National Cancer Institute

measurement ERROR webinar series

Combining self-report dietary intake data and biomarker data to reduce the effects of measurement error

Laurence Freedman, PhD, MA
Gertner Institute for Epidemiology

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Slide 1

Hello and welcome to the 11th webinar in the Measurement Error Webinar Series. I'm Kevin Dodd with the Division of Cancer Prevention at the U.S. National Cancer Institute. Today we will be hearing from Dr. Larry Freedman about combining self-report instruments and biomarkers, but before we get started, please note that the webinar is being recorded so that we can make it available on our Web site. All phone lines have been muted and will remain that way throughout the webinar. Following the presentation, there will be a question and answer session; please use the Chat feature to submit a question. A reminder: You can find the slides for today's presentation on the Web site that has been set up for series participants.

Now it's my pleasure to introduce the presenter for today's webinar. Dr. Laurence Freedman is Director of the Biostatistics Unit at the Gertner Institute for Epidemiology, where he directs a research and consulting program in biostatistics and advises the government on public health policy. Larry has previously worked for the British Medical Research Council and the U.S. National Cancer Institute, where he was Acting Branch Chief of the Biometry Branch from 1993-1996 and was part of the team that developed the Women's Health Initiative and the AARP Nutritional Cohort Study. He was founding co-editor of *Statistics in Medicine*, and has also served as co-Editor of *Biometrics*. As I mentioned, today Dr. Freedman will discuss combining self-report dietary intake data and biomarker data to reduce the effects of measurement error. Dr. Freedman.

measurement ERROR webinar series

Today's presentation will be a LIVE audiocast

You must join the teleconference to listen to the session

(To join, click the telephone icon in the top right of your screen; *audio will not be broadcast through computer speakers*)

Slide 2

Thank you, Kevin.

Good morning, afternoon, or evening to everyone. For me it is late afternoon. Today, I will be describing methods for combining data from dietary intake self-reports with dietary biomarkers, with the object of reducing the errors in dietary intake measurement.

measurement ERROR webinar series

This series is dedicated to the memory of

Dr. Arthur Schatzkin

In recognition of his internationally renowned contributions to the field of nutrition epidemiology and his commitment to understanding measurement error associated with dietary assessment.

Slide 3

This series is dedicated to the memory of our dear colleague, Arthur Schatzkin.

Presenters and Collaborators

Sharon Kirkpatrick
Series Organizer

Regan Bailey	Laurence Freedman	Douglas Midthune
Dennis Buckman	Patricia Guenther	Amy Subar
Raymond Carroll	Victor Kipnis	Fran Thompson
Kevin Dodd	Susan Krebs-Smith	Janet Tooze

Slide 4

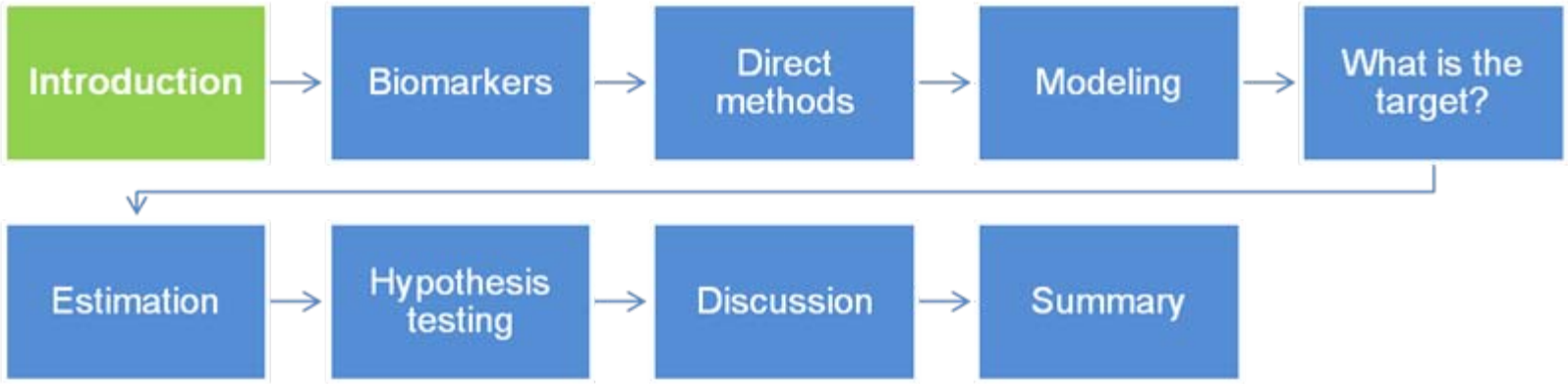
And here is a list of the presenters and collaborators in this webinar series.

Learning objectives

- Understanding the motivation for combining dietary self-reports and biomarkers
- Understanding different methods of combining self-reports and biomarkers, their aims and the knowledge required for implementing each method
- Understanding the potential gains of such combination and the limitations to the methods

Slide 5

My aims today are, firstly, to give you an understanding of the motivation for combining self-reports with biomarkers; secondly, to describe methods of combining these sources of data, together with their aims and the information that is required to implement each method; and, thirdly, to show you the potential gains as well as the limitations of each method.



INTRODUCTION

Slide 6

We'll start with a general introduction.

Main results on impact of measurement error

- When a dietary exposure measured with error is included in a disease outcome regression model:
  a) Risk estimates are factored down (attenuated)
  b) Study power is decreased (see lectures 6-7)
- These problems are caused by a **loss of information** about usual dietary intake caused by the measurement error
- In the previous lecture and in this lecture we deal with this loss of information

Slide 7

To understand why we would want to combine self-report data with dietary biomarker data, we need to go back to some of the lessons that we learned in lectures 6 and 7 of this webinar series. We learned that when self-reports are used in a study to investigate a diet-health relationship, the measurement error in the report causes: estimates of relative risk or odds ratios to be attenuated towards the null value; and, secondly, study power to be decreased.

These effects are a direct result of the loss of information about usual dietary intake that occurs due to measurement error. Both last week's lecture by Doug Midthune and my lecture today discuss how we might recover some of the lost information.

Supplying further information about intake

- In Lecture 10 we described how combining self-report instruments could increase information about usual intake and thereby help with relative risk estimation and power
- In this lecture we focus on combining dietary self-reports with biomarkers, to increase information about intake

Slide 8

Last week Doug Midthune explained how combining different self-report instruments can help to recover some of the lost information and thereby alleviate the attenuation of estimated relative risks and loss of study power. Today, we will learn how combining self-reports and biomarkers can also help, although we will see that some of the conceptual and statistical issues are different from those involved in combining self-reports.

Background

- Suppose we have a nutritional cohort study in which we want to relate usual intake, T, of a specific nutrient to a health outcome, Y
- We will consider the case where Y indicates whether an individual develops a specific disease (Y=1) or not (Y=0)
- We cannot measure T exactly and in its place we obtain a self-report from a food frequency questionnaire, Q

Slide 9

To place our discussion in a clear context, suppose we are conducting a nutritional cohort study in which we have particular interest to relate usual intake, T, of a specific nutrient or food group to a health outcome, Y. We'll suppose that the health outcome is a binary variable indicating diagnosis of a specific disease. And, as we have noted before in this webinar series, we cannot measure T exactly and we obtain, instead, a self-report. Here we will assume that it is a food frequency questionnaire, and that the reported intake is denoted by Q.

Disease model - logistic regression

Disease model:

$$\log\{\text{Odds}(Y = 1)\} = \alpha_0 + \alpha_T T + \alpha_{Z1} Z_1 \ldots + \alpha_{Zp} Z_p$$

Y = health outcome variable (0 or 1)

T = dietary exposure (true usual intake)

$Z_1 \ldots, Z_p$ = other exposures, confounders, effect modifiers or intermediate variables

α's = log odds ratios for the explanatory variables

Slide 10

We'll also suppose that the association between the health outcome and the usual intake of interest is a logistic regression, as described in this equation.

As just mentioned, Y is a binary health outcome, and T is the true intake of the dietary component of interest. The Z's in the equation are other exposures, confounders, effect modifiers, or mediators. In particular, we'll be talking quite a bit today about confounders and mediators. The alpha coefficients in the equation are the log odds ratios corresponding to each explanatory variable, and we will be particularly interested in estimating the log odds ratio, alpha subscript T, for the dietary intake.

Attenuation

Disease model:

$$\log\{\text{Odds}(Y = 1)\} = \alpha_0 + \alpha_T T + \alpha_{Z1} Z_1 \ldots + \alpha_{Zp} Z_p$$

- Instead of T, we obtain a report Q
- If we use Q instead of T in the regression, then our estimate of α_T will be attenuated

Slide 11

We cannot obtain an exact measure of usual intake, T. Our problem is that if we use our food frequency report, Q, in its place, then our estimate of the log odds ratio alpha-T will be attenuated.

Regression calibration to adjust the estimate

Regression calibration:

$$\log\{\text{Odds}(Y = 1)\} = \alpha_0 + \alpha_T T + \alpha_{Z1} Z_1 \ldots + \alpha_{Zp} Z_p$$

- Instead of using Q in the regression, use $E(T|Q,\underline{Z})$
- $E(T|Q,\underline{Z})$ is the value of true intake that is **predicted** when the report is Q and the other explanatory variables are $Z_1, \ldots, Z_p$

Slide 12

In Lecture 7 of this series Doug Midthune explained the method of regression calibration which is used to deattenuate the estimate of the log odds ratio and thereby give an unbiased estimate of alpha-T. The method entails using not Q as our measure of usual intake, but the expected value of usual intake conditional on Q and on the other explanatory variables in the model, denoted by this symbol E, (for expectation), of T (the usual intake) given the values of Q and Z.

You can think of this expression as the predicted value of true intake given our knowledge of the reported intake and the other explanatory variables.

Usual regression calibration does not increase power

- Regression calibration removes bias from the estimate, but usually makes little or no change to the result of the test of the null hypothesis that the log odds ratio is zero
  - *Occasionally a result that was significant using the unadjusted method will become non-significant - see Lecture 7*
- This is because usual regression calibration uses the **same** information, Q, about dietary intake as does the unadjusted method
- In this lecture, we will consider using together with Q, a biomarker value, M

Slide 13

Although regression calibration successfully removes bias from our estimate of the log odds ratio, it unfortunately usually has no impact on the second problem caused by dietary measurement error, the loss of study power. In fact, regression calibration usually makes little or no change to the result of the test of significance of the log odds ratio, and if the result is nonsignificant using the attenuated estimate arising from entering the value Q into the logistic regression model, then it is usually also nonsignificant using the regression calibration adjustment. In a very small proportion of cases, a result that is statistically significant using the unadjusted method can become nonsignificant using regression calibration.

The reason that it usually has no impact on the problem of lost power is because regression calibration uses the same information about dietary intake (essentially the value Q) as does the unadjusted method. In order to improve the study power, we need to bring in additional information about dietary intake, and in this lecture we will consider doing that through a dietary biomarker whose value we will denote by M.

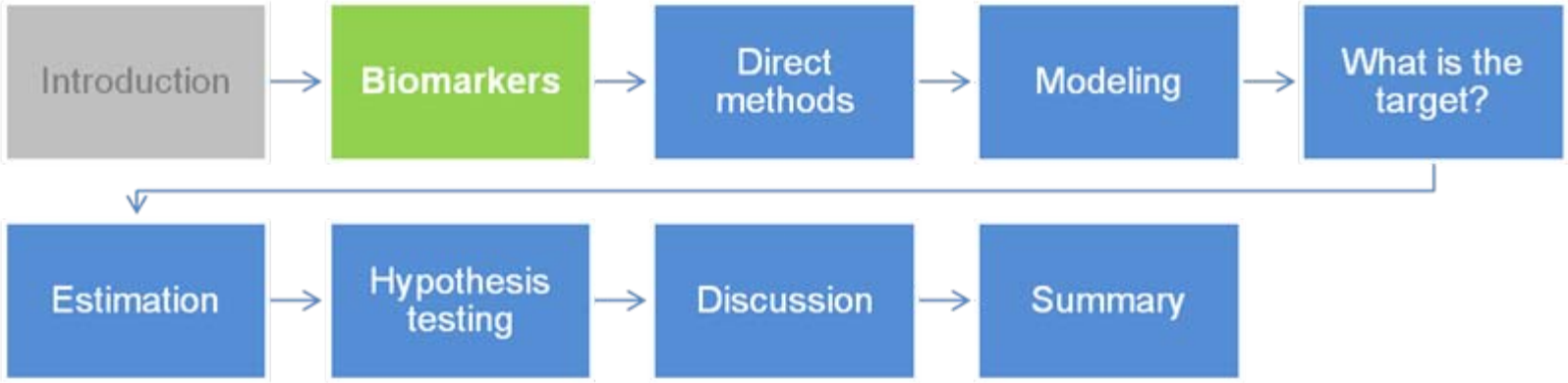
Methods of combining self-report and biomarker

- Two main approaches to combining self-reports and biomarkers:
  - **Direct** methods, that can sometimes recover lost power but do not yield unbiased estimates of relative risk
  - A more complex **modeling-based** method, that recovers lost power and gives unbiased relative risk estimates, but that requires more information about the biomarker's relation to true usual intake

Slide 14

We're going to consider two rather different approaches to combining information about dietary intake from self-reports and biomarkers. The first, and simpler, approach we call the "direct" approach. These methods are relatively easy to use and can sometimes recover lost power, but are not guaranteed to do so. In addition, they do not yield unbiased estimates of relative risks or odds ratios.

The second approach we call the "modeling-based" approach, and it will normally recover some of the power lost through dietary measurement error and will also give unbiased estimates of relative risks or odds ratios. It therefore carries substantial advantages over the direct approach. However, unlike with the direct approach, to implement the modeling-based method we need to supply information on the relation between the biomarker and true usual intake, and the necessary information may not always be available.


Introduction
Biomarkers
Direct methods
Modeling
What is the target?
Estimation
Hypothesis testing
Discussion
Summary

BIOMARKERS

Slide 15

Before we get into the methods of combining biomarkers with self-reports, we need to review briefly some background regarding dietary biomarkers.

