Composite Indices for Health Outcomes in Epidemiologic Studies: Multi-dimensional Coding Using SAS

Nadia Chung, MPH, Huiyan Ma, PhD, Jenna Voutsinas, MPH, Leslie Bernstein, PhD, Sophia Wang, PhD, James V. Lacey, Jr., PhD,
Beckman Research Institute, City of Hope, Duarte, CA

ABSTRACT

Analyses in health studies often focus on one outcome at a time. In certain situations, such as prospective epidemiologic studies, the outcome of interest will be a summary measure of multiple potential outcomes, such as all-cause-mortality (i.e., death from any cause) or first diagnosis of any cancer. These types of composite indices are particularly useful when studying an intervention or exposure that may reduce the incidence of some outcomes but increase the incidence of other outcomes. Creating a composite index within a single data source is fairly straightforward, but creating a composite index across multiple data sources is analytically more complex. For an analysis of the health effects of menopausal hormone therapy in the prospective California Teachers Study (CTS) cohort, we created a composite index that was defined as the first diagnosis of breast, colorectal or endometrial cancer, coronary heart disease, stroke, pulmonary embolism, hip fracture or death from any cause. To identify the individual events and calculate the composite index, we combined data from four independent databases: in-patient hospitalizations from the California Office of Statewide Health Planning and Development (OSHPD), the California Cancer Registry, national and state mortality files, and self-reported questionnaires from CTS participants. This paper will provide tips and techniques to efficiently analyze multi-dimensional data in order to create a composite index of cancer and non-cancer endpoints.

INTRODUCTION

A common approach in most research on health outcomes is for each individual analysis to focus on a single event of interest, such as heart disease, breast cancer, or hip fracture. Doing so typically means ignoring other endpoints, even though many diseases and health conditions occur together, can affect each other, and, from a practical standpoint, are often present in the same source data. Generating a summary measure based on multiple endpoints can be particularly useful when studying an exposure that is known to simultaneously but differentially affect multiple endpoints, such as when an intervention is beneficial for some outcomes but harmful for others. In addition, these types of summary measures can also be an efficient way to boost statistical power in smaller studies, where individual endpoints might be rare but the combination of endpoints is sufficient to achieve statistically valid results more quickly.

A recent example of innovative use of a composite index was the Women’s Health Initiative (WHI) estrogen plus progestin clinical trial (Rossouw et al., 2002). The objective of that double-blinded, placebo-controlled, randomized clinical trial was to determine whether use of conjugated equine estrogens (CEE; 0.625 mg/day) plus medroxyprogesterone acetate (MPA; 2.5 mg/day) could prevent chronic disease among healthy postmenopausal women. Because previous studies had shown that CEE plus MPA likely increased the risk of breast cancer but decreased the risk of hip fracture and potentially coronary heart disease (CHD), the WHI trial created a “Global Index,” which was defined as the first occurrence of breast cancer, CHD, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, or death from any other cause. The primary clinical endpoints in the WHI trial included each of those individual outcomes and the Global Index. The WHI study investigators then utilized active surveillance methods with regular follow-up to identify each occurrence of these endpoints among all study participants. At the end of the trial, that Global Index provided key data to support the overall conclusion, which was that, contrary to what many observers had expected to see, use of CEE plus MPA was associated with more harms than benefits (Rossouw et al., 2002).

The WHI clinical trial demonstrated that, when an investigator is planning a research study that includes a composite index, data collection procedures can be designed and individually tailored to efficiently generate data on the multiple individual component endpoints that contribute to the composite index. However, retrospectively assembling information on multiple endpoints using existing data sets, which may not have originally been constructed for that purpose, can be more challenging. We recently conducted a statistical analysis within the California Teachers Study (CTS), in which we wished to create a replica of the WHI’s Global Index. Like the WHI, our analysis in the CTS would include both the individual endpoints and the composite index, but unlike the WHI, the CTS data collection was not originally designed to actively capture each of those endpoints. In this paper, we will describe techniques for creating that single summary composite measure based on multiple endpoints of interest that were obtained from the multiple separate endpoint data sets within the CTS.
SOURCE SETS

The CTS is a prospective observational epidemiologic cohort study of 133,479 active or retired female public school teachers and administrators who have been followed from 1995-1996 through the present. The CTS participants were between the ages of 22 and 104 years at baseline, and completed and returned a mailed, self-administered questionnaire. Since the baseline questionnaire, 3 additional follow-up questionnaires have been mailed (in 1997, 2000-2001, and 2005-2006), and a 4th follow-up questionnaire will be mailed in August, 2012.

