Cancer Etiology and Natural History: A Web Tool for Age-Period-Cohort Analysis

Philip S. Rosenberg
David P. Check
William F. Anderson
&
The NCI Center for Biomedical Informatics and Information Technology (CBIIT)

Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute

06 March 2014, DCEG Seminar
### Input

<table>
<thead>
<tr>
<th>Age Deviations</th>
<th>Per Deviations</th>
<th>Coh Deviations</th>
<th>Long Age</th>
<th>Cross Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long 2 CrossRR</td>
<td>Fitted Temporal Trend</td>
<td>PeriodRR</td>
<td>CohortRR</td>
<td>Local Drifts</td>
</tr>
</tbody>
</table>

### Table

<table>
<thead>
<tr>
<th>Age</th>
<th>Count</th>
<th>Population</th>
<th>Count</th>
<th>Population</th>
<th>Count</th>
<th>Population</th>
<th>Count</th>
<th>Population</th>
</tr>
</thead>
</table>

Copy and paste into table on right or upload a csv with population and count information.

- [Browse...](#)
- Clear
- Calculate
• **What does it do?**
• **Why is that important?**
• **Who built it, and how?**
• **How do I use it?**
Acknowledgments

CBIIT

• Robert Shirley, NCI CBIIT
• Sue Pan, NCI CBIIT
• Larry Brem, Leidos Biomedical Research, Inc. (NCI CBIIT Dev Team Contractor)
• Brent Coffey, Leidos Biomedical Research, Inc. (NCI CBIIT Dev Team Contractor)
• Shaun Einolf, Leidos Biomedical Research, Inc. (NCI CBIIT Dev Team Contractor)
• Sula Rajapakse, Leidos Biomedical Research, Inc. (NCI CBIIT Dev Team Contractor)
• Cuong Nguyen, SRA International, Inc. (NCI CBIIT Systems Team Contractor)

Division of Cancer Epidemiology & Genetics

Discovering the causes of cancer and the means of prevention
• Carl McCabe, Office of Division Operations and Analysis
• Sholom Wacholder, Nicolas Wentzensen, Christine Fermo
  • http://analysistools.nci.nih.gov/meanstorisk/
Outline

• The APC Model
  • Overview
  • Examples from the literature

• The Web Tool
  • What’s in it (and why)
  • How it works
APC Model: Overview

• Macro-epidemiological model for population-based cancer surveillance data
  o Incidence and Mortality
  o SEER, IARC, other large-scale open cohorts

• Parametric approach
  o complements traditional descriptive approaches

• Quantification (via parameters and functions)
  o Burden
  o Trends
  o Natural History
  o Etiology
  o Disparity
Overview

Examples from the literature

APC Model: Data

A registry is a cohort of cohorts . . .

Rate matrix or Lexis diagram for invasive female breast cancer.

Rosenberg P S , and Anderson W F Cancer Epidemiol Biomarkers Prev 2011;20:1263-1268

©2011 by American Association for Cancer Research
APC Model: Parameters from Data

**Longitudinal Form**

\[ \rho_{ac} = \mu + (\alpha_L + \pi_L)(a - \bar{a}) + (\pi_L + \gamma_L)(c - \bar{c}) + \tilde{\alpha}_a + \tilde{\pi}_p + \tilde{\gamma}_c \]

**Cross-sectional Form**

\[ \rho_{ap} = \mu + (\alpha_L - \gamma_L)(a - \bar{a}) + (\pi_L + \gamma_L)(p - \bar{p}) + \tilde{\alpha}_a + \tilde{\pi}_p + \tilde{\gamma}_c \]
**APC Model: Putting the pieces together**

Through independent and collaborative descriptive studies, we developed a panel of standard and novel functions** and corresponding hypothesis tests that appear to be effective in identifying signatures or patterns in disease rates for many types of cancers.

** linear combination of estimable parameters in the APC Model
APC Model: Key Parameters, Functions, and Tests

• **Net Drift is the single most important parameter!**
  - Model analogue of EAPC adjusted for cohort effects
  - Determines ratio of Longitudinal to Cross-sectional Age Curves

• **Age effects** (Longitudinal and Cross-sectional Age Curves)
• **Period effects** (Fitted Temporal Trends, Period RR)
• **Cohort effects** (Cohort RR; Local Drifts = age-specific EAPC)

• **The Significance Test for Local Drifts is the second most important APC statistic!**
  - Tells you if you have important cohort effects
Outline

• The APC Model
  • Overview
  • Examples from the literature

• The Web Tool
  • What’s in it (and why)
  • How it works
Trimodal age-specific incidence patterns for Burkitt lymphoma in the United States, 1973–2005

Sam M. Mbulaiyeye¹, William F. Anderson², Kishor Bhatia¹, Philip S. Rosenberg², Martha S. Linet³ and Susan S. Devesa²

