General Multiple Mediation Analysis to Examine Ethnic Differences in Anxiety and Depression in Cancer Survivors Using the MY-Health Survey

Kaelen L. Medeiros

Louisiana State University Health Sciences Center - New Orleans
Biostatistics MS, 2016
Thesis w/Dr. Qingzhao Yu

NU Biostats Presentation
May 15, 2018
1 Motivation
2 Mediation Methods
   - Simple and multiple mediation analysis
   - Total effect, direct effect, indirect effect
   - MMA for GLM and MART
   - Delta Method and Bootstrapping
   - Statistical Tests on Mediation Effects
   - Statistical Software
3 PROMIS and MY-Health overview
4 Our Study using MY-Health
   - Goals, Exposure, Outcomes, and Candidate Mediators
5 Results
   - Anxiety (in detail)
6 Conclusions
7 Limitations and Future Research
Motivation

- There is a well-documented difference in the experiences of cancer survivors by ethnicity. Hispanic survivors tend to be younger, less educated, and report higher levels of “psychosocial distress” (Apollo 2007, Culver 2002).
- For my thesis, we sought to find out if there was an ethnic difference in anxiety and depression as reported in the MY-Health survey, and to see what factors affect the disparity if it exists.
- Multiple Mediation Analysis (MMA) can answer this question.
Mediation Methods

Intro

- Mediation analysis is a type of analysis which allows examination of the mediating factors (mediators) in the causal pathway between an exposure and an outcome.

- A *mediator* is a "risk factor" in the casual pathway that can partially or totally transmit the effect of the exposure to the outcome variable.
To begin to understand simple mediation analysis, let’s examine these three regression equations, given that $X$ is some exposure, $M$ is a mediator, and $Y$ is an outcome:

1. $M = \alpha X + e_1$
2. $Y = \beta M + c_1 X + e_2$
3. $Y = c_2 X + e_3$
Mediation Methods
“The Four Steps”

Then, there are three widely accepted fundamental assumptions that must be met in order to conduct a mediation analysis (Judd & Kenny 1981, James & Brett 1984, Baron & Kenny 1986):

1. The exposure variable (X) should be associated with the response variable (Y);
2. Any proposed mediator (M) should be associated with the exposure X;
3. M is associated with Y, controlling for X.

Further research into mediation analysis established the “fourth step”:

4. To establish complete mediation, the effect of X on Y controlling for M must be zero.
Mediation Methods
Partial and Complete Mediation

- A *complete mediation* refers to the situation in which the mediator(s) transmit all of the effect of X on Y (where the “fourth step” is satisfied.)

- A *partial mediation* is the case in which only the first three steps are met, and there will still be some non-zero effect of X on Y directly.
There are then two main methods of finding the *indirect effect (IE)*, or the effect of X on Y transmitted by the mediators:

- Coefficients difference (CD) method: $c_2 - c_1 = IE$
- Coefficients product (CP) method: $\alpha \cdot \beta = IE$

These two methods have been proven to be equivalent (MacKinnon and Dwyer 1985).
However, these two methods have their limitations.

- CD method assumes the 2nd and 3rd models (from slide 5) to be simultaneously true, doesn’t allow for separation of mediation effects for multiple mediators, plus the scale of coefficients will be different with logistic regression.
- CP method is difficult to interpret for anything besides a linear regression.

Thus, the desire to expand on mediation analysis.
Here, $X$ is some exposure.

$Y$ is an outcome.

$M_1,\ldots,M_p$ are the candidate mediators in the path between the exposure and the outcome.

