National Cancer Institute

measurement ERROR webinar series

Combining self-report dietary assessment instruments to reduce the effects of measurement error

Douglas Midthune, MS
National Cancer Institute

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

Slide 1

Hello and welcome to the tenth session in the Measurement Error Webinar Series. I'm Amy Subar, a nutritionist with the U.S. National Cancer Institute. In today's webinar, we'll focus on combining instruments as a strategy to reduce measurement error in dietary intake data.

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

Because today's presenter, Doug Midthune, is on travel, we had initially planned to play back a prerecorded version of his presentation with a live question and answer period. Fortunately, Doug is now able to join us live for the entire webinar and so today's presentation will be broadcast over the phone as with the previous webinars. The audio will not be broadcast over your computer speakers. To hear the audio, please stay on the teleconference line throughout the webinar.

Before we get started with today's presentation, 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. Please use the Chat feature to submit a question for the question and answer period following the presentation.

A reminder that you can find the slides for today's presentation on the Web site that has been set up for series participants. The URL is available in the Notes box at the top left of the screen. Other resources available include the glossary of key terms and notation, and the recordings of the preceding webinars.

Now I'd like to introduce the presenter for today's webinar. Doug Midthune is a mathematical statistician in the Biometry Research Group, Division of Cancer Prevention, at the National Cancer Institute. Doug has worked with other members of the Surveillance Measurement Error Group at the National Cancer Institute over the past several years to develop the NCI method and to extend the method to more complex applications. As I mentioned, today Doug will discuss combining self-report instruments to reduce the effects of measurement error. Doug.

Welcome to today's webinar about combining self-report dietary assessment instruments to reduce the effects of measurement error. In particular, we'll be talking about reducing the effects of measurement error for assessing diet-health relationships. *(D. Midthune)*

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 Arthur Schatzkin, a colleague who worked with us for many years on the problem of measurement error in dietary assessment.

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

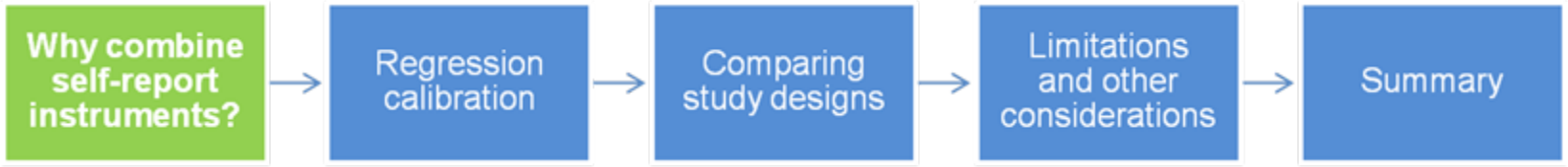
This is a list of everyone involved in the webinar project.

Learning objectives

- Understanding how measurement error leads to loss of precision in estimating diet-health associations
- Learning how to combine self-report dietary instruments to regain precision and improve power to detect associations
- Understanding the limitations of such an approach

Slide 5

Today's learning objectives are: to understand how measurement error leads to loss of precision in estimating diet-health associations; to learn how to combine self-report dietary instruments to regain precision and improve power to detect associations; and, finally, to understand the limitations of such an approach.



WHY COMBINE SELF-REPORT INSTRUMENTS?

Slide 6

First, we're going to talk about why we would want to combine self-report instruments.

Impact of measurement error

- Measurement error (ME) in self-report dietary assessment instruments leads to:
  - Bias (attenuation) in estimated diet-health associations
  - Loss of precision in estimated associations

$$E(\alpha_Q) / s.e.(\alpha_Q) < E(\alpha_T) / s.e.(\alpha_T)$$

  - Loss of power to detect associations

Slide 7

As we've learned in previous webinars, measurement error in self-report dietary instruments causes three problems: first, it leads to bias, or attenuation, in estimated diet-health associations. Second, it leads to loss of precision in estimated associations. By loss of precision, I mean that the ratio of the expected value of the estimated association to its standard error decreases. This makes it more likely that the confidence interval will include the value 0, which leads to the third problem—the loss of power to detect diet-health associations.

Impact of measurement error

- Statistical methods such as regression calibration can correct for bias due to measurement error
- These methods do not typically recover lost precision or power

Slide 8

We've also learned that statistical methods like regression calibration can correct for bias due to measurement error, but that these methods do not typically recover lost precision or power.

Impact of measurement error

- Ways to improve precision and power
  - Increase the sample size
  - Decrease the measurement error
    - Improve existing dietary instruments
    - Develop new instruments
    - **Combine different self-report dietary instruments**

Slide 9

There are ways to improve precision and power. One approach is to increase the sample size. This approach has been taken in a few very large cohort studies, including the NIH-AARP Diet and Health study in the U.S. and the EPIC study in Europe, each of which has over half a million participants.

Another approach is to try to decrease the measurement error in reported intake. One can do this by improving existing dietary instruments or developing new ones, and a lot of work is being done in this area. Alternatively, one could try to combine different dietary instruments to obtain a better measure of true intake, which is what we are going to do today.

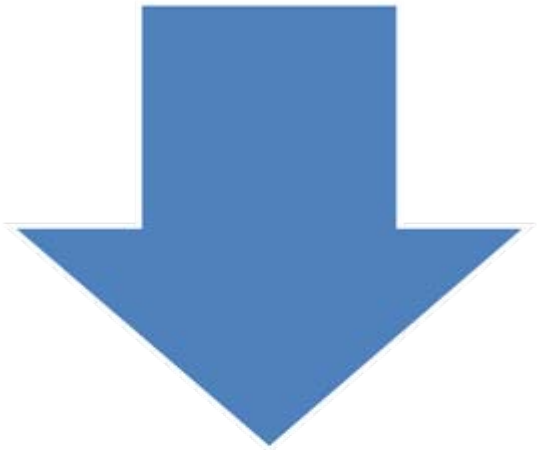
Combining instruments

- Examples of self-report instruments that could be combined:
  - Food frequency questionnaire (FFQ)
  - 24-hour dietary recall (24HR)
  - Multiple-day food record (FR)
- Each instrument has its own strengths and weakness

Slide 10

Examples of self-report dietary instruments that could be combined are food frequency questionnaires, 24 hour dietary recalls, and multiple-day food records. Each instrument has its own strengths and weaknesses.

Self-report dietary instruments

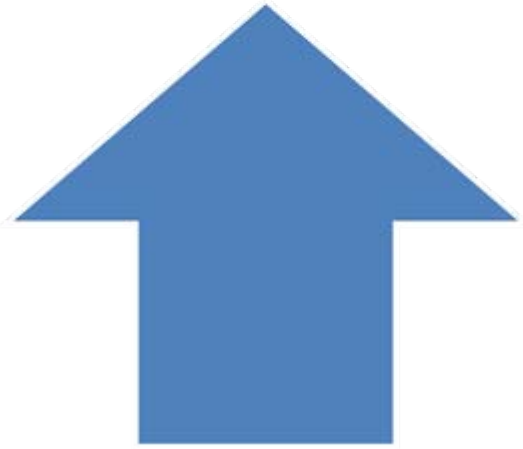


FFQ

Cognitive tasks required are more difficult

24HR / FR

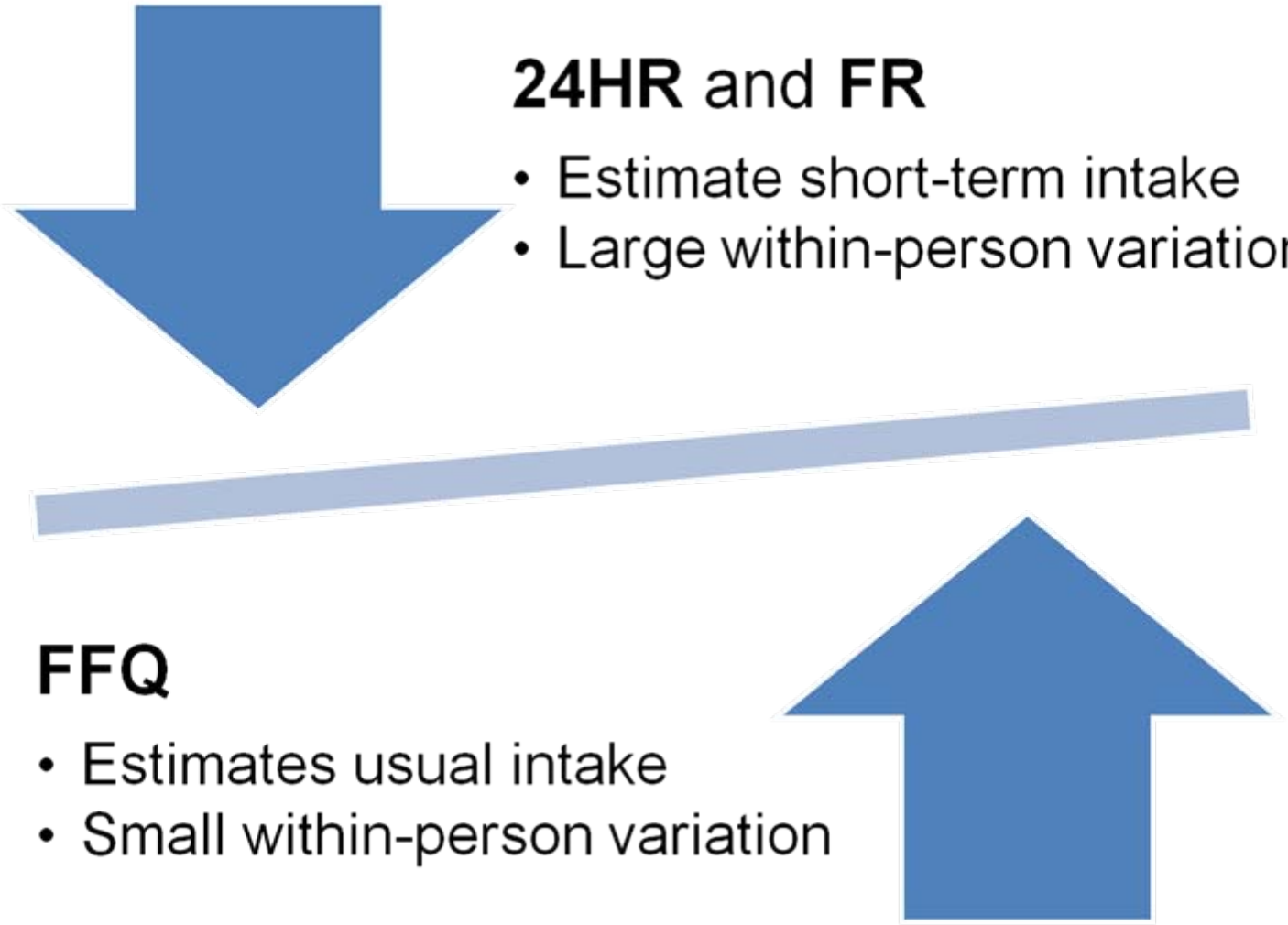
Less biased for estimating intake



Slide 11

One of the problems with food frequency questionnaires is that the cognitive tasks required to complete them are more difficult than those required for 24 hour recalls or food records. Because of this, food frequency questionnaires tend to be more biased for estimating intake than 24 hour recalls or food records.

Self-report dietary instruments



Slide 12

On the other hand, 24 hour recalls and food records are designed to estimate short-term intake, and as a result have large within-person variation, or day-to-day variability. Food frequency questionnaires are designed to estimate usual intake, and so they have smaller within-person variation.

Self-report dietary instruments

- Potential for significant gain in precision by combining instruments with different types of information
  - FFQ measures long-term diet
  - 24HR/FR less bias, measure short-term diet
- Problem:
  - Traditional 24HR and FR are expensive to administer and/or process
  - Not practical for use in large cohort studies

Slide 13

We believe there is potential for significant gain in precision by combining instruments that have different types of information. For example, one can combine a food frequency questionnaire, which measures long-term diet but can be biased, with a 24 hour recall, which is less biased but has large day-to-day variability because it measures intake on a given day.

A problem with such an approach is that traditional 24 hour recalls and food records are expensive to administer and/or process. As a result, they have not been practical for use in large cohort studies.

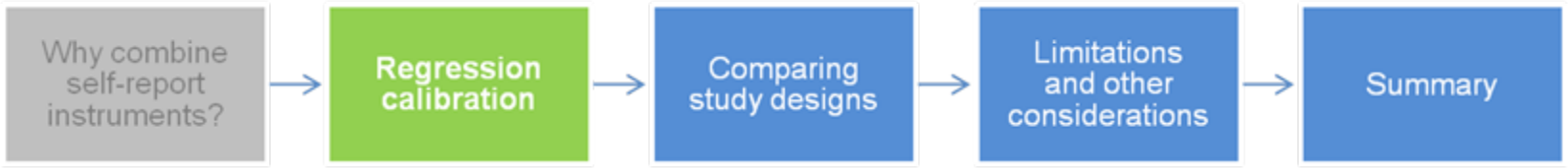
Self-report dietary instruments

- Recent developments in dietary assessment
  - Self-administered automated 24HR, such as the ASA24 (NCI)
  - Automated FR, some using mobile phone technology
  - Much less expensive than traditional 24HR/FR
  - Practical for use in large cohort studies

Slide 14

Recent developments in dietary assessment, however, have made this less of a problem. New automated self-administered 24 hour recalls have been developed, such as the ASA24 developed at NCI. There are also automated food records, including some that use mobile phone technology to allow participants to photograph their food before it is eaten, and then photograph the leftovers.

