TYPE Methods
PUBLISHED 09 February 2024
DOI 10.3389/fonc.2023.1332429

Advances in statistical methods for cancer surveillance research: an age-period-cohort perspective

Philip S. Rosenberg* and Adalberto Miranda-Filho

Division of Cancer Epidemiology and Genetics, Biostatistics Branch, National Cancer Institute, Bethesda, MD, United States

OPEN ACCESS

EDITED BY
Jennie L. Williams,
Stony Brook University, United States

REVIEWED BY
Robert O'Brien,
University of Oregon, United States
Benny Ren,
Stony Brook University Hospital (SBUH),
United States

*CORRESPONDENCE
Philip S. Rosenberg
rosenbep@mail.nih.gov

RECEIVED 02 November 2023
ACCEPTED 28 December 2023
PUBLISHED 09 February 2024

CITATION
Rosenberg PS and Miranda-Filho A (2024) Advances in statistical methods for cancer surveillance research: an age-period-cohort perspective.
Front. Oncol. 13:1332429.
doi: 10.3389/fonc.2023.1332429

Background: Analysis of Lexis diagrams (population-based cancer incidence and mortality rates indexed by age group and calendar period) requires specialized statistical methods. However, existing methods have limitations that can now be overcome using new approaches.

Methods: We assembled a "toolbox" of novel methods to identify trends and patterns by age group, calendar period, and birth cohort. We evaluated operating characteristics across 152 cancer incidence Lexis diagrams compiled from United States (US) Surveillance, Epidemiology and End Results Program data for 21 leading cancers in men and women in four race and ethnicity groups (the "cancer incidence panel").

Results: Nonparametric singular values adaptive kernel filtration (SIFT) decreased the estimated root mean squared error by 90% across the cancer incidence panel. A novel method for semi-parametric age-period-cohort analysis (SAGE) provided optimally smoothed estimates of age-period-cohort (APC) estimable functions and stabilized estimates of lack-of-fit (LOF). SAGE identified statistically significant birth cohort effects across the entire cancer panel; LOF had little impact. As illustrated for colon cancer, newly developed methods for comparative age-period-cohort analysis can elucidate cancer heterogeneity that would otherwise be difficult or impossible to discern using standard methods.

Conclusions: Cancer surveillance researchers can now identify fine-scale temporal signals with unprecedented accuracy and elucidate cancer heterogeneity with unprecedented specificity. Birth cohort effects are ubiquitous modulators of cancer incidence in the US. The novel methods described here can advance cancer surveillance research.

KEYWORDS

cancer surveillance research, Lexis diagram, statistical methods, age-period-cohort model, SEER program

1 Introduction

Cancer Surveillance Research (CSR) (1) is an observational science of cancer occurrences ascertained in population-based cohorts, notably, cancer registries. CSR is dedicated to tracking cancer incidence and mortality; quantifying cancer differences; characterizing cancer's natural history and its evolution over time; uncovering etiologic clues; gauging effectiveness of screening and therapy; and informing cancer control programs.

To date, most CSR studies have relied on specialized nonparametric statistical tools that are effective and popular (2, 3). The parametric age-period-cohort (APC) model provides a complementary approach (4–7). Even so, large scale studies covering many populations or outcomes (8, 9) are labor intensive and demand technical expertise, thereby pushing the boundaries of feasibility.

Advances in biostatistics and data science have the potential to usher a 'golden age' where high-quality data are universally accessible, and contemporary methods from biostatistics and data science are rapidly and freely deployable. To contribute to this vision, we survey a "toolbox" of newly developed biostatistical methods for analyzing population-based cancer incidence and mortality data. The unique focus of this toolbox is its age-period-cohort perspective.

This is an opportune time to propose such an upgrade. In the United States (US), the cancer landscape has evolved over the last half-century as the US population grew, aged, and changed (10). Throughout this period, the Surveillance, Epidemiology, and End Results (SEER) Program accumulated authoritative population-based data on cancer outcomes (11). Globally, cancer is rapidly rising in many countries (9). Fortunately, the number of high-quality population-based cancer registries has also increased over time (12, 13).

In Section 2, we assemble a panel of examples and illustrate limitations and pitfalls of traditional methods. In Section 3, we present promising new methods that complement the traditional approaches. In Section 4, we provide a summary and outline avenues for future research.

The new methods leverage four core principles. First, the Lexis diagram (14) is a fundamental construct that provides a unifying schema for the data. Second, nonparametric smoothing techniques for the Lexis diagram (15, 16) enhance our ability to quantify trends. Third, no analysis of a Lexis diagram is complete without considering the effects of birth cohort: This is most easily accomplished using APC models (6, 17, 18). Fourth, newly developed methods for comparative analysis (19–25) can elucidate heterogeneity between Lexis diagrams ascertained within strata defined by factors such as sex, race and ethnicity, geographic region, and tumor characteristics. We present an overview of these approaches in Figure 1.

2 Materials and methods

2.1 Lexis diagrams

The Lexis diagram (18) is a rectangular grid with binned age groups along one axis and binned calendar periods along the other. Individuals from a surveilled population contribute person-years (number of people and the amount of time at risk) and events (incident cancers, or deaths by cause) to each cell. The observed event counts are modeled as independent Poisson random variables with or without overdispersion. Cells along the diagonals represent persons born in the same period (birth cohorts). Lexis diagrams can be obtained from hundreds of population-based cancer registries worldwide, from the Surveillance, Epidemiology, and End Results (SEER) Program (26), the North American Association of Central Cancer Registries (NAACCR (13)), and the International Agency for Research on Cancer (Cancer Incidence in Five Continents, CI5 (27)).

Using SEER's Thirteen Registries Database (28), we constructed a panel of 152 cancer incidence Lexis diagrams covering 50 single-years of age (ages 35 – 84), 27 calendar years (1992 – 2018), and 76 single-year birth cohorts (1908 – 1983) for 21 leading cancers in women and men within four race and ethnicity categories: non-Hispanic White (NHW), non-Hispanic Black (NHB), Hispanic (HIS), and Asian and Pacific Islander (API). The 21 cancer sites are: esophagus, stomach, gallbladder, liver, pancreas, colon, rectum, kidney, bladder, leukemia, non-Hodgkin Lymphoma (NHL), myeloma, brain, thyroid, lung, melanoma, breast, ovary, corpus, cervix, and prostate.

2.2 Classic methods

Lexis diagrams are analyzed using four classic methods: canonical plots for visualization of age-specific rates (29, 30); age-standardized rates (31) (ASRs) for dimension reduction; estimated annual percentage change (EAPC) of the ASRs for trend estimation (32); and JoinPoint analysis for gradient estimation (33, 34), e.g., to identify changes in the EAPC of the ASR over time. These popular statistical tools have limits that warrant attention, summarized in Figure 2.

2.2.1 To lump or to split?

The most granular possible Lexis diagrams obtainable from population-based cancer registries encapsulate the rates for single-years of age within single calendar years (1x1s). If the data are sparse, we can bin the 1x1s to 2x2s or 5x5s. The CI5 database (12) provides five-by-ones (5x1s): five-year age groups within single calendar years. The novel methods described in this report require *equal* bin widths for age and period. So, for 5x1s, we must bin the single calendar years into five-year periods or interpolate to single years of age from the age quinquennium within each calendar year. While feasible, interpolation can introduce bias, complicating the interpretation of the results.

Hence, we face a choice: We can analyze 5x5s, 2x2s, or 1x1s. Going one way or the other makes an implicit bias-variance trade-off. Opting to lump may introduce bias, but the granular data are noisy.

2.2.2 ASRs and EAPCs: more than one

There are four widely recognized standard populations (e.g., US 2000 Census, Canadian, WHO World 2000, and Uniform), and

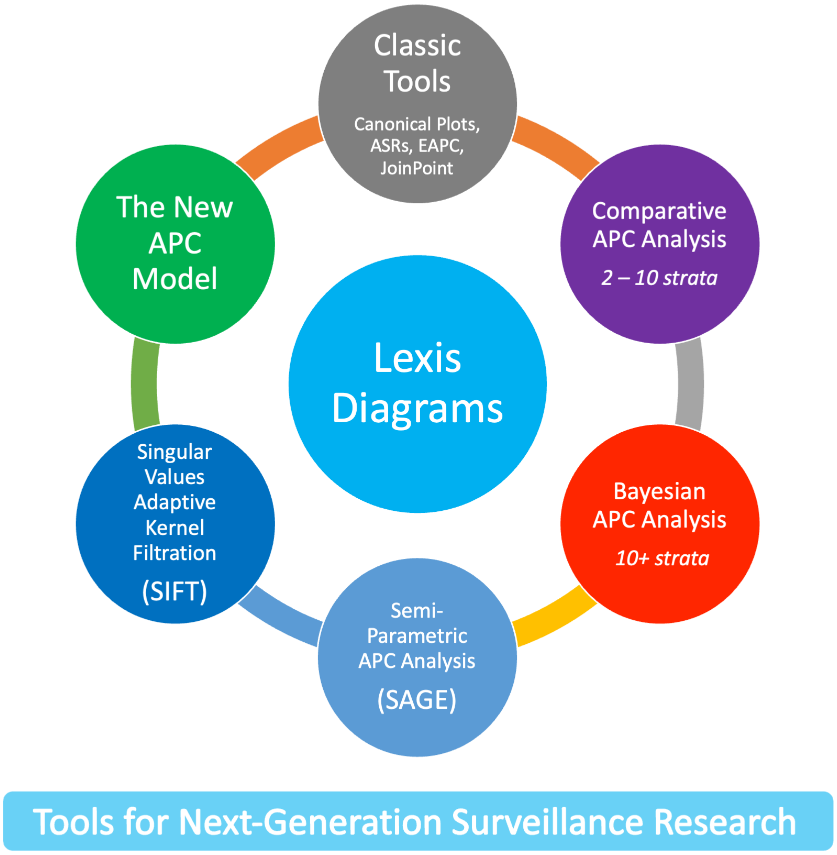

Classic Tools
Canonical Plots, ASRs, EAPC, JoinPoint
Comparative APC Analysis
2 – 10 strata
Bayesian APC Analysis
10+ strata
Semi-Parametric APC Analysis
(SAGE)
Singular Values Adaptive Kernel Filtration
(SIFT)
The New APC Model
Lexis Diagrams
Tools for Next-Generation Surveillance Research

FIGURE 1
New Tools for Next-Generation Surveillance Research.

four well-posed estimators of trend (32). Essentially all studies select only one of these 16 possibilities. Are conclusions sensitive to this choice?

