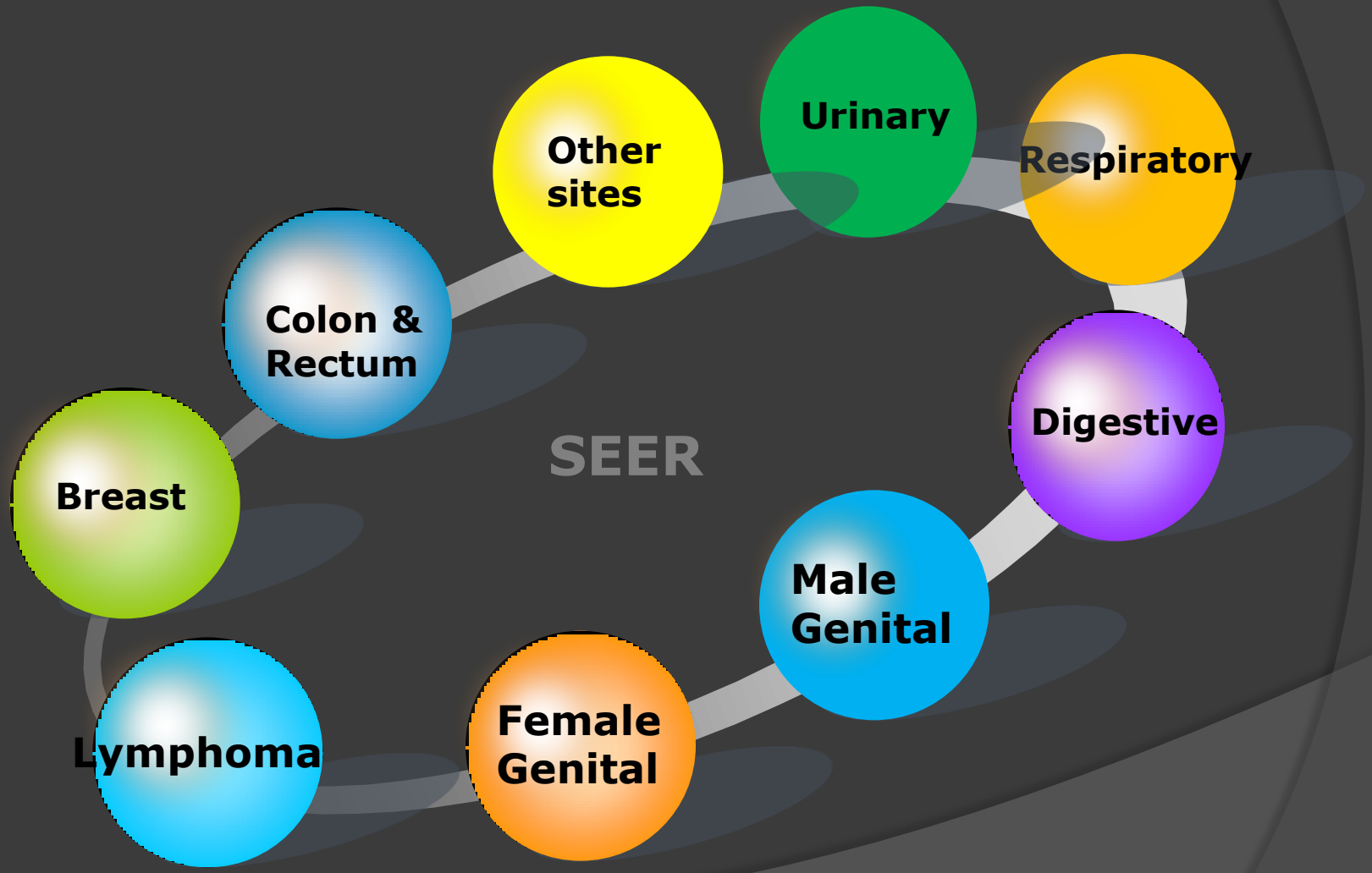
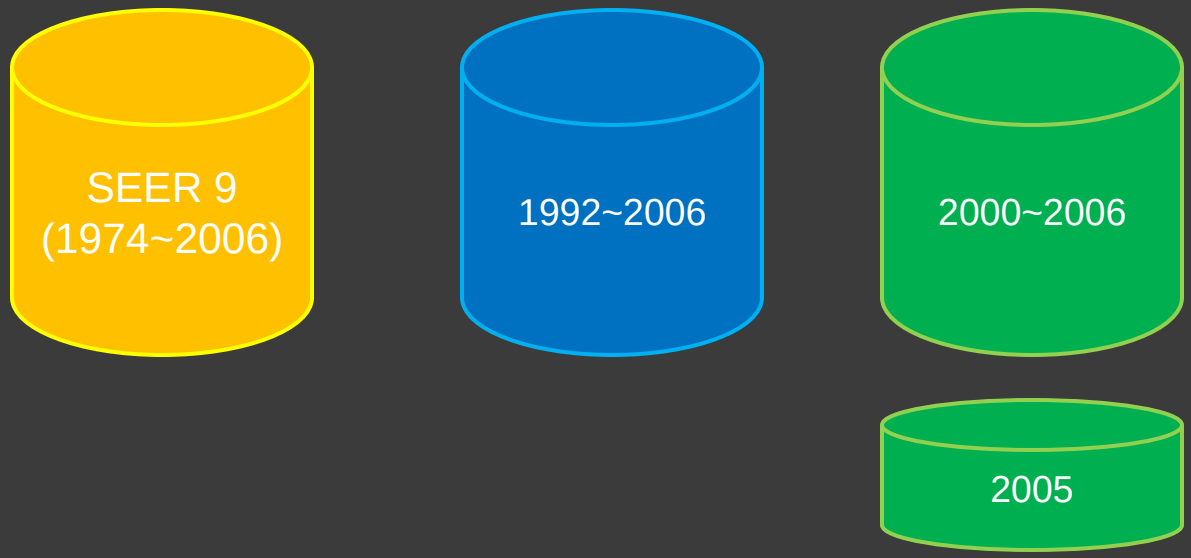
SEER DATABASE DATA MANIPULATION

USF CANCER RESEARCH TEAM

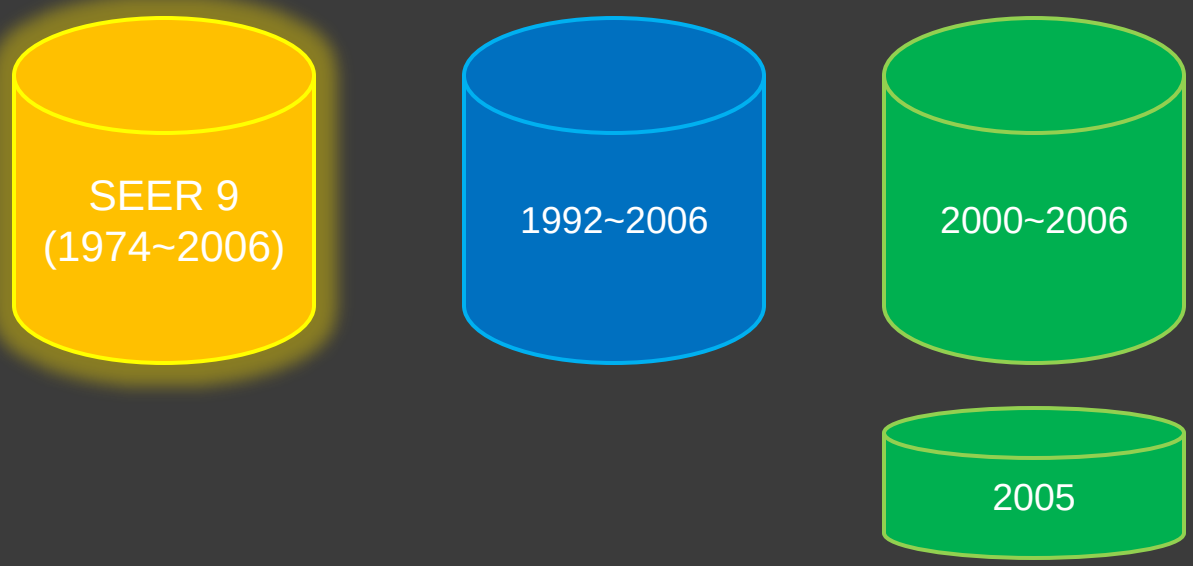
SEER Cancer Data



Understanding SEER database



Understanding SEER database



Understanding SEER database

The SEER 9 registries are Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, Seattle-Puget Sound, and Utah. Data are available for cases diagnosed from 1973 and later for these registries with the exception of Seattle-Puget Sound and Atlanta. The Seattle-Puget Sound and Atlanta registries joined the SEER program in 1974 and 1975, respectively.

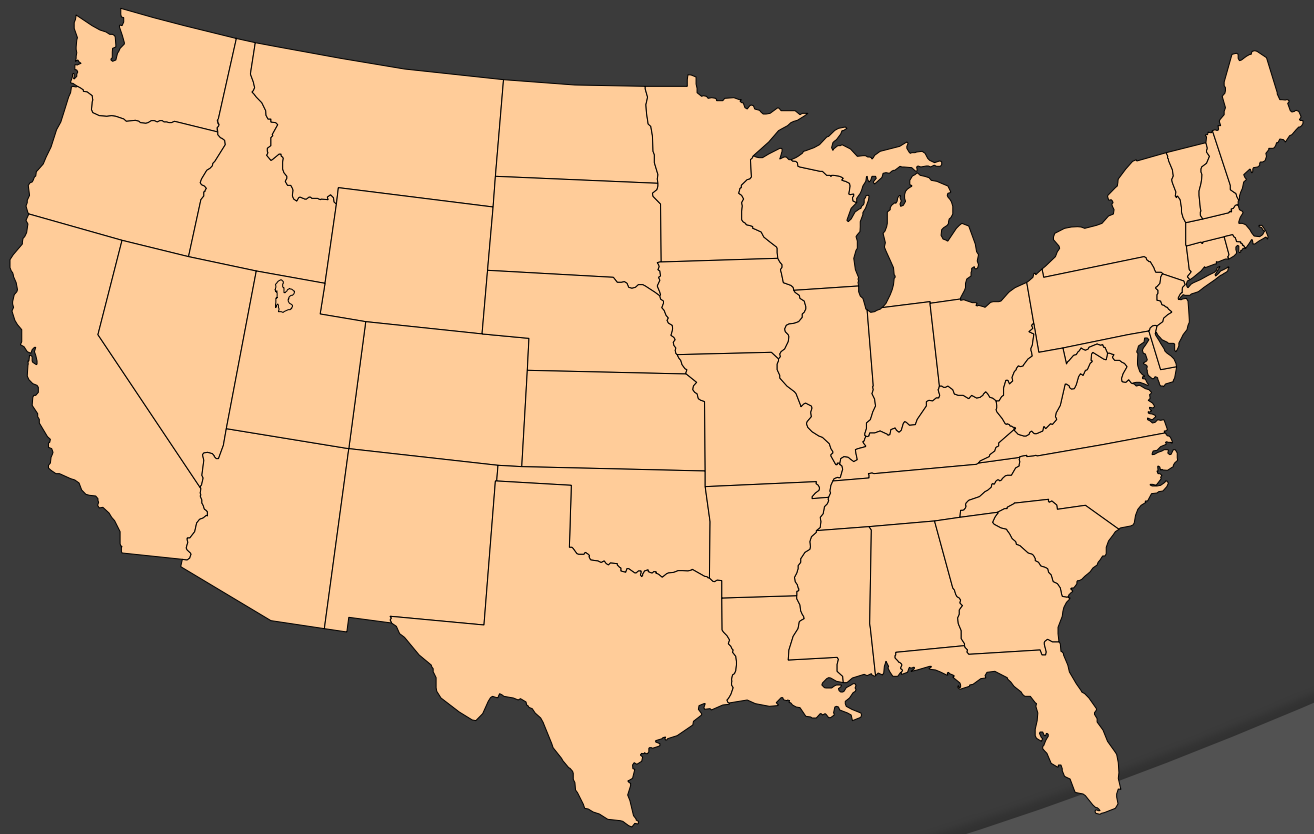
http://www.cancer.org

Understanding SEER database

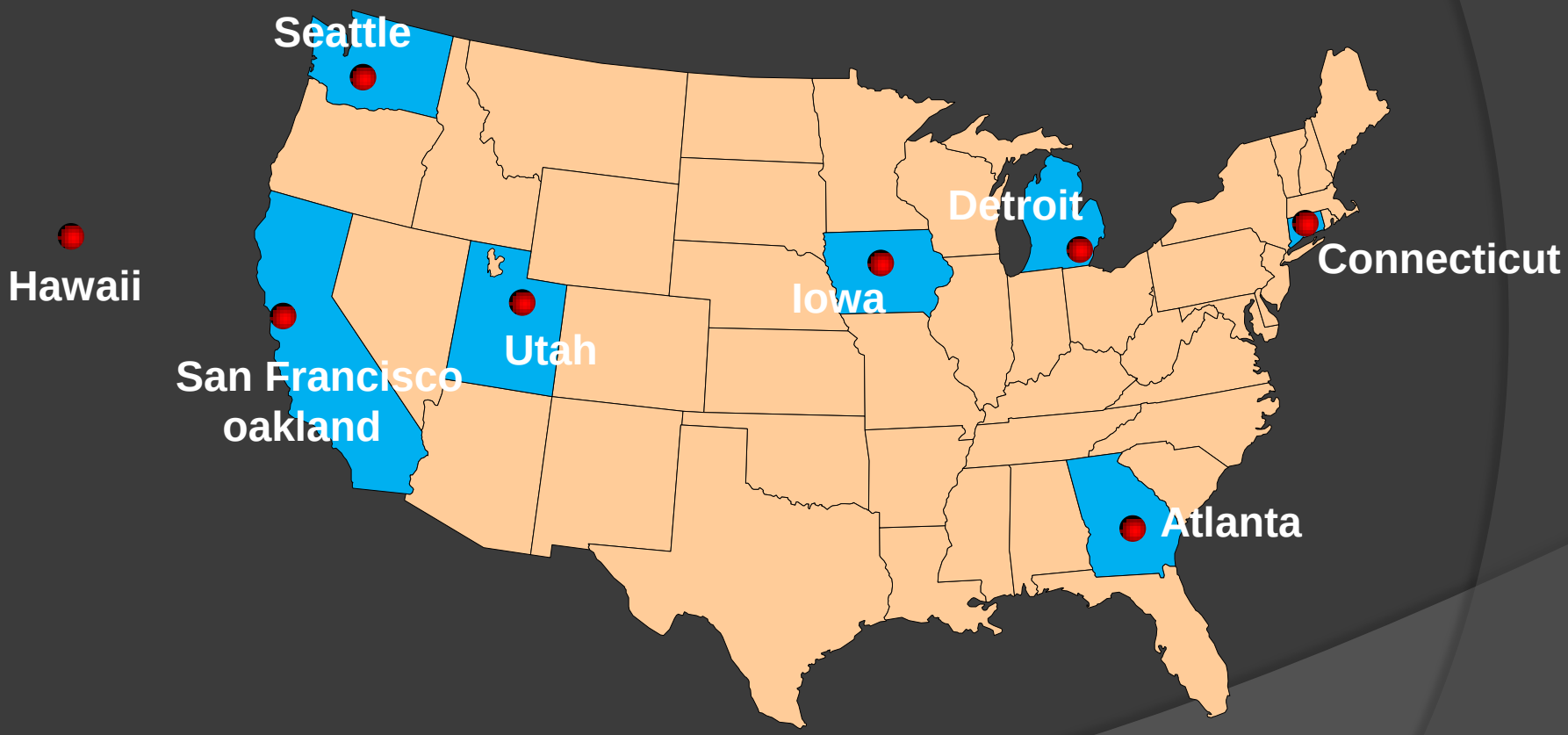
The SEER 9 registries are Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, Seattle-Puget Sound, and Utah. Data are available for cases diagnosed from 1973 and later for these registries with the exception of Seattle-Puget Sound and Atlanta. The Seattle-Puget Sound and Atlanta registries joined the SEER program in 1974 and 1975, respectively.

http://www.cancer.org

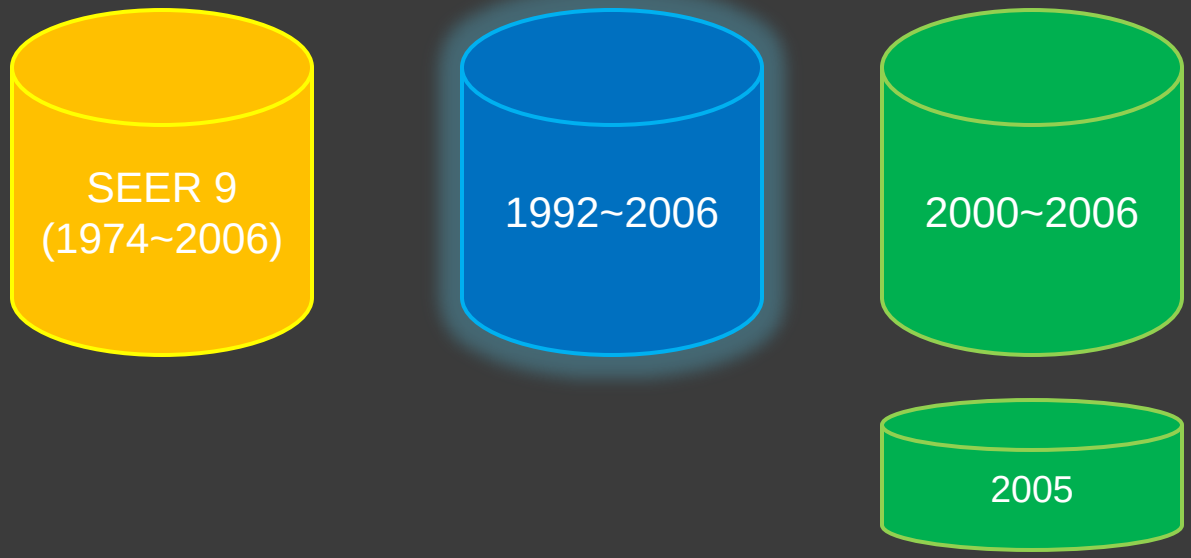
SEER 9 (1973~2006)



SEER 9 (1973~2006)



Understanding SEER database



Understanding SEER database

This directory contains the SEER November 2008 Limited-Use Data files from the San Jose-Monterey, Los Angeles, Rural Georgia and Alaska Natives SEER registries for 1992-2006.