Biomarkers (1)

Dietary biomarkers:
Biological measurements related to dietary intake

- Recovery biomarkers
  - Ideal measures of intake that have no (or minimal) bias
  - Only a few are known
- Concentration biomarkers
  - Other biomarkers that are correlated with dietary intake; these comprise the vast majority of biomarkers

Slide 16

A dietary biomarker is any biological measurement that is related to a dietary intake.

Two main classes of biomarkers have been recognized. The first class, which has been termed recovery biomarkers, are those that are in a sense ideal in that they measure a certain dietary intake with little or no bias. Unfortunately, only a few are known.

The remaining biomarkers, which form the great majority of those available, have been termed concentration biomarkers.

Biomarkers (2)

- Recovery biomarkers

  i. Based on recovery of specific biological products directly related to intake, and not subject to substantial inter-individual differences in metabolism

  ii. Measure short-term intake

  iii. Only a few are known:

  – Doubly-labeled water for energy intake*

  – Urinary nitrogen for protein intake

  – Urinary potassium for potassium intake

  iv. Measure intake directly with minimal bias. The error is independent of true intake

** Under assumption that person is in energy balance*

Slide 17

The recovery biomarkers are called that because they are based on the recovery of specific biological products directly related to intake and are not subject to large interindividual differences in metabolism. By their nature, they measure short-term intake.

The three that are known are doubly-labeled water for energy intake, 24-hour urinary nitrogen for protein intake, and 24-hour urinary potassium for potassium intake. Besides measuring intake with minimal bias, they also have the important property that their errors are independent of true intake.

Biomarkers (3)

- Concentration biomarkers

  i. Concentrations in blood, urine or tissues of specific chemicals or compounds

  ii. Related to dietary intake but not in a straightforward manner

  iii. Could depend on factors that affect metabolism (e.g., gender, smoking, other dietary intakes)

  iv. Very many are known:
  e.g., Serum lipids, carotenoids, vitamins, metals

Slide 18

The concentration biomarkers are called that because they are usually based on concentration of specific chemicals or compounds found in blood, urine, or tissues. They are related to dietary intake, but not in a straightforward manner. Because of the complex metabolic processes underlying them, they could depend on dietary factors other than the target component, or on nondietary factors such as hormone levels, gender, or smoking.

This is a large class including serum lipids, carotenoids, vitamins, and metals, among others.

Biomarkers (4)

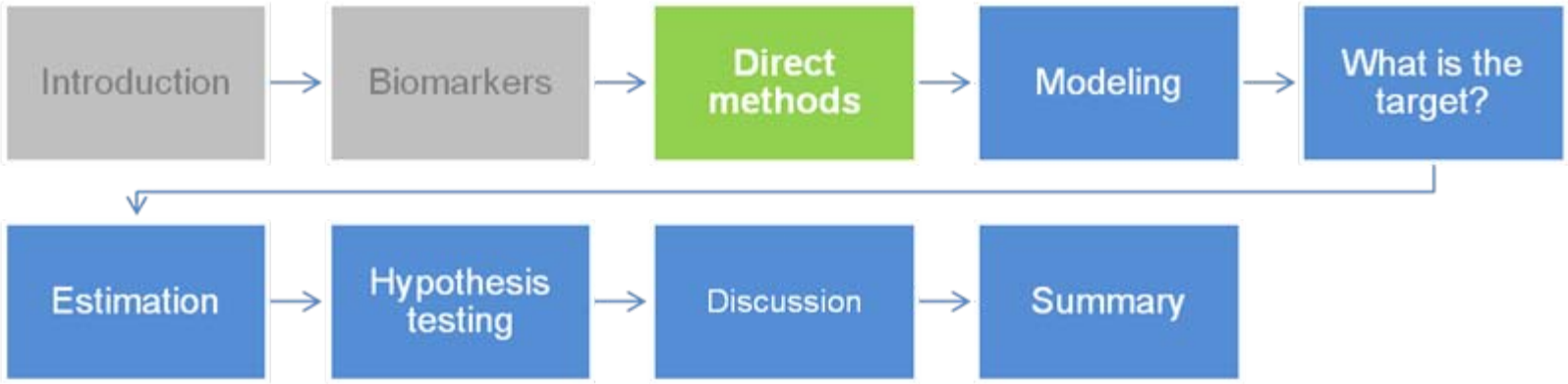
Use of biomarkers

- Recovery biomarkers:
  i. As the reference instrument in validation studies (see Lectures 6 and 7)
  ii. Combined with self-reports, using the same methodology as described in lecture 10
- Concentration biomarkers:
  i. Combined with self-reports using methods we will describe in this lecture

Slide 19

The recovery biomarkers have been extremely useful in the evaluation of self-report instruments, as we have already explained in lectures 6 and 7 of this series. They could also be used, potentially, in combination with self-reports using the same methods as were described by Doug Midthune in the previous lecture, with the object of increasing precision in measuring dietary intake. However, this latter use is limited by their high cost or difficulty of specimen collection.

The concentration biomarkers can be combined with self-reports, also with the object of increasing the precision in measuring dietary intake, and that's what I'm going to describe to you now.



DIRECT METHODS

Slide 20

As I mentioned earlier, there are two main types of methods for combining concentration biomarkers with self-reports. The first type of method, which is the simpler one, we call the direct method.

Introduction

Suppose that we conduct to investigate the association between a dietary intake T and a health outcome Y.

We measure the dietary intake using a self-report instrument, e.g., an FFQ, value denoted by Q. We also measure a dietary biomarker for the intake, with value M.

In the usual unadjusted method, we regress:

Outcome Y on (i) FFQ reported intake Q, and
(ii) confounders Z

leading to loss of power, because of measurement error in the FFQ-report Q.

Slide 21

Suppose that we are interested in the association of a dietary component whose true usual is T and a health outcome, Y. We cannot measure T, but we obtain a self-report from a food frequency questionnaire denoted by Q, and also a dietary marker value for that component, denoted by M.

In the usual approach, we investigate the association by regressing health outcome Y on the food frequency questionnaire value, Q, together with some known confounders, Z. This approach involves loss of statistical power due to the measurement error in the self-reported intake.

Direct methods (1)

Instead, we can incorporate the dietary biomarker value M into the analysis, as follows:

In the principal component (PC) method, we regress:

Outcome Y on:
(i) first principal component of (Q,M), and
(ii) confounders Z

In Howe's method we regress:

Outcome Y on:
(i) sum of the ranks of Q and M, and
(ii) confounders Z

Slide 22

Instead, because we have the biomarker measurement available as well as the self-report, we can combine them in one of two ways. Firstly, we may use principal components. The principal components method is a way of forming a set of linear combinations of a set of correlated variables so that the combinations themselves are uncorrelated. The first component is always the one that has the highest variance possible, and is often used to serve as a summary measure of the full set of variables. Accordingly, we take the first principal component of the pair of variables Q and M as our measure of dietary intake, and use that together with the confounders, Z, in the regression model for the health outcome, Y.

Secondly, we can use Howe's method. This involves taking the sum of each participant's ranks according to each variable Q and M. This sum is then used as the measure of dietary intake, and is entered together with the confounders into the regression for the health outcome Y. I will spell both of these methods out in greater detail in just a minute, but first I want to make a few general points about them.

Direct methods (2)

Note that these two methods (PC and Howe's):

1. Do not adjust for the attenuation in the estimated relative risk

2. Will in some circumstances recover some of the lost power caused by measurement error

3. Do not require knowledge of the quantitative relationship between marker level and true dietary intake

4. Do require that the marker (as well as the FFQ) is measured in all participants*

** A small amount of missing data may be accommodated*

Slide 23

There are several advantages and disadvantages of these two approaches. Firstly, the estimated coefficient of the combined intake variable is not adjusted for measurement error, so it will likely be attenuated. However, note that the combined variable itself has no recognized units and so the actual value of the estimated coefficient is not of intrinsic interest. What is of interest is its sign and whether or not it is statistically significant.

Secondly, in some but not all circumstances, the method will recover some of the power lost due to measurement error in the FFQ.

Thirdly, we do not need any external information regarding the relationship of the FFQ or the marker to true usual intake. The method can be used just on the data observed in the study.

And, fourthly, we need all participants to have a marker value, M, as well as the self-report, Q, and the health outcome, Y, although small amounts of missing information can be accommodated.

Direct methods (3)

Details – PC method:

1. The first principal component is given by:

   PC = Q/sd(Q) + M/sd(M)

   if Q and M are positively correlated

2. The first principal component is given by:

   PC = Q/sd(Q) - M/sd(M)

   if Q and M are negatively correlated

3. Regress Y on PC and confounders, Z

4. Test the statistical significance of the coefficient of PC

Slide 24

Before we proceed to an example that illustrates these methods, there are a few notes on each that will be helpful to those who want to use them. We start with the principal components method.

In the case where there are just two variables, the self-report and the marker, the first principal component takes a very simple form. It is the weighted sum of the two variables where the weight is the inverse of the standard deviation of the variable. This is shown in the formula presented here. The formula holds if the two measures are positively correlated.

If they are negatively correlated, then instead of the sum we take the difference, as shown here. And then, as previously explained, once we have formed the principal component variable and calculated it for each participant, we enter it together with the confounders into a regression model for the health outcome and test whether its coefficient is significantly different from zero.

Direct methods (4)

Details – Howe's method:

1. The method is a non-parametric procedure
2. Rank the Q's according to their values from lowest to highest
3. Rank the M's according to their values from lowest to highest
4. For each individual calculate H = Q-rank + M-rank (or, if Q and M are negatively correlated, H = Q-rank - M-rank)
5. Regress Y on H and confounders, Z
6. Test the statistical significance of the coefficient of H

Slide 25

Howe's method for two variables may be thought of as a nonparametric version of the principal components method. The way it is performed is as follows.

First, the participants are ranked according to their self-report values, Q. And then the same is done according to their marker values, M. Then, each participant's ranks on the two variables are summed, giving a result denoted by H for Howe. Finally, as before, H is entered with the confounders into the regression model for the health outcome, and the regression coefficient for H is tested.

Direct methods (5)

Example: Carotenoids in Eye Disease Study (CAREDS)

1. Ancillary study of the Women's Health Initiative Observational Study
2. 1802 women were recruited to CAREDS during 2001-4
3. Disease of interest, Y: nuclear eye cataract; defined according to current eye examination or reported previous treatment for cataract

Slide 26

We'll now look at an illustrative example taken from an ancillary study of the Women's Health Initiative known as CAREDS, where the interest was in associations between carotenoids and eye disease. There were 1,802 women in this ancillary case-control study and the eye disease of interest in our example is nuclear cataract.

Direct methods (6)

Example: Carotenoids in Eye Disease Study (CAREDS)

4. Dietary intake of interest, Q:
   FFQ-reported lutein plus zeaxanthin

5. Biomarker, M:
   serum level of lutein plus zeaxanthin

6. Confounders, Z:
   age (y)
   smoking (0=never, 1=past, 2=current)

Slide 27

The carotenoid intakes of particular interest are lutein and zeaxanthin, which are combined for our analysis. Available in the study were: a food frequency questionnaire report on these carotenoids, and also the serum levels. The most important confounders were age and smoking, and we restrict our analyses to these in our example.

Direct methods (7)

Example – Carotenoids in Eye Disease Study (CAREDS):

Logistic regression analyses relating nuclear cataract to dietary lutein/zeaxanthin

Method	Estimated Odds Ratio†	95% CI	Z-value	Sample size ratio*
Unadjusted	0.75	(0.57,0.99)	-2.04	-
PC	0.65	(0.49,0.86)	-3.05	0.45
Howe	0.65	(0.49,0.85)	-3.11	0.43

† Comparing the 90th percentile to the 10th percentile

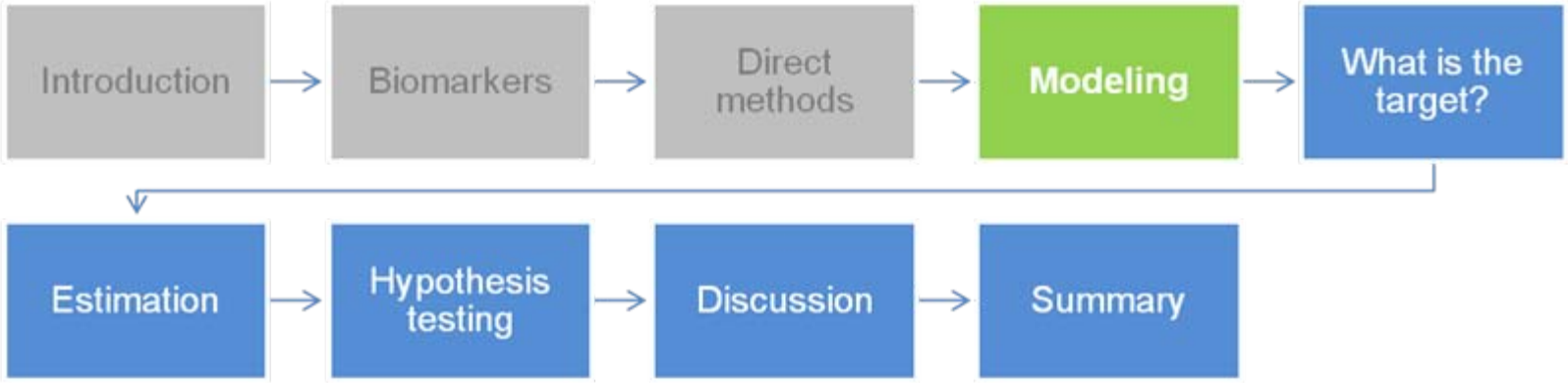
* Compared to the unadjusted method:
Sample size required is proportional to $1/z^2$
So sample size ratio is the inverse ratio of the z^2 values

Slide 28

This slide shows a table with the results of applying the usual regression analysis to the data in this study, based on diet report alone, compared to the principal components and Howe methods.

The second column shows the estimated odds ratios of nuclear cataract for lutein and zeaxanthin. Remember that I mentioned earlier that the units of the combined measures were not really meaningful. To overcome that, we have expressed the odds ratios to compare the risk of individuals at the 90^{th} percentile of measured intake with the risk of those at the 10^{th} percentile (of the control group). You can see that with the introduction of the marker information, either using principal components or Howe's method, the estimated odds ratios become stronger. And in the fourth column you can see that they are more highly statistically significant, with larger negative z-values.