These new automated tools are much less expensive than the traditional 24 hour recalls and food records, so they are practical for use in large cohort studies.



USING REGRESSION CALIBRATION TO COMBINE SELF-REPORT INSTRUMENTS

Slide 15

Now we're going to talk about using regression calibration to combine self-report instruments.

Regression calibration

- **Regression calibration** corrects estimated diet-health associations for bias due to ME in reported intake
  - (Relatively) simple and intuitive
  - Applicable in many situations (e.g., linear and logistic regression, survival analysis)
  - Often nearly as efficient as maximum likelihood estimation
  - Extends naturally to combine multiple instruments

Slide 16

Regression calibration is a method to correct estimated diet-health associations for bias due to measurement error in intake. The advantages of regression calibration include: it's relatively simple and intuitive to use; it's applicable in many situations, such as linear and logistic regression and survival analysis; it's often nearly as efficient as maximum likelihood estimation; and, especially important for our purpose, it extends naturally to combine multiple instruments.

Review of regression calibration

- Diet-health model:

$$\text{Log}\{\text{Odds}(Y=1)\} = \alpha_0 + \alpha_T T$$

  - Y = health outcome variable (0 or 1)
  - Odds(Y=1) = Prob(Y=1) / Prob(Y=0)
  - T = **true** usual dietary intake (**unobserved**)
  - α_T = log odds ratio (quantifies diet-health association)
  - R = **self-reported** dietary intake (**observed**)

Slide 17

We begin with a short review of regression calibration before describing how it can be used to combine multiple instruments.

We'll assume that our diet-health model is a logistic regression model. In logistic regression, the health outcome Y is a binary variable that equals 0 or 1 to indicate whether or not some health event has occurred. The odds that Y=1 is defined as the probability that the event has occurred divided by the probability that it has not occurred. The logistic regression model assumes that the log of the odds is equal to a linear function of true dietary intake, T.

The parameter α_T is called the log odds ratio, and it quantifies the relationship between true dietary intake and the health outcome. We are unable to observe true intake, however, and instead observe reported intake, which we call R.

Review of regression calibration

- Diet-health model:

$$\text{Log}\{\text{Odds}(Y=1)\} = \alpha_0 + \alpha_T \cancel{T}$$

E(T|R)

Prediction equation: $E(T \mid R) = \lambda_0 + \lambda_1 R$

- Regression calibration: **replace T** with its **predicted value E(T | R)** in diet-health model and perform standard analysis
- E(T | R) is the **conditional expectation** (mean) of true intake T given reported intake R
- E(T | R) is the **best** predictor of T given R

Slide 18

To perform regression calibration, we need a prediction equation. Here, we have an example of a linear prediction equation, where the predicted value of true intake, T, is a linear function of reported intake, R.

The regression calibration method is to replace T with its predicted value in the diet-health model, and then perform the standard analysis. In statistical terminology, the predicted value is the conditional expectation, or conditional mean, of true intake T given reported intake R. The conditional expectation is known to be the best predictor of T given R, in the sense that it minimizes the mean squared error.

Review of regression calibration

- Diet-health model:

$$\text{Log}\{\text{Odds}(Y=1)\} = \alpha_0 + \alpha_T T$$

E(T|R)

- **Assumption**:
  - R has "**nondifferential error**" with respect to disease Y
  - R provides no information about disease Y beyond that provided by T
- Under this assumption, regression calibration estimates are (approximately) **unbiased**

Slide 19

The main assumption of regression calibration is that reported intake, R, has nondifferential error with respect to disease, Y. This means that reported intake provides no information about disease Y beyond that already provided by true intake T.

Under this assumption, regression calibration estimates are approximately unbiased.

Review of regression calibration

- Diet-health model:

$$\text{Log}\{\text{Odds}(Y=1)\} = \alpha_0 + \alpha_T \cancel{T}$$

E(T|R)

Prediction equation: $E(T \mid R) = \lambda_0 + \lambda_1 R$

- Predicted value E(T | R) provides no more **information** about true intake than R

- As a result, regression calibration does **not** recover **power** lost due to measurement error

Slide 20

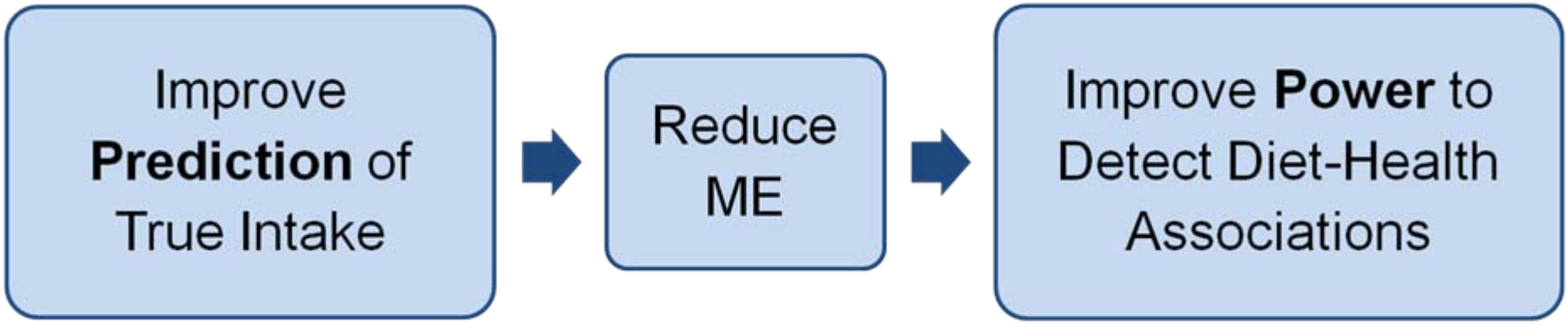
The predicted value of T given R, however, provides no more information about true intake than R itself.

As a result, regression calibration does not recover power that is lost due to measurement error.

Let me clarify this a little. For linear regression calibration with a single dietary exposure, it is strictly true that you can't recover any of the lost power. If you have multiple dietary exposures, however, you may sometimes increase power by reducing residual confounding, but this is not something that generally happens. So, in general, regression calibration does not recover power lost due to measurement error.

Can RC be made more powerful?

- If we can **improve prediction** of true intake, we can increase precision and power



Slide 21

Can we make regression calibration more powerful? We know that if we can improve prediction of true intake, then we can increase precision and power. By improving prediction of intake, we reduce measurement error, which in turn improves the power to detect diet-health associations.

Can RC be made more powerful?

- **Conditional expectation** is the best predictor of true intake T
  - **Q**: How can we improve prediction if the RC predictor is already the "best"?
  - **A**: Conditional expectation is the best predictor of true intake given reported intake (given the information provided)
- Can improve prediction by adding information

Slide 22

Now, one of the things I said earlier was that conditional expectation is the best predictor of true intake, T. So how can we improve prediction if the regression calibration predictor is already the best?

Well, the answer is that the conditional expectation is actually the best predictor of true intake given reported intake, or given the information that's provided.

So we can improve prediction by adding information.

Enhanced regression calibration

- Enhanced RC: predict T using E(T| R,C), where C is an additional variable that:
  1. Helps to **predict** true intake, but
  2. Not related to health outcome given **true** intake
     - Not a confounder
     - Has nondifferential error
- Requirement 2) crucial: if C is related to intake, estimated diet-health association will be **biased**
- Additional **self-report** instruments seem to be perfect candidates for enhanced regression calibration

Slide 23

This leads to the idea of enhanced regression calibration, which is sometimes called "extended" regression calibration. In enhanced regression calibration, we add an additional variable, C, to the prediction equation for T.

This additional variable has to fulfill two criteria. First, it helps predict true intake; second, it's not related to the health outcome given true intake. Another way of saying this is to say that C is not a confounder. We can also say that C has nondifferential error with respect to the health outcome.

This second requirement is crucial, because if C *is* related to intake, then the estimated diet-health association will be biased.

Since we usually assume that self-report dietary instruments have nondifferential error, additional self-report instruments seem to be perfect candidates for enhanced regression calibration.

Example: Enhanced RC

- Enhanced RC with two 24HR (R) and FFQ (Q)
  - Assumption: 24HR **unbiased** for true intake
  - Prediction equation:

$$E(T \mid R_1, R_2, Q) = w \times \bar{R} + (1-w) \times E(T \mid Q)$$

  - $\bar{R}$ = mean of two 24HR
  - w = var(u) / {var(u) + var(e) / 2}
  - var(u) = between-person variance in 24HR
  - var(e) = within-person variance in 24HR
- Parameters estimated in **linear mixed effects model**

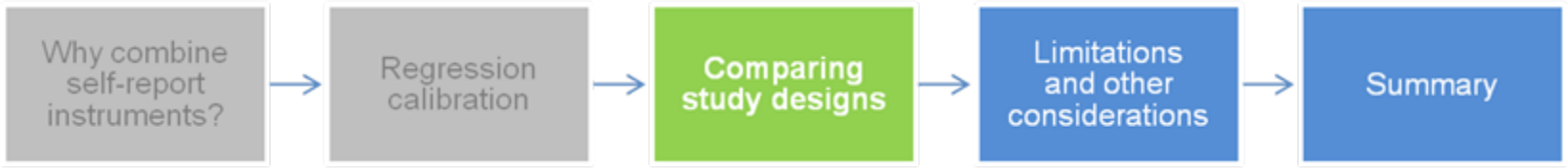
Slide 24

Here is an example of enhanced regression calibration. In this example, we're going to combine two 24 hour recalls and a single FFQ. Our main assumption is that the 24 hour recall is unbiased for true intake.

In earlier webinars, we described regression calibration when the FFQ is the main instrument and the 24 hour recall is used as a reference measure to calibrate the FFQ. Here, both the 24 hour recall and FFQ are considered main instruments; that is, they are completed by everyone in the study and are used to predict true intake. In this case, the 24 hour recall is going to be both a main instrument and the reference measure, so things are a little more complicated than before.

Like before, we have a prediction equation, but in this case the prediction equation is a weighted average of the mean of the two 24 hour recalls and the predicted value of T given Q. The weights, w, are a function of the within- and between-person variance in the 24 hour recalls.

Since we need to estimate the between- and within-person variances, we need to use a linear mixed effects model in order to estimate all the parameters in the prediction equation.



COMPARING DIFFERENT COMBINATIONS OF SELF-REPORT INSTRUMENTS

Slide 25

[No notes.]

Comparing study designs

- In remainder of talk, we compare 3 possible **study designs** (or dietary assessment **strategies**) for estimating dietary intake:
  - **FFQ** alone: one FFQ per subject
  - **24HR** alone: one or more 24HR per subject
  - **24HR** and **FFQ**: one FFQ and one or more 24HR per subject
- **Carroll et al.** Taking advantage of the strengths of two different dietary assessment instruments to improve intake estimates for nutritional epidemiology. *Am J Epidemiol*. (in press)

Slide 26

In the remainder of this talk, we're going to compare three possible study designs, or dietary assessment strategies, for estimating true dietary intake. I call them "study designs" because the dietary assessment strategy must be decided at the design phase of a study.

The first strategy is to use a single FFQ to assess diet; the second is to use one or more 24 hour recalls per subject; and the third is to combine the FFQ with one or more 24 hour recalls per subject.

This comparison is based on a paper by Raymond Carroll and colleagues that will appear in the *American Journal of Epidemiology*.

Comparing study designs

- Study designs evaluated by ability to predict true intake (detect diet-health associations)
- **R-squared** value of the predictor
  - The R-squared value of a predictor is defined as the **squared correlation coefficient** between true and predicted intake
  - Equivalently, it can be thought of as the **proportion of variation** in true intake that is explained by the predictor

Slide 27

We will evaluate the study designs by their abilities to predict true intake, which is equivalent to their abilities to detect diet-health associations when regression calibration is used to correct for measurement error.

Ability to predict true intake will be measured by a predictor's R-squared value. The R-squared value is defined as the squared correlation between true and predicted intake. Equivalently, it can be thought of as the proportion of variation in true intake that is explained by the predictor.

Why is the R-squared value important?

- R-squared value is a direct measure of the ability to predict true intake
- R-squared value determines:
  - **Variance** (precision) of estimated diet-health association
  - **Power** to detect the association
  - **Sample size** needed to obtain desired power

Slide 28

Why is the R-squared value important? The R-squared value is a direct measure of the ability to predict true intake, and, in particular, determines the following properties: the variance of the estimated diet-health association; the power to detect the association; and the sample size required to obtain the desired power for testing an association.