$Z$ are additional covariates, which are significantly associated with the outcome only.
Necessary Notation

- Let $\mathbf{M} = (M_1, ..., M_p)^T$, where $M_j$ is the $j^{th}$ mediator.
- $\mathbf{Z}$ is the vector of other independent covariates that relate to $Y$, but do not interact with or relate to $X$. 
From Yu et al. 2014:

Given $Z$, the **total effect** ($TE$) of $X$ on $Y$ at $X = x^*$ is defined as the changing rate of $E(Y|(X,Z))$ when $X$ changes by a $u^*$ unit at $x^*$:

$$TE_{|Z}(x^*) = \lim_{u \to u^*} \frac{E[Y(x^*+u)−y(x^*)]|Z]}{u}.$$ 

The total effect is defined in terms of average rate of change of $Y$ with the exposure $X$.

Generally, the total effect is thought of as the effect transmitted both by $X$ directly on $Y$, and also on $Y$ by $X$ through $M$.

The **average total effect** is defined as $ATE_{|Z} = E_{x^*} [TE_{|Z}(x^*)]$.  

Direct Effect

- From Yu et al., 2014:
  
  For given $Z$, the **direct effect** (DE) of $X$ on $Y$ not from $M_j$ is defined as

  $$DE_{M_j|Z}(x^*) = E_{m_j} \lim_{u \to u^*} \frac{E(Y(x^*+u,M_j=m_j,M_{-j}(x^*+u))|Z) - E(Y(x^*,M_j=m_j,M_{-j}(x^*))|Z)}{u}$$

  - The **average direct effect** of $X$ on $Y$ not from $M_j$ is

  $$ADE_{M_j|Z} = E_{x^*} DE_{M_j|Z}(x^*)$$

  where $M_{-j}$ denotes the vector $M$ without $M_j$.

- Direct effect is the effect of $X$ on $Y$, holding all the candidate mediators constant.
From Yu et al., 2014:
*Given Z, the indirect effect of X on Y through M_j is defined as*
\[ IE_{M_j|Z}(x^*) = TE_{Z}(x^*) - DE_{M_j|Z}(x^*). \]

Indirect effect is the effect of X on Y transmitted through the candidate mediators.

It is defined through the definitions of total effect and direct effect.

The average indirect effect through M_j can be similarly defined:
\[ AIE_{M_j|Z} = ATE_{Z} - ADE_{M_j|Z}. \]
Without loss of generality, assume two mediators and that the true relationships between $X$, $Y$, $M_1$, and $M_2$ are:

1. $M_{1i} = a_{01} + a_1 X_i + e_{1i}$
2. $M_{2i} = a_{02} + a_2 X_i + e_{2i}$
3. $Y_i = b_0 + b_1 M_{1i} + b_2 M_{2i} + c X_i + e_{3i}$

where $(e_{1i}, e_{2i})$ are i.i.d. bivariate normal distributed, and are independent with $e_{3i}$, which are i.i.d. $N(0, \sigma_3^2)$, for $i = 1, \ldots, n$. 
Then, it is helpful to have the following Lemma 3.4, from Yu et al. 2014:

- The total effect of $X$ on $Y$ is $a_1 b_1 + a_2 b_2 + c$
- The average indirect effect through $M_1$ is $a_1 b_1$ and through $M_2$ is $a_2 b_2$.
- Of all the effect from $X$ to $Y$, a $\frac{a_1 b_1}{a_1 b_1 + a_2 b_2 + c}$ fraction is directly from $M_1$, $\frac{a_2 b_2}{a_1 b_1 + a_2 b_2 + c}$ is directly from $M_2$, and $\frac{c}{a_1 b_1 + a_2 b_2 + c}$ is directly from $X$. 

![Diagram of causal relationships between X, M1, M2, and Y]
Multiple additive regression trees, or MART, is another name for the general boosted additive modeling approach developed and outlined in Friedman 2001.

For mediation analysis, it will allow us to model non-linear relationships between $X$, $M$, and $Y$. 
Yu 2016 outlines two algorithms for MMA: one linear using GLM methods, and one non-linear using MART methods.

They describe calculation of mediation effects with relaxed model assumptions, and the algorithms are derived directly from the definitions of mediation effects.
The delta method can be utilized to estimate the variances of the mediation effect estimators in the GLM setting.

