Preponderance of the evidence: an example from the issue of calcium intake and body composition

Robert P Heaney and Karen Rafferty

Meta-analysis is typically applied to studies developed in fairly mature fields, but may be ill-suited for younger fields in which most of the evidence comes from studies that were designed for other endpoints entirely and that are often significantly underpowered for the effect in question. Here, there are no generally accepted methods for getting a grasp on the preponderance of the evidence. In this review, one way of doing so is proposed and the recently emergent literature concerning calcium intake and body composition is used as an illustration of how such an approach might be used.

© 2009 International Life Sciences Institute

INTRODUCTION

Few things in life are quite as certain as death and taxes. For many issues, whether or not they are hotly disputed, there is often evidence supporting both sides of the matter. Sometimes we have the luxury of suspending judgment until the issue becomes clearer, occasionally even indefinitely, but when a course of action is called for, prudent individuals act on the preponderance of the evidence. This is a concept that may be more familiar in the context of ethical or legal decision-making than it is in science, but it has applicability in most spheres of human activity. Since there is often evidence supporting both sides of many issues, even policy determinations, such as the elimination of high-fat, high-salt, and low-nutrient-density foods from school vending machines, are usually based on someone’s reading of the preponderance of the evidence.

Evidence itself comes in various levels of quality. We now recognize, for example in forensics, that DNA evidence is a much better identifier than eyewitness testimony. So, too, in clinical research, we tend to accord a higher level of credibility to the results of a randomized controlled trial (RCT) than to an observational study. But no criterion is absolute. In clinical research, RCTs are most applicable in the evaluation of drugs and much less well suited for the evaluation of nutrient effects. Furthermore, RCTs have limited generalizability, while observational studies are subject to various forms of selection, healthy volunteer, and ascertainment bias. The strongest evidence actually comes from a combination of the two types. When the two are concordant, the RCT establishes the causal connection, while the observational study contributes to the generalizability of the RCT’s results.

Recent interest in the possible relationship between calcium intake and body composition (and obesity), for which there is today no clear consensus, provides a possibly useful example of how to get a grasp on the preponderance of the evidence and how to craft a prudent course of action based on that assessment. The hypotheses concerned can take several forms; typically they...
are reduced to “inadequate calcium intake predisposes to overweight, obesity, or increased percentage of body fat”.

In situations of this sort, when studies are of varying quality and are spread over a breadth of literature, it is often difficult for busy practitioners, or even for investigators not working directly in the field concerned, to get a useful grasp of the totality of the evidence. In this forum article, we attempt to show how some apparent sense can be drawn from scattered and often confusing evidence.

**SOME PRELIMINARIES**

Before attempting to assess the evidence in this or any other field, it is essential to have some idea of what one might expect. This will always involve an understanding of basic probabilistic issues, as well as consideration of features that are unique to the particular problem and to the biology of the nutrient being evaluated.

**Probabilistic issues**

Assume, purely theoretically, that a certain population is known to have a 20% risk of developing diabetes in any given 5-year period; assume also that the population concerned has a low average intake of nutrient X; finally, assume that both animal and epidemiological data suggest that inadequate intake of nutrient X increases the risk of diabetes substantially. Thus, there is biological plausibility to the hypothesis that intake of nutrient X influences risk of diabetes in this population. Assume, further, that the actual fact (unknown, of course, at the time of an investigation) is that the nutrient does function as expected, and that a population replete with nutrient X has only a 10% chance of developing diabetes over the same 5-year period, i.e., a 50% reduction in risk. If we were to mount a randomized controlled trial involving 200 subjects, 100 assigned to a low intake of nutrient X and 100 to a higher intake, what might we expect to find? Would we find the actual difference? How often would such a trial succeed? How often would it fail?

