Constructing Physician Networks from Medical Records

James Moody
Sociology
Duke Network Analysis Center

Based on work funded by RWJ with

Lawrence Peter Casalino, Cornell Medical School
Theodore Iwashyna, U Michigan
David Nyweide, HHS
Andrew Ryan, Cornell Medical School
Matthew Press, Cornell Medical School
Michael Pesco, Cornell Medical School
Jayme Mendelsohn, Cornell Medical School
Xuming Sun, Cornell Medical School
Introduction

Rapid growth in network science & approaches to health:

Papers on Networks and Health
as a proportion of all papers on health

Year
Introduction
And scientists are starting to take network seriously: why?
Specific Aims of current NIA proposal:

**AIM 1: To identify PPCs and their characteristics by:**

1. Optimizing social network community detection algorithms to Medicare claims data:
   - a) to identify PPCs
   - b) to measure PPCs’ structural network characteristics (e.g. density of relationships among physicians; centrality of primary care physicians)

2. Surveying PPC physicians to characterize PPCs in terms of:
   - a) extent of care coordination
   - b) propensity of physicians to provide services likely to be unnecessary (as defined by the American Board of Internal Medicine Foundation’s widely endorsed Choosing Wisely list)

**AIM 2: To determine the quality and cost of care provided to Medicare patients by PPCs and to determine how quality and cost vary with PPC characteristics by:**

1. Using multiple measures to determine the quality and cost of care provided by PPCs
2. Determining the relationship between PPC characteristics and the quality and cost of care
3. Estimating the extent to which the variation in quality and cost are explained by the PPC versus explained by the individual physician.

**AIM 3: To advance methodological understanding of physician care networks by determining:**

1. The extent to which PPCs are stable over time
2. The minimum sample size for Medicare beneficiaries in a state needed to reliably identify PPCs and PPC structural characteristics
3. The extent to which PPCs are based on mutual referrals among physicians versus the extent to which they are based on patients choosing to see physicians in the same geographic area.
Basic Process:

Use the information recorded by Medicare (part B & outpatient) on patient visits to identify physicians who share patients. Initial work is based on 3 years of data for 5 states.
Basic Process:

Use the information recorded by Medicare on patient visits to identify physicians who share patients. Initial work is based on 3 years of data for 5 states.

This is a network of physicians connected by shared patients.

The weight of the connection between any two physicians is the sum over all patients of the minimum times either physician saw the patient (visits limited to one per day).

We excluded physicians who are typically hospital-based and see all patients, without specific referrals – specifically anesthesiologists, emergency physicians, hospitalists*, pathologists, and radiologists. I first present the full linkages here to get a sense of the scope.

*We defined hospitalists as physicians who provided more than 90% of eligible services in the inpatient setting and whose specialty was internal medicine, geriatrics, general practice, or family practice.
Basic Process:

Use the information recorded by Medicare on patient visits to identify physicians who share patients. Initial work is based on 3 years of data for 5 states, a 39% sample.

Table 1. Network Sample Construction

<table>
<thead>
<tr>
<th></th>
<th>Patient Visits</th>
<th>Unique Patients</th>
<th>Unique Physicians</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>12,263,448</td>
<td>922,189</td>
<td>138,375</td>
</tr>
<tr>
<td>2009</td>
<td>12,977,008</td>
<td>924,387</td>
<td>134,863</td>
</tr>
<tr>
<td>2010</td>
<td>12,167,013</td>
<td>871,993</td>
<td>135,212</td>
</tr>
<tr>
<td>Total</td>
<td>37,407,477</td>
<td>963,899</td>
<td>190,785</td>
</tr>
</tbody>
</table>

Yields a network with ~23M year-specific edges and about 15M uniquely connected dyads (because many of the dyads are connected all three years).
Results: Basic Network Descriptive Statistics, combining all states/years

Distribution of edge values (summed over all years), Log-Log scales

Mean = 4.4
Median = 2
Std.Dev = 11.6
Max = 2139
N = 15.1M
Results: Basic Network Descriptive Statistics, combining all states/years

Edges within 50 miles:
Mean = 4.93
Median = 2
Std.Dev = 12.6
Max = 2139
N = 13.1M

