Adventist Health Study (AHS2)
Pilot linkage study

US and Canada-wide
Special population with diet/cancer focus
Research opportunity for NAACCR members

Susan Preston-Martin, PhD¹
Gary Fraser, MD, PhD², Larry Beeson, DrPH²

¹ University of Southern California, Los Angeles, CA
² Loma Linda University, Loma Linda, CA
Goals of the Research

1. Effect of consuming soy products on risk of colon, breast and prostate cancers.

2. Effect of dietary and supplemental calcium on risk of these cancers.

3. Effect of meat and fat consumption on risk of these cancers.

4. Particular focus on African-Americans
Soy Milk and Incidence of Prostate Cancer

Relative Risk

<table>
<thead>
<tr>
<th>Soy Milk</th>
<th># men</th>
<th>Never</th>
<th>&lt;Daily</th>
<th>Daily</th>
<th>&gt;Daily</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10875</td>
<td>902</td>
<td>395</td>
<td>223</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0.9</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5-1.4</td>
<td>0.4-1.4</td>
<td>0.1-0.9</td>
</tr>
</tbody>
</table>

P = 0.02
AHS 2 Methods

1. To enroll **70,000 White and 35,000 Black Adventists** across US and Canada.

2. Recruitment **church-by-church** (>4000 churches)

3. Also National promotional campaigns using church media.

4. Enrollment over a 6-year period.
Enrollment simply requires completion of the questionnaire

- Past medical history
- **Diet**
- Physical activity
- Female questions
- Demographic questions
- Tear-off section for personal information
ADVENTIST HEALTH STUDY - 2

Connecting Lifestyle to Disease and Longevity

- a continuation of previous research on Adventists

Please return to:

Adventist Health Study - 2
Loma Linda University
Evans Hall - Room 203
Loma Linda, CA 92350

IT'S NOT AS LONG AS IT LOOKS!
See note on next page

Please return to: Adventist Health Study - 2
Loma Linda University
Evans Hall - Room 207
Loma Linda, CA 92350

Are you age 30 or older? If so, please complete this card now, tear it out, and give it to the Adventist Health Study coordinator. Thank you.

Name: 
Address: 
Telephone: 
Date of Birth: 
Local church at which you hold membership:
A. MEDICAL HISTORY

1. Would you say, in general, your health is:
   - Excellent
   - Fair
   - Good
   - Poor

2. During the past 12 months, have you had influenza (flu)?
   - Yes
   - No

3. During the past 12 months, how many times have you had:
   - Upper respiratory infection (cold, sore throat, sinusitis):
     - None
     - 1-2
     - 3-4
     - 5-6
     - 7+
   - Gastroenteritis (stomach flu, food poisoning):
     - None
     - 1-2
     - 3-4
     - 5-6
     - 7+
   - Bronchitis:
     - None
     - 1-2
     - 3-4
     - 5-6
     - 7+

4. Has a physician ever told you that you had any form of cancer (including leukemia, lymphoma, myeloma and skin cancer)?
   - No
   - Yes
   - Go to Question 5, on the next page
   - If YES, write-in site of cancer(s) (e.g. lung, colon, etc.) and approximate year it was first diagnosed in the space(s) below. Also fill in the matching circles below the year first diagnosed. Notice that there is space for up to three different cancers.

Please use CAPITAL LETTERS.

<table>
<thead>
<tr>
<th>Site of cancer:</th>
<th>Year first diagnosed:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>1987</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site of cancer:</th>
<th>Year first diagnosed:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>2000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site of cancer:</th>
<th>Year first diagnosed:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lungs</td>
<td>2001</td>
</tr>
</tbody>
</table>
Smoking

Percentage

Black Subjects (N=10622)
White Subjects (N=39376)
Nurses' Health Study (N=78206)

Never smoked
Past smokers
Current smokers

Frequency
Alcohol Use

Frequency

Percentage

- Black Subjects: N=10622
- White Subjects: N=39376
- Nurses' Health Study: N=89046

Never drank
≤ 1 per month
≤ 1 per week
1+ per week
Coffee

Black Subjects (N=11407)
White Subjects (N=39948)
Nurses' Health Study

Frequency

Never
≤ 1 per week
2-6 per week
≥ 1 per day

Percentage
Beef, Lamb as Main Dish

Percentage

Frequency

Black Subjects (N=11407)
White Subjects (N=39948)
Nurses' Health Study

Never < 1 per week 2-4 per week 5-6 per week > 1 per day
Pilot study of linkage to registry files

• Why do?
  – To work out details of linking AHS2 to registry files –
    develop common procedures
  – Confirm self-reported prevalent cancers
  – Show we can do for grant renewal application