From these z-values one can estimate the sample size savings that could accrue from use of the marker information. This last column tells us that one could obtain approximately the same power as a study with just self-report data from a study that had about 45 percent of the number of participants in that study but that included marker data on all participants as well as self-report data.


Introduction
Biomarkers
Direct methods
Modeling
What is the target?
Estimation
Hypothesis testing
Discussion
Summary

MODELING

Slide 29

Although the previous example shows that the direct approach to combining self-reports and markers can yield useful results, to get the full benefits of adding the marker information, one needs to take a modeling approach.

Modeling the intake-marker-disease relationship (1)

Disadvantages of the direct methods:

1. They do not always increase statistical power, and sometimes decrease it*
2. The estimated odds ratios are attenuated
3. The combined measure (PC or Howe) does not have any recognized units

* *For example, when the marker is poorly correlated with intake, or has a weaker relationship with disease than the self-reported intake*

Slide 30

As I've already mentioned, the direct approach carries several disadvantages. Firstly, it does not always lead to increased statistical power, and sometimes its use can lead to further loss of power. For example, if the marker has a weaker relationship with the health outcome than does the self-report, then power will be lost by combining them.

Secondly, the estimated odds ratios are attenuated and no correction is made for that. And, thirdly, as I mentioned earlier, the combined measure of marker and self-report does not have any recognized units.

Modeling the intake-marker-disease relationship (1a)

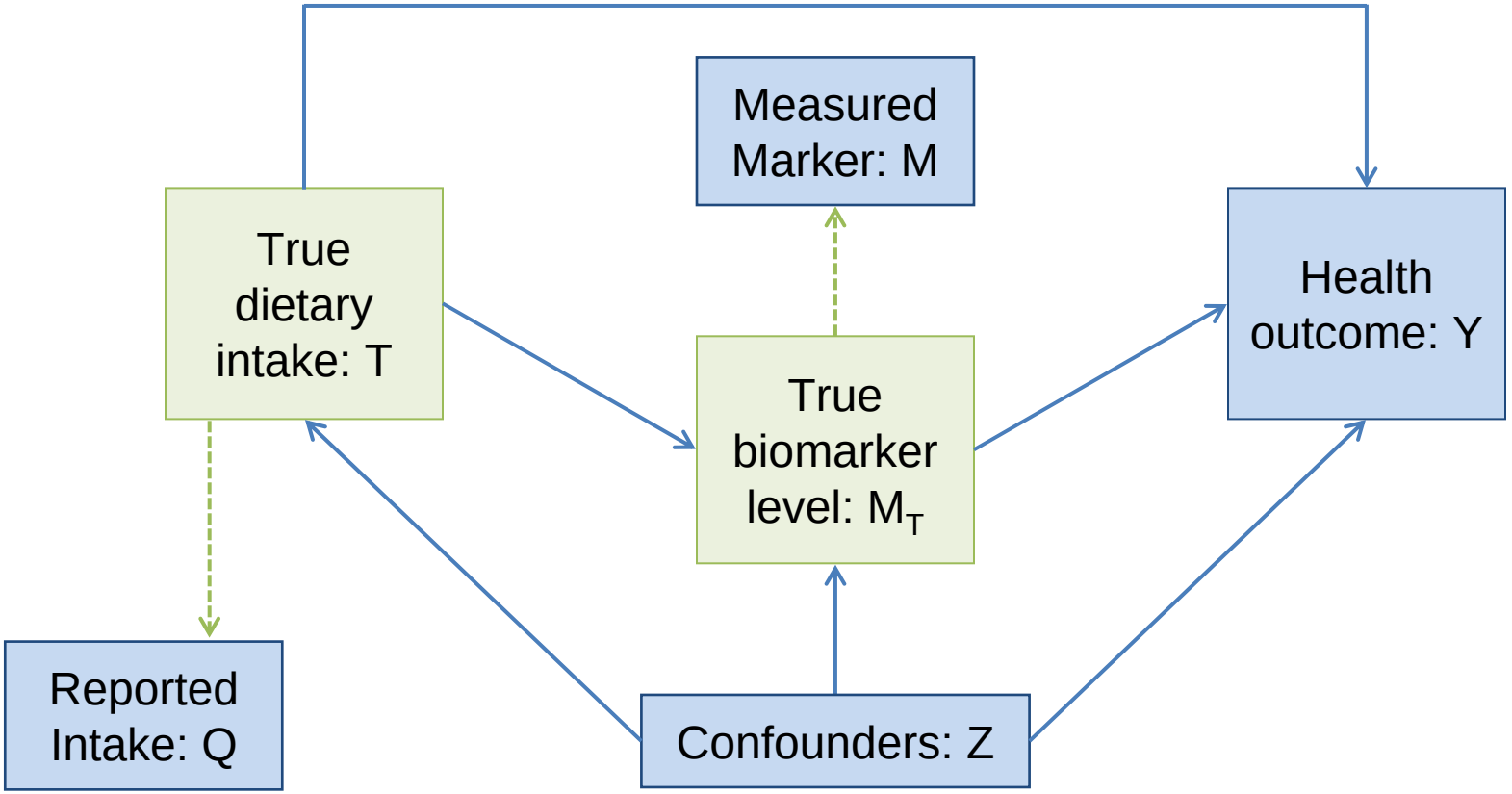
- To make progress in addressing these deficiencies, we have to consider models of diet, their markers and health outcomes, including aspects of causality

Slide 31

All of these difficulties can be overcome in a modeling approach. But in doing so, we have to consider some more complex models, including issues of causality. And that's what we will do now.

Modeling the intake-marker-disease relationship (2)

Causal pathways: dietary intake, biomarkers, and disease



Slide 32

This is a causal pathway diagram that describes relationships between dietary intake, biomarker and health outcome. Starting with the box for true dietary intake, T, and following its arrows, we see that the model postulates that the intake, T, causes a change in the biomarker, M, which in turn affects the health outcome, Y. In addition, the intake, T, can also affect the health outcome through pathways that do not involve the marker.

As with previous models, we have considered there are confounders, Z, and these can affect the true intake, the marker, and the health outcome.

Lastly, the true intake is not observed but is reflected by the self-report, Q, and, similarly, the true value of the biomarker, M subscript T, is not observed but is reflected by the measurement of the marker, denoted as before by M that we obtain from our assay.

Modeling the intake-marker-disease relationship (3)

Causal pathways:

Dietary intake, biomarkers, and disease

- Main assumptions:
  - Dietary intake T causally affects the biomarker level M_T
  - The biomarker level M_T may (at least partially) mediate the effect of dietary intake T on disease Y
  - The main confounders Z are known and are measured exactly

Slide 33

The pathway diagram you've just seen actually includes a number of important assumptions that we are making in adopting such a model. The main assumptions are:

- Firstly, the dietary intake causally affects the biomarker level.
- Secondly, the marker may at least partially mediate the effect of dietary intake on the health outcome. This seems reasonable when the marker is a serum or tissue level of the dietary component of interest.
- And, thirdly, the main confounders are known and measured exactly.

This last assumption is the strongest of the three. And, indeed, the greatest challenge to the reliability of results that use dietary concentration biomarkers is the question of whether we can identify the important confounders.

Modeling the intake-marker-disease relationship (4)

Statistical Models that describe the causal pathways:

1. Health outcome model:

$$\text{logit}(P(Y = 1)) = \alpha_0 + \alpha_1 T + \alpha_2 M_T + \alpha_Z Z$$

2. Marker-Intake model:

$$M_T = \gamma_0 + \gamma_1 T + \gamma_Z Z + \varepsilon_{MT}$$

3. Reported intake model:

$$Q = \beta_0 + \beta_1 T + \varepsilon_Q$$

4. Measured marker model:

$$M = M_T + \varepsilon_M$$

Slide 34

The pathway diagram that you've seen can be described by four different statistical models, all acting together. They are shown on this slide.

The principal model is the health outcome model, shown here. It describes the way that the dietary intake, T; the biomarker, M subscript T; and the confounders, Z, jointly act to influence the health outcome, Y, in the form of logistic regression, assuming here that the health outcome is a binary variable indicating disease status.

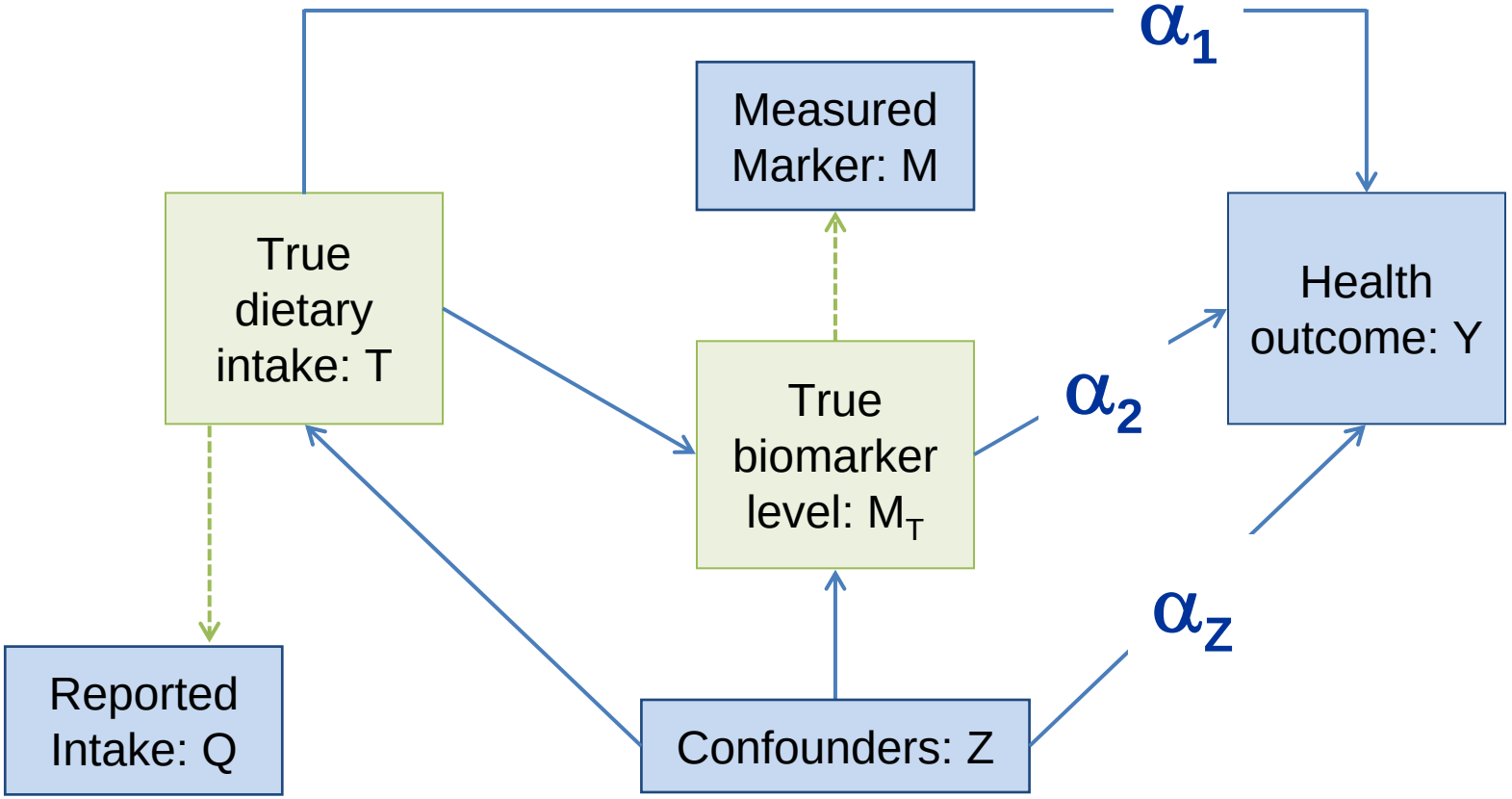
The second model describes how the dietary intake, T, and confounders, Z, influence the biomarker, M subscript T. Here, as in the health outcome model, it is important to know and measure the main confounders.

The third model describes how self-reported intake, Q, is related to true intake, T, and is similar to the measurement error models that were described in lectures 6 and 7 of this webinar series.

And the fourth model describes the measurement error model for the measured biomarker. Note that here we assume classical measurement error. In other words, we assume that the measured biomarker, M, is an unbiased measurement of the true biomarker level, M subscript T, with random errors that are independent of the true biomarker level.

