differences between how the sites were run
    - There's an advantage to multi-sites because in theory it brings more heterogeneity. It should cancel out the effect of any one study over another
- But it's harder to control and therefore you can have, whether it be deliberately or non-deliberately rogue sites introducing more bias

-How big is the association or effect? … *Is it clinically meaningful?*

- You can have something that is statistically significant in that sense the study is "a success," but it's clinically irrelevant because the effect is not that big
- Example: Say you've tested a new drug for blood pressure
    - Turns out that it lowers systolic blood pressure by one millimeter of mercury after a year of use
    - That might be statistically significant if the study was large enough
    - But *is it clinically significant*? — Almost assuredly not.

–**Adverse events:** You want to pay attention to what the adverse events were, both in frequency, severity and distribution

-Who funded the trial:

- You may to pay attention to who funded the trial

- A lot of trials are funded by drug companies
    - Now, they're usually done with very clear data monitoring and data analytics
    - Despite all of the fear mongering out there, it's not like pharma really gets to put their hand on the scale of these pharma studies
    - What they can maybe do is get certain findings they don't like buried in supplemental journals

-Do the authors have any conflicts of interest?

- Even more important than who funded the trial is understanding what the conflicts of interest are of the authors
- Nowadays, those have to be declared
- However, there have been some very famous examples of people who were on editorial boards of journals and not declaring that, "*Hey, I'm a paid consultant of these 10 pharma companies and I'm writing or doing experiments on drugs by these people*," or, "*I'm an editor on journals that are commenting on this*."

-Is the study adequately powered? [56:15]

- you really want to understand if the study was adequately powered
- this becomes very important if the study has a null outcome

Power, p-values, and statistical significance [56:15]

> "*You really want to understand if the study was adequately powered, and this becomes very important if the study has a null outcome.*

⇒ *For more on this subject, check out: [Studying Studies: Part V – power and significance]*

What is power?

- **Power** is defined as 1-β where β is defined as the probability of a **false *negative***
- You can contrast that by talking about what a false *positive* is

What is p-value?

- A **false *positive*** is defined as α, and that's also known as the **p-value**
- everybody's heard of a p-value, but most don't think of it as a false positive rate
- Most people have not heard of the false negative rate being β and then 1-β being the power
- Peter's frequently talks about about p-values being 0.05 or less and that it's very difficult to make a case that we're going to look at a study that has a p-value of 0.1 and say it's significant

		What is true in the real world?	
		There is no effect (null = true)	**There is an effect (null = false)**
What conclusion is reached?	**No effect**	Correct conclusion ($p = 1 - \alpha$) [True Negative]	Type II error ($p = \beta$) [False Negative]
	An effect	Type I error ($p = \alpha$) [False Positive]	Correct conclusion ($p = 1 - \beta$) [True Positive]

Abbreviations: p: power; α: alpha or probability of a false positive result (i.e., Type I error); β: beta or probability of a false negative result (i.e., Type II error)

Figure 3. The four test outcomes. Credit: Studying Studies: Part V

So what does that mean?

The p-value is the probability that the effect you've seen is a false *positive* (i.e., it's not the true effect)

- For instance, you do a study and you're trying to determine if coffee changes eye color
- Let's say the results of the study appears to show that coffee DID make the eyes of the subjects darker
- If the p-value is 0.17, it means there's a 17% chance that this was a false *positive*

To restate this: a p-value is basically trying to answer the question, "*What's the probability of rejecting the null hypothesis when it is in fact true?*"

- The p-value is trying to answer the question: What is the probability—0 being that it is impossible and 1 being that it is certain—of rejecting the null hypothesis when it is, in fact, true? The p-value is the probability of obtaining a statistic as extreme or more extreme than the one observed in the experiment if the null hypothesis is true. [Benjamin et al., 2017]
- Obviously, we want p-values that are as small as possible—it can never be zero, but you want them to be as close to zero as possible
- Peter says 5% (0.05) is the maximum threshold
- *Remember, the default position is that the hypothesis is correct, that there is no difference between the groups

Statistical significance

- This term **statistical significance** basically means that the *null hypothesis is rejected* if the p-value is less than that pre-stated level (<0.05)

- **Statistical significance**: A result is deemed to be (or not to be) statistically significant based on a significance level that is preselected (i.e., before the study) by the investigators. This level is typically set at or below 5% (or <0.05). The significance level is also referred to as alpha, or α, where α is the probability of a false positive.
- If the p-value is less than 0.05, or 5%, the results are deemed "statistically significant" and the null hypothesis is rejected. If not, the results are deemed not statistically significant, and the null hypothesis is accepted

Why the 0.05 threshold?

