By Alan Flanagan

Hello, and welcome to this Nutrition Science Explained segment, where today we're going to be focusing on the GRADE system of evaluating evidence. GRADE stands for the grading of recommendations, assessment development, and evaluations, and while not specific to nutrition science necessarily. The application of the GRADE framework to assess the certainty or quality of evidence has posed issues for the interpretation of nutrition research, particularly in relation to epidemiology and the findings from cohort studies.

We'll specifically talk about the concept of effect size. And how systems like GRADE, which are born out of the biomedical model, in effect by default bias, the findings from nutrition research towards being rated as "low certainty" and quality evidence overall. So first off, what is GRADE? GRADE is a framework that was developed for presenting a summary of evidence as part of informing clinical practice recommendations and in the biomedical world GRADE has become the most widely adopted tool for the grading of the quality of evidence. Like many issues that pertain to nutrition science methodology, the difficulty is we have the application of a framework with GRADE that has largely been developed to consider evidence either from epidemiological exposures in a biomedical hazard context, for example, the risk of smoking cigarettes and any cancer or cardiovascular related

outcomes. And the difficulty with this application as well is posed in relation to randomized control trials because of the use of this system to rate trials based on biomedical criteria, but where the friction arises for nutrition research is primarily in relation to epidemiology and prospective cohort study outcomes, which will be much of our focus of today's segment.

So how does GRADE work and what is the process of evaluation? Like a systematic review or indeed a meta analysis., usually the GRADE system will look to a specific research question. For example, what is the evidence for the consumption of soy products and breast cancer risk? Or what is the evidence for the consumption of red meat products and colorectal cancer risk? So it starts with that particular research question and the evidence is then rated at the individual study level overall. So for example, let's say we are proceeding with this question of the associations between soy food intake and breast cancer.

And let's say hypothetically, that we identify 10 cohort studies, we would take each of those cohort studies and based on certain aspects of their methodological quality, we would then give each of these studies a rating. And then from that overall pool of studies with their individual ratings, we would then come to an overall certainty rating.

That would then apply to the level of evidence. And so at this point, it's important to note that the concept of "certainty" of evidence is not necessarily any reflection of the findings of that evidence. And I'm going to highlight a specific example of the application of GRADE as a certainty rating for evidence and how that can stand against the actual findings.

For a given exposure outcome relationship and how the use of GRADE can often mean that people overlook other ways of thinking about evidence, such as whether the findings are congruent with wider evidence in a given field. So GRADE, like we've said, rates the certainty of evidence and it does so primarily by rating four different levels of certainty: within the GRADE system, we can go from very low certainty and that means that a true effect size is probably substantially different from the estimated effect.

OK. So for example, let's say that we have our 10 cohort studies that we have included in an analysis. And the average estimated relative risk is 1.2. Now in the GRADE framework that could easily attract a certainty rating of low, to

very low. And what that would mean practically, if we assume for our example that the certainty rating was very low.

What that would mean was that there would be an assumption that the actual true effect would be different to that perhaps even to the point of not even being statistically significant or associated in the same direction of effect. So in the example, we used a relative risk of 1.2 or 20% relative risk increase that would make an assumption that could even be not the true direction of effect.

And then we have from very low to low, which is where the true effect might be again, markedly different from the estimated effect. So the difference here is the choice of language. So let's be clear about this and repeat these two certainty ratings very low is that the true effect is probably markedly different from the estimated effect low, is that the true effect might be markedly different from the estimated effect.

So you can see in the choice of language here between probable and might, that we are very much in a balance of probabilities type of discourse, as it relates to certainty as a particular outcome, moderate evidence. Remember, what we mean by moderate is the moderate certainty rating would be that the authors believe that the true effect is probably close to the estimated effect.

Again, note the use of the word probable in the rating of moderate and high certainty GRADE rating would mean that the authors have a lot of confidence that the true effect is similar to the estimated effect. So generally speaking, most of these ratings are in between very low to low, similar in principle, but differ in the words probable or might.

So one is probably different from the estimated effect low. The other is, might be different from the estimated. And then moderate and high differ in again, moderate being the true effect is probably close to the estimated effect and then high being a lot of confidence that the true effect is similar to the estimated effect.

Now it's important to consider the various domains in which evidence is assessed for each of the qualities of the included studies. And there are a number of main factors that are considered when dealing with the evidence

assessment and coming to an ultimate certainty rating conclusion. The first is risk of bias.

