Using health care data to study and improve colorectal cancer screening

3rd Seattle Symposium on Health Care Data Analytics

Jessica Chubak, PhD
October 23, 2018
Disclosures and funding

- No Disclosures
- Funding (NIH):
  - R01CA121125 (SOS)
  - U24CA171524 (CRN)
  - U54CA163261, UM1CA222035 (PROSPR)
Outline

- Part 1: CRC screening as a case study
- Part 2: Practical considerations
- Part 3: Summary and recommendations
Part 1

COLORECTAL CANCER SCREENING: A CASE STUDY
Why care about colorectal cancer screening?

- CRC is 3rd most common cancer in US
- Screening can prevent incidence and death
- Gaps in screening persist
- Comparative effectiveness of screening regimens is largely unknown
Types of CRC screening

- Fecal Immunochemical Test (FIT)

Adenoma – carcinoma sequence

Role of health care data in CRC screening

- **What we should do**
  - Evidence for guidelines

- **What we are doing**
  - Identification of care gaps

- **What we can do**
  - Foundation for interventions
Example 1: Evidence for guidelines
(What we should do)

Evidence ➔ USPSTF Recommendations ➔ Coverage

AMERICAN GASTROENTEROLOGICAL ASSOCIATION

The New England Journal of Medicine

A CASE–CONTROL STUDY OF SCREENING SIGMOIDOSCOPY AND MORTALITY FROM COLORECTAL CANCER

Joe V. Selby, M.D., M.P.H., Gary D. Friedman, M.D., M.S., Charles P. Quesenberry, Jr., Ph.D., and Noel S. Weiss, M.D., Dr.P.H.
Example 2: Identification of care gaps
(What we are doing)

Positive screening stool test  Diagnostic colonoscopy
Follow-up colonoscopy after positive stool test
Example 3: Platform for interventions (What we can do)

1. Usual care
2. Automated
3. Assisted
4. Navigated

<table>
<thead>
<tr>
<th></th>
<th>Usual care</th>
<th>Automated</th>
<th>Assisted</th>
<th>Navigated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Usual care</td>
<td>Automated</td>
<td>Assisted</td>
<td>Navigated</td>
</tr>
<tr>
<td>2</td>
<td>Usual care</td>
<td>Automated</td>
<td>Assisted</td>
<td>Automated</td>
</tr>
<tr>
<td>3</td>
<td>Usual care</td>
<td>Assisted</td>
<td>Automated</td>
<td>Usual care</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>Usual care</td>
</tr>
</tbody>
</table>
Effect of intervention on CRC testing

<table>
<thead>
<tr>
<th>Intervention</th>
<th>% Current for CRC testing over 2 years (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usual care</td>
<td>26.3 (23.4-29.2)</td>
</tr>
<tr>
<td>Automated</td>
<td>50.8 (47.3-54.4)</td>
</tr>
<tr>
<td>Assisted</td>
<td>57.5 (54.5-60.6)</td>
</tr>
<tr>
<td>Navigated</td>
<td>64.7 (62.5-67.0)</td>
</tr>
</tbody>
</table>
Part 2

PRACTICAL CONSIDERATIONS
Advantages of claims- and EHR-based research

- Mitigation of selection bias
- Generalizability
- Not subject to recall bias
Health care data

What you want

https://www.marthastewart.com

What you get

https://www.foodnetwork.com

http://figareau.blogspot.com
Data sources

Claims
- Procedure codes
- Diagnosis codes

Structured clinical data
- Forms
- Laboratory results

Unstructured clinical data
- Pathology report text
- Colonoscopy reports
Data needed for CRC screening research

- Test indication
- Tests results
- Cancer incidence
- Cancer mortality
- Cancer risk factors
### Challenges with using health care data

<table>
<thead>
<tr>
<th></th>
<th>Claims</th>
<th>Structure clinical</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test indication</td>
<td>No</td>
<td>Sometimes</td>
<td>Yes</td>
</tr>
<tr>
<td>Tests results</td>
<td>No</td>
<td>Sometimes</td>
<td>Yes</td>
</tr>
<tr>
<td>Cancer incidence</td>
<td>Limited accuracy</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Cancer mortality</td>
<td>No</td>
<td>No</td>
<td>Sometimes</td>
</tr>
<tr>
<td>Cancer risk factors</td>
<td>Sometimes</td>
<td>Sometimes</td>
<td>Sometimes</td>
</tr>
</tbody>
</table>
Opportunities for getting the necessary data