- This guy who established this 0.05 level, Ronald Fisher, more or less said for the purposes of a single trial they're willing to accept a level of false positive in their results and still make that claim that they rejected that hypothesis.
- Because if you make the p-value so low, say for example "*No, my threshold is 0.000001*", then you really run the risk of discarding a lot of information that turns out to be relevant (i.e., false negative)

False negatives

- For the false negative rate, it's typical to allow it to be a larger number (between 10 and 20%)
- The flip side of that is we have 80 to 90% power because one minus accepted false negative rate is called your power
- This is one of the most important concepts to understand in designing any sort of clinical trial, whether it's humans, animals, any sort of intervention

Number of Patients in Each of Two Treatment Groups (Two-Sided Test)

Smaller Success Rate	*Larger Minus Smaller Success Rate*									
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.05	620*	206	113	74	54	42	33	27	23	19
	473†	159	88	58	43	33	27	22	18	16
0.10	956	285	146	92	64	48	38	30	25	21
	724	218	112	71	50	38	30	24	20	17
0.15	1250	354	174	106	73	53	41	33	26	22
	944	269	133	82	57	42	32	26	21	18
0.20	1502	411	197	118	79	57	44	34	27	22
	1132	313	151	91	62	45	34	27	22	18
0.25	1712	459	216	127	84	60	45	35	28	23
	1289	348	165	98	65	47	36	28	22	18
0.30	1880	495	230	134	88	62	46	36	28	22
	1414	375	175	103	68	48	36	28	22	18
0.35	2006	522	239	138	89	63	46	35	27	22
	1509	395	182	106	69	49	36	28	22	18
0.40	2090	537	244	139	89	62	45	34	26	21
	1571	407	186	107	69	48	36	27	21	17
0.45	2132	543	244	138	88	60	44	33	25	19
	1603	411	186	106	68	47	34	26	20	16
0.50	2132	537	239	134	84	57	41	30	23	17
	1603	407	182	103	65	45	32	24	18	14

* Upper figure: significance level 0.05, power 0.90.
† Lower figure: significance level 0.05, power 0.80.

[Cancer: Principles & Practice of Oncology 5th, 1999]

Figure 4. Power table.

- What this table is saying is you want to presuppose you know what the difference is between the treatment groups
- You have to say, "*I believe that the difference between the success rate in the treatment between group A and group B is going to be X percent, and the smaller of the two is Y percent*."

Let's come up with a scenario: "*I think that we are going to look at how this drug impacts your rate of being cured of a urinary tract infection*"

- I think that the placebo group is going to have a success rate of 25%
- I think that the treatment group is going to have a success of 35%
- So I think there's a 10% gap
- And I think the lower of those two is 25%.
    - So you go to 0.25 on the horizontal axis
    - you go to 0.1 over on the column
    - you'll see there's two numbers there, 459 and 358
    - And the upper of those two is if you want 90% power, i.e. 10% false negative
    - And the lower of those two is for 80% power or 20% false negative rate
    - And those numbers basically tell you how many people you need in each of the two treatment groups if you want to be significant at a level of 0.05%.
- W*hat do you notice when you look at this*?
    - notice that the bigger the gap, the bigger the effect size between the two groups, the fewer subjects you need
    - if you look from from left to right in this table, holding that effect size at 0.25, if you say, "*Well, the difference is 15%*," and now you only need 216 or 165
    - If the difference is 30%—so one group is going to have a 25% success rate one group's going to have a 55% success rate—you're down to needing 60 and 47
    - And if you go out to a 50% difference—so one group is going to have a 25% response rate, the other group's a 75% response rate—you're now down to needing somewhere between 18 and 23 people per arm
    - And by the way, if you go down to 5%—one group responds at 25%, the other at 30%—you're at 1700 or nearly 1300 depending on your level of power.

> "*It's the single most important table you should ever familiarize yourself with if you want to be in the business of designing clinical trials. . .because it is just so easy to get this wrong and over or under power an experiment.*

Over-and under-powered studies

Under-powered studies

- To under power experiment is the more common mistake
- You simply don't have enough people in the study to appreciate a difference if it is there
- the study ends up being null

- The p-value does not exceed the threshold the 0.05 and you say, “Look, there is no difference between treatment A and treatment B
- when in reality, there may well have been, but you didn’t have the power to determine it and therefore you don’t actually know if you should have rejected the null hypothesis or accepted it

Over-powered studies

- An equally sinister, but perhaps not as common, is when a study is overpowered
- you have more people in the study than you should have had for the effect size
- you start to find things that are statistically significant, but are probably irrelevant clinically
- That’s when you start to pick up an effect size of 1% when you’re dealing with something clinically that should never be thought of as being relevant below 10% detection threshold
- you typically don’t see this as much with clinical trials, but you’ll see this more with kind of data dump trials, data mining studies, where they’re grossly overpowered

> “When you look at a study and it’s not significant, you should ask a question, ‘Was this study powered correctly?’” —Peter Attia

Measuring effect size: Relative risk vs. absolute risk, hazard ratios, and “Number Needed to Treat” [1:07:45]

What are some of the ways in which researchers measure the association or the “effect size” in these studies?

> A lot of times it’s only reported as a **relative risk,** but you can’t really talk about relative risk without knowing **absolute risk**

Hypothetical example:

- At the end of study, group 1 had a 5% risk of dying and group 2 had a 3% risk of dying
- The absolute risk reduction (ARR) is therefore 2% (5%-3%=2%)
- That’s important to know because often what’s only reported is the relative risk reduction, which is the absolute risk reduction over the non-exposure absolute risk
- In this case, the relative risk reduction would be 2% divided by 5% (the non-exposure risk) which comes out to be a 40% relative risk reduction
- Both of those things are important, but it’s really critical that you know both

Real world example: Women’s Health Initiative (WHI)

- This study was looking at the increase in the risk of breast cancer for the women who were receiving the estrogen and synthetic progesterone treatment
- The study reported that the women receiving the hormone replacement therapy had a 25% increase in breast cancer
- So if you went from 4 cases of breast cancer per thousand women to 5 cases of breast cancer per thousand women, that is indeed an increase of 25%
- But what’s the absolute risk increase? ⇒ It’s one over a thousand or **0.1%**

- Peter says, "*what I usually say to women when we're talking about hormone replacement therapy is you can kind of use that as your ceiling for the true risk increase of this therapy even if you discount the 12 mistakes in that study that make it hard to believe that that effect size would hold*."

