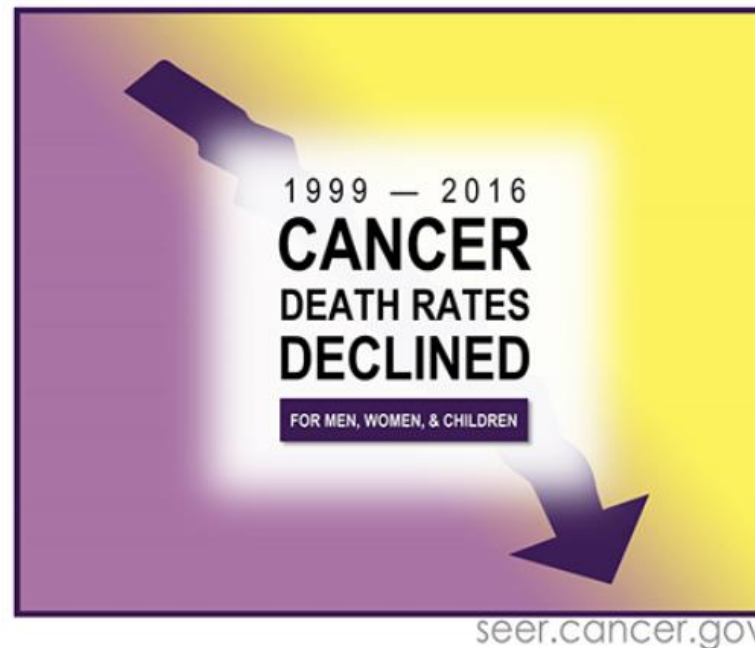# JOINPOINT – Analysis of Trends in Cancer Rates

Huann-Sheng Chen, Rocky Feuer, NCI
Don Green, Jeff Byrne, Steve Scoppa, IMS

NCI Analytic Tools SEERies
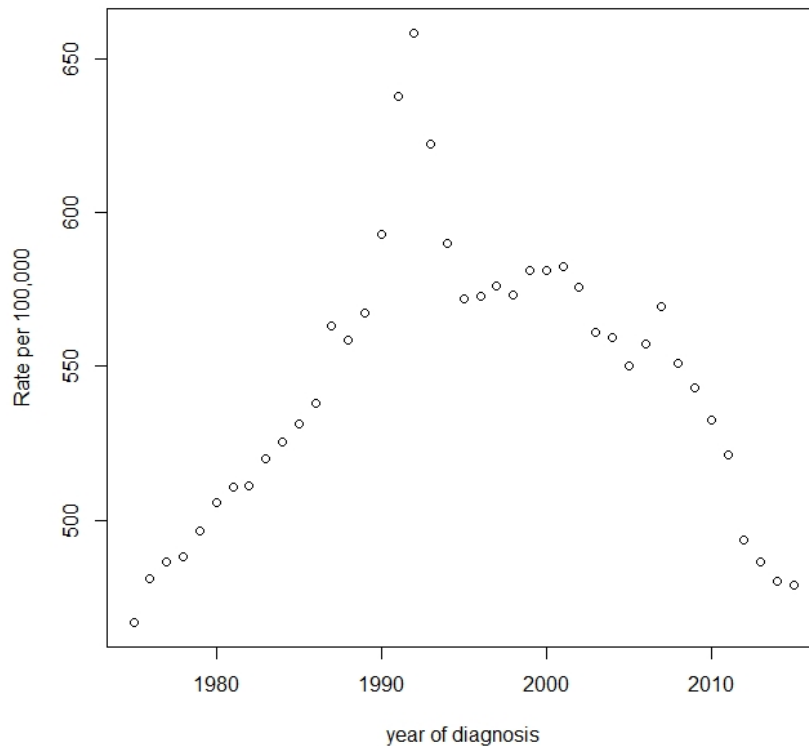
NIH **NATIONAL CANCER INSTITUTE**

SEER

How Should Trends be Summarized?