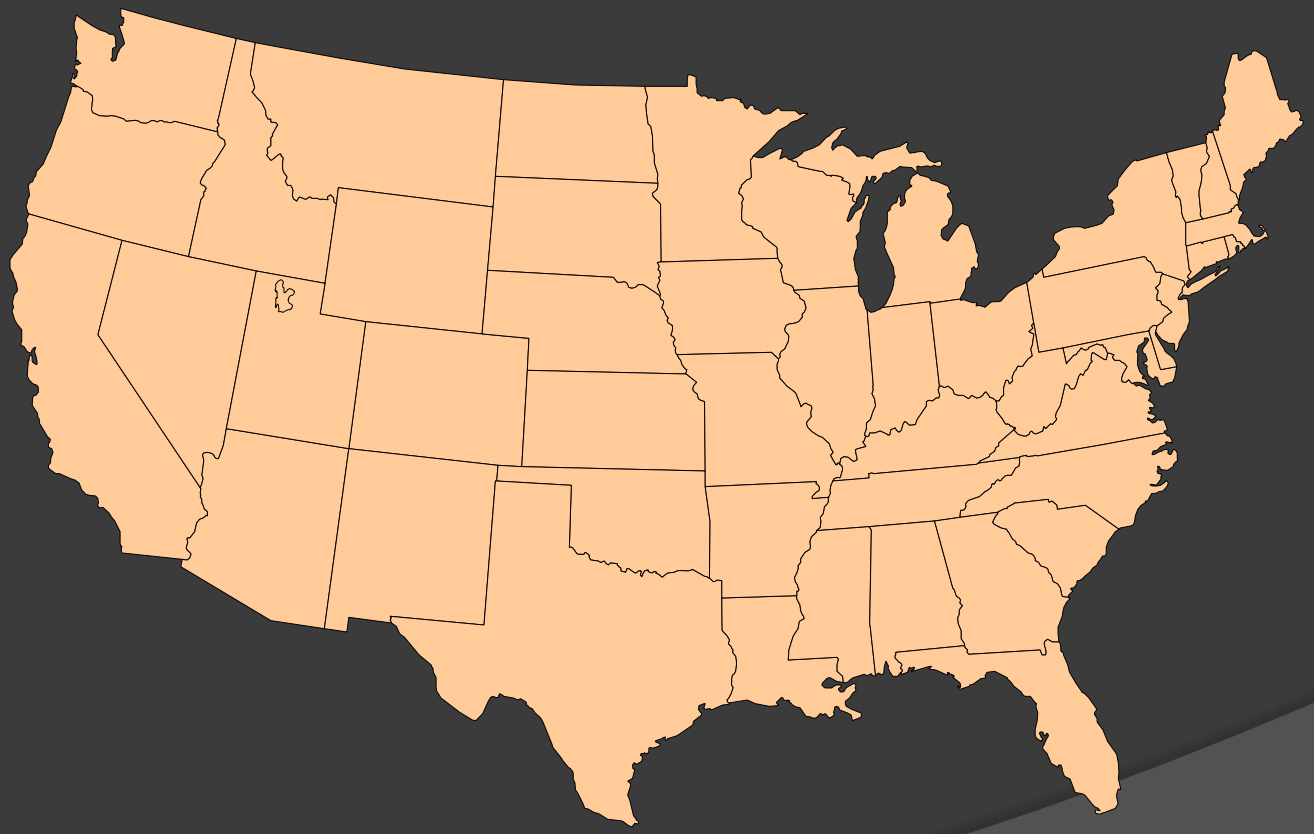
http://www.cancer.org

Understanding SEER database

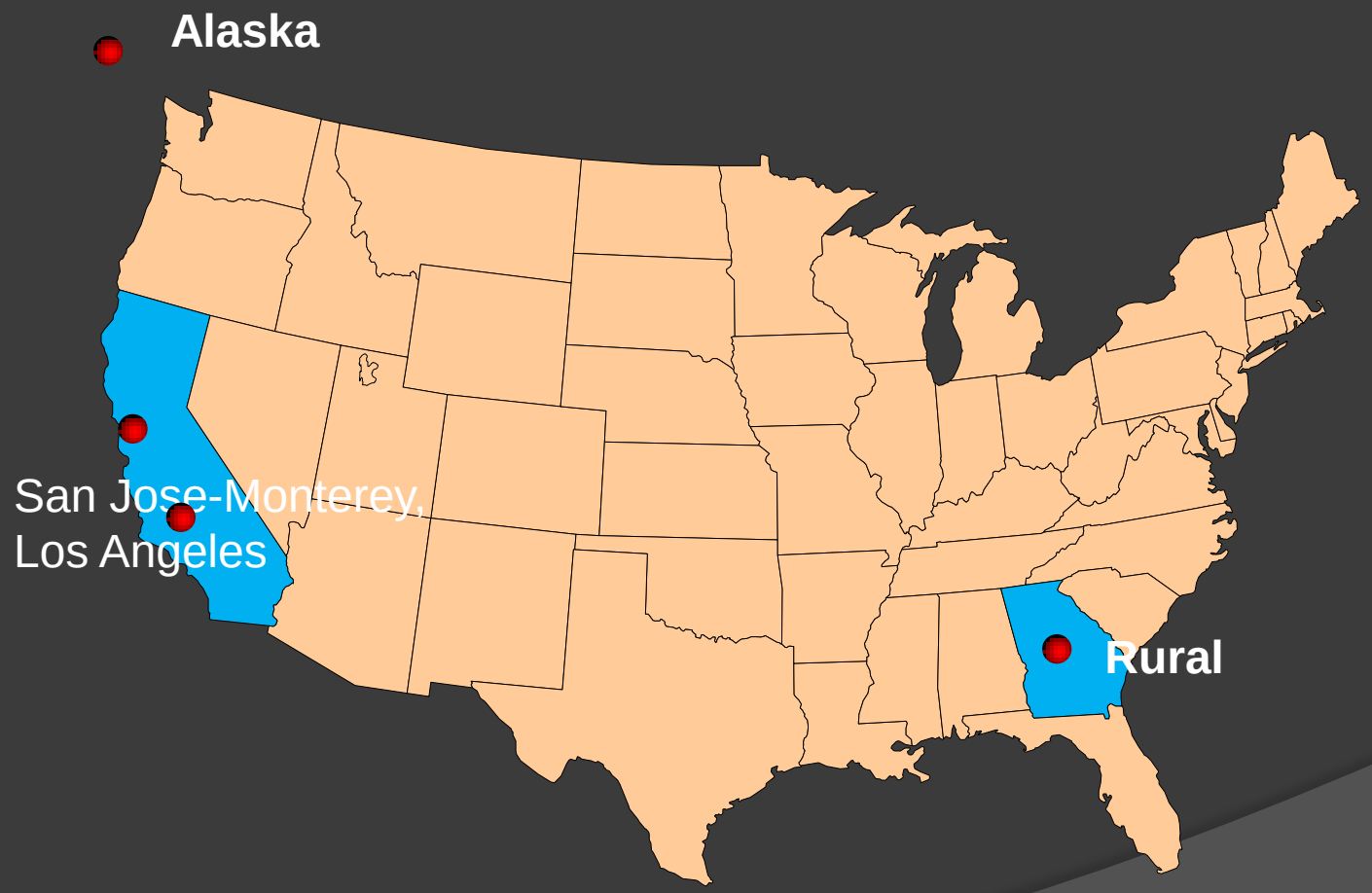
This directory contains the SEER November 2008 Limited-Use Data files from the San Jose-Monterey, Los Angeles, Rural Georgia and Alaska Natives SEER registries for 1992-2006.

http://www.cancer.org

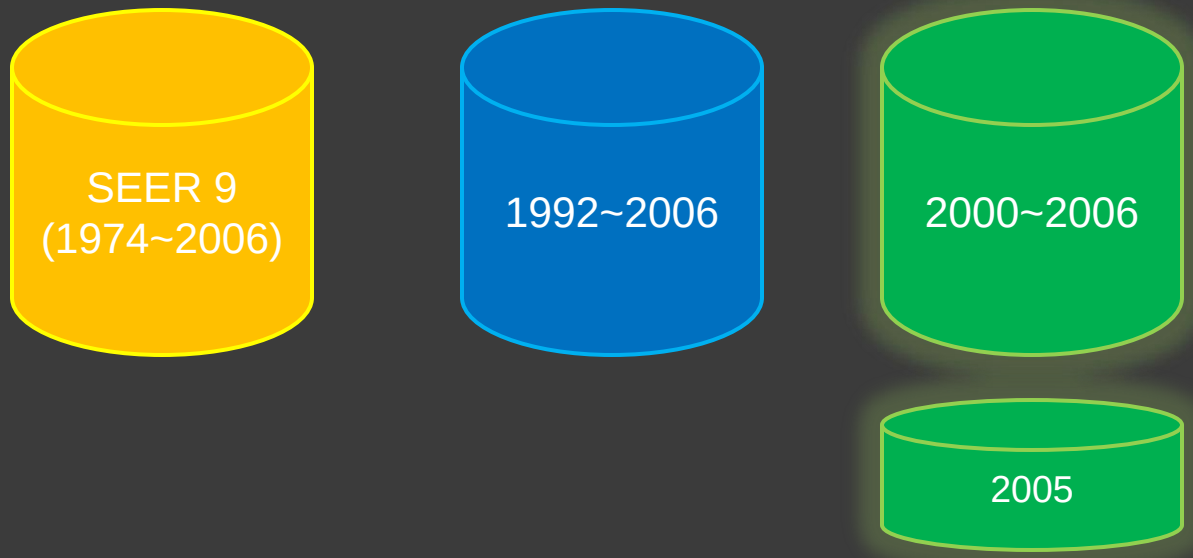
SEER (1992~2006)



SEER (1992~2006)