1. Link data sources
2. Develop and validated EHR-based algorithms
3. Find or create structured data
4. Use unstructured data (i.e., text)
1. Data linkages

https://seer.cancer.gov/registries/

*Subcontract under New Mexico
**Three regions represent the state of California: Greater Bay, Los Angeles, and Greater California
Caveats with linkages

- Not everyone can be linked
- Requires sharing direct identifiers
- Linkages aren’t always perfect.
2. EHR algorithms for cancer screening research

- Px / Dx codes, lab results, etc.
- Classification trees, regression, etc.
- Binary variables, probabilities, etc.
## Example: colonoscopy indication

<table>
<thead>
<tr>
<th>Has the patient had...</th>
<th>Exam Indication:</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the last 10 years: inflammatory bowel disease diagnosis</td>
<td>Yes → Diagnostic indication</td>
</tr>
<tr>
<td>In the last 12 months: positive fecal test</td>
<td>Yes → Diagnostic indication</td>
</tr>
<tr>
<td>In the last 6 months: gastrointestinal symptoms (abdominal pain; iron-deficiency anemia; gastrointestinal bleeding or blood in stools; diarrhea or change in bowel habits; unexplained weight loss; diverticulitis; referral due to abnormal abdominal imaging [CT, MRI, or barium enema])</td>
<td>Yes → Diagnostic indication</td>
</tr>
<tr>
<td>In the last 6 months: initial CRC diagnosis from cancer registry or CRC diagnostic code from pathology</td>
<td>Yes → Diagnostic indication</td>
</tr>
<tr>
<td>In the last 12 months: most recent diagnostic code for adenoma from pathology</td>
<td>Yes → Diagnostic indication</td>
</tr>
<tr>
<td>In the last 12 months: sigmoidoscopy or colonoscopy with ICD-9 diagnosis of polyp (+1 week of procedure)</td>
<td>Yes → Diagnostic indication</td>
</tr>
<tr>
<td>More than 6 months before: initial CRC diagnostic code from pathology or cancer registry</td>
<td>Yes → Surveillance indication</td>
</tr>
<tr>
<td>Only has ICD-9 diagnosis of CRC or history of CRC</td>
<td>Yes → Surveillance indication</td>
</tr>
<tr>
<td>More than 12 months before: most recent diagnostic code for adenoma from pathology</td>
<td>Yes → Surveillance indication</td>
</tr>
<tr>
<td>Only has prior ICD-9 diagnosis of colon adenoma or history of colon adenoma</td>
<td>Yes → Surveillance indication</td>
</tr>
<tr>
<td>Only has prior ICD-9 diagnosis of colon polyp or history of colon polyp</td>
<td>Yes → Surveillance indication</td>
</tr>
<tr>
<td>Family history of CRC</td>
<td>Yes → Screening indication (high risk)</td>
</tr>
</tbody>
</table>

**Diagnostic**

**Surveillance**

**Screening**

Caveats with EHR-based algorithms

- Misclassification
  - Missing data
  - Coding errors
- Window of data availability
- Different coding practices in different settings
- Switch from ICD-9 to ICD-10
Find or create structured data

- Leverage reporting requirements
  - Mammography Quality Standards Act
  - CMS-approved lung cancer screening registries
- Partner with care providers and delivery systems
  - Develop reporting systems
Example: Colonoscopy Reporting System

**Procedure details**

**Findings**

**Recommendations**

Use unstructured data

Chart abstraction

Natural language processing
Example: NLP for high grade dysplasia

**Algorithm**

1. Look for key words (e.g., dysplasia, dysplastic)
2. Exclude if preceded by negation key words (e.g., no evidence)
3. Include if preceded by “high grade” key words
Caveats with text

- Not always available
- Expensive to work with
- Charting practices (and definitions) can be inconsistent
Part 3

SUMMARY AND RECOMMENDATIONS
Summary

- Healthcare data are critical for improving CRC screening
- Healthcare data offer many advantages
- Data are not always “research ready”
Recommendations

1. Leverage the advantages of health care data
2. Know your source data and its limitations
3. Consider different data collection approaches
4. Validate your approaches
5. Don’t go it alone: find networks and consortia
Thank you

jessica.chubak@kp.org