Incidence Rates for Male, All Sites, from 1975-2015

Incidnece rates of new cases of all cancer, Male, 1975-2015

A polynomial model

Incidnece rates of new cases of all cancer, Male, 1975-2015

A Joinpoint model

Incidnece rates of new cases of all cancer, Male, 1975-2015
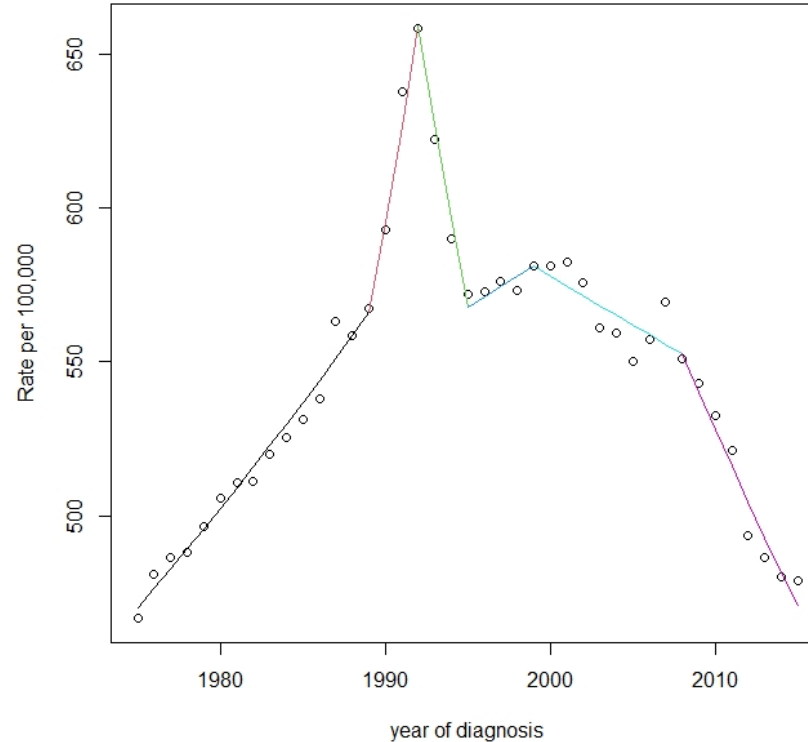
A comparison

Incidnece rates of new cases of all cancer, Male, 1975-2015
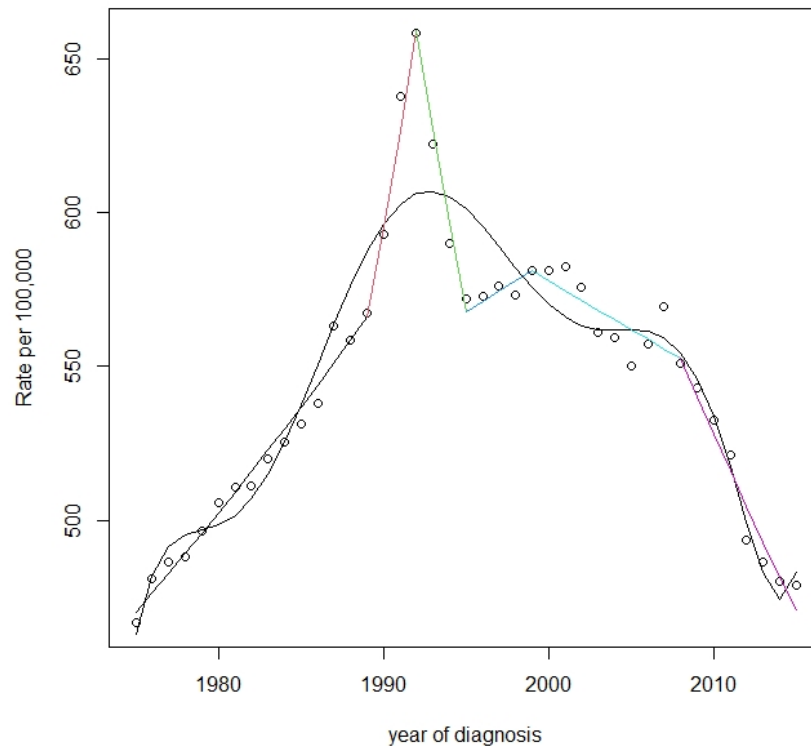
What is a Joinpoint model?
Tries to answer the question "Is the trend changing?"

