Predictive Model for Detection of Colorectal Cancer in Primary Care by Analysis of Complete Blood Counts


Slides: Dan Coster
Colorectal Cancer (CRC)

- Globally more than 1 million people get colorectal cancer every year (In 2012, 1.4 million new cases).
- 2nd most common cause of cancer in women (9.2% of diagnoses)
- 3rd most common in men (10% of diagnoses)
- Screening Tests:
  - Colonoscopy (recommended above age 50)
  - gFOBT (guaiac fecal occult blood test)

- Many avoid both tests and CRC discovered too late. Can regular blood tests help?
Data

• **Data Sets:**
  - **Maccabi Health Care** - persons of age 40+ between 01/2003 - 06/2011
  - **The Health Improvement Network (THIN)** - UK data, persons of age 40+ between 01/2007 - 05/2012

• **CRC Labeling:**
  - **Maccabi** - via Israel National Cancer Registry
  - **THIN** - based on patients records (documented Tumors / medications)
# Data

## Input data

**Derivation (training) set data (Israeli dataset)**

466,107 individuals with CBC / 2,437 CRC patients

<table>
<thead>
<tr>
<th>Patient (#)</th>
<th>Date of birth</th>
<th>Gender</th>
<th>CRC diagnosis date</th>
<th>CBC date</th>
<th>CBC parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>5.8.39</td>
<td>M</td>
<td>10.10.04</td>
<td>8.1.03</td>
<td>MCH 30.4, WBC 6.0, MCV 89</td>
</tr>
<tr>
<td>#2</td>
<td>2.3.44</td>
<td>F</td>
<td>Healthy</td>
<td>5.1.06</td>
<td>MCH 29, WBC 6.7, MCV 87</td>
</tr>
<tr>
<td>#</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>#466,107</td>
<td>7.2.57</td>
<td>F</td>
<td>Healthy</td>
<td>8.8.06</td>
<td>MCH 28, WBC 6.8, MCV 90</td>
</tr>
</tbody>
</table>
Table 2: CBC characteristics of females in the Israeli and UK datasets

<table>
<thead>
<tr>
<th>CBC values, median (Q1–Q3)</th>
<th>CRC patients</th>
<th>Cancer-free individuals&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Israel&lt;sup&gt;a&lt;/sup&gt;</td>
<td>UK&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Red Blood Cell Count (RBC)</td>
<td>4.2 (3.9–4.6)</td>
<td>4.2 (3.8–4.5)</td>
</tr>
<tr>
<td>White Blood Cell Count (WBC)</td>
<td>7 (5.8–8.4)</td>
<td>8 (6.6–9.9)</td>
</tr>
<tr>
<td>Mean Platelet Volume (MPV)</td>
<td>10.6 (10–11.3)</td>
<td>9 (8.1–10)</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>11.7 (9.6–12.8)</td>
<td>11 (9.4–12.8)</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>36.2 (31.4–39.6)</td>
<td>34.5 (30.3–39.2)</td>
</tr>
<tr>
<td>Mean Corpuscular Volume (MCV)</td>
<td>85 (80–89)</td>
<td>85 (77.5–89.3)</td>
</tr>
<tr>
<td>Mean Corpuscular Hemoglobin (MCH)</td>
<td>27.4 (24.8–29.1)</td>
<td>27.5 (24.4–29.4)</td>
</tr>
<tr>
<td>Mean Corpuscular Hemoglobin Concentration (MCHC)</td>
<td>31.8 (30.8–32.7)</td>
<td>32.2 (30.9–33.1)</td>
</tr>
<tr>
<td>Red blood cell Distribution Width (RDW)</td>
<td>14.6 (13.7–16)</td>
<td>NA (NA–NA)</td>
</tr>
<tr>
<td>Platelets</td>
<td>296 (246.8–366)</td>
<td>362 (284–444)</td>
</tr>
<tr>
<td>Eosinophils (#)</td>
<td>0.16 (0.1–0.26)</td>
<td>0.13 (0.1–0.23)</td>
</tr>
<tr>
<td>Eosinophils (%)</td>
<td>2.3 (1.4–3.6)</td>
<td>2.2 (1.7–3)</td>
</tr>
<tr>
<td>Neutrophils (#)</td>
<td>4.1 (3.3–5.3)</td>
<td>5.3 (4.2–7)</td>
</tr>
<tr>
<td>Neutrophils (%)</td>
<td>60 (53.7–66.5)</td>
<td>68.7 (63.1–72.2)</td>
</tr>
<tr>
<td>Monocytes (#)</td>
<td>0.58 (0.46–0.72)</td>
<td>0.6 (0.5–0.8)</td>
</tr>
<tr>
<td>Monocytes (%)</td>
<td>8.2 (7.1–9.9)</td>
<td>7.9 (6.8–9.9)</td>
</tr>
<tr>
<td>Basophils (#)</td>
<td>0.03 (0.02–0.04)</td>
<td>0.02 (0–0.09)</td>
</tr>
<tr>
<td>Basophils (%)</td>
<td>0.4 (0.3–0.6)</td>
<td>0.35 (0.2–1)</td>
</tr>
<tr>
<td>Lymphocytes (#)</td>
<td>1.9 (1.5–2.4)</td>
<td>1.7 (1.3–2.2)</td>
</tr>
<tr>
<td>Lymphocytes (%)</td>
<td>28.5 (22.9–34)</td>
<td>20.3 (17–26.5)</td>
</tr>
</tbody>
</table>