In the CTS, outcome data are obtained via regular linkages with cancer, hospitalization, and mortality databases. In-patient hospitalization data are obtained from the California Office of Statewide Health Planning and Development (OSHPD). Cancer data are obtained from the California Cancer Registry (CCR). National and state mortality files provide information on deaths (Figure 1). For our analysis, we used those multiple data sources to create a composite index that was defined as the first diagnosis of breast, colorectal or endometrial cancer; CHD; stroke; pulmonary embolism; hip fracture; or death from any other cause.

Figure 1. Diagram of source data sets used for the California Teachers Study endpoint analysis.

Hospitalization Data

The OSHPD captures information on in-patient hospital admissions including principal hospital diagnosis and up to 24 other diagnosis reported by the physician in addition to a principal procedure and up to 20 other procedures for each hospital admission in California since 1991. Admissions dates are reported and each of the procedure variables has corresponding procedure dates for each patient. From 1991 (the earliest date of OSHPD coverage) through 2009, over 218,000 in-patient hospitalization records are available for CTS participants. For our analysis, endpoints were based on International Classification of Diseases (ICD) version 9 clinical modification (CM) diagnosis and procedure codes listed in the principal position for each recorded hospitalization.

Cancer Data

The California Cancer Registry (CCR) captures essentially all incident cancer cases diagnosed in California from 1988 onward reported using National Cancer Institute (NCI) Surveillance, Epidemiology and End Results (SEER) or International Classification of Diseases for Oncology, Third Edition (ICD-O-3) codes. Through the end of 2009, 14,882 incident cancers have been identified within the CTS cohort. In our analysis we only considered invasive cancers cancer endpoints of interest (excluding in situ and borderline cases).

Mortality Data

Data for all-cause and cause-specific mortality between 1995 and 2009 are available through linkage of the CTS to available mortality databases. Deaths are identified by ICD codes (version 9 through 1998 and version 10 codes thereafter). On an annual basis, the CTS is linked to the state Department of Public Health mortality files to ascertain dates and causes of death for all deceased cohort participants resident to California. Annual linkage with the national Social Security Administration (SSA) Death Master File to ascertain date of death for CTS participants living elsewhere in the U.S. is also performed. Periodically, a deterministic linkage is performed with the National Death Index (NDI) to ascertain dates and causes of death information, as well as to augment cause of death information for those identified through the SSA Death Master File. Between 1995 and 2009, there were 18,901 deaths among CTS
participants.

DEFINING THE COMPOSITE INDEX ENDPOINT

Hospital, cancer and mortality data files were linked to the CTS baseline questionnaire to create a data set restricted to eligible participants. For this analysis, eligible participants were defined as participants who were California residents at cohort entry, were at least 40 years old, and who did not have any previous history of the endpoints of interest (i.e., breast, endometrial or colorectal cancer; stroke; CHD; pulmonary embolism; or hip fracture) prior to their joining the cohort. Participants with a previous history of these endpoints were identified based on self-reports from the baseline questionnaire and information from the linked OSHPD data and CCR data. Participants were followed from the date of baseline questionnaire completion to the earliest of the following events:

- The date of first endpoint diagnosis
- The date of death
- The date they moved out of California
- December 31, 2009 (end of study follow-up date)

Participants who were under age 40 years at CTS baseline became eligible for this analysis and started contributing person-time when they reached age 40 years (i.e., on the date of their 40th birthday).

IDENTIFYING INDIVIDUAL ENDPOINTS WITHIN THE SOURCE DATA SETS

In order to generate a single data set for each outcome of interest, we first extracted variables for each eligible participant from the source data sets. We then restructured the data set from multiple observations per participant to a single observation with multiple columns for each participant. The end result is a data set of only the variables of interest and one record per study participant that captures all relevant codes for study analysis purposes. Table 1 shows two examples, one for breast cancer and one for CHD.

Table 1. Sample output after resolving multiple records and only keeping variables and codes of interest for the study participant. This example includes a participant with a cancer diagnosis (reported by SEER or ICD-O-3 codes), and a participant with the non-cancer endpoint CHD (reported by ICD codes).