EATS study

- Comparisons based on data from the Eating at America's Table Study (EATS)
  - Conducted 1997-1998
  - Representative sampling of U.S. population
  - 965 men and women, aged 20-70

Slide 29

We're going to base our comparisons on data from the Eating at America's Table Study, or EATS. The study used representative sampling of the U.S. population, and included 965 men and women between the ages of 20 and 70.

EATS study

- Dietary instruments:
  - Four 24HR

    Administered 3 months apart

    By telephone

    Multiple-pass methodology (USDA)
  - One FFQ

    Diet History Questionnaire (NCI)

Slide 30

Two dietary instruments were administered in the EATS study: a 24 hour dietary recall and a food frequency questionnaire. Four 24 hour recalls were collected from each subject. The recalls were administered three months apart, by telephone, using a multiple-pass methodology developed at the USDA. The FFQ was the Diet History Questionnaire developed at NCI.

EATS study

- Dietary variables:
  - Total fat
  - Whole grains
  - Dark-green vegetables
- Dietary variables are energy-adjusted (residual method)
- Carroll et al. looked at 10 dietary components, both unadjusted for energy and energy-adjusted

Slide 31

We're going to look at the effect of study design on the prediction of three dietary variables: total fat intake, whole grain intake, and dark green vegetable intake. We will energy-adjust these dietary variables, using the residual method to adjust for energy intake.

I want to mention that Carroll and his colleagues performed a more complete analysis, looking at a total of ten dietary components, both unadjusted and adjusted for energy.

Assumptions

Assumptions:

- **24HR** provides an **unbiased** estimate of true usual intake for each individual
- **24HR** and **FFQ** have **non-differential** error with respect to health outcome

Slide 32

In our comparisons, we make the following assumptions: first, we assume that the 24 hour recall provides an unbiased estimate of true usual intake for each individual.

As in previous webinars, we call this first assumption a “working” assumption, and we note that there is evidence that, for at least some dietary components, the 24 hour recall is in fact biased; nevertheless, we need to make such an assumption in order to make any progress toward solving these problems. I will talk more about this assumption at the end of this lecture.

The second assumption we make is that the 24 hour recalls and food frequency questionnaires have nondifferential error with respect to health outcome. As I mentioned earlier, this is usually considered a reasonable assumption in cohort studies; that’s because in cohort studies diet is assessed in the beginning of the study before any health events have occurred.

Comparing study designs

- Comparison does not explicitly consider **cost** of study designs (will be discussed later)
- To simplify comparison, will ignore uncertainty due to estimating parameters in the prediction equation

Slide 33

I want to mention that the comparison will not explicitly consider the cost of the study designs, but I will talk about cost later.

Also, to simplify the comparisons, we're going to ignore the uncertainty due to estimating the parameters in the prediction equation. This is because two of the study designs require a calibration study in order estimate the prediction equations, while the other designs don't require a calibration study. Because we didn't want to take into account such things as the size of the calibration study, we decided to make this simplifying assumption.

Comparing study designs

- Study designs (dietary assessment strategies):
  1. Single **FFQ**
  2. From one to twelve **24HR**
  3. Single **FFQ plus** from one to twelve **24HR**
- Since subjects in EATS completed only four 24HR, must simulate 5 or more
- **FFQ plus twelve 24HR** is the "best" study design
  - Adding information always improves prediction
  - More than twelve 24HR may impose unreasonable burden

Slide 34

The study designs we're going to look at are: a single FFQ, from one to twelve 24 hour recalls, and, third, the combination of a single FFQ plus one or more 24 hour recalls.

Since the subjects in EATS only completed four 24 hour recalls, we're going to have to simulate the case of five or more 24 hour recalls per subject.

I also wanted to mention that FFQ plus twelve 24 hour recalls represents the "best" study design, and by this I mean two things: First, it's the best design among those we are considering. This follows from the simple fact that adding information always improves prediction, or at least never makes prediction worse. Second, we think that an FFQ plus twelve 24HR is probably about the best one could hope for in practice without imposing an unreasonable burden on the participants, although even this might be overly optimistic.

Research questions

- Research questions:

1. Does a single **FFQ** work better or worse than (one or more) **24HR**?
2. How many **24HR** per subject?
3. How much does adding the **FFQ** improve the performance of the **24HR** (and vice versa)?
4. Is it better to add another **24HR** or add the **FFQ**?

Slide 35

We're going to consider four research questions. First, does a single FFQ work better or worse than one or more 24 hour recalls? Second, how many 24 hour recalls per subject should one collect? Third, how much does adding the FFQ improve the performance of the 24 hour recall, and vice-versa? And, finally, if you already have a 24 hour recall, is it better to add another 24 hour recall or to add the FFQ?

Graphical comparisons

- Three ways of looking at the data:
  - R-squared value
  - Power to detect diet-health associations
  - Sample size needed to achieve 90% power
- Comparisons relative to the “best” predictor = **FFQ plus twelve 24HR**
- Results presented for women (results for men are similar)

Slide 36

In our graphical comparisons, we're going to look at the data in three different ways. First, we'll compare the R-squared values for the different study designs. As I mentioned earlier, the R-squared value is inversely proportional to the variance of the estimated diet-health association. So if one design has an R-squared value that is twice as large as that for another design, then the variance of the estimated association for that design will be half as large as for the other design.

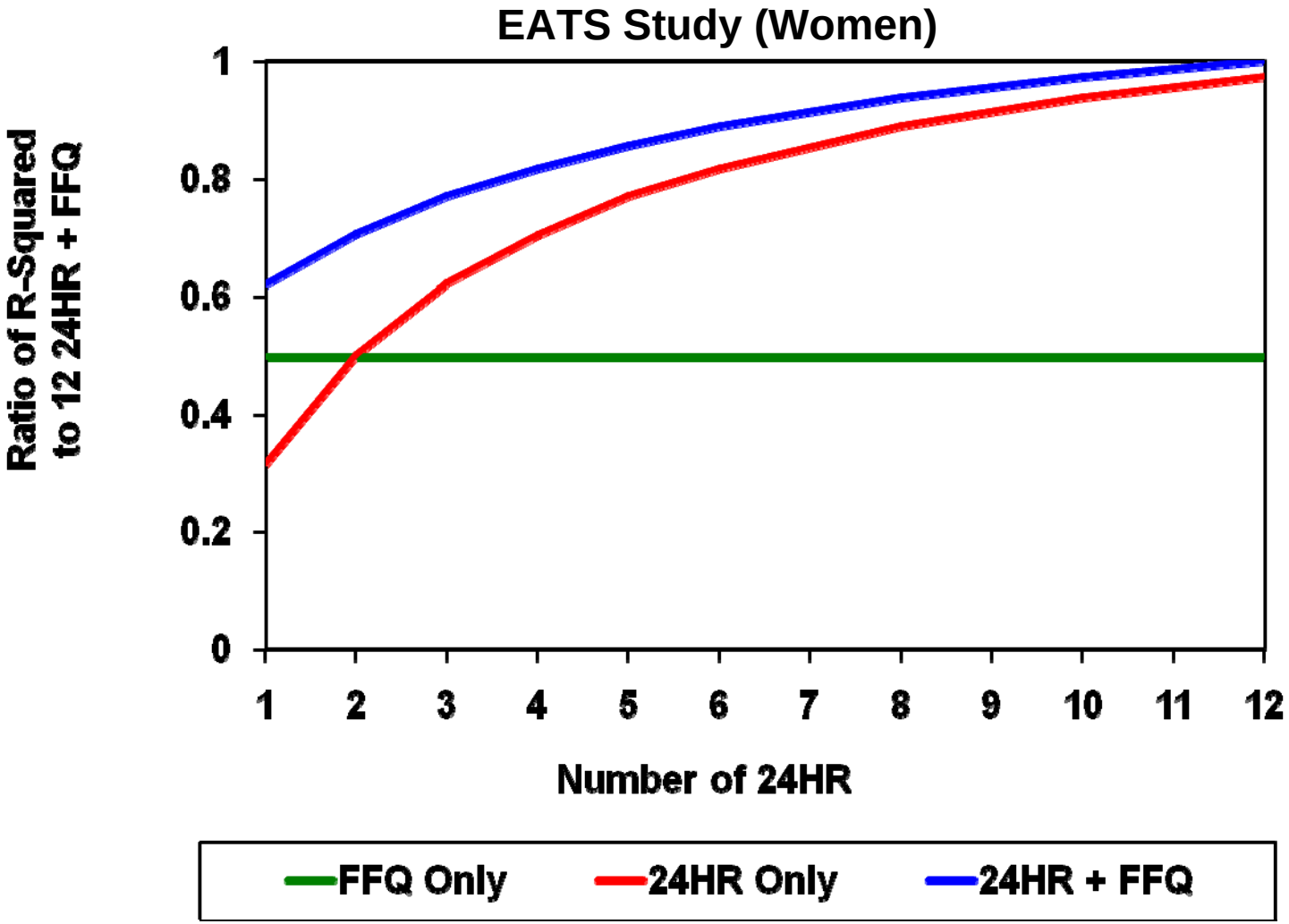
Second, we'll compare the powers to detect diet-health associations. And, third, we'll compare the sample sizes needed in order to achieve a 90 percent power.

As I said earlier, power and sample size are both functions of the R-squared value, so in some sense the different representations are equivalent, but each provides a different way of interpreting the data.

The comparisons are going to be relative to the best predictor, which, as I said before, is the FFQ plus twelve 24 hour recalls. So when we look at R-squared values, we'll actually look at relative R-squared values; that is, the R-squared value of each predictor divided by the R-squared value for the best predictor.

I'll present the results for women. The results for men were similar, and are available in the paper by Carroll and his colleagues.

Ratio of R-squared values: total fat



Slide 37

Here is the first graph, which shows the relative R-squared values for total fat intake in women. The graph has a lot of information, so I'm going to go through it in some detail.

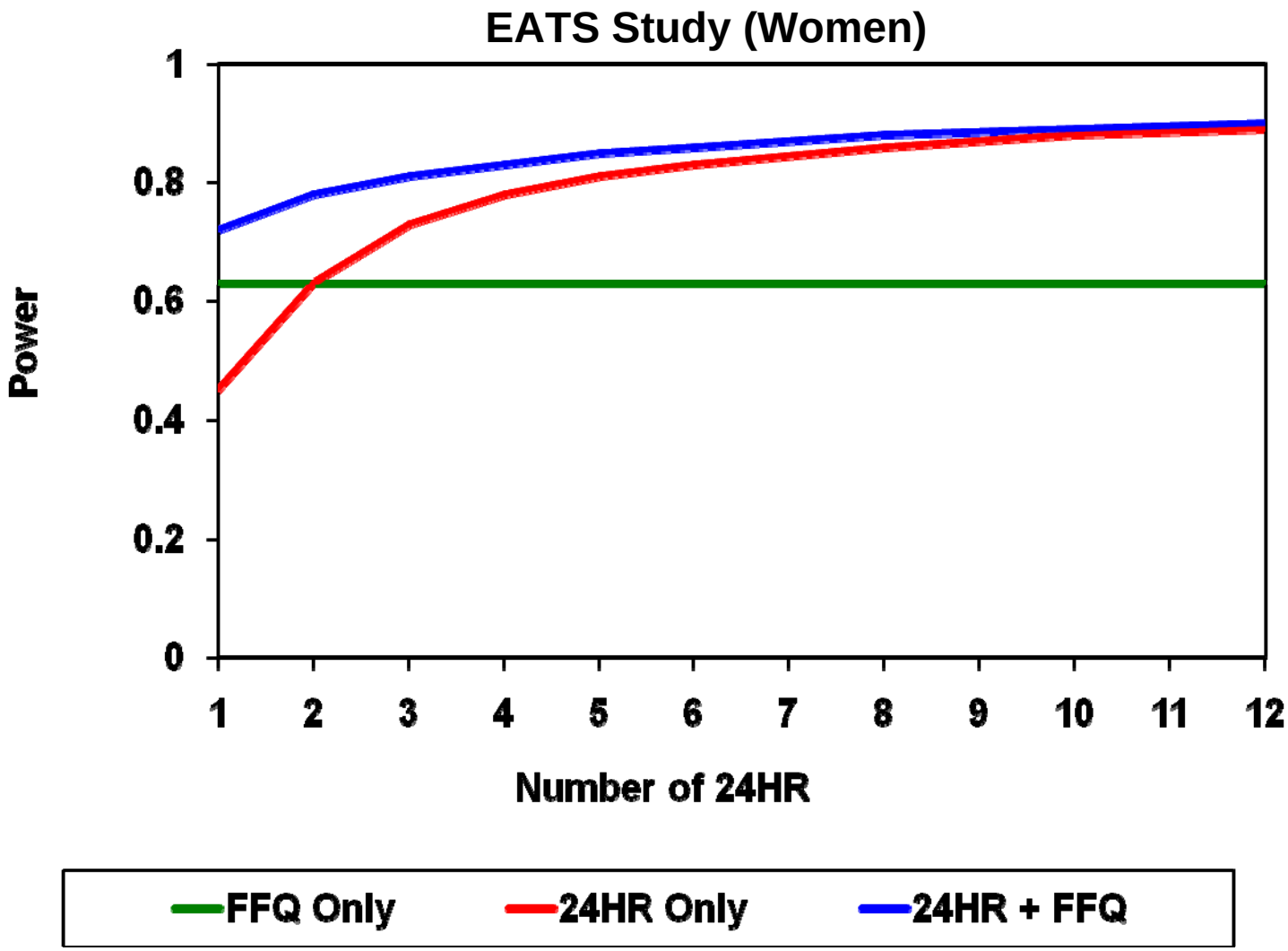
The X axis represents the number of 24 hour recalls in the study design, and the Y axis represents the relative R-squared value. The green line in the middle is the R-squared value for the FFQ alone; the red curve on the bottom represents 24 hour recall alone; and the blue line on top represents the combined FFQ and 24 hour recall

The relative R-squared value for a single 24 hour recall is equal to about 0.35, compared to about 0.5 for the FFQ, and about 0.6 for the combined FFQ and single 24 hour recall. We asked whether adding the FFQ would improve the performance of the 24 hour recall, and we see that in this example it does. The R-squared value almost doubles, increasing from 0.35 to a little over 0.6. This is equivalent to cutting the variance of the estimated diet-health association in half.

