Data harmonization across large consortia: analytic challenges

Donna Spiegelman, Sc.D.
Departments of Epidemiology, Biostatistics, Nutrition and Global Health
Harvard School of Public Health
stdls@channing.harvard.edu

Stephanie Smith-Warner
Molin Wang
Harvard School of Public Health
Methods for Summarizing the Evidence

♦ Narrative review

♦ Systematic review

♦ Meta-analysis of published data

♦ Pooled analysis of primary data (meta-analysis of individual data)
  - Retrospectively planned
  - Prospectively planned
Pooled Analyses


Methods for pooling results of epidemiologic studies: the Pooling Project of Prospective Studies of Diet and Cancer.

Outline for Conducting Pooled Analyses

- Search strategy
- Study inclusion criteria
- Obtain primary data
- Prepare data for pooled analysis
- Estimate study-specific effects
- Examine whether results are heterogeneous
- Estimate pooled result
- Conduct sensitivity analyses

Friedenreich 1993
Pooling Project of Prospective Studies of Diet and Cancer

- Collaborative project to re-analyze the primary data in multiple cohort studies using standardized analytic criteria to generate summary estimates
  - *Retrospectively-planned* pooling project of individual patient data
- Established in 1991

http://www.hsph.harvard.edu/poolingproject/index.html
Pooling Project of Prospective Studies of Diet and Cancer: Inclusion Criteria

- Prospective study with a publication on diet and cancer
- Usual dietary intake assessed
- Validation study of diet assessment method
- Minimum number of cases specific for each cancer site examined
Cohort Studies in the Pooling Project of Prospective Studies of Diet and Cancer

Canadian National Breast Screening Study

Sweden Mammography Cohort Study

New York University Women’s Health Study

Netherlands Cohort Study

CA Teacher’s Study

Health Professionals Follow-up Study, Nurses’ Health Study, Women’s Health Study, Nurses’ Health Study II

Adventist Health Study

New York State Cohort

Iowa Women’s Health Study

ORDET

Cancer Prevention Study II Nutrition Cohort

Alpha-Tocopherol Beta-Carotene Cancer Prevention Study

Breast Cancer Detection Demonstration Project
Current Cohort Studies in the Pooling Project of Prospective Studies of Diet and Cancer

n = 32 studies, > 2.8 million participants
Analytic Strategy

- **Data collection**
  - Nutrients
  - Food groups
  - Foods
  - Non-dietary risk factors

- **Study-specific analyses**
  - Main effect
  - Population subgroups
  - Effect modification

- **Pooling**

- **Annual meeting**
Pooling Project: Primary Analysis Programs

♦ Read data
  • Study name, exposure, covariates

♦ Data management

♦ Analysis - Cox proportional hazards model (SAS)
  • Continuous
  • Categorical
  • Splines
  • Interactions
  • Automatically screen out covariates from the model with no variation or perfectly correlated with another covariate
Pooling Project: Pooling Programs

- Read study-specific output data sets
- Calculate summary relative risk by using random effects model
- Test for between studies heterogeneity
- Test for effect modification by sex and other key risk factors
- Output table
Fixed Effects

- Assumes that all studies are estimating the same underlying effect

- Variability only from sampling of people within each study

- Precision depends mainly on study size
Fixed Effects Model

\[ \hat{\beta}_s = \beta + \varepsilon_s, \ s = 1, \ldots, S \text{ studies} \]

\[ \beta = \text{true common log relative risk} \]
\[ \varepsilon_s = \text{random within-study (sampling variability)} \]
\[ \hat{\beta}_s = \text{estimated log relative risk in study } s \]
\[ E(\varepsilon_s) = 0, \ s = 1, \ldots, S \]
\[ \text{Var}(\varepsilon_s) = \text{Var}(\hat{\beta}_s) \]
\[ Pooled \hat{\beta} = \frac{\sum_{s=1}^{S} w_s \hat{\beta}_s}{\sum_{s=1}^{S} w_s}, \ w_s = [\text{Var}(\hat{\beta}_s)]^{-1} \]
\[ \text{Var}(\hat{\beta}) = \left( \sum_{s=1}^{S} [\text{Var}(\hat{\beta}_s)]^{-1} \right)^{-1} = [\sum_{s=1}^{S} w_s]^{-1} \]