<table>
<thead>
<tr>
<th>IDNUMBER</th>
<th>EVENT</th>
<th>ENDDATE</th>
<th>SEER</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXXX08</td>
<td>2</td>
<td>09/14/1998</td>
<td>41001</td>
</tr>
</tbody>
</table>

Before merging the multiple single data sets to form our final summary data set we created four variables for each endpoint included within the global index: \( x_{\text{event}} \), which denotes the category of the outcome that caused follow-up to be stopped; \( x_{\text{enddate}} \), which denotes the date that corresponds to their ‘event’ if the event was the endpoint of interest; \( x_{\text{dtmvd}} \), which denotes the date they moved out of California if their follow-up ended for that reason; and \( x_{\text{dth_dt}} \), which corresponds to the date they died if the reason for ending follow-up was due to death. The ‘x’ will be an abbreviation that corresponds to the endpoint data set; for example, \( b_{\text{event}} \), \( b_{\text{dtmvd}} \), \( b_{\text{dth_dt}} \) and \( b_{\text{enddate}} \) are variables within the ‘breast’ data set. Additional variables were created in the multi-level endpoint data sets below that follow essentially the same structure and format.

- \( x_{\text{event}} \)
  - Values: 1= moved out of California
  - 2= had the event of interest
  - 3= died from any other cause

We generated a separate data set for each outcome of interest. For example, we used the following SAS code to output a data set called ‘breast’ for breast cancer.

```sas
data breast;
b_event=0;
if event eq 2 then do;
  b_event=1; b_dtmvd=enddate; moved out of CA
```
end;
else if event eq 3 then do;
b_event=3; b_dth_dt=enddate; died from another cause
end;
else if event eq 1 then do;
b_event=2; b_enddate=enddate; breast cancer
end;
keep idnumber b_event b_dtmvd b_dth_dt b_enddate;
run;

Similar codes were created for the other cancer or other endpoints (i.e., cr_event, cr_dtmvd, cr_dth_dt, cr_enddate for colorectal cancer event, date moved outside of California, date of death and event date, respectively).

Cancer is a binary outcome, but the same approach can be used to operationally define a multi-level endpoint, such as CHD, which can be further divided into confirmed CHD based on ICD-9 codes, unconfirmed CHD based on ICD-9 codes and duration of hospital stay, or potential CHD based on ICD-9 procedure codes that strongly suggest that a patient has CHD (e.g., catheterization or angioplasty) but that are not considered as valid as CHD based on the specific ICD-9 discharge diagnosis codes.

Below is the code to create summary variables for non-cancer endpoints, using CHD as an example. There is an additional DATA step for non-cancer endpoints incorporating mortality, move and procedure data. We matched idnumber with death dates for all-cause mortality and cause-specific mortality using ICD-9 and ICD-10 codes. Procedure codes were used to identify ‘potential chd’ events by looking at procedure codes common for CHD (i.e., catheterization or angioplasty). Our ‘initial’ data set was utilized to extract move dates.

```sas
data chd_a;
enddate=min(dth_dt, chddate, proc_date, dtmvd, mdy(12,31,2009));
if enddate eq chddate and chd eq 1 then n_event=1; confirmed CHD
else if enddate eq proc_date and proc ne . then n_event=3; potential CHD
else if enddate eq chddate and chd eq 2 then n_event=2; unconfirmed CHD
else if enddate eq dtmvd then n_event=4; moved out of CA
else if enddate eq dth_dt and event eq 1 then n_event=5; died from CHD
else if enddate eq dth_dt and event eq 3 then n_event=6; died from another cause
else if enddate eq mdy(12,31,2009) then n_event=7; participants remained in the study until end of follow-up (12,31,2009)
run;
```

Table 2 shows an annual frequency distribution using the date variables and the variable n_event for each data set by utilizing a PROC FREQ macro (not presented in this paper).
Table 2. Sample spreadsheet created to look at annual frequencies of events and eligibility distribution for each individual data set.

<table>
<thead>
<tr>
<th>CORONARY HEART DISEASE BY YEAR</th>
<th>1997</th>
<th>1998</th>
<th>...</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants in CTS CHD data set w no events prior to 1/1/xxxx</td>
<td>121,243</td>
<td>119,492</td>
<td></td>
<td>98,735</td>
</tr>
<tr>
<td>Participants age ≥40 years during that year</td>
<td>103,946</td>
<td>104,037</td>
<td></td>
<td>96,979</td>
</tr>
<tr>
<td>Participants who ended follow up due to incident confirmed CHD in that year</td>
<td>73</td>
<td>98</td>
<td></td>
<td>87</td>
</tr>
<tr>
<td>Participants who ended follow up due to incident unconfirmed CHD in that year</td>
<td>5</td>
<td>0</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Participants who ended follow up due to potential CHD in that year</td>
<td>202</td>
<td>263</td>
<td></td>
<td>178</td>
</tr>
<tr>
<td>Participants who left California that year</td>
<td>657</td>
<td>727</td>
<td></td>
<td>782</td>
</tr>
<tr>
<td>Participants who ended follow up due to fatal CHD in that year</td>
<td>34</td>
<td>52</td>
<td></td>
<td>68</td>
</tr>
<tr>
<td>Participants who died from another cause in that year</td>
<td>566</td>
<td>618</td>
<td></td>
<td>1058</td>
</tr>
<tr>
<td>Total # of participants who end follow-up that year</td>
<td>1537</td>
<td>1758</td>
<td></td>
<td>2181</td>
</tr>
<tr>
<td>Total # of participants who remained event-free at end of year</td>
<td>102,409</td>
<td>102,279</td>
<td></td>
<td>94,798</td>
</tr>
</tbody>
</table>