Hazard ratios [1:11:25]

- Another way that we tend to measure effect size or association is using something called a hazard ratio
- It actually involves some really complicated math called a Cox proportional-hazard
- The magic of the hazard ratio is it is temporal so it captures the risk of something, i.e. the hazard over time, and that differentiates it from something called an odds ratio which only can measure risk over the entire period of time
- At the risk of oversimplifying this, let's talk about the hazard ratio over a given period of time but acknowledging that its real magic is its ability to tell you what's happening at any point in time

Pretend we're talking about a cancer drug trial:

- The hazard rates, i.e. the rates of disease progression, were 20% in one group and 30% in another group
- So the people getting the drug progressed 20% of the time, the people not getting the drug progressed 30% of the time
- So the hazard ratio is the ratio of 0.2 to 0.3, which is 0.667—in other words, the treatment group was 67% as likely to experience disease progression as the control group
- You flip the math and say, "Well, what if you saw the exact same rates, but in something that was desirable?"
    Then it would be the 0.3 over the 0.2 which equates to 1.5—which means there's a 50% increase in the benefit or the harm if it's something that's harmful
- Hazard ratios are ubiquitous in clinical trials…the thing you just have to know is how to do the math on it

The math:

- Say there's a hazard ratio is 0.82
    - You're comparing the experimental group to the control
    - The experimental group is at about 18% reduced risk of whatever the event is you're talking about of progression
    - The math is 1-0.82=0.18 which translates to a reduction of 18%.
- Now say there's a hazard ratio of 2.2
    - If we said 1.8, it would be an 80% increase
    - 2.2 would be 120% increase
    - The math: 2.2-1=1.2 and then you multiply by 100 so you get 120%

Back to absolute risk reduction (ARR):

- There's another common theme you'll hear about in trials called the number needed to treat or the NNT analysis
- And this gets back to the importance of absolute risk reduction

Let's say there's an example where you're testing a drug to see if it reduces heart attacks:

- The people who take it have four heart attacks per thousand people over a five year period
- Then the placebo, they have five events over that same period of time per thousand people
- The drug reduces the events from five out of a thousand to four out of a thousand
- *What's the relative risk reduction there?* ⇒ The relative risk reduction is 20%, (4-5)/5 — in this case is a 20% relative risk reduction
    - You might see this and say, "This is something we should be putting in the drinking water. This is such an important thing."
    - But, you want to calculate *how many people do you need to treat* to prevent the event
    - To do that, you have to take 1 and divide it by the **absolute risk reduction**, not the relative risk reduction
    - And the absolute risk reduction here is 0.01% — so 1 divided by 0.01% is 1000
        So this means you have to treat 1,000 people to achieve the effect, which means *you better figure out what the side effects are of that thing, what the cost of that thing is, what the complexity of it to justify it*
    - There may be certain things for which a NNT of a thousand is valuable, but you wouldn't say that the board
- Conversely, if you have a drug that reduces the risk of death from…
    - 4% to 3%, then you would say 4-3=1% and 1 divided by 1% is an NNT of 100
    - If it took it from 4% to 2%, it would be 1 divided by 2% is an NNT of 50
    - If it went from a reduction of death from 4% to 1%, your difference is 3%, you're now talking about an NNT of 33

> "*As a general rule, we love to see drugs in that sub 100 range of NNT. We tend to not get that impressed when the NNT of something is like a thousand*

How to interpret confidence intervals [1:17:30]

What are they?

- Confidence intervals are technically intervals in which the population statistic could lie
- Typically, what you see on a paper is this 95% CI and it's usually a reported next to the hazard ratio
- Say the hazard ratio 0.5, which means having of the risk in the experimental group versus the control group
- And then you'll see this 95% confidence interval
- So they'll give you these two numbers, for example, let's just say 0.2 to 1.2 is your confidence interval

- That really the flip side of the significance level, which is 1-α (α being the p-value, but also being the false positive rate)
- So it's the flip side, so when you see 0.05 for your p-value, that's a tip off that your confidence interval is 95%.

Confusion around the term confidence interval:

- A lot of people think about the word confidence in this definition, and they take it to mean the probability that a specific confidence interval—so in the example is 0.4 to say 1.2—that interval between those two ratios contains the population parameter
- They think, "Okay, we could be 95% confident that the true effect, say meat consumption and cancer, is between these two numbers," but that's not really what the confidence interval suggests
- Really, it means that if you were to take a hundred different samples and compute this confidence interval, then approximately 95 out of those 100 will contain the true mean value.
- "*It's been described by some as an **uncertainty interval** rather than a confidence in interval*." says Peter

Another way to interpret this:

- A "quick and dirty" way to do this is just to look at the confidence interval and ask if the interval contains "1" or not
- Bob's example had a hazard ratio of 0.5 with a confidence interval of 0.4 to 1.2—So that would not be significant
- So even though your hazard ratio, you might look at that and say, "Oh, look, that was a big reduction. 0.5 hazard ratio means a 50% reduction," but your confidence interval was very wide
- Going from 0.4 up to 1.2 means it crosses over unity
- Conversely, if you had a hazard ratio of 0.5 but your confidence interval was 0.4 to 0.6 or even up to 0.9, you would say, "indeed, that is at the level of not 95%, that is confidence"