Figure 1 depicts graphically the results of not just one trial, but simulations of 200 such trials, all conducted in perfectly adherent populations, with zero dropouts, and each incorporating the chance variations in individual susceptibility and responsiveness that always accompany such trials. (We have chosen ideal, if unattainable, circumstances so as to allow us to focus exclusively on the probabilistic component of the variability in trial outcome.) With our behind-the-scenes knowledge, we expect that the low-intake group would have a diabetes incidence/proportion at the end of the trial of 0.20 (i.e., the value in the untreated population), and the high-intake group would have an incidence/proportion of 0.10, for a difference in proportion of 0.10. We recognize that we would probably not get exactly such a difference in any given trial, that sometimes it might be less, sometimes it might be more; Figure 1 shows how often trials would produce any given difference in proportion between the samples. As the figure shows, most of the trials would produce results with a difference in proportion near what we would have expected, i.e., in the range of 0.05 to 0.15. However, the figure also shows, for the sample sizes we employed, that most of the studies performed would produce a result that was not significantly different from zero (i.e., they would fall in the “null” range of Figure 1). Just about 25% of all the trials would produce a significantly positive result, i.e., would correctly show that an adequate intake of nutrient X did, in fact,
reduce the risk of diabetes. Thus, this trial design had an empirical power of 0.25 (25%); that is, it would correctly identify the presence of an actual difference 25% of the time, but miss it 75% of the time. (To achieve a power of 0.85 [85%], it can be shown that two samples of 230 each would be required. Even so, with these much larger sample sizes, about one-sixth of these trials would fail to find a statistically significant difference between the two treatments.)

It may seem contrived to illustrate this point with a simulation of underpowered studies; however, several considerations make this example apt. First, while in the example of Figure 1, effect size could be known in advance, in studies of new, theoretical relationships, effect size can often be only guessed. Second, a trial of 460 individuals is expensive and time consuming and may actually be beyond the capacity of many research units. Hence, underpowered studies may be more common than one might like to believe. Third, studies of relationships such as that between calcium and body composition are often secondary endpoints in studies designed and powered for another endpoint entirely. Hence, many such otherwise well-designed studies will often be underpowered for this particular endpoint. Moreover, in the early stages of investigating a research question, it makes sense to look at RCTs that have already been performed, as there will not usually be enough information to design an adequately powered trial testing the hypothesis as a primary endpoint.

Inspection of Figure 1 shows a further interesting feature, namely that, although not statistically significant, seven of the trials actually found negative differences, i.e., the risk of diabetes was greater in the individuals receiving more of nutrient X than in the low-intake group. It is sometimes hard to appreciate that chance alone can do this; nevertheless, it is easy to guess what an investigator might conclude from results of an isolated trial that were, by chance, not only null, but seemingly in a direction opposite to the hypothesis. Many investigators, finding such a result, would conclude not just that the hypothesis had not been supported, but that there was no protection afforded by nutrient X and, if anything, the effect might be to increase the risk of diabetes.

Finally, inspection of Figure 1 shows that, even though about three-fourths of the trials failed to show a statistically significant difference between the treatment groups, the distribution of the trials is situated in a positive direction relative to zero, a point we shall come back to later when we formally consider the preponderance of the evidence.

Figure 2, by contrast, shows an identical set of simulated trials, but this time involving a truly null-effect nutritional intervention. While some individual trials will have produced seemingly positive (or negative) results, the distribution taken as a whole shows that it is, in fact, centered around a difference in proportions of zero, as would be expected. Nevertheless, about 5% of all trials produced significantly positive (or negative) results, as must inevitably occur when we use $P < 0.05$ as the criterion for statistical significance. Investigators conducting trials that turn out, by chance, to be located at the tails of the distribution might easily be persuaded that the nutrient concerned was actually beneficial (or harmful, as the case may be). Only when the preponderance of the evidence is evaluated does it become clear that their experience was atypical and likely due purely to chance.