*This total corresponds to the sum of ‡ through †*
*This total corresponds to ‡ minus the total from †*

SUMMARY DATA SET

After creating our individual endpoint data sets, we merged each of the data sets into a summary file that included the event variables for each of the global index endpoints. We then created two summary variables for global index. Variable gindex (global index event values 1-3: 1=moved outside of California; 2=global index event; 3=remained in the study until 12, 31, 2009) and gl_event with values from 1-8 based on the designated endpoint of interest (1=breast cancer; 2=stroke; 3=CHD; 4=pulmonary embolism; 5=hip fracture; 6=colorectal cancer; 7=endometrial cancer; 8=death). Coding the endpoints in this way allowed us to easily restrict analysis to a specific endpoint based on the value assigned to the endpoint of interest and to easily identify the specific contribution, in either absolute or relative terms, of each endpoint to the overall global index. If a participant happened to develop more than one global index event on the same day, her gl_event status was assigned based on the most severe event. Severity was based on qualitative judgment of the endpoints.

To create our global index variable we used the x_event, x_enddate, x_dtmvd, and x_dth_dt variables in existing individual endpoint data sets created with code from the prior section. We merged each endpoint data set ('question' consisted of variables from the baseline questionnaire, 'breast' included breast cancer case variables, 'endo' included endometrial cancer case variables, 'colr' included colorectal cancer case variables, 'chd' included CHD case variables, 'stroke' included stroke case variables, 'pe' included pulmonary embolism case variables and 'hipf' included hip fracture case variables) to the smallest data set, which in this case was coronary heart disease ('chd' data set). We needed to exclude participants with any personal history of any of the individual endpoints of interest, and therefore restricting the eligible population to the data set that had the most exclusions was an efficient way to accomplish that. All the individual data sets were sorted by idnumber prior to merging.

```
data glbinx;
merge chd (in=a) question breast endo colr stroke pe hipf;
by idnumber;
if a;
run;
```

From each endpoint data set we want to take the earliest date of an event and define that as the global index event that ends follow-up. For example, if a participant had a record of breast cancer on August 10, 1998 and a record of a stroke on December 17, 1999 then using the MIN function for the dates, we would take the date of the breast cancer diagnosis to be the date on which follow-up ends; breast cancer would then be the event that ends follow-up.

```
data global1;
```
set glbinx;
gindex=0;
dtmvd=min(b_dtmvd, e_dtmvd, cr_dtmvd, chd_dtmvd, s_dtmvd, pe_dtmvd, hf_dtmvd);
dth_dt=min(b_dth_dt, e_dth_dt, cr_dth_dt, chd_dth_dt, s_dth_dt, pe_dth_dt, hf_dth_dt);
enddate=min(b_enddate, e_enddate, cr_enddate, chd_enddate, s_enddate, pe_enddate, hf_enddate);
gindexdate=min(dtmvd, dth_dt, enddate, mdy(12,31,2009));
if gindexdate eq mdy(12,31,2009) then gindex=3; participants remained in the study until end of follow-up (12, 31, 2009)
else if gindexdate eq dtmvd then gindex=1; moved out of CA
else if gindexdate eq enddate or dth_dt then gindex=2; global index event
run;
The following DATA step creates gl_event that contains different values (1-8) that denotes which global index event the participant was censored for (i.e., breast cancer, stroke, etc.) based on the gindexdate variable created in the prior DATA step.