• Why these six states?
  – Recruitment finished by 2004; #s of both black and
    white Adventists
  – Range of states by size, region, age of registry
Percent of U.S. Adventists by state and race, AHS2 pilot linkage to state cancer registries planned for late 2004; completed late May 2005
Obtaining approval at pilot registries

• Time (first contact to final) - 4-16 months
• Longer than expected in all 6 states
• Reasons for unexpected delay
  – IRB form changed
  – Registry director changed
  – Additional approvals needed after IRB
  – Staff change
  – Application forgotten in drawer
  – Additional info requested after initial review
Traveling Linkage Programmers

• **AHS-2 programmers** work at each registry
• On-site linkage on **AHS laptop** PC
  • using Link Plus software
  • AHS-2 files on PC
  • **PC left for registry use**
• Only cancer data on matches taken to LLU
• Linkage done with or without supervision by registry personnel
Link Plus record linkage software

• Developed at CDC

• Computes probabilistic linkage scores ( < 5; 5-24; 25+)

• Facilitates a simple and efficient “OR” blocking mechanism by matching pairs on at least one key variable
  - AHS2 used SS#, surname, and birthdate

• Output is match +/- text files easily imported into Excel

• AHS2 supplementary program run on all files with scores 5-24 and displays linked pairs on the screen for review:
  - Definite match or non-match
  - Still undecided (i.e. query) – often missing data
Link Plus Comparisons

- **Frequency-based sets weights** for matching score:
  - **Frequent** value $\rightarrow$ **low** weight
  - **Rare** value $\rightarrow$ **high** weight

- **Last and first name** comparison uses **both partial and frequency-based matching** - typos, misspellings, hyphens

- **Middle name** either full name or initial OK

- **SSN** comparison **allows partial matching** to account for typos and transposition of digits

- **Date** comparison **allows partial matching** if missing values
AHS2 Custom Matching Code

The matching code is composed of 12 digits. Each digit represents result from comparing a pair of fields.

Example Code: 210191117891

<table>
<thead>
<tr>
<th>Codes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No match</td>
</tr>
<tr>
<td>1</td>
<td>Good match</td>
</tr>
<tr>
<td>2</td>
<td>Probable match</td>
</tr>
<tr>
<td>7</td>
<td>Missing CR data</td>
</tr>
<tr>
<td>8</td>
<td>Missing AHS data</td>
</tr>
<tr>
<td>9</td>
<td>Missing both</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position / Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) First name</td>
</tr>
<tr>
<td>2) Last name</td>
</tr>
<tr>
<td>3) Middle name</td>
</tr>
<tr>
<td>4) Gender</td>
</tr>
<tr>
<td>5) Date of birth</td>
</tr>
<tr>
<td>6) Social security number</td>
</tr>
<tr>
<td>7) Address number</td>
</tr>
<tr>
<td>8) Address street name</td>
</tr>
<tr>
<td>9) City</td>
</tr>
<tr>
<td>10) Zip</td>
</tr>
<tr>
<td>11) Phone area code</td>
</tr>
<tr>
<td>12) Phone local code</td>
</tr>
</tbody>
</table>
Clerical review of uncertain matches

<table>
<thead>
<tr>
<th>ID's</th>
<th>Link Plus Score</th>
<th>Matching Code</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>100000001</td>
<td>20.8</td>
<td>120191117891</td>
<td></td>
</tr>
<tr>
<td>10000002</td>
<td></td>
<td>1 of 1 records</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Middle Name</th>
<th>DOB</th>
<th>SSN</th>
<th>Address</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gary</td>
<td>Frazer</td>
<td>E</td>
<td>9</td>
<td>213978038</td>
<td>25169 Huron St</td>
<td>7961626</td>
</tr>
<tr>
<td>Dr. Gary</td>
<td>Fraser</td>
<td>D</td>
<td></td>
<td>8038</td>
<td>25169 Huron Street</td>
<td>7961626</td>
</tr>
<tr>
<td>City</td>
<td>Zip</td>
<td>Phone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loma Linda</td>
<td>92354</td>
<td>7961626</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Years linked at six pilot registries

Range: all years **1981-2003**
(only MI link used all these years)

first year (**1981-1995**)
most recent year (**2001-2004**)
Possible matches still unresolved

• **Range** [possible/(definite+possible)]
  across states (<5% - >80%)

• **Key missing data** (>50% of unresolved)
  
  **Address** (3 states)
  **Phone #** (3 states)
  **SS#** (AHS2 and 2 states)*
  **Middle name** or initial (3 states)

* 4 states if missing in AHS2, didn’t record if also missing in registry file
Cancers that would have been missed if only matched at one state

- **132 Adventists** with cancer *(8.5%)* total
  (range: 1-82; 3-14%)

- **144 Cancers** missed *(8.4%)* across 6 states

- **Need to match with entire AHS2 file**
How good is self report of cancer?