Figure 3 calculates 16 estimators of EAPC for colon cancer incidence in NHW, NHB, API and HIS women and men. The estimates in each stratum are heterogeneous (Panels A – H), and the EAPC spread – the range between the left- and right-facing triangles – ranges from 1.5 to 2 percent across the panels. Similar heterogeneity is seen across the Cancer Incidence Panel (Figure 4): the EAPC spread varies by around 2% on average in both females (Panel A) and males (Panel B). This amount of heterogeneity is substantial, given that EAPCs and EAPC differences in excess of ± 0.5% are generally considered notable. One appeal of the APC Net Drift parameter described in Section 2.3.1 is there is only one.

2.2.3 The problem with JoinPoint is scalability

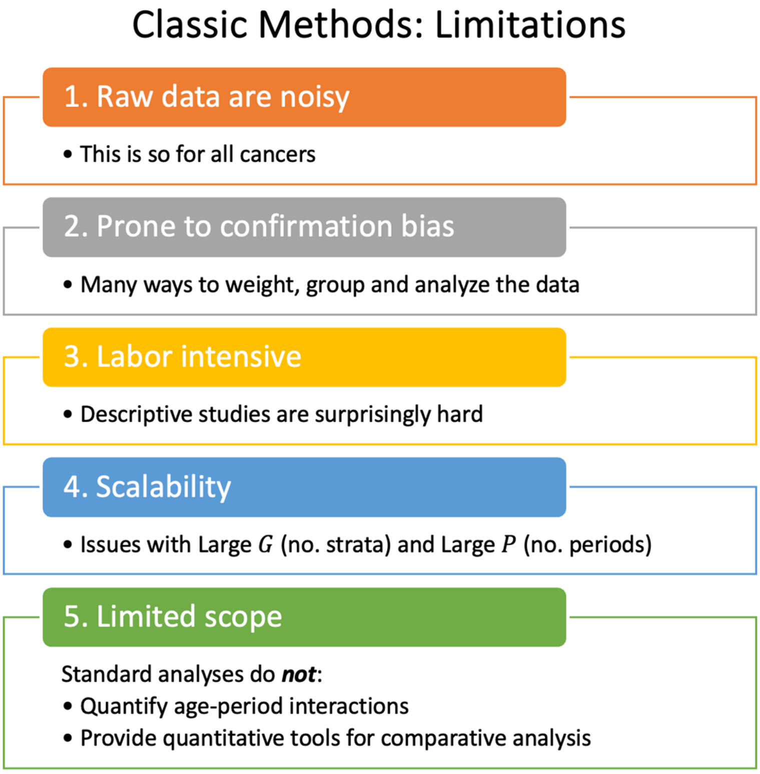
JoinPoint is a signature method of CSR (33). Whereas the EAPC estimates the *average* rate of change over time, JoinPoint estimates the *gradient*, i.e., the *instantaneous* rate of change. Typically, JoinPoint is applied to age-standardized or age-group-specific (a.k.a. truncated) rates over time (35). JoinPoint can also be used in conjunction with APC models, for example, to identify changes in birth cohort effects. In principle, JoinPoint can be applied to *any* series of n observations y_i at time point t_i, $i = 1, \ldots, n$, with a full-rank variance-covariance matrix Σ. In the context of CSR, the time series are *equally spaced.*

JoinPoint fits a piecewise linear spline to the data, where the number and locations of the knots, or join-points are estimated from the data. The corresponding gradient curve, a step function, obtains from the *slopes* of the fitted linear spline. To fit a JoinPoint model, we must specify 4 constraints: 1) the minimum number of segments k_{min}, the maximum number of segments k_{max}, the minimum number of time points per segment a, and the maximum number of time points per segment b.

For knot locations restricted to t_i, $i = 1, \ldots, n$, the set of possible JoinPoint models corresponds to the set of doubly restricted integer combinations of n, $RIC(n, k_{min}, k_{max}, a, b)$ (36). Efficient formulas and code exist for enumerating *RICs* (37). As n increases, it becomes increasingly difficult to fit the model without imposing strong restrictions on k_{max} and a, because the *RIC* numbers become too big.

Suppose we wish to fit a JoinPoint model for 76 single-year birth cohorts, e.g., 1908 – 1983, as in Section 2.1. To fit up to 5 segments each with 10 or more cohorts, JoinPoint must evaluate $RIC(76, 1, 5, 10, 76) = 37,730$ models. To allow for up to 10 segments each with 5 or more cohorts – an interesting and plausible scenario – the number is 177,817,540, which is not feasible.

JoinPoint was designed to analyze ASRs for epochs up to several decades long. For this purpose, JoinPoint provides a popular and



Classic Methods: Limitations

1. Raw data are noisy
- This is so for all cancers

2. Prone to confirmation bias
- Many ways to weight, group and analyze the data

3. Labor intensive
- Descriptive studies are surprisingly hard

4. Scalability
- Issues with Large G (no. strata) and Large P (no. periods)

5. Limited scope

Standard analyses do ***not***:
- Quantify age-period interactions
- Provide quantitative tools for comparative analysis

FIGURE 2
Classic Methods: Limitations.

enduring standard that has recently been improved (34). For applications to longer time series, for example, daily COVID counts, the scalability issue abrogates its appeal as a flexible and adaptive nonparametric estimator of gradients. Fortunately, recent work in this area using stochastic optimization is promising (38).

2.2.4 Standard methods are not designed to detect interactions

Tailored statistical approaches to identify age-period interactions are limited. One exception is the method of Kim et al. (39) for comparing two JoinPoint models.

Savvy epidemiologists have discovered several notable age-period interactions using classic methods alone (40–43). New methods could accelerate the pace of discovery.

2.3 The age-period-cohort model

The APC model is a standard in the field. Fundamentally, it expands the scope of inference. Using the APC model, we can quantify age-period interactions and characterize the longitudinal experience of birth cohorts. Even so, its use in cancer incidence studies has been relatively limited compared to studies that use classic descriptive methods alone, despite freely available software (6, 18). Why is this so? There are several concerns, summarized in Figure 5.

Perhaps the biggest are:

1. What about the “identifiability problem”?
2. When is the model appropriate?
3. How can you determine whether the model’s fit is adequate?

2.3.1 Identifiability

The statistical identifiability problem arises because an individual’s year of birth can be determined by subtracting their attained age from the current calendar year. This relationship has an important consequence: when we model event rates in a population, it is impossible to separate the log-linear trend associated with the year of birth, the parameter γ_L, from the log-linear trend associated with calendar year, the parameter π_L. We can, however, estimate their sum, $(\pi_L + \gamma_L)$, which is called the Net Drift. In our view, the impossibility of estimating the constituents π_L and γ_L in $(\pi_L + \gamma_L)$ reflects an intrinsic limitation of observational epidemiologic cohort studies (44). Similarly, the identifiable cross-sectional age trend is $(\alpha_L + \pi_L)$ not α_L, and the identifiable longitudinal age trend is $(\alpha_L - \gamma_L)$ not α_L.

Estimable Functions (EFs) are linear combinations of model parameters that are invariant with respect to the particular identifiability constraints imposed on the parameters to fit the model (4, 5, 45).

Despite the identifiability problem, the New APC Model (7) provides an expansive array of informative EF based on the intercept μ, the identifiable trend parameters $(\alpha_L - \gamma_L)$, $(\alpha_L + \pi_L)$