1. Each segment is linear (usually on a log scale),

2. The trend or slope changes at the joinpoint,

3. The outcome is continuous -- two segments join at the joinpoints,

4. The number and locations of the joinpoints are determined by the data,

5. The minimum and maximum number of joinpoints are selected by the user.

Aliases:

- Piecewise linear model, segmented line model, broken line model, spline regression.
- Joinpoints, change-points, changeover points, knots

We usually use a log scale in cancer rates

- A linear segment on a log scale changes with a constant Annual Percent Change (APC)

- The APC is a relative metric, so it can be meaningfully compared across a broad spectrum of cancers (from very common to very rare)

- We do not necessarily believe that cancer rates change at a constant rate for a fixed period, and then suddenly change direction, but it is a useful simplification of reality (like all models)

NATIONAL CANCER INSTITUTE

k-joinpoint model

Model: $y = \beta_0 + \beta_1 x + \delta_1 (x - \tau_1)^+ + \cdots + \delta_k (x - \tau_k)^+ + \text{error}$

Here $(x - \tau_1)^+ = 0$ if $x - \tau_1 < 0$
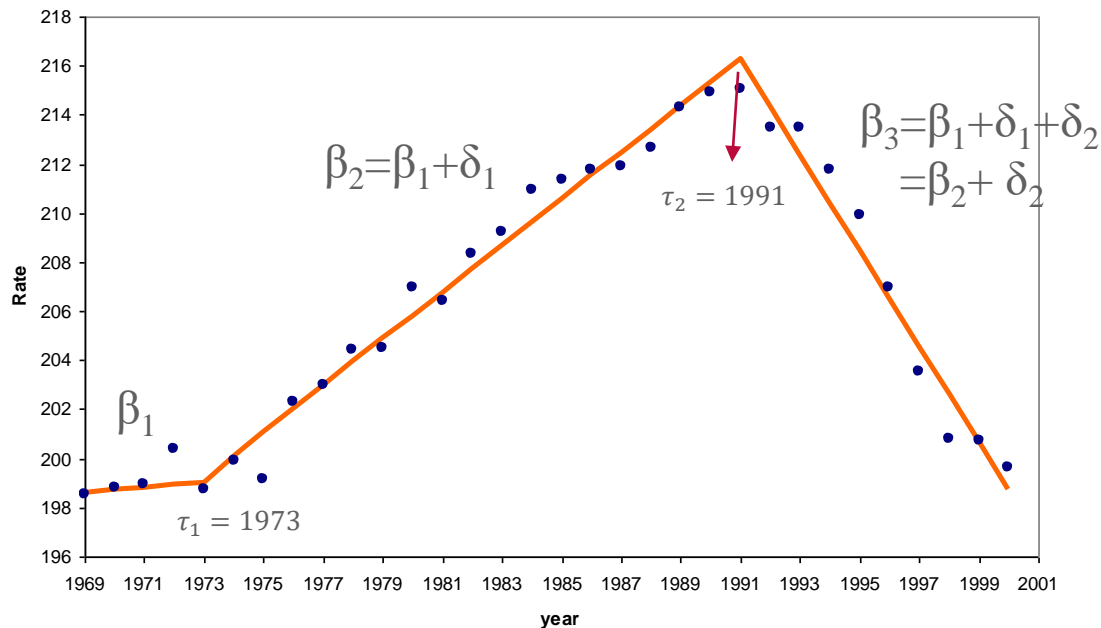
Interpretation of parameters

- Slope of each segment

  - When the dependent variable is in log-scale, the slope is corresponding to annual percent change (APC)

  - $APC = (exp(\beta_i) - 1) \times 100\ \%$

Liver & Intrahepatic Bile Duct Cancer
SEER Age-Adjusted Incidence Rates 2000-2018

By Sex, Observed SEER Incidence Rate, White (includes Hispanic), All Ages, All Stages