- minimum variance weights
Random Effects

♦ Studies allowed to have different underlying effects, which vary around a mean over “all studies”

♦ Allows variation between studies as well as within studies
Random Effects Model

\[ \hat{\beta}_s = \beta + b_s + \varepsilon_s, \ s = 1, \ldots, S \text{ studies} \]

\[ \beta = \text{true common log relative risk} \]

\[ b_s = \text{random between-studies variability} \]

\[ \varepsilon_s = \text{random within-study (sampling variability)} \]

\[ \hat{\beta}_s = \text{estimated log relative risk in study } s \]

\[ E(\varepsilon_s) = E(b_s) = 0, \ s = 1, \ldots, S \]

\[ \text{Var}(\varepsilon_s) = \text{Var}(\hat{\beta}_s), \ s = 1, \ldots, S \]

\[ \text{Var}(b_s) = \sigma_B^2, \ s = 1, \ldots, S \]

\[ \text{Pooled } \hat{\beta} = \frac{\sum_{s=1}^{S} w_s \hat{\beta}_s}{\sum_{s=1}^{S} w_s}, \ w_s = \left[ \text{Var}(\hat{\beta}_s) + \hat{\sigma}_B^2 \right]^{-1} \]

\[ \text{Var}(\hat{\beta}) = \left( \sum_{s=1}^{S} \left[ \text{Var}(\hat{\beta}_s) + \hat{\sigma}_B^2 \right]^{-1} \right)^{-1} = \left( \sum_{s=1}^{S} w_s \right)^{-1} \]

- minimum variance weights
Under random effects model,
\[ \hat{\beta}_s = \beta + b_s + \varepsilon_s, \quad s = 1, \ldots, S \text{ studies} \]
\[ \text{Var}(b_s) = \sigma_B^2 \]
\[ \hat{\sigma}_B^2 = \max \left\{ 0, \frac{Q - (S - 1)}{\sum_{s=1}^{S} w_s - \frac{\sum_{s=1}^{S} w_s^2}{\sum_{s=1}^{S} w_s}} \right\} \]
\[ w_s = [\text{Var}(\varepsilon_s)]^{-1} \approx [\text{Vâr}(\hat{\beta}_s)]^{-1} \]
Pooling Study-Specific Results vs. Combining Primary Data into One Dataset

- Difficult to distinguish population-specific differences in true intake from artifactual differences due to differences in dietary assessment methods
  - exception: when the unit of measurement is standard (e.g. alcohol, body mass index)
- Pooling allows for study-specific differences of effects of confounders (otherwise residual confounding in aggregated analysis)
- When there is no confounding or effect modification, efficiency of 2 approaches is equivalent (Basagana, http://cdn1.sph.harvard.edu/wp-content/uploads/sites/271/2012/08/optitxs-The-Design-of-Observational-Longitudinal-Studies-APPENDIX.pdf, Appendix A.3)
Test for Heterogeneity

Under random effects model,

\[ Q = \sum_{s=1}^{S} w_s (\hat{\beta} - \hat{\beta}_s)^2, \]

\[ \hat{\beta} = \text{pooled log relative risk, fixed effects model} \]

\[ \hat{\beta}_s = \text{study-specific estimated log relative risk} \]

\[ w_s = [\text{Var}(\hat{\beta}_s)]^{-1} \]

\[ Q \sim \chi^2_{S-1} \text{ under } H_0 : \sigma_B^2 = 0 \]
Fixed Effects vs Random Effects Model

♦ Random effects generally yield larger variances and CI
  • Why? Incorporate $\sigma^2_B$

♦ If heterogeneity between studies is large, $\sigma^2_B$ will dominate the weights and all studies will be weighted more equally

♦ Model weight for large studies less in random vs fixed effects model

♦ Controversial issue – area for further research – more interpretable weights under heterogeneity
Sources of Between Study Heterogeneity