And so bias, particularly in, in research can result when the actual findings of a study. Cannot be taken to necessarily represent a quote true effect because of limitations in the design or the conduct and execution of that study. an example in nutrition research is if GRADE was being applied to a randomized controlled trial and that randomized controlled trial used a food based intervention.

For example, we were comparing the addition of two tablespoons of olive oil per day, compared to continuing with a usual cooking fat or oil. However by nature, this type of intervention would be single blind because the participants obviously know that they have to take in olive oil every day in the biomedical framework, the lack of full double blinding would mean that a type of intervention such as described would be rated as having a higher risk of bias, because it was only single blind rather than double.

In the observational research context, and this is particularly important for nutrition research and prospective cohort studies. The simple fact that the study is a prospective cohort design would automatically mean that it would carry an assessment of risk of bias in relation to the the findings of the study itself.

So it's important at the outset to see that although many of these tools in the biomedical realm are offered as objective work throughs of an evidence assessment. It should be obvious to you at this point that in fact, there is an inherent degree of subjectivity in the application of these types of certainty ratings.

The second potential consideration in evaluating evidence through the GRADE approach is in relation to inprecision. Now, generally we would think about imprecision as the variance around the mean outcome i.e. That would typically be represented by the standard deviation or the 95% confidence intervals around the estimated effect. So for example, if we take a cohort study with our previously used example of a relative risk of 1.2. And so let's say that's 1.20 representing a 20% relative risk increase of whatever our outcome is. And let's say our confidence intervals are 95% confidence intervals around that point estimate of 1.20 are 1.13. To 1.27. Now they would

be relatively precise as far as the spread of confidence intervals around that mean estimate, but where those confidence intervals to be, for example, 1.0 to 2.31, then we would have a much wider. Estimate of precision around our point estimate. And we would also have that lower bound, that 1.02 being very close to the 1.0 mark, the null or no significant association mark. And so the level of precision in estimates is quite important and this doesn't necessarily inherently biased nutrition research per se because well conducted cohort studies with large sample sizes and a large amount of event outcomes.

For example, coronary heart disease events. Can often have quite precise confidence interval estimates around the single point estimate or the relative risk or hazard ratio summary. However, inprecision also comes into the effect sizes themselves. And so this is a factor that does play into bias against nutrition research, because.

Such types of magnitude of relative risk as we are discussing in our examples are in fact considered to be very imprecise and often are dismissed as being low effect sizes. The third consideration is consistency. Okay. And this is where several studies show a consistent direction of effect. and what they're typically looking for in this is also similarity in the point estimates.

So for example, let's take our relative risk of 1.20 or a 20% increase in risk where another study to find that the same exposure outcome relationship had a 1.27 relative risk. And the other study had a 1.18. And another study had a 1.2, three. These would all be within a fairly similar and indeed consistent point estimate of effect.

And it would also be desirable in each of those contexts that the confidence intervals overlap with each other. However, this does present some challenges for nutrition research because the effect of the same exposure is not necessarily homogenous. I E it's not necessarily the same across all populations factors, like the background diet, the baseline health status of the participants, the actual dose or level of intake of the food itself.

How that level of intake is stratified and divided to compare a high versus low intake and other factors can play into the ultimate estimate of. So if we use red meat, as an example here, if you look at some of the us cohorts, for

example, the nurses health study, the health professionals follow up study, or the national institutes of health retired person studies.

What you'll often see is with high levels of intake of over a 150/160/170 grams a day. See relative risks of often at least 1.2 to 1.3 and even over 1.4. If we came to a European cohort and on paper, it's nominally comparing high versus low, we might find a relative risk of 1.05. That is not significant. And we might have confidence intervals from 0.9, six to point to 1.08, just as an example. And so at a superficial level, people using the GRADE system would say, this is inconsistent, but they haven't done due diligence to go further and reconcile that the actual dose compared in each respective study was the same where they to do that.

They might find that in the European cohorts, for example, the very highest level of intake is maybe around a hundred grams. And so where we to actually do subgroup. As some studies have in European populations to find higher levels of intake that reflect what we see in the us cohorts. We would actually see relative risks similar to those us cohorts emerge.

The problem with the application of a system like GRADE is that it essentially encourages lazy epistemic and evidence reconciliation. And this, again becomes a problem when. The application of GRADE results in a loss of confidence in the veracity and accuracy of certainly nutritional epidemiological findings.

There are a number of other factors that will play into the overall GRADE assessment. Such as the publication bias, which is generally a statistical and visual method of plotting evidence for publication bias. This is likely, mostly observed when studies are funded by industry.

