Section XXIX
Cancer Incidence Rates Adjusted for Reporting Delay

Limin X. Clegg, Douglas N. Midthune, Eric J. Feuer, Michael P. Fay, Benjamin F. Hankey

Timely and accurate calculation of cancer incidence rates is hampered by reporting delay, the time elapsed before a diagnosed cancer case is reported to the NCI. Currently, the NCI allows standard delays of approximately 13 and 19 months between the time cases are diagnosed and the time they are first reported to the NCI. The submission of data in August (i.e., 19 months after diagnosis) is released to the public (e.g. cases diagnosed in 1997 were reported to the NCI in August 1999 and released to the public in April 2000). However, in each subsequent release of the SEER data, all prior diagnosis years (e.g. diagnosis years 1996 and earlier in the 1999 submission to the NCI) are updated as either new cases are found or errors are detected in existing data. The submissions for the most recent diagnosis year are, in general, about two percent below the number of cancers that will be submitted in later years, although this varies by cancer site and other factors. The idea behind modeling reporting delay is to adjust the current case count to account for anticipated future corrections to the data (both additions and deletions). These adjusted counts and the associated delay model are valuable in more precisely determining current cancer trends, as well as monitoring the timeliness of data collection (an important aspect of quality control). Reporting delay models have been previously used in the reporting of AIDS cases (1-3).

In this, our first report of delay adjusted incidence rates, we show SEER age-adjusted incidence rates and trends for five cancer sites: melanoma (for whites only), lung/bronchus, colon/rectum, prostate, and female breast. We chose melanoma because it has a longer reporting delay than other cancer sites, presumably because of the difficulties associated with a cancer which is increasingly diagnosed in a non-hospital setting. We chose the four other cancer sites because of their high incidence rates and importance in cancer control.

A delay distribution models the probability of a cancer being reported after a delay of (d) years (d = 2, 3, ...18). The number of cancers reported at each delay year is assumed to follow a Poisson distribution. Cases are removed as corrections to the data are made, and the probability of removing cases is modeled as a binomial distribution. To reduce the number of parameters which have to be estimated and to achieve stability in the tails of the delay distributions, an assumption is made that all cancer cases will be reported within 18 years of diagnosis.

The delay distributions were modeled as a function of covariates using a discrete-time proportional hazards model, or as a stratified variable if for particular variables one does not believe that the proportional hazards assumption is met. For the models presented here diagnosis year, delay times, race/ethnicity, and registry were included as potential covariates. Reporting source (in-hospital and non-hospital) was designated as a potential stratification variable rather than a covariate, because of large possibly non-proportional differences in the delay distribution. Diagnosis year was modeled either as a continuous covariate or as categorized variables: 1981-1985, 1986-1990, or 1991-1997. Only blacks and whites were analyzed. For melanoma, only whites were analyzed because melanoma is rare for blacks.

Maximum likelihood estimates of delay probabilities were obtained using Newton-Raphson algorithm. For each of the cancer sites, models of many combinations of covariates and a stratification variable for reporting source were considered. We evaluated the models by fitting the models using data from each of the annual August data submissions between 1983 and 1997, and then predicting the counts for the 1999 submission. For each cancer site, the model that minimized the sum of squared prediction errors was chosen as the final model. The chosen model was then refit using all data (1983-1999 submissions, i.e. 1981-1997 diagnosis years) to estimate delay distributions and calculate delay-adjusted estimates of the cancer counts.

Age-adjusted (directly adjusted using 1970 US population as the standard) cancer incidence rates

were then calculated with and without adjusting for reporting delay. Joinpoint linear regression was used to obtain the annual percentage changes for the 1973-1997 incidence rates for the data series with and without delay adjustment. The delay model was not fit to diagnosis years prior to 1981 because data submissions before 1983 were not available. Thus in fitting the joinpoint models prior to 1981 the delay adjusted rates are set equal to the observed rates. Up to 3 change points were allowed in the joinpoint regression model (4), which was fit using the option of unweighted least squares in Version 2.5 of the joinpoint regression program developed by the NCI (the latest version of the software can be freely downloaded at the web site www.dccps.ims.nci.nih.gov/SRAB).