- Different study designs
- Different length of follow-up
- Different distributions of effect modifiers
- Different statistical methods/models used
- Different sources of bias
- Study quality
Meta-Analyses: Sensitivity Analyses

- Exclude studies with particular heterogeneous results
- Conduct separate analyses based on study-specific features
  - Study design
  - Geographic location
  - Time period
  - Study quality
  - Key risk factors hypothesized to be possible modifiers
Meta-regression (Stram, Biometrics, 1996)

**Purpose**: to identify heterogeneity of effects by covariates that are constant within study (e.g. gender, smoking status)

**Model**: \( \hat{\beta}_s = \beta_0 + \beta_1 \times \text{GENDER}_s + \beta_2 \times \text{CURRENT}_s + \beta_2 \times \text{PAST}_s + b_s + \epsilon_s \)

- \( \text{GENDER}_s = 1 \) if study \( s \) is male; 0 if female
- \( \text{CURRENT}_s = 1 \) if study \( s \) has current smokers only, 0 otherwise
- \( \text{PAST}_s = 1 \) if study \( s \) has past smokers only, 0 otherwise

\[ H_0: \beta_1 = 0 \Rightarrow \text{no effect-modification by gender} \]

Standard methods for mixed effects models can be used to test hypotheses and estimate parameters.
Analysis of between-studies heterogeneity

- *p*-value for test for heterogeneity is a function of the power of the pooled analysis to detect between-studies differences. This power is believed to be low.

- A simulation study was conducted and published (Takkouche, Cardoso-Suárez, Spiegelman, AJE, 1999) which investigated the power of several old and some newly developed test statistics to detect heterogeneity of different plausible magnitudes, as quantified by $CV_B$ and $R^2$ (to be defined).

- $CV_B = \sigma_B / |\beta|$, with S=ranging from 7 to 33
Analysis of between-studies heterogeneity

- Explored maximum likelihood methods for estimation of \( \sigma_B^2 \) and testing \( H_0: \sigma_B^2 = 0 \) (i.e. no between-studies heterogeneity)

  - ML methods have power roughly equivalent to D&L’s, but assume \( b_s \sim N(0, \sigma_B^2) \) and \( \varepsilon_s \sim N(0, \text{Var}_s(\beta_s)) \)

  - LRT for \( H_0: \sigma_B^2 = 0 \) has no known asymptotic distribution because hypothesis is on the boundary of the parameter space

  - A simulation-based bootstrap approach for constructing the empirical distribution function of the test statistic was developed
Quantification of heterogeneity

\[ CV_B = \frac{\sigma_B}{|\beta|} \]

- between-studies variance expressed relative to the magnitude of the overall association
- if the association is small, CV 'blows up'

\[ R^2 = \frac{\sigma_B^2}{\sigma_B^2 + \sum_{s=1}^{S} \text{Var}(\hat{\beta}_s) / S} \approx I^2 = \frac{(Q - df)}{Q} \]

- proportion of the variance of the pooled estimate due to between-studies variation

- Useful for when \( \beta \) is near 0 as well as when it is far from it
- Heterogeneity is evaluated relative to within-studies contribution to the variance, and can appear large if the participating studies yield precise estimates

- Further experience with these measures will give us more insight as to their relative merits

- Confidence intervals for \( CV_B \), \( R^2 \) (Takkouche et al., 2013); \( I^2 \) (Higgins et al., 2013)
## Power of Test of Heterogeneity