Understanding SEER database

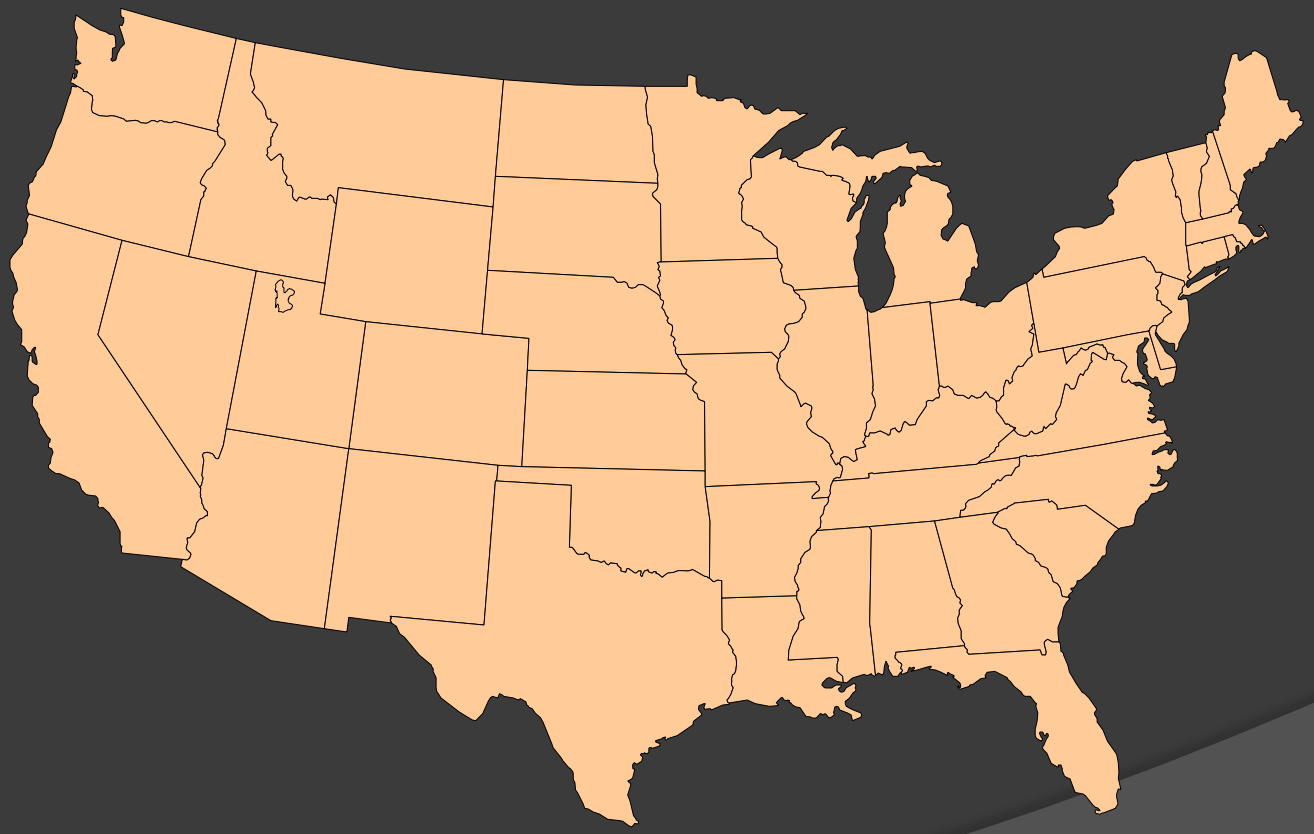


Understanding SEER database

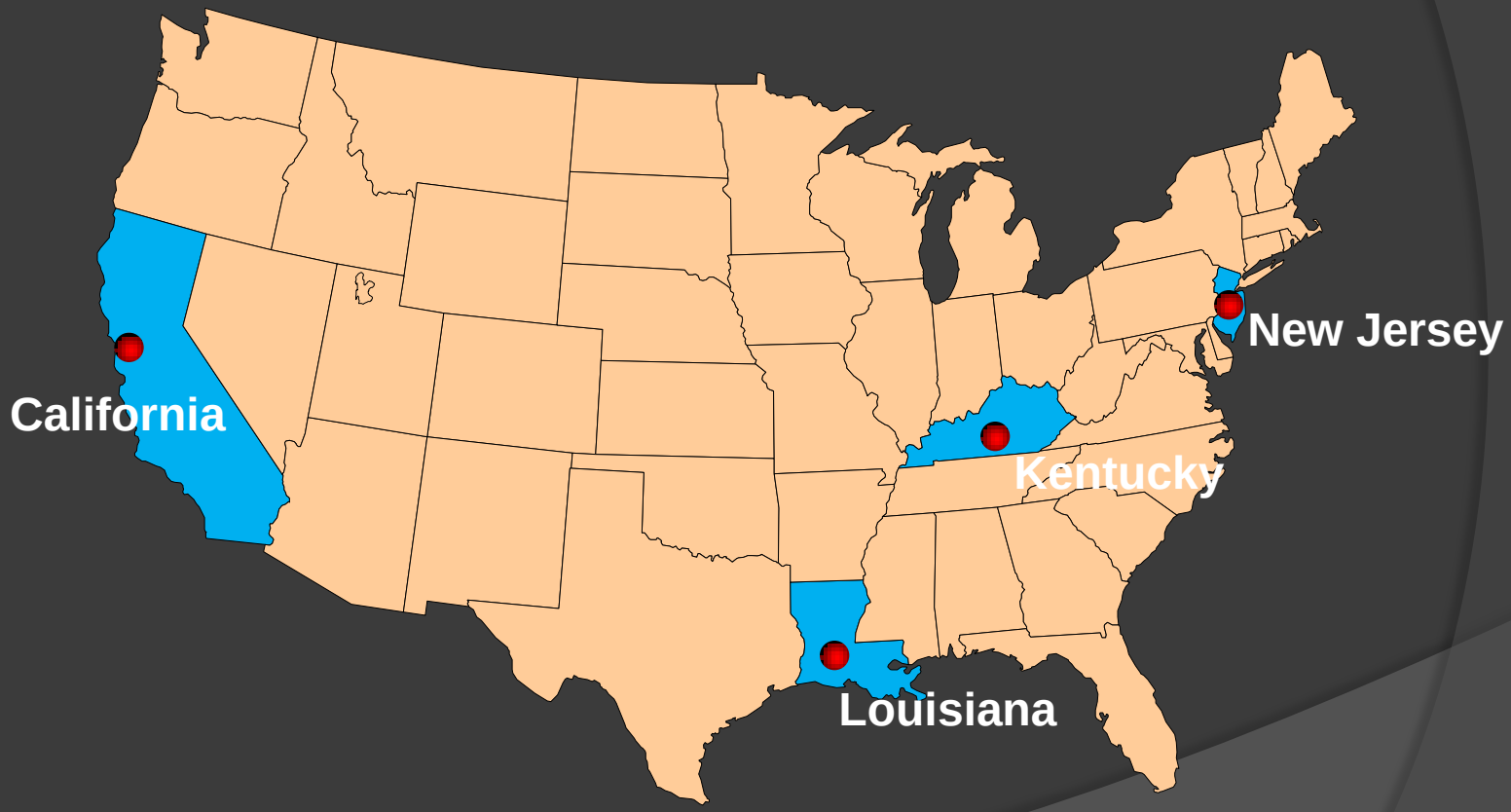
This Data files from the Greater California, Kentucky, Louisiana, and New Jersey SEER registries for 2000-2006. For the year 2006, only January – June. diagnoses are included for Louisiana. Hurricane Katrina had a large impact on Louisiana's population for the July - December 2005 time period. For most SEER reporting, Louisiana cases diagnosed in the latter half of 2005 are not analyzed.

http://www.cancer.org

SEER 9 (1973~2006)



SEER (2000~2006)



Data Dictionary

Item Name	Applicable Years	NAACCR Item #	Positions	Length
Patient ID number		20	01-08	8
Registry ID		40	09-18	10
Marital Status at DX		150	19-19	1
Race/Ethnicity		160	20-21	2
Spanish/Hispanic Origin		190	22-22	1
NHIA Derived Hispanic Origin		191	23-23	1
Sex		220	24-24	1
Age at diagnosis		230	25-27	3
Year of Birth		240	28-31	4
Birth Place		250	32-34	3
Sequence Number--Central		380	35-36	2
Month of diagnosis		390	37-38	2
Year of diagnosis		390	39-42	4

Data Dictionary

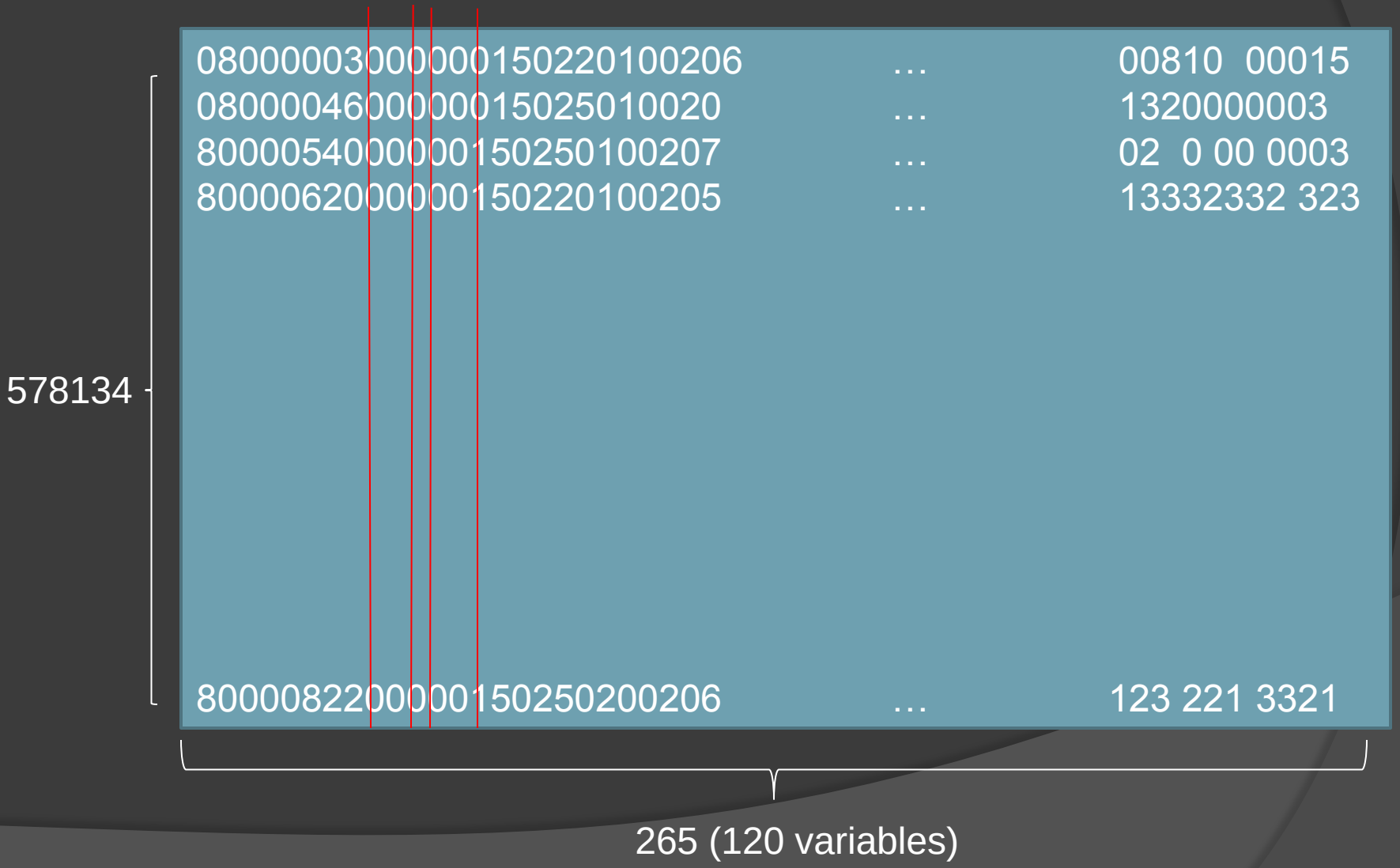
EOD—Tumor Size	1988-2003	780	61-63	3
EOD—Extension	1988-2003	790	64-65	2
EOD—Extension Prost Path	1985-2003	800	66-67	2
EOD—Lymph Node Involv	1988-2003	810	68-68	1

RX Summ—Surg Prim Site	1998+	1290	148-149	2
RX Summ—Scope Reg LN Sur	2003+	1292	150-150	1
RX Summ—Surg Oth Reg/Dis	2003+	1294	151-151	1
RX Summ—Reg LN Examined	1998-2002	1296	152-153	2
RX Summ—Reconstruct 1st	1998-2002	1330	154-154	1
Reason for no surgery		1340	155-155	1
RX Summ—Radiation		1360	156-156	1
RX Summ—Rad to CNS	1988-1997	1370	157-157	1
RX Summ—Surg / Rad Seq		1380	158-158	1
RX Summ—Surgery Type	1973-1997	1640	159-160	2
RX Summ—Surg Site 98-02	1998-2002	1646	161-162	2
RX Summ—Scope Reg 98-02	1998-2002	1647	163-163	1
RX Summ—Surg Oth 98-02	1998-2002	1648	164-164	1

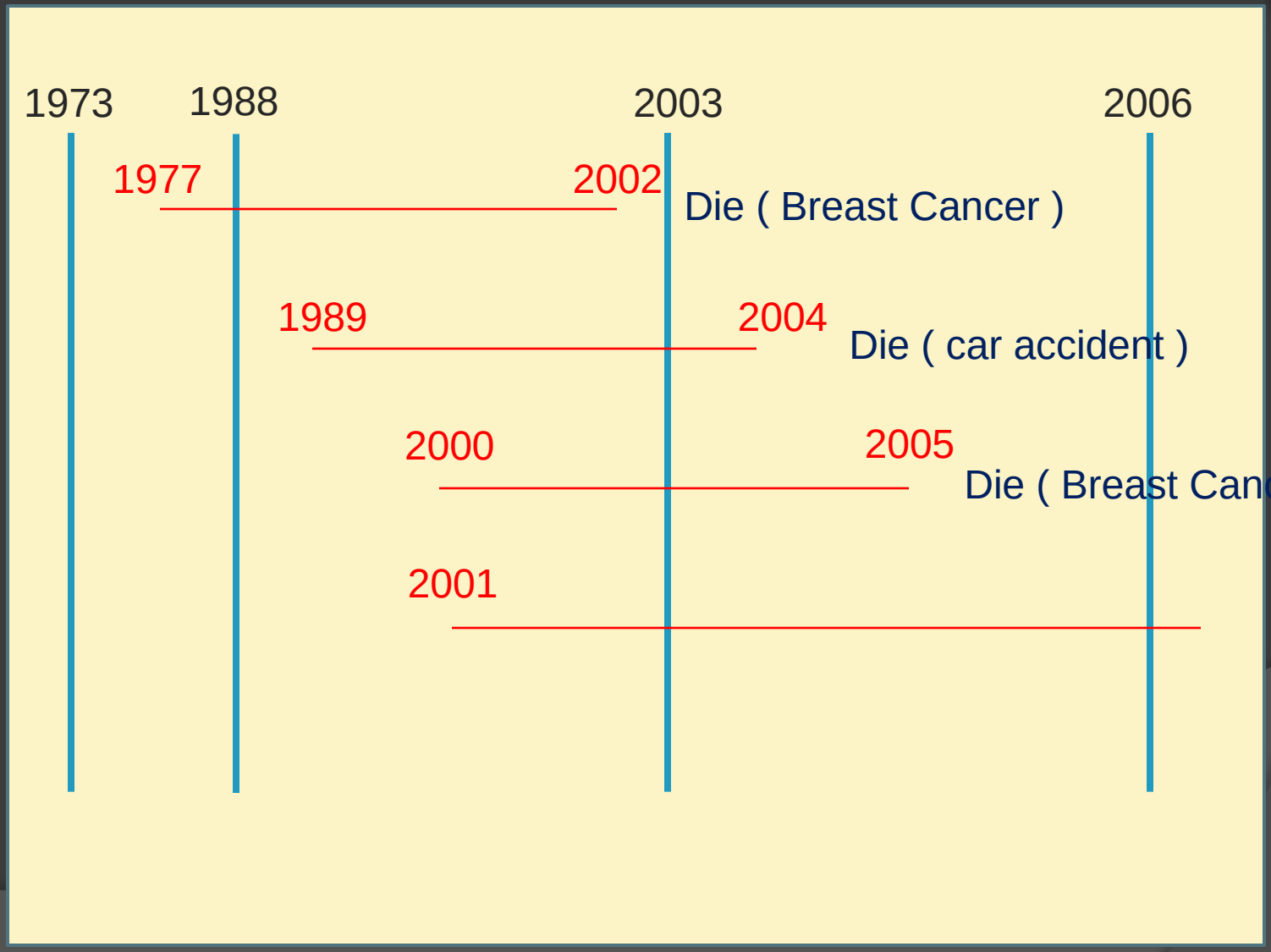
Data Dictionary

Survival time recode		N/A	241-244	4
Cause of Death to SEER site recode		N/A	245-249	5
COD to site rec KM		N/A	250-254	5
Vital Status recode		N/A	255-255	1
Race--NAPIIA		193	256-257	2
IHS Link		192	258-258	1
Summary stage 2000 (1998+)	1998+	N/A	259-259	1
AYA site recode		N/A	260-261	2
Lymphoma subtype recode		N/A	262-263	2
Vital status relative to site at dx		N/A	264-264	1