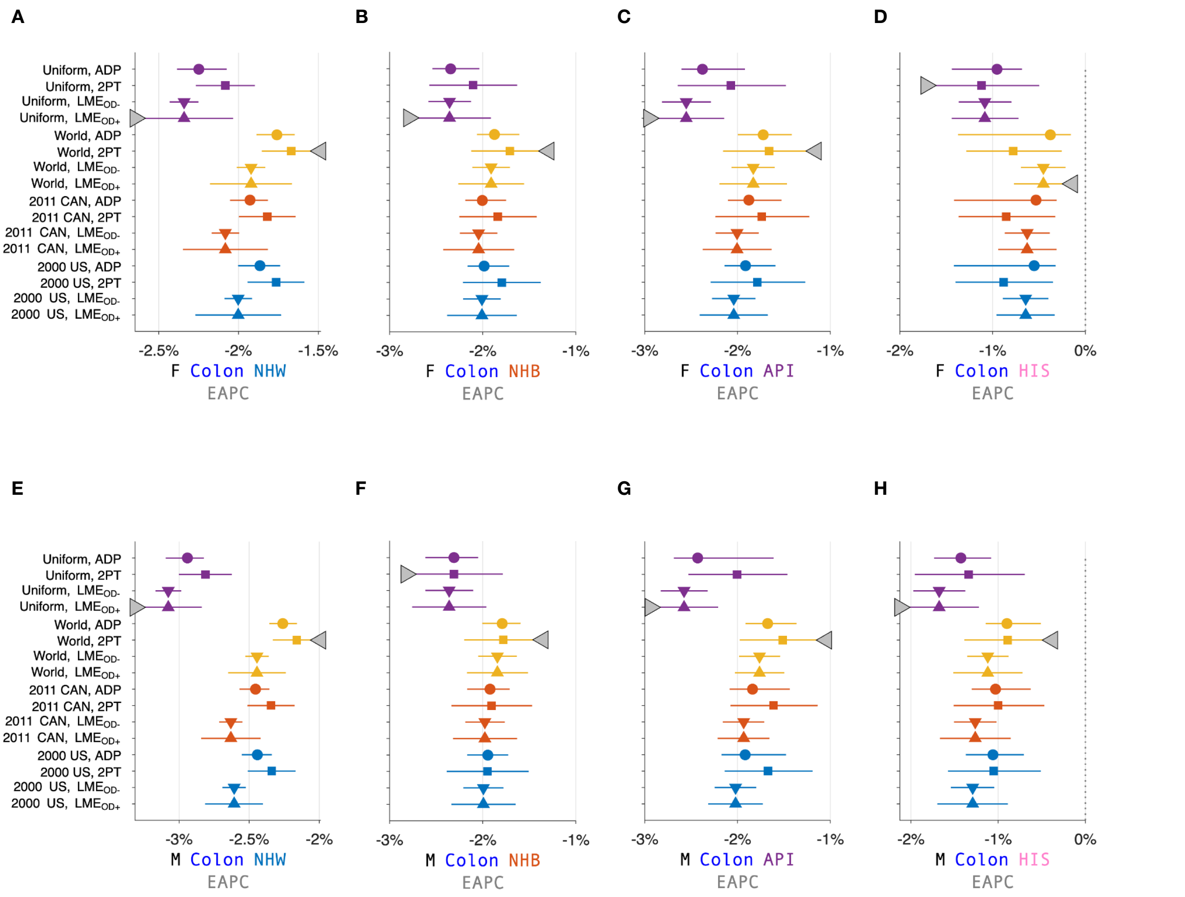


FIGURE 3
Estimated Annual Percentage Change (EAPC) for Colon Cancer. EAPC for colon cancer stratified by sex and race and ethnicity in Panels **(A–H)**. Each panel presents a Forest Plot of 16 estimators based on four standard populations (Uniform, World, 2011 CAN, and 2000 US) and four trend estimators: Adaptive (ADP); Two-Point (2PT); Linear Model Estimator (LME) without overdispersion (LME_{OD-}); and LME with overdispersion (LME_{OD-}). Females (F): **(A–D)**. Males (M): **(E–H)**. Non-Hispanic White (NHW) **(A, E)**, non-Hispanic Black (NHB) **(B, F)**, Asian and Pacific Islander (API) **(C, G)**, and Hispanic **(D, H)**. Lowest lower limits and highest upper limits are highlighted by left- and right-facing triangles, respectively.

and $(\pi_L + \gamma_L)$, the global curvature parameters for age, period and cohort, θ_α, θ_π, and θ_γ, respectively, and the corresponding higher order deviations $\breve{\gamma}_{a*}$, $\breve{\gamma}_{p*}$, and $\breve{\gamma}_{c*}$. Local Drifts (model-based estimates of the age-specific trends over time) are especially valuable. Please refer to Sections 3 and 5 of the introductory paper for a summary of the parameters, and Table 1 for a summary of essential EF (7).

2.3.2 When is the model appropriate?

"All models are wrong, some are useful" (46). In our context, lack-of-fit (LOF) implies that some birth cohort effects vary over time and age, for example, one generation has higher risk than another for early onset of a cancer, but lower risk for late onset.

In principle, the APC model is well suited for cancer *incidence* if one accepts "the primacy of birth cohort effects." This concept asserts that: 1) Most cancers (47) have exogenous risk factors (or endogenous risk factors modulated by environmental exposures) and long latency periods from initiation to promotion and progression (48); 2) Exposures in a population typically wax and wane over time. 3) The interplay between biology and tumor natural history induces risk heterogeneity across generations. From this perspective, the APC model is a natural choice for modeling cancer incidence because estimable birth cohort effects quantify net changes in incidence from one birth cohort to the next.

2.3.3 Current methods to assess lack-of-fit are limited

Current methods to assess LOF include estimating overdispersion parameters, comparing observed and fitted values, and examining residuals (49). In those cases where the LOF is notable, one remedy is to split the rate matrix into blocks within which the LOF is nominal. See the supplement to Best et al. (49) for details. These methods are labor intensive and may not be sensitive, especially for cancers with relatively few events.

3 Results: tools for next-generation surveillance research

Recent advances overviewed in Figure 1 mitigate the limits and concerns summarized in Figures 2, 5. In brief: The SIFT method

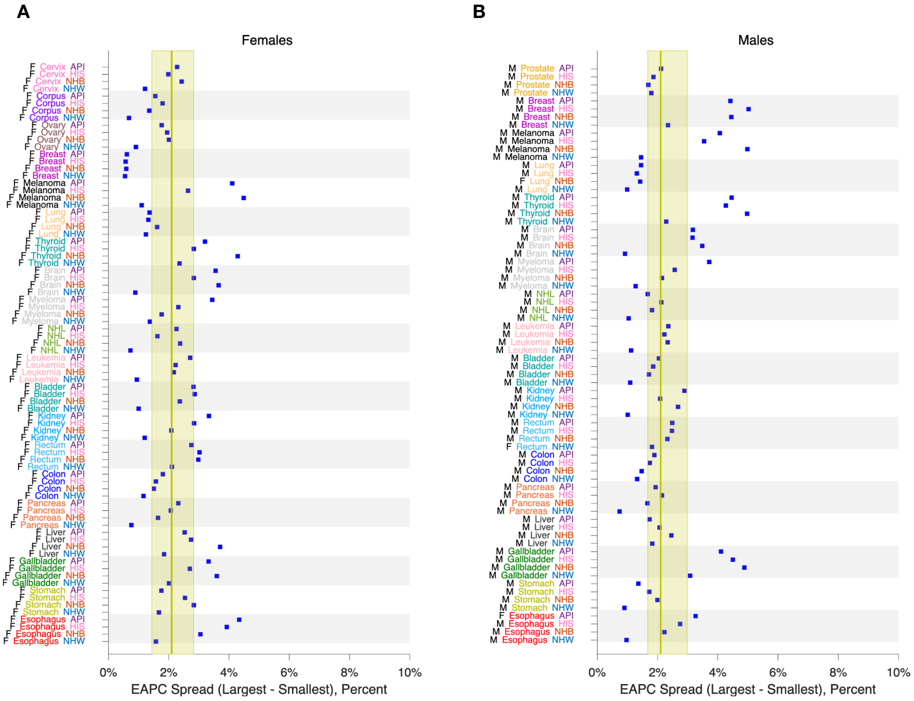
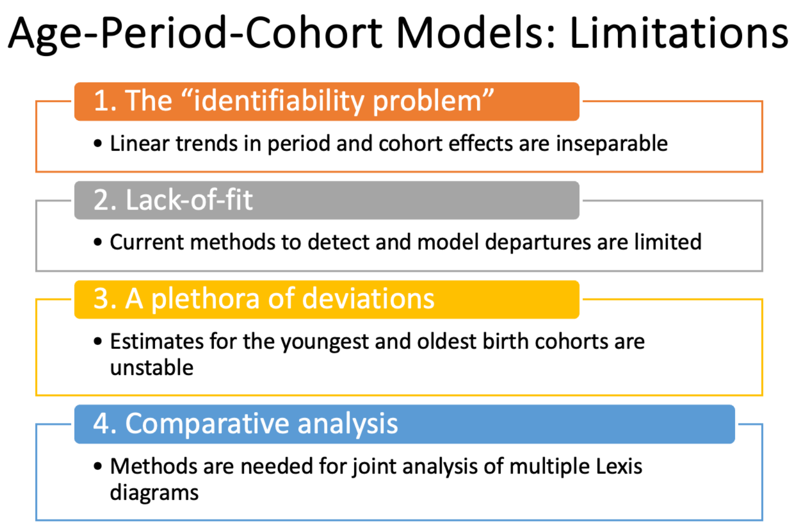


FIGURE 4
EAPC Spread in 152 Cancers. EAPC spread: Range between lowest lower limit and highest upper limit over 16 EAPC estimators as described in the legend to Figure 3. Panel **(A)**, Females. Panel **(B)**, Males. Values by race and ethnicity (API, HIS, NHB, NHW) within cancer sites as labelled. Yellow reference intervals show median and inter-quartile range for 80 Lexis diagrams in females **(A)** and 72 in males **(B)**. Cancer sites: Cervix, Corpus, Ovary, Breast, Melanoma, Lung, Thyroid, Brain, Myeloma, NHL, Leukemia, Bladder, Kidney, Rectum, Colon, Pancreas, Liver, Gallbladder, Stomach, Esophagus, Prostate.

(Section 3.1 – 3.2) mitigates the limits noted in Figure 2. The New APC Model (Section 3.3) eases concerns about identifiability (Figure 5.1); SAGE (Section 3.4) addresses worry about lack-of-fit (Figure 5.2) and instability (Figure 5.3); and sophisticated methods are now available for comparative analysis (Figure 5.4).

3.1 Sifting through the data

Cancer rates are intrinsically "noisy" (31), and this random variation can mask important signals. The newly developed SIFT (singular values adaptive kernel filtration) method produces



Age-Period-Cohort Models: Limitations

1. The "identifiability problem"
- Linear trends in period and cohort effects are inseparable

2. Lack-of-fit
- Current methods to detect and model departures are limited

3. A plethora of deviations
- Estimates for the youngest and oldest birth cohorts are unstable

4. Comparative analysis
- Methods are needed for joint analysis of multiple Lexis diagrams

FIGURE 5
Age-Period-Cohort Models: Limitations.

smoothed Lexis diagrams with an optimal bias-variance trade-off (50). SIFT incorporates two key innovations. First, for any candidate kernel function, SIFT discards superfluous "high-frequency" basis vectors from the corresponding smoothing matrix based on the bias-corrected Akaike information criterion. Second, because the optimal kernel for any given rate matrix is unknown, SIFT estimates the optimal kernel by model averaging over a panel of candidate kernels with diverse shapes and bandwidths.

SIFT has excellent performance for 1x1 and 2x2 rate matrices (50). Sifted Lexis diagrams are *much* more accurate on a cell-by-cell basis. How much better is it to analyze sifted data versus raw data? We can answer this question more definitively using the Cancer Incidence Panel described in Section 2.1.

For any given Lexis diagram, denote the expected rate per 100,000 person-years in age group a during calendar period p as $E(\lambda_{ap}) = 10^5 \times \frac{E(y_{ap})}{PY_{ap}}$, where y_{ap} is the observed number of events and PY_{ap} is the corresponding person-years. For the raw data, the Poisson signal-to-noise ratio is $SNR_{Raw} = \frac{E(\lambda_{ap})}{Var(\lambda_{ap})^{1/2}} = E(y_{ap})^{1/2}$. Hence, the noise-to-signal percent or relative error is $NSP_{Raw} = 100 \times SNR_{Raw}^{-1}$ %.