The graphs show that adjusting for delay tends to raise cancer incidence rates in more current reporting years. While this adjustment increases the rate of change over the most recent diagnosis years, it probably will only rarely cause the detection of a new joinpoint, although this is possible. For example, the delay adjusted rates for prostate cancer show some early signs of the development of a recent change in the trends for blacks, although this change is not yet statistically significant. Publications on the precise statistical formulation of the reporting delay model, and its application to SEER cancer rates are forthcoming**.**

**References**

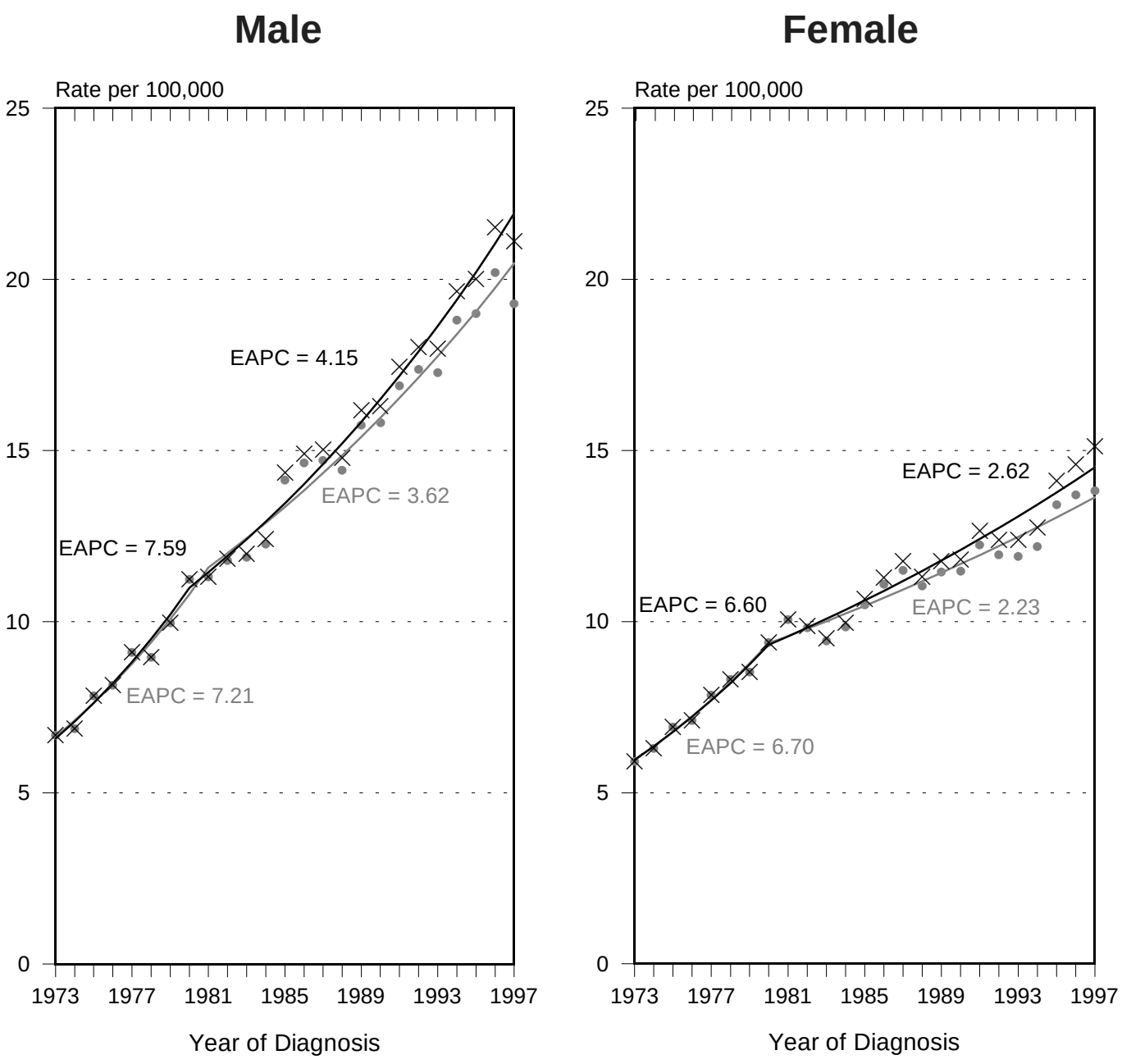

1. Brookmeyer, R. and Damiano, A. (1989). Statistical methods for short-term projections of AIDS incidence. *Statistics in Medicine* **8**, 23-34.