<table>
<thead>
<tr>
<th></th>
<th>Q</th>
<th>LRT†</th>
<th>Q*</th>
<th>LRT*</th>
<th>$\tau^2$-bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_I$§ = 0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S = 7$</td>
<td>14.38</td>
<td>7.85</td>
<td>12.18</td>
<td>14.68</td>
<td>15.32</td>
</tr>
<tr>
<td>$S = 20$</td>
<td>25.90</td>
<td>17.90</td>
<td>24.18</td>
<td>25.97</td>
<td>24.92</td>
</tr>
<tr>
<td>$S = 40$</td>
<td>38.12</td>
<td>30.64</td>
<td>35.10</td>
<td>37.81</td>
<td>37.94</td>
</tr>
<tr>
<td>$R_I$ = 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S = 7$</td>
<td>37.54</td>
<td>27.14</td>
<td>27.60</td>
<td>39.10</td>
<td>38.50</td>
</tr>
<tr>
<td>$S = 20$</td>
<td>70.56</td>
<td>64.11</td>
<td>58.54</td>
<td>69.21</td>
<td>68.62</td>
</tr>
<tr>
<td>$S = 40$</td>
<td>90.88</td>
<td>88.61</td>
<td>79.76</td>
<td>90.21</td>
<td>91.14</td>
</tr>
<tr>
<td>$R_I$ = 0.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S = 7$</td>
<td>77.66</td>
<td>70.21</td>
<td>58.84</td>
<td>78.31</td>
<td>75.02</td>
</tr>
<tr>
<td>$S = 20$</td>
<td>98.66</td>
<td>97.51</td>
<td>93.64</td>
<td>99.12</td>
<td>98.34</td>
</tr>
<tr>
<td>$S = 40$</td>
<td>100.00</td>
<td>99.93</td>
<td>99.64</td>
<td>99.94</td>
<td>99.94</td>
</tr>
</tbody>
</table>

* Parametric bootstrap version of the test.
† Odds ratio = 2.
‡ WLS, weighted least squares; LRT, likelihood ratio test.
§ $R_I$, proportion of the total variance due to between-study variance: $\sigma^2_B/((\sigma^2_B + (S \times \text{Var}(\beta))))$. 
Multivariate meta-analysis for data consortia, individual patient meta-analysis, and pooling projects

Ritz J, et al., *Journal of Statistical Planning and Inference*, 2008;

- Maximum likelihood and estimating equations methods for combining results from multiple studies in pooling projects

- The univariate meta-analysis model is generalized to a multivariate method, and efficiency advantages are investigated

- The test for heterogeneity is generalized to a multivariate one
**Multivariate Pooling**

Let $\mathbf{\hat{\beta}}_s = \mathbf{\beta} + \mathbf{b}_s + \mathbf{e}_s + +$, $s = 1, \ldots, s$ studies

$\text{dim} (\mathbf{\beta}) = p \text{ model covariates}$

$E(\mathbf{b}_s) = E(\mathbf{e}_s) = 0$, $\text{Var}(\mathbf{b}_s) = \Sigma_B$, $\text{Var}(\mathbf{e}_s) = \text{Cov}(\mathbf{\hat{\beta}}_s)$

$p \times p$

- If $\Sigma_B$ and $\text{Cov}(\mathbf{\hat{\beta}}_s)$ are not diagonal, more efficient estimates of $\mathbf{\beta}$ can be obtained

- Weighted estimating equation ideas are used

- Test statistics for $H_0: \Sigma_B = 0$

and major submatrices are given assuming normality in $\mathbf{b}_s$ and $\mathbf{e}_s$ and asymptotically
Table 1: Score functions for $\beta$ and $\Sigma_b$