From the same data, SIFT produces smoothed rates $\lambda_{SIFT}(a,p)$ and corresponding variances $\hat{v}^{\lambda}_{SIFT}(a,p)$. Hence, the median estimated NSP for the sifted data is $NSP_{SIFT} = 100 \times \underset{\{\text{cells } (a,p)\}}{\text{median}} \frac{[\hat{v}^{\lambda}_{SIFT}(a,p)]^{1/2}}{\lambda_{SIFT}(a,p)}$ %.

Figure 6 compares NSPs for raw 1x1 data (solid red line) versus sifted data (females, light blue circles; males, magenta squares) for all 152 Lexis diagrams in the Cancer Incidence Panel. NSPs are plotted versus the Lexis diagram's mean number of events per cell on a log-log scale. For typical 1x1 Lexis diagrams with around 5 events per cell, the NSP is ~50% for the raw data versus ~5% for the sifted data – a 90% reduction. As indicated by the least squares lines for Females (light blue line) and males (magenta line), substantial reductions are expected regardless of the mean number of events per cell. On average, The NSP was reduced by 86% across the panel.

Suppose we eschew sifting, and instead chunk the data from 1x1s to 5x5s. Chunking will indeed reduce the NSP – by 80% - almost as much as SIFT. To see this, compare the red reference line when the mean number of events is 1 versus 25, or 5 versus 125, etc. Unfortunately, we lose temporal resolution. By aggregating 25 cells into one, we throw away four-fifths of our information about age and period effects (one 5-year time point versus five 1-year time points), and eight-ninths of our information about birth cohort effects (1 diagonal in a 1x1 cell versus 9 diagonals in a 5x5 cell). Fortunately, as demonstrated in Figure 6, there is no need to do so.

For 5x1 data, rather than chunking up to 5x5s, one might consider interpolating down to 1x1s (51). In our view, this approach merits development: At this point, the optimal interpolation scheme remains unclear.

3.2 An abundance of features

A *Feature* is a linear or log-linear combination of the rates. The class of features includes averages, gradients, and trends, in any combination (50). ASRs and EAPCs are features, as are the curves graphed in canonical plots. Features can be calculated from observed data or sifted data. A key point is, Features calculated from sifted data are much more accurate. Furthermore, one way to overcome the scalability issue of the JoinPoint approach (Section 2.2.3) is to extract empirical gradients from the sifted data.

Features can also be calculated from fitted rates obtained via APC models. The essential distinction between Features and EFs is, Features describe expected values of *observed* rates, whereas EFs are estimated from model parameters and therefore describe expected values of *adjusted* rates.

3.3 Best practices for APC analysis

Despite the limitations noted in Section 2.3, the APC model greatly expands the scope of inference. When birth cohort effects are present time trends *necessarily* vary by age (7). Since many, perhaps most cancers are influenced by birth cohort effects (Sections 2.3.2 and 3.4), this implies that ASRs and ASR features (EAPC, JoinPoint) at best describe the average trend, which may not provide a reasonable summary of the trends within any given age group. In our view, one should *always* examine either Local Drifts (an EF) or age-specific temporal trends (a Feature; Section 3.2).

When the effects of LOF are modest, one can emphasize conclusions based on EFs, including Local Drifts. Indeed, under the model, Local Drifts are a *consequence* of changes in the gradient of the Fitted Cohort Pattern (FCP; the rate at arbitrary reference age a_0 in each birth cohort). Hence, the latter provide an *explanation* for the former.

In 1987, Clayton and Schifflers (52) presented a popular "checklist" for fitting classic APC models. In Figure 7 we present a checklist for interpreting model outputs from the New APC Model. A key distinction is our checklist puts Local Drifts front and center.

3.4 Semi-parametric age-period-cohort analysis

We have in hand two powerful and complementary approaches – the New APC Model and SIFT – parametric and nonparametric. Can we combine them?

One natural way to do so is to de-noise the raw data using SIFT, and then partition the sifted values into a component arising from the APC model plus a residual component that represents the LOF. We will call this procedure SAGE, an acronym for Semi-Parametric Age-Period-Cohort Analysis. Algorithm 1 presents the details.

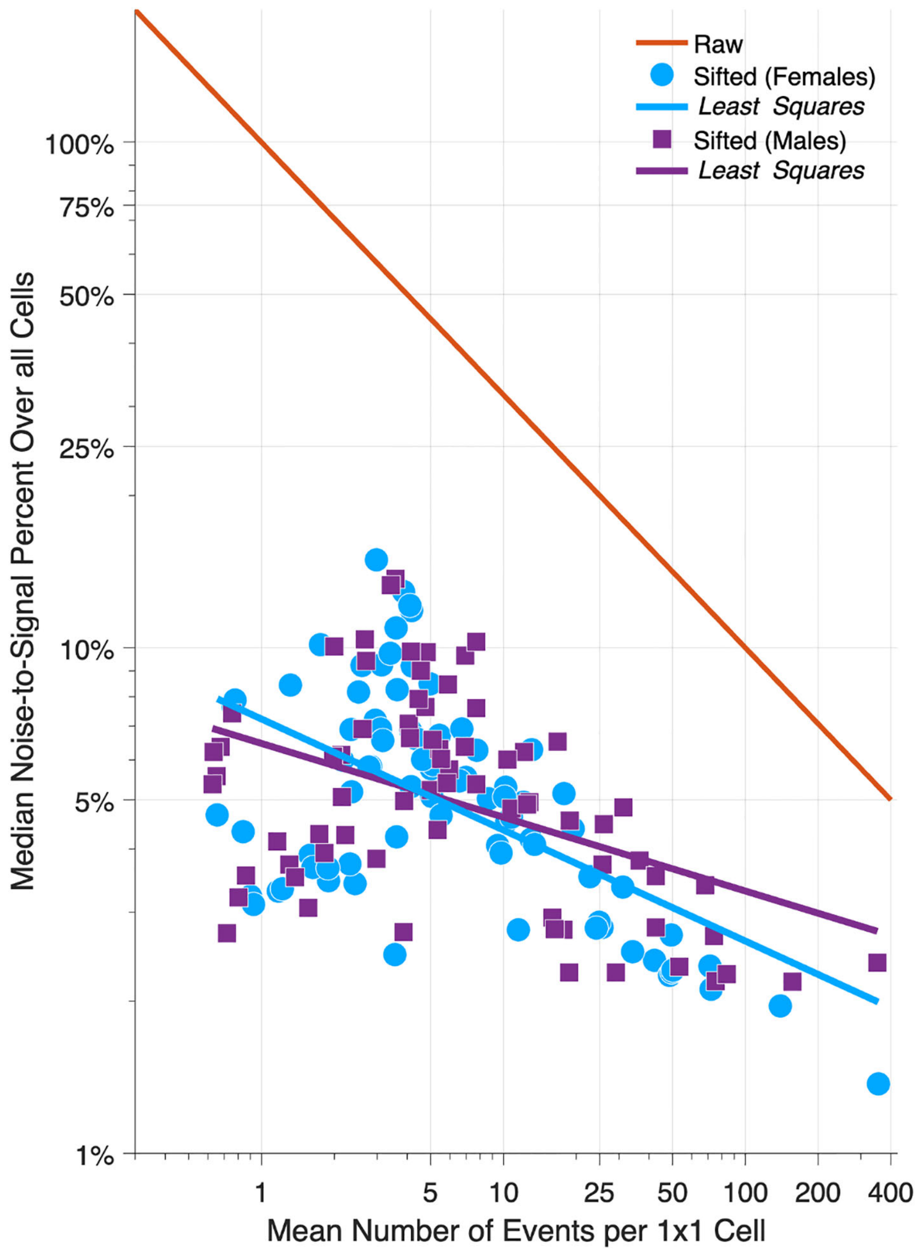


FIGURE 6
Relative Error in 152 Cancers: Raw versus Sifted Data. Scatter Plot of median noise-to-signal percent versus mean number of events per cell for 80 Lexis diagrams in females (light blue circles) and 72 in males (magenta squares). Light blue and magenta dashed lines: least squares fit. Solid red line: reference curve for raw data assuming Poisson error. Data are plotted on a log-log scale.

From the raw data y with variance-covariance matrix V^y:

1. Calculate $v = \log y$ and variance-covariance matrix $V^v = V^y \circ (yy')^{\circ -1}$

2. SIFT the log-transformed data to obtain v_{SIFT} and $\hat{V}^v_{SIFT}$

3. Construct an APC design matrix $\mathcal{X}$ incorporating identifiability constraints

Partition $\mathcal{X}$ into two sets of columns, $\mathcal{X} = [\mathcal{X}_\Theta \vdots \mathcal{X}_{(\breve{\alpha},\breve{\pi},\breve{\gamma})}]$.