Also worth pointing out…

- The closer one edge of the confidence interval comes to 1, the closer the p-value is to 0.05
- When you have a confidence interval that runs from 1.01 up to 2.0, your p-value is probably about 0.049 or something like that
- Whereas when you have confidence intervals that are miles away from 1.0, the p-values tend to be very small

"uncertainty in intervals"

- Andrew Gelman, a statistician, talked about uncertainty in intervals
- Reason why he says that–imagine you've got a huge confidence interval meaning instead of 0.4 to 1.2, it was like 40% reduction at 0.4 or it went all the way out to say a thousand

- He would say that's a huge uncertainty interval but the way that we think about confidence is that's a huge confidence interval and it's maybe intuitively backwards for some people
- The tighter the interval, the more confidence you actually have in it (And obviously it can't cross 1.0)

Example:

- Say you see a hazard ratio was 1.4 in a paper
- You might say, "Oh wow, 40% increase"
- But then there's a 95% confidence interval that went from 1.1 to 17
- Do I really have a lot of confidence in that? ⇒ No, that's an enormous uncertainty interval.

> "If you make the decision that you want to pay attention to science, you just have to roll up your sleeves and accept the fact you're not going to be able to read these things in the bathtub on a lazy Sunday morning, you kind of have to roll up your sleeves and pay attention to all of this little stuff." —Peter Attia

***Beware of where you get your information on studies**

"It's unfortunate that I think a lot of people in the media don't know how to do this and yet they're the ones that are reporting on these things…

…So if you're getting your science info from Twitter and from the news, there's a little bit of a buyer beware, you have to understand the fact that it's very likely that the people that are reporting these things, not because they're necessarily not well intentioned, but they themselves might not be doing the type of analysis that's necessary."

Why a study might be stopped before its completion [1:23:45]

Do studies ever stop mid-way through? If so, why?

> Yes. Due to safety, benefit, or futility.

Safety

- the first and most important of these is safety
- A phase 1 trial looks at safety
- Phase 2 is about efficacy and safety
- Anytime there's a safety breach, which means there is a statistically significant difference between an important safety metric between the groups, that'll just stop the study.

Benefit

- The second thing that will stop a study is benefit

- The PREDIMED study is an example:
    when it was first done, is it stopped two thirds of the way through, because it was deemed that there was such a benefit to the group on the mediterranean diet relative to the low-fat diet that would've been unethical to let those people on the low-fat diet continue for another two and a half years on a diet that was so clearly increasing their risk of mortality

Table 3. (Continued.)

End Point	Mediterranean Diet with EVOO (N=2543)	Mediterranean Diet with Nuts (N=2454)	Control Diet (N=2450)	P Value†	
				Mediterranean Diet with EVOO vs. Control Diet	Mediterranean Diet with Nuts vs. Control Diet
Hazard ratio for Mediterranean diets combined vs. control (95% CI)					
Primary end point					
Unadjusted	0.70 (0.55–0.89)		1 (ref)	0.003	
Multivariable-adjusted 1§	0.71 (0.56–0.90)		1 (ref)	0.004	
Multivariable-adjusted 2¶	0.71 (0.56–0.90)		1 (ref)	0.005	
Secondary end points‖					
Stroke	0.61 (0.44–0.86)		1 (ref)	0.005	
Myocardial infarction	0.77 (0.52–1.15)		1 (ref)	0.20	
Death from cardiovascular causes	0.83 (0.54–1.29)		1 (ref)	0.41	
Death from any cause	0.89 (0.71–1.12)		1 (ref)	0.32	

* CI denotes confidence interval, and ref reference.
† All P values were calculated with the use of Cox proportional-hazards models with robust variance estimators and stratification according to recruiting center.
‡ The primary end point was a composite of myocardial infarction, stroke, and death from cardiovascular causes.
§ The primary end point was stratified according to recruiting center and adjusted for sex, age (continuous variable), family history of premature coronary heart disease (yes or no), and smoking status (never smoked, former smoker, or current smoker).
¶ The primary end point was additionally adjusted for body-mass index (continuous variable), waist-to-height ratio (continuous variable), hypertension at baseline (yes or no), dyslipidemia at baseline (yes or no), and diabetes at baseline (yes or no).
‖ The secondary end points were stratified according to recruiting center and adjusted for sex, age (continuous variable), family history of premature coronary heart disease (yes or no), smoking status (never smoked, former smoker, or current smoker), body-mass index (continuous variable), waist-to-height ratio (continuous variable), hypertension at baseline (yes or no), dyslipidemia at baseline (yes or no), and diabetes at baseline (yes or no).