Modeling the intake-marker-disease relationship (5)

1. Health outcome model

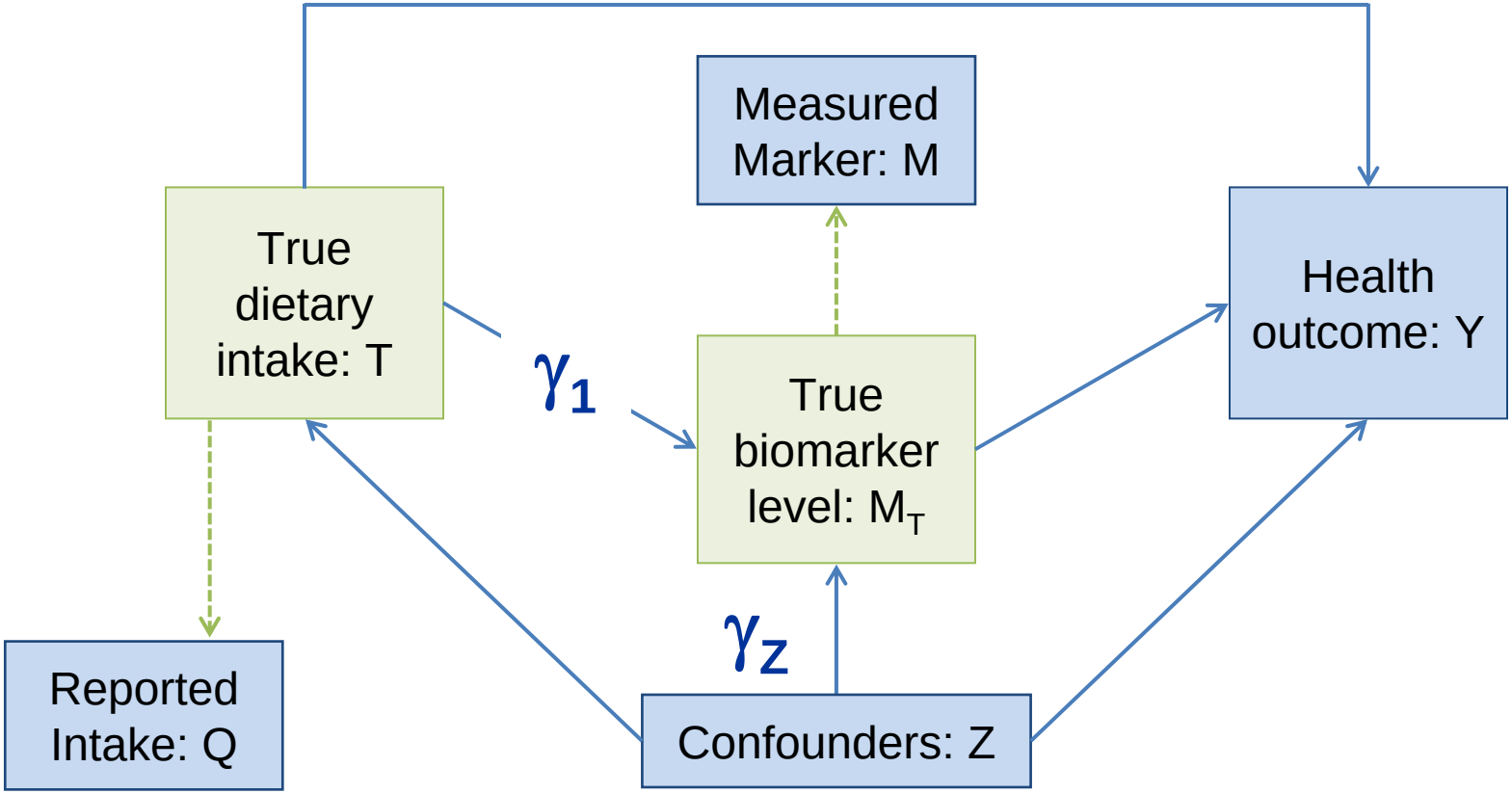


Slide 35

The next set of four slides will show which parts of the causal pathway diagram represents each of the four component models whose equations you have just seen. In this first diagram, you see the health outcome model component comprising all of the arrows leading to the health outcome box. Each of the arrows carries the coefficient of the variable in the regression model, so that the arrow from true intake, T, to health outcome, Y, carries the coefficient alpha_1, etc.

Modeling the intake-marker-disease relationship (6)

2. Marker-intake model

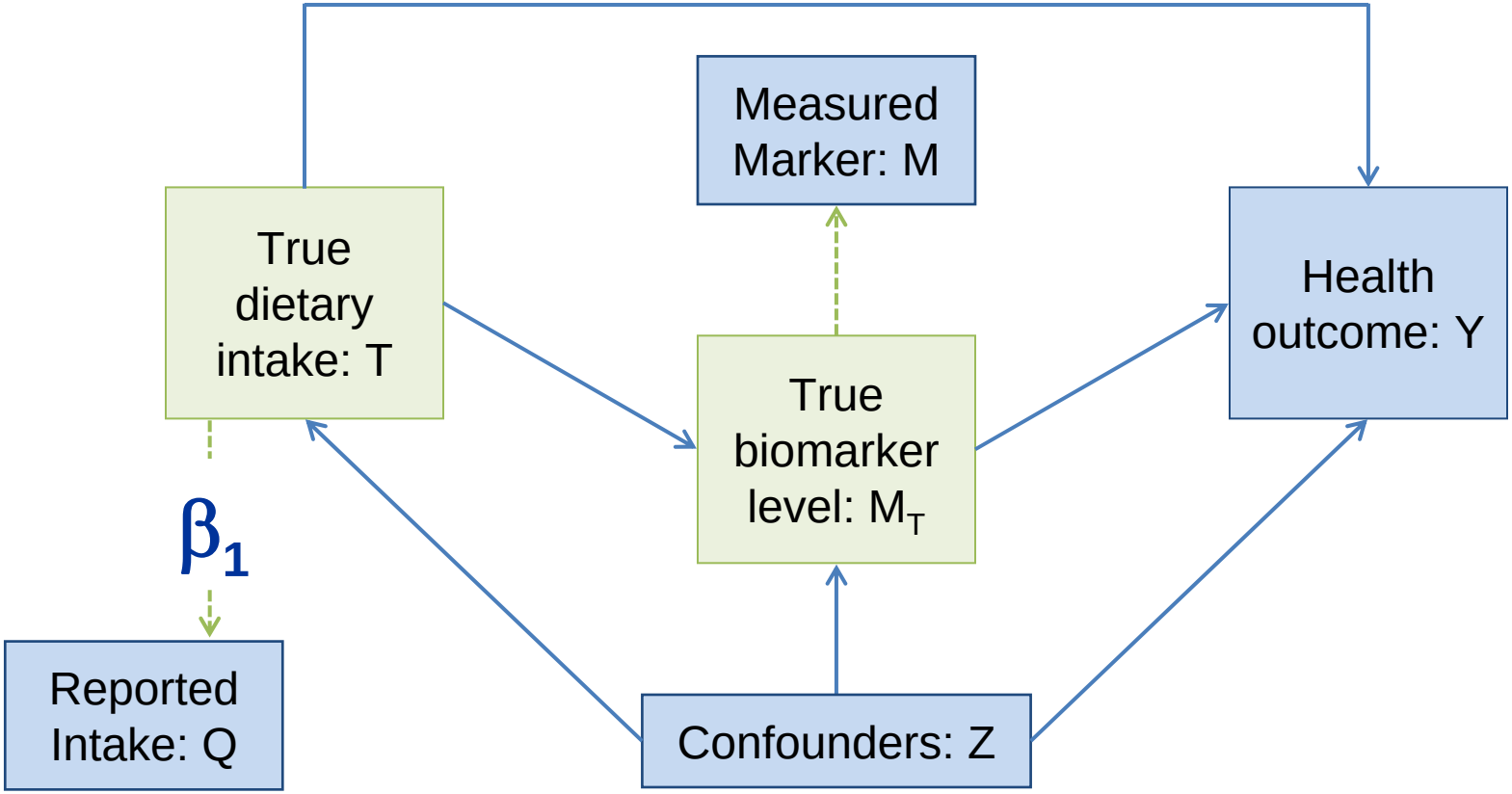


Slide 36

And this slide shows the marker-intake model with the arrows leading from true intake and the confounders to the true biomarker level.

Modeling the intake-marker-disease relationship (7)

3. Reported intake model

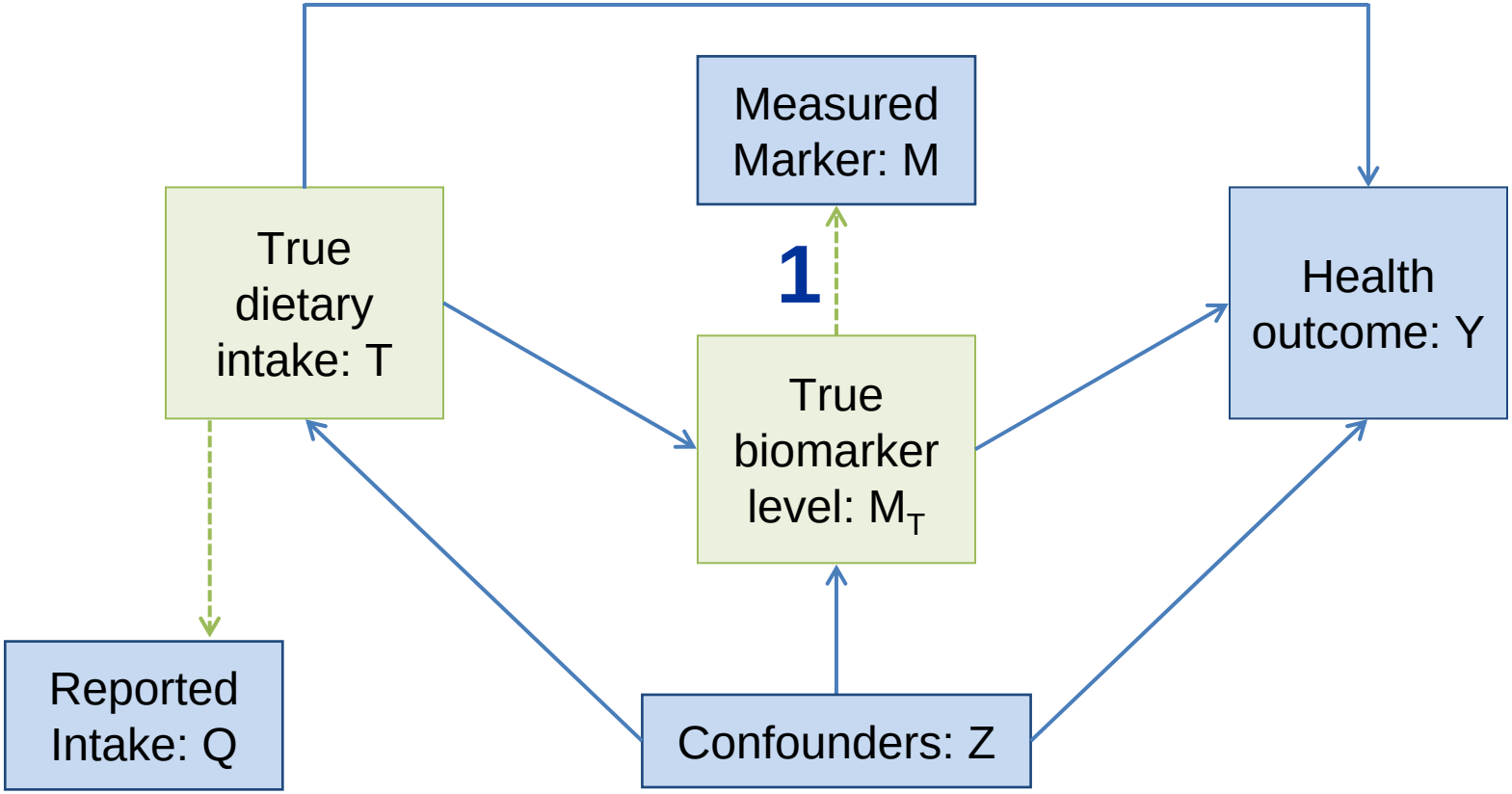


Slide 37

This slide shows the model for dietary self-report.

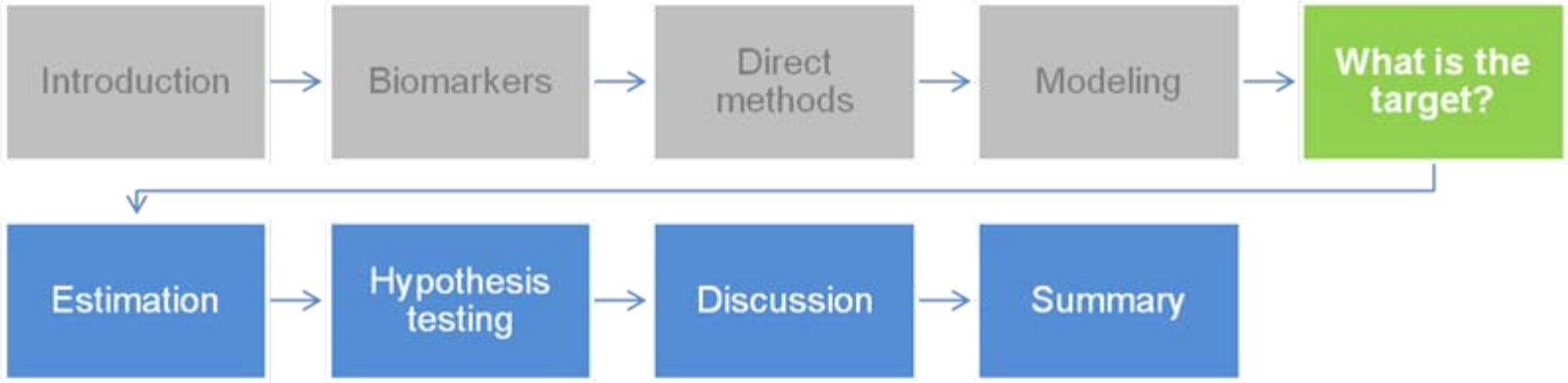
Modeling the intake-marker-disease relationship (8)

4. Measured marker model



Slide 38

And this last slide of the set shows the classical measurement error model for the measured biomarker level, the figure 1 denoting the coefficient for M subscript T in that model.