2. Pagano, M., Tu, X.M., De Gruttola, V. and MaWhinney, S. (1994). Regression analysis of censored and truncated data: estimating reporting-delay distributions and AIDS incidence from surveillance data. *Biometrics* **50**, 1203-1214

3. Harris, J.E. (1990). Reporting delays and the incidence of AIDS. Journal of the American Statistical Association **85**, 915-924.

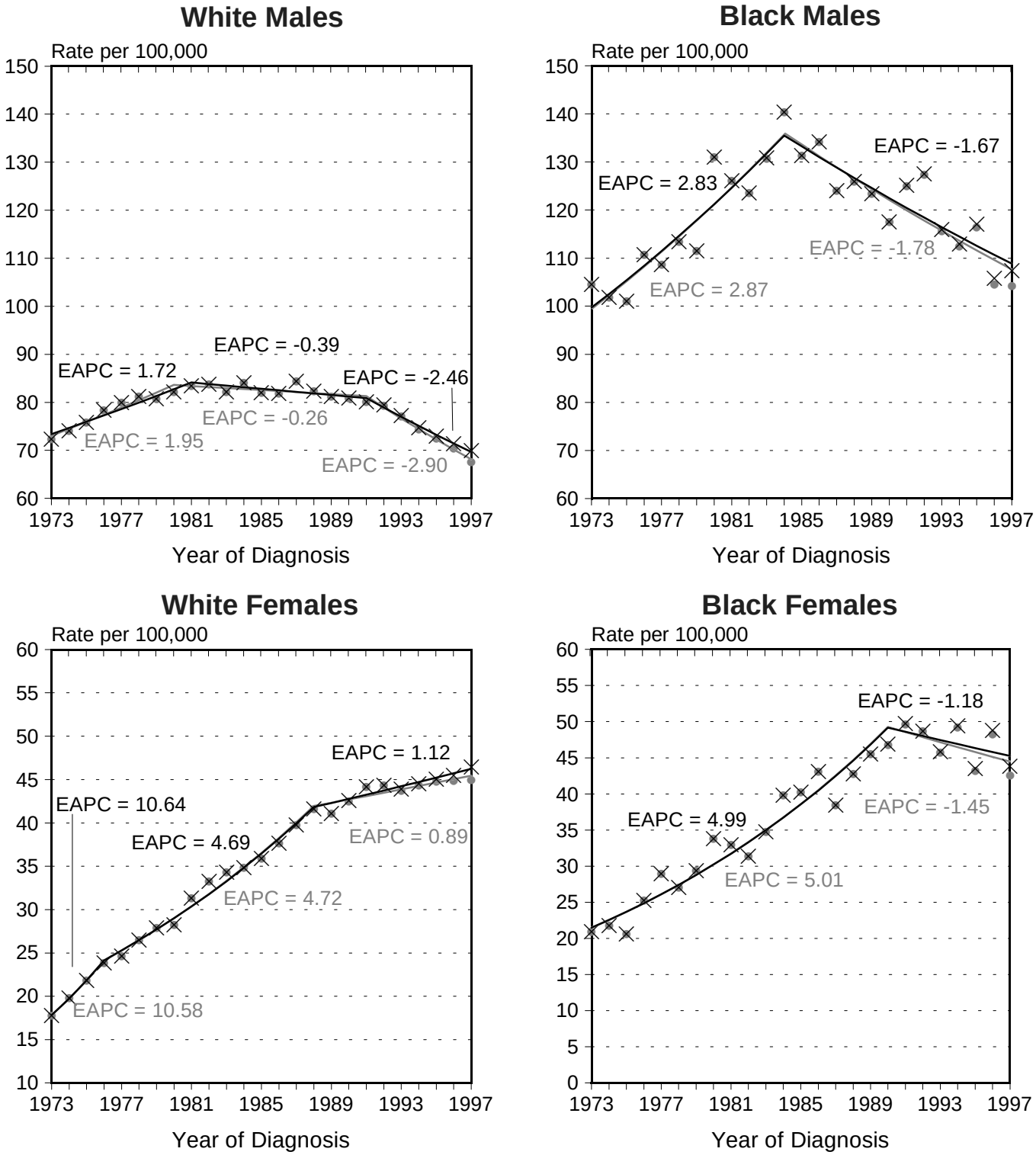
4. Kim, H.J., Fay, M.P., Feuer, E.J. and Midthune, D.N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine* **19**, 335-351.