Surprisingly, this tends to be not necessarily that much of an issue for nutritional research, despite what people would claim large prospective cohort studies tend not to be funded by industry. This is of course the potential limitation for randomized control trials, which are more likely to be funded by industry.

So there are these factors that feed into the assessment of the certainty of evidence. Risk of bias inprecision or precision in the findings, inconsistency or consistency in the findings, publication bias, and also a concept known as

indirectness. And that is where studies comparing the same exposure where you are looking to make sure that the.

Actual outcomes are the same that your studies are matched for the type of exposure, outcome relationship, and that no other indirect factor could necessarily be feeding into the overall results. As a result. These factors also then come into. The overall assessment when it comes to factoring in low or very low levels of certainty.

Okay. If there is evidence of risk of bias, if the findings are imprecise or the estimate of effect is imprecise. If the estimate of effect is inconsistent in different studies, if there's evidence of publication bias, these are all factors that will go into what is known as rating. These were typically coalesce in an assessment to result in a certainty rating of very low or low.

However, there are factors within the GRADE system that can allow for a study to be rated up. And there are three in particular. The first is that there is a very large magnitude of effect. The second is that there is a dose response gradient. And the third is that all other possible residual confounding that could decrease a magnitude of effect has been accounted for in the analysis.

So there are a number of things in each of these that we can consider in turn. The first is what is a large magnitude of. This is where there does carry a default implication of the application of these considerations for nutrition research, particularly for epidemiology in order to rate up a study to move it.

For example, from low evidence to moderate or from moderate to high certainty, a study is required in this context to have a large magnitude of effect of over 2.0 to 5.0, these are quite significant and substantial magnitudes of effect. The types of magnitudes of effect in observational research often only observed with high levels of cigarette smoking.

Even if you were to compare smoking 10 cigarettes a day. None. You may not even find a relative risk of over 2.0 where you to compare 20 cigarettes a day to none. You might very well find relative risks of over 2.0, the difficulty this holds for nutrition research is that for nutrition research, most effect sizes range.

If there increases in risk of between one to 1.5. It's rare to have effect sizes in nutrition that are over 1.6, 1.7 I E 60, 70% increases in risk. This isn't a knock on nutrition. This is simply to say that the nature of the exposure is different. These are our potentially important findings because diet, we tend to consume in gram amounts.

And our exposure is daily and our exposure is also continual across the lifespan. So the effect of consuming, for example, 150 grams of unprocessed red meat isn't necessarily occurring that day or the next day or that week or that month, or particularly even that year. It's a cumulative exposure that adds up over.

And so to dismiss magnitudes of effect, simply because of their actual point estimate. And without the context of thinking about what the nature of the exposure is shortsighted and results in a default bias against the findings from nutrition research, which never reached that magnitude of effect.

And therefore cohort study findings are never rated up. Irrespective of whether the cohort was well executed in its dietary assessment possessed a very large sample size had, or encompassed a large number of events for its end point and all other methodological criteria that we would consider to assess the quality of a prospective cohort study in nutrition, the dose response gradient.

Is also really important. It somewhat relates to the magnitude of effect, but what a dose response gradient essentially is describing is that for each increase in the dose of your exposure, there is a corresponding increase in risk. So dose response curves are hugely illustrative and supportive of causal relationships in epidemiological research.

If we go back to our cigarette comparison, although the magnitude of effect for someone smoking five cigarettes compared to none, or maybe 10 to none would not necessarily be as great as someone. Comparing 20 cigarettes to none where we to stratify that exposure and look at five to none, 10 to none, 15 to none, 20 to none 25 to none.

We would still see that the risk increased from one to the other category. As we sequentially increased the actual dose of the exposure. We do see dose response gradients in nutritional epidemiology. We do see dose response

gradients in nutrition research. However, that dose response gradient is highly dependent on having a sufficiently wide variation in the levels of intake of a given exposure in the population.

One of the challenges for nutrition research is. In typical populations in Western industrialized countries, dietary intake at the population level is fairly homogenous. And the variation that exists in people's intake is low variability. And this can often mean that the lack of evidence of a dose response gradient isn't because there is not a dose response gradient, but it may simply be a reflection of the.

that within that given study, there wasn't a sufficiently large variation and contrast in that exposure to create evidence of a dose response gradient. So again, this is something that will typically by default result in an automatic bias against nutritional epidemiology research. But it often comes back to the fact that people are just assuming that a dose response gradient exists or does not often independent of the actual levels of intake that have been stratified in a given study.