data g1;
set global1;
gl_event=0;
if gindex eq 2 then do;
if ((b_event eq 3 or e_event eq 3 or cr_event eq 3 or chd_event eq 3 or s_event eq 3 or pe_event eq 3 or hf_event eq 3) and gindexdate eq dth_dt)then gl_event=8;
end;
run; death

data g2;
set g1;
if b_index eq 2 then do;
if b_event eq 2 and b_enddate eq gindexdate then gl_event=1; breast cancer
else if s_event eq 2 and s_enddate eq gindexdate then gl_event=2; stroke
else if chd_event eq 2 and chd_enddate eq gindexdate then gl_event=3; chd
else if pe_event eq 2 and pe_enddate eq gindexdate then gl_event=4; pulmonary embolism
else if hf_event eq 2 and hf_enddate eq gindexdate then gl_event=5; hip fracture
else if cr_event eq 2 and cr_enddate eq gindexdate then gl_event=6; colorectal cancer
else if e_event eq 2 and e_enddate eq gindexdate then gl_event=7; endometrial cancer
end;
run;
Using the variable gl_event we are able to generate a table of annual frequencies of events that define the global index (Table 3).

Table 3. Sample spreadsheet created to look at annual frequencies of global index events.

<table>
<thead>
<tr>
<th>GLOBAL INDEX EVENTS BY YEAR</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
<th>...</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death</td>
<td>330</td>
<td>325</td>
<td>367</td>
<td>...</td>
<td>548</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>353</td>
<td>350</td>
<td>360</td>
<td>...</td>
<td>308</td>
</tr>
<tr>
<td>Coronary Heart Disease</td>
<td>228</td>
<td>297</td>
<td>284</td>
<td>...</td>
<td>187</td>
</tr>
<tr>
<td>Stroke</td>
<td>164</td>
<td>181</td>
<td>200</td>
<td>...</td>
<td>222</td>
</tr>
<tr>
<td>Hip Fracture</td>
<td>117</td>
<td>118</td>
<td>127</td>
<td>...</td>
<td>168</td>
</tr>
<tr>
<td>Endometrial Cancer</td>
<td>80</td>
<td>71</td>
<td>70</td>
<td>...</td>
<td>62</td>
</tr>
<tr>
<td>Colorectal Cancer</td>
<td>64</td>
<td>65</td>
<td>67</td>
<td>...</td>
<td>66</td>
</tr>
</tbody>
</table>
CONCLUSION

In this analysis, the outcome of interest was a summary measure of multiple outcomes. Our global index was defined as the first diagnosis of breast, colorectal or endometrial cancer, CHD, stroke, pulmonary embolism, hip fracture or death from any cause.

We extracted detailed hospitalization, morbidity and mortality data from various data sets and linked them to our CTS data set. From the individual endpoints we created separate individual output data sets with identical structures and then merged those data sets into an overall analytic data set. We set coding rules to prioritize events and event dates that fell on the same day. For multiple events occurring at different times, we prioritized the global index event that ended follow-up to be the event that occurred first. Alternatively, the approach we have described here could be easily modified to create a composite index that assigns weights to each of the individual endpoints based on their clinical importance that sums the total number of endpoints that occur over time, or that combines both severity and number of endpoints to produce a different type of composite index.

Our analysis utilized a composite index as a study outcome. Similar composite indices are also utilized as exposure variables in studies of health and disease. For example, the presence of comorbidities—that is, other chronic diseases or conditions that are present in a patient—can significantly and independently predict short and long-term major health outcomes, such as mortality. The Charlson Comorbidity index is a commonly used metric that counts the number of comorbidities in a patient to generate a “score,” with higher scores indicating more comorbidities or more severe comorbidities. Within longitudinal studies, the analytic approach that we described above can be utilized to continuously scan multi-dimensional data to identify comorbidities and create indices, such as the Charlson Comorbidity Index. Whereas our analysis used the date of a qualifying individual endpoint to end follow-up, an analysis that uses a composite index as an exposure could instead code, classify, and analyze that type of date as a time-dependent exposure, e.g., to identify the point in time at which a person becomes exposed or has a change in exposure status. Those types of composite exposures can then be used for survival analyses or analyses of the interval between a change in exposure and cause-specific mortality.

Global measures that summarize health status based on a number of factors can provide more information than single outcomes do, especially in studies of factors that have different effects on different outcomes, or when studying the effects of competing risks on an endpoint of interest. Although data analysis in these situations can be complex, our experience demonstrates that fairly straightforward analytic procedures can be used to efficiently identify, create, and analyze composite indexes out of multi-dimensional data.

REFERENCE


CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Nadia Chung, MPH
Beckman Research Institute, City of Hope National Medical Center
1500 East Duarte Road
Duarte, CA 91010-3000
626-256-4673 Ext. 62615
nchung@coh.org

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.