**Unique features of the problem**

Overweight and obesity are multifactorial problems. Although the transition from normal to overweight to obesity may involve no more than an energy imbalance of ~70 calories per day, the causes for that imbalance are multiple. Altering one of those causal factors, without affecting any of the others, would be expected to reduce body weight (or the risk of weight gain) by only a relatively small amount, or would be expected to work in some individuals but not in others. The reason is straightforward: if there were, for example, six contributing factors, and if each had the same degree of influence, then altering only one of them would reduce the overall effect by perhaps only one-sixth. One might even conclude that such an effort, even though successful, would not be very useful. But that would not be correct. For example, in NHANES-III mean BMI in non-Hispanic white women aged 40–59 years was $28.3 \text{kg/m}^2 \pm 1.8$. About 40% were obese (BMI > 30 kg/m$^2$). A weight reduction across this cohort to a mean BMI of 25.0 kg/m$^2$ would be sufficient.
to reduce the prevalence of obesity by 50%, despite the fact that the average per-person weight loss would amount to less than 12%. In brief, in multifactorial disorders, small individual effects can result in big population-level benefits.

**Nutrient biology**

Unlike most drugs, many nutrients (perhaps most) exhibit threshold behavior. That means the specific nutrient effects increase as intake rises from deficient to replete levels, but then no further response occurs as intake continues to increase. One well-known example is the rise in hemoglobin as iron is administered in patients with iron deficiency anemia, but no increase in hemoglobin beyond individual normal hemoglobin values. Calcium behaves similarly. During growth, skeletal retention rises as intake increases, up to a plateau level, beyond which even large increases in calcium intake produce no further retention. Thus, in evaluating evidence from studies testing the possible effect of increased intake of most nutrients, it is essential that the low-intake (control) group have a study intake sufficiently below the plateau threshold to allow a nutrient effect in the high-intake group to be detectable.

While other, similar considerations could be mentioned, these suffice to illustrate the kinds of expectations one should bring to any attempt to synthesize the results of multiple studies.

**THE CASE IN POINT: CALCIUM AND BODY COMPOSITION**

What we have attempted to do with these preliminaries is to elucidate a set of expectations that one should bring to an analysis of the evidence. First, we should expect some studies of a truly efficacious agent to have a null outcome, even when adequately powered. Moreover, when one is evaluating secondary outcome measures in studies that were designed for a different primary endpoint, underpowered studies should be expected. Second, in multifactorial situations, one should expect the effect of a single factor to be relatively small. Third, one should expect null results in most or all studies in which there is not a contrast group with low nutrient intake.

Although the evidence relating calcium intake to body composition can be traced back as far as McCarron’s original analysis of the NHANES-I data, as well as isolated weight loss studies, such as that of Summerbell et al., attention of the scientific community to this topic did not really begin until the publication by Zemel et al. in 2000. Since that time, there have been roughly 80 human studies relating calcium and/or dairy intake to body composition. Thirty-one of the total were randomized controlled trials or controlled metabolic experiments, and 61 were observational and/or epidemiological studies. The source of calcium varied among studies, with dairy foods being the calcium source in the majority of the studies, and supplemental calcium salts providing the calcium in the rest. Although a few reviews have been published, there has been no systematic attempt to look at the large body of evidence, taking it as a whole. When evaluating the individual studies, it becomes clear that many (indeed most) suffer from one or more of the shortcomings outlined briefly in the foregoing paragraphs. Many were small, indeed quite small; and many failed to have low-calcium contrast groups. In short, they exhibited the features that are common at the early stages of investigation of a field when the relevant conditions may not be fully understood or ensured. Nevertheless, in keeping with the approach we have taken in this forum, it may be useful to look at them in their totality. Taken as a whole, what do they seem to show?

Figure 3 sets forth the frequency distribution for the 31 RCTs and metabolic studies, and Figure 4 the corresponding distribution for the 61 observational and epidemiological studies. In both cases, the studies assembled had many different measurement schemes and endpoints; hence, it was not possible to group the results along an interval continuum (as in Figures 1 and 2). Accordingly, we have used a categorical variable for displaying the results. There were five logical possibilities: 1)
the study showed essentially zero difference (Z), 2) the study was negative (weight gain) or 3) positive (weight loss), but not statistically significantly so (N and P, respectively), or the study was 4) significantly negative (SN) or 5) significantly positive (SP). As Figure 3 shows, none of the 32 randomized controlled trials or metabolic studies was negative, let alone significantly so; 15 showed either zero difference or were positive to a nonsignificant degree, i.e., null-effect studies; and 17 were significantly positive. Thus, ignoring entirely the differences in quality among the various studies, the preponderance of the evidence nevertheless points toward the hypothesized effect of calcium intake on body composition.