The final one is that all residual confounding that could decrease a magnitude of effect are ruled out now modern nutritional epidemiology is becoming very adept at using sophisticated adjustment models to consider the potential role of both non dietary lifestyle, confounding factors and dietary lifestyle, confounding factors in order to fully isolate the effects of a given exposure on the particular outcome that is being looked at in that.

However, all of this ties together with the fact that the magnitude of effect is often not as large, and there may not always be evidence of a dose response gradient. In fact, where in the GRADE system of the evidence assessment, a relative risk, the magnitude of effect is for example, 1.2, this will often lead to then an assumption of this third.

Criteria for rating a study up that we cannot rule out residual confounding that could decrease this magnitude of effect, which in this evidence assessment would be labeled small and very prone to residual confounding. However, this again is shortsighted one way to think about. Confounding in nutritional epidemiological research is to think, as afore mentioned, in terms of non dietary and dietary lifestyle factors in relation to non dietary factors,

these are factors that we largely know factors like smoking alcohol intake, BMI, and other non dietary, but lifestyle related factors, even to.

Factors like hours of sleep per night. If these are measures that are taken in the study are all measures that can be included in the statistical adjustment model. And this is really important because a common pushback that people will hear in relation to observational nutrition findings is well, the people in the high category of red meat, for example, also were the highest smokers and had the highest alcohol intake and had the highest BMI.

And this is. True, but this is why statistical adjustment for these non dietary factors then allows you to see what the average weighted effect of that factor. For example, smoking or alcohol was on your outcome. And if after adjusting for these factors, the relationship between the exposure and the outcome remains statistically significant and has a precise estimate of effect in the confidence intervals around the mean summary estimate of effect.

Then this is not a reason to then think that those factors have unduly influenced the. And then for dietary confounders, there are different ways of approaching this, but one of the most common methods now is what is known as a substitution analysis. And this is for example, where in the model for the study, you assume you that total energy intake is matched by adjusting for total energy.

So that's a potential dietary confounder that would be adjusted for as a matter of standard practice in nutritional epidemiology. And then you look at the effect of replacing, for example, 5% of your intake from red meat. With 5% of intake from legumes or 5% of intake from whole milk dairy, with 5% of intake from low fat dairy.

And so on and so forth. And this is a way of inherently accounting for the potential of dietary confounders, because you are simply doing an isocaloric food substitution. The reality is that many of the large ongoing well executed current prospective cohort studies in nutrition, like the European perspective investigation into cancer or the EPIC cohort, the NIH RP; the retired person's cohort in the us, the Nurses' Health and Health Professionals follow up study in the US, the Japanese collaboration cohort in Japan, and a number of these studies.

In their analyses have quite sophisticated adjustment models that have factored in a range and multiplicity of both non dietary factors and also dietary factors in trying to tease out and isolate the independent effects of whatever the exposure of interest in that study. However as stated the issue is that if the ultimate estimate after those adjustments remains less than 2.0, the study will still fall short of the criteria using the GRADE assessment to rate up the evidence.

And so this is a default bias against nutrition research because all observational research will immediately start from a certainty rating of low. And then the criteria to rate that study's quality in terms of its certainty assessment and the overall body of evidence in its GRADE certainty rating from low to moderate.

Is going to inevitably not be met because of the criteria for application to nutrition science. Okay. So a lot of that has been discussing the methods applied to GRADE the factors considered in downgrading an evidence space, or to rate up both an individual study and an overall evidence base into one of the four GRADE certainty ratings, which as we discussed before are very low.

Moderate high. And the semantics of each of these assessments typically relate to the language of probability, either probably different or might be different either probably close to an estimated effect or confidence that the true estimate is similar to the estimate of effect we've discussed the nutrition specific issues that might face the barriers of the GRADE system for rating up, such that common nutritional dose responses may be non-linear. And that effect sizes may commonly in the increased range from one to 1.5 or in the decreased range 0.7 or 0.6. We've seen that despite the ever increasing sophistication of adjustment models to account for potential confounding in epidemiology, if the actual point estimate is low, This will typically be used to default back to say that residual confounding cannot be ruled out for that study and therefore to keep it at a low to very low certainty rating.

I want to finish this by tying this all together for you in a concrete example that we can use to picture these concepts many of which have been rather abstract in our discussions as an example in late 2020. a number of meta analyses were published by a group known as the nutritional recommendations consortium or NutriRECS.