2000-2002
APC=0.2%

2002-2006
APC=6.5%*

2006-2015
APC=2.8%*

2015-2018
APC=-2.3%*
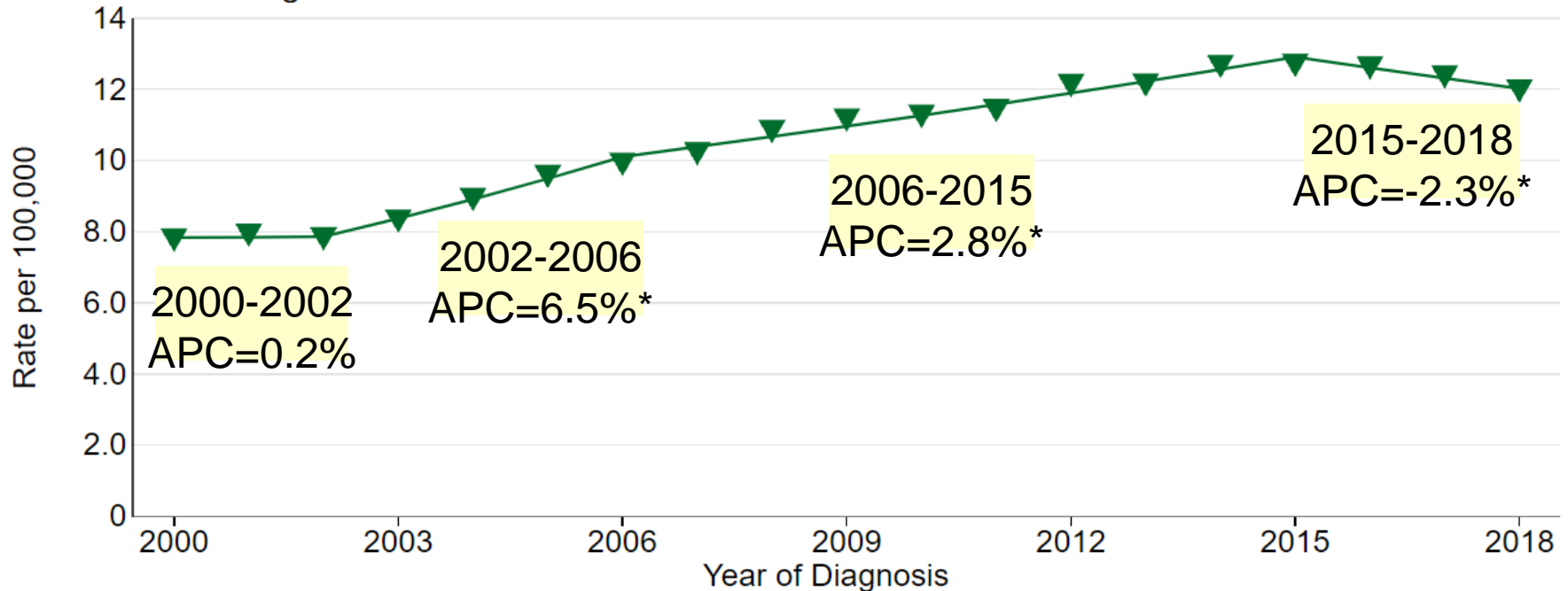
* APC is statistically significant (p≥.05)

For a joinpoint model

- We are interested in

  - Where is/are the change-point(s)?

  - What is the rate of change in a segment?

  - How reliable is the estimate?

  - How to interpret the trend?

Model Estimation

- Fitting the model: Find the best fit among all possible combinations of Joinpoints, assuming that the number of joinpoints is given

  - Grid Search:

    - Creates a "grid" of all possible locations for joinpoints specified by the settings, and calculates the SSE (sum of squared errors) at each one to find the best possible fit.

    - The grid points are on the observations – the default

    - A better fit can be achieved by using a finer grid - by changing the setting for "Number of points to place between adjacent observed x values in the grid search" to something larger than the default of zero.

Model Selection: Determine the number of Joinpoints

- Traditional statistical testing (such as F-test) not possible
- Permutation Test: Computationally intensive (default)
- Traditional BIC Methods:
  - Bayesian Information Criterion (BIC)
  - BIC3
  - Modified BIC
- Data Driven BIC Methods
  - Weighted BIC (WBIC)
  - Weighted BIC Alternative (WBIC-Alt)
  - Data Dependent Selection (DDS)

NATIONAL CANCER INSTITUTE

Permutation Test

- Permute residuals multiple times (N=4499) to ensure that the approximate probability of overall Type I error is less than the specified significance level (also called the alpha level, default = .05).
- Assuming that the default value of the minimum number of joinpoints is 0, "the overall Type I error" is the probability of incorrectly concluding that the underlying model has one or more joinpoints when, in fact, the true underlying model has no joinpoints.