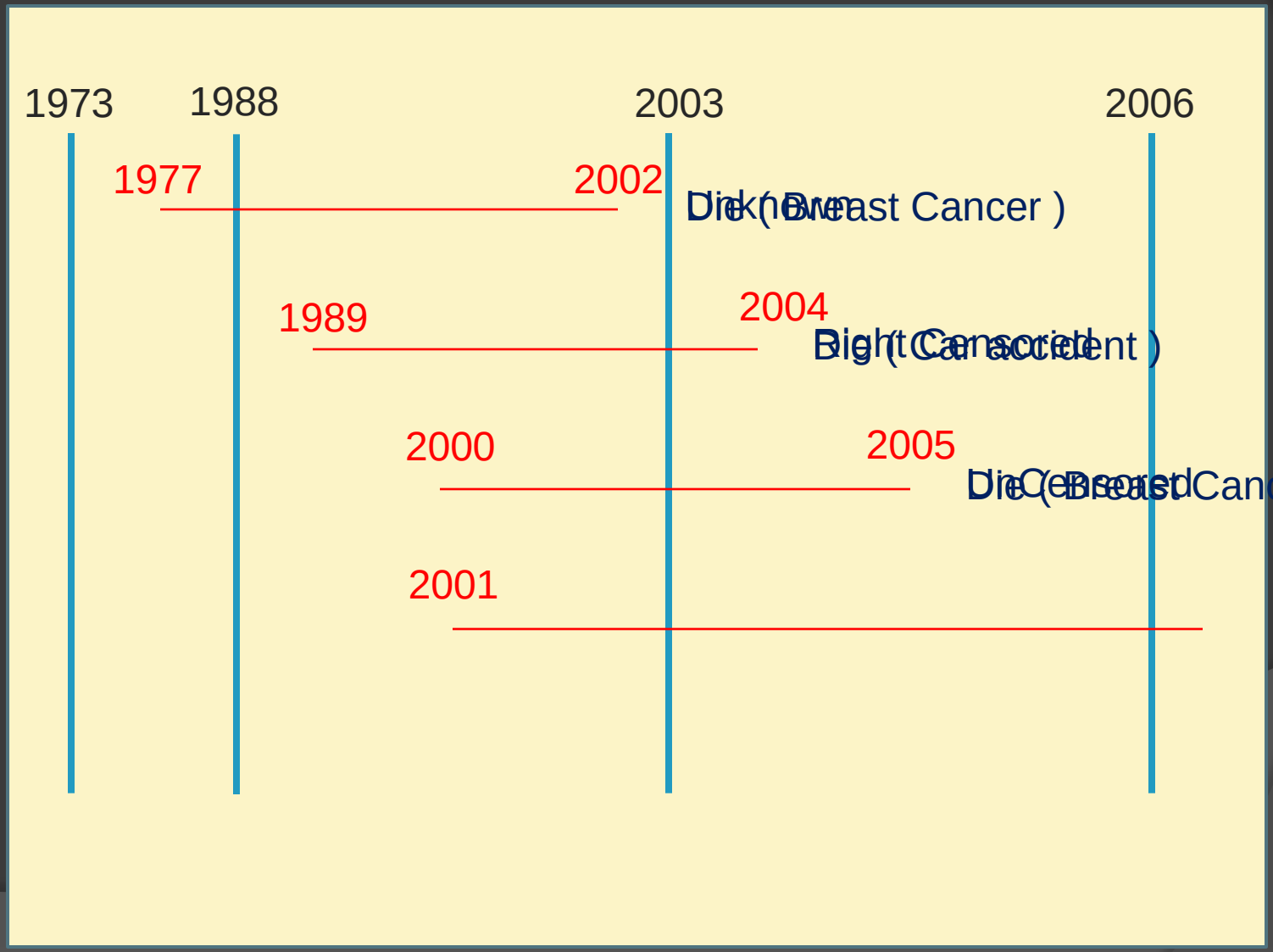
Original Data



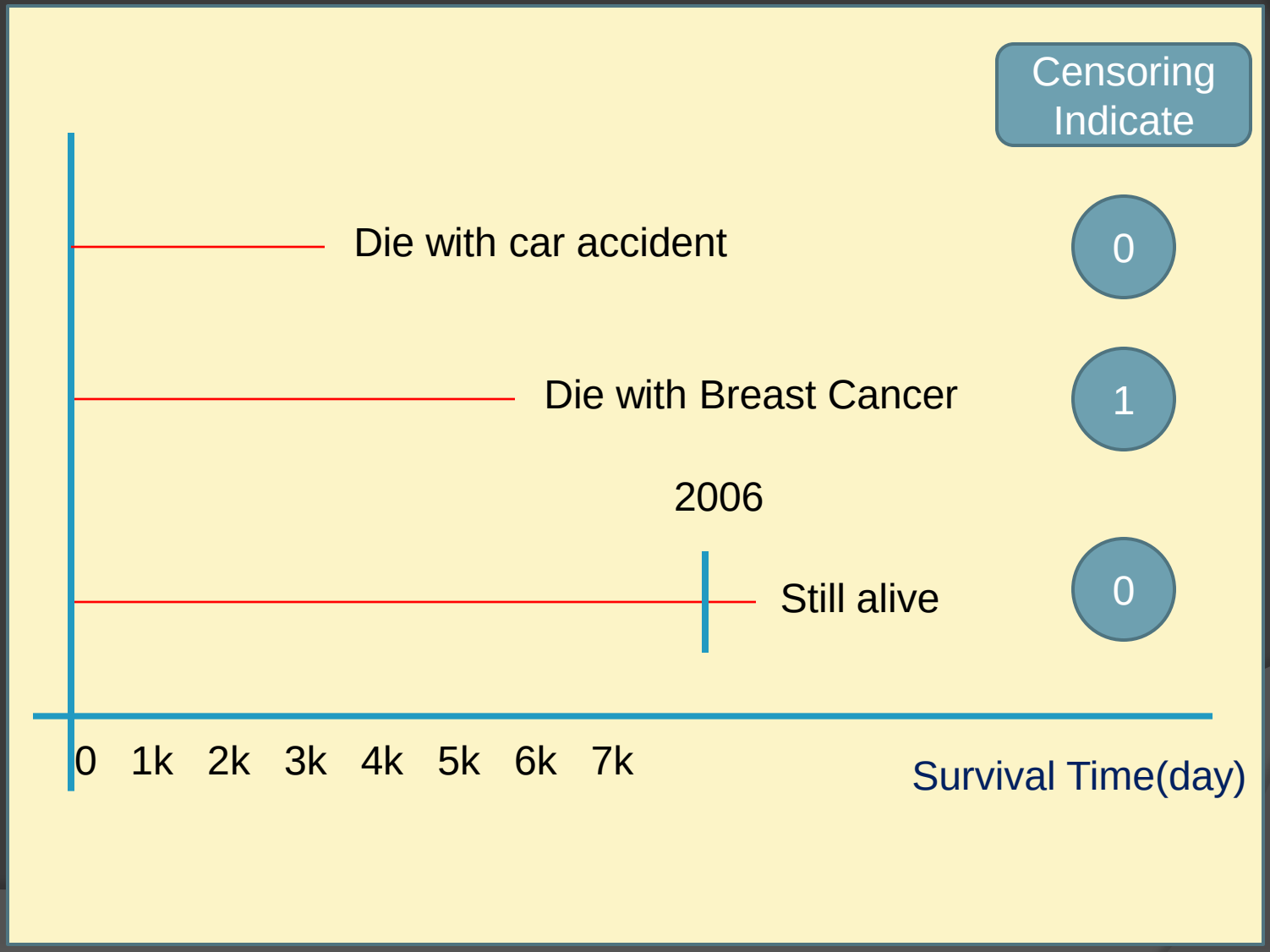
Censoring Indicate



Censoring Indicate



Censoring Indicate



Type of Censoring

Type I censoring :

This type arises in engineering applications. In such situations there are transistors, tubes, chips, etc.; we put them all on test at time t=0 and record their times to failure. Some items may take a long time to “burn out” and we will not want to wait that long to terminate the experiment. Therefore, we terminate the experiment a pre specified time tc. We call tc the fixed censoring time.

Type of Censoring

Type I censoring :

This type arises in engineering applications. In such situations there are transistors, tubes, chips, etc.; we put them all on test at time t=0 and record their times to failure. Some items may take a long time to “burn out” and we will not want to wait that long to terminate the experiment. Therefore, we terminate the experiment a pre specified time tc. We call tc the fixed censoring time.

Type of Censoring

Type II censoring :

In similar engineering applications as above, the censoring time may be left open at the beginning. Instead, the experiment is run until a pre specified fraction r/n of the n items has failed. By plan, observations terminate after the rth failure occurs.

Likelihood function

- Type I Censoring :

$$L = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(t_c)^{1-\delta_i}$$

- Type II Censoring :

$$L = \frac{n!}{(n-r)!} \prod_{i=1}^{r} f(t_{(i)}) S(t_{(r)})^{n-r}$$

Likelihood function

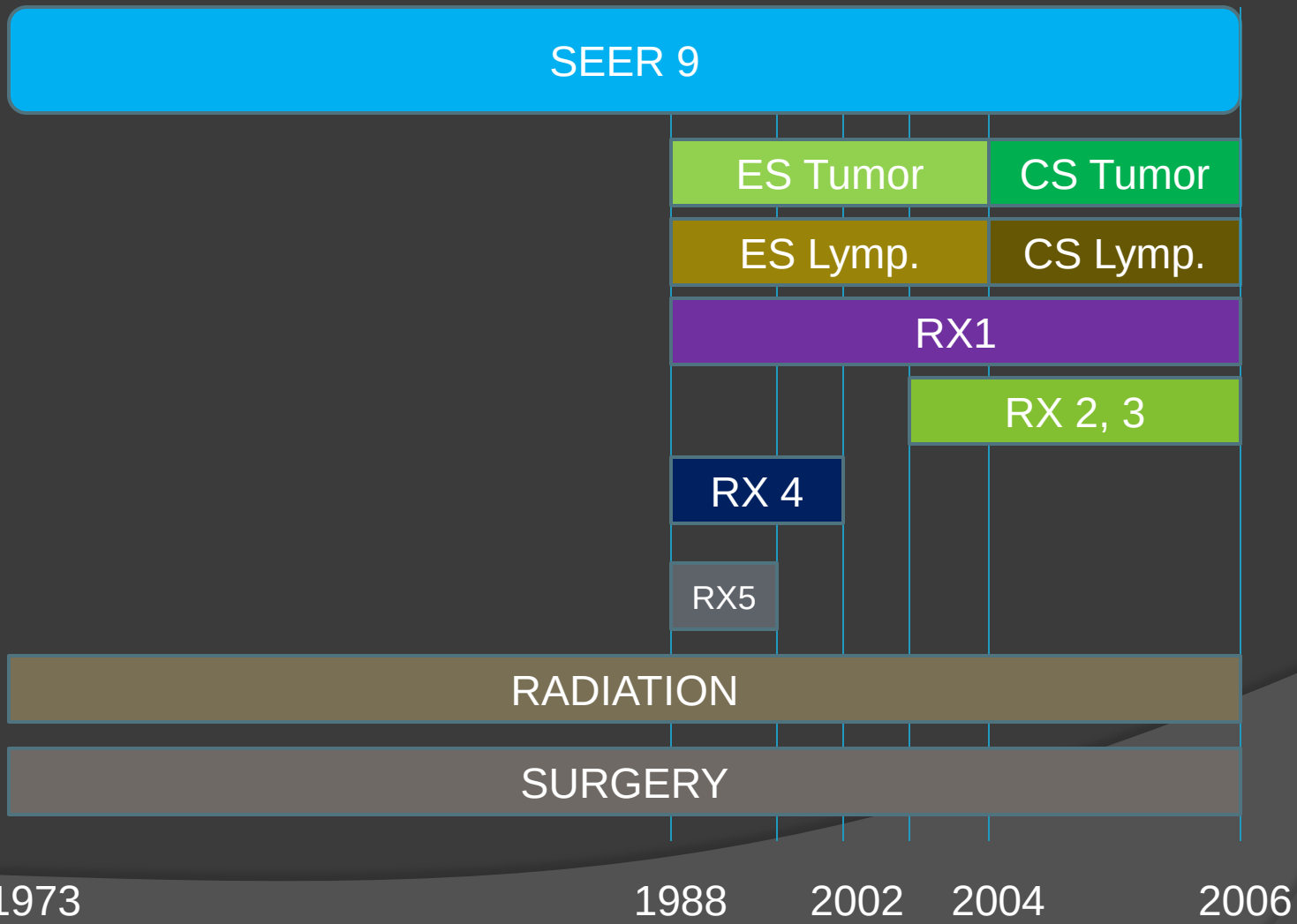
- Type I Censoring :

$$L = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(t_c)^{1-\delta_i}$$

Survival Time~ variables(original)

	id	agedig	yeardig	tumorsize	lymphnode	race	invasion	metastasis	survtime	state	stage	histologic
1	8000003	60	1992	8	0	1	1	0	1120	0	1	1
2	8000046	76	1996	20	0	1	1	0	3100	0	1	1
3	8000054	70	1994	12	0	1	1	0	3010	0	1	1
4	8000062	59	1977	NA	NA	1	NA	NA	1260	0	2	1
5	8000082	61	1986	NA	NA	2	NA	NA	1000	0	0	1
6	8000085	45	1977	NA	NA	2	NA	NA	1291	0	1	1
7	8000090	65	1975	NA	NA	1	NA	NA	5090	0	1	1
8	8000112	78	2005	NA	NA	1	NA	NA	1010	0	0	1
9	8000121	76	2000	31	6	1	0	0	3060	0	2	1
10	8000132	76	1993	21	6	1	1	0	3020	0	2	1
11	8000139	67	1980	NA	NA	1	NA	NA	3261	0	1	1
12	8000148	87	1988	28	9	1	1	0	3081	0	1	1
13	8000168	70	1979	NA	NA	1	NA	NA	5031	0	2	4
14	8000180	68	1974	NA	NA	1	NA	NA	1210	0	2	1
15	8000185	58	1973	NA	NA	1	NA	NA	1010	0	2	1
16	8000188	78	1990	15	6	1	1	0	1161	0	2	1
17	8000204	78	1977	NA	NA	1	NA	NA	9050	1	1	1
18	8000208	88	1999	10	0	1	0	0	1060	0	1	1
19	8000215	72	1973	NA	NA	1	NA	NA	9030	0	2	1
20	8000222	69	2001	12	0	1	0	0	3050	0	1	1

Risk Factors Period



Radiation

- 0 None; diagnosed at autopsy
- 1 Beam radiation
- 2 Radioactive implants
- 3 Radioisotopes
- 4 Combination of 1 with 2 or 3
- 5 Radiation, NOS – method or source not specified
- 6 Other radiation (1973-1987 cases only)
- 7 Patient or patient‘s guardian refused radiation therapy
- 8 Radiation recommended, unknown if administered
- 9 Unknown if radiation administered

Radiation Sequence with Surgery

- 0 No radiation and/or surgery as defined above
- 2 Radiation before surgery
- 3 Radiation after surgery
- 4 Radiation both before and after surgery
- 5 Intraoperative radiation therapy
- 6 Intraoperative radiation with other radiation given before or after surgery
- 9 Sequence unknown, but both surgery and radiation were given

Survival Time~ variables(Finial)

id	agedig	yeardig	tumorsize	lymphnode	race	invasion	metastasis	survtime	state	stage	histologic	rx6f
8000003	60	1992	8	0	1	1	0	1120	0	100	1	0
8000046	76	1996	20	0	1	1	0	3100	0	100	1	1
8000054	70	1994	12	0	1	1	0	3010	0	100	1	0
8000121	76	2000	31	6	1	0	0	3060	0	200	1	0
8000132	76	1993	21	6	1	1	0	3020	0	200	1	0
8000148	87	1988	28	9	1	1	0	3081	0	100	1	1
8000188	78	1990	15	6	1	1	0	1161	0	200	1	0
8000208	88	1999	10	0	1	0	0	1060	0	100	1	1
8000222	69	2001	12	0	1	0	0	3050	0	100	1	1
8000291	76	1991	35	0	1	0	0	1061	0	200	1	0
8000302	79	1989	15	0	1	1	0	9071	0	100	1	0
8000423	79	1992	20	0	1	1	0	1080	0	100	1	0
8000451	55	2000	12	0	1	0	0	1010	0	100	3	0
8000477	68	1998	8	0	2	0	0	5040	0	100	1	1
8000482	80	1998	10	0	1	0	0	9060	0	100	1	0
8000501	50	1998	55	4	2	0	1	9011	1	500	1	1
8000521	72	1988	20	6	1	0	0	7110	0	200	1	0
8000523	65	1995	10	0	1	1	0	9110	0	100	1	1
8000542	81	1995	1	0	1	0	0	1090	0	500	1	0
8000542	81	1995	15	0	1	1	0	1090	0	100	1	0

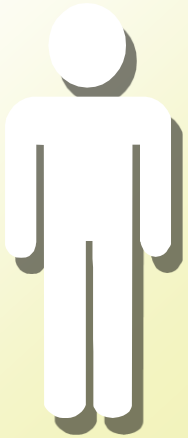
SEER 9 BREAST CANCER

USF CANCER RESEARCH TEAM

SEER 9 Breast Cancer by Race (1988~2006)

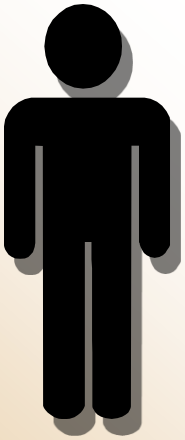
White

496,153



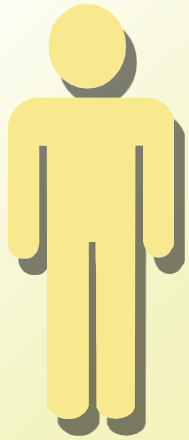
African American

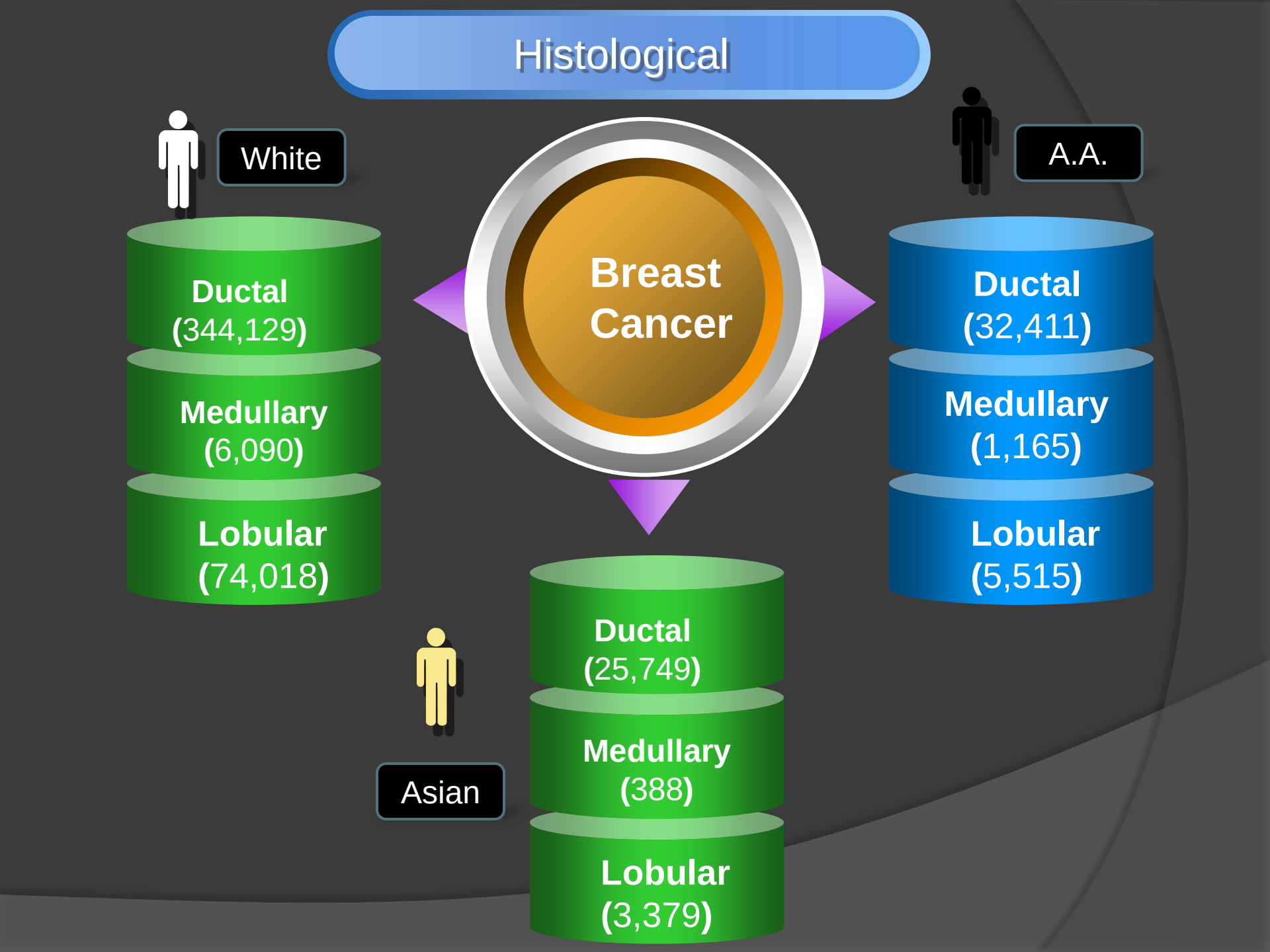
17,207



Asian

33,434

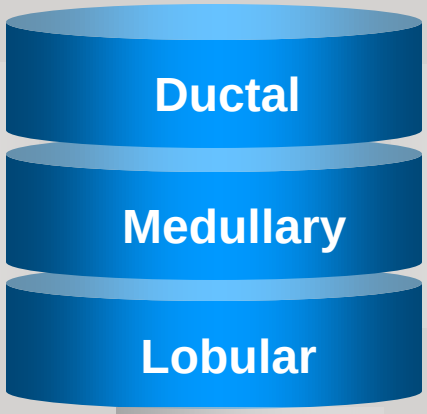



Histological
White
Ductal
(344,129)
Medullary
(6,090)
Lobular
(74,018)
Breast
Cancer
A.A.
Ductal
(32,411)
Medullary
(1,165)
Lobular
(5,515)
Asian
Ductal
(25,749)
Medullary
(388)
Lobular
(3,379)