This was a self styled and titled group who conducted a number of meta analyses on red meat consumption and health outcomes. And this included both prospective cohort studies and randomized control trials. they also then separately published a paper, which the nutrient X group termed quote, new dietary guidelines in which they recommended that individuals continue with habitual levels of red meat intake.

Now the results, the actual findings for mortality outcomes in the meta analysis of cohort studies. When we do the usual comparison of high versus low levels of intake. Included for example, a 14% relative risk reduction for cardiovascular disease. So that would be a point estimate of 0.86 I E in GRADE terms, a small magnitude of effect.

The confidence intervals around that estimate were 7 9 0 0.79 to 0.94. This was not an imprecise estimate of. But it was viewed through the lens of the actual magnitude of effect. There were other findings such as a 24% lower risk for type two diabetes. The point estimate, therefore for that, the relative risk reduction was 0.76.

The confidence intervals were 0.68 to 0.86. This. A much more robust finding from a nutritional epidemiological perspective, not just in the point estimate, but in the precision and indeed the upper bound of the confidence interval at 0.86. Now what you would have to then explain is that effect that, that upper bound could somehow be non-significant or that some residual confounding could actually explain a way.

That low of an estimate of effect. Of course, for us in nutrition, we would see that as a robust finding within the GRADE system, they would say that not only is the magnitude of effect still small, but that residual confounding could easily explain away the actual totality of the finding and the outcome and its precision with its confidence intervals.

And although much isn't made in. Nutrition, epidemiology and the criticisms thereof of the use of relative risk measures let's consider the absolute risk differences per 1000 people for overall cancer incidents. It was 18 fewer diagnosis, a confidence interval range of 11 to 26 fewer diagnoses for cardiovascular mortality.

It was six fewer diagnoses and a range of two to. now, these are not insignificant numbers when we consider a whole population approach to disease risk. So we have two outcomes here in terms of risk assessment. We have our relative risk reductions. Which are fairly robust in their precision from a nutrition perspective, but from a great perspective, the effect sizes are small, but we also have the absolute risk difference, which is quite substantial.

When we look at outcomes like overall cancer incidents, for example, however, ultimately this group assessed the certainty of evidence from these findings through use of the GRADE criteria. For certainty ratings. As a result, the evidence was rated low certainty, strength of evidence. I E recall our definition of low certainty that the true effect is probably much different from the estimated effect.

Now, in order to maintain that claim, you would want to factor in consistency. The issue with this analysis is that those findings both in terms of magnitude of effect and direction of effect, I E that compared to high levels of intake, low levels of intake were associated with these reductions in risk was entirely consistent with the wider evidence for this particular exposure.

I E red meat and outcome relationships. And yet the certainty of the evidence couldn't ever be rated up because of the actual effect size themselves. There was some evidence of a dose response gradient depending on the analysis, but it wasn't consistent overall. And again, that reflects the fact that the primary included studies used different definitions of what was high versus what was low.

And while many of the included studies used sophisticated adjustment models to account for potential confounding by non dietary and dietary factors. The fact that the magnitude of effect was what it was meant that from the GRADEs application, the application of GRADE itself, the magnitude of effect meant that in that assessment, residual confounding would not be ruled out.

Ultimately, this leads to the recommendations made by the group were not based on the actual outcomes and findings of the research itself. They were made on the basis of the low certainty, strength of evidence as assessed,

according to the GRADE criteria yet this certainty rating of low certainty evidence.

Was an inevitability of using the GRADE framework. Observational research carries a default low certainty rating to consider it moderate certainty. It was going to have to meet those criteria that we discussed for rating up, which as we can see from the findings in this study, it was never going to reach.

And in this analysis, the conclusions amounted to considering evidence. Through a GRADE framework, which has default implications for nutrition research, which resulted in the downgrading of the actual findings of the research itself, the congruence of those findings with the wider literature and the biological plausibility in the findings that we have from experimental studies and indeed from randomized control trials.

There have been proposals to improve the GRADE system, such as nutrient GRADE, for example, which moves to consider more nutrition, specific issues in the rating of the quality or certainty of nutrition, research, cohort studies and randomized control trials. But unfortunately the application of these systems.

Really hasn't taken hold in a way that it should have, and nutrition ultimately remains yet to develop more widespread use of its own standards of rating, evidence, certainty, and quality. So I hope that was helpful. Any further questions in relation to the GRADE system, its methods of assessment or its application in nutrition.

Please feel free to submit them to Danny or myself. And we will be sure to get around to answering them in the future.