Another question we asked was: how well does the FFQ perform in comparison with the 24 hour recall? We can see that in this example the FFQ performs about as well as two 24 hour recalls, but not as well as three or more 24 hour recalls.

Another question we asked was: How many 24 hour recalls per subject should we collect? This is a somewhat subjective question because, as you can see, the R-squared value continues to improve whenever you add more 24 hour recalls. But we can also see that the slope of this curve becomes flatter as the number of 24 hour recalls increases, so that the improvement becomes smaller and smaller. In their paper, Carroll and colleagues concluded that four to six 24 hour recalls seemed to capture most of the information that was available in the recalls.

Power to detect association: total fat



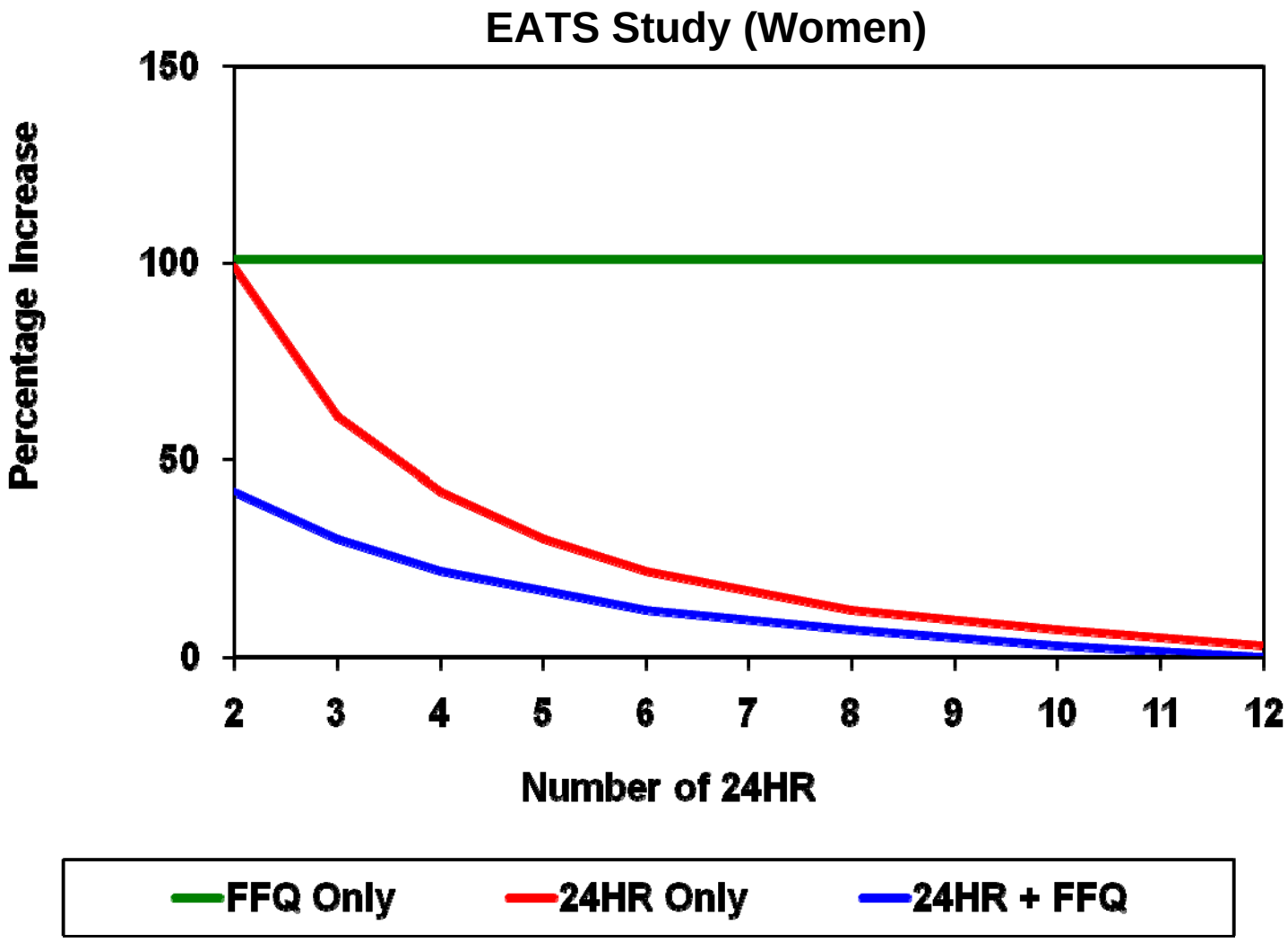
Slide 38

In this second graph, we're looking at the power to detect a diet-health association, again for total fat intake. The Y axis represents the power to detect an association given that the power for the best predictor has 90 percent power. So, again, our comparison is relative to the best predictor.

We see that the power for a single 24 hour recall is about 45 percent, compared to about 60 percent for the FFQ, and about 70 percent for the combined FFQ and single 24 hour. Again, we see that the FFQ performs about as well as two 24 hour recalls.

One interesting thing to note is that the slopes of the curves are much flatter than they were in the graph of R-squared values; this is especially true for the slope of the curve for the FFQ plus 24 hour recalls. This indicates that the increase in R-squared value isn't translating into a significant increase in power. That's something to consider when you're thinking about how many 24 hour recalls to collect.

Percentage increase in sample size: total fat



Slide 39

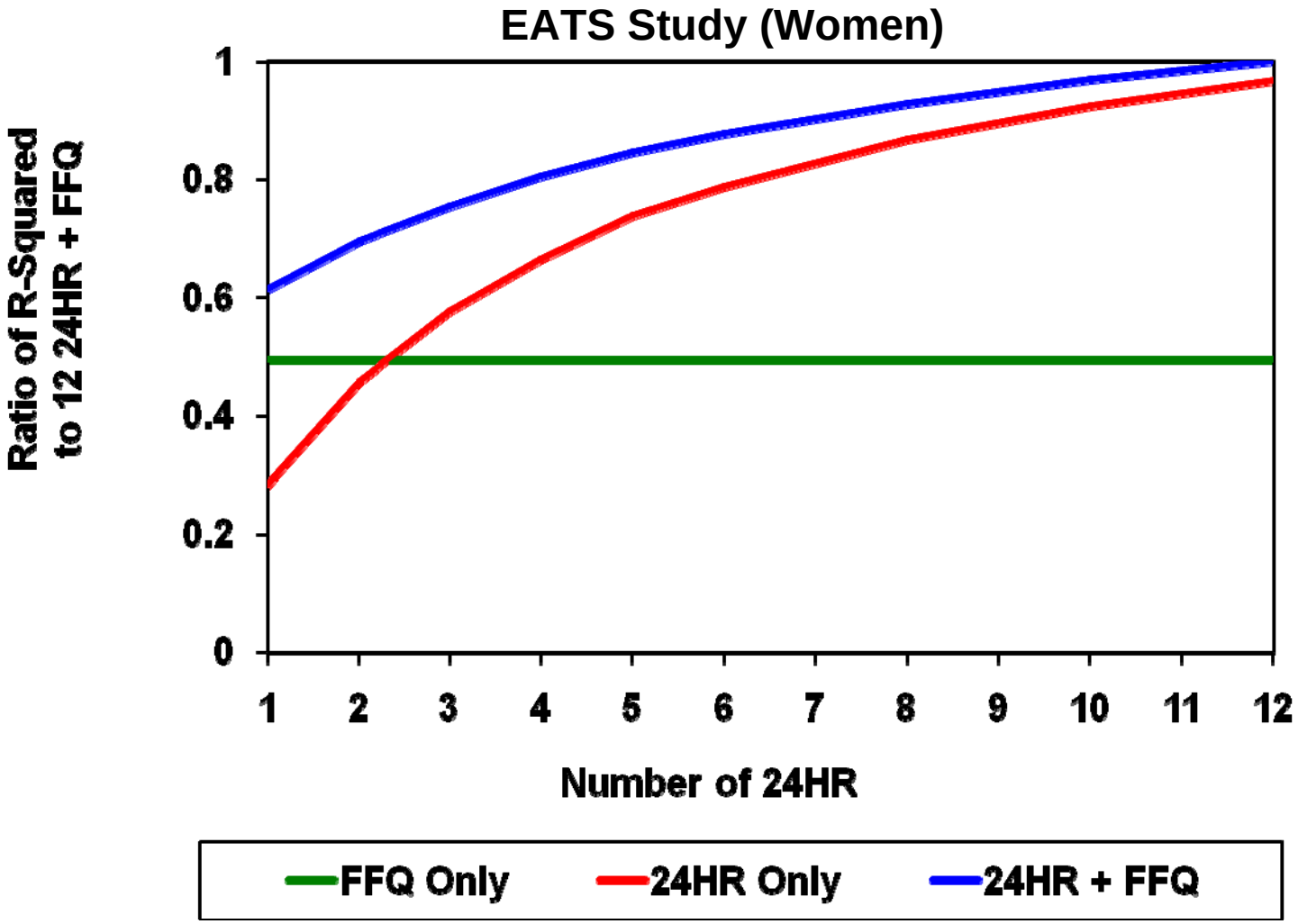
In this third graph, we're looking at the percentage increase in sample size needed to obtain 90 percent power to detect an association. Again, the dietary variable is total fat intake, and we're comparing each design to the combination of the FFQ and twelve 24 hour recalls.

We see that if you used the FFQ alone or two 24 hour recalls alone, you would need a 100 percent increase in sample size in order to get the same power as a study that used a combination of the FFQ and twelve 24 hour recalls.

As I said earlier, we're not going to explicitly consider the cost of each study design, but this is a plot that would be useful when you are actually considering cost. In this case, a single FFQ would require about twice the sample size as the design that combines the FFQ and twelve 24 hour recalls. And depending on the cost of administering twelve 24 hour recalls, the larger sample size may or may not be cost-effective.

In between these two extremes, we see that for the combined FFQ and two 24 hour recall design, we would need a 50 percent increase in sample size, while for the combined FFQ and six 24 hour recalls, we would need only a 10 percent increase.