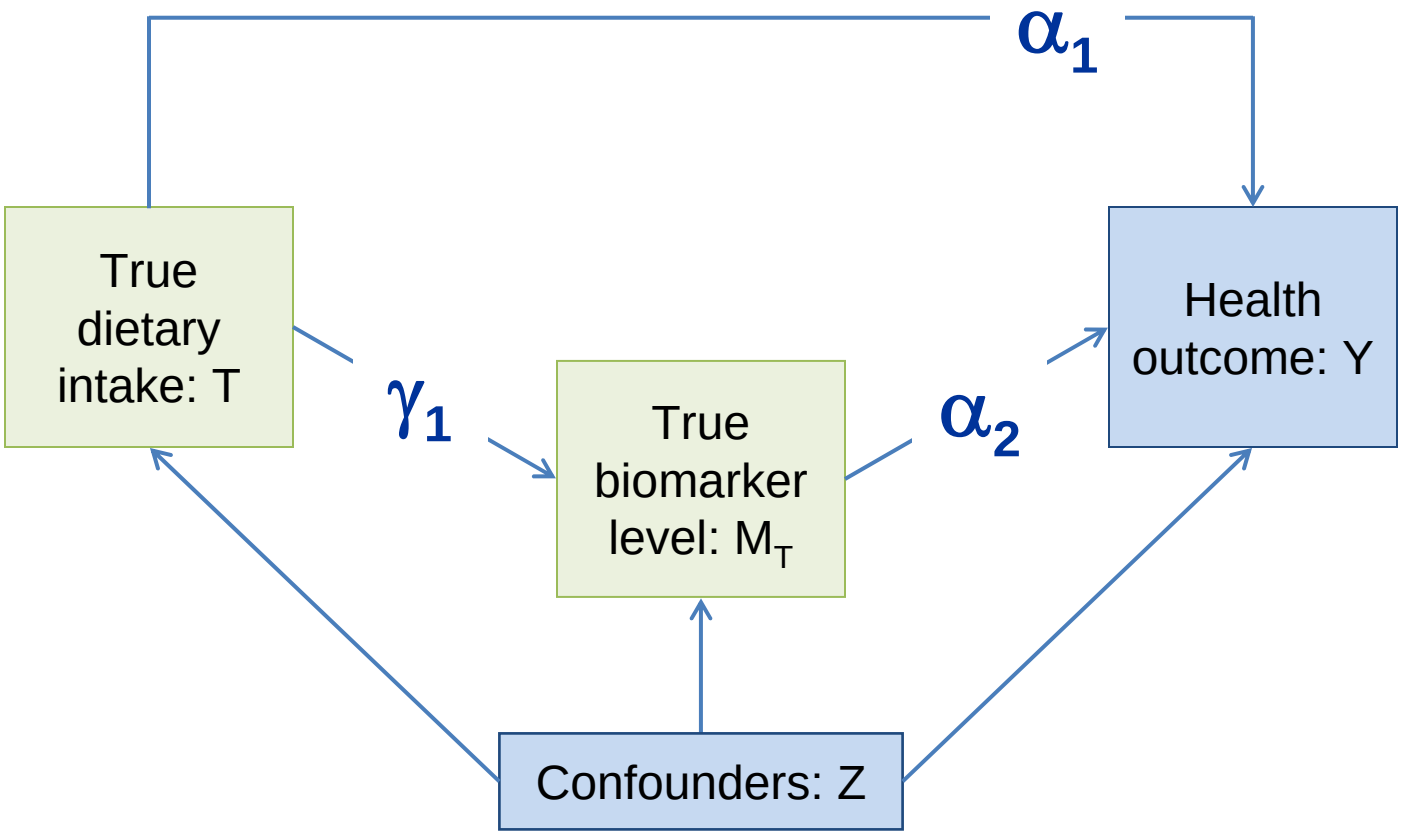
WHAT IS THE TARGET?

Slide 39

Having described the postulated causal model that underlies our measurements, we must now address what risk parameter we are interested in estimating. Because of the complex causal model, this is not quite as straightforward as in simpler situations where we usually want to estimate a simple odds ratio or relative risk.

What is the target? (1)

Model without measurement error



Direct (non-mediated) effect of diet on disease = α_1

Indirect (mediated) effect of diet on disease = $\gamma_1\alpha_2$

Total effect of diet on disease = $\alpha_1 + \gamma_1\alpha_2$

Slide 40

To make the discussion simpler, we'll for the moment forget that we have error-prone measurements and suppose that we can measure true intake and the true biomarker level. This slide shows such a model. The coefficients in the health outcome model (the alphas) and the marker-intake model (the gamma) are shown beside the relevant arrows, as before. Also, for simplicity, we have omitted the coefficients of the confounders.

You can see that there are two separate pathways from intake, T, to health outcome, Y. One goes directly from T to Y, and is known as the direct effect. Its magnitude is given by the coefficient alpha_1.

The other pathway goes through the marker, and is known as the indirect or mediated effect. Its magnitude is the product of the coefficient for the arrow going from T to M subscript T and the coefficient for the arrow going from M subscript T to Y; in other words, gamma_1 times alpha_2.

According to the model, when a unit change is made to intake, T, the total effect on outcome Y is the sum of the direct and indirect effects; in other words, alpha_1 plus gamma_1 times alpha_2. We call this the total effect of diet on health outcome. For a logistic health outcome model, it is the log odds ratio for a change in intake of one unit.

What is the target? (2)

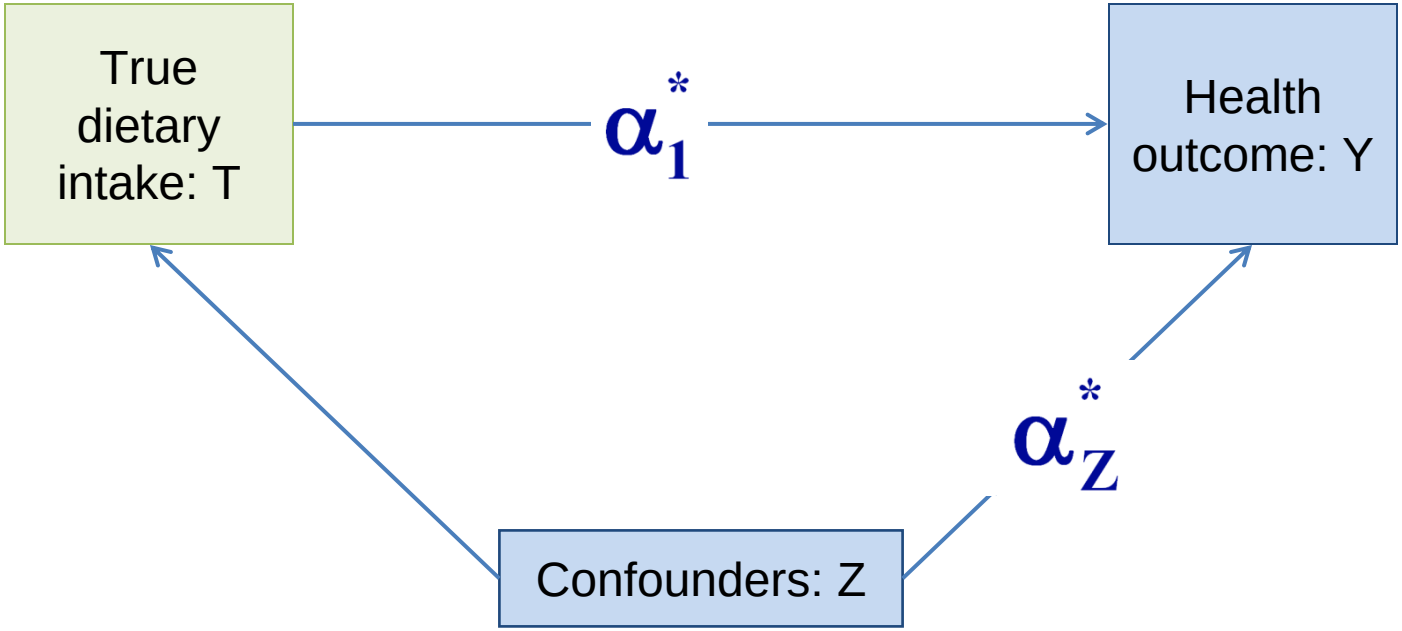
- Total effect of diet on disease $= \alpha_1 + \gamma_1 \alpha_2$
- We will denote this quantity by α_1^*
- Our object is:
  - to estimate α_1^*
  - and, to test whether $\alpha_1^* = 0$

Slide 41

It is this total effect that we are usually most interested in estimating. We will denote it by alpha_1_ star. Our object is to estimate it and also test whether or not it is equal to zero.

What is the target? (3)

Note that when there is no measurement error we can estimate α_1^* by dropping M_T from the model

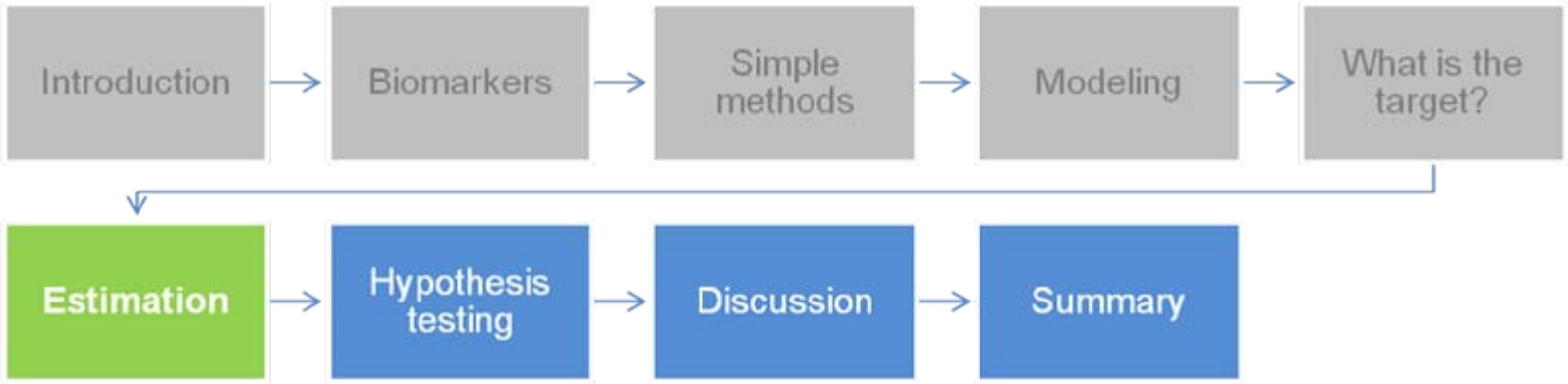


$$\text{logit}(P(Y=1)) = \alpha_0^* + \alpha_1^* T + \alpha_Z^* Z$$

Slide 42

When there is no measurement error, there is a much simpler way of estimating the direct effect than estimating each of the parameters alpha_1, gamma_1, and alpha_2, and then calculating alpha_1 plus gamma_1 times alpha_2, according to the formula shown earlier. Instead, all we have to do is ignore the biomarker and use a simple model relating dietary intake to health outcome, retaining the confounders, as shown in this slide. If we do that, then the coefficient for the intake is indeed the total effect of intake on health outcome.

Unfortunately, when there is measurement error in the dietary intake, matters are no longer so simple, as we will see.


Introduction
Biomarkers
Simple methods
Modeling
What is the target?
Estimation
Hypothesis testing
Discussion
Summary

ESTIMATION

Slide 43

So in the next section we will consider various methods of estimating our target parameter, the log odds ratio, for the total effect of the dietary intake on disease risk.

Estimating the total dietary effect (1)

Methods:

1. **Unadjusted:**
   Regress health outcome on Q and Z and take the coefficient of Q

2. **Regression Calibration**
   Regress health outcome on E(T|Q,Z) and Z and take the coefficient of E(T|Q,Z)

3. **"Enhanced" Regression Calibration:**
   Regress health outcome on E(T|Q,M,Z) and Z and take the coefficient of E(T|Q,M,Z)

4. **New method:**
   Regress health outcome on E(T|Q,M,Z), $E(M_T|Q,M,Z)$ and Z and calculate $\alpha_1 + \gamma_1 \alpha_2$

Slide 44

We will consider four different estimates of our target parameter.

The first, that we call the unadjusted method, is the usual one of entering the self-reported intake, Q, together with confounders, Z, into the logistic regression model. We learned in Lecture 6 that this gives attenuated estimates of the log odds ratio, and also loses power, because of the measurement error in Q.

The second estimate comes from the usual regression calibration method , where we enter into the logistic regression not the self-reported intake, Q, but the expected (or predicted) true intake given the values of Q and the values of the confounders, which we denote by the symbol E (for expectation) of true intake, T, given Q and Z. The coefficient of this variable in the logistic regression then serves as our estimate of the log odds ratio. We learned in lecture 7 that this method will give us an unbiased—that is, deattenuated—estimate of the log odds ratio, but it will usually not recover any of the lost statistical power.

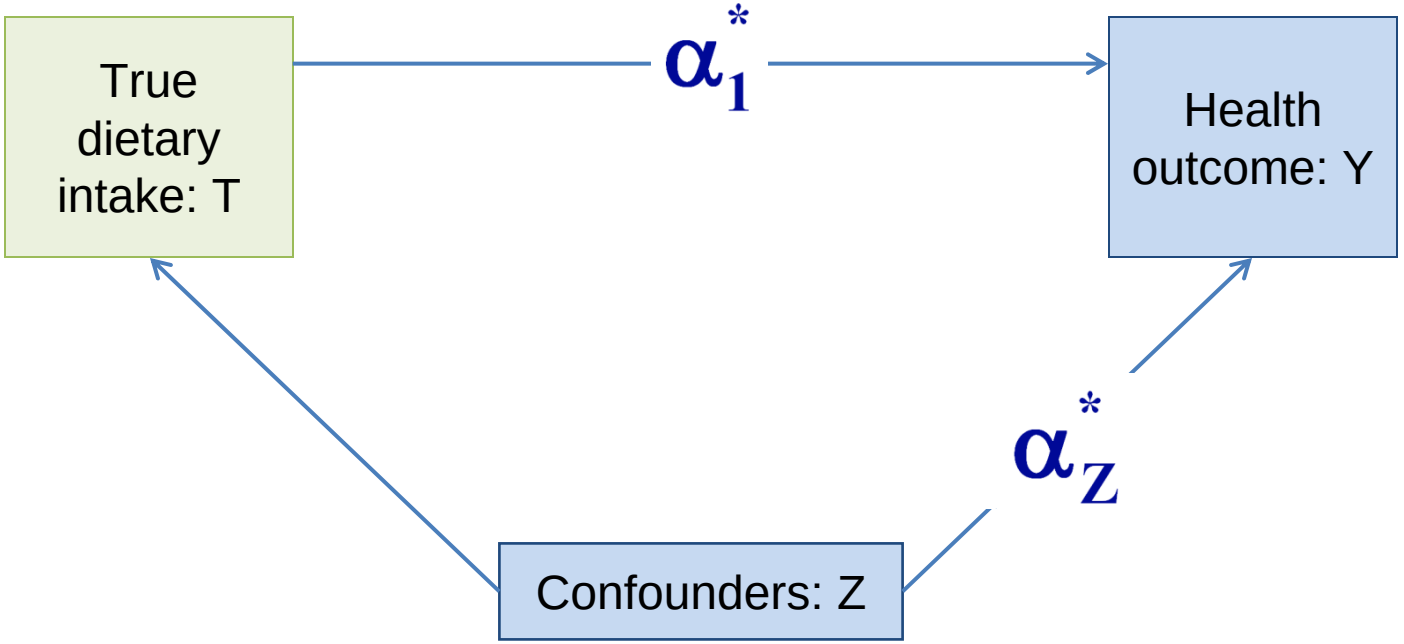
The third method we call enhanced regression calibration, and it is very similar to the usual regression calibration method, except that we use the biomarker value, M, as well as the self-report, Q, and confounders, Z, in order to predict the true intake, T. This is denoted by the expression E of T given Q, M, and Z. As before, the coefficient of this variable in the logistic regression gives the estimate of our targeted log odds ratio. This method was suggested by Prentice and colleagues in a paper in the *American Journal of Epidemiology* in 2009.