Bayesian Information Criteria (BIC)

The value of BIC is the loglikelihood value with penalizing the cost of extra parameters. The model with the minimum value of BIC is selected as the optimal model.

BIC3

A modification of traditional BIC with a harsher penalty.

Modified BIC

A modification of traditional BIC proposed to improve its performance.

Weighted BIC (WBIC)   [Beta Version]

The weighted BIC combines BIC and BIC3 using a weighted penalty term based on the data characteristic.

Weighted BIC Alternative (WBIC-Alt)   [Beta Version]

A modification of BIC3 that is less conservative than WBIC.

Data Dependent Selection (DDS) [Beta Version]

This procedure internally determines the model selection method, BIC or BIC3, based on the characteristics of data, and its basic idea is to use BIC if change sizes are relatively small and BIC3 otherwise.

Overall recommendations

- Use the permutation test if the user prefers the method that has the longest track record and generally produces parsimonious results (i.e. will detect fewer joinpoints than other approaches, especially if the slope changes are small).

- BIC is good in situations where there are many joinpoints and slope changes are small

- Use the BIC3 method if the user would like to produce results similar to the permutation test, but computation time is an issue.

- Data driven BIC methods combine the best properties of BIC and BIC3

  - Use the Weighted BIC method if the user prefers a method that on average performs best across a wide range of situations. (maybe the default method in the future)

  - The DDS method is a worthy competitor to the Weighted BIC

- Other methods (MBIC, WBIC-Alt) are included in the software for historical continuity, and we do not necessarily recommend their use.

Model Specification

- It does not make sense to have a joinpoint located at the beginning or the end of the data

  - **NumObsEnd**: The minimum number of observations from a Joinpoint to either end of the data, excluding the first or last Joinpoint if it falls on an observation.

  - Default = 2 (means minimum number of data points of first and last segment =3)

- It does not make sense to have two joinpoints next to each other

  - **NumObsBetween**: The minimum number of observations between two joinpoints, excluding any Joinpoint if it falls on an observation.

  - Default = 2 (means minimum number of data points of middle segments =4)

- Only allows a maximum number of joinpoints.

NATIONAL CANCER INSTITUTE

Default maximum number of joinpoints.

Number of Data Points	Default Maximum Number of Joinpoints
0 - 6	0
7 - 11	1
12 - 16	2
17 - 21	3
22 - 26	4
27+	5

- The software allows a maximum of 9 joinpoints
- The permutation test can take a long time to run, which is why the default maximum is no greater than 5
- WBIC method is more computationally efficient and it is one reason we are moving toward using it as default method

Interpretation of results

- How to summarize the recent trends across both sexes, many cancer sites, and race/ethnicities?

Melanoma (Whites)- Incidence

APC = 8.9*

APC = 0.2

Males　　**APC = 3.7***

APC = 5.8*

APC = 3.5*

Females

APC = 2.2*

APC = 6.2*

NATIONAL CANCER INSTITUTE
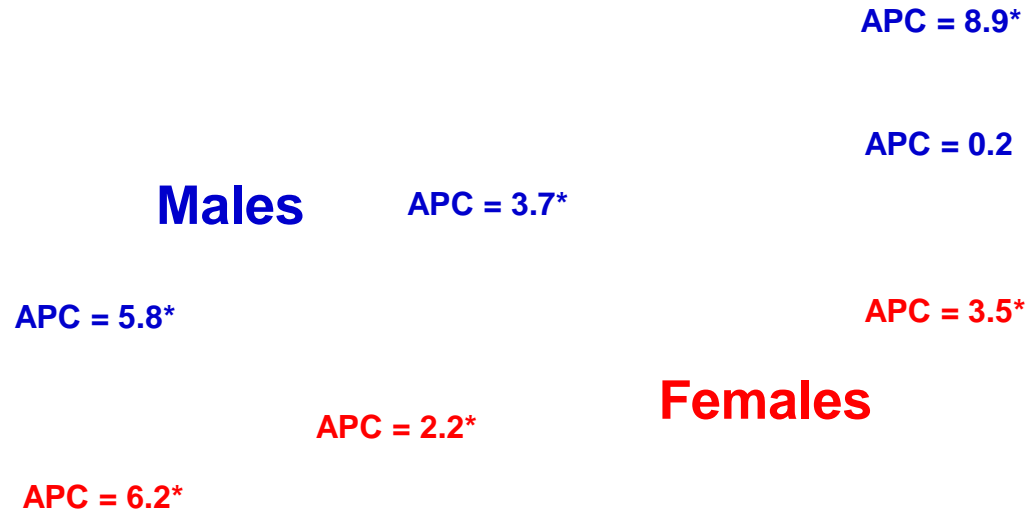
Summarize the Recent Trends Across many cohorts