Ratio of R-squared values: whole grains



Slide 40

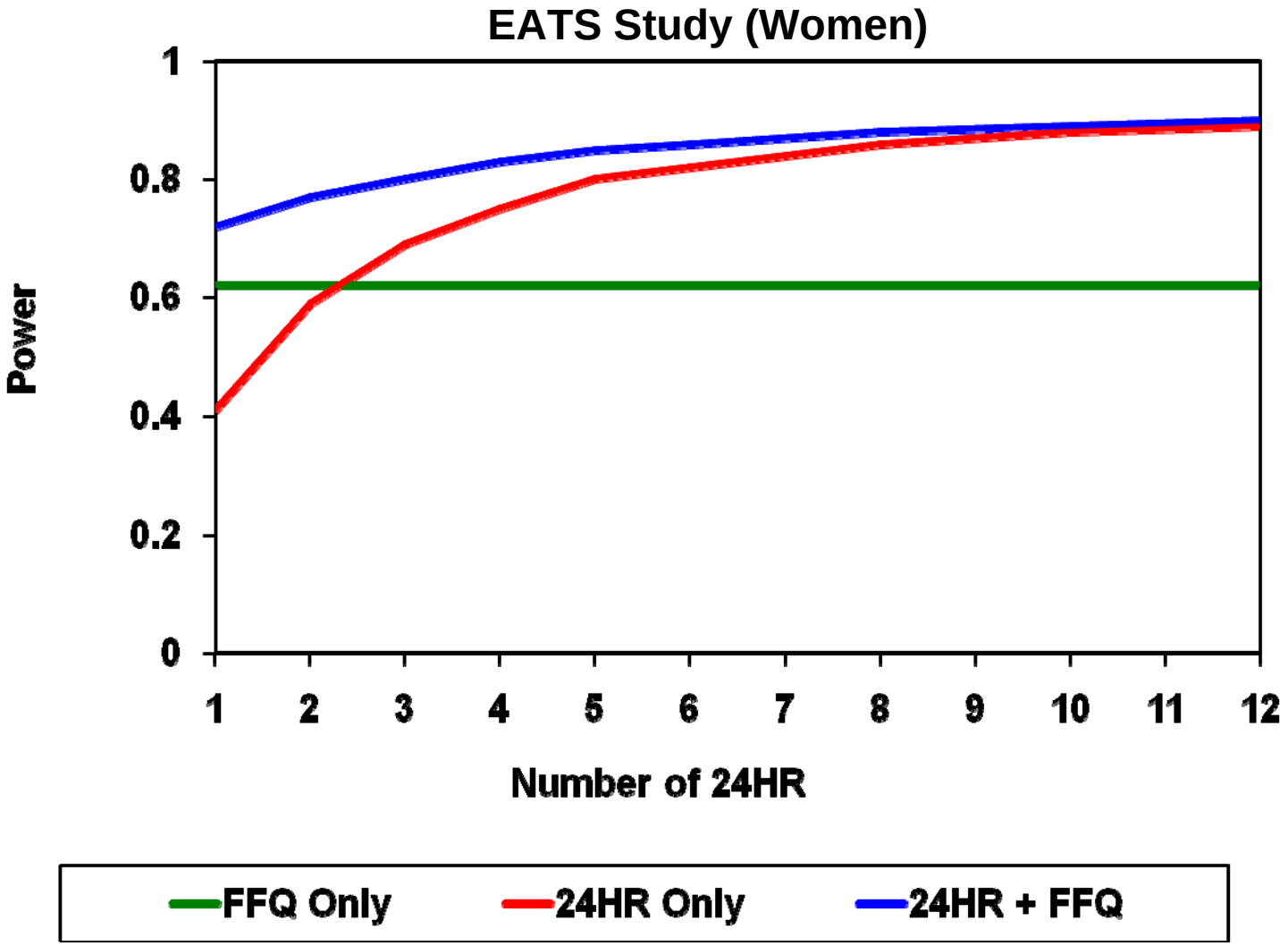
We've looked at three graphs for total fat intake and now we're going to look at the same three graphs for whole grain intake. Total fat is a nutrient that is consumed nearly every day by nearly everyone in the population, while whole grains is an episodically consumed food group. On any given day, about 70 percent of the population consume whole grains.

This graph shows the relative R-squared values for whole grain intake. It is fairly similar to the graph we saw for total fat intake. The R-squared for a single 24 hour recall is about 0.3, compared to about 0.5 for the FFQ, and about 0.6 for the combined FFQ and 24 hour recall. Again, adding the FFQ to a single 24 hour recall approximately doubles the R-squared value for the predictor.

Comparing the FFQ to the 24 hour recall, we see that the FFQ is performing a little better than two 24 hour recalls but not as good as three.

Something I didn't mention before is that if you compare the blue and red curves, you see that as the number of 24 hour recalls increases, the lines become closer and closer together. This indicates that the more 24 hour recalls you have, the less the FFQ adds to the prediction.

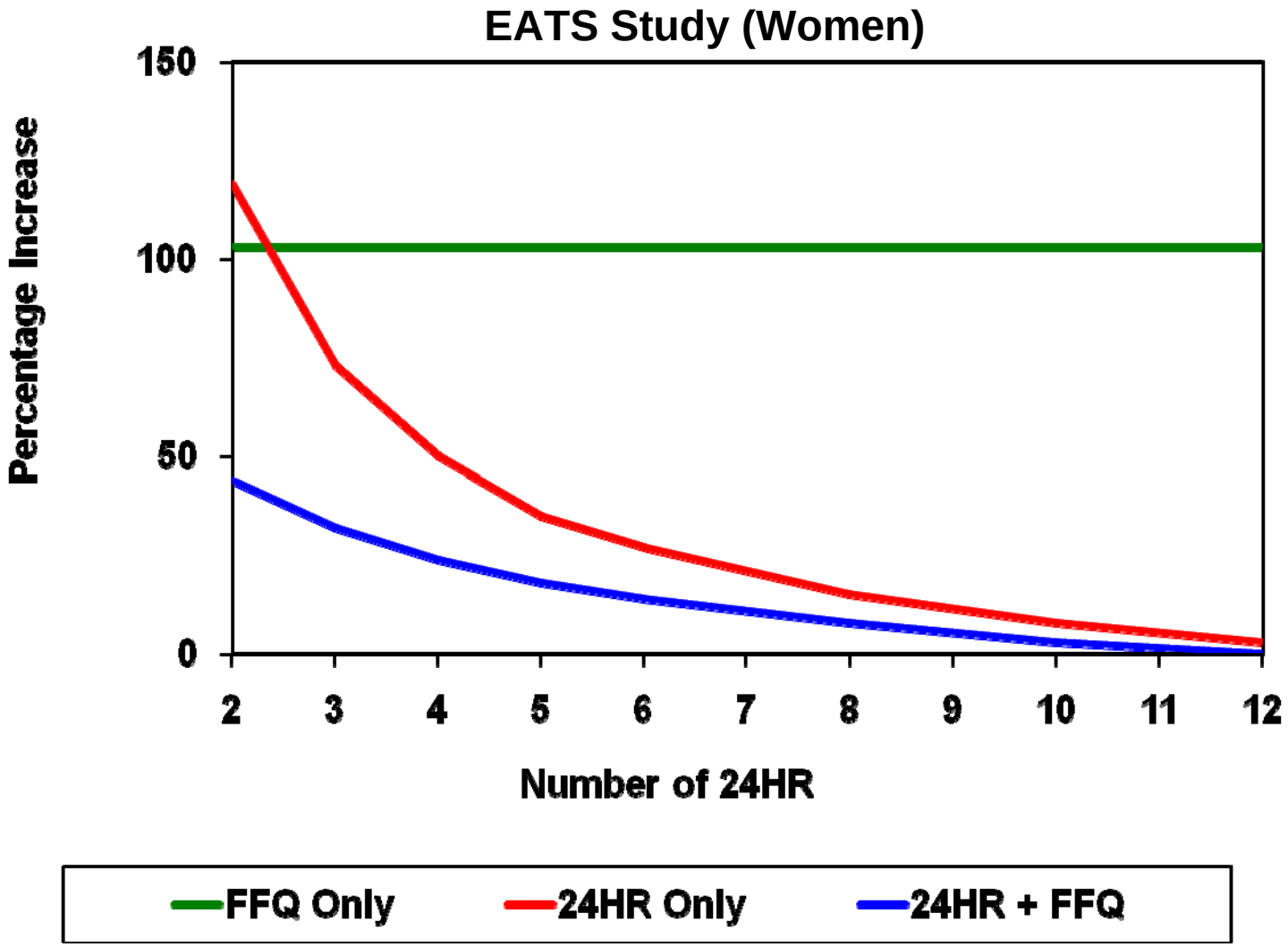
Power to detect association: whole grains



Slide 41

Here is the power to detect diet-health associations for whole grain intake. A single 24 hour recall has about 40 percent power, compared to 60 percent for the FFQ, and about 70 percent for the combined FFQ and 24 hour recall. And, again, the slopes of the curves are pretty flat after four or six 24 hour recalls, indicating that additional recalls will not substantially increase power.

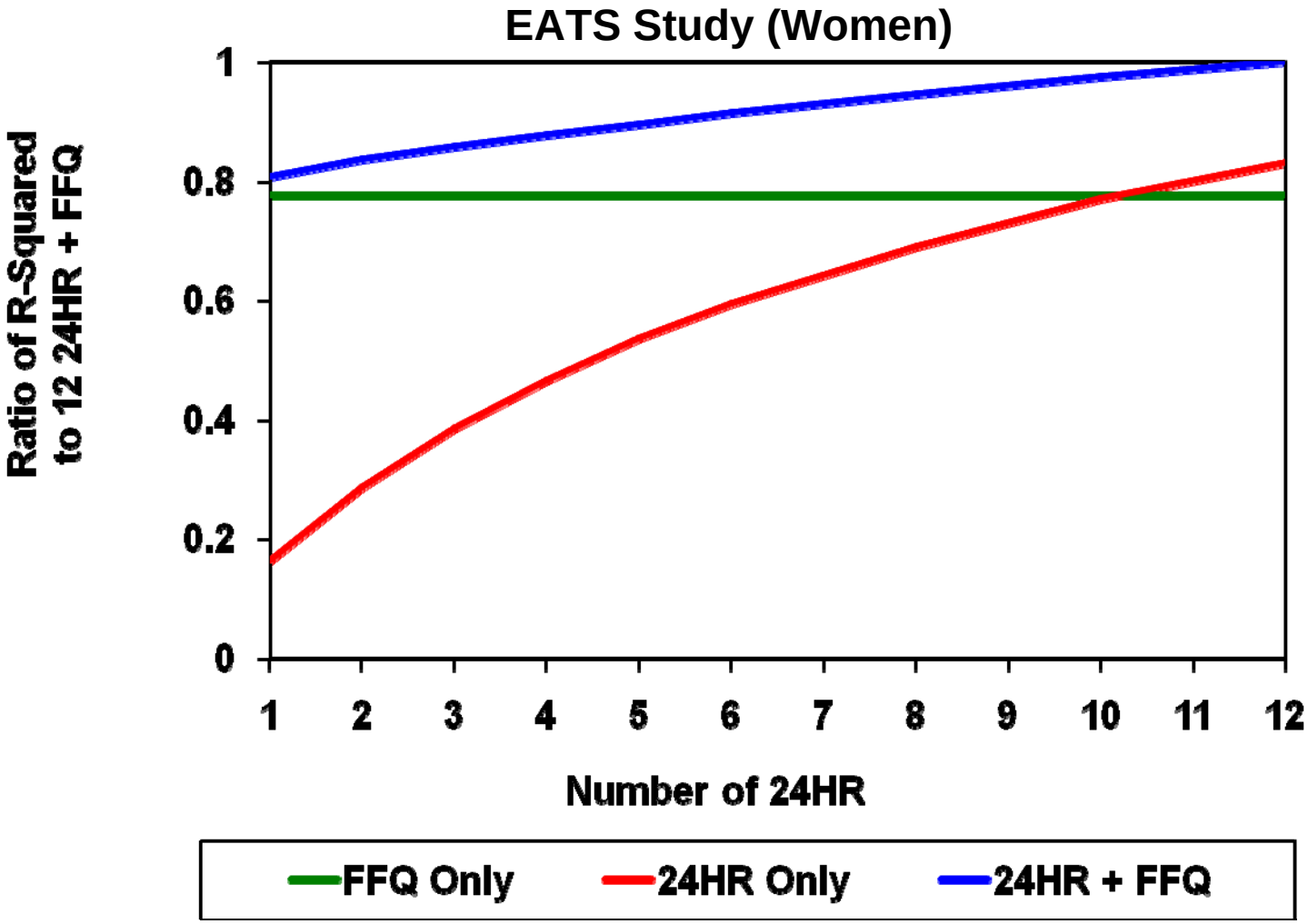
Percentage increase in sample size: whole grains



Slide 42

Here is the percentage increase in sample size needed to obtain 90 percent power for whole grain intake. We see, again, that the FFQ alone requires about twice as many subjects in the study in order to have the same power as the design that combines the FFQ and twelve 24 hour recalls. In comparison, the study design that combines the FFQ and six 24 hour recalls would need only about a 10 percent increase in sample size.