The fourth method, which we call the new method, uses the full health outcome model that we considered in an earlier section, which includes as explanatory variables the dietary intake, the biomarker, and confounders. To obtain unbiased estimates of odds ratios from this model, we use the regression calibration version of it; that is, we enter for the dietary intake the predicted true intake given self-reported intake, measured biomarker, and confounders, and for the biomarker we enter the predicted true biomarker given the measured biomarker, self-reported intake, and confounders.

Having obtained the estimates of the log odds ratios alpha_1 and alpha_2 for the dietary intake and biomarker, respectively, we then calculate the log odds ratio of the targeted parameter as alpha_1 plus gamma_1 times alpha_2, as explained previously.

Estimating the total dietary effect (2)

The first three methods take the following model and substitute different quantities for T: Q or E(T|Q,Z) or E(T|Q,M,Z)



$$\text{logit}(P(Y=1)) = \alpha_0^* + \alpha_1^* T + \alpha_Z^* Z$$

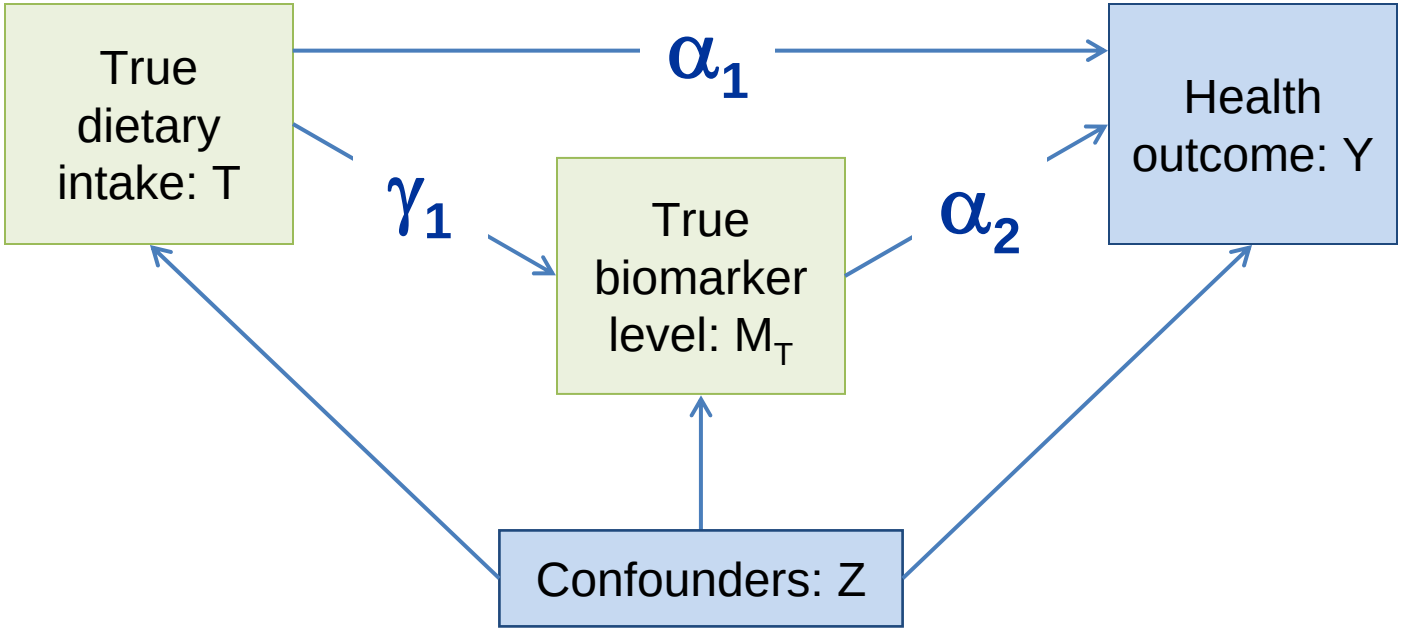
Slide 45

The slide illustrates the first three of these methods graphically. Essentially, we are leaving the biomarker out of the model and relating true dietary intake to the disease, with control for confounders. The three methods differ only in the measure used for the true dietary intake, T.

Estimating the total dietary effect (3)

The new method takes the full model and substitutes E(T|Q,M,Z) for T and $E(M_T|Q,M,Z)$ for M_T

The parameters α_1 and α_2 are estimated and then $\alpha_1 + \gamma_1 \alpha_2$



$$\text{logit}(P(Y=1)) = \alpha_0 + \alpha_1 T + \alpha_2 M_T + \alpha_Z Z$$

Slide 46

The new method uses the full causal model that we described earlier, estimates the relevant log odds ratio parameters, the alphas and the gamma, and computes the targeted log odds ratio in the manner already described.

Estimating the total dietary effect (4)

Which of these methods estimates α_1^* without bias?

- **Unadjusted**
  - Unbiased only if Q has no measurement error
- **Regression Calibration**
  - Unbiased
- **"Enhanced" Regression Calibration**
  - Unbiased only if marker does not mediate the effect of diet (α_2=0)
- **New method**
  - Unbiased

Slide 47

So we first need to understand which of these methods gives us biased estimates of our target parameter and which gives us unbiased estimates. This question can be answered by some theoretical statistical work, and we have also checked it out with simulations.

The theory tells us that the unadjusted method is biased unless there is no measurement error in the self-report, as we have already learned. It also tells us that usual regression calibration is unbiased, which we have also already learned.

The new results are, firstly, that enhanced regression calibration gives biased estimates of the targeted log odds ratio whenever there is some mediation of the dietary effect through the biomarker, and that the estimate obtained from the new method is unbiased.

Implementing the estimation (1)

- Data available: Q, M, Z
  Example – (CAREDS):
  - Y: eye cataract (yes/no)
  - Q: log FFQ-reported lutein plus zeaxanthin
  - M: log serum level of lutein plus zeaxanthin
  - Z: age (y)
    smoking (0=never, 1=past, 2=current)

Slide 48

So far, we have been talking quite theoretically. Let's see how these methods are implemented in a real example. We're going back to the CAREDS study that I described to you earlier when we were looking at the simpler direct approach to combination of self-reports and biomarkers.

Just to remind you, the disease outcome, Y, is nuclear cataracts of the eye; the self-reported intake, Q, is the logarithm of a food frequency questionnaire report of lutein and zeaxanthin intake combined; the biomarker, M, is the logarithm of their combined serum levels; and the confounders are age and smoking.

Implementing the estimation (2)

Methods:

1. **Unadjusted:**
   Directly implemented: logistic regression of Y ~ Q, Z

2. **Regression Calibration:**
   First determine E(T|Q,Z); then Y ~ E(T|Q,Z), Z

3. **"Enhanced" Regression Calibration:**
   First determine E(T|Q,M,Z); then Y ~ E(T|Q,M,Z), Z

4. **New method:**
   First determine E(T|Q,M,Z), $E(M_T|Q,M,Z)$;
   then Y ~ E(T|Q,M,Z), $E(M_T|Q,M,Z)$, Z ;
   then calculate $\alpha_1 + \gamma_1 \alpha_2$

Slide 49

And here, once more, are the definitions of the four methods that we have just described. To implement the second, third, and fourth methods, we have to compute the predicted true intake, T, using just the self-report and confounders, or, for enhanced regression calibration, using also the biomarker.

For the new proposal we also have to predict the true biomarker level. These predictions are made through what we call calibration equations.

Implementing the estimation (3)

Determining the calibration equations:

- Usually one needs feeding studies to relate biomarker to dietary intake, and population studies of the biomarkers and the dietary instruments to obtain population means and SDs

CAREDS:

- **Feeding studies:**
  Van het Hoff et al, *Am J Clin Nutr* 1999, 70:261
  Brevik et al, *Eur J Clin Nutr* 2004, 58:1166

- **Population studies:**
  Delcourt et al, *Invest Opthalmol Vis Sci* 2006, 47:2329
  Dixon et al, *J Nutr* 2006, 136:3054
  Mares et al, *Am J Clin Nutr* 2006, 84:1107

Slide 50

Whether these calibration equations can be built will depend on the external information that is available. Usually, what are required are one or more feeding studies to relate the biomarker level to the true intake, and one or more population-based studies to obtain population means and standard deviations of biomarker levels, and self-reported intakes. Fortunately, these are available in the case of lutein and zeaxanthin, and references to these studies are shown here.

Implementing the estimation (4)

Determining the calibration equations *(cont'd)*

- Using data from these studies, we built three models (all measurements were transformed to the log scale)

Marker-intake model:
$M_T = 5.29 + 0.60T + e$, $var(e) = 0.10$

Reported intake model:
$Q = 0.35 + 0.71T + e_Q$, $var(e_Q) = 0.36$

Measured marker model:
$M = M_T + e_M$, $var(e_M) = 0.05$

* *Note that in these models it is assumed that the confounders Z have no bearing on measurement error*

Slide 51

Using the information reported from these studies, we first built three models describing the relation of marker to true intake and measurement error in the self-report and in the measured biomarker.

A while ago in this lecture, you may remember that when describing the full causal pathway graph underlying the relationships between dietary intake, biomarker, and disease, we saw that the graph actually comprised four models. The three models shown in this slide are actually specific versions of three of those four models, with the health outcome as yet being unspecified.

So we see here a marker-intake model, a self-reported intake measurement model, and a marker measurement error model. The parameters of these models, including the variances of the random error terms, are all based upon the information from the references shown in the previous slide.

Note that we assume that these models are not modified by the confounders in the health outcome model—age and smoking. Although we cannot be certain that this assumption is true, the available references did not provide information on this issue.

Implementing the estimation (5)

Determining the calibration equations *(cont'd)*

- The final step is to turn these measurement error models into calibration equations

Regression Calibration

$E(T \mid Q, Z) = 0.355\ Q + 0.00560\ age - 0.101\ smoking$

Enhanced Regression Calibration

$E(T \mid Q, M, Z) = 0.242\ Q + 0.515\ M + 0.00692\ age - 0.0954\ smoking$

New Method

$E(T \mid Q, M, Z) = 0.242\ Q + 0.515\ M + 0.00692\ age - 0.0954\ smoking$

$E(M_T \mid Q, M, Z) = 0.051Q + 0.769M + 0.00059\ age - 0.0023\ smoking$

Slide 52

Once those three models have been constructed, it is a relatively simple statistical task to convert the models into calibration equations using also information from the CAREDS study itself on the relation between the confounders and the self-reported intake and measured biomarker. The calibration equations are shown in this slide.

You can see here that each is a linear equation including self-report, Q; confounders, Z; and except for usual regression calibration, also the biomarker, M.

It can be noticed in the equations for the new method that when predicting true dietary intake, the biomarker and the self-reported intake are both influential; you can see that their coefficients are large (0.515 and 0.242, respectively). However, when predicting the true biomarker level, the measured biomarker level is highly important (with a coefficient of 0.769) and the self-reported intake has little influence (with a coefficient of 0.051). The confounders have relatively little contribution in these predictions.

Implementing the estimation (6)

Results

- Logistic Regression Analyses Relating Nuclear Cataracts to Dietary Lutein and Zeaxanthin in the CAREDS study

Method	Log Odds Ratio	Standard Error	z-value
Unadjusted	-0.16	0.08	-2.07
Regression Calibration	-0.46	0.22	-2.07
Enhanced Regression Calibration	-0.51	0.16	-3.15
New Method	-0.44	0.22	-2.00

Slide 53

This table shows the results obtained from the four methods when applied to the CAREDS data. You can see in the first row of the table that the estimated log odds ratio from the unadjusted method is much smaller than those from the other three methods, displaying the attenuation that we would expect to see.

The two unbiased methods, usual regression calibration and the new method, yield similar estimates, whereas enhanced regression calibration gives a slightly larger value.

We will consider the other columns of the table in the next section.

Implementing the estimation (7)

Conclusion

- For this study either regression calibration or the new method could be used, since theoretically both are unbiased.
- Therefore, the point estimate for the log odds ratio could be taken as **-0.45** (midway between the two estimates). This translates into an odds ratio of **0.73*** corresponding to a doubling of lutein/zeaxanthin intake.

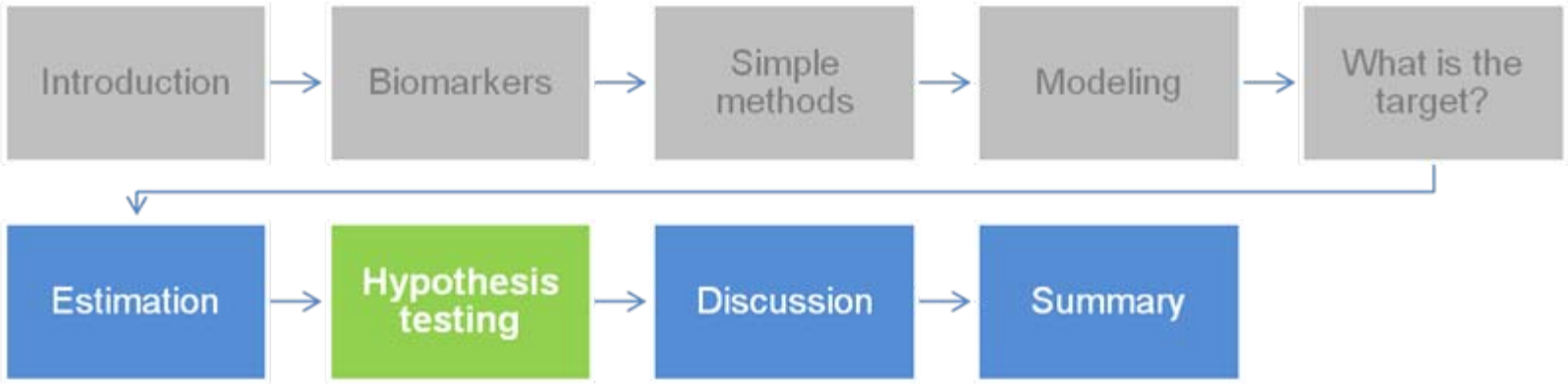
* $0.73 = \exp(-0.45 \times \ln(2))$

Slide 54

So the conclusions from this analysis, in relation to the strength of the association between lutein and zeaxanthin intake and eye cataracts, are as follows.