<table>
<thead>
<tr>
<th>Method</th>
<th>Score for $\beta$</th>
<th>Score for $vech(\Sigma_b)$</th>
<th>Computational options</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVML</td>
<td>$\sum_{s=1}^{S} W_s (\hat{\beta}_s - \beta)$</td>
<td>$-\frac{1}{2} \cdot \sum_{s=1}^{S} [(\hat{\beta}<em>s - \beta)' V^{(s)}</em>{ij} (\hat{\beta}<em>s - \beta) + v^{(s)}</em>{ij}]$</td>
<td>FSQP, EM</td>
</tr>
<tr>
<td>REML</td>
<td>$\sum_{s=1}^{S} W_s (\hat{\beta}_s - \beta)$</td>
<td>$-\frac{1}{2} \cdot \sum_{s=1}^{S} (\hat{\beta}<em>s - \beta)' V^{(s)}</em>{ij} (\hat{\beta}_s - \beta)$</td>
<td>FSQP, EM</td>
</tr>
<tr>
<td>MVEE1</td>
<td>$\sum_{s=1}^{S} W_s (\hat{\beta}_s - \beta)$</td>
<td>$\sum_{s=1}^{S} [(\hat{\beta}<em>{si} - \beta_i)(\hat{\beta}</em>{sj} - \beta_j) - (\sigma_{bij} + \sigma_{sij})]$</td>
<td>Minimize SS via FSQP</td>
</tr>
<tr>
<td>MVEE2</td>
<td>$\sum_{s=1}^{S} W_s (\hat{\beta}_s - \beta)$</td>
<td>$\sum_{s=1}^{S} [(\hat{\beta}<em>{si} - \beta_i)(\hat{\beta}</em>{sj} - \beta_j) - (S - 1)(\sigma_{bij} + \sigma_{sij}/S)]$</td>
<td>Minimize SS via FSQP</td>
</tr>
<tr>
<td>MVEE3</td>
<td>$\sum_{s=1}^{S} W_s (\hat{\beta}_s - \beta)$</td>
<td>$\sum_{s=1}^{S} w_s [(\hat{\beta}<em>{si} - \beta_i)(\hat{\beta}</em>{sj} - \beta_j) - (\sigma_{bij} + \sigma_{sij})]$</td>
<td>Minimize SS via FSQP</td>
</tr>
<tr>
<td>MVEE4</td>
<td>$\sum_{s=1}^{S} (\hat{\Sigma}_b + \Sigma_s)^{-1}(\hat{\beta}_s - \beta)$</td>
<td>N/A</td>
<td>Direct solution</td>
</tr>
<tr>
<td>MVEE5</td>
<td>$\sum_{s=1}^{S} (\hat{\Sigma}_b + \Sigma_s)^{-1}(\hat{\beta}_s - \beta)$</td>
<td>N/A</td>
<td>Direct solution</td>
</tr>
</tbody>
</table>

Notes:

$W_s = (\Sigma_b + \Sigma_s)^{-1}$

$-W_s F_{ij} W_s, F_{ij} = E_{ij} + E_{ji} - \delta_{ij} E_{jj}, E_{ij} = e_i e_j'$

$ij$th element of $V^{(s)}_{ij}$

$w_s^{-1} = \text{var}[(\hat{\beta}_{si} - \beta_i)(\hat{\beta}_{sj} - \beta_j)] = (\sigma_{sii} + \sigma_{sii})(\sigma_{sjj} + \sigma_{sij}) + (\sigma_{bij} + \sigma_{sij})^2$ (normal assumption)

$SS = S(\gamma)' S(\gamma)$, $S(\gamma)' = (S(\beta), S(\theta))$, $\gamma' = (\beta, \theta)$

$\hat{\Sigma}_b = \left\{ \sum_{s=1}^{S}(\hat{\beta}_s - \beta)(\hat{\beta}_s - \beta)' - \Sigma_s \right\} / S$

$\hat{\Sigma}_b^* = \sum_{s=1}^{S} \left\{ \frac{1}{S - 1}(\hat{\beta}_s - \beta)(\hat{\beta}_s - \beta)' - \frac{1}{S} \Sigma_s \right\}$
Results
From Smith-Warner, et al. 2001

“Types of dietary fat and breast cancer: a pooled analysis of cohort studies”

ARE (compared to corresponding univariate estimate)

<table>
<thead>
<tr>
<th></th>
<th>MLE</th>
<th></th>
<th>EE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=3</td>
<td>p=18*</td>
<td>p=3</td>
<td>p=18*</td>
</tr>
<tr>
<td>Saturated fat</td>
<td>86%</td>
<td>81%</td>
<td>93%</td>
<td>84%</td>
</tr>
<tr>
<td>Mono-unsaturated fat</td>
<td>76%</td>
<td>72%</td>
<td>36%</td>
<td>32%</td>
</tr>
<tr>
<td>Poly-unsaturated fat</td>
<td>91%</td>
<td>88%</td>
<td>1.35%</td>
<td>1.21%</td>
</tr>
</tbody>
</table>