Bootstrapping can also be used for the GLM and MART situations. A general algorithm for this:

1. Randomly draw a sample of n observations from the data with replacement;
2. Estimate $DE$, $IE$, and $TE$;
3. repeat steps 1, 2 B times. Obtain B sets of estimates on each of the resampled sets of observations;
4. Obtain the empirical variances of mediation effects plus the $\frac{\alpha}{2}$th and $(1 - \frac{\alpha}{2})$th percentiles, based on the B sets of estimates.
We can perform hypothesis testing on the indirect effects separately:
- $H_0$: $IE_j = 0$ vs. $H_a$: $IE_j \neq 0$
- then we check the confidence interval from a given significance level to evaluate if 0 is in the CI.

We can also perform hypothesis testing on the direct effect:
- $H_0$: $DE = 0$ vs. $H_a$: $DE \neq 0$
- and similarly, we check the CI from a given significance level to evaluate if 0 is in the CI.
Multiple Mediation Analysis (MMA) as developed in Yu et al. 2014

- MMA is beneficial because it generalizes previous mediation analysis methods and allows for flexibility: exposure, mediators, and outcome can be binary, multicategorical, or continuous.
- MMA also allows for flexibility in analysis, allowing for GLM and MART, the latter of which allows non-linear assumptions about the relationships between exposure, mediators, and outcome to be made.
- Finally, MMA allows us to separate out the effect of each mediator, allowing for comparison of effects.
The mma package, available on the Comprehensive R Archive Network (CRAN), can be downloaded and utilized to perform all MMA calculations.

- The package contains functions that will do the entire mediation analysis in one step, using the mma() function, or it can be completed step by step using other functions in tandem.
- mma also includes many built-in plots.
PROMIS instrument
as developed by the National Institutes of Health (NIH)

- PROMIS is a registered trademark of the NIH and stands for "Patient Reported Outcomes Measurement Information System."
- It was created to allow for compilation and validation of a set of questions (item bank) about patient reported health outcomes that can be used on diverse populations in similarly diverse healthcare settings.
- It measures information the domains of physical, mental, and social health for both adults and children to build information about Health Related Quality of Life (HRQoL) of patients.
The outcome scores on the domains measured in PROMIS, which include:

- physical function
- pain intensity
- pain interference
- fatigue
- sleep disturbance
- depression
- anxiety
- ability to participate in social roles and activities

The scores are reported as t-scores with a mean of 50 and a standard deviation of 10, and are calibrated in such a way that the mean score of 50 reflects a reference to the U.S. general population.

Higher scores indicate more of some domain.
The MY-Health Study was conducted at Georgetown University, in conjunction with investigators at the Fred Hutchinson Cancer Research Center and the Louisiana Tumor Registry.

The study’s aim is to evaluate validity of PROMIS items in a racially, ethnically, and age-diverse cancer survivor population.

Stratified sampling was employed to ensure a diverse population, ensuring that representative samples were taken from four racial/ethnic subgroups (white, black, Asian-Pacific Islander, and Hispanic) and three age-based subgroups (21-49, 50-64, 65-84).
MY-Health overview

- MY-Health was conducted with a mailed survey, sent to the patients recruited from four tumor registries in three states who were at least 6 months from their initial diagnosis.
- Participants were scored on all possible PROMIS domains.
Mainly, this included variable creation, including:
- social support
- self-reported comorbidities
- surgery of the primary site group

We also examined missing data.
- Most of the variables had no missing values.
- Those that had an extreme amount missing (>3000) were usually variables that did not apply to people born in the U.S., and therefore these questions were not asked of them.
Our Study using MY-Health data

- We hypothesized that Hispanic versus non-Hispanic survivors would have significantly different anxiety and depression scores (separate analyses) as measured by MY-Health.

- Our main "exposure" in the study is ethnicity, coded as either Hispanic or Not Hispanic (including all white, black, and Asian/Pacific-Islander individuals).

