Age-Period-Cohort Models in Cancer Surveillance Research: Ready for Prime Time?

Philip S. Rosenberg, PhD
William F. Anderson, MD, MPH

May 5, 2011

Raw text ~ 2055 words

**DISCLAIMER:** None of the co-authors has a financial conflict of interest that would have affected this research.

**ACKNOWLEDGEMENTS:** This research was supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute. All of the authors had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
ABSTRACT (146 words, unstructured)
Standard descriptive methods for the analysis of cancer surveillance data include canonical plots based on the lexis diagram, directly age-standardized rates (ASR), estimated annual percentage change (EAPC), and joinpoint regression. The age-period-cohort (APC) model has been used less often. Here, we argue that it merits much broader use. Firstly, we describe close connections between estimable functions of the model parameters and standard quantities such as the ASR, EAPC, and joinpoints. Estimable functions have the added value of being fully adjusted for period and cohort effects, and generally more precise. Secondly, the APC model provides the descriptive epidemiologist with powerful new tools, including rigorous statistical methods for comparative analyses and the ability to project the future burden of cancer. We illustrate these principles using invasive female breast cancer incidence in the United States, but these concepts apply equally well to other cancer sites for incidence or mortality.

Key words: Cancer Surveillance Research, age standardized rates (ASR), estimated annual percentage change (EAPC), joinpoint regression, Age-period-cohort (APC) model, APC estimable parameters, APC linear trends, APC deviations, APC drifts, and APC fitted age at onset curve
Introduction

Cancer incidence and mortality rates are closely monitored to track the burden of cancer and its evolution in populations (1-4), provide etiological clues (5-11), reveal disparity (12-14), and gauge the dissemination of screening modalities (15-17) and therapeutic innovations (18, 19). A standard “toolbox” of graphical and quantitative methods has evolved to handle the needs of cancer surveillance researchers. Perhaps the most widely used methods include classical descriptive plots based on the lexis diagram (20-22), directly age-standardized rates (ASR) (23), estimated annual percentage change (EAPC) (24), and the joinpoint regression method (25). The underlying philosophy is agnostic and empirical; hence standard tools are particularly well suited to descriptive, exploratory, and hypothesis-generating studies.

At the same time, the age-period-cohort (APC) model has been developed in the statistics literature as a mathematical counter-point to purely descriptive approaches (20, 26-33). The APC model is based on fundamental generalized linear model theory (34); in principle, it allows the descriptive epidemiologist to both generate and test hypotheses. However, although the APC model is generally accepted, our sense is it remains more of a niche methodology than an integral part of mainstream practice.

We believe two misunderstandings have slowed the uptake of the APC approach. Firstly, there are concerns about the “identifiability problem” of the APC model (27, 28). Secondly, close connections between the classical toolbox and the APC model have not been clearly spelled out in the literature. In this commentary,
we will attempt to clarify both misunderstandings and thereby make the case that the APC model merits much wider use.

**Data, Methods, and Results**

*Example: Breast cancer incidence data*

We will develop this commentary using as a concrete example the incidence of invasive female breast cancers in the United States. For this purpose, we obtained age-specific case and population data from the National Cancer Institute’s Surveillance, Epidemiology, and End Results 9 Registries Database (SEER9) for the 36-year time period from 1973 through 2008 (November 2010 submission)(35).

In general, for any given cancer and population group, the matrix \( Y = [Y_{pa}, p = 1, \ldots, P, a = 1, \ldots, A] \) contains the number of cancer diagnoses in calendar period \( p \) and age group \( a \), and the matrix \( O = [O_{pa}, p = 1, \ldots, P, a = 1, \ldots, A] \) contains the corresponding person-years. The observed incidence rates per 100,000 person-years are \( \lambda_{pa} = 10^5 Y_{pa} / O_{pa} \), and the expected log rates are \( \rho_{pa} = \log(E(Y_{pa}) / O_{pa}) \).

It is instructive to think of the rate matrix in terms of its corresponding Lexis diagram (Figure 1), which makes visually clear how the diagonals of matrices \( Y \) and \( O \), from upper right to lower left, represent successive birth cohorts indexed by \( c = p - a + A \), from the oldest \( (c = 1) \) to the youngest \( (c = C = P + A - 1) \). From this perspective, it becomes clear that a new cohort enters prospective follow-up with each consecutive calendar period. For this reason, one can think of a registry as a “cohort of cohorts.” Because cancer registries are operated in perpetuity, over
time, a substantial number of birth cohorts are followed. Our example includes $C = 24$ nominal 8-year cohorts born from 1892 through 1984 (referred to by mid-year of birth).