Israel: median of 8 CBC tests in 7 years

UK: 3
Design

• Derivation (training) Set:
  - 80% of Maccabi’s data (466,107 patients, 2,437 with CRC)
  - Patients with other types of cancer were excluded.
  - Each individual was assigned a single score to avoid bias of over-representing individuals with many blood counts.
  - **Cases**: The most recent CBC within a defined **time window** (3-6 months/ 0-30 days) prior to CRC diagnosis was used
  - Individuals with no CBC in the time window were excluded.
  - **Controls**:
    • Individuals aged 50-75 (per current CRC screening guidelines)
    • A single randomly selected CBC was used *(age bias?)*

• Validation Sets:
  - **Israel validation set**: other 20% - 139,205 patients, 698 with CRC
  - **UK validation set**: THIN data - 25,613 patients, 5,061 with CRC.
Data Preparation

- Generated 62 features per patient:
  - Age (on CBC date)
  - Gender
  - 20 CBC values on CBC date
  - 20 CBC estimated values* 18 + 36 months before current CBC date

*Computed by linear regression on past CBCs of the individual
Algorithm

- Using Random Forest & Gradient Boosting.
Decision Trees (reminder)

Split (node, {Examples}):
1. $A \leftarrow$ The best feature for splitting the {examples}
2. Decision feature for this node $\leftarrow A$
3. For each value of $A$, create new child node
4. Split training examples to child nodes
5. For each child node / subset:
   1. If subset is pure: STOP
   2. Else: Split (child_node, {subset})

Splitting criterion:
Information gain

Random Forest

1. Select $ntree$, no. of trees to grow, and $mtry$, no. of vbls to use.
2. For $i = 1$ to $ntree$:
   1. Draw a bootstrap sample (subsample with replacement) from the data. Call those not in the sample the out-of-bag data.
   2. Grow a tree, where at each node, the best split is chosen among $mtry$ randomly selected variables. The tree is grown to maximum size and not pruned.
   3. Use the tree to predict out-of-bag data.
6. In the end, use the predictions on out-of-bag data to form majority votes.
7. Prediction of test data is done by majority votes from predictions from the ensemble of trees.

www.rci.rutgers.edu/~cabrera/587/
Results

- On Israel validation set:
  - AUC $0.82 \pm 0.01$
  - OR at FP=0.5% $26 \pm 5$
  - Specificity at 50% sensitivity $88\% \pm 2\%$

- On UK validation set:
  - AUC $0.81$
  - OR at FP=0.5% $40$
  - Specificity at 50% sensitivity $94\%$
Sensitivity at 0.5% specificity
Variable importance

• Want to find out which vbls are most important for the prediction.
• Age is the most important contributing parameter ($AUC = 0.72$ with age alone)
• What about the contribution of blood variables?
• Problem: Some of the blood variables are highly correlated
• Soln here:
  - Compute two parameters **Contribution & Redundancy**
  - Use **iterative removal of parameters** from the model
  - Evaluate performance by AUC
Computing variable importance

Repeat for each CBC variable \( p \): (example: Hg)

1. Let \( \Delta_0 \) be the decrease of AUC between the full model and the model without \( p \).
2. Sort the other 19 vbls by their correlation with \( p \).
3. For \( K = 1, 19 \):
   1. Remove from the data the \( K \) vbls most correlated with \( p \) (but not \( p \) itself!). Construct a new partial model on these data.
   2. Let \( \lambda_k \) be the decrease in AUC between the full model and the new partial model.
   3. Let \( \Delta_k = \lambda_k - \Delta_0 \)
4. Define the contribution of \( p \) as \( \max_i \{ \Delta_i \} \).
5. Define the redundancy of \( p \): \( \min \{ i | \Delta_i > \text{Threshold} \} \).

An important vbl: high contribution, low redundancy
Variable importance

**Contribution of parameters to AUC at 90–180 days before diagnosis**

- Hgb
- MCH
- MCHC
- Hct
- MCV
- RDW
- Plts
- RBC
- Mon#
Using the model + gFOBT

• Compared the method’s CRC detection rate to that of gFOBT on the Israeli cohort.
• Data: 75,822 gFOBT tests for 63,847 individuals, compared to 210,923 individuals with CBCs.
• The gFOBT positive rate was 5%, detecting 170 CRC cases.
• The model discovered 252 CRC cases.
• By considering individuals who were identified either by the model or by gFOBT the number of CRC cases detected would increase by 115%: from 70 to 365.
Performance on other cancers

- Does the same model detect other cancer types?
- Applied the model; examined the sensitivity at a FP rate of 3% (shifted from 0.5% to get reliable results on less common cancers).
Early predictions?

- Results using only CBCs performed 0-2 months, 2-4, ..., 20-22 months prior to CRC diagnosis
בריאותו

111 איש אחרי כנפיים סוכן לחoldt בטרון

המוות הנכון

מודרב בדיקה חסינת פיתוחה בקופת חוליםобщаיגי • נשטת על המשכילים שנהנו
הבריאת בדיקה כיוןMohammad של ספורט עלי הרג

משה יונה בירור //

http://www.israelhayom.co.il/article/436995 12/2016

ABDBM © Ron Shamir

http://www.israelhayom.co.il/article/436995 12/2016
Recap

• CRC predictor based on age, gender, temporal blood measurements
• Uses decision tree ensemble methods: RF, gradient boosting
• “Feature engineering” key to the results
• Results improve over FOBT, anemia guidelines
• Can detect risk of CRC 3-6 (12?) months ahead of time
• Great potential