Stage

White

- Stage 1 (34,772)
- Stage 2 (61,957)
- Stage 3 (0)
- Stage 4 (1)

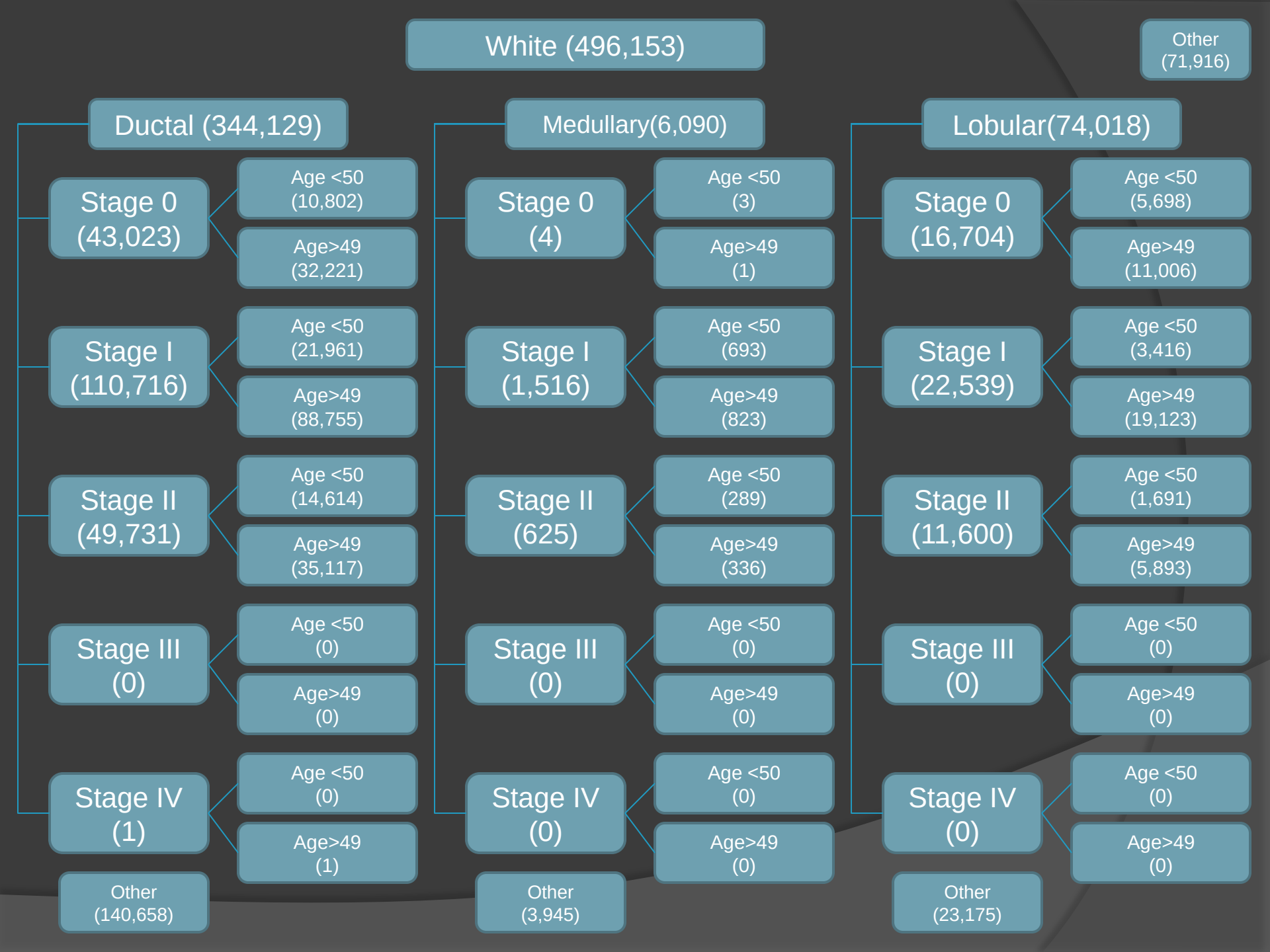


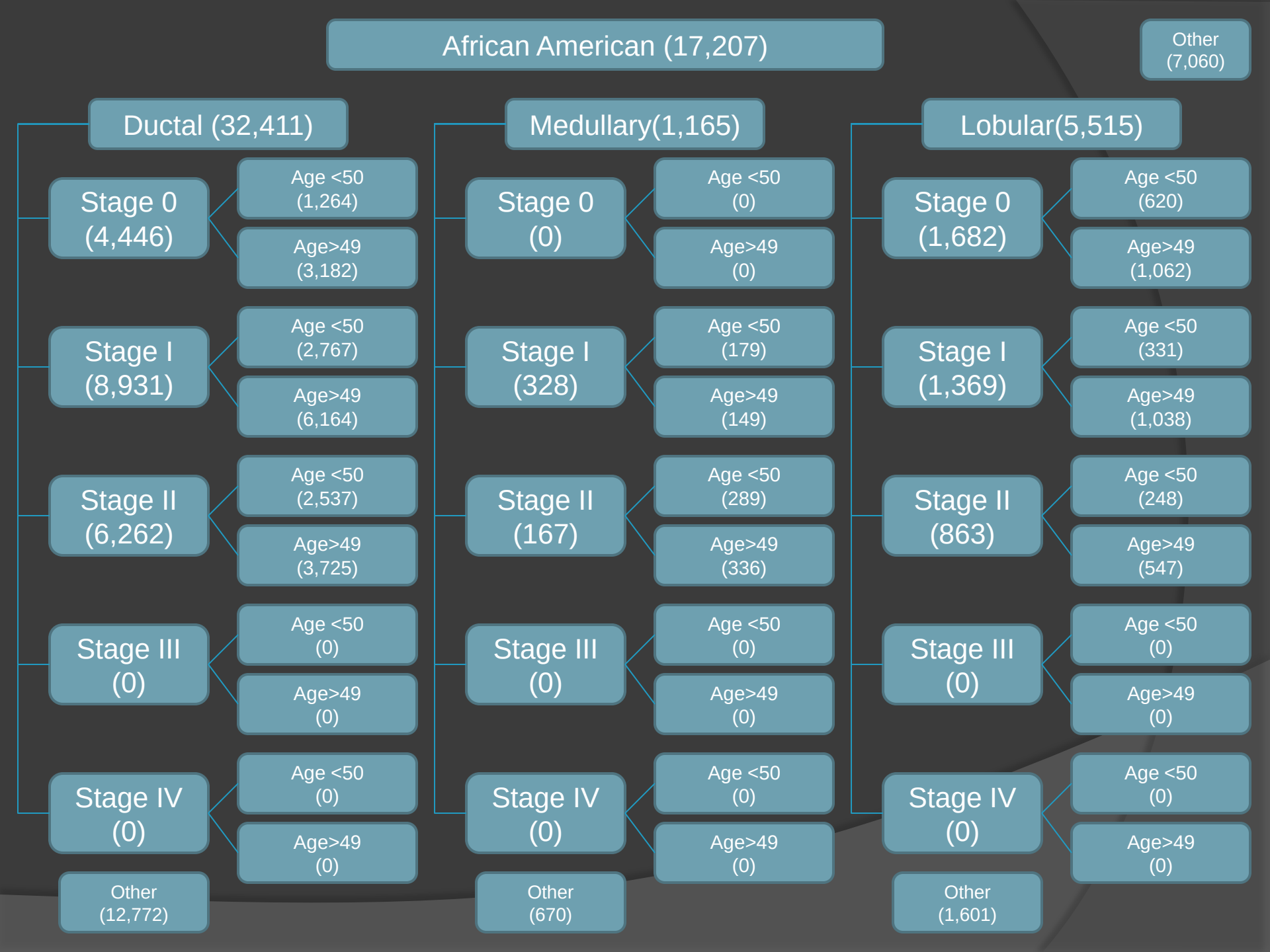
African American

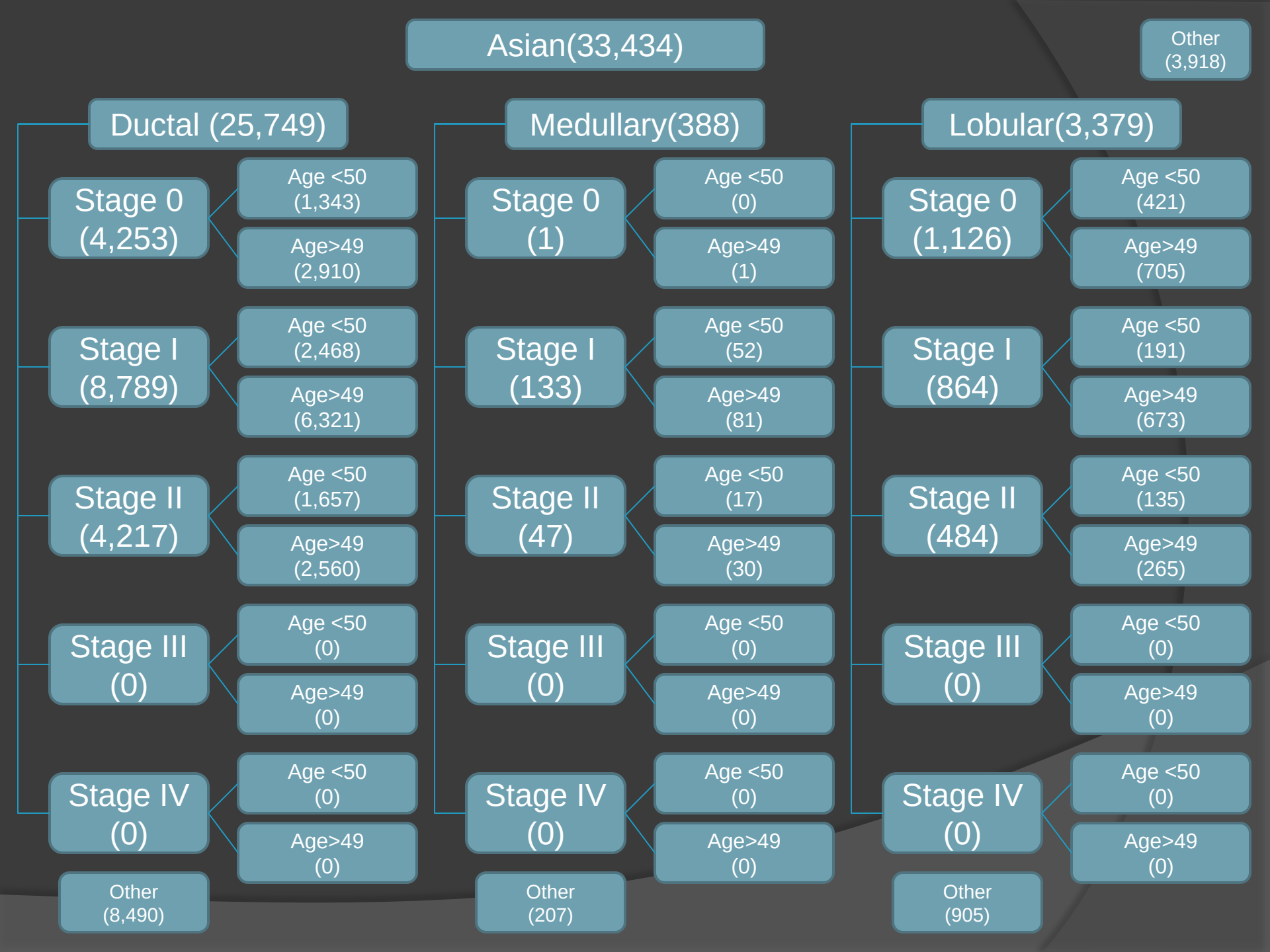
- Stage 1 (10,628)
- Stage 2 (7,292
- Stage 3 (0)
- Stage 4 (0)

Asian

- Stage 1 (34,772)
- Stage 2 (61,957)
- Stage 3 (0)
- Stage 4 (1)


White (496,153)
Other (71,916)
Ductal (344,129)
Medullary(6,090)
Lobular(74,018)
Stage 0 (43,023)
Age <50 (10,802)
Age>49 (32,221)
Stage I (110,716)
Age <50 (21,961)
Age>49 (88,755)
Stage II (49,731)
Age <50 (14,614)
Age>49 (35,117)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (1)
Age <50 (0)
Age>49 (1)
Other (140,658)
Stage 0 (4)
Age <50 (3)
Age>49 (1)
Stage I (1,516)
Age <50 (693)
Age>49 (823)
Stage II (625)
Age <50 (289)
Age>49 (336)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (3,945)
Stage 0 (16,704)
Age <50 (5,698)
Age>49 (11,006)
Stage I (22,539)
Age <50 (3,416)
Age>49 (19,123)
Stage II (11,600)
Age <50 (1,691)
Age>49 (5,893)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (23,175)


African American (17,207)
Other (7,060)
Ductal (32,411)
Medullary(1,165)
Lobular(5,515)
Stage 0 (4,446)
Age <50 (1,264)
Age>49 (3,182)
Stage I (8,931)
Age <50 (2,767)
Age>49 (6,164)
Stage II (6,262)
Age <50 (2,537)
Age>49 (3,725)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (12,772)
Stage 0 (0)
Age <50 (0)
Age>49 (0)
Stage I (328)
Age <50 (179)
Age>49 (149)
Stage II (167)
Age <50 (289)
Age>49 (336)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (670)
Stage 0 (1,682)
Age <50 (620)
Age>49 (1,062)
Stage I (1,369)
Age <50 (331)
Age>49 (1,038)
Stage II (863)
Age <50 (248)
Age>49 (547)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (1,601)