- Since many are interested in the most recent trend, one possibility would just give the APC for the final segment, the year it started, and its statistical significance

- Problem:
  - Comparing segments of unequal length presents compatibility issues, because the standard error of the APC is related to the length of segment

- Alternatively, we can ask "what is the last 5 year trend?", or "what is the last 10 year trend"? This can be answered by "Average APC (AAPC).

Melanoma (Whites)- Incidence
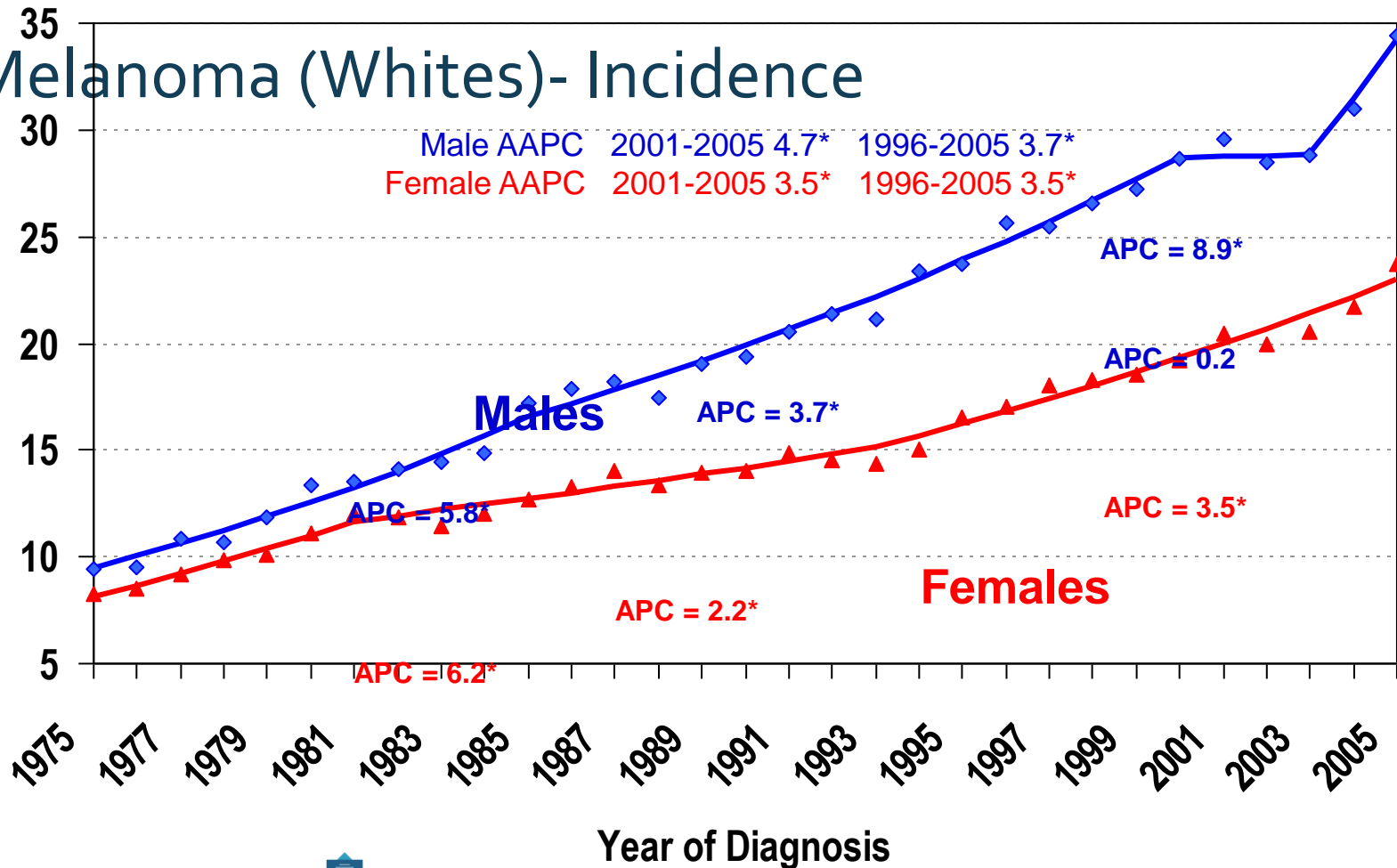
AAPC – average APC