**The APC model: formulation**

APC analysis is based on a log-linear model for the expected rates with additive effects for age, period, and cohort:

$$p_{pa} = \alpha_a + \pi_p + \gamma_c \quad (1).$$

The generic additive effects in equation (1) can be partitioned into linear and non-linear components (28). There are number of equivalent ways to make this partition while incorporating the fundamental constraint that $c = p - a$. Two of the most useful (36) are the age-period form

$$\rho_{pa} = \mu + (\alpha_L - \gamma_L)(a - \bar{a}) + \tilde{\alpha}_a + (\pi_L + \gamma_L)(p - \bar{p}) + \tilde{\pi}_p + \tilde{\gamma}_{p-a+\Delta} \quad (2)$$

and the age-cohort form

$$\rho_{ca} = \mu + (\alpha_L + \pi_L)(a - \bar{a}) + \tilde{\alpha}_a + (\pi_L + \gamma_L)(c - \bar{c}) + \tilde{\pi}_{c+a-\Delta} + \tilde{\gamma}_c \quad (3).$$

Notation and parameters are summarized in Table 1. Importantly, all the parameters in equations (2) and (3) can be estimated from the data without imposing additional constraints, and fitted rates from both forms are identical.

There is a close correspondence between APC parameters and estimable functions in Table 1 and fundamental aspects of the data investigated using the standard descriptive toolbox. Before highlighting some of these connections below, we hopefully can shed further light on the much discussed identifiability problem.
Identifiability: “problem” or uncertainty principle?

The aspect of identifiability in question concerns whether log-linear trends in rates can uniquely be attributed to the influences of age, period, or cohort, quantified by parameters $\alpha_L, \pi_L$, and $\gamma_L$. Mathematically, it has been shown by Holford (28) that one cannot do this without imposing additional unverifiable assumptions, because the three time scales are co-linear (cohort equals period minus age, $c = p - a$). This issue has often implicitly been held out as a unique and unfortunate limitation of the APC model. In fact, the same issue affects time-to-event analysis of any cohort study.

To see this, consider the following thought experiment. Suppose one enrolls a cohort of exchangeable persons of identical age (e.g., the 1956 birth cohort in Figure 1) and follows them longitudinally over a decade for cancer. At the end of the study, one observes that the log incidence rate increases linearly with age. It is natural to attribute this trend entirely to the effects of ageing, and equate the age-associated slope to the value of a parameter $\alpha_L$.

However, suppose one had also assembled an identical cohort of persons of the same age, but this study had been conducted ten years earlier. It is possible that the age-associated slopes of the two studies would be very different, if disease-causing exposures out of experimental control had been increasing or decreasing in prevalence over time. Hence, the observed age-associated slope actually estimates parameter $(\alpha_L + \pi_L)$ or longitudinal age trend (LAT in Figure 1) (32), where $\alpha_L$ is the component of the trend that is attributable to aging and $\pi_L$ is the component of
the trend due to the net impact of unknown and uncontrollable exposures over successive calendar-periods.

A similar issue affects any cross-sectional analysis. To “control” for the effects of ageing, suppose one studied in succession over time an event rate in persons of the same age (e.g., age group 65-69 years in Figure 1), to estimate the slope of the time-trend $\pi_L$. By definition, each successive group in this cross-sectional study was born a year later. Hence, both unknown factors and factors out of experimental control associated with birth cohort could also play a role. Therefore, the observed slope over time actually estimates a parameter $(\pi_L + \gamma_L)$ or net drift in Figure 1 (29, 30), where $\pi_L$ is the component of the trend that is attributable to calendar time and $\gamma_L$ is the component of the trend attributable to the successive cohorts enrolled in the study.

These simple thought experiments, Figure 1, and Table 1 illustrate an important ‘uncertainty principle’ regarding the measurement of absolute rates in cohorts. Interestingly, this principle is seldom considered in the context of most epidemiological cohort and case-control studies, perhaps because these studies have a fairly narrow accrual window and often focus on relative rates rather than absolute rates. In contrast, this issue is often central in the analysis of registry data, because the follow-up has sufficient breadth and depth to reveal long-term secular trends in the population associated with age, period, and cohort. Indeed, a unique role of registry studies is to identify and quantify such trends, thereby providing direction and guidance regarding the needs for targeted analytical studies.
**Estimable functions: separating signal from noise**

The APC model provides a unique set of best-fitting log incidence rates, $\hat{\rho}_{pa}$ or equivalently $\hat{\rho}_{ca}$, obtained by plugging in maximum likelihood estimators into equations (2) or (3), respectively. The corresponding variances are readily calculated. In our experience the fitted rates have an appealing amount of smoothing, and we use them routinely in our studies (36-45), especially for rare cancer outcomes. Experience suggests that for “moderate” sized rate matrices (in terms of $A$ and $P$), the APC model smoothes the data conservatively, about as much as a 3-point moving average, yielding around a 40-60% reduction in the width of the confidence intervals. Of course, the precise amount of noise reduction depends on a number of technical details including whether over-dispersion is present or accounted for.