Figure 5. Hazard ratios for cardiovascular outcomes. Credit: Estruch et al., 2013

Futility

- The final thing that will stop a study prematurely, is futility
- This is a little bit harder to understand, but it actually comes down to that hazard ratio concept which is able to measure risk temporarily in an aggregate fashion
- if two thirds of the way through a study, there's no benefit and statistically, you know that nothing that's going to happen in the remainder of the study is going to change that, you stop the study

Trial that was stopped due to safety: The CETP inhibitor torcetrapib [Honey, 2007; NYT]

- On December 2, 2006, Pfizer halted the development of torcetrapib, after a little more than a year of treatment [Tall et al., 2007], a decision made by the sponsor on the recommendation of the trial's independent steering committee, which was acting on advice from the independent data and safety monitoring board.
- In this case, the trial was set up with 7,500 patients in each group
    - One group was on a CETP inhibitor plus a statin (Lipitor)
    - The other group was on Lipitor alone
- The background on this: CETP inhibitors raised HDL cholesterol so the idea was to pair it with Lipitor which lowers LDL cholesterol
- They found that 82 patients receiving the drug combination had died, compared with only 51 on Lipitor alone and so they halted the study a little over 1 year in (it was intended to go 4.5 years)
- In this case, they had this pre-specified P value of less than 0.01 based on a test for death from any cause and that's what alerted them
- And there is still a published paper, even though the trial only went for 12 months in the New England Journal of Medicine, and they report those endpoints where the study was stopped

Table 3. Estimated Hazard Ratios for Protocol-Specified Cardiovascular Outcomes.*

Variable	Atorvastatin Only (N=7534)	Torcetrapib plus Atorvastatin (N=7533)	Hazard Ratio (95% CI)	P Value†
	number (percent)			
Primary composite outcome‡	373 (5.0)	464 (6.2)	1.25 (1.09–1.44)	0.001
Secondary outcome				
Death from coronary heart disease	33 (0.4)	40 (0.5)	1.21 (0.77–1.92)	0.41
Nonfatal myocardial infarction§	118 (1.6)	142 (1.9)	1.21 (0.95–1.54)	0.13
Stroke	40 (0.5)	43 (0.6)	1.08 (0.70–1.66)	0.74
Hospitalization for unstable angina	201 (2.7)	270 (3.6)	1.35 (1.13–1.62)	0.001
Death from any cause	59 (0.8)	93 (1.2)	1.58 (1.14–2.19)	0.006
Tertiary outcome				
Composite of death from coronary heart disease, nonfatal myocardial infarction, and stroke§	185 (2.5)	214 (2.8)	1.16 (0.95–1.41)	0.14
Stroke				
Hemorrhagic	2 (<0.1)	5 (0.1)	2.50 (0.49–12.91)	0.26
Ischemic	30 (0.4)	31 (0.4)	1.03 (0.63–1.71)	0.89
Embolic	9 (0.1)	7 (0.1)	0.78 (0.29– 2.09)	0.62
Not classified	0	0	NA	NA
Coronary revascularization procedure	403 (5.3)	505 (6.7)	1.27 (1.11–1.44)	<0.001
Peripheral vascular disease¶	159 (2.1)	110 (1.5)	0.69 (0.54–0.88)	0.003
Transient ischemic attack	13 (0.2)	23 (0.3)	1.77 (0.90–3.50)	0.09
Hospitalization with primary diagnosis of congestive heart failure	50 (0.7)	84 (1.1)	1.69 (1.19–2.39)	0.003
Major coronary event‖	147 (2.0)	179 (2.4)	1.22 (0.98–1.52)	0.07
Major cardiovascular event and coronary revascularization procedure	589 (7.8)	738 (9.8)	1.27 (1.14–1.42)	<0.001
Major cardiovascular event, coronary revascularization procedure, and peripheral vascular disease	723 (9.6)	820 (10.9)	1.15 (1.04–1.27)	0.008
Stroke and transient ischemic attack	53 (0.7)	65 (0.9)	1.23 (0.85–1.77)	0.27
Major coronary event, stroke, and transient ischemic attack	197 (2.6)	234 (3.1)	1.19 (0.99–1.44)	0.07
Procedure-related myocardial infarction	8 (0.1)	11 (0.1)	1.38 (0.55–3.42)	0.49

* Data were censored on December 2, 2006, the date of the termination of the study. NA denotes not applicable.
† P values were calculated with the use of the log-rank test.
‡ The primary composite outcome was the time to the first occurrence of a major cardiovascular event, a composite that included four components: death from coronary heart disease, nonfatal myocardial infarction (excluding procedure-related events), stroke, and hospitalization for unstable angina.
§ Procedure-related myocardial infarction was excluded from this category.
¶ Peripheral vascular disease includes either first diagnosis or any procedure.
‖ A major coronary event was the time to the first occurrence of death from coronary heart disease or nonfatal myocardial infarction (excluding procedure-related events).

Figure 6. HRs for cardiovascular outcomes and death. Credit: Barter et al., 2007

Comments about CETP inhibitor

> It turned out that CETP inhibitors in general are not a good thing. At best, they do nothing. And at worst, they kill people and probably had to do with the fact that they're altering HDL function
>
> Peter spoke to Tom Dayspring about this on the podcast