*Using estimates adjusted for total energy, % protein intake, bmi, parity x age at first birth, height (4 levels), dietary fiber (quintiles)
Measurement error correction for pooled estimates

- Measurement error correction study-specific estimate using the measurement error model developed from each study's validation data

- Pool measurement error corrected $\beta$'s in the usual way -- note that pooled variance will reflect additional uncertainty in the study-specific measurement error corrected estimates

- Works best for continuous exposures
Additional analytic issues

Rare cancers (e.g. kidney, ovarian)
  - Include study if 50 or more cases
  - Use propensity score to control for confounding if the number of cases in the study is < 500
  - Propensity score is $\Pr_i (E = 1 | all\ confounders)$ from logistic regression
  - Continuous generalization $\widehat{E}_i (E | all\ confounders)$ from linear

Additional analytic issues

- How to pool smooth exposure-response curves?
  - Aggregate – has all the disadvantages of aggregation
  - Meta-analysis for dose-response – categories are limited, could miss important effects in the tails
  - New methods needed
Additional analytic issues

- **Zero cell problem** --- an extreme category has no cases
  - **Collapse with closest category** – bias effect towards the null
  - **Delete data in this category from analysis** – unbiased if exposure-response is non-linear but lose power
  - **Exact analysis**  
    Tian L, Cai T, Pfeffer M, Piankov N, Cremieux P and Wei LJ. Exact and Efficient Inference Procedure for Meta-analysis And Its Application to the Analysis of Independent 2 x 2 Tables With All Available Data But Without Artificial Continuity Correction. Biostatistics, 10:275-281, 2009
Example: Fruits and vegetables in relation to breast cancer risk

- Meta-analysis
- Pooled analysis
Fruit & Vegetable and Breast Cancer Meta-Analysis

♦ Objective: analyze published results that explore the relationship between breast cancer risk and the consumption of fruits and vegetables

♦ Search strategy
  • MEDLINE search of studies published January 1982 – April 1997
  • Review of reference lists

Gandini, 2000
Fruit & Vegetable and Breast Cancer Meta-Analysis: Inclusion Criteria

- Relative risks and confidence intervals reported or could be estimated
  - Comparisons: tertiles, quartiles, quintiles
- Studies were independent
- Diet assessed by food frequency questionnaire
- Populations were homogeneous, not limited to specific subgroup
Selected risk estimate for total fruits and total vegetables, when possible
  • Otherwise, selected nutrient dense food

Extracted the most adjusted relative risk comparing the highest vs. lowest intake
  • Comparisons: tertiles, quartiles, quintiles

Used random effects model to calculate summary estimate

Sensitivity analyses
<table>
<thead>
<tr>
<th></th>
<th>RR (95% CI)</th>
<th>p for het</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>high vs low</td>
<td></td>
</tr>
<tr>
<td>Total fruits</td>
<td>0.94 (0.79-1.11)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Total vegetables</td>
<td>0.75 (0.66-0.85)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
## Fruits and Vegetables and Breast Cancer Meta-Analysis: Sensitivity Analysis

<table>
<thead>
<tr>
<th>Study design</th>
<th># of Studies</th>
<th>RR</th>
<th>Significance of factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-control</td>
<td>14</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>Cohort</td>
<td>3</td>
<td>0.73</td>
<td>0.30</td>
</tr>
<tr>
<td>Validated FFQ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>6</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>11</td>
<td>0.66</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Gandini, 2000
Fruits and Vegetables and Breast Cancer: Conclusion of Meta-Analysis

- The quantitative analysis of the published studies...suggests a moderate protective effect for high consumption of vegetables...For fruit intake, study results were less clear. Only two studies show a significant protective effect of high fruit intake for breast cancer.