Ratio of R-squared values: dark-green vegetables



Slide 43

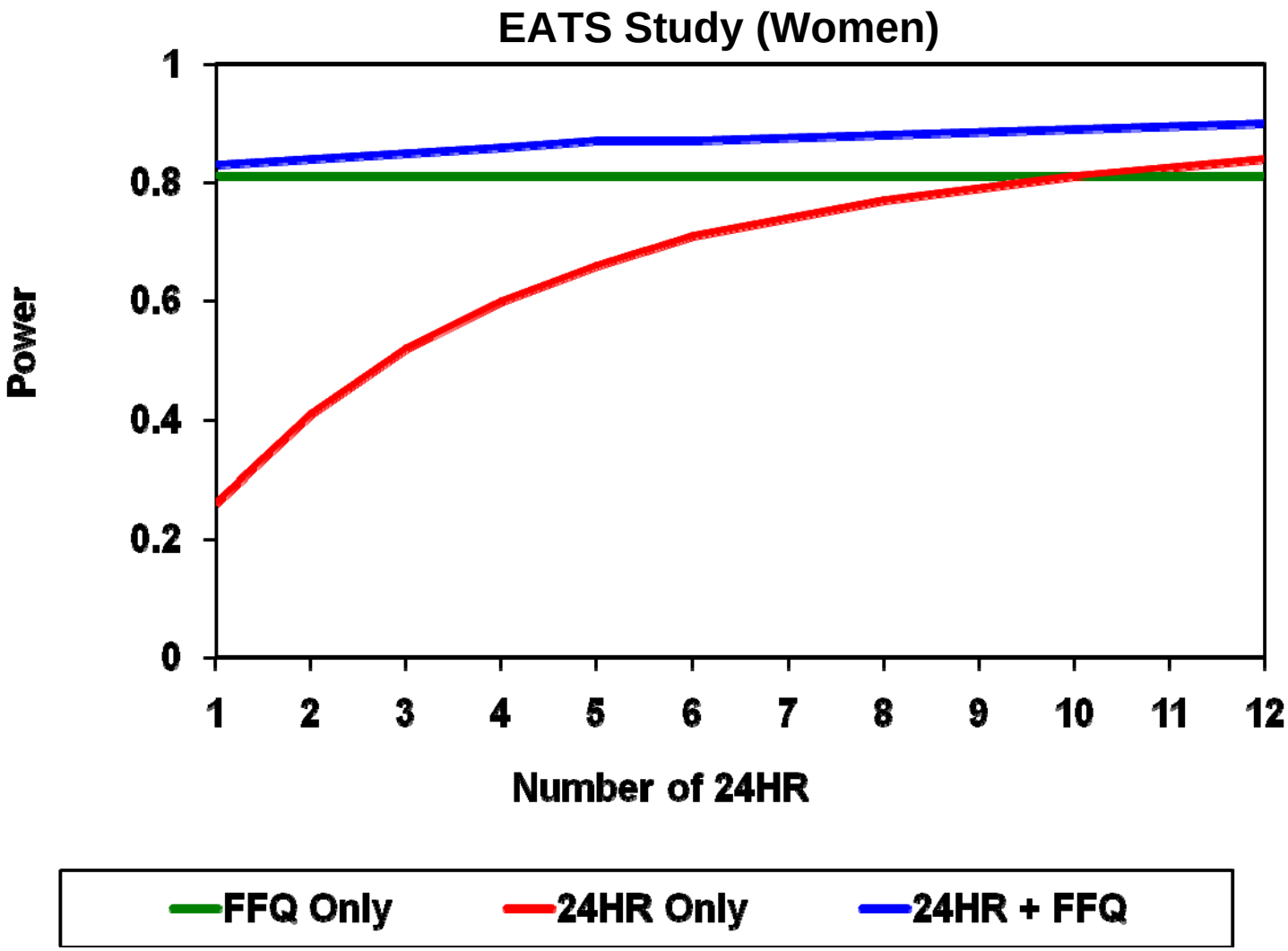
Finally, we're going to look at dark green vegetable intake. These graphs are quite a bit different from those we saw for whole grains and total fat. I want to note that dark green vegetables are very episodically consumed. On any given day they are consumed by only about 20 percent of the population.

In this graph, we see that the single FFQ has a relative R-squared value of about 0.8, substantially higher than we saw for total fat or whole grains. I want to emphasize that this is a relative measure for comparing different study designs rather than different dietary components. So we're not saying that the FFQ estimates dark green vegetable intake more accurately than it estimates intake of total fat or whole grains. What is actually happening is that the 24 hour does not estimate intake for dark green vegetables as accurately as it estimates for total fat and whole grains, so that by comparison the FFQ seems to be performing much better.

Of course, this is very reasonable, since dark green vegetables are consumed by only 20 percent of the population on any given day. As a result, 80 percent of the 24 hour recalls would have zero intake of dark green vegetables, so a single 24 hour recall would provide relatively little information. In fact, we can see that it would take ten 24 hour recalls to perform as well as the single FFQ.

I want to point out that we saw a similar pattern when we looked at fish intake and red meat intake. Fish is another food group that is very episodically consumed, but red meat is not as episodic; it's consumed by about 50 percent of the population on a given day.

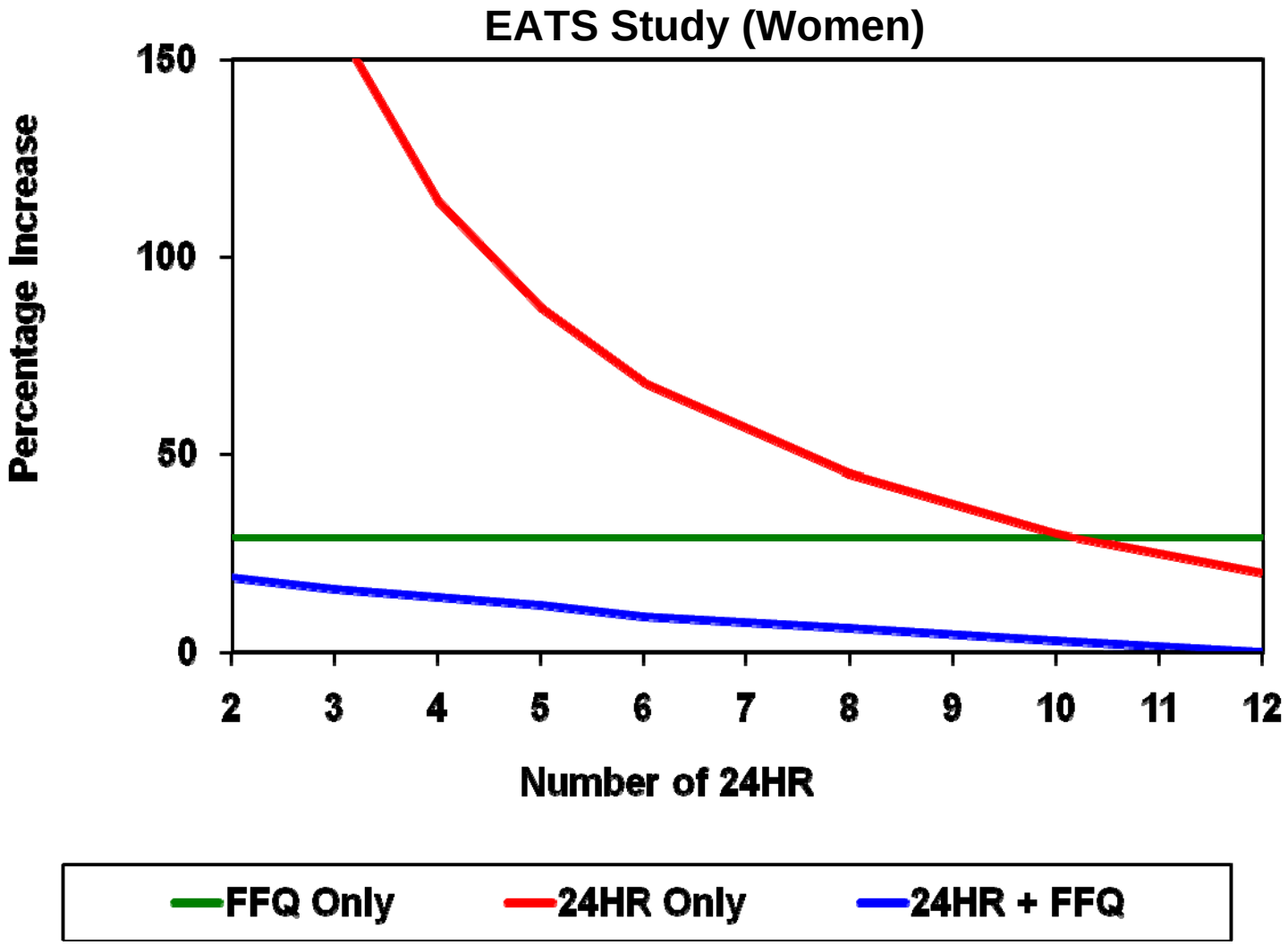
Power to detect association: dark-green vegetables



Slide 44

This graph shows the power to detect association for dark green vegetables. We see that a single FFQ gives you 80 percent power compared to 90 percent power for the FFQ plus twelve 24 hour recalls. So adding twelve 24 hour recalls only increases the power by about 10 percent. The power for a single 24 hour recall is quite low, and it takes about ten 24 hour recalls to have the same power as a single FFQ.

Percentage increase in sample size: dark-green vegetables



Slide 45

Finally, this graph shows the percentage increase in sample size needed to obtain 90 percent power for dark green vegetables. We see that you only need about a 30 percent increase in sample size to obtain the same power as the combination of the FFQ and twelve 24 hour recalls. For the 24 hour recall alone, you have to have four 24 hour recalls before you even make it onto this plot. So for dark green vegetables, the FFQ is quite important in prediction of true intake.

Summary of comparisons

- In general, calibrated FFQ performs about as well as two 24HR
- For some dietary variables (e.g. dark-green veg.) FFQ performs better than 6 or more 24HR
- Using 4-6 24HR seems to capture most of the information available in 24HR

Slide 46

In summary, we can say that, in general, the FFQ performs about as well as two 24 hour recalls. For some dietary variables, such as dark green vegetables, the FFQ performs better than six or more 24 hour recalls.

We can also say that four to six 24 hour recalls seems to capture most of the information that's available in the 24 hour recall.

Summary of comparisons

- Combining FFQ and 24HR can lead to substantial gains over either alone
- Adding an FFQ to a 24HR is usually better than adding a second 24HR
- For **episodically-consumed** dietary components, it may be especially important to include an FFQ

Slide 47

We also found that combining the FFQ and the 24 hour recall can lead to substantial gains over either alone. And in most cases, adding one or more 24 hour recalls to an FFQ also provided significant additional precision.

We can also say that adding an FFQ to one or more 24 hour recalls improves precision more than adding an additional 24 hour recall.

Finally, for dietary components that are very episodically consumed, such as dark green vegetables, it is especially important to include an FFQ.

LIMITATIONS AND OTHER CONSIDERATIONS

Slide 48

Now we're going to talk about some limitations and other considerations.

Limitations of comparisons

- Had to simulate 5 or more 24HR (assumes quality will not drop off)
- Study designs with FFQ alone or just a single 24HR require a **calibration sub-study** of participants who complete two 24HR
- Did not take into account the **uncertainty** due to estimating parameters in prediction equation
- Assumed that the 24HR provided an **unbiased** estimate of true intake for each individual

Slide 49

The first limitation of these comparisons, as I mentioned earlier, is we had to simulate five or more 24 hour recalls, and when we did this we assumed that the quality of the recall would not drop off; that is, we assumed that the respondents would continue to respond and continue to diligently complete their 24 hour recalls.

A second limitation is that, for the study designs that include just the FFQ or a single 24 hour recall, you need a calibration substudy of participants who completed at least two 24 hour recalls in order to estimate the prediction equations. And for our comparisons to be valid, the calibration substudy must be large enough that the prediction equations can be accurately estimated.

But the most important limitation is that we assumed that the 24 hour recall provided an unbiased estimate of true intake for each individual.

Limitations of comparisons

- Studies with **reference biomarkers** of intake (doubly-labeled water for total energy, urinary nitrogen for protein) have shown that 24HR are biased for these nutrients

- In general, incorrectly assuming that the 24HR is unbiased leads to:
  - **Biased** estimates of diet-health associations
  - **Invalid comparisons** of precision and power, unless bias is the same for all instruments

Slide 50

From studies with reference biomarkers, such as doubly labeled water for energy intake and urinary nitrogen for protein intake, we know that 24 hour recalls are in fact biased for these nutrients.

In general, if you incorrectly assume that the 24 hour recall is unbiased, two things happen. First, estimated diet-health associations will be biased. Second, comparisons of different study designs will be invalid, unless the bias is the same for all study designs.

Since we used the 24 hour recall as the reference measure in all our studies, it may be reasonable to think that the bias for each of the study designs might be the same or at least similar.