Trial that was stopped due to futility: Look AHEAD, 2013

- Randomly assigned about 5,000 overweight or obese patients with type 2 diabetes to participate in an intensive lifestyle intervention that promoted weight loss through decreased caloric intake and increased physical activity (intervention group) or to receive diabetes support and education (control group). The primary outcome was a composite of death from cardiovascular causes, nonfatal myocardial infarction, nonfatal stroke, or hospitalization for angina during a **maximum follow-up of 13.5 years.**
- The trial was **stopped early on the basis of a futility analysis when the median follow-up was 9.6 years**.
- "On September 14, 2012, on the basis of a futility analysis and recommendation from the data and safety monitoring board, the study's primary sponsors instructed the study investigators to terminate the intervention. All data were censored on this date. **At that time, the probability of observing a significant positive result at the planned end of follow-up (i.e., a hazard ratio of 0.82 in the intervention group) was estimated to be 1%.**" [Look AHEAD, 2013]
- "The intervention was stopped for futility by the DSMB [data and safety monitoring board] on September 14, 2012, at which point the trial was converted to an observational study. … **The DSMB determined, in August 2012, that the chance of detecting a difference in the primary outcome between treatment assignments was small and requested that the intensive lifestyle intervention be ended and Look AHEAD be converted to an observational study. This change took effect on September 14, 2012 with no further intensive lifestyle intervention sessions being offered at the sites**. [protocol]

Table 2. Primary and Secondary Outcomes and Other Cardiovascular Outcomes.*

Outcome	Patients with Event	Control Group	Intervention Group	Hazard Ratio (95% CI)	P Value
	no.	*no. of events (rate/100 person-yr)*			
Primary outcome					
Death from cardiovascular causes, nonfatal myocardial infarction, nonfatal stroke, or hospitalization for angina	821	418 (1.92)	403 (1.83)	0.95 (0.83–1.09)	0.51
Secondary outcomes					
Death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke	550	283 (1.25)	267 (1.17)	0.93 (0.79–1.10)	0.42
Death from any cause, nonfatal myocardial infarction, nonfatal stroke, or hospitalization for angina	1025	529 (2.43)	496 (2.25)	0.93 (0.82–1.05)	0.23
Death from any cause, nonfatal myocardial infarction, nonfatal stroke, hospitalization for angina, CABG, PCI, hospitalization for heart failure, carotid endarterectomy, or peripheral vascular disease	1177	600 (2.81)	577 (2.67)	0.94 (0.84–1.05)	0.29
Other cardiovascular outcomes					
Death					
Any cause	376	202 (0.86)	174 (0.73)	0.85 (0.69–1.04)	0.11
Cardiovascular cause	109	57 (0.24)	52 (0.22)	0.88 (0.61–1.29)	0.52
Myocardial infarction					
Fatal or nonfatal†	354	191 (0.84)	163 (0.71)	0.84 (0.68–1.04)	0.11
Fatal	16	11 (0.05)	5 (<0.02)	0.44 (0.15–1.26)	0.13
Nonfatal	342	183 (0.80)	159 (0.69)	0.86 (0.69–1.06)	0.16
Hospitalization for angina	390	196 (0.87)	194 (0.85)	0.97 (0.80–1.19)	0.79
Stroke	165	80 (0.34)	85 (0.36)	1.05 (0.77–1.42)	0.78
Heart failure	218	119 (0.51)	99 (0.42)	0.80 (0.61–1.04)	0.10
CABG	525	269 (1.21)	256 (1.14)	0.93 (0.78–1.10)	0.41
Carotid endarterectomy	54	25 (0.11)	29 (0.12)	1.10 (0.64–1.87)	0.74

* CABG denotes coronary-artery bypass grafting, and PCI percutaneous coronary intervention.
† Patients who had both nonfatal and fatal myocardial infarctions were counted only once.

Figure 7. Hazard ratios for cardiovascular outcomes. Credit: Look AHEAD, 2013

Why only a fraction of studies are ever published and how to combat publication bias [1:31:30]

The process of getting a study published:

- Once the study is done and they've done their analysis and they write up a manuscript, they'll submit it to a journal for publication

- That journal will have an editor who will look to see if the paper meets their criteria and if they think it's original and interesting. "*Is this paper adding something to the body of knowledge?*"
    - At that point, the editor might just say, "*Hey, this is not really a good fit for our journal*" Or for whatever reason, "*This is something we're not interested in any further*"
    - But otherwise, the editor is going to invite individuals that are typically part of an editorial board to peer review the manuscript
- You hear this term all the time: "*Is this a peer reviewed publication?*"
    - That's important, because not all things that get published have been peer reviewed and that's obviously the highest standard
    - The reviewers are basically invited, not randomly, but because they have expertise in this area
- Other things are important such as **conflicts of interest** *(financial or philosophical)*
    - They might have to decline if they're conflicted
    - There's a whole deeper discussion about when you have sort of philosophical conflicts of interest with the person
- The concept of "blinded" reviewers:
    - Peer reviews can be blinded or not blinded
    - It can be single blinded (most common), where the reviewer knows who the author is but the author doesn't know who the feedback is from
    - It can be double blinded where the reviewer doesn't know who it's being written by and vice versa, and they can be completely open
- Review process:
    - You'll typically have three reviewers review something and they can either accept it outright, reject it outright, or make recommendations for revisions
    - the most common thing where they say, "We're still interested in this paper, but did you actually consider this hypothesis?"
    - Sometimes the revisions are just, "*Repeat your analysis."* Sometimes, it's, *"Do another experiment.*"
        - Peter says: "I've had papers where that happened where I'd done a series of experiments and I'd written it all up and I'd submitted, and the reviewer came back and said, 'Well, you really should have done this experiment as well. Because this would've served as another control'"
        - So you go and repeat that experiment. And of course, when you're working in cell culture or something like that, it's not that onerous.
- This process can go on several times, but ultimately, the editor makes a decision to accept that paper and publish it, or reject it again
- As an author…
    - you're going to try to get your paper published in the most prestigious journal
    - you'll sort of keep going down the pecking order until you can get it into the right journal

Do all studies get published?