Cancers identified by *registry only* – 268/1657 (16%) (46 breast, 31 colon, 50 prostate, 141 other)

<table>
<thead>
<tr>
<th>Site</th>
<th>Site same</th>
<th>Site different</th>
<th>Cancer not self-reported</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1389</td>
<td>56</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td>84%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy of confirmed self reported cancer 1657

<table>
<thead>
<tr>
<th>Year of diagnosis</th>
<th>Year exact</th>
<th>+/- 1 year</th>
<th>&gt; 1 year</th>
<th>Cancer not self-reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>exact</td>
<td>978</td>
<td>278</td>
<td>203</td>
<td>198</td>
</tr>
<tr>
<td>+/- 1 year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 1 year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer not self-reported</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>59%</td>
<td>59%</td>
<td>59%</td>
<td>59%</td>
</tr>
</tbody>
</table>
## Sensitivity by cancer site

(\#self reports / \#registry matches)

<table>
<thead>
<tr>
<th>Cancer Site</th>
<th>Sensitivity</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>91.0%</td>
<td>(466/512)</td>
</tr>
<tr>
<td>Colon, rectum, anus</td>
<td>80.3%</td>
<td>(126/157)</td>
</tr>
<tr>
<td>Prostate</td>
<td>88.6%</td>
<td>(387/437)</td>
</tr>
<tr>
<td>All other (except skin)*</td>
<td>72.6%</td>
<td>(321/442)</td>
</tr>
<tr>
<td>Total cancers</td>
<td>83.8%</td>
<td>(1389/1657)</td>
</tr>
</tbody>
</table>

* i.e., other cancers excludes malignant melanoma and other skin cancer
Positive Predictive Value by Site
(#registry matches / # self reports)*

- Breast
- Colon
- Prostate
- All other (except skin)
- Total cancers

* among those who lived within 10 miles of current residence during match period
Variation by Demographics
(race, age, sex, year, and state)
Schedule for Recruitment

2002
2003
2004
2005-6
2005-7
Years we hope to do links

• First link
  – already done in 2005 in 6 pilot states
    2006-2007 in rest of US and Canada

• Link to identify incident cancers
  2008-2010
Measures to assure Confidentiality during the linkage process

• Only personal identifiers we’ll take away will be ones we brought on AHS2 file
• Will bring AHS2 back-up file to resolve dubious matches
• Portable AHS2 PC used for match will be left at registry as payment for assistance
• Disk made will only have cancer data on matches and AHS id #s; reg. staff can check.
Confidentiality of cancer data taken to LLU

- No personal identifiers on any electronic file that also has lifestyle or cancer data
- Study **ID # can be linked** to name and address **only by 2** senior data managers.
- When **junior staff** work with data only out of range or unlikely responses show on PC.
- Thousands of subjects contacted about this; none have expressed concern
What have we learned?

1) IRB approval takes longer than expected
2) Traveling programmer model works
3) Missing data is key issue
4) Level of missing registry data varies widely
5) Cancer self-reports need registry details
6) PPV
7) Registry and IRB staff most helpful

THANK YOU ALL!
Summary

• A diet-cancer study in a special population whose results can be generalized.

• Participants across the U.S. and Canada.

• Opportunity to use NAACCR member registry data in a uniform manner to greatly enhance validity and efficiency of this important research.

• A new era for cohort studies of NAACCR member collaboration continent-wide.
# Acknowledgments

<table>
<thead>
<tr>
<th>AZ</th>
<th>Georgia Yee</th>
<th>MO</th>
<th>Jeanette Jackson-Thompson</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chris Newton</td>
<td></td>
<td>Mary Jane King</td>
</tr>
<tr>
<td></td>
<td>Keith Laubham</td>
<td></td>
<td>Sue Vest</td>
</tr>
<tr>
<td>CA</td>
<td>Bill Wright</td>
<td>NE</td>
<td>Kurt Weiss</td>
</tr>
<tr>
<td></td>
<td>Mark Allen</td>
<td></td>
<td>Saba Yemane</td>
</tr>
<tr>
<td></td>
<td>David Harris</td>
<td></td>
<td>Judy Paradies</td>
</tr>
<tr>
<td>MI</td>
<td>Glenn Copeland</td>
<td>NY</td>
<td>Maria Schymura</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Amy Kahn</td>
</tr>
<tr>
<td>LLU</td>
<td>Larry Beeson</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lars Sveen</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>