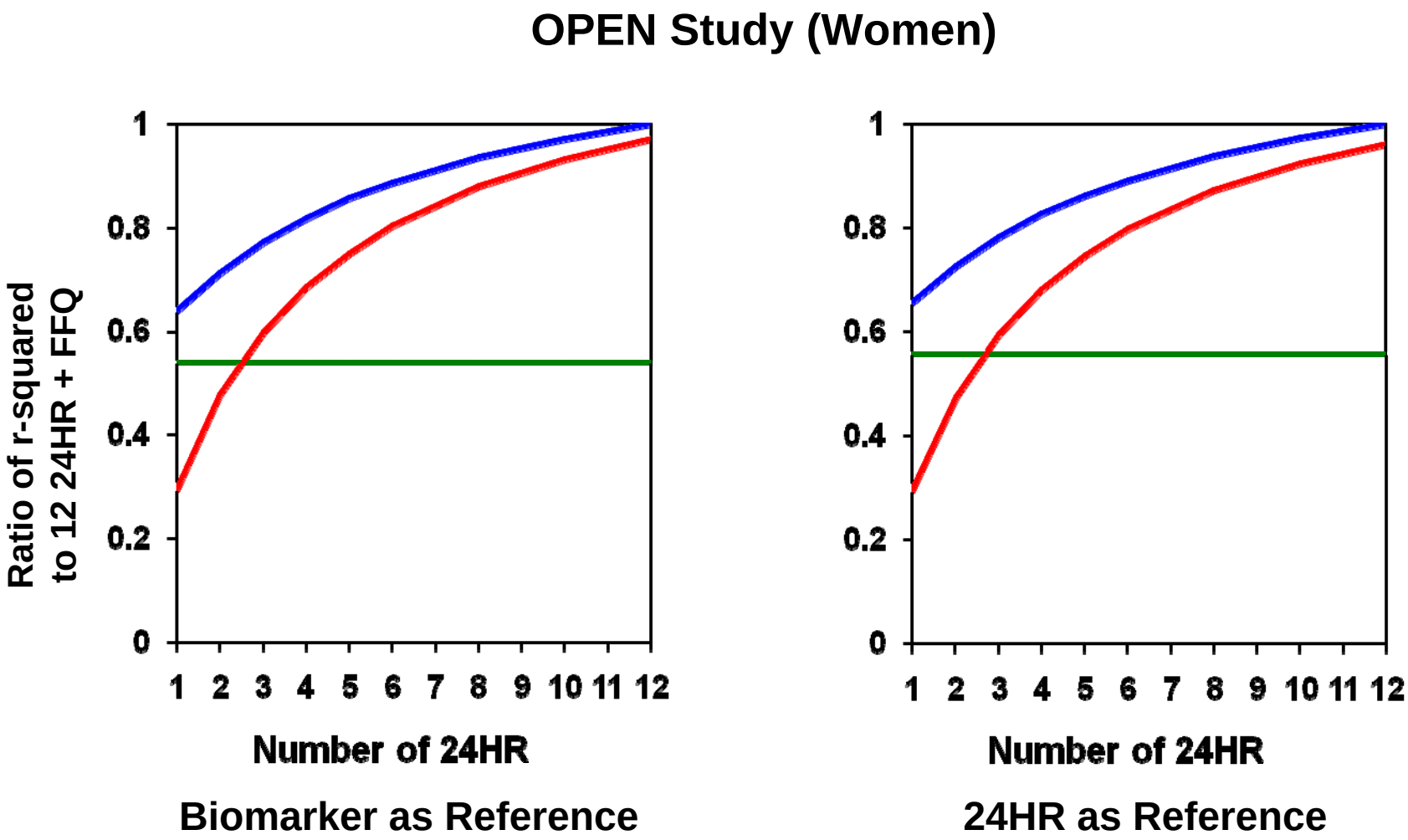
OPEN study

- Use OPEN to examine effect of **biased** 24HR
- OPEN study (1999-2000)
- 484 men and women, aged 40-69
- Dietary Assessment:
  - **FFQ** (2 per subject)
  - **24HR** (2 per subject)
  - **Reference biomarkers** for energy, protein and potassium

Slide 51

We're going to look at the effect of assuming that the 24 hour recall is unbiased when it's really biased. We'll use the OPEN biomarker study that included 484 men and women between the ages of 40 and 69. The dietary assessment instruments in OPEN included a food frequency questionnaire, two 24 hour recalls per subject, and reference biomarkers for energy, protein, and potassium. These biomarkers have been shown in feeding studies to provide approximately unbiased estimates of true intake.

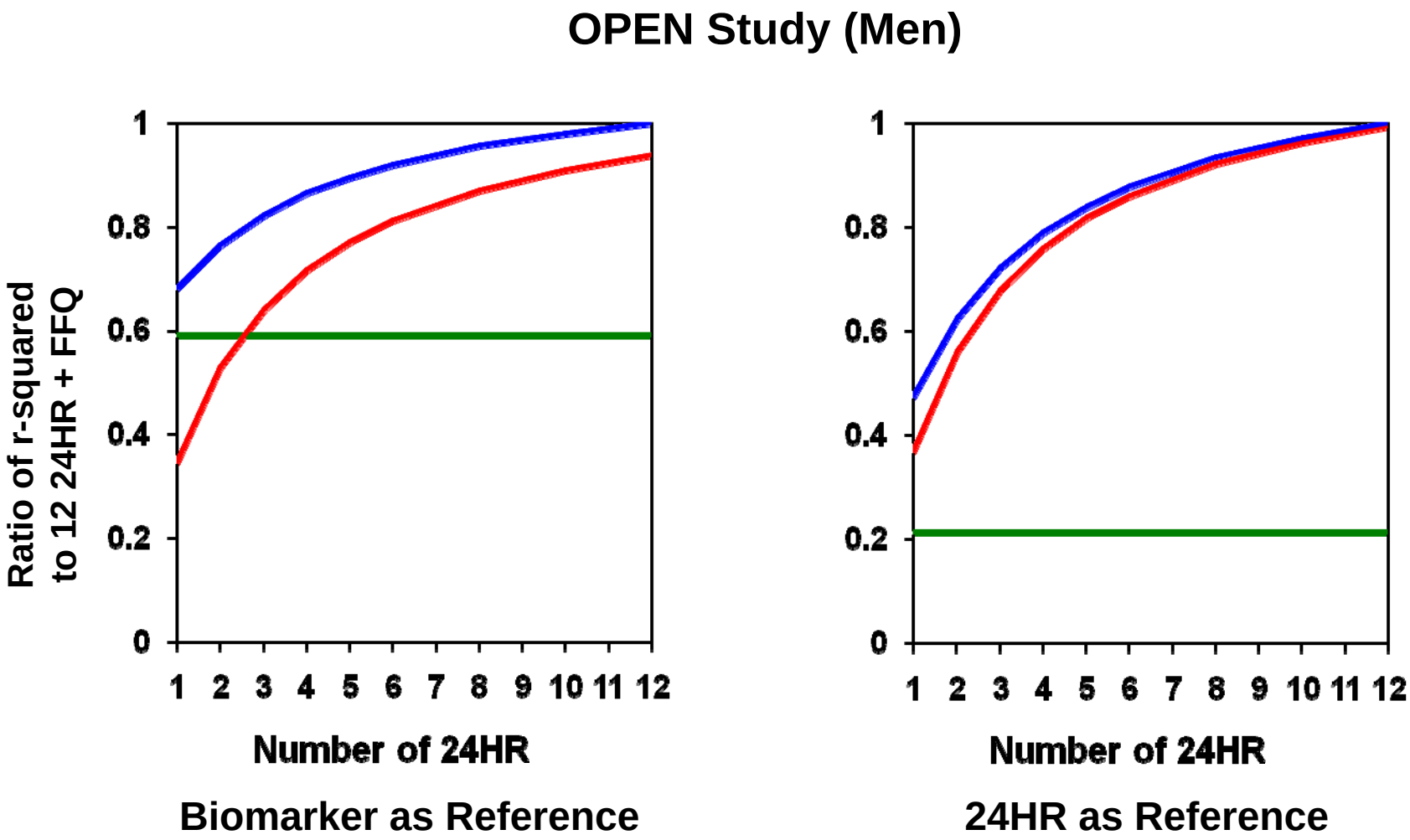
Ratio of R-squared values: protein



Slide 52

Here are two graphs of estimated R-squared values for energy-adjusted protein in women. On the left-hand side are the R-squared values estimated using the biomarker as the reference instrument, while on the right-hand side are R-squared values estimated using the 24 hour recall as the reference instrument. As we can see, the comparisons based on the 24 hour recall are quite similar to those based on the reference biomarker.

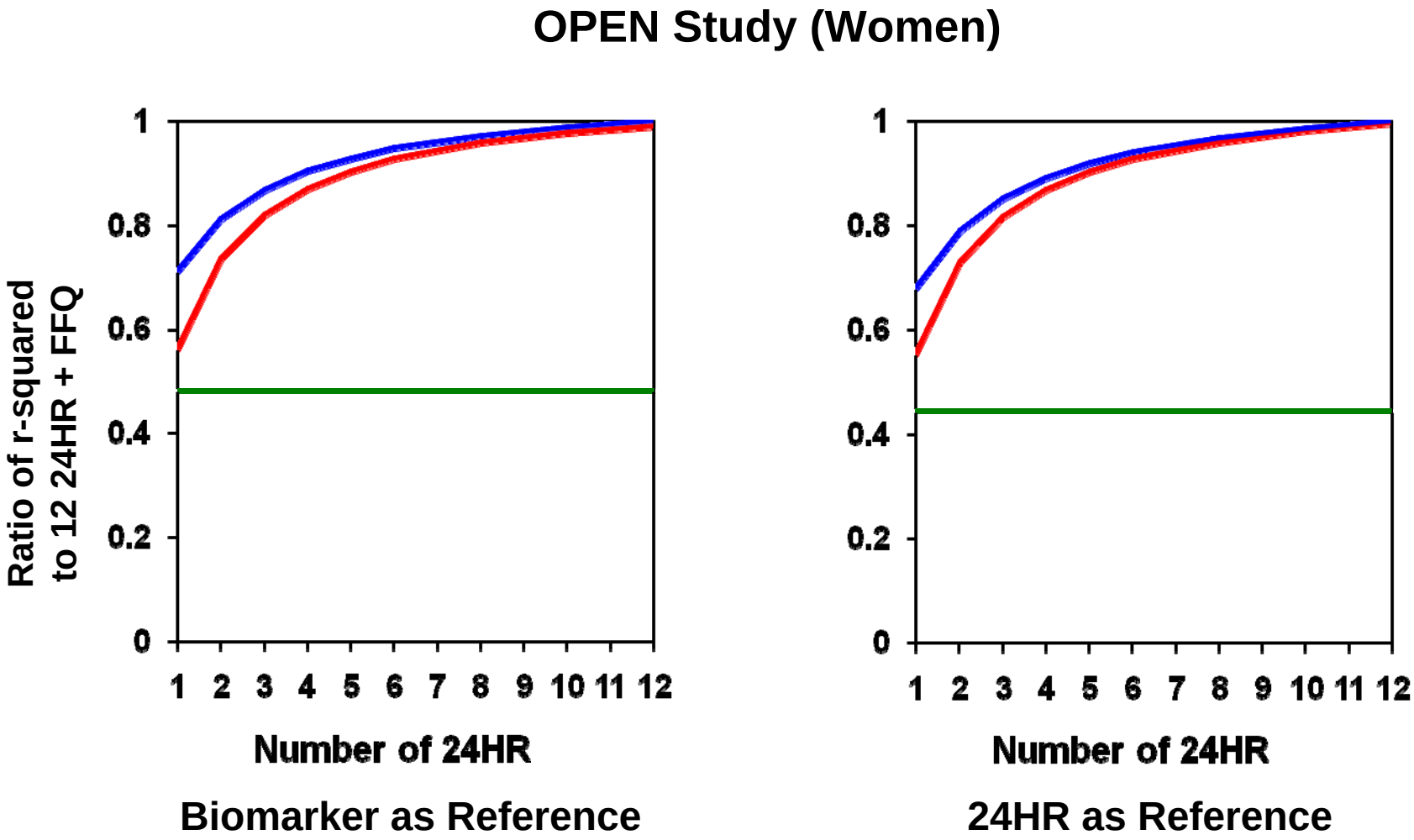
Ratio of R-squared values: protein



Slide 53

Here are the same two graphs for energy-adjusted protein in men. In this case, we see that the two graphs are not similar. In particular, the graph on the left that uses the biomarker as reference has a much higher R-squared value for the FFQ than the plot on the right that uses the 24 hour recall as reference. In this is case, assuming that the 24 hour recall is unbiased leads to underestimation of the predictive ability of the FFQ.