- Our outcome was the continuous t-score for anxiety or depression.
## Our Study
using MY-Health data

<table>
<thead>
<tr>
<th>Candidate Mediators</th>
<th>Binary</th>
<th>Categorical</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment</td>
<td></td>
<td>Insurance</td>
<td>Income</td>
</tr>
<tr>
<td>Married</td>
<td></td>
<td>Spirituality</td>
<td>Education</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td>AJCC Stage</td>
<td>Age at Diagnosis</td>
</tr>
<tr>
<td>Kids &lt;18 in home</td>
<td></td>
<td>Radiation treatment</td>
<td>Age to U.S.</td>
</tr>
<tr>
<td>Born in U.S.</td>
<td></td>
<td>Surgery of the Primary Site</td>
<td>Social Support</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Grade</td>
<td>Comorbidities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cancer site</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diagnosis to Survey Completion group</td>
<td></td>
</tr>
</tbody>
</table>
The MY-Health dataset includes 5,506 observations.

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>n=5506</th>
<th>n(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic</td>
<td>1064</td>
<td>(19.32)</td>
</tr>
<tr>
<td>Not Hispanic</td>
<td>4442</td>
<td>(80.68)</td>
</tr>
</tbody>
</table>
The mean scores for the two populations are as follows:

<table>
<thead>
<tr>
<th>Score</th>
<th>Hispanic mean</th>
<th>Non-Hispanic mean</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety</td>
<td>52.71</td>
<td>49.26</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>Depression</td>
<td>51.39</td>
<td>48.32</td>
<td>&lt;2e-16</td>
</tr>
</tbody>
</table>
For reasons of time, I will only outline in detail the results of the Anxiety Score analysis.

The outcomes for Depression Score are very similar, with the addition of cancer site as a mediator.
## Results

### Anxiety Score

<table>
<thead>
<tr>
<th>Mediators</th>
<th>Covariates</th>
<th>Dropped from model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>Employment</td>
<td>Married</td>
</tr>
<tr>
<td>Age at Diagnosis</td>
<td>Sex</td>
<td>Diagnosis to Survey Completion group</td>
</tr>
<tr>
<td>Age came to U.S.</td>
<td>DAJCC Stage</td>
<td>Chemotherapy treatment</td>
</tr>
<tr>
<td>Insurance</td>
<td></td>
<td>Radiation treatment</td>
</tr>
<tr>
<td>Social Support</td>
<td></td>
<td>Grade</td>
</tr>
<tr>
<td>Spirituality</td>
<td></td>
<td>Kids (&lt;18) in the Household</td>
</tr>
<tr>
<td>Comorbidities</td>
<td></td>
<td>Born in the U.S.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Surgery of the Primary Site</td>
</tr>
</tbody>
</table>
We found that there is still an ethnic disparity in anxiety score even after examining the effects of all of the potential mediators.

Specifically, the estimated average direct effect of ethnicity on anxiety score was:

- 1.224 (-2.047, -0.450) using GLM;
- 1.003 (-1.679, -0.369) using MART.
The indirect effects are given on the next slides, plus we can calculate relative effects for ease of discussion.