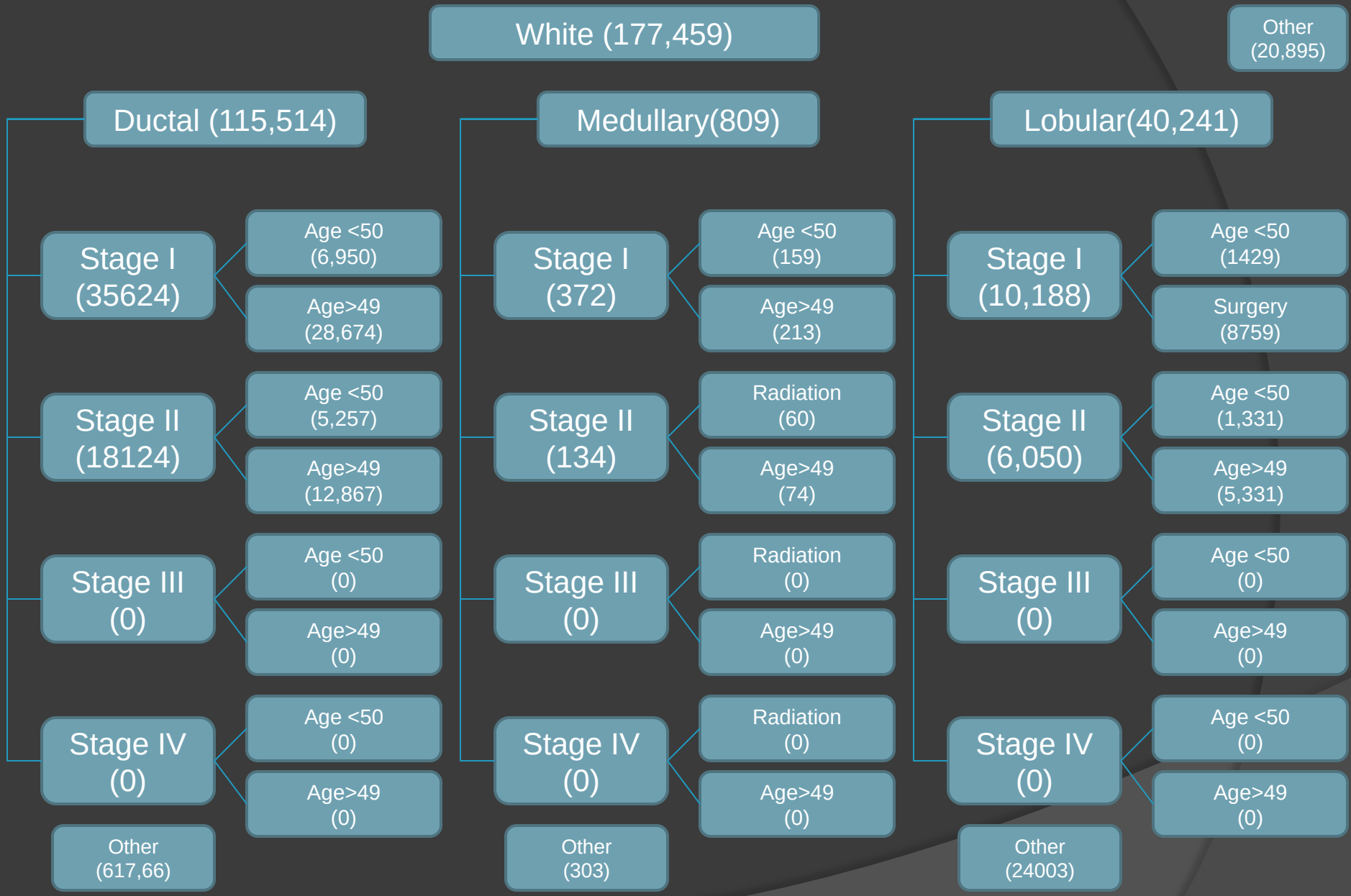

Asian(33,434)
Other (3,918)
Ductal (25,749)
Medullary(388)
Lobular(3,379)
Stage 0 (4,253)
Age <50 (1,343)
Age>49 (2,910)
Stage I (8,789)
Age <50 (2,468)
Age>49 (6,321)
Stage II (4,217)
Age <50 (1,657)
Age>49 (2,560)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (8,490)
Stage 0 (1)
Age <50 (0)
Age>49 (1)
Stage I (133)
Age <50 (52)
Age>49 (81)
Stage II (47)
Age <50 (17)
Age>49 (30)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (207)
Stage 0 (1,126)
Age <50 (421)
Age>49 (705)
Stage I (864)
Age <50 (191)
Age>49 (673)
Stage II (484)
Age <50 (135)
Age>49 (265)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (905)

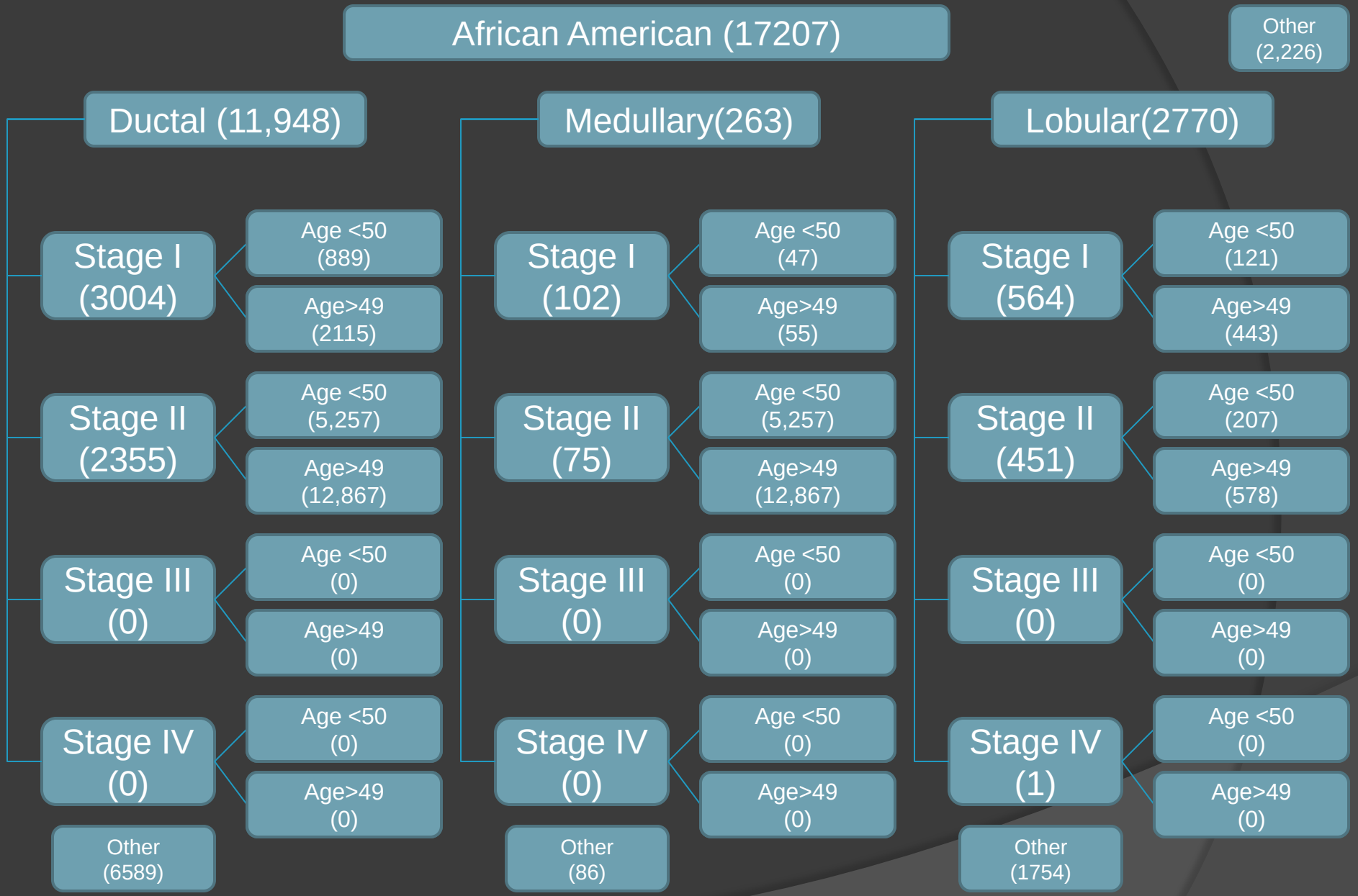
SEER 2000-2006 BREAST CANCER

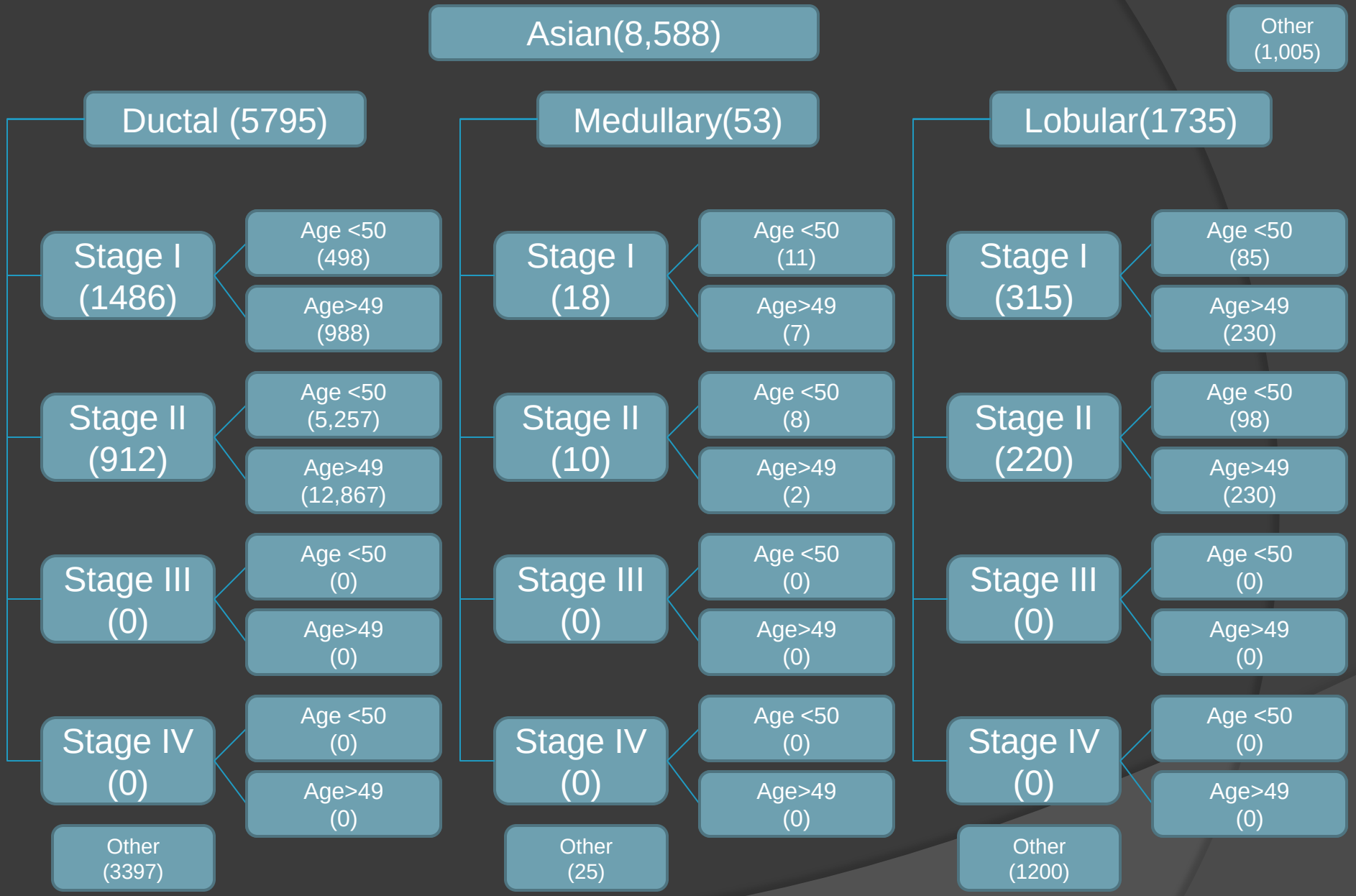
USF CANCER RESEARCH TEAM

SEER 2000-2006

- This directory contains the SEER November 2008 Limited-Use Data files from the Greater California, Kentucky, Louisiana, and New Jersey SEER registries for 2000-2006. For the year 2006, only January – June diagnoses are included for Louisiana. Hurricane Katrina had a large impact on Louisiana's population for the July - December 2005 time period. For most SEER reporting, Louisiana cases diagnosed in the latter half of 2005 are not analyzed.


White (177,459)
Other (20,895)
Ductal (115,514)
Medullary(809)
Lobular(40,241)
Stage I (35624)
Age <50 (6,950)
Age>49 (28,674)
Stage II (18124)
Age <50 (5,257)
Age>49 (12,867)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (617,66)
Stage I (372)
Age <50 (159)
Age>49 (213)
Stage II (134)
Radiation (60)
Age>49 (74)
Stage III (0)
Radiation (0)
Age>49 (0)
Stage IV (0)
Radiation (0)
Age>49 (0)
Other (303)
Stage I (10,188)
Age <50 (1429)
Surgery (8759)
Stage II (6,050)
Age <50 (1,331)
Age>49 (5,331)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (24003)

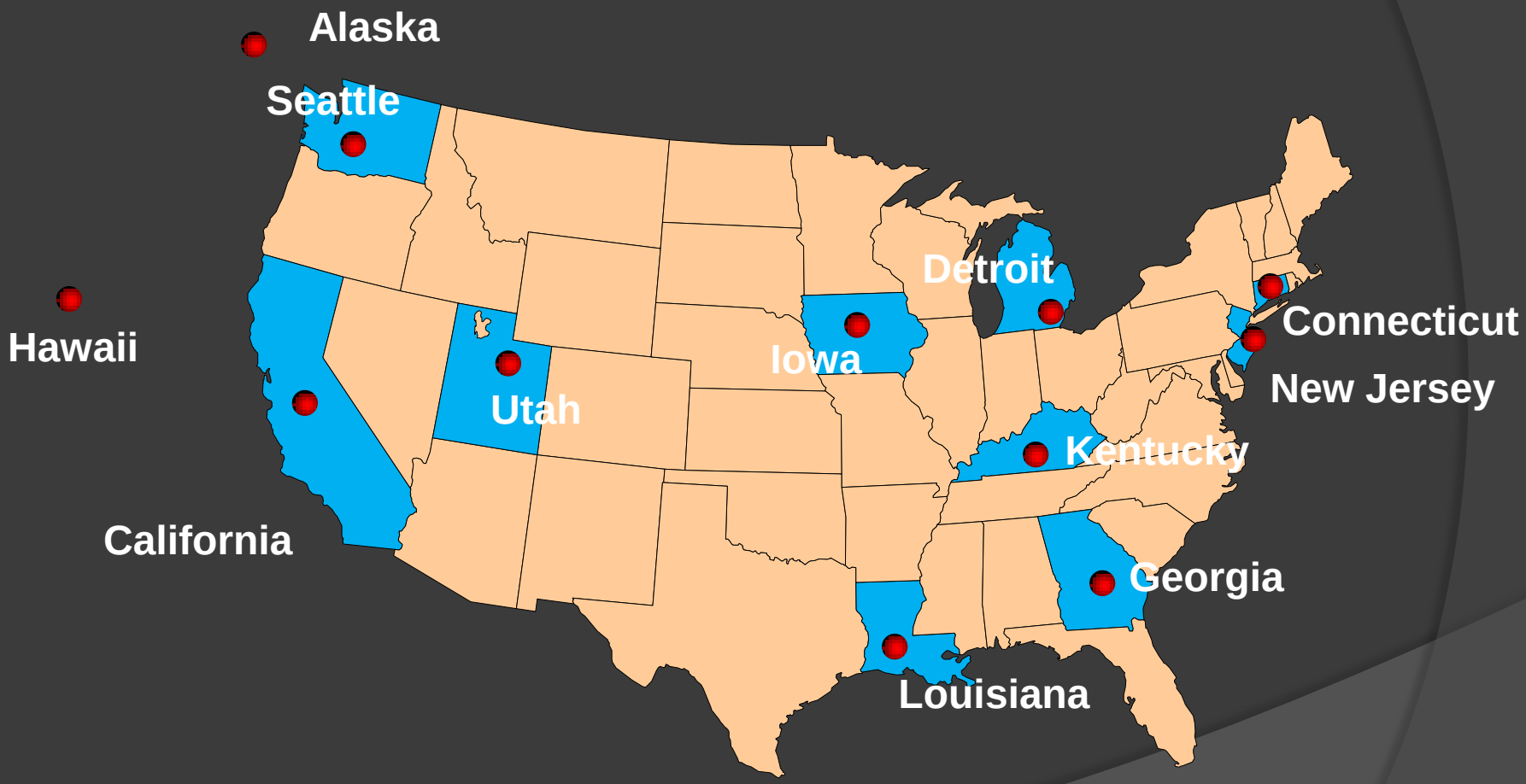

African American (17207)
Other (2,226)
Ductal (11,948)
Medullary(263)
Lobular(2770)
Stage I (3004)
Age <50 (889)
Age>49 (2115)
Stage II (2355)
Age <50 (5,257)
Age>49 (12,867)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (6589)
Stage I (102)
Age <50 (47)
Age>49 (55)
Stage II (75)
Age <50 (5,257)
Age>49 (12,867)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (86)
Stage I (564)
Age <50 (121)
Age>49 (443)
Stage II (451)
Age <50 (207)
Age>49 (578)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (1)
Age <50 (0)
Age>49 (0)
Other (1754)