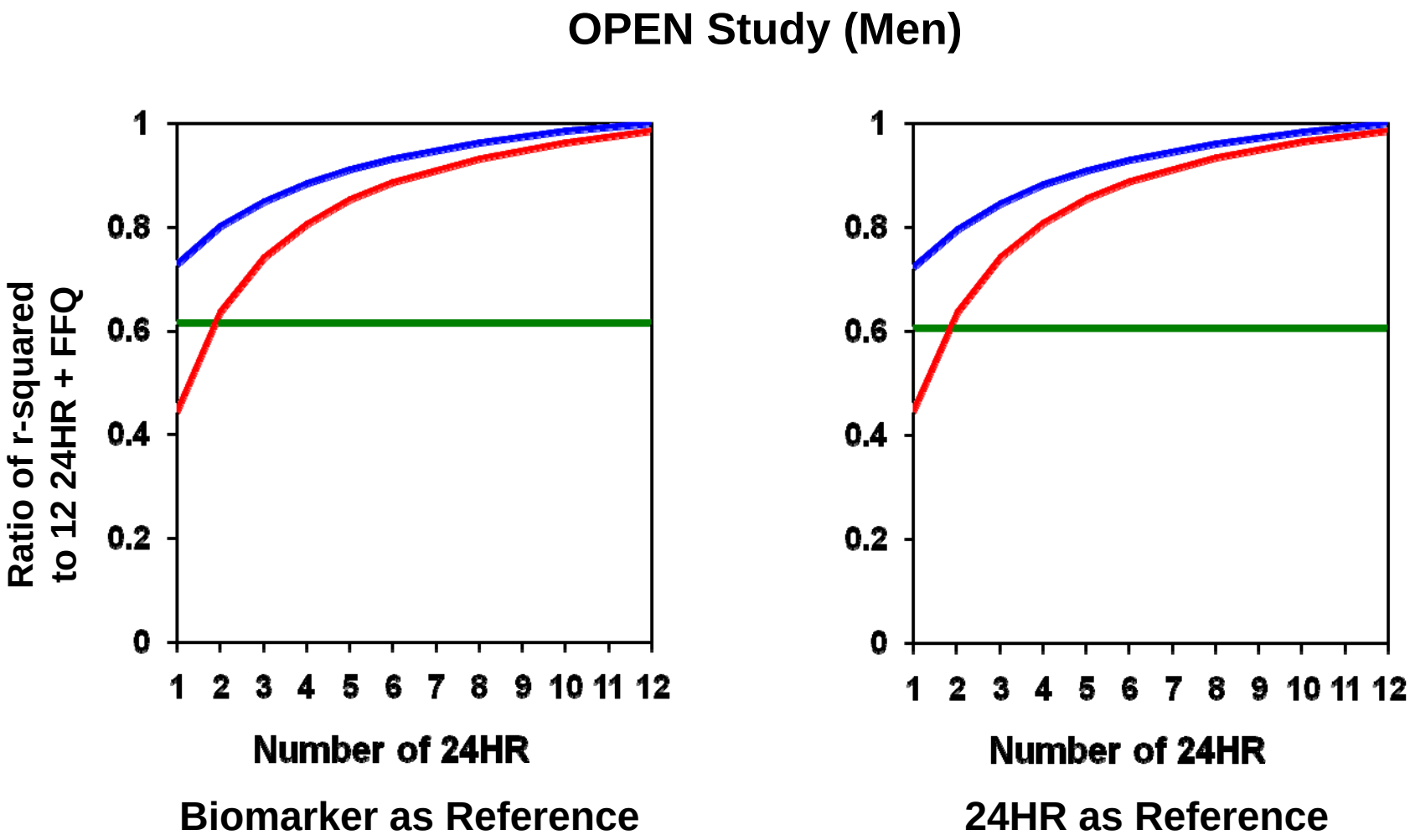
Ratio of R-squared values: potassium



Slide 54

Here are graphs of R-squared values for energy-adjusted potassium in women. In this example, the two graphs are again quite similar.

Ratio of R-squared values: potassium



Slide 55

Finally, here are graphs of R-squared values for energy-adjusted potassium in men. Again, we see they are quite similar.

Summary of OPEN study

- In 3 out of 4 cases, assuming 24HR is unbiased produces very similar comparisons as **reference biomarkers** known to be unbiased
- When comparing study designs assuming 24HR is unbiased
  - Conclusions about any **particular** dietary component may or may not be valid
  - Conclusions about **general patterns** that are consistent over many dietary components are probably valid

Slide 56

In summary, we found that in three out of four cases, using the 24 hour recall as a reference instrument led to results very similar to those using the unbiased biomarker as reference. In the fourth case, however, the results were quite different.

These results indicate that we need to be careful when using the 24 hour recall as a reference instrument. Conclusions about any particular dietary component may or may not be valid. Conclusions about general patterns that are consistent over many dietary components, however, are likely to be valid.

Is it worth the cost?

- Gain in precision vs. cost
- 24HR and FR impose substantial burden on participants
- New automated 24HR/FR reduce cost but not burden

Slide 57

As I mentioned earlier, our comparisons didn't consider the cost of the different study designs. When you are actually designing a study, however, you will need to ask: Is the gain in precision worth the cost of collecting more information? And by cost, we mean both the monetary cost of conducting the study and the burden imposed on the participants in the study. Twenty-four hour recalls and food records impose substantial burden on participants, and while new automated versions of these instruments may reduce cost, they won't reduce the burden on participants.

Other questions

- How many 24HR can we reasonably expect participants to complete?
  - Response rates?
  - Declining quality?
- Will automated 24HR perform as well as the traditional 24HR?
- Will FR perform similarly to 24HR?
  - Does a 4-day FR = four 24HR?

Slide 58

Here are a few other questions that we haven't addressed. First, how many 24 hour recalls can we reasonably expect a participant to complete before we start to get declining response rates or declining quality?

Second, our study design comparisons were based on the EATS study, which used traditional interviewer-administered 24 hour recalls. Will the new automated 24 hour recalls perform as well as traditional 24 hour recalls?

And, finally, how will food records perform in combination with other instruments? We weren't able to assess this, since the EATS study didn't have food records; we only had 24 hour recalls and FFQs. Will the food records give us results similar to those for the 24 hour recall? In particular, can we equate a four-day food record with four 24 hour recalls? These are questions we haven't addressed and can't really answer.

Looking for answers

- Biomarker studies designed to answer these questions (and more)
  - Six ASA24
  - Two FR
  - Two FFQ
  - Biomarkers of energy, protein and potassium
  - Also: ACT24 (physical activity), accelerometers, blood

Slide 59

But there are biomarker studies currently under way that have been designed to answer these questions. These studies will have six automated self-administered 24 hour recalls, two four-day food records, and two FFQs. They will also have biomarkers of energy, protein, and potassium. They will also measure physical activity using an automated 24 hour physical activity recall and accelerometer, and collect fasting bloods and a lot of other information.

These are exciting studies that are going to help us answer questions about response rates and whether automated 24 hour recalls perform as well as traditional 24 hour recalls, and whether or not food records perform similarly to ASA24.

Repeat FFQ?

- What about using more than one FFQ?
  - Less within-person variation, so less potential for gain in precision
  - Challenges in interpretation:

    Do differences in two FFQ taken 1 year apart reflect random within-person error or a real change in diet?

    How to define true usual intake if diet is changing over time?

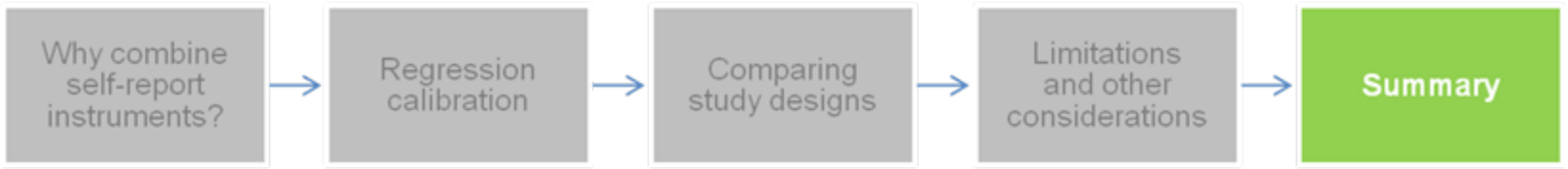
Slide 60

Finally, I wanted to mention the possibility of using more than one FFQ for each subject. We talked about having repeat 24 hour recalls, perhaps repeat food records, but what about repeating FFQs?

As I said earlier, one of the virtues of the FFQ is that it has less within-person variation than the 24 hour recall or food record. As a consequence, there is less potential for gain in precision by adding additional FFQs. This is because repeat application of an instrument is only useful for decreasing the within-person variation.

Also, the use of repeat FFQs leads to some challenges in interpretation. First, if two FFQs are administered one year apart, does the difference between them reflect a random, within-person error, or does it reflect a real change in diet?

Second, if the true intake is not static but is always changing over time, how should we define true usual intake? What is the time period that is relevant to diet-health associations?


Why combine self-report instruments?
Regression calibration
Comparing study designs
Limitations and other considerations
Summary

SUMMARY

Slide 61

I’d like to summarize the main points of this talk.

Summary

- Combining self-report dietary instruments can lead to significant improvement in estimating diet-health associations
- Regression calibration is an effective way to combine instruments

Slide 62

First, we saw that combining self-report dietary instruments can lead to significant improvement in estimating diet-health associations. We also saw that regression calibration is an effective way to combine instruments.

Summary

- 4-6 24HR capture most of the information available in 24HR
- Adding FFQ to 1 or more 24HR generally improves prediction more than adding another 24HR

Slide 63

We saw that four to six 24 hour recalls seem to capture most of the information that's available in 24 hour recalls. We also saw that adding an FFQ to one or more 24 hour recalls generally improves prediction more than adding an additional 24 hour recall.

Summary

- When designing diet-health studies, one should consider using FFQ plus 4-6 24HR to measure diet
- Other factors such as **cost** and participant **burden** must also be considered and balanced with need for **precision** and **power**

Slide 64

When designing a diet-health study, one should consider using a food frequency questionnaire and four to six 24 hour recalls to measure diet. Other factors such as cost and participant burden must also be considered and balanced with the need for precision and power.

Summary

- These conclusions:
  - Apply to estimating diet-health relationships (predicting individual intake)
  - Do not apply to estimating population distributions of dietary intake
- Tooze et al. (*J Am Diet Assoc*, 2006) found that adding FFQ to two 24HR did not improve estimated population distributions

Slide 65

And, finally, I want to emphasize that these conclusions apply to studies designed to estimate diet-health relationships. They don't apply to studies designed to estimate population distributions of dietary intake. These are two very different tasks that need to be evaluated differently.

In fact, in a 2006 article, Janet Tooze and her colleagues found that adding a food frequency questionnaire to two 24 hour recalls did not substantially improve estimated population distributions. They didn't do a comprehensive analysis, and there may be situations where adding an FFQ could improve estimates of the population distributions. Nevertheless, their results indicate that a food frequency questionnaire will have, at most, a limited role in estimating population distributions.

QUESTIONS & ANSWERS

Moderator: Amy Subar

Please submit questions using the *Chat* function

Slide 66

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

Measurement Error Webinar 10 Q&A

Question: **I believe you just answered the first question, which was: Will adding a food frequency questionnaire to two 24 hour recalls improve the precision of estimated population distributions? Am I correct in saying you just answered that?**

Yes, I would say that there is a possibility of getting a moderate gain, but it's not going to be as large as when you add 24 hour recalls, just because FFQs don't have as much within-person variation. *(D. Midthune)*

So in the example you showed from the OPEN study, did you also look at the R squares? Or were the R squares for energy also assessed?

Well, actually, no because we were concentrating on energy-adjusted nutrients and of course you can't energy-adjust energy, so we didn't look at that. *(D. Midthune)*

For cross-sectional data like in NHANES, can it also be assumed that self-report instruments have nondifferential error? And if not, what impact does this have for methods?

That's a difficult question. Typically, researchers believe that the assumption of nondifferential error is reasonable for cohort studies. For case-control studies, it may be unreasonable because of the possibility of recall bias; that's when people who have had some health event remember their past diet differently than those who haven't had that same event. And cross-sectional studies probably fall in-between [cohort and case-control studies], because the health event and the dietary assessments are obtained contemporaneously. So there are definitely problems where someone might have developed a disease and subsequently changed his/her diet. So I would have to say cross-sectional studies are more like case-control studies in that you have to worry about differential error. So you have to be very careful when you examine a cross-sectional study [for diet-health associations]. *(D. Midthune)*

When you were dealing with the EATS study and you talked about truth, how did you actually estimate truth in that study?

In the EATS study we had to assume that the 24 hour recall was unbiased because we didn't have any kind of biomarkers. And when we say it's unbiased, we don't mean that equals truth, but that equals truth plus some random within-person error, so that if you averaged over many, many 24 hour recalls for the same individual, you would get true intake. We used a measurement error model in which we assumed that the 24

hour recall equaled true intake plus within-person error, and we estimated all the parameters [in the measurement error model]. And from that, we were able to estimate, not truth itself, but the relationship between true and reported intake*. (D. Midthune)*

Does adding a screener to a 24 hour recall or a number of recalls have the same potential to improve precision and power as what you showed for adding an FFQ?

I would say it definitely has potential to improve the 24 hour recall. We'd have to look at it. We'd have to actually look at it in a study to see if it performed as well as a food frequency questionnaire, but certainly there is potential. *(D. Midthune)*

National Cancer Institute

measurement ERROR webinar series

Next Session

Tuesday, November 29, 2011
10:00-11:30 EST

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

Laurence Freedman
Gertner Institute

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

Slide 67

Thank you very much, Doug, and thanks to our audience for joining today's webinar. Please join us next week for webinar 11, when Dr. Laurence Freedman will discuss combining self-report intake data with biomarker data to reduce the effects of measurement error.