# SEER Incidence and Delay Adjusted Incidence Rates[+]
# Melanoma of the Skin
# White, By Sex



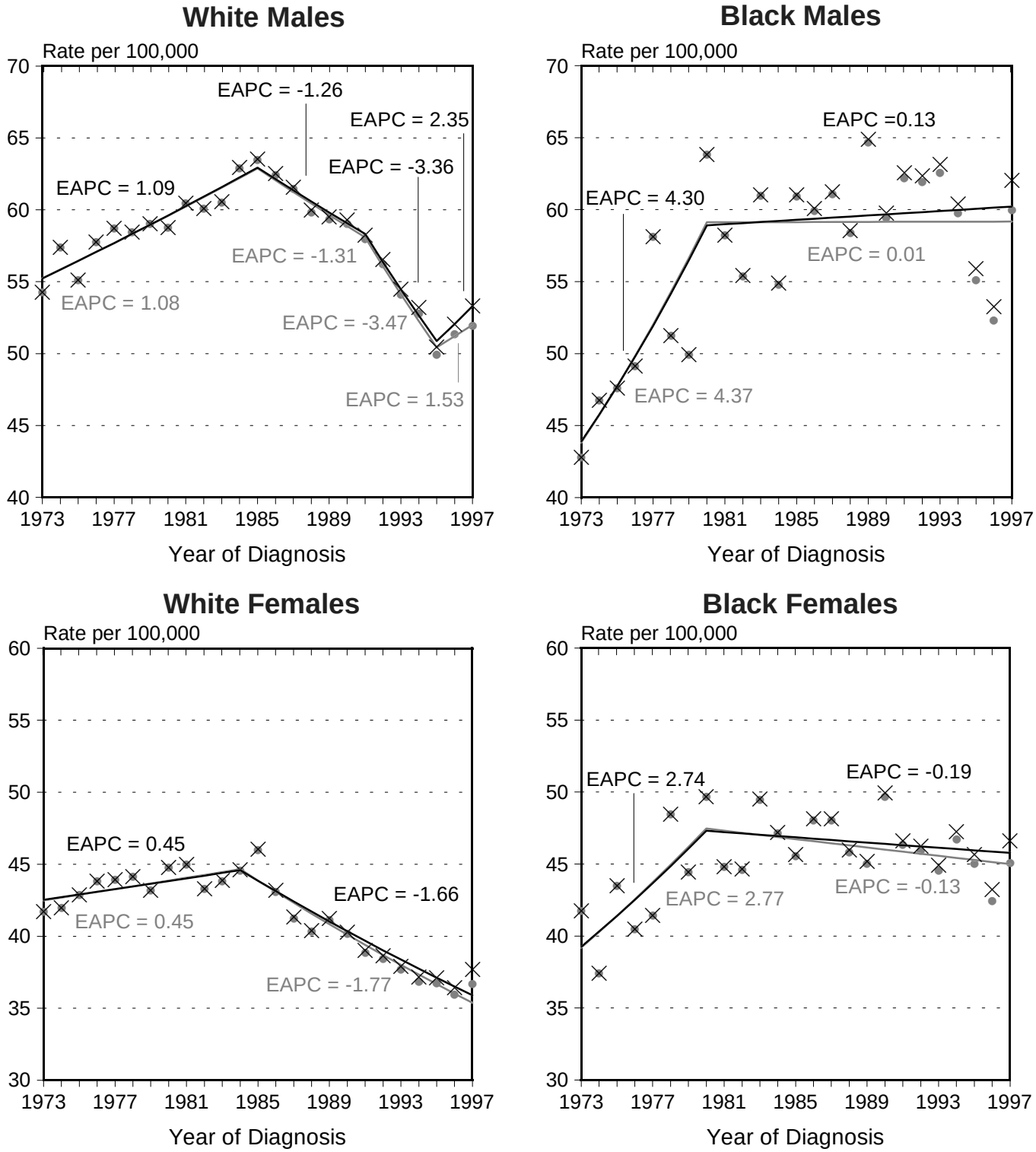

[+] Rates are age-adjusted to the 1970 U.S. standard million population. Regression lines are calculated using the Joinpoint Regression Program. The EAPC is the Estimated Annual Percent Change for the regression line segments and is generated by the Joinpoint Regression Program.