Asian(8,588)
Other (1,005)
Ductal (5795)
Medullary(53)
Lobular(1735)
Stage I (1486)
Age <50 (498)
Age>49 (988)
Stage II (912)
Age <50 (5,257)
Age>49 (12,867)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (3397)
Stage I (18)
Age <50 (11)
Age>49 (7)
Stage II (10)
Age <50 (8)
Age>49 (2)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (25)
Stage I (315)
Age <50 (85)
Age>49 (230)
Stage II (220)
Age <50 (98)
Age>49 (230)
Stage III (0)
Age <50 (0)
Age>49 (0)
Stage IV (0)
Age <50 (0)
Age>49 (0)
Other (1200)

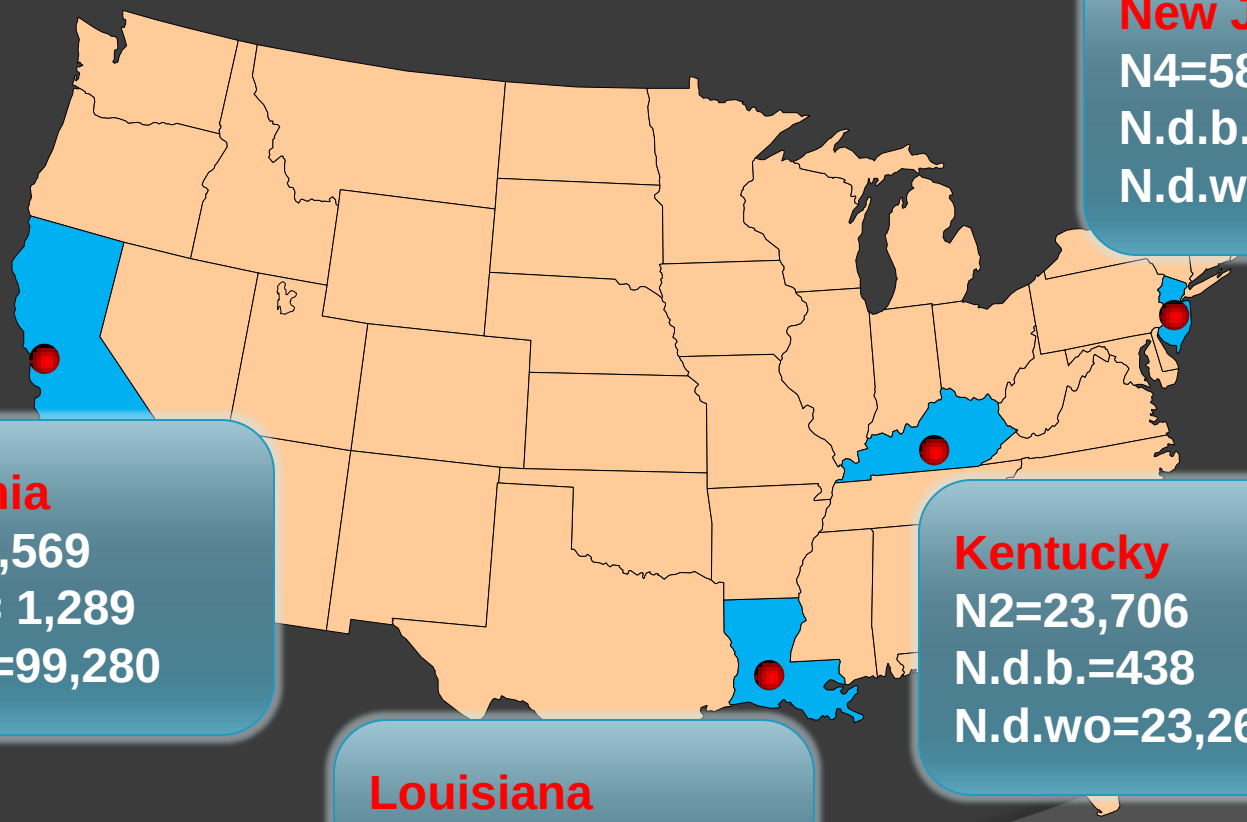
PROPOSE REGIONAL ANALYSIS OF BREAST CANCER

USF CANCER RESEARCH TEAM

SEER (10 States Information)



SEER (2000~2006)



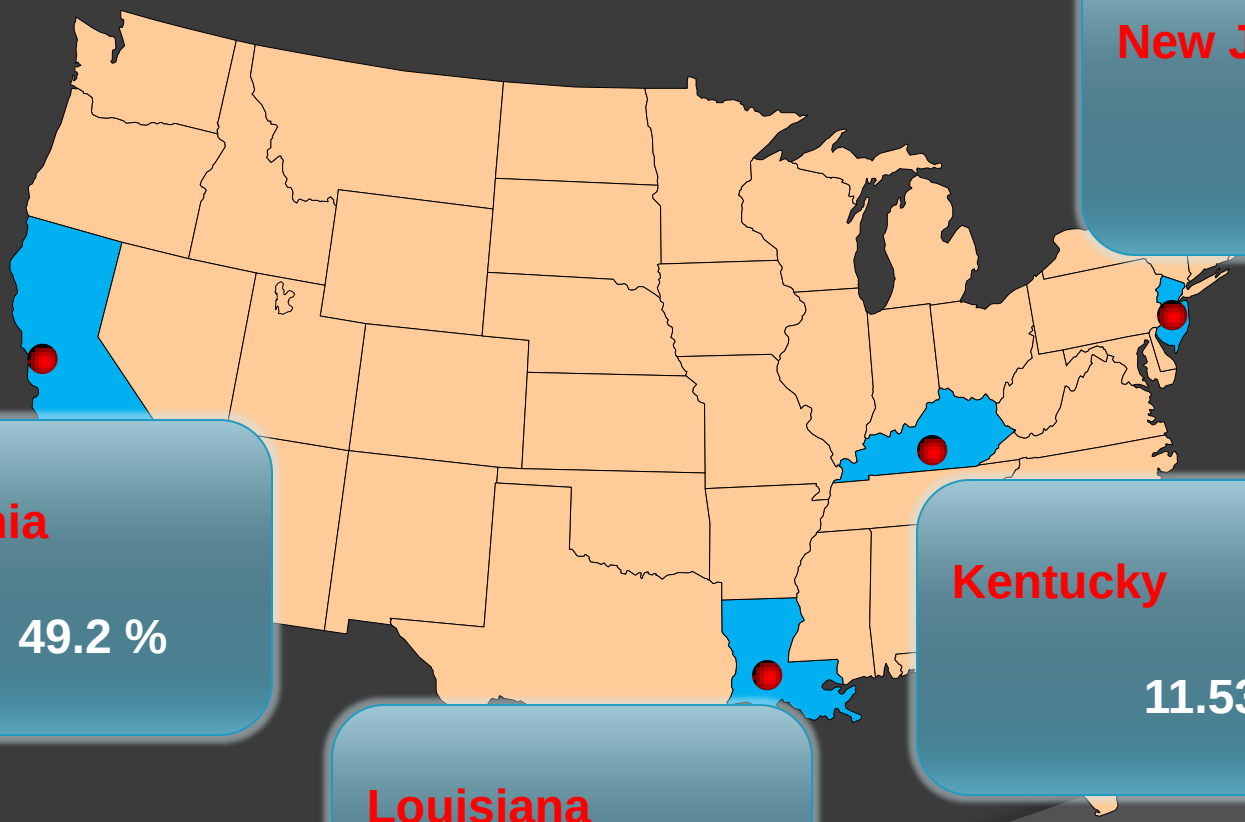
New Jersey
N4=58,242
N.d.b.=1,015
N.d.wo=57,227

California
N1=100,569
N.d.b. = 1,289
N.d.wo=99,280

Kentucky
N2=23,706
N.d.b.=438
N.d.wo=23,268

Louisiana
N3=22,432
N.d.b.=446
N.d.wo=21,986

Death Rates with Other Reasons



New Jersey

28.36%

California

49.2 %

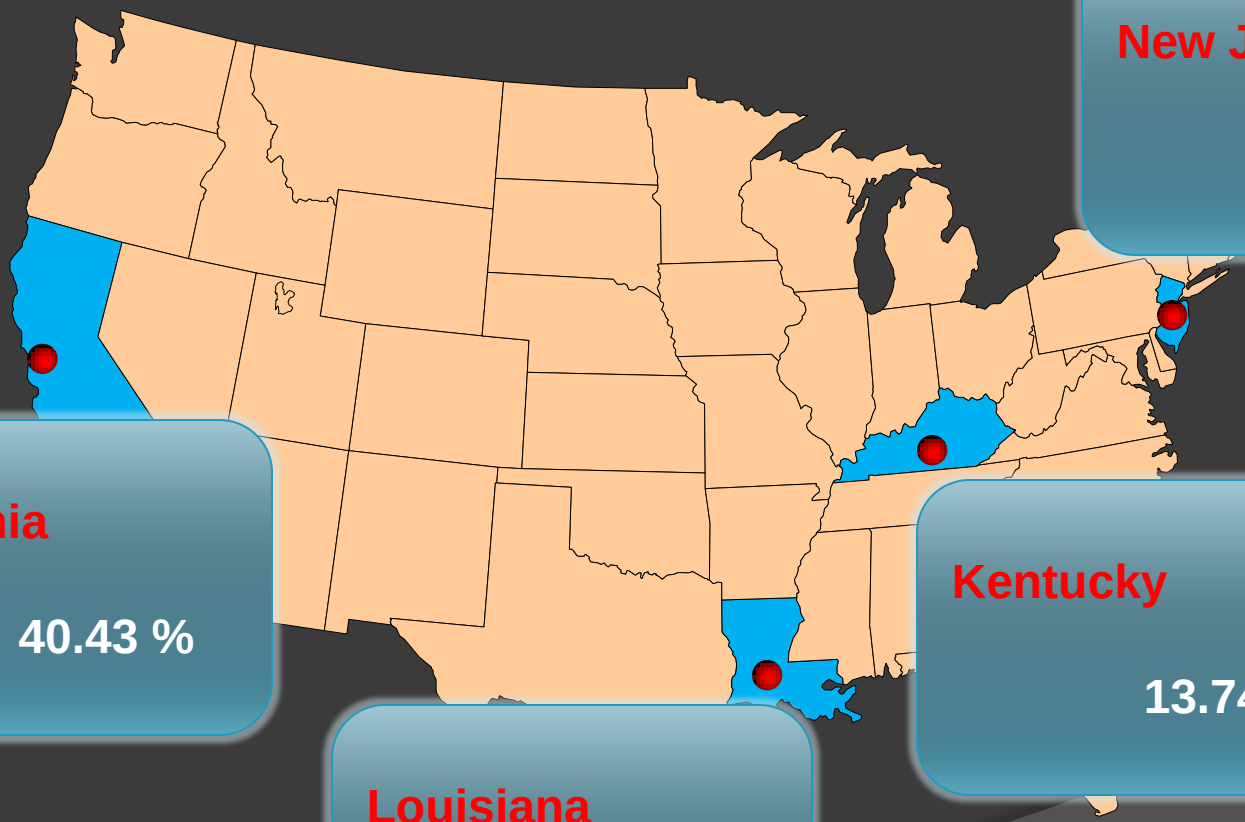
Kentucky

11.53%

Louisiana

10.9%

Death Rates with Breast Cancer



New Jersey

31.84%

California

40.43 %

Kentucky

13.74%

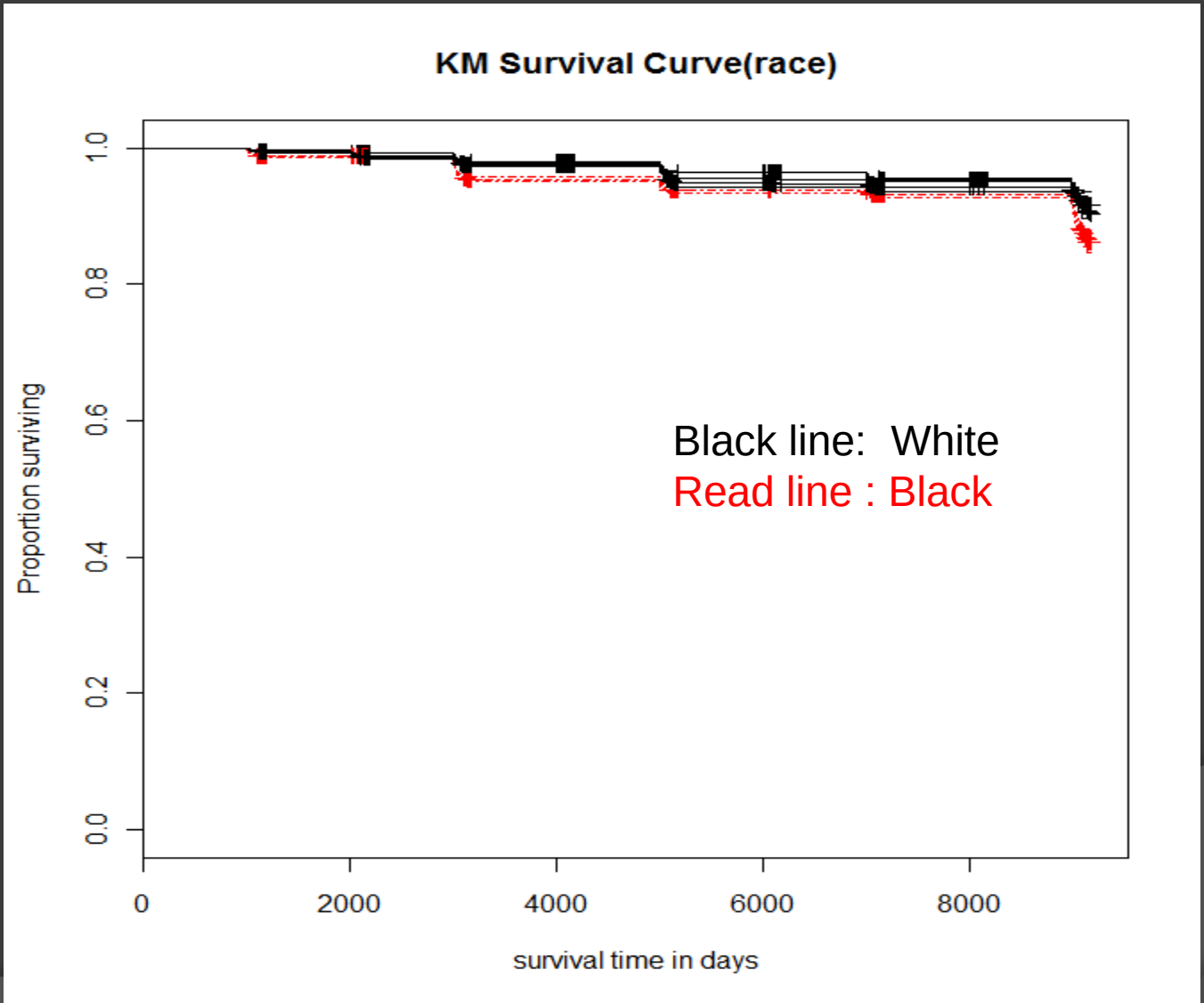
Louisiana

13.99%

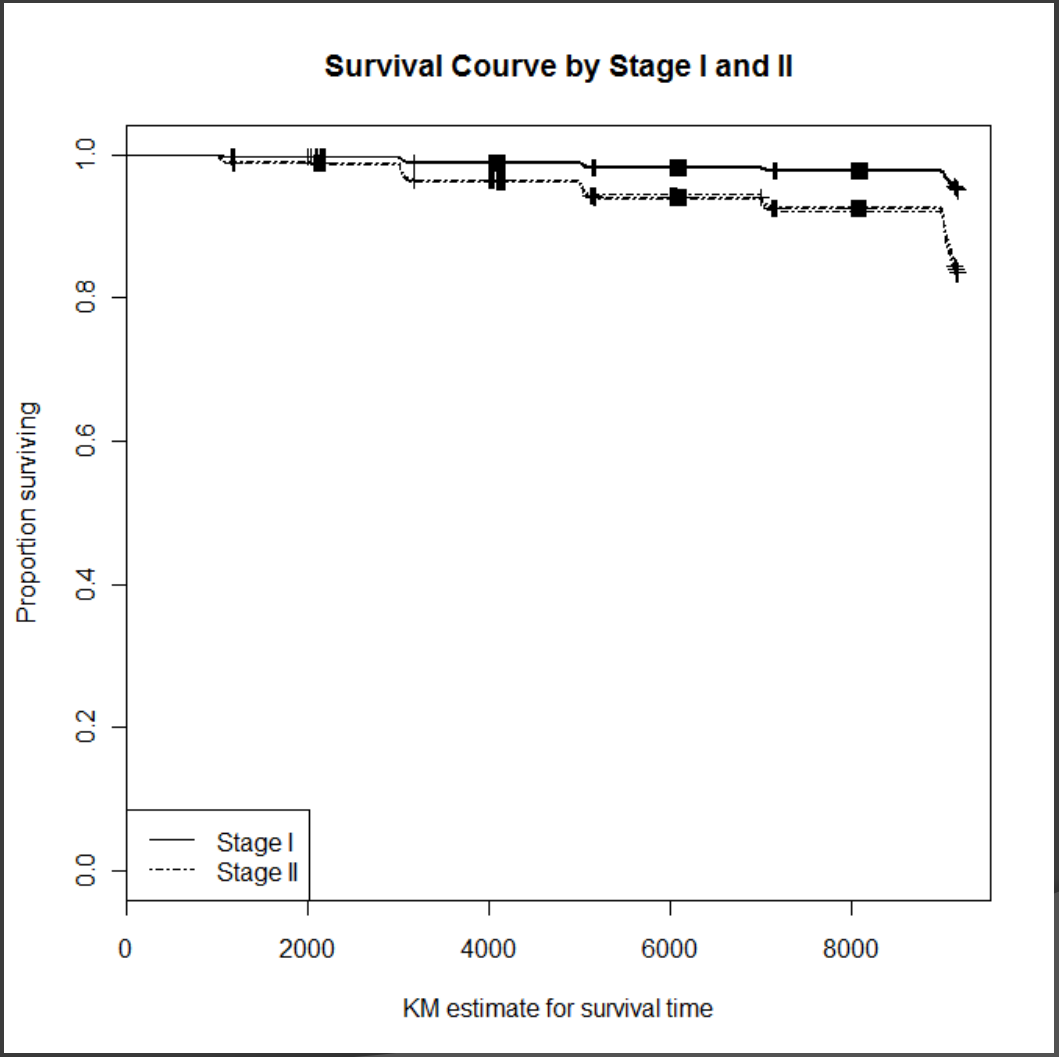
SEMI PARAMETRIC WAY TO ANALYSIS OF BREAST CANCER DATA

USF CANCER RESEARCH TEAM

Survival Curve(race)