Edges > 50 miles:
Mean = 2.16
Median = 1
Std.Dev = 4.27
Max = 735
N = 2.8M

Distribution of edge values by distance
Results: Basic Network Descriptive Statistics, combining all states/years

Distribution of edge distances (in miles)
Geographic distribution (limited to US lower 48) of physicians seen by Ohio patients

Physician Networks of shared patients compiled from Medicare Files
Physician Networks of shared patients compiled from Medicare Files

Geographic distribution (limited to US lower 48) of physicians seen by Ohio patients

(node size=degree)
Results: Basic Network Descriptive Statistics, combining all states/years

Degree distribution
(Degree=number of other docs a doc shares at least one patient with)

<table>
<thead>
<tr>
<th>Physicians</th>
<th>Patients</th>
<th>Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1, 2, 3</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1, 2, 4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>5, 7, 6</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>5, 8, 13</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>3, 4, 9</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>2, 3, 5, 7, 8, 9</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>10, 11</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>10, 12</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>11, 12</td>
<td></td>
</tr>
</tbody>
</table>

- **Doc** | **Degree** | **Weighted Degree** |
- A       | 3          | 5                |
- B       | 3          | 4                |
- C       | 3          | 4                |
- D       | 2          | 3                |
- E       | 3          | 4                |
- F       | 5          | 9                |
- G       | 1          | 1                |
- H       | 2          | 2                |
- I       | 2          | 2                |
- J       | 2          | 2                |
Results: Basic Network Descriptive Statistics, combining all states/years

Degree Distribution: Number of other docs each doc is connected to.

Mean: 168.42
Median: 35
Std.Dev.: 267
Max: 3272
Results: Basic Network Descriptive Statistics, combining all states/years

Weighted Degree Distribution: Sum of number of patients seen in common

Mean: 748.6
Median: 61
Std.Dev.: 1665
Max: 35,278
Figure 2. Geographic distribution of physicians by patient residence
Nodes colored by degree: number of adjacent physicians (colorramp logged)
Restrict network to focus on substantive exchanges:
   a) limit by distance (drop ties longer than the 90th percentile of the state’s distribution) and look only within the largest connected component.
   b) cap degree to top 75 strongest ties
   c) use a two-stage clustering method
      stage 1: largest clusters that maximize modularity → effectively geographic
      stage 2: re-cluster within these geographic.
   d) based on balance of ties, identify core, periphery & semi-periphery members

Finding Physician Practice Communities (PPCs)
This tree outlines the case-dispensation flow for our PPC routine. The numbers in parentheses are for Ohio in 2008, just as an example.

1. We start with a full network based on overlapping patients, but rest ties by distance & degree.

2. We identify the largest connected component from 1 and then cluster at a broad “top” level to identify a small number of large (regionally-based) clusters.

3. We then apply a 2nd-level clustering within each top-level cluster to identify possible PPCs, if top-level cluster is small treat it as potential PPC. Set a single resolution parameter within each state.

4. Then adjust fit by local tie patterns (percent of ties within group, etc.).

5. Check non-capped network to see if disconnected nodes can be uniquely assigned to a group (if not, moved to background).

6. After identifying groups, we only retain groups where at least 80% of tie value is within state.

7. Distinguish core, semi-periphery & periphery based on within group tie volume.
Base network

First level clustering
Mixing Pattern of two groups with common admission to one PA hospital

(a) Map of Pennsylvania showing node distribution.

(b) Mixing Matrix: Sum of edges within and between groups

<table>
<thead>
<tr>
<th></th>
<th>PPC 1 (n=51)</th>
<th>PPC 2 (n=71)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPC 1</td>
<td>19,330</td>
<td>3,425</td>
</tr>
<tr>
<td>PPC 2</td>
<td>16,732</td>
<td></td>
</tr>
</tbody>
</table>

(c) Sociogram, layout based on tie distribution.