# SEER Incidence and Delay Adjusted Incidence Rates[+]
# Lung and Bronchus
# By Race and Sex



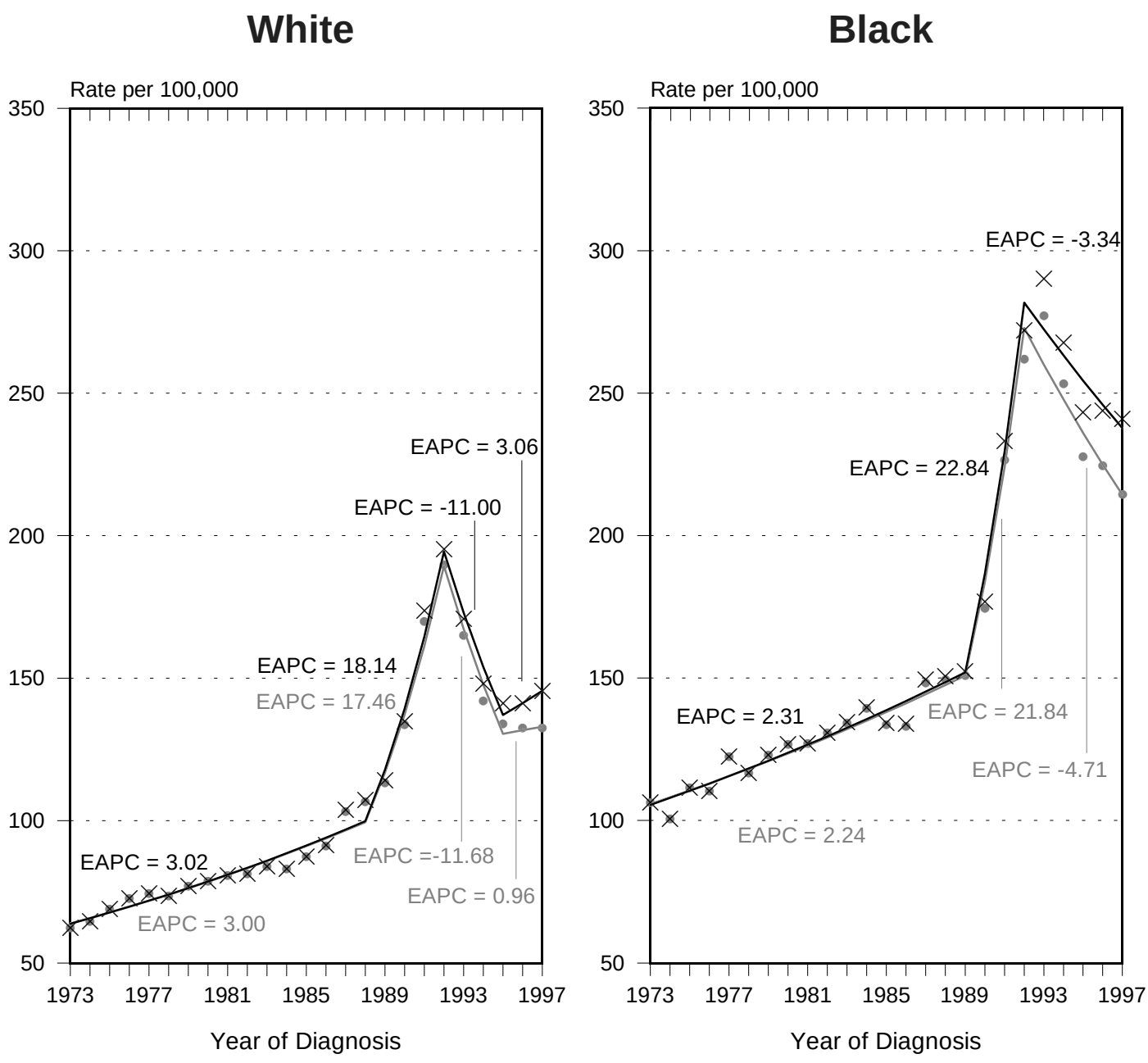

[+] Rates are age-adjusted to the 1970 U.S. standard million population. Regression lines are calculated using the Joinpoint Regression Program. The EAPC is the Estimated Annual Percent Change for the regression line segments and is generated by the Joinpoint Regression Program.

