Semi-Automating Knowledge Base Construction for Cancer Genetics

Somin Wadhwa      Kanhua Yin      Kevin S. Hughes      Byron C. Wallace

Northeastern University
Khoury College of Computer Sciences

UMassAmherst
College of Information & Computer Sciences

Massachusetts General Hospital
Harvard Medical School
Penetrance papers Medical literature describing risk of cancer with a particular pathogenic variant in cancer susceptibility gene.
Key study information

- What population was studied? **Ascertainment**!
- How many patients were in the study?
- What cancer was the patient at-risk for? What was the associated risk?
- Ideally: Synthesize the key elements from papers into a database.
  - e.g. ask2me.org

<table>
<thead>
<tr>
<th>PMID</th>
<th>Gene</th>
<th>Cancer</th>
<th>Race</th>
<th>OR</th>
<th>RR</th>
<th>HR</th>
<th>Max Age</th>
<th>Total Carriers</th>
</tr>
</thead>
<tbody>
<tr>
<td>29922827</td>
<td>BRCA2</td>
<td>Pancreatic</td>
<td>Multiple</td>
<td>6.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>370</td>
</tr>
<tr>
<td>29922827</td>
<td>TP53</td>
<td>Pancreatic</td>
<td>Multiple</td>
<td>6.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>31</td>
</tr>
<tr>
<td>27595995</td>
<td>CHEK2</td>
<td>Breast</td>
<td>White</td>
<td>3.39</td>
<td>-</td>
<td>-</td>
<td>75</td>
<td>11</td>
</tr>
<tr>
<td>21145788</td>
<td>MSH2</td>
<td>Colorectal</td>
<td>Multiple</td>
<td>-</td>
<td></td>
<td>0.49</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Key study information

- What population was studied? **Ascertainment**!
- How many patients were in the study?
- What cancer was the patient at-risk for? What was the associated risk?
- Ideally: Synthesize the key elements from papers into a database.
  - e.g. [ask2me.org](http://ask2me.org)

<table>
<thead>
<tr>
<th>PMID</th>
<th>Gene</th>
<th>Cancer</th>
<th>Race</th>
<th>OR</th>
<th>RR</th>
<th>HR</th>
<th>Max Age</th>
<th>Total Carriers</th>
</tr>
</thead>
<tbody>
<tr>
<td>29922827</td>
<td>BRCA2</td>
<td>Pancreatic</td>
<td>Multiple</td>
<td>6.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>370</td>
</tr>
<tr>
<td>29922827</td>
<td>TP53</td>
<td>Pancreatic</td>
<td>Multiple</td>
<td>6.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>31</td>
</tr>
<tr>
<td>27595995</td>
<td>CHEK2</td>
<td>Breast</td>
<td>White</td>
<td>3.39</td>
<td>-</td>
<td>-</td>
<td>75</td>
<td>11</td>
</tr>
<tr>
<td>21145788</td>
<td>MSH2</td>
<td>Colorectal</td>
<td>Multiple</td>
<td>-</td>
<td>-</td>
<td>0.49</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
The problem: Too many studies

We seek to semi-automate data extraction from cancer genetics literature to aid KB construction
A control population was defined from the National Danish Civil Registration System, matched for sex, year of birth, mutation carriers as well as first degree relatives.

For age adjusted analysis, the projected U.S. population was used (year 2000); 84% of the 3499 individuals were white.
These included **CDKN2A**, with mutations in 0.30% of cases and 0.02% of controls (OR, **12.33**; 95% CI, 5.43-25.61);

**TP53**, with mutations in 0.20% of cases and 0.02% of controls (OR, **6.70**; 95% CI, 2.52-14.95);

**MLH1**, with mutations in 0.13% of cases and 0.02% of controls (OR, **6.66**; 95% CI, 1.94-17.53);

**BRCA2**, with mutations in 1.90% of cases and 0.30% of controls (OR, **6.20**; 95% CI, 4.62-8.17);

**ATM**, with mutations in 2.30% of cases and 0.37% of controls (OR, **5.71**; 95% CI, 4.38-7.33);
A KRAS-Variant in Ovarian Cancer Acts as a Genetic Marker of Cancer Risk

Abstract: Ovarian Cancer is the single most deadly form of women's cancer, typically presented as an advanced disease at diagnosis in part due to a lack of known risk factors or known genetic markers of risk.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure
A KRAS-Variant in Ovarian Cancer Acts as a Genetic Marker of Cancer Risk

Abstract: KRAS-Cancer is the most common form of women's cancer, typically presenting as an identified disease. The isolation of the disease is often due to a lack of early risk factors or advanced genetic markers.

Full-text PDFs of Gene-Cancer Studies

<germline-mutation, risk>  <germline-mutation, risk>  ...  <germline-mutation, risk>
F1 Scores

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.78</td>
<td>0.44</td>
</tr>
<tr>
<td>0.79</td>
<td>0.49</td>
</tr>
<tr>
<td>0.77</td>
<td>0.61</td>
</tr>
<tr>
<td>0.78</td>
<td>0.62</td>
</tr>
</tbody>
</table>

- d-BERT
- d-SciBERT
- j-BERT
- j-SciBERT

d – disjoint; j – joint
Thank you!

sominwadhwa.com || @sominw || sominwadhwa@cs.umass.edu