The 61 observational studies point to a similar conclusion. Here, there were four studies that were actually significantly negative, i.e., calcium or dairy intake was associated with significantly greater weight or weight gain. Nine showed essentially zero effect, four were positive, i.e., they showed weight loss, but not at significant levels, and 45 were significantly positive. It is generally understood that observational studies are often confounded by uncontrolled variables. For example, the significantly negative studies in Figure 4 could be the result simply of the fact that a higher caloric intake might be expected to lead to weight gain, altogether apart from dietary calcium intake. Conversely, the significantly positive studies could be influenced by such factors as the consumption of dairy foods or calcium supplements potentially being a marker for health consciousness, one of the features of which would be more attention to weight control.

It is not our intention to explain away inconvenient results or to argue for or against one or the other conclusion, but simply to call attention to the inescapable fact that uncontrolled factors may strongly influence the outcomes of observational studies. Nevertheless, the general similarity of the pattern of the observational studies to that of the randomized trials is compatible with what one would expect from a group of studies of mixed quality for an intervention with a real effect.

Comment

It is important to stress that the foregoing review of the evidence is not a meta-analysis and is not a substitute for one. In fact, the extreme heterogeneity of the studies would make a suitable meta-analysis difficult to perform and could easily lead to charges of bias if the selection factors for study inclusion were such as to include predominantly positive or predominantly negative studies (as many such systematic reviews of various topics appear to have done). Instead, we have assembled, so far as we were able, all of the relevant human studies. This approach has shown that the totality of the evidence tilts distinctly in favor of the hypothesis. Even the RCTs, providing level 1 evidence, are of widely varying quality. Perhaps most persuasive are the results from the Women’s Health Initiative (WHI) Study, involving over 36,000 women — more than in all the other controlled trials combined. The results of this trial strongly support the hypothesis, if for no other reason than the failure of the trial to include a truly low-intake contrast group, suggesting a real effect size larger than actually found. If there were no other studies to consider, this result might be taken as definitive. Additionally, some of the other RCTs that were smaller in size but had low-intake control groups, were, nevertheless, distinctly positive as well, suggesting considerable robustness of the result.

The fact that there is not yet a consensus on this matter is likely due to several factors: 1) the large number of null-effect studies; 2) the smallness of the effect in individuals; and 3) the understandable attachment of investigators to the results of their own studies. As already noted, small effects are what should be expected in studies of multifactorial problems. Moreover, a small effect is often hard to find, and seemingly unimportant even when conclusively shown to be real. Once again, this latter reaction is likely because, when we evaluate effects, we tend to think in terms of individuals rather than populations. A 10–15 lb weight difference in a 250 lb woman seems trivial, but as already noted, the same reduction across the population reduces the prevalence of overweight and obesity to a major extent.
CONCLUSION

At the outset we mentioned the frequent need to craft a prudent course of action based upon the preponderance of the evidence when no true scientific consensus yet exists. In the example we have employed, the preponderance of the evidence makes it seem highly likely that a high calcium intake affects body composition by some combination of favoring lean over fat mass, reducing the tendency to gain weight, and augmenting weight loss on energy-restricted diets. Since calcium intakes of adults in the United States are generally well below the recommended values, since increasing calcium intake into the ranges used in the studies we have summarized here is both safe and inexpensive, and since the obesity problem is not getting better by itself, it would seem that the prudent course of action is to push vigorously to improve calcium intake, not just for its skeletal benefits, but for its effect on body composition as well.

Acknowledgment

Declaration of interest: Dr. Heaney works occasionally on projects the Osteoporosis Research Center undertakes for the National Dairy Council; he also gives occasional talks at the Speakers Bureau. Karen Rafferty has no interests to declare.

REFERENCES

27. Major GC, Alarie F, Doré J, Phouttama S, Tremblay A. Supplementation with calcium + vitamin D enhances the beneficial