- The AAPC over any fixed interval is calculated using a weighted average of the slope coefficients of the underlying joinpoint regression line with the weights equal to the length of each segment over the interval.

For each segment (i)

$$\text{APC}_i = \left(e^{b_i} - 1 \right) \times 100$$

Summary measure over many segments

$$\text{AAPC} = (e^{\sum w_i b_i / \sum w_i} - 1) \times 100$$

NATIONAL CANCER INSTITUTE

Is the change statistically significant?

Confidence interval of APC and AAPC -- decide whether the change is statistically significant

- Parametric Method: the AAPC confidence interval is based on a Normal distribution; the APC confidence interval is based on a t distribution.
  - Using simulation we found the confidence intervals sometimes are too wide
- Empirical Quantile:
  - We cannot improve the methods using standard parametric inference
  - We developed a computationally intense resampling method which improves the test properties
- Parametric method is the current default, but we are thinking of changing to empirical quantile
- The software also use these methods for confidence interval of the locations of Joinpoints

APC and AAPC Summary

- The 5-year AAPC (and its statistical significance) is now the established summary measure we provide to indicate if the rates for a cancer are going up or down
  - When possible it is best to provide both the full Joinpoint series and the AAPC

Cancer of the Thyroid (Invasive)

SEER 9 Delay-Adjusted Incidence[a], 1975–2015

Race/Ethnicity	Joinpoint Trend 1		Joinpoint Trend 2		Joinpoint Trend 3		Joinpoint Trend 4		Joinpoint Trend 5		Joinpoint Trend 6		AAPC[d]	
	Years	APC	Years	APC	Years	APC	Years	APC	Years	APC	Years	APC	2006-15	2011-15
All Races	1975-77	6.3	1977-80	-6.5	1980-97	2.4*	1997-09	6.7*	2009-15	1.5*			3.2*	1.5*
White	1975-77	7.5	1977-80	-6.7	1980-97	2.7*	1997-09	7.0*	2009-15	1.4*			3.2*	1.4*
Black	1975-95	0.8	1995-12	6.3*	2012-15	-1.9							3.5*	0.1

Some advanced topics

- Testing whether trends of two cohorts are parallel or identical – pairwise comparison

- Jump model/comparability ratio model – Joinpoint model for data with coding change

  - Example: Coding change from ICD-9 to ICD-10 in 1999

NATIONAL CANCER INSTITUTE

Summary: JOINPOINT - Software Product for the Analysis of Trends in Cancer Rates

- A series of connected line segments

  - For each segment the trend rises or falls at a constant Annual Percent Change (APC) until it changes abruptly at "joinpoints"

  - Dependent variable is the ln(rate) – linear segment on a log scale changes at a constant APC

  - Statistical problem is to determine the number and location of the joinpoints (standard parametric inferential methods are not possible – uses a computationally intensive permutation test)

- Answers the seemingly simple question "Is the trend changing", which is surprisingly difficult to answer using standard statistical software

  - The APC is a relative metric, so it can be meaningfully compared across a broad spectrum of cancers (from very common to very rare)

  - We do not necessarily believe that cancer rates change at a constant rate for a fixed period, and then suddenly change direction, but it is a useful simplification of reality (like all models) that has proven useful in the interepretation of trends

- Used widely for the analysis of cancer trends and other health metrics (over 5000 downloads a year – also a callable version to imbed in other software)

- Continuing updates and enhancements based on user feedback (e.g., Jump model to incorporate coding changes, use with complex survey data)

www.cancer.gov   **www.cancer.gov/espanol**