- "*No, many don't. I think this is a really big problem*." says Peter
- There is a massive problem called **publication bias**—It occurs when the outcome of an experiment or research study influences the decision whether to publish or otherwise distribute it.

    An egregious example of publication bias is with "negative" studies
    - Positive results—finding a relationship between variables or an effect of treatments on outcomes—are rewarded more than negative results—failing to find a relationship or effect
    - There are a lot of studies that don't get published because they're negative — "*And that's a shame because when something doesn't work, it is just as important as when it does work.*"

> "*If you want to go out and do an experiment and 10 people have done that experiment before you and it's always failed, wouldn't it be great to know that? Would that impact your decision on whether or not you want to do the experiment a certain way or would you want to try something a little bit different? So you can see very quickly this becomes problematic when papers don't get published*

Are there ways that can combat publication bias?

- One of the most important steps is **pre registration—**Which is you force investigators to pre register their experiments on clinicaltrials.gov
- On that pre registration that list out things like statistical methods, number of subjects, primary and secondary outcomes, etc.
- That basically makes it a lot harder to say, "I'm not going to publish this when it comes out if it doesn't turn out the way I wanted it to."
- There's both requirements of journals as well as requirements of funding entities, which say, "We won't fund you unless the study is pre registered"
- Registered reports is a publishing format created by an organization called the Center for Open Science founded by Brian Nosek

Why certain journals are more respected than others [1:40:30]

The annual Journal Citation Reports (JCR) impact factor is a ratio between the number of citations to articles published in the journal over the previous year to the number of articles published in that same time period.

- In other words, *how often are articles from a given journal cited*?
- It's often used as a proxy for the relative importance of a journal within its field.
- By this metric, journals like NEJM, Nature, Cell, The Lancet, Science, are considered among the most prestigious medical journals.

    Out of the 13,000 journals included, about 98% have an IF <10. About 95% have an IF <5. About half of the journals have an IF <2.

Rank	Full Journal Title	Total Cites	Journal Impact Factor
1	CA-A CANCER JOURNAL FOR CLINICIANS	39,917	292.278
2	NEW ENGLAND JOURNAL OF MEDICINE	347,451	74.699
3	Nature Reviews Materials	12,657	71.189
4	NATURE REVIEWS DRUG DISCOVERY	33,154	64.797
5	LANCET	256,199	60.392
6	WHO Technical Report Series	3,560	59.000
7	NATURE REVIEWS MOLECULAR CELL BIOLOGY	46,307	55.470
8	Nature Reviews Clinical Oncology	12,384	53.276
9	NATURE REVIEWS CANCER	52,053	53.030
10	CHEMICAL REVIEWS	200,014	52.758
11	Nature Energy	17,747	46.495
12	ASSOCIATION (JAMA)	158,632	45.540
13	REVIEWS OF MODERN PHYSICS	51,122	45.037
14	CHEMICAL SOCIETY REVIEWS	150,703	42.846
15	NATURE	767,209	42.778
16	SCIENCE	699,842	41.845
17	Nature Reviews Disease Primers	7,567	40.689
18	World Psychiatry	6,486	40.595
18	World Psychiatry	6,486	40.595
20	NATURE REVIEWS IMMUNOLOGY	42,168	40.358
21	NATURE MATERIALS	99,502	38.663
22	CELL	258,178	38.637
23	NATURE BIOTECHNOLOGY	63,979	36.558
24	NATURE MEDICINE	85,220	36.130
25	Living Reviews in Relativity	3,074	35.429
26	Nature Reviews Chemistry	3,209	34.953
27	NATURE REVIEWS MICROBIOLOGY	32,027	34.209
28	LANCET ONCOLOGY	53,592	33.752

Figure 8. Impact factor 2019 rankings: Journal Citation Reports. (Journals with over 100,000 citations highlighted) Credit: Mandal, 2020

- Highlighted are the journals that have more than 100,000 citations
- So you've got the New England Journal of Medicine—which is kind of staggering—nearly 350,000 citations
    - Because if you divide 347,000 by that number, you get 74.699
    - that's the impact factor for the New England Journal of Medicine
    - The Lancet, by comparison, has 250,000 citations with an impact factor 60
- There's kind of an outlier: The Cancer Journal for Clinicians, which has an impact factor of 292, despite only having 40,000 citations.
    - That's a little bit of a skew and Peter doesn't really consider that to be in the same league, because it's basically the Global Cancer Society Statistic article and therefore, it reports on tons of cancer statistics
    - It doesn't really publish that much but it gets referenced so much
    - Because anytime someone is basically referencing a cancer statistic, they're going to reference that
- Notice that the WHO technical report series has an impact factor of 59 but it's only cited like 3,500 times, it's cited a lot for a very few number of publications

- But the ones that really matter here clinically are the New England Journal of Medicine, The Lancet, and JAMA

Peter's process when reading a scientific paper [1:43:45]

Overview of Peter's process:

1 – Abstract (first)

1b. Introduction (I only read if I'm not familiar with subject matter)

2 – Methods (third)

3 – Results (fourth)

Figures and tables first, then the prose

4 – Discussion (fifth)

Abstract

- The title of the paper is usually not sufficient for Peter to know if he's going to be interested, but the abstract usually indicates whether he wants to read further
- "*I could read 10 abstracts in a matter of minutes and decide, 'Do I want to read three of these papers?*"

Introduction

If not really familiar with subject matter, he will read the intro

Methods

- If the paper passes the abstract test, he'll move to the methods section where it gets into the details
- For example, say it's an exercise study with muscle biopsies
  - So how many subjects were there?
  - How were they randomized?
  - What were the interventions?
  - When were the biopsies taken?
  - Was there a crossover?
  - And so on.