# SEER Incidence and Delay Adjusted Incidence Rates[+]
# Colon & Rectum
# By Race and Sex



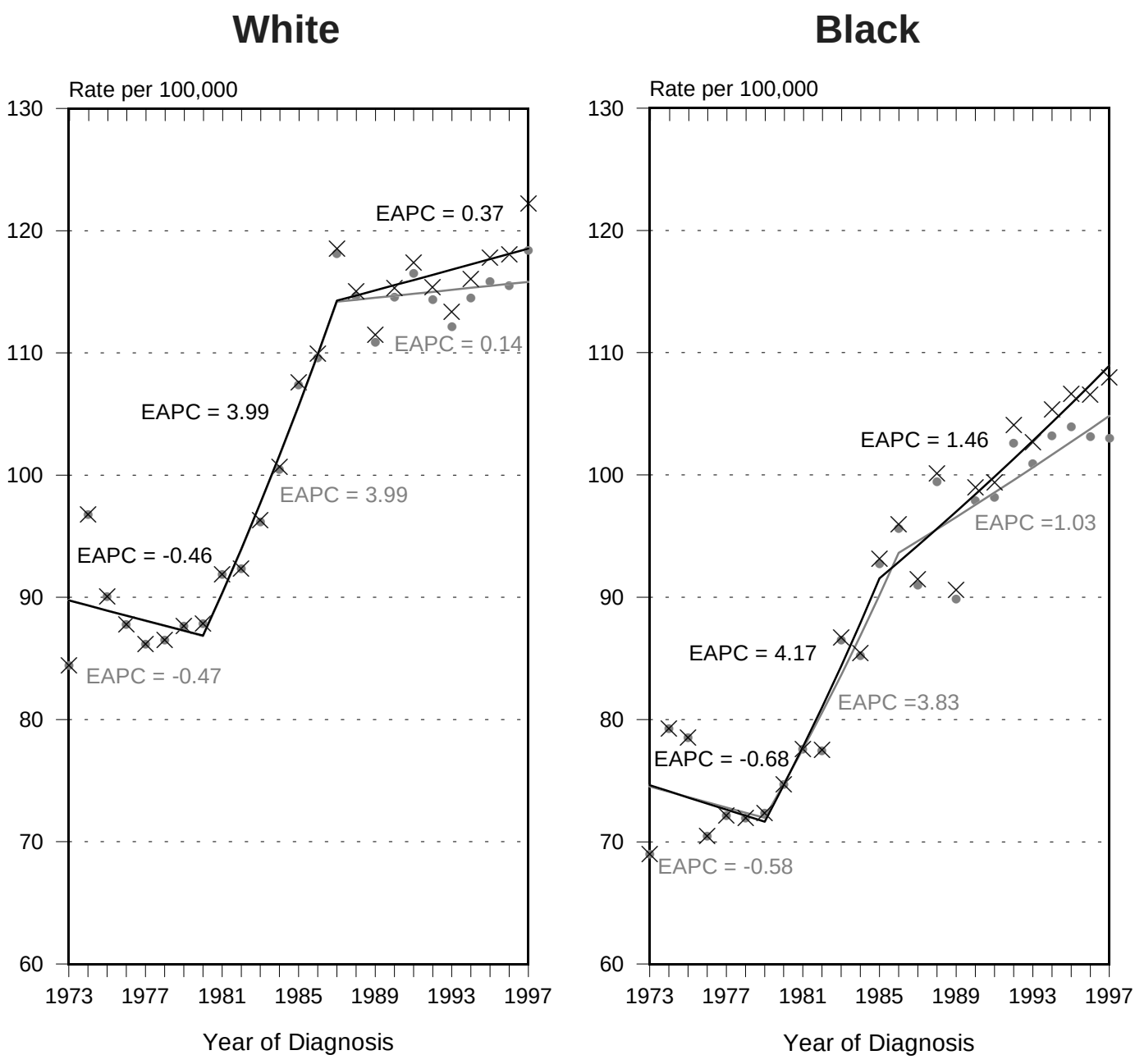

[+] Rates are age-adjusted to the 1970 U.S. standard million population. Regression lines are calculated using the Joinpoint Regression Program. The EAPC is the Estimated Annual Percent Change for the regression line segments and is generated by the Joinpoint Regression Program.

# SEER Incidence and Delay Adjusted Incidence Rates[+]
# Prostate
# By Race




[+] Rates are age-adjusted to the 1970 U.S. standard million population. Regression lines are calculated using the Joinpoint Regression Program. The EAPC is the Estimated Annual Percent Change for the regression line segments and is generated by the Joinpoint Regression Program.

# SEER Incidence and Delay Adjusted Incidence Rates[+]
# Female Breast Cancer
# By Race




[+] Rates are age-adjusted to the 1970 U.S. standard million population. Regression lines are calculated using the Joinpoint Regression Program. The EAPC is the Estimated Annual Percent Change for the regression line segments and is generated by the Joinpoint Regression Program.