- This analysis confirms the association between intake of vegetables and, to a lesser extent, fruits and breast cancer risk from published sources.
Pooling Project: Exposures

- Total fruits
  - Fruits, excluding fruit juice
  - Fruit juice
- Total vegetables
- Total fruits and vegetables
- Botanical groups
- Specific foods
# Pooled Multivariate Relative Risks for Breast Cancer and Fruits and Vegetables

<table>
<thead>
<tr>
<th>Quartile</th>
<th>Total Fruits RR (95% CI)</th>
<th>Total Vegetables RR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
</tr>
<tr>
<td>2</td>
<td>0.94 (0.87-1.01)</td>
<td>0.99 (0.90-1.08)</td>
</tr>
<tr>
<td>3</td>
<td>0.92 (0.86-0.99)</td>
<td>0.97 (0.90-1.05)</td>
</tr>
<tr>
<td>4</td>
<td>0.93 (0.86-1.00)</td>
<td>0.96 (0.89-1.04)</td>
</tr>
</tbody>
</table>

p for trend: 0.08 0.54
p for hetero.: 0.94 0.73

Smith-Warner, et al. 2001
### Contrast to Vegetable and Breast Cancer Meta-Analysis: Results

<table>
<thead>
<tr>
<th></th>
<th>RR (95% CI)</th>
<th>p for het</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total fruits</strong></td>
<td>0.94 (0.79-1.11)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Total vegetables</strong></td>
<td>0.75 (0.66-0.85)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fruit and Vegetable Group</td>
<td>RR (95% CI)</td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>-------------------</td>
<td></td>
</tr>
<tr>
<td>Cruciferae</td>
<td>0.96 (0.87-1.06)</td>
<td></td>
</tr>
<tr>
<td>Leguminosae</td>
<td>0.97 (0.87-1.08)</td>
<td></td>
</tr>
<tr>
<td>Rutaceae</td>
<td>0.99 (0.97-1.01)</td>
<td></td>
</tr>
<tr>
<td>Carrots</td>
<td>0.95 (0.81-1.12)</td>
<td></td>
</tr>
<tr>
<td>Potatoes</td>
<td>1.03 (0.98-1.08)</td>
<td></td>
</tr>
<tr>
<td>Apples</td>
<td>0.96 (0.92-1.01)</td>
<td></td>
</tr>
</tbody>
</table>

Increment=100 g/d

Smith-Warner 2001
## Pooled Multivariate Relative Risks of Breast Cancer and F & V by Menopausal Status

<table>
<thead>
<tr>
<th></th>
<th>Multi RR (95% CI) (increment=100 g/d)</th>
<th>p-for interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total fruits</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premeno.</td>
<td>0.98 (0.94-1.02)</td>
<td></td>
</tr>
<tr>
<td>Postmeno.</td>
<td>0.99 (0.98-1.01)</td>
<td>0.80</td>
</tr>
<tr>
<td><strong>Total vegetables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premeno.</td>
<td>0.99 (0.93-1.06)</td>
<td></td>
</tr>
<tr>
<td>Postmeno.</td>
<td>1.00 (0.97-1.02)</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Smith-Warner, et al. 2001
Pooled Analyses vs. Meta-Analyses

♦ Strengths
  • Increased standardization
  • Can examine rare exposures
  • Can analyze population subgroups
  • Reduce publication bias

♦ Limitations
  • Expensive
  • Time-consuming
  • Requires close cooperation with many investigators
  • Errors in study design multiplied (prospective)
Why Are Pooled Analyses Time-Consuming?

- Data management
- Add updated case information
- Errors found in data
- Individual study wants to publish their findings prior to submitting pooled results
- Many large data sets, computational time to complete the many analyses can be long
- Manuscript review
- Signature sheets
Challenges for Individual Cohorts Participating in Pooled Analyses

- Overlap with specific aims of primary cohort grant
- Inadequate resources for participating cohorts
  - Data management
  - Cohort maintenance
  - Investigator time
- Data harmonization – differences in assessment of variables leads to use of the least common denominator
- Recognition of contribution of individual investigators
- Conserving biobank specimens

From presentations at Cohorts and Consortia meeting, 2009
REFERENCES