- $\mathcal{X}_\Theta$ for the intercept and the linear and quadratic terms

$$\Theta \equiv (\mu, (\alpha_L - \gamma_L), (\pi_L + \gamma_L),\ \theta_\alpha, \theta_\pi, \theta_\gamma)$$

- $\mathcal{X}_{(\breve{\alpha},\breve{\pi},\breve{\gamma})}$ for the higher order deviations $\breve{\gamma}_{a_*}$, $\breve{\gamma}_{p_*}$, and $\breve{\gamma}_{c_*}$

4. Evaluate regression matrix $\mathcal{R} = (\mathcal{X}'\mathcal{X})^{-1}[\mathcal{X}_\Theta \vdots \mathcal{X}_{(\breve{\alpha},\breve{\pi},\breve{\gamma})}]'$

- Calculate $\hat{\beta}_{SIFT} = \mathcal{R}v_{SIFT} = (\Theta, \breve{\alpha}, \breve{\pi}, \breve{\gamma})'$ and $\hat{V}^\beta_{SIFT} = \mathcal{R}\hat{V}^v_{SIFT}\mathcal{R}'$
- From these outputs calculate *Estimable Functions*

5. From the Hat matrix $\mathcal{H} = \mathcal{X}\mathcal{R}$ calculate

- Fitted values $\hat{\eta}_{SIFT} = \mathcal{H}v_{SIFT} = \hat{\eta}^\Theta_{SIFT} + \hat{\eta}^{(\breve{\alpha},\breve{\pi},\breve{\gamma})}_{SIFT}$ with $\hat{V}^\eta_{SIFT} = \mathcal{H}\hat{V}^v_{SIFT}\mathcal{H}'$
- Lack-of-Fit $LOF_{SIFT} = (I - \mathcal{H})v_{SIFT}$ with $\hat{V}^{LOF}_{SIFT} = (I - \mathcal{H})\hat{V}^v_{SIFT}(I - \mathcal{H})'$

```
    • Note that $v_{SIFT} \equiv \hat{\eta}_{SIFT} + LOF_{SIFT}$
    • From $\hat{\eta}_{SIFT}$ and $v_{SIFT}$ calculate $Features$ of interest
```

Algorithm 1. Semi-Parametric Age-Period-Cohort Analysis (SAGE).

SAGE advances our understanding of the data in two ways. First, we are better able to examine LOF and gauge its impact on Features. Second, when the model appears adequate, we can draw conclusions from the EF. These estimates will be smoother and have narrower confidence limits than corresponding estimates obtained by fitting the APC model to the raw data.

To illustrate, we applied SAGE to colon cancer incidence among NHW women and visualized the outputs using heat maps (Figure 8, Panel A). Panel A.1 shows the raw data, A.6 the sifted values ("SAGE"), and A.7 the SIFT residuals ("Pure Error"). Panels A.2 – A5 present the partitions described in Step 5 of the SAGE algorithm. The full APC model (A.4) is the sum of contributions from the key parameters (A.2) and higher-order deviations (A.3).

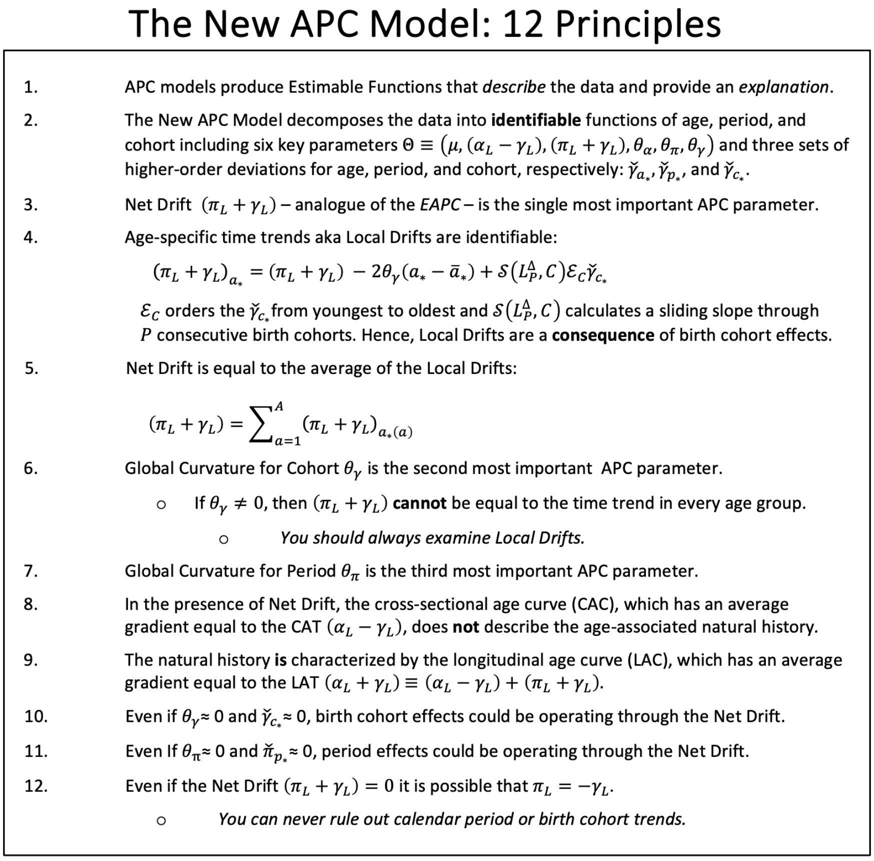
Panel B plots higher-order deviations and LOF using surface plots to better gauge their relative magnitudes. The former is substantially larger. Panel C shows the estimated period trend by age from the APC model (solid blue) and the APC model plus LOF (dash red). There are small gaps between the curves, especially at around age 50, when the model appears to under-estimate the empirical trend by around 0.5% per year. The median absolute deviation (MAD) between the parametric and nonparametric curves is 0.17% per year.

We applied SAGE to all Lexis diagrams in the Cancer Incidence Panel. Comparing the period trends by age (APC model versus APC Model plus LOF), the MAD never exceeded 0.6% per year in Females (Figure 9A) or 0.8% per year in Males (Figure 9B). On average, the MAD was 0.16% per year in Females and 0.19% per year in Males.

We also fitted JoinPoint models to the FCPs from SAGE, allowing up to 5 segments each with 10 or more birth cohorts. Figure 10 presents the number of segments identified by the JoinPoint permutation test (33) in Females (Figure 10A) and Males (Figure 10B). In every case, the number of segments was 3 or more.

3.5 Comparative age-period-cohort analysis

The APC model describes a single Lexis diagram. Most studies involve ensembles of $G \geq 2$ Lexis diagrams defined by strata such as sex, race and ethnicity, geographic region, tumor characteristics, etc. A key goal is to compare and contrast EF *between* strata. One can think of the strata as covariates.



The New APC Model: 12 Principles

1. APC models produce Estimable Functions that *describe* the data and provide an *explanation*.
2. The New APC Model decomposes the data into **identifiable** functions of age, period, and cohort including six key parameters $\Theta \equiv \left(\mu, (\alpha_L - \gamma_L), (\pi_L + \gamma_L), \theta_\alpha, \theta_\pi, \theta_\gamma\right)$ and three sets of higher-order deviations for age, period, and cohort, respectively: $\check{\gamma}_{a_*}, \check{\gamma}_{p_*}$, and $\check{\gamma}_{c_*}$.
3. Net Drift $(\pi_L + \gamma_L)$ – analogue of the *EAPC* – is the single most important APC parameter.
4. Age-specific time trends aka Local Drifts are identifiable:

$$(\pi_L + \gamma_L)_{a_*} = (\pi_L + \gamma_L) - 2\theta_\gamma(a_* - \bar{a}_*) + \mathcal{S}\left(L_P^\Delta, C\right)\mathcal{E}_C\check{\gamma}_{c_*}$$

$\mathcal{E}_C$ orders the $\check{\gamma}_{c_*}$ from youngest to oldest and $\mathcal{S}\left(L_P^\Delta, C\right)$ calculates a sliding slope through P consecutive birth cohorts. Hence, Local Drifts are a **consequence** of birth cohort effects.

5. Net Drift is equal to the average of the Local Drifts:

$$(\pi_L + \gamma_L) = \sum_{a=1}^{A} (\pi_L + \gamma_L)_{a_*(a)}$$

6. Global Curvature for Cohort θ_γ is the second most important APC parameter.
   - If $\theta_\gamma \neq 0$, then $(\pi_L + \gamma_L)$ **cannot** be equal to the time trend in every age group.
     - *You should always examine Local Drifts.*
7. Global Curvature for Period θ_π is the third most important APC parameter.
8. In the presence of Net Drift, the cross-sectional age curve (CAC), which has an average gradient equal to the CAT $(\alpha_L - \gamma_L)$, does **not** describe the age-associated natural history.
9. The natural history **is** characterized by the longitudinal age curve (LAC), which has an average gradient equal to the LAT $(\alpha_L + \gamma_L) \equiv (\alpha_L - \gamma_L) + (\pi_L + \gamma_L)$.
10. Even if $\theta_\gamma \approx 0$ and $\check{\gamma}_{c_*} \approx 0$, birth cohort effects could be operating through the Net Drift.
11. Even If $\theta_\pi \approx 0$ and $\check{\pi}_{p_*} \approx 0$, period effects could be operating through the Net Drift.
12. Even if the Net Drift $(\pi_L + \gamma_L) = 0$ it is possible that $\pi_L = -\gamma_L$.
    - *You can never rule out calendar period or birth cohort trends.*

FIGURE 7
The New Age-Period-Cohort Model: 12 Principles.

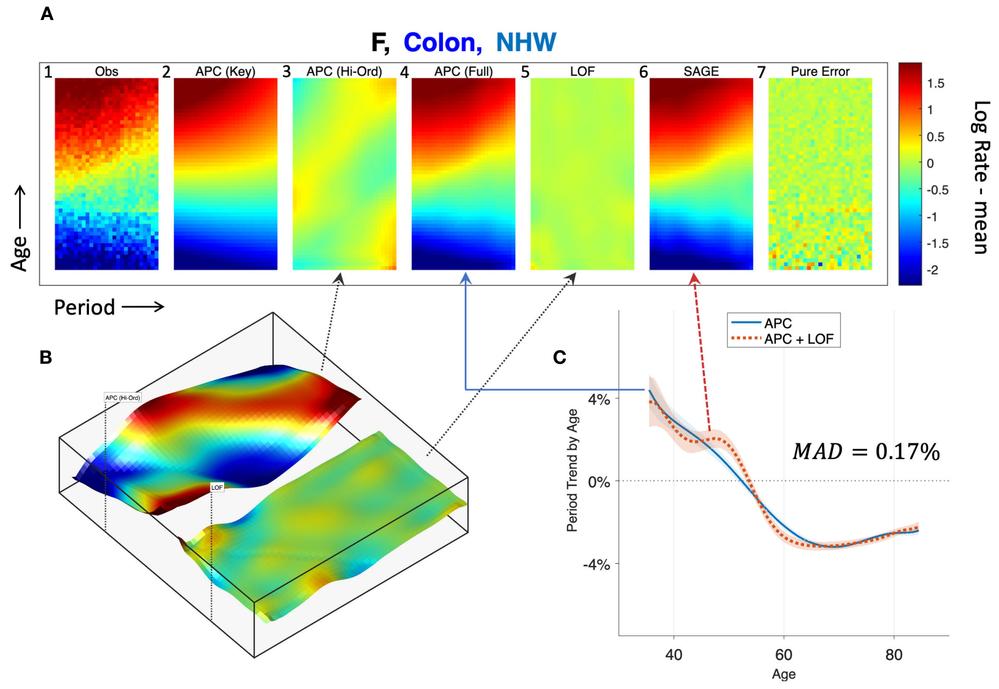


FIGURE 8
Semi-parametric Age-Period-Cohort (SAGE) Analysis. SAGE analysis of colon cancer incidence in Non-Hispanic White Females. **(A)**: Heat Maps of raw data **(A.1)**, pure error **(A.7)**, and decomposition of sifted data **(A.6)** into APC model components **(A.2–A.4)** and Lack-of-Fit (LOF; **A.5**). **(B)**: Surface Plots of APC higher-order deviations (left panel) versus LOF (right panel). **(C)**: Age-specific period trends with (dashed red curve) and without (solid blue curve) LOF. The median absolute deviation (*MAD*) between the curves in 0.17%.