**Overview**

**Examples from the literature**

**Age Effects**
(Longitudinal Age Curve)

**Cross-Sectional**

**Longitudinal**

Figure 2, Burkitt lymphoma APC model-based expected period- and cohort-specific age-specific incidence rates by sex, SEER 9, 13 and 17 registries, 1982–2005. Panels c and d included “fitted” age-at-onset curves (see Methods). Cases diagnosed during 1973–1981 were excluded because of sparse numbers. LAT = longitudinal age trend.
Period Effects
(Deviations, Period RR)
Cohort Effects (Cohort RR)
Cohort Effects (Local Drifts)

Age-specific temporal trends

Local Drifts
Outline

• The APC Model
  • Overview
  • Examples from the literature

• The Web Tool
  • What’s in it (and why)
  • How it works
All of the APC functions (and corresponding hypothesis tests) shown in *Examples from the literature* are produced by the Web Tool.
What’s in it (and why)

Key Functions

The most important functions calculated by the web tool are summarized in the Table of Key Functions using the following conventions:

- The APC model is defined over $A$ age groups and $P$ calendar periods with equal intervals.
- The central age group, calendar period, and birth cohort define standard reference values $a$, $p$, and $c$, respectively.
- When there is an even number of age, period, or cohort categories, the reference value is the lower of the two central values.
- All values labeled CI Lo are lower 95% confidence limits. All values labeled CI Hi are upper 95% confidence limits.

<table>
<thead>
<tr>
<th>Nomenclature</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitted Temporal Trend</td>
<td>Expected age-adjusted rates over time</td>
</tr>
<tr>
<td>Net Drift</td>
<td>Annual percentage change of the expected age-adjusted rates over time</td>
</tr>
<tr>
<td>Local Drifts</td>
<td>Annual percentage change of the expected age-specific rates over time</td>
</tr>
<tr>
<td>Cross-Sectional Age Curve (Cross Age)</td>
<td>Expected age-specific rates within a given calendar period</td>
</tr>
<tr>
<td>Longitudinal Age Curve (Long Age)</td>
<td>Expected age-specific rates for a given number of birth cohorts</td>
</tr>
<tr>
<td>Period Rate Ratios (PeriodRR)</td>
<td>Ratio of age-specific rates in period relative to reference period</td>
</tr>
<tr>
<td>Cohort Rate Ratios (CohortRR)</td>
<td>Ratio of age-specific rates in cohort relative to reference cohort</td>
</tr>
</tbody>
</table>

Other parameters and functions calculated by the web tool and not listed above are described in greater mathematical detail elsewhere (click here for a PowerPoint presentation).
Hypothesis Tests

Statistical hypothesis tests calculated by the web tool are summarized in the Table of Hypothesis Tests.

The Wald Tests follow a Chi-Square distribution when the Null Hypothesis is true. The df (degrees of freedom) count the number of free parameters included in each test. The web tool reports P-values; values less than 0.05 are often considered 'statistically significant', meaning there is statistical evidence that the Null Hypothesis is unlikely to be correct.

### Table of Hypothesis Tests

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Implications</th>
<th>Degrees of Freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net drift = 0</td>
<td>Fitted temporal trends are stable (i.e., flat with no change) over time. Fitted longitudinal and cross-sectional age curves are equal.</td>
<td>1</td>
</tr>
<tr>
<td>All age deviations = 0</td>
<td>Fitted longitudinal and cross-sectional age curves are log-linear (i.e., log-additive).</td>
<td>A - 2</td>
</tr>
<tr>
<td>All period deviations = 0</td>
<td>Fitted temporal trends and period rate ratios are log-linear (i.e., log-additive).</td>
<td>P - 2</td>
</tr>
<tr>
<td>All cohort deviations = 0</td>
<td>Cohort rate ratios are log-linear, all local drifts equal the net drift.</td>
<td>C - 2</td>
</tr>
<tr>
<td>All period rate ratios = 1</td>
<td>Net drift is 0 and fitted temporal trends are constant; Cross-sectional age curve describes age incidence pattern in every period.</td>
<td>P - 1</td>
</tr>
<tr>
<td>All cohort rate ratios = 1</td>
<td>Net drift is 0 and all local drifts are 0; Longitudinal age curve describes age incidence pattern in every cohort.</td>
<td>C - 1</td>
</tr>
<tr>
<td>All local drifts = the net drift</td>
<td>Temporal patterns are the same in every age group.</td>
<td>A</td>
</tr>
</tbody>
</table>

*For APC model defined over A age groups, P calendar periods, and C = P + A - 1 birth cohorts.*
Web Tool: Usability

• We paid a lot of attention to **workflow**.

• The Web Tool promotes **reproducible research**.