Firstly, both usual regression calibration and the new method are for this study theoretically unbiased and produce similar estimates. We could therefore take the estimate of the log odds ratio as -0.45, midway between the two estimates. And this translates into an odds ratio of 0.73 associated with a doubling of the lutein/zeaxanthin intake.

The calculation for this translation to the odds ratio associated with a doubling of the intake is shown on the bottom line of the slide.



HYPOTHESIS TESTING

Slide 55

So far, we have concentrated on how to use modeling to estimate the targeted odds ratio. We now turn to hypothesis testing, and particularly to the question of statistical power to detect the odds ratio as statistically significant.

Testing the null hypothesis of a zero total dietary effect (1)

- Besides estimating the odds ratio, we also want to test whether $\alpha_1^* = \mathbf{0}$
- The four methods of estimation each lead to a test of this null hypothesis:
  - Compare z = estimate/SE with the standard normal distribution

Slide 56

We want to test whether the targeted log odds ratio is zero or not. The four methods of estimation that we have considered all lead naturally to a method of testing this hypothesis. All that we have to do to convert the estimate into a test is to consider the ratio of the estimate to its standard error, and compare the ratio to the standard normal distribution.

Testing the null hypothesis of a zero total dietary effect (2)

- **Example:**
  Logistic Regression Analyses Relating Nuclear Cataracts to Dietary Lutein and Zeaxanthin in the CAREDS study

Method	Log Odds Ratio	Standard Error	z-value	P-value (2-sided)
Unadjusted	-0.16	0.08	-2.07	0.038
Regression Calibration	-0.46	0.22	-2.07	0.038
Enhanced Regression Calibration	-0.51	0.16	-3.15	0.002
New Method	-0.44	0.22	-2.00	0.046

Slide 57

Here is the same table that I showed you just before for the CAREDS study example. Previously, we concentrated on the estimates of the log odds ratio in the second column. The third column shows the standard errors of each estimate, and the fourth column, their ratio.

I have added an extra column at the end that shows the P-value. We'll consider the actual values in just a minute.

Testing the null hypothesis of a zero total dietary effect (3)

- Which of these methods is valid?
  - i.e., yields a test that has the correct probability of rejecting the null hypothesis when it's true
- Answer:
  - All four methods yield valid tests!
- Why?
  - Because each estimation method is unbiased when the total dietary effect α_1^* is zero, even though the unadjusted and enhanced RC methods are otherwise biased

Slide 58

But, first, we have to ask the question: Which of these tests is a valid test of the null hypothesis? In other words, for which of these tests is the probability of rejecting the null hypothesis actually equal to more or less 5 percent when the nominal 5 percent level is used and the null hypothesis is true? One might expect that the test would be valid only for those methods that give an unbiased estimate of the log odds ratio.

The surprising answer is that all four tests are valid. And the reason is that when the null hypothesis is true, and the log odds ratio is truly zero, then all of the estimates will be unbiased and will on average equal zero, even the unadjusted estimate and enhanced regression calibration estimate that are otherwise biased.

Testing the null hypothesis of a zero total dietary effect (4)

- Since all of these methods of testing the null hypothesis are valid, which is the most powerful?
- Answer:
  - The enhanced RC method

Slide 59

Since all four tests are valid, the next question to ask is: Which of them is statistically the most powerful? In other words, which of them is the most likely to detect an odds ratio that is truly different from unity?

The simple answer to this question is that the enhanced regression calibration method is the most powerful of the four tests.

Testing the null hypothesis of a zero total dietary effect (5)

- Logistic Regression Analyses Relating Nuclear Cataracts to Dietary Lutein and Zeaxanthin in the CAREDS study:
  - The method leading to the largest z-value and smallest P is Enhanced Regression Calibration

Method	Log Odds Ratio	Standard Error	z-value	P-value (2-sided)
Unadjusted	-0.16	0.08	-2.07	0.038
Regression Calibration	-0.46	0.22	-2.07	0.038
Enhanced Regression Calibration	-0.51	0.16	-3.15	0.002
New Method	-0.44	0.22	-2.00	0.046

Slide 60

We see this fact reflected in the table of results for the CAREDS study example. The enhanced regression calibration method leads to the most extreme of the z-values and the smallest p-value. The intuitive reason for this gain in power over usual regression calibration is that using the marker in addition to self-reported intake to predict true dietary intake has increased the precision of our predictions, and the more precise estimates of true dietary intake then increase the power of detecting the association with disease.

Testing the null hypothesis of a zero total dietary effect (6)

- **Sample size savings:**
  Estimated sample size required is proportional to $1/z^2$
  Estimated sample size ratio for Enhanced RC versus RC = $(2.07/3.15)^2 = 0.43$
  Required sample size is reduced by >50%!

Method	Log Odds Ratio	Standard Error	z-value	P-value (2-sided)
Unadjusted	-0.16	0.08	-2.07	0.038
Regression Calibration	-0.46	0.22	-2.07	0.038
Enhanced Regression Calibration	-0.51	0.16	-3.15	0.002
New Method	-0.44	0.22	-2.00	0.046

Slide 61

If we want to translate this advantage in power to potential savings in sample size, then we can consider the inverse ratio of the squares of the z-values for the two methods. For usual regression calibration the z-value is -2.07 and for enhanced regression calibration it is -3.15. The inverse ratio of their squares is 0.43, meaning that it is estimated that in order to achieve the same statistical power, a study that incorporates the marker in the prediction of true intake would require only 43 percent of the sample size required in a study that does not incorporate the marker information.

Computer simulations that we have conducted indicate over a range of circumstances that enhanced regression calibration is the most powerful approach, and these results are presented in a paper in press in the *American Journal of Epidemiology*.

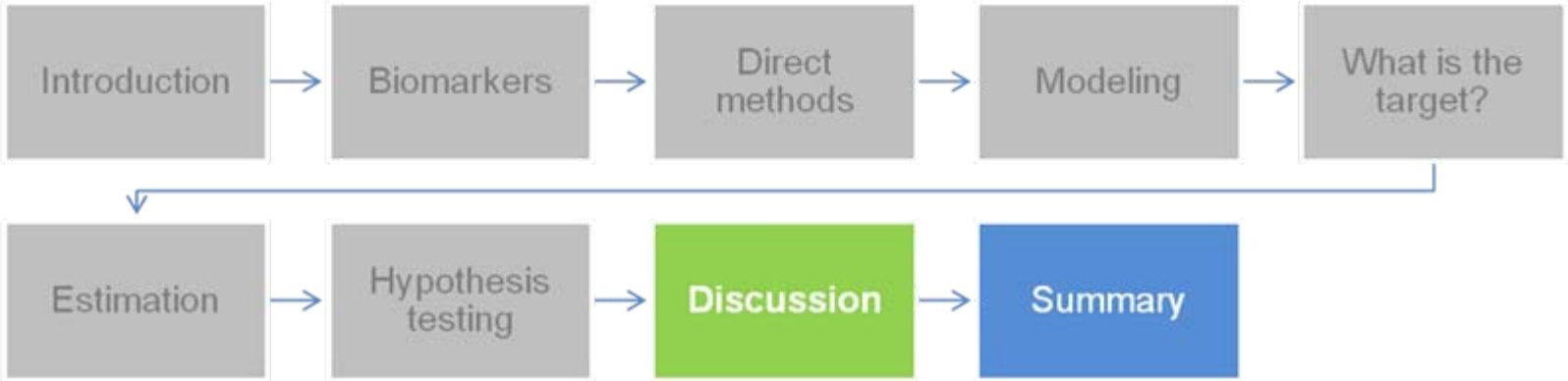
Testing the null hypothesis of a zero total dietary effect (7)

- Recommended overall strategy:
  - Estimate the odds ratio using an unbiased method—either the **new method** or, for cases like CAREDS, **the RC method**
  - Test the odds ratio using **Enhanced RC** that incorporates marker information and thus increases power

Slide 62

So our overall strategy that we recommend based on this work is that for estimation the targeted odds ratio should be estimated by usual regression calibration or the new method we propose, and for testing the null hypothesis of no association between dietary intake and health outcome, the enhanced regression method should be used.

One question that arises from these results is: Why should we bother using the new method when usual regression calibration gives an unbiased estimate and enhanced regression calibration gives a more powerful test? In fact, we will see in the next section that there is a situation where the new method is the only one of the four methods that provides an unbiased estimate of the risk parameter.


Introduction
Biomarkers
Direct methods
Modeling
What is the target?
Estimation
Hypothesis testing
Discussion
Summary

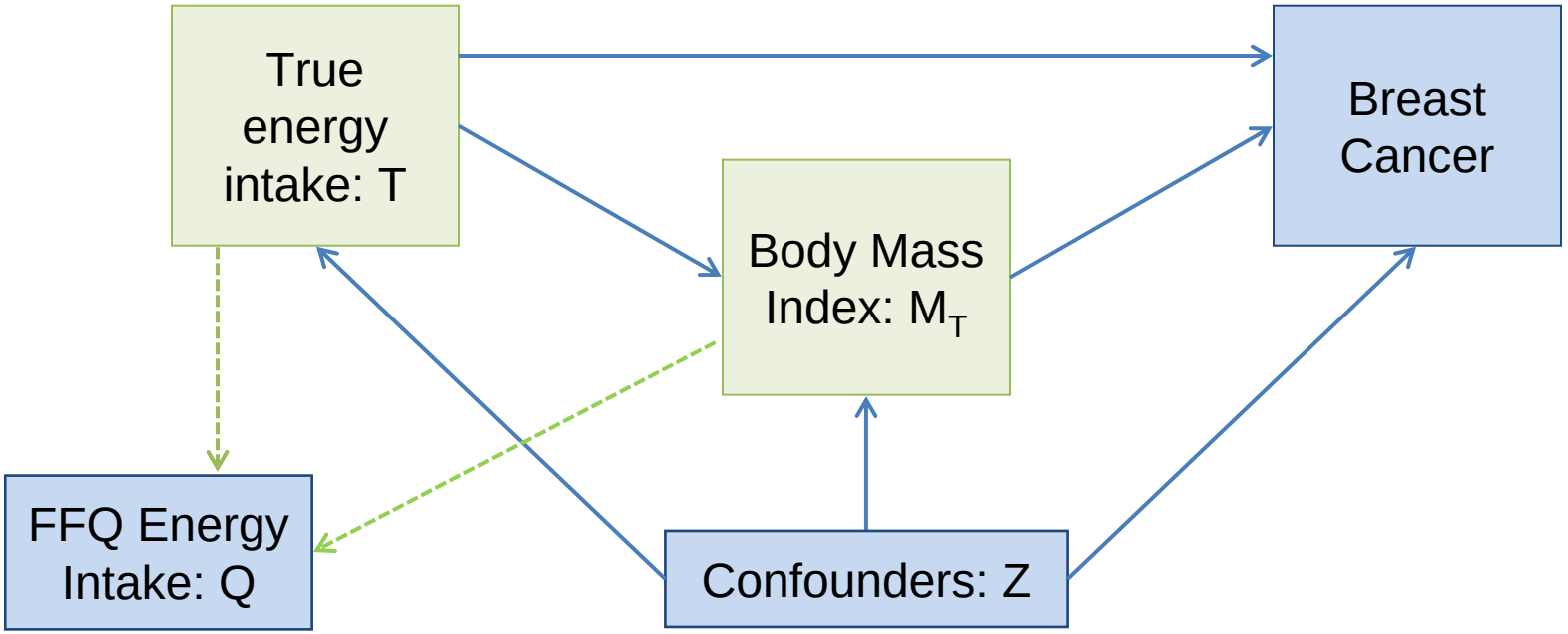
DISCUSSION

Slide 63

Actually, there are a number of important issues that need to be considered when thinking about using biomarkers in the manner we have described.

An example where the RC method gives biased estimates

- Prentice et al: *Am J Epidemiol*. 2009;169:977, consider using body mass index (BMI) to help predict energy intake
- The "biomarker" BMI is related to error in the FFQ energy report. Obese persons under-report more



Slide 64

The first two points are raised by some work by reported Prentice and colleagues in the *American Journal of Epidemiology* in 2009. They considered a situation where they wished to use body mass index rather than a typical biomarker in their prediction of true dietary intake. Their context was the association of total energy intake to a variety of cancers. In the diagram here, we take breast cancer as one of their primary interests.