This application of the APC model is illustrated in Figure 2 for the breast cancer data. The age-standardized rates (ASRs) over time calculated using the observed rates are nearly identical to the ASRs calculated using the APC fitted rates. However, the point-wise confidence intervals for the fitted rates are substantially narrower, by around 40% averaged over the 10-year time period.

**Estimable functions: connections to the classical approaches**

The APC parameter called the net drift (Table 1 and equations (2) and (3)) estimates the same quantity as the EAPC of the ASR, i.e. the overall long-term secular trend. The point estimates for these quantities are almost identical for the breast cancer data in Figure 2; net drift = 0.83% per year (95% CI: 0.78 to 0.87).
0.85%/yr) and EAPC = 0.78% percent per year (0.18 to 1.39%/yr). However, for this example, the estimated confidence bands are much narrower for the net drift.

We introduced a novel estimable function called the fitted age-at-onset curve to summarize the longitudinal (i.e. cohort-specific) age-associated natural history (Table 1 and figure 3) (46). By construction, the fitted curve extrapolates from observed age-specific rates over the full range of birth cohorts to estimate past, current, and future rates for the referent cohort, e.g., the 1932 cohort in this example. The fitted age-at-onset curve provides a longitudinal age-specific rate curve that is adjusted for both calendar-period and birth-cohort effects. We view it as an improved version of the cross-sectional age-specific rate curve, improved because the cross-sectional curve is not adjusted for period and cohort effects (47). The fitted curve has proven very useful in practice (38-40, 42-44, 46, 48).

Finally, period deviations in the APC model (Table 1) identify changes over time; such change points are often analyzed non-parametrically using joinpoint regression methods (25). Similarly, cohort deviations can provide an explanation for joinpoint patterns in age-specific rates over time.

**APC analysis: beyond the basics**

There are many useful extensions to the basic APC model. Estimable functions are amenable to formal hypothesis tests (29, 30). Parameters associated with age, period, and cohort can be smoothed (49). Parametric assumptions about the shape of the age incidence curve derived from mathematical models of carcinogenesis can be incorporated (50). Other extensions have included parametric
(33) and nonparametric (51, 52) assessments of changes in period and cohort deviations, and simultaneous modeling of a moderate or large number of strata, such as geographic areas, using Bayes and Empirical Bayes methods (53).

Recently, we developed novel methods to compare age-related natural histories and time-trends between distinct event rates assuming that separate APC models hold for each (36). Using this approach one can formally contrast the incidence of a given tumor such as breast cancer in two populations, say Black versus White women (46), or the incidence of two tumor subtypes in the same population, say, ER positive versus ER negative breast cancers (46, supplemental Figure). We demonstrated that two event rates are proportional over age, period, or cohort if and only if certain sets of APC parameters are all equal across the respective event-specific models (36). We also developed corresponding tests of proportionality and estimators of rate ratios.

A number of authors have forecast future cancer rates using the APC model (54-58). Projections quantify the future implications of current trends, for example, the impact of a net drift of 1% versus 2% over time, or the future impact of recent changes in birth cohort patterns.

Discussion

Successful technological evolution builds on effective design. This is just as true for statistical methods as for computers and cellular phones. We have argued here that the APC model provides a useful evolutionary extension to the standard armamentarium of methods available to the descriptive epidemiologist. The APC
model is not a replacement for existing methods, which are popular and successful. Rather, it provides a refined means of estimating the same quantities, while also adding useful new capabilities, such as formal methods for comparing two sets of rates or projecting the future cancer burden.

Using the APC model, cancer registry data can be analyzed in the same spirit as any other epidemiological cohort using the same concepts, such as proportional hazards, confounding, and effect modification/interaction. Importantly, because cancer registries follow a cohort of cohorts, analysis of registry data can reveal fundamental changes in population rates that are not usually discernable in standard cohort or case-control studies.