The *relative effect* (RE) is defined as \( RE = \frac{IE}{IE} \times 100 \)
<table>
<thead>
<tr>
<th>Mediators</th>
<th>IE(95% CI)</th>
<th>RE(95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education*</td>
<td>-0.643(-0.875,-0.428)</td>
<td>20.35(12.9,29.6)</td>
</tr>
<tr>
<td>Age at Diagnosis*</td>
<td>-0.689(-0.454,-0.924)</td>
<td>21.76(13.4,32.4)</td>
</tr>
<tr>
<td>Age came to U.S*</td>
<td>-0.266(-0.498,-0.043)</td>
<td>8.38(1.4,16.2)</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>-0.176(-0.433,0.105)</td>
<td>5.40(-3.4,13.1)</td>
</tr>
<tr>
<td>Insurance</td>
<td>-0.159(-0.329,0.017)</td>
<td>5.02(-0.5,11.1)</td>
</tr>
<tr>
<td>Social Support*</td>
<td>-0.256(-0.504,-0.001)</td>
<td>7.96(0.02,15.5)</td>
</tr>
<tr>
<td>Spirituality*</td>
<td>0.248(0.088,0.413)</td>
<td>-7.93(-14.3,-2.5)</td>
</tr>
<tr>
<td>Mediators</td>
<td>IE(95% CI)</td>
<td>RE(95% CI)</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----------------------------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>Education*</td>
<td>-0.607(-0.839,-0.385)</td>
<td>21.73(13.1,34.2)</td>
</tr>
<tr>
<td>Age at Diagnosis*</td>
<td>-0.595(-0.812,-0.377)</td>
<td>21.24(13.0,31.6)</td>
</tr>
<tr>
<td>Age came to U.S.*</td>
<td>-0.300(-0.502,-0.095)</td>
<td>10.71(3.5,18.9)</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>-0.195(-0.458,0.029)</td>
<td>6.77(-1.3,14.5)</td>
</tr>
<tr>
<td>Insurance*</td>
<td>-0.221(-0.380,-0.066)</td>
<td>7.93(2.3,14.4)</td>
</tr>
<tr>
<td>Social Support</td>
<td>-0.168(-0.408,0.063)</td>
<td>5.77(-2.5,13.0)</td>
</tr>
<tr>
<td>Spirituality*</td>
<td>0.178(0.040,0.315)</td>
<td>-6.46(-12.6,-1.4)</td>
</tr>
</tbody>
</table>
Results
Anxiety Score

- One good example is the results for Age at Diagnosis. We can say from the table that it explains:
  - 21.76% of the ethnic disparity in anxiety score in GLM and
  - 21.24% of the ethnic disparity in anxiety score in MART.

- This makes sense, given that the Non-Hispanic group tends to be diagnosed older (median age 62) and have lower anxiety scores, while the Hispanic group is diagnosed younger (median age 58) with higher anxiety scores.

- On average, people diagnosed at an older age have lower anxiety scores.
Results

Anxiety Score

- This can also be illustrated graphically, first with the GLM results:
Results
Anxiety Score

Then the MART:
The other variables can be interpreted similarly, with one interesting exception: Spirituality.

Spirituality is an interesting result. We’ve shown that it reduces anxiety score by 7.68% in GLM and 6.30% in MART.
Conclusions

- Our results show that education, age at diagnosis, age a person came to the U.S., insurance, comorbidities, and social support help explain the ethnic disparity in anxiety score in the MY-Health survey.

- Any developed target intervention should be paid to those survivors (particularly Hispanics) with less education, private or government insurance, more comorbid conditions, with less social support, who were diagnosed at a younger age, and who came to the U.S. at a later age.
Limitations and Future Research

- Limitations included:
  - leaving out the other domains for reasons of time;
  - choice to combine the three racial groups (white, black, API) into one Ethnic group;
  - plus survey limitations, such as low response rate and no information about comorbidity severity.

- Interesting future research would be an extension of the MMA method to accommodate multivariate outcomes.

- A version of this study was published in *Psychometrika* in April 2018. It included:
  - An analysis of only Hispanic white vs. non-Hispanic white individuals
  - Updates to the mma R package, including newly implemented graphs
  - Income, insurance, and employment were treated as joint mediators due to high correlation
Acknowledgements

- The advisor to this thesis, Dr. Qingzhao Yu, plus committee members Dr. Julia Volaufova and Dr. Xiao-Cheng Wu, all from LSUHSC-NO.
- Dr. Roxanne Jensen, Dr. Arnold Potosky, Dr. Carol Moinpour, plus the rest of the MY-Health investigative team.
- Amy Yang and NU Biostats for inviting me to come speak about my thesis today.
References


References