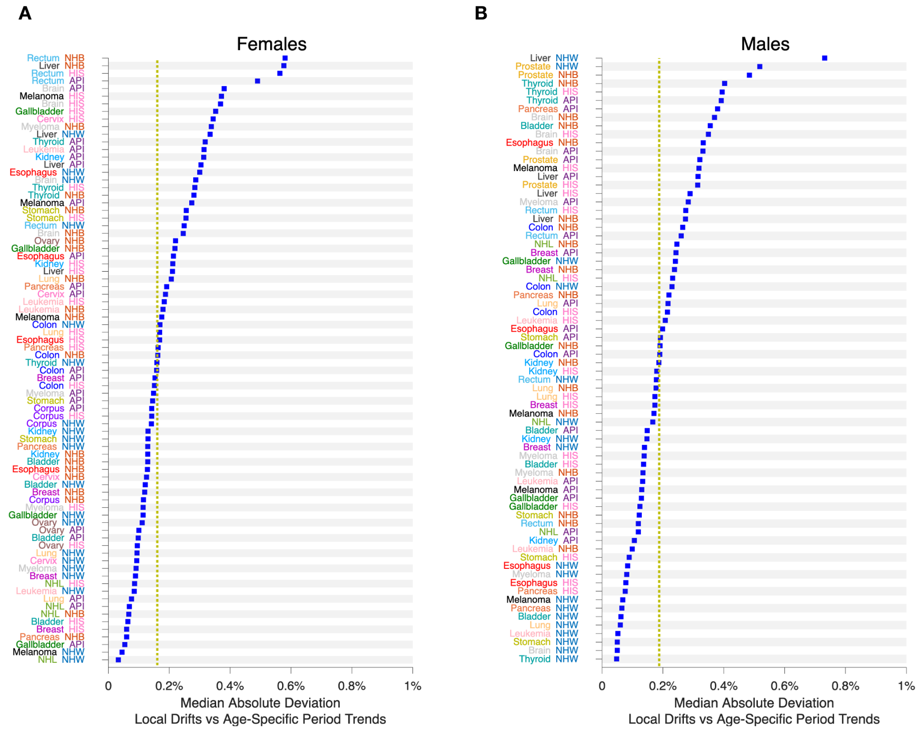


FIGURE 9
Impact of Lack-of-Fit (LOF) on Local Drifts in 152 Cancers. Median Absolute Deviation (*MAD*) between Local Drifts and Age-Specific Period Trends from SAGE analysis. See the Legend to Figure 8 for details. **(A)**, Females. **(B)**, Males. Values by race and ethnicity (API, HIS, NHB, NHW) within cancer sites as labelled. Yellow reference lines show median of *MAD* values for Females and Males.

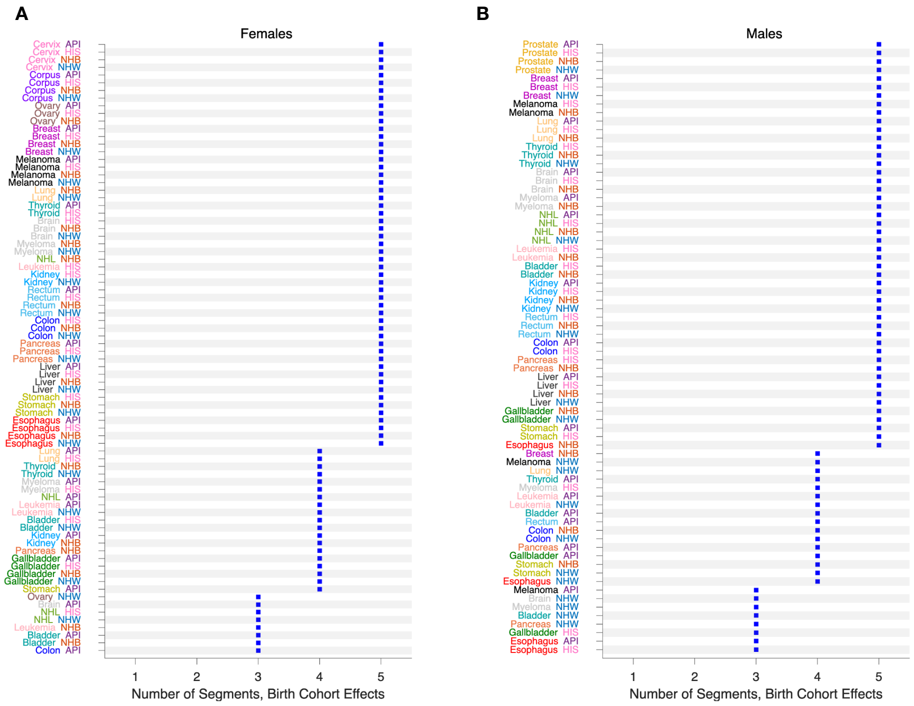


FIGURE 10
Birth Cohort Effects in 152 Cancers. Number of segments identified by JoinPoint analysis of Fitted Cohort Patterns from SAGE. **(A)**, Females. **(B)**, Males. Values by race and ethnicity (API, HIS, NHB, NHW) within cancer sites as labelled.

However, the Lexis diagram can be analyzed on four different time scales (6), and the event rates can be proportional with respect to one time scale but not the others (53).

Recently, we developed a comparative method that can identify whether the stratum-specific hazard rates in an ensemble of $G \geq 2$ Lexis diagrams are *proportional* overall, or within calendar periods, age groups, or birth cohorts (25). Proportionality imposes meaningful constraints on the stratum-specific EF. For example, when the hazard rates are proportional within calendar periods, the Local Drifts for each stratum are all equal. Alternatively, when the hazards are proportional within age groups, the stratum-specific Local Drifts are parallel. Such constraints can highlight important signals that otherwise might be missed by inspection of outputs from separate models.

To illustrate, we carried out an exploratory comparative analysis of colon cancer incidence by sex, race, and ethnicity. The analysis partitioned the 8 strata into 4 subsets: non-proportionality in NHW women and men, age proportionality in NHB, API and HIS women, and absolute proportionality in NHB, API and HIS men. From these partitions we extracted FCPs and ran JoinPoint models (Figure 11). In each stratum, colon cancer incidence bottomed out among Baby Boomers (1946 – 1964 birth cohorts), then increased year-over-year among members of Generation X (1965 – 1980 birth cohorts).

It is computationally feasible to evaluate partitions when the number of strata is small to moderate, $2 \leq G \leq 10$. When $G > 10$, Bayesian methods provide a valuable approach. Bayesian spatial age-period-cohort analysis is one promising application (21). Bayesian methods can be used to characterize the distribution of Features. In practice it appears essential to take birth cohort effects into account (Figure 10). One way to do so is to carry out the Bayesian analyses separately within age strata (22).

3.6 Cancer forecasts

Forecasts of cancer incidence obtained from APC models are popular because the underlying assumptions of the model are often reasonable (Section 2.3.2) (54). APC-based forecasts extrapolate parameter estimates from observed to future age and period cells (44, 54). Consequently, Incidence forecasts are EF. Different scenarios can be modeled by varying the extrapolation scheme used to account for future period and cohort deviations. Best et al. (49) extrapolate future period effects using the global curvature for period, and future cohort effects using the most recent segment from a join-point analysis of the FCP, i.e., by extrapolating from the experience of the youngest observed birth cohorts.

4 Discussion

The novel statistical approaches we describe here do not replace classic analytical tools and methods for cancer rates: they build upon them. Each method starts with one or more Lexis diagrams. SIFT and SAGE increase accuracy by leveraging the power of contemporary nonparametric smoothing. Indeed, SIFT is our recommended smoother for 1x1 and 2x2 data because it offers remarkable increases in precision across a broad spectrum of