Currently, software for APC analysis is available only through fairly specialized packages (SAS, R, Matlab). Development of good stand-alone software, in addition to education and training, are needed if the full potential of the APC model is to be exploited by descriptive epidemiologists.
Table 1. Some estimable parameters and functions in the APC model*

<table>
<thead>
<tr>
<th>Quantity*</th>
<th>Nomenclature</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>Grand mean</td>
</tr>
<tr>
<td>( \alpha_a, \pi_p, ) and ( \gamma_c )</td>
<td>Age, period, and cohort deviations</td>
</tr>
<tr>
<td>( (\alpha_L + \pi_L) )</td>
<td>Longitudinal age trend</td>
</tr>
<tr>
<td>( (\alpha_L - \gamma_L) )</td>
<td>Cross-sectional age trend</td>
</tr>
<tr>
<td>( (\pi_L + \gamma_L) )</td>
<td>Net drift ≈ EAPC of the ASR**</td>
</tr>
<tr>
<td>( \mu + (\alpha_L + \pi_L)(a - \bar{a}) + \tilde{\alpha}_a )</td>
<td>Fitted longitudinal age-at-event curve</td>
</tr>
<tr>
<td>( \mu + (\alpha_L - \gamma_L)(a - \bar{a}) + \tilde{\alpha}_a )</td>
<td>Fitted cross-sectional age-at-event curve</td>
</tr>
<tr>
<td>( \mu + (\pi_L + \gamma_L)(p - \bar{p}) + \tilde{\pi}_p )</td>
<td>Fitted temporal trends</td>
</tr>
</tbody>
</table>

*The APC model is defined over a \( P \times A \) event matrix \( Y \) and corresponding matrix of person-years \( O \). The referent age, period, and cohort are \( \bar{a} = [(A+1)/2] \), \( \bar{p} = [(P+1)/2] \), and \( \bar{c} = \bar{p} - \bar{a} + A \), respectively, where \( P \) and \( A \) are the total numbers of period and age groups and \([\cdot]\) is the greatest integer function.

**EAPC, estimated annual percentage change; ASR, direct age standardized incidence rate (ASR) (23).
Figure legends:

**Figure 1:** Rate matrix or Lexis diagram (20-22) for invasive female breast cancer. Data from the National Cancer Institute’s Surveillance, Epidemiology, and End Results 9 Registries Database (SEER 9) for cases diagnosed between 1973 through 2008 (35). Sixteen 4-year age groups (21-24, 25-29, ..., 81-84 years) and nine 4-year time periods (1973-1976, 1977-1980, ..., 2005-2008) span 24 partially overlapping 8-yr birth cohorts, with age groups in the columns, time-periods in the rows, and birth-cohorts along the diagonals. Conditioned upon time period (e.g., 1973-1977), the cross sectional age trend (CAT) cuts across increasing age groups and decreasing birth cohorts, \( \text{CAT} = (\alpha_i - \gamma_i) \). Conditioned upon birth cohort (e.g., 1956), the longitudinal age trend (LAT) cuts across increasing age groups and time periods, \( \text{LAT} = (\alpha_i + \pi_i) \). Conditioned upon age group (e.g., 65-69 years), the net drift cuts across increasing time periods and birth cohorts, \( \text{Net drift} = (\pi_i + \gamma_i) \).

**Figure 2:** Age standardized rates (ASRs, 2000 standard US population) for invasive female breast cancer. Data from the National Cancer Institute's SEER 9 Database. ASRs calculated using observed rates (grey) and age-period-cohort fitted rates (red). Point estimates (solid circles) and 95% confidence limits (shaded areas) are shown.

**Figure 3:** Cohort-specific age-specific incidence rates for invasive female breast cancer. Data from the National Cancer Institute's SEER 9 Database, stratified by 8-year birth cohorts. The age-period-cohort (APC) fitted age at onset curve (red line
and shaded 95% confidence intervals) ‘stitches together’ the experience of 11 cohorts observed over staggered age intervals. By construction, the fitted (or longitudinal) curve is centered on the referent-cohort; e.g., the 1932 birth-cohort. Because these breast cancer incidence rates are increasing over time with a net drift of 0.83% per year (figure 2 and text), the fitted rates for the referent-cohort lie below the observed rates for the younger cohorts and above the rates for the older cohorts. In other words, with rising incidence rates from older to younger generations (figure 2 and text), the referent-cohort rates on average are lower than rates for the younger cohorts and higher than rates for the older cohorts, respectively. See the text and Table 1 for additional details.
References


**APC fitted rates**

**Observed rates**
Age-Period-Cohort Models in Cancer Surveillance Research: Ready for Prime Time?

Philip S Rosenberg and William F Anderson

Cancer Epidemiol Biomarkers Prev Published OnlineFirst May 24, 2011.