• We think it is really simple to use.
Web Tool: Architecture

**What’s in it (and why)**

- **R** (Back-End Server)
  - MATLAB (Prototyping)
  - Python + JavaScript (Front-End Server)

**How it works**

- SEERSTAT
- Excel
- csv
- R
- Excel
- Text

**User** (Browser)
Data Input

Getting started

Input data for the web tool consist of **Count** and **Population** data for particular age groups over calendar time, in the form of a matrix of rows with paired columns. Rows correspond to particular age groups and columns correspond to calendar time periods. The age and period intervals must all be equal, i.e. if 5-year age groups are used then 5-year calendar periods must also be used. The data can be input by copy-and-paste from an Excel worksheet or file upload of a comma-separated-values (csv) file.

**To input from Excel:**

1. Copy the paired columns of data you want to analyze from your spreadsheet, right-click inside the empty matrix on the Input tab, and paste your selection.

2. Fill in the information (meta-data) on the left hand side of the Input page:
   - **Title** - describe your data
   - **Description** - add optional details
   - **Start Year** - list the first calendar year of the first calendar period of your data, for example, use 1990 for the interval 1990 - 1994
   - **Start Age** - list the first age of the first age group of your data, for example use 30 for the interval 30 - 34
   - **Interval (Years)** - the width of the age and period intervals, for example use 1 for single-year data, 2 for two-year data, 5 for five-year data (e.g., 1990 - 1994), etc.

3. Click the calculate button
Data Input Using Excel

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 - 29</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 - 34</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35 - 39</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 - 44</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45 - 49</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 - 54</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55 - 59</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 - 64</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65 - 69</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70 - 74</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75 - 79</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


23
Data Input Using CSV

What’s in it (and why)

### Title: Belgium Female Lung Cancer Mortality

<p>| | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2</td>
<td>Title: Belgium Female Lung Cancer Mortality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Start Year: 1955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Start Age: 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Interval (Years): 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>3</td>
<td>1578947</td>
<td>1538462</td>
<td>7</td>
<td>1400000</td>
<td>3</td>
<td>1578947</td>
<td>10</td>
<td>1428571</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>1666667</td>
<td>162653</td>
<td>11</td>
<td>1527778</td>
<td>10</td>
<td>1408451</td>
<td>7</td>
<td>1228070</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1410256</td>
<td>22</td>
<td>1666667</td>
<td>24</td>
<td>1632653</td>
<td>25</td>
<td>1524390</td>
<td>15</td>
<td>1136364</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1348315</td>
<td>44</td>
<td>1392405</td>
<td>42</td>
<td>1660079</td>
<td>35</td>
<td>1658047</td>
<td>48</td>
<td>1221374</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1590009</td>
<td>74</td>
<td>1321429</td>
<td>68</td>
<td>1379310</td>
<td>99</td>
<td>1636364</td>
<td>88</td>
<td>1288433</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1600061</td>
<td>131</td>
<td>1541176</td>
<td>99</td>
<td>1294118</td>
<td>142</td>
<td>1340888</td>
<td>134</td>
<td>1285988</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>157</td>
<td>1515444</td>
<td>184</td>
<td>1533333</td>
<td>189</td>
<td>1490536</td>
<td>180</td>
<td>1255230</td>
<td>177</td>
<td>986072.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>193</td>
<td>1307588</td>
<td>232</td>
<td>1417227</td>
<td>262</td>
<td>1455556</td>
<td>249</td>
<td>1414773</td>
<td>239</td>
<td>999581.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>219</td>
<td>1066732</td>
<td>267</td>
<td>1181146</td>
<td>323</td>
<td>1297189</td>
<td>325</td>
<td>1335799</td>
<td>334</td>
<td>1048930</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>223</td>
<td>849847.6</td>
<td>250</td>
<td>902527.1</td>
<td>308</td>
<td>1010830</td>
<td>412</td>
<td>1115322</td>
<td>358</td>
<td>930595.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>198</td>
<td>591574.5</td>
<td>214</td>
<td>636715.3</td>
<td>253</td>
<td>688060.9</td>
<td>338</td>
<td>773632.4</td>
<td>312</td>
<td>690265.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

How it works
Click on this button . . .
What’s in it (and why)

How it works

Scroll down . . .
Hypothesis Tests...
What’s in it (and why)

How it works

Division of Cancer Epidemiology and Genetics
Biostatistics Branch

Input  Age Deviations  Per Deviations  Coh Deviations  Long Age  Cross Age
Long 2 CrossRR  Fitted Temporal Trend  PeriodRR  CohortRR  Local Drifts

Local Drifts with Net Drift

Percent per Year

Age
What’s in it (and why)
Model Outputs in Excel
Conclusions

• **What does it do?**
  The tool fits the APC Model and serves up Model Outputs.

• **Why is that important?**
  Many cancers present complicated patterns. The outputs complement and extend standard descriptive methods.

• **Who built it, and how?**
  BB – concept, design, computations
  CBIIT – “Webification”

• **How do I use it?**