Results

- Next, he goes the results section, but her first looks at the **figures and tables** before reading
- Figures and tables should, in Peter's opinion, be standalone, so the legend should explain everything you need to know.
- Then reading the prose of the results section kind of adds a little bit more color to that

Discussion

- This is the last thing he'll do
- By this point, he'll have formulated his own thoughts on what the strengths and weaknesses of the studies are, what questions remain, etc.
- Oftentimes, the authors will have thought of things that he hasn't thought of or they'll thought of things that he disagrees with

When writing a paper:

- Peter learned in Steve Rosenberg's lab when doing experiments and writing up the results
- Peter says, when you finished an experiment, the very first thing you did was you made the figures and tables and the legends that go with them
    - You wouldn't really take pen to paper to write anything until you head that down
    - You had to sort of know, "*What are the relevant figures? What are the relevant tables? Can I explain them very concisely in a legend?*"
    - The last thing you would write would be the intro in the abstract

§

Selected Links / Related Material

Peter interviewed on the Tim Ferriss Podcast: Dr. Peter Attia on Longevity Drugs, Alzheimer's Disease, and the 3 Most Important Levers to Pull (#517) | (tim.blog) [2:25]

Peter's five-part series on studying studies: [2:30]

- Studying Studies: Part I – relative risk vs. absolute risk
- Studying Studies: Part II – observational epidemiology
- Studying Studies: Part III – the motivation for observational studies
- Studying Studies: Part IV – randomization and confounding
- Studying Studies: Part V – power and significance

The Drive episode with Steve Rosenberg: #177 – Steven Rosenberg, M.D., Ph.D.: The development of cancer immunotherapy and its promise for treating advanced cancers

AMA episode discussing a phase IV trial with semaglutide: #184 – AMA #29: GLP-1 Agonists – The Future of Treating Obesity?

The Drive episode with John Ioannidis where he argued that it's really not possible to identify, let alone eliminate, all confounders: #143 – John Ioannidis, M.D., D.Sc.: Why most biomedical research is flawed, and how to improve it

PREDIMED study that was stopped due to benefit: [44:30]

- First paper was published in New England Journal of Medicine in 2013: "The Mediterranean Diet, its Components, and Cardiovascular Disease" (Widmer et al., 2013)

- A reanalysis was published in 2018: Primary Prevention of Cardiovascular Disease with a Mediterranean Diet Supplemented with Extra-Virgin Olive Oil or Nuts (Estruch et al., 2018) [45:45]

The New York Times article discussion the flaws of the PREDIMED study: That Huge Mediterranean Diet Study Was Flawed. But Was It Wrong? | Gina Kolata (nytimes.com) [46:15]

On The Drive Podcast, Rick Doblin mentioned that one of the huge challenges of studying psychedelics is it's very difficult to blind anybody, most of all, the subject: #65 – Rick Doblin, Ph.D.: MDMA — the creation, scheduling, toxicity, therapeutic use, and changing public opinion of what is possibly the single most important synthetic molecule ever created by our species

The Women's Health Initiative: Women's Health Initiative | (wikipedia.org) [1:09:45]

Andrew Gelman's piece that talked about uncertainty in intervals: Are confidence intervals better termed "uncertainty intervals"? | (Gelman and Greenland, 2019) [1:20:45]

CETP inhibitor trial that was stopped due to safety: Drug designed to raise HDL levels falls down (Honey, 2007) [1:25:25]

NYT article about the CETP study that was stopped: Pfizer Ends Studies on Drug for Heart Disease | Alex Berenson [1:25:25]

The published paper that still remains on the CETP inhibitor trial that was stopped: Effects of Torcetrapib in Patients at High Risk for Coronary Events (Barter et al., 2007) [1:27:30]

The Drive Podcast with Tom Dayspring where he discusses CETP inhibitors: #22 – Tom Dayspring, M.D., FACP, FNLA – Part III of V: HDL, reverse cholesterol transport, CETP inhibitors, and apolipoproteins

The Drive Podcast with Eric Topol: #91 – Eric Topol, M.D.: Can AI empower physicians and revolutionize patient care?

The Look AHEAD Trial that was stopped due to futility: Cardiovascular Effects of Intensive Lifestyle Intervention in Type 2 Diabetes (The Look AHEAD Research Group, 2013) [1:30:15]

Registered reports is a publishing format created by an organization called the Center for Open Science: Center for Open Science | (wikipedia.org) [1:37:50]
§

People Mentioned

- Tim Ferriss [2:25, 1:22:45]
- Ronald Fisher [1:00:00]
- Steve Rosenberg [10:15]
- James Yang [17:45]

- John Ioannidis [43:45, 48:15]
- John Carlisle [44:45]
- David Allison [44:45]
- Rick Doblin [52:15]
- Andrew Gelman [1:20:45]
- Tom Dayspring [1:28:15]
- Ron Krauss [1:28:15]
- Eric Topol [1:29:30]
- Brian Nosek [1:37:50]

§