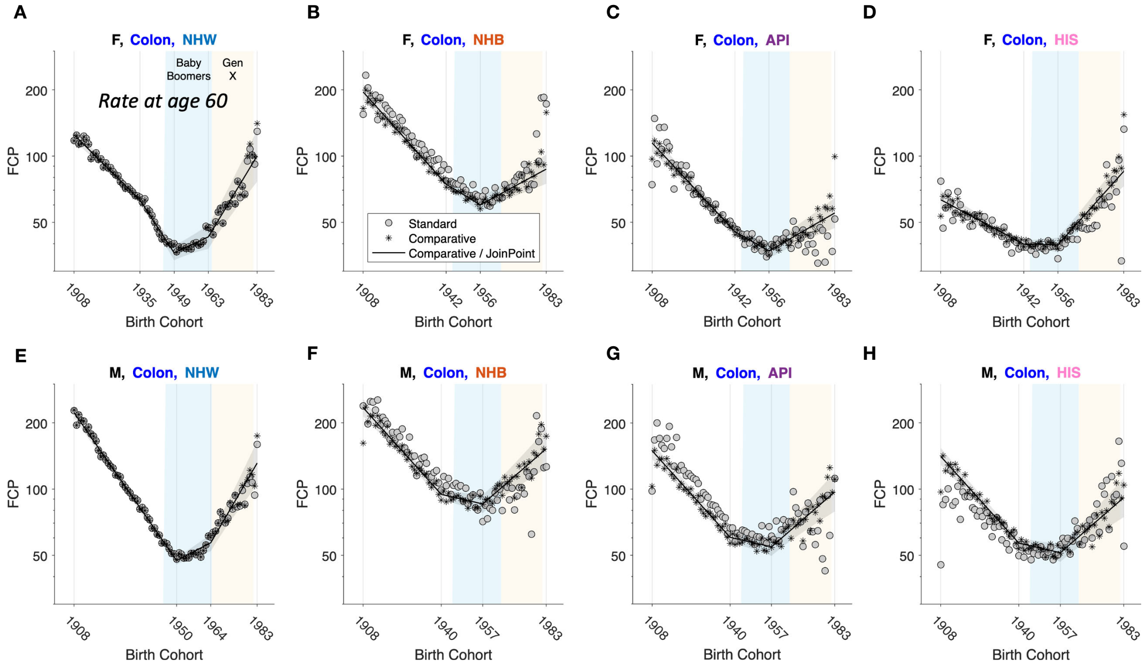


FIGURE 11
Fitted Cohort Patterns (FCPs) for Colon Cancer. FCP: Expected rate at age 60 by birth cohort. Separate APC models (grey circles); joint APC model identified by exploratory comparative analysis (stars); JoinPoint analysis of the joint model (black curves with grey pointwise 95 percent confidence limits). Females (F): **(A–D)**. Males (M): **(E–H)**. Non-Hispanic White (NHW) **(A, E)**, non-Hispanic Black (NHB) **(B, F)**, Asian and Pacific Islander (API) **(C, G)** and Hispanic (HIS) **(D, H)**. In each panel, x-axis ticks show estimated join-points. Baby-Boomer (blue) and Gen-X (yellow) birth cohorts are highlighted.

cancers (Figure 6). The New APC Model and SAGE elucidate birth cohort effects: The latter provides appealingly smooth Estimable Functions (EF). Comparative APC Analysis ascertains cancer heterogeneity across ages, periods, and birth cohorts for a small to moderate number of strata, and Bayesian methods when the number of strata is moderate to large. Taken together, the new methods summarized in Figure 1 mitigate the limits highlighted in Figures 2–5.

With these new tools in hand, our ability to detect fine-scale temporal signals in granular data with one- or two-year age and period intervals is greatly enhanced. Indeed, smoothing Lexis diagrams up front using a contemporary non-parametric procedure such as SIFT has compelling advantages. Accuracy is *greatly* increased (Figure 6), and you can extract all of the standard Features from the sifted data (e.g., canonical plots, ASRs, EAPC, and JoinPoint). You can also extract novel Features defined by averages, trends, and gradients. One particularly valuable Feature is the sifted estimates of the age-specific trends over time, a model-free analogue of the APC Local Drifts.

The APC Model provides a conceptual framework for interpreting cancer incidence based on a principle we call “the primacy of birth cohort effects”. The APC model can be applied to cancer mortality, with the proviso that trends in mortality reflect changes in both cancer incidence and cancer survival. For mortality analysis it is especially crucial to assess LOF because many treatments are used to a greater or lesser extent according to the patient’s age at diagnosis.

The SAGE method illustrated in Figure 8 provides a valuable new tool for “stress-testing” an APC model. If the LOF is large relative to the higher-order deviations, one can step away from model-based EF and base conclusions on Features, which are model-free constructs. This strategy improves the overall reliability of the analysis.

Using SAGE, we surveyed 152 cancer incidence Lexis diagrams across 21 leading sites in men and women in four race and ethnicity groups. The LOF had remarkably little impact on the Local Drifts (Figure 9). This is not surprising if the underlying expected rates are smooth functions, which is a standard assumption. In that case, it is straightforward to show that the 6 key parameters of the New APC Model describe a second-order Taylor expansion of the Lexis diagram around the middle cell located at $(\bar{a}_{\star}, \bar{p}_{\star})$. Furthermore, given the “primacy of birth cohort effects,” it is not entirely unexpected that birth cohort effects are statistically significant in every case (Figure 10).

When birth cohort effects are present – which, for US incidence, appears to be most of the time – the EAPC **cannot** represent the time trend in every age group. Consequently, unless the LOF is substantial, one should always examine birth cohort effects. This is easily done using the New APC Model, but it is much harder to do so using canonical plots or other classic descriptive methods. This is because we observe the oldest cohorts only at older ages and the youngest cohorts only at younger ages: the data are not balanced over cohorts. For the practitioner, our synopsis of “best practices” (Figure 7) should provide a handy “cheat sheet” for applications using the New APC Model or SAGE.

Frequentist statistical methods are now available for comparative studies with a small to moderate number of stratum-specific Lexis diagrams, two through around 10, and Bayesian methods when the number of strata is larger (10 up to several hundred). As illustrated in Figure 11 for colon cancer, Comparative Analysis can identify patterns and signals in birth cohort effects within and between strata that would otherwise be difficult or impossible to detect. Using this approach, we discovered that members of Generation X born between 1965 – 1980 are at increased risk of colon cancer compared to Baby Boomers born between 1946 – 1964. This unfavorable trend was seen across both sexes and in all four race and ethnicity groups.

For each new tool in Figure 1, the requisite statistical software is now available or soon will be. It is now technically possible to integrate advances in statistical methodology and data science and put powerful new tools in the hands of cancer surveillance researchers. Doing so could facilitate a 'golden age' in CSR. Furthermore, these tools can be combined in novel ways, effectively making new tools. Feature extraction from sifted data (50) and SAGE (Figure 8) are two examples.

Whereas forecasts of cancer *incidence* are EF and therefore fall within the purview of the methods summarized in Figure 1, forecasts of cancer *burden* – the absolute numbers of new cases – requires a combination approach that integrates population forecasts from the Census Bureau with incidence forecasts from APC models (49, 54–56). In principle, cancer *prevalence* (past, current, and future numbers of persons living with a cancer) can also be estimated using combination methods that integrate survival analysis of cancer cases with APC models of cancer incidence. This would require a separate toolbox of survival methods including cause-specific hazard functions (57, 58) and cumulative incidence of competing risks (59, 60).

Knowing that something is possible does not make it happen – that will require serious work in the area of implementation science, for example, to accelerate computational algorithms for SIFT, SAGE, and Comparative Analysis, scale up the JoinPoint method for longer time series, streamline access to data using Findable, Accessible, Interoperable, Reusable (FAIR) principles (61), and develop interfaces that integrate FAIR data, analysis tools, and workflows (62). In light of recent increases in cancer incidence occurring in the US (63) and globally (9), we believe such efforts are warranted to advance cancer control and cancer research.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://seer.cancer.gov/data/.

Ethics statement

Our analysis is based entirely on publicly available data.

Author contributions

PR: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. AM-F: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. PR and AM-F are supported by the Intramural Research Program of the National Cancer Institute, Division of Cancer Epidemiology and Genetics. AM-F is also supported through an appointment to the National Cancer Institute (NCI) ORISE Research Participation Program under DOE contract number DE-SC0014664.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. American Cancer Society. *Cancer Surveillance Research* (2023). Available at: https://www.cancer.org/research/surveillance-and-health-equity-science/surveillance-research.html (Accessed September 9, 2023).

2. National Cancer Institute. *SEER*Stat Statistical Software*. (2023). Available at: https://seer.cancer.gov/seerstat.

3. Chen HSM, Zhu L, Kim HJ, Cho H, Feuer EJ. Developments and challenges in statistical methods in cancer surveillance. *Stat And Its Interface* (2014) 7:135–51. doi: 10.4310/SII.2014.v7.n1.a14

4. Holford TR. The estimation of age, period and cohort effects for vital rates. *Biometrics* (1983) 39:311–24. doi: 10.2307/2531004

5. Holford TR. Understanding the effects of age, period, and cohort on incidence and mortality rates. *AnnuRevPublic Health* (1991) 12:425–57. doi: 10.1146/annurev.pu.12.050191.002233

6. Rosenberg PS, Check DP, Anderson WF. A web tool for age-period-cohort analysis of cancer incidence and mortality rates. *Cancer Epidemiol Biomarkers Prev* (2014) 23:2296–302. doi: 10.1158/1055-9965.EPI-14-0300

7. Rosenberg PS. A new age-period-cohort model for cancer surveillance research. *Stat Methods Med Res* (2019) 28:3363–91. doi: 10.1177/0962280218801121

8. Cronin KA, Scott S, Firth AU, Sung H, Henley SJ, Sherman RL, et al. Annual report to the nation on the status of cancer, part 1: National cancer statistics. *Cancer* (2022) 128:4251–84. doi: 10.1002/cncr.34479

9. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: Cancer J Clin* (2021) 71:209–49. doi: 10.3322/caac.21660

10. Smith BD, Smith GL, Hurria A, Hortobagyi GN, Buchholz TA. Future of cancer incidence in the United States: burdens upon an aging, changing nation. *J Clin Oncol* (2009) 27:2758–65. doi: 10.1200/JCO.2008.20.8983

11. NCI. *SEER 50th Anniversary* (2023). Available at: https://seer.cancer.gov/about/50-anniversary.html.

12. Bray FC, Mery L, Pineros M, Znaor A, Zanetti R, Ferlay J. *Cancer incidence in five continents* Vol. Vol XI. . Lyon, France: International Agency for Research on Cancer (2017).

13. NAACCR. *North American Association of Central Cancer Registries* (2023). Available at: https://www.naaccr.org (Accessed September, 2023).