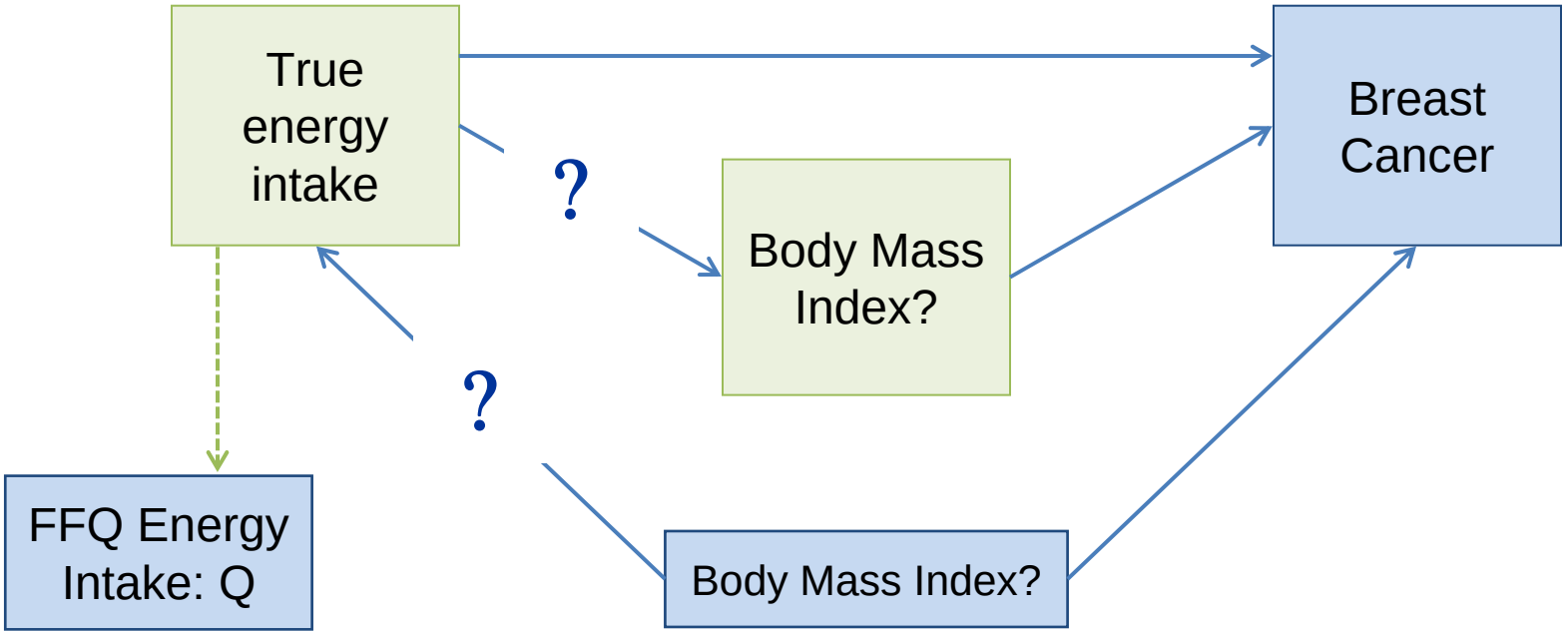
The causal pathway diagram shown here is almost the same as the one that we have considered in this lecture, but there is one important difference. Notice that there is now an arrow leading from the true marker (or true body mass index) to the self-reported intake. This is because of the often-observed phenomenon that obese persons tend to underreport their dietary intake more than the nonobese. In our models we have assumed no relation between the biomarker value and the measurement error.

In these circumstances, as indicated in the title of this slide, the usual regression calibration method gives a biased estimate of the targeted log odds ratio. So when this situation occurs, the only unbiased method of estimation available to us is the new method, the fourth and last of the estimation methods that we described earlier.

I should also mention that all four methods of testing the null hypothesis remain valid in this situation and that enhanced regression calibration remains the most powerful.

Sometimes the biomarker may be a confounder as well as a mediator!

- Prentice et al: *Am J Epidemiol*. 2009;169:977
- The “biomarker” BMI could affect energy intake or could mediate its effect. In such circumstances, it is unclear what to do



Slide 65

Another point raised by the application considered by Prentice et al. is that there may sometimes be uncertainty regarding the direction of the causality. In the example where body mass index is in the position of the biomarker, it is not entirely clear whether it acts as a mediator of the effect of energy intake on breast cancer or whether it acts as a confounder. This dilemma is portrayed by the diagram in this slide.

In this case it is unclear how to proceed. If body mass index were a mediator, then our target risk parameter should be the total effect of dietary intake on breast cancer and we could use the methods described in this lecture. However, if body mass index were a confounder we should be interested in a different target risk parameter—the effect of diet adjusted for body mass index—and then the estimation procedure would be different. If we cannot resolve the causal pathway question, then we cannot decide what to estimate!

Costs of including a biomarker

- The methods described (except unadjusted and RC) all require that biomarker values can be obtained for any individual in the study
- This requires storing biological samples on all individuals. The cost of taking the sample and storing it needs to be reckoned against the increased power that could accrue from their use
- Cost of the assay is less crucial, since nested case-control designs can be used to analyze the data
- Many prospective studies now incorporate biobanks allowing the use of the methods described

Slide 66

An important practical question is that of the cost of including a biomarker in the study.

To implement the methods described today, the investigator needs to be able to obtain a biomarker measurement from any person in the study. This means that biological specimens have to be stored for all participants. The cost of taking and storing the specimens needs to be weighed against the sample size savings that could accrue from their use.

It's the taking and storing of the specimens that is crucial, and the cost of performing the assay is less crucial since, typically, the number of assays that are done can be drastically reduced without loss of power, by using a case-control design nested within the cohort study.

Many prospective studies today are designed to include a blood or tissue bank, and in such cases the methods I have described may give very useful increases in statistical power for addressing questions of diet-health associations.

Can we measure all the important confounders?

- The methods described all require that all important confounders of both dietary intake and the biomarker are identified and measured
- Unfortunately, however hard we try, we can never be sure that we have identified and measured all of these confounders
- Introducing the marker into the analysis introduces a new set of potential confounders
- For this reason, extra care in the interpretation of results is required

Slide 67

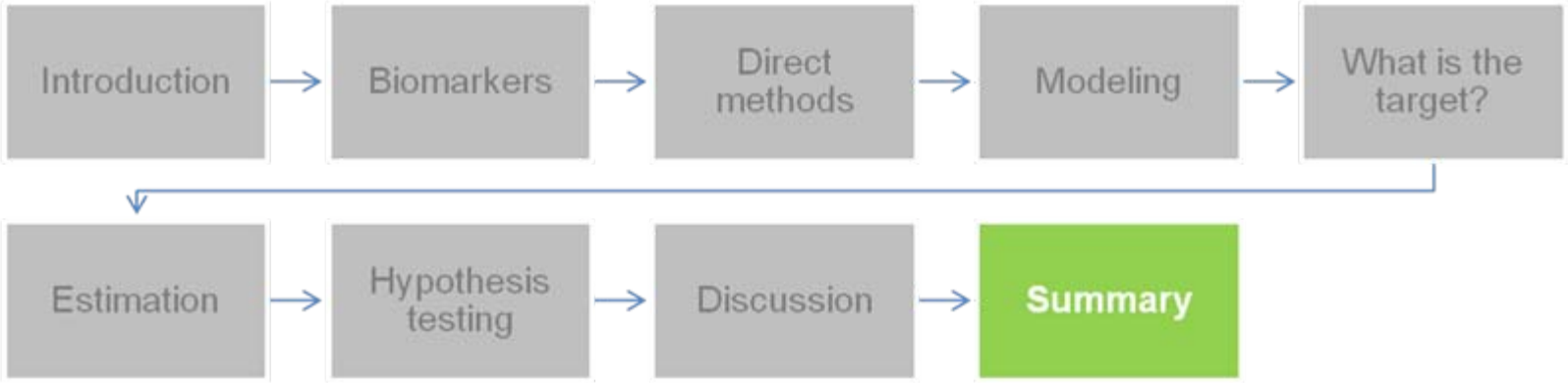
Perhaps the greatest challenge to the successful use of concentration biomarkers in the manner we have described is the challenge of adjusting for confounding. As soon as we enter a biomarker into the model, we introduce a potential new set of confounders, and we can never be sure that we know what these are. Thus, associations that we see may sometimes be spurious, as we will find out, to our cost if we try to build intervention programs based upon them. So extra care is needed in interpreting the results.

Do we have the necessary information to execute enhanced RC or the new method?

- As shown in the CAREDS example, it is not a simple matter to set up the calibration equations needed to implement Enhanced RC or the new method.
- Sometimes, as in that example, previous feeding studies and population studies may be available. Otherwise, special feeding or calibration studies will be required.
- In addition the number of biomarkers known to provide good prediction of true usual intake are limited.
- Prentice et al are currently conducting a large feeding study to identify new biomarkers and develop calibration equations for several foods and nutrients, as part of the WHI.

Slide 68

Finally, as mentioned earlier, the modeling methods, enhanced regression calibration and the new method, require extra knowledge beyond what will be available from the study in question. In many cases, that knowledge may not exist and new feeding studies may be required to provide it. Prentice and colleagues are currently conducting a large feeding study to identify new biomarkers of dietary intake and to enable the development of new calibration equations for predicting a wider range of dietary intakes. If that study is successful, then the methods we have described today may become central to future nutritional epidemiology.


Introduction
Biomarkers
Direct methods
Modeling
What is the target?
Estimation
Hypothesis testing
Discussion
Summary

SUMMARY

Slide 69

So, to summarize…

Summary

1. Usual Regression Calibration does not usually increase the power to detect diet-health outcome relationships.

2. Using biomarkers can sometimes increase power.

3. Simple methods such as Howe's method or principal components can be used, and are sometimes successful, but (a) do not guarantee increase in power, and (b) can sometimes even reduce power!

4. More complex methods such as Enhanced Regression Calibration can yield important gains in power, but require considerable extra information regarding the relationship between the biomarker and dietary intake.

5. The methods require availability of biological specimens for the individuals in the study, and may be feasible in prospective studies that have incorporated biobanks.

Slide 70

[No notes.]

QUESTIONS & ANSWERS

Moderator: Kevin Dodd

Please submit questions using the *Chat* function

Slide 71

Thank you Dr. Freedman. We'll now move on to the question and answer period of the webinar.

Measurement Error Webinar 11 Q&A

Question: **As related to the choice of confounders, Z, in the various models and how all of those work together, can the set of confounders be different for different parts of the modeling exercise?**

I think there would be a possible situation where you had different confounders for the marker intake model and from the health outcome model. In other words, there may be confounders which are important in the health outcome model which are not important in the marker intake model. And, in fact, that was a situation which we used in the CAREDS example, where we had confounders in the health outcome model, age and smoking, and they were not used in the marker intake model. We would have liked to use smoking in the marker intake model but we just didn't have enough information about it. So in summary, yes, it is quite possible that there would be confounders which are relevant to the health outcome model but not to the marker intake model. *(L. Freedman)*

And following in that same vein, are there ever some cases where there might be a role for some additional covariates that help with the prediction of, say, the biomarker but don't actually come into the health outcome model?

Yes, indeed, that is a possibility. And, of course, the prime example is the questionnaire reported intake, which comes into the prediction for the dietary intake, but is not included as a variable in itself in the health outcome model. If there are several variables like that, they could all be entered into the prediction equation for dietary intake . The important condition is that the risk of disease is conditionally independent from them given the true intake. And in that case, they can be used for predicting true intake and not included in the health outcome model. *(L. Freedman)*

And that was motivated by two of the questions that came up. One, I think, was a good one, where the supposition that the same confounders might affect—I'm talking about markers or covariates for metabolism that might affect—concentration biomarkers might not affect the dietary intake. And so those additional metabolic covariates would probably be used in the biomarker model but not used to predict diet from the FFQ. Is that correct?

That's correct; they would not be used to predict diet from the FFQ. (However, if you wanted to use enhanced regression calibration and include the biomarker in the prediction of dietary intake, then those

metabolic covariates might also enter such a prediction equation.) *(L. Freedman)*

And then another question that's related to this is: You had mentioned in a footnote that all of these models assume that the measurement error is not affected by the confounder, Z, and I think that kind of goes with the fact that you're using regression calibration for these sorts of things. Can you give a little discussion of what might happen or how you might have tried to adjust these models if the measurement error in Q is affected by the confounders?

If the measurement error, through Q, is affected by the confounders and you could measure the confounders, then you would be able to take it into account by including those confounders in the measurement error model. And if you did that, then you would largely overcome this problem. If you were not aware of it but they nevertheless did affect the measurement error in the self-report, then I would think that there would be a potential bias. But we haven't studied how large those biases might be, and it is a concern that needs to be looked at. *(L. Freedman)*

Earlier on, when you were talking about the direct method, Howe's and the PC method unadjusted, you showed a table—I think it was slide 27—that showed that you were looking at odds ratios in going from the 10^{th} to the 90^{th} percentile of measured intake. Now, did those percentiles change depending on what you were using, whether you were using the principal component or whether you were using the Howe's sum of the ranks as your variable? Or was that all based upon some standardized value of intake?

For each variable, the variable itself was used, in the control group, to find out empirically what the percentiles were, the 10^{th} and 90^{th} percentiles, in the population. That's why the control group was used, on the assumption of a fairly rare disease. And of course because in one case you're using the self-report and in another case you're using the Howe measure, which is a sum of ranks, and in another case you're using principal components, which is a weighted sum of the biomarker and the self-report, each of them has different percentiles, but what's common to them is that every time you are comparing the risk of disease of someone sitting at the 90^{th} percentile compared to someone at the 10^{th} percentile in the population, if people are ordered according to each of these variables. *(L. Freedman)*

Can you talk a little bit about the possibility of extending these methods to apply to some non-FFQ self-report instruments like food records or 24 hour recalls as the measured dietary exposure that's not the biomarker?

I think they could be used equally well in that situation. The only proviso is that the external information is available, if we're talking about the modeling methods. And generally speaking, when it comes to doing population studies and we need to know what the distributions of these intakes reported on these instruments are in the population, generally speaking, apart from NHANES, where we have 24 hour recalls, I don't think there are so many population studies which use other sorts of self-reports. So we have 24 hour recalls. It could certainly be done for that. And food records I think would be a bit more problematic, but there may be some studies which have that sort of information as well. So, in principle, yes, it could be applied equally well to other sorts of instruments. Whether or not it could be used in the case where there are episodically consumed foods, I don't know. One may have to do more sophisticated things along the lines that Victor Kipnis has already talked about and will talk about in the next lecture as well. *(L. Freedman)*

[This page intentionally blank.]

National Cancer Institute

measurement ERROR webinar series

Next Session

Tuesday, December 6, 2011
10:00-11:30 EST

Assessing diet-health relationships using a short-term unbiased dietary instrument: focus on risk models with multiple dietary components

Victor Kipnis
National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Slide 72

Thank you very much, Larry, and thanks to our audience for joining today's webinar. Please join us next week for the last webinar in our series, in which Dr. Victor Kipnis will discuss assessing diet and health relationships using a short-term unbiased dietary instrument, with a focus on risk models with multiple dietary components.