Caption: Panel (a): all Physicians in PA, 2008. Highlighting nodes from two groups. (b) Mixing matrix: sum of edges within & between the two groups & close-up map.
### PPC Size Distributions

<table>
<thead>
<tr>
<th>State</th>
<th>Year 2008</th>
<th>Year 2009</th>
<th>Year 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>OH</td>
<td>Mean 174</td>
<td>Mean 172</td>
<td>Mean 188</td>
</tr>
<tr>
<td></td>
<td>25% 80.0</td>
<td>25% 60.0</td>
<td>25% 101</td>
</tr>
<tr>
<td></td>
<td>Median 145</td>
<td>Median 138</td>
<td>Median 151</td>
</tr>
<tr>
<td></td>
<td>max 863</td>
<td>max 887</td>
<td>max 925</td>
</tr>
<tr>
<td></td>
<td>N 95.0</td>
<td>N 96.0</td>
<td>N 89.0</td>
</tr>
<tr>
<td>PA</td>
<td>Mean 158</td>
<td>Mean 152</td>
<td>Mean 159</td>
</tr>
<tr>
<td></td>
<td>25% 69.0</td>
<td>25% 64.0</td>
<td>25% 62.0</td>
</tr>
<tr>
<td></td>
<td>Median 127</td>
<td>Median 116</td>
<td>Median 116</td>
</tr>
<tr>
<td></td>
<td>max 933</td>
<td>max 820</td>
<td>max 804</td>
</tr>
<tr>
<td></td>
<td>N 126</td>
<td>N 133</td>
<td>N 135</td>
</tr>
<tr>
<td>TN</td>
<td>Mean 148</td>
<td>Mean 138</td>
<td>Mean 130</td>
</tr>
<tr>
<td></td>
<td>25% 82.0</td>
<td>25% 70.0</td>
<td>25% 66.0</td>
</tr>
<tr>
<td></td>
<td>Median 135</td>
<td>Median 108</td>
<td>Median 99.0</td>
</tr>
<tr>
<td></td>
<td>max 548</td>
<td>max 531</td>
<td>max 582</td>
</tr>
<tr>
<td></td>
<td>N 58.0</td>
<td>N 61.0</td>
<td>N 66.0</td>
</tr>
<tr>
<td>WV</td>
<td>Mean 120</td>
<td>Mean 116</td>
<td>Mean 107</td>
</tr>
<tr>
<td></td>
<td>25% 40.0</td>
<td>25% 46.0</td>
<td>25% 44.0</td>
</tr>
<tr>
<td></td>
<td>Median 85.0</td>
<td>Median 85.0</td>
<td>Median 79.5</td>
</tr>
<tr>
<td></td>
<td>max 533</td>
<td>max 341</td>
<td>max 344</td>
</tr>
<tr>
<td></td>
<td>N 71.0</td>
<td>N 74.0</td>
<td>N 84.0</td>
</tr>
<tr>
<td>WM</td>
<td>Mean 121</td>
<td>Mean 130</td>
<td>Mean 133</td>
</tr>
<tr>
<td></td>
<td>25% 52.0</td>
<td>25% 66.0</td>
<td>25% 68.0</td>
</tr>
<tr>
<td></td>
<td>Median 89.0</td>
<td>Median 97.0</td>
<td>Median 98.0</td>
</tr>
<tr>
<td></td>
<td>max 520</td>
<td>max 515</td>
<td>max 508</td>
</tr>
<tr>
<td></td>
<td>N 67.0</td>
<td>N 69.0</td>
<td>N 65.0</td>
</tr>
</tbody>
</table>
Types of models: variability in Pneumonia readmission rates by group

Estimates of network effect on 30 day PN readmission

Percentage point difference from mean readmission rate

Network effects

Estimate 95% CI

Preliminary results: These are not stable and should not be quoted.
Population recovery by patient sample rate

Node Recovery (PA, 2009)

Edge Recovery (PA, 2009)
Conclusions/Challenges

How do we avoid Durkheim’s particularism & share ideas?

FIELDS ARRANGED BY PURITY
MORE PURE

Sociology is just applied psychology
Psychology is just applied biology.
Biology is just applied chemistry.
Which is just applied physics. It's nice to be on top.
Oh, hey, I didn't see you guys all the way over there.

Sociologists
Psychologists
Biologists
Chemists
Physicists
Mathematicians