14. Keiding N. Statistical-inference in the lexis diagram. *Philos T Roy Soc A* (1990) 332:487–509. doi: 10.1098/rsta.1990.0128

15. Chien L-C, Wu Y-J, Hsiung CA, Wang L-H, Chang IS. Smoothed lexis diagrams with applications to lung and breast cancer trends in Taiwan. *J Am Stat Assoc* (2015) 110:1000–12. doi: 10.1080/01621459.2015.1042106

16. Camarda CG. MortalitySmooth: an R package for smoothing poisson counts with P-splines. *J Stat Softw* (2012) 50:1–24. doi: 10.18637/jss.v050.i01

17. Smith TR, Wakefield J. A review and comparison of age-period-cohort models for cancer incidence. *Stat Sci* (2016) 31:591–610. doi: 10.1214/16-Sts580

18. Carstensen B. Age-period-cohort models for the Lexis diagram. *Stat Med* (2007) 26:3018–45. doi: 10.1002/sim.2764

19. Riebler A, Held L, Rue H. Estimation and extrapolation of time trends in registry data-borrowing strength from related populations. *Ann Appl Stat* (2012) 6:304–33. doi: 10.1214/11-Aoas498

20. Riebler A, Held L. The analysis of heterogeneous time trends in multivariate age-period-cohort models. *Biostatistics* (2010) 11:57–69. doi: 10.1093/biostatistics/kxp037

21. Chernyavskiy P, Little MP, Rosenberg PS. Spatially varying age-period-cohort analysis with application to US mortality, 2002-2016. *Biostatistics* (2020) 21:845–59. doi: 10.1093/biostatistics/kxz009

22. Chernyavskiy P, Kennerley VM, Jemal A, Little MP, Rosenberg PS. Heterogeneity of colon and rectum cancer incidence across 612 SEER counties, 2000-2014. *Int J Cancer* (2019) 144:1786–95. doi: 10.1002/ijc.31776

23. Chernyavskiy P, Little MP, Rosenberg PS. Correlated Poisson models for age-period-cohort analysis. *Stat Med* (2018) 37:405–24. doi: 10.1002/sim.7519

24. Chernyavskiy P, Little MP, Rosenberg PS. A unified approach for assessing heterogeneity in age-period-cohort model parameters using random effects. *Stat Methods Med Res* (2017) 2017:962280217713033. doi: 10.1177/0962280217713033

25. Rosenberg PS, Miranda-Filho A, Whiteman DC. Comparative age-period-cohort analysis. *BMC Med Res Method* (2023) 23(1):238. doi: 10.1186/s12874-023-02039-8

26. National Cancer Institute. *Surveillance, Epidemiology, and End Results (SEER 13, Plus) Program Populations (1992-2018)* Vol. 2022S. National Cancer Institute, DCCPS, Surveillance Research Program (2022). Available at: www.seer.cancer.gov/popdata.

27. Parkin DM, Whelan SL, Ferlay J eds. *Cancer Incidence in Five Continents* Vol. Vol. VIII. Lyon, France: International Agency for Research on Cancer (2002).

28. Surveillance E, and End Results (SEER) Program. *Registry Groupings in SEER Data and Statistics* (2023). Available at: https://seer.cancer.gov/registries/terms.html.

29. Robertson C, Boyle P. Age-period-cohort models of chronic disease rates. *II: Graphical approaches. Stat Med* (1998) 17:1325–39. doi: 10.1002/(SICI)1097-0258(19980630)17:12<1325::AID-SIM854>3.0.CO;2-R

30. Devesa SS, Donaldson J, Fears T. Graphical presentation of trends in rates. *Am J Epidemiol* (1995) 141:300–4. doi: 10.1093/aje/141.4.300

31. Breslow NE, Day NE. *Statistical Methods in Cancer Research, Volume 2, The Design and Analysis of Cohort Studies*. Oxford: International Agency for Research on Cancer (1987).

32. Fay MP, Tiwari RC, Feuer EJ, Zou Z. Estimating average annual percent change for disease rates without assuming constant change. *Biometrics* (2006) 62:847–54. doi: 10.1111/j.1541-0420.2006.00528.x

33. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. *Stat Med* (2000) 19:335–51. doi: 10.1002/(SICI)1097-0258(20000215)19:3<335::AID-SIM336>3.0.CO;2-Z

34. Kim HJ, Chen HS, Byrne J, Wheeler B, Feuer EJ. Twenty years since Joinpoint 1.0: Two major enhancements, their justification, and impact. *Stat Med* (2022) 41:3102–30. doi: 10.1002/sim.9407

35. Boyle P, Parkin DM. Statistical methods for registries. In: Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, editors. *Cancer Registration: Principles and Methods*. Lyon, France: International Agency for Research on Cancer (1991).

36. Opdyke JD. A unified approach to algorithms generating unrestricted and restricted integer compositions and integer partitions. *J Math Model Algorithms* (2010) 9:45. doi: 10.1007/s10852-009-9116-2

37. Raptis T. *Restricted Integer Composition*. MATLAB Central File Exchange: The Mathworks, Inc. (2023).

38. Kim S, Lee S, Choi JI, Cho H. Binary genetic algorithm for optimal joinpoint detection: Application to cancer trend analysis. *Stat Med* (2021) 40:799–822. doi: 10.1002/sim.8803

39. Kim HJ, Fay MP, Yu BB, Barrett MJ, Feuer EJ. Comparability of segmented line regression models. *Biometrics* (2004) 60:1005–14. doi: 10.1111/j.0006-341X.2004.00256.x

40. Cook MB, Dawsey SM, Freedman ND, Inskip PD, Wichner SM, Quraishi SM, et al. Sex disparities in cancer incidence by period and age. *Cancer Epidemiol Biomarkers Prev* (2009) 18:1174–82. doi: 10.1158/1055-9965.EPI-08-1118

41. Anderson WF, Camargo MC, Fraumeni JF Jr., Correa P, Rosenberg PS, Rabkin CS. Age-specific trends in incidence of noncardia gastric cancer in US adults. *JAMA* (2010) 303:1723–8. doi: 10.1001/jama.2010.496

42. Zhou CK, Check DP, Lortet-Tieulent J, Laversanne M, Jemal A, Ferlay J, et al. Prostate cancer incidence in 43 populations worldwide: An analysis of time trends overall and by age group. *Int J Cancer* (2016) 138:1388–400. doi: 10.1002/ijc.29894

43. Jemal A, Miller KD, Ma J, Siegel RL, Fedewa SA, Islami F, et al. Higher lung cancer incidence in young women than young men in the United States. *New Engl J Med* (2018) 378:1999–2009. doi: 10.1056/NEJMoa1715907

44. Rosenberg PS, Anderson WF. Age-period-cohort models in cancer surveillance research: ready for prime time? *Cancer Epidemiol Biomarkers Prev* (2011) 20:1263–8. doi: 10.1158/1055-9965.EPI-11-0421

45. O'Brien RM. A simplified approach for establishing estimable functions in fixed effect age-period-cohort multiple classification models. *Stat Med* (2021) 40:1160–71. doi: 10.1002/sim.8831

46. Box GEP. Science and statistics. *J Am Stat Assoc* (1976) 71:791–9. doi: 10.1080/01621459.1976.10480949

47. Thun ML, Cerhan JR, Haiman CA, Schottenfeld D. *Cancer Epidemiology and Prevention. 4th edn*. Oxford University Press (2017).

48. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* (2000) 100:57–70. doi: 10.1016/S0092-8674(00)81683-9

49. Best AF, Haozous EA, Berrington de Gonzalez A, Chernyavskiy P, Freedman ND, Hartge PT, et al. Premature mortality projections in the USA through 2030: a modelling study. *Lancet Public Health* (2018) 3(8):E374–84. doi: 10.1016/S2468-2667(18)30114-2

50. Rosenberg PS, Filho AM, Elrod J, Arsham A, Best AF, Chernyavskiy P. Smoothing Lexis diagrams using kernel functions: A contemporary approach. *Stat Methods Med Res* (2023) 32:1799–810. doi: 10.1177/09622802231192950

51. De Kerf JLF. The interpolation method of Sprague-Karup. *J Comput Appl Mathematics* (1975) I:101–10. doi: 10.1016/0771-050X(75)90027-3

52. Clayton D, Schifflers E. Models for temporal variation in cancer rates. II: Age-period-cohort models. *StatMed* (1987) 6:469–81. doi: 10.1002/sim.4780060406

53. Rosenberg PS, Anderson WF. Proportional hazards models and age-period-cohort analysis of cancer rates. *Stat Med* (2010) 29:1228–38. doi: 10.1002/sim.3865

54. Bray F, Moller B. Predicting the future burden of cancer. *Nat Rev Cancer* (2006) 6:63–74. doi: 10.1038/nrc1781

55. Rosenberg PS, Barker KA, Anderson WF. Estrogen receptor status and the future burden of invasive and in situ breast cancers in the United States. *J Natl Cancer Institute* (2015) 107. doi: 10.1093/jnci/djv159

56. Petrick JL, Kelly SP, Altekruse SF, McGlynn KA, Rosenberg PS. Future of hepatocellular carcinoma incidence in the United States forecast through 2030. *J Clin Oncol* (2016) 34:1787–94. doi: 10.1200/JCO.2015.64.7412

57. Rosenberg PS. Hazard function estimation using B-splines. *Biometrics* (1995) 51:874–87. doi: 10.2307/2532989

58. Breslow NE, Lubin JH, Marek P, Langholz B. Multiplicative models and cohort analysis. *J Am Stat Assoc* (1983) 78:1–12. doi: 10.1080/01621459.1983.10477915

59. Dinse GEL. A note on semi-markov models for partially censored data. *Biometrika* (1986) 73:379–86. doi: 10.1093/biomet/73.2.379

60. Gaynor JJ, Feuer EJ, Tan CC, Wu DH, Little CR, Strauss DJ, et al. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *J Am Stat Assoc* (1993) 88:400–9. doi: 10.1080/01621459.1993.10476289

61. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* (2016) 3:160018. doi: 10.1038/sdata.2016.18

62. Grossman RL. Ten lessons for data sharing with a data commons. *Sci Data* (2023) 10:120. doi: 10.1038/s41597-023-02029-x

63. Sung H, Siegel RL, Rosenberg PS, Jemal A. Emerging cancer trends among young adults in the USA: analysis of a population-based cancer registry. *Lancet Public Health* (2019) 4:e137–47. doi: 10.1016/S2468-2667(18)30267-6