Stage I and II



CANCER PREVENTION STUDY II

CANCER PREVENTION STUDY II

CPS II

Cancer Prevention Study II

- The Cancer Prevention Study II (CPS-II) is a prospective cohort study funded and conducted by the American Cancer Society (ACS). The goal of the study is to examine the impact of environmental and lifestyle factors on cancer etiology in a large group of American men and women. To achieve this goal, approximately 1.2 million men and women were enrolled in 1982 with the help of 77,000 American Cancer Society volunteers in 50 states, the District of Columbia, and Puerto Rico. Many of the participants were friends, neighbors, family members, or acquaintances of the volunteers.
- Study participants (known as the CPS-II Baseline Cohort) completed an initial study questionnaire in 1982 that obtained information on a range of lifestyle factors such as diet, use of alcohol and tobacco, occupation, medical history, and family cancer history. These data have been examined extensively in relation to cancer mortality. Vital status of study participants is updated biennially through computerized linkage with the National Death Index. Cause of death has been documented for 99% of all deaths that have occurred. Mortality follow-up of the CPS-II Baseline Cohort is complete through 2006 and is expected to continue for many years. Over 488,000 deaths have occurred in this cohort from 1982 to 2006.

- http://www.cancer.org

CPS II Nutrition Cohort

- In 1992, a new questionnaire was mailed to a subgroup of the CPS-II Baseline Cohort to obtain detailed information on diet, to update other lifestyle factors, and to conduct prospective cancer incidence follow-up in addition to mortality follow-up. This subgroup was chosen among baseline cohort members, aged 50-74, who resided in 21 states with population-based state cancer registries (California, Connecticut, Florida, Georgia, Illinois, Iowa, Louisiana, Maryland, Massachusetts, Michigan, Minnesota, Missouri, New Mexico, New Jersey, New York, North Carolina, Pennsylvania, Utah, Virginia, Washington, and Wisconsin).
- Known as the CPS-II Nutrition Cohort, this subgroup of 184,194 men and women received additional mailed questionnaires in 1997, 1999, 2001, 2003, 2005, and 2007, to update exposure information and to obtain self-reported cancer diagnoses. With permission from study participants, self-reported cancer diagnoses are verified by medical record review. Computerized linkage with state cancer registries is used to supplement self-reported information on cancer incidence. Future questionnaires are planned on a biennial basis.
  - http://www.cancer.org

CPS II Biospecimen Repository

- In 1998, the CPS-II Lifelink Cohort was initiated to obtain blood samples from 40,000 surviving members of the CPS-II Nutrition Cohort that resided in urban and suburban areas. Blood collection was coordinated by American Cancer Society staff and volunteers and performed by hospital staff at community hospitals (approximately 312 hospitals in 20 states, recruited mainly from American College of Surgeons [ACOS] Commission on Cancer approved programs). Collection of blood samples for LifeLink was completed in June, 2001.
- A total of 39,380 Nutrition Cohort members gave a single blood sample. The biospecimen repository was expanded to obtain buccal cell samples by mail from those participants who were unable or unwilling to give a blood sample. Collection began in January 2001, and was completed in May 2002. Buccal cell samples were received from approximately 67,000 cohort members. These blood and buccal cell samples are being stored in liquid nitrogen for epidemiologic investigations, including the role of nutritional, hormonal, and genetic factors in the development of cancer and other diseases.

- http://www.cancer.org

CPS II (N=370,264)

White

347,202

African American

18,905

Asian

1,074

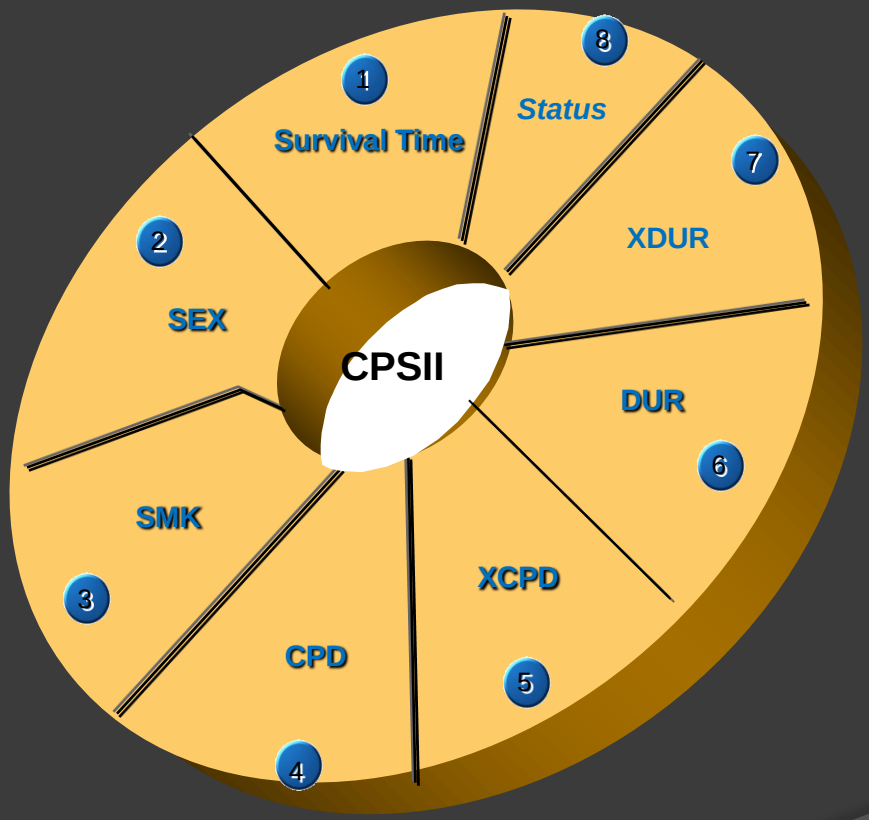
Hispanic

2,217

Others

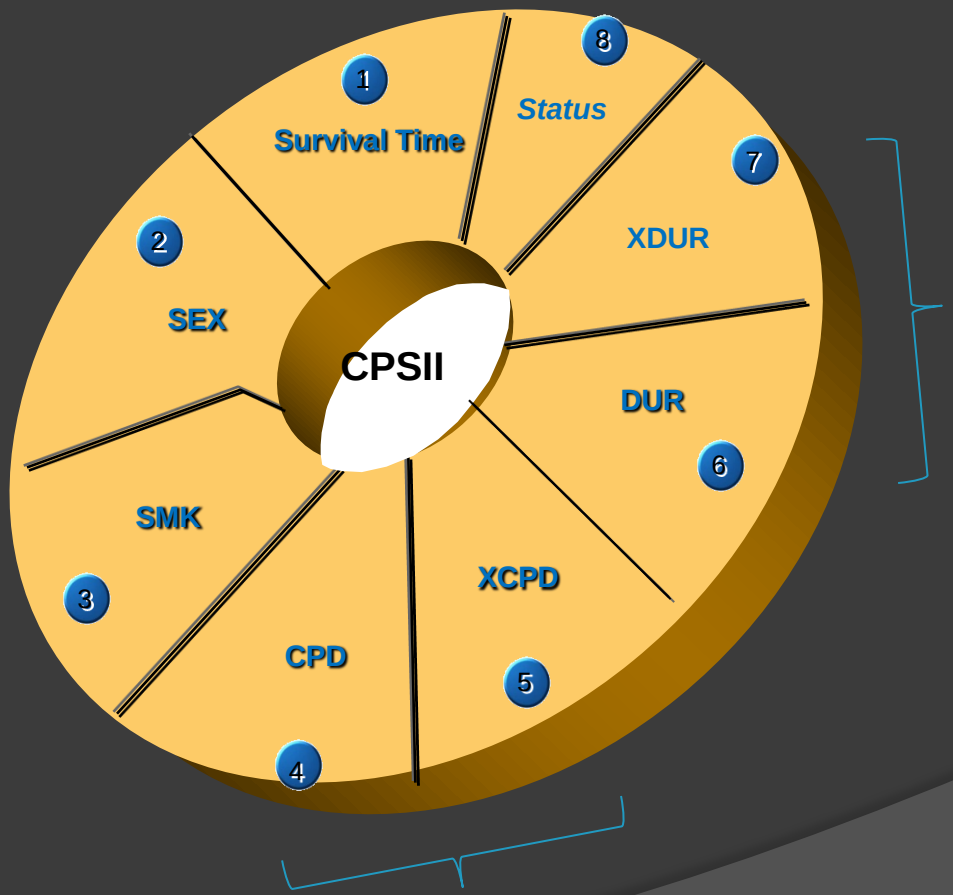
866

Data Network



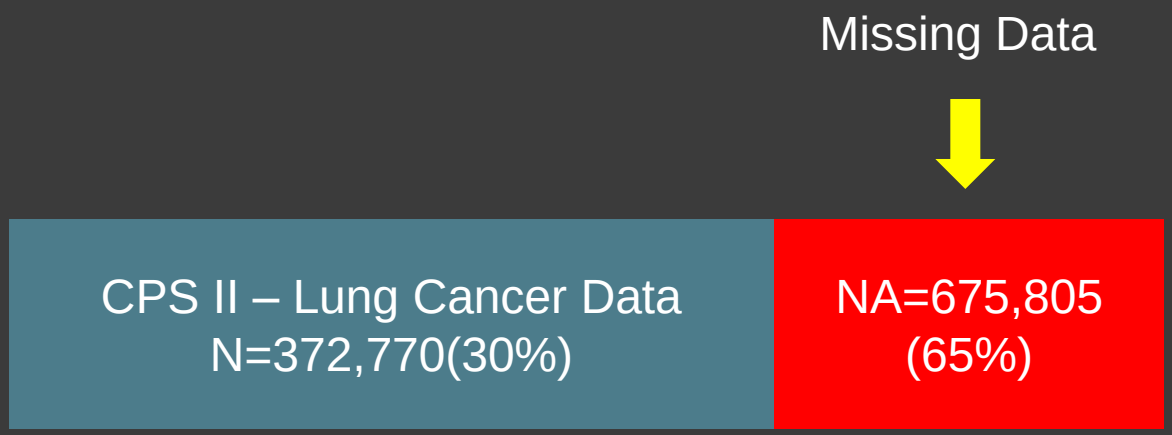
"1" = "Nonsmoker“
"2" = "Current Smoker, Complete Info"
"3" = "Former Smoker, Complete Info"
"4" = "Current Smoker, Incomplete Info"
"5" = "Former Smoker, Incomplete Info"
"6" = "Ever Smoker, Unclassified"
"7" = "Bad Data"
"8" = "Ever Pipe/Cigar Smoker"

Data Network

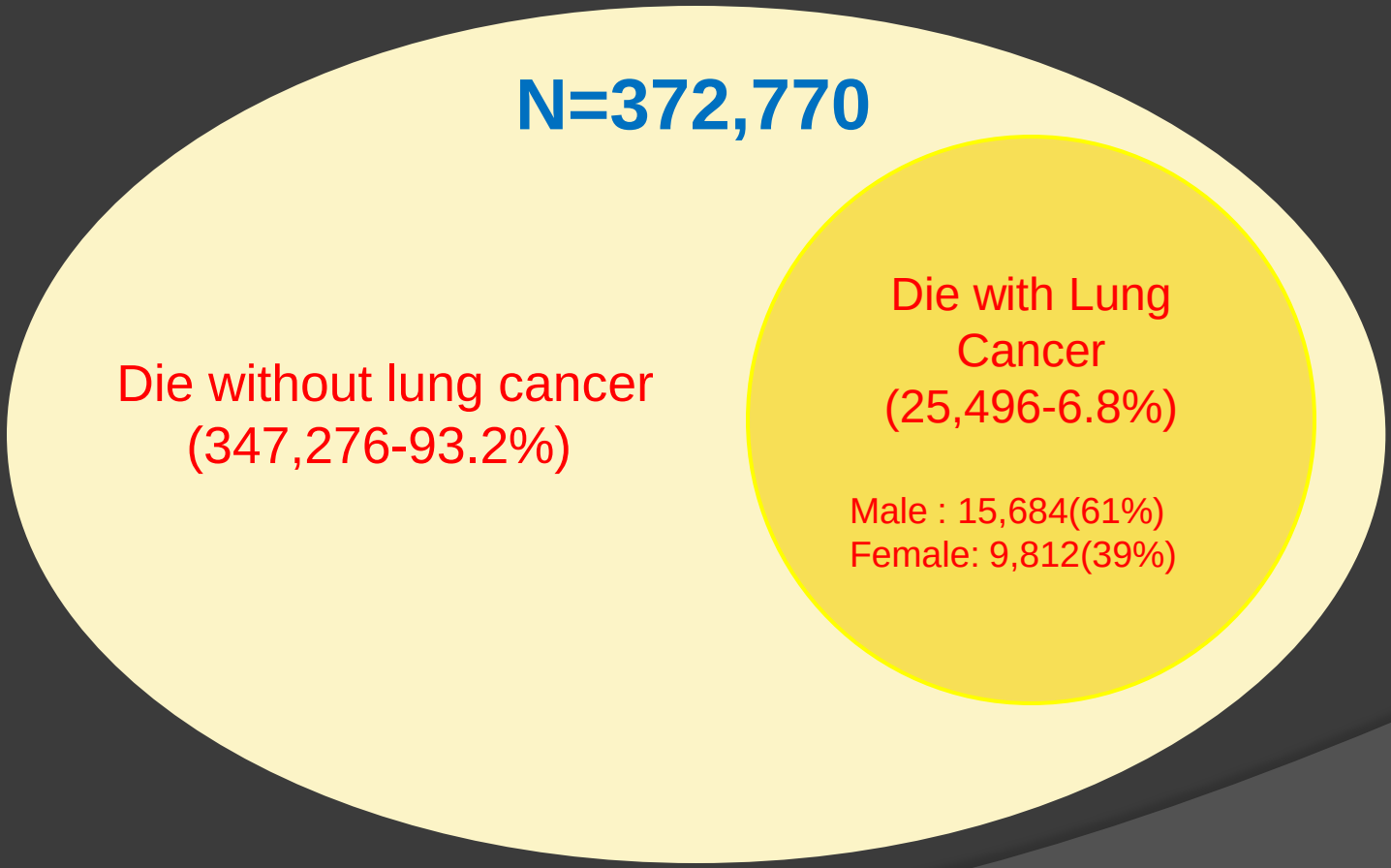


"1" = "Nonsmoker“
"2" = "Current Smoker, Complete Info"
"3" = "Former Smoker, Complete Info"
"4" = "Current Smoker, Incomplete Info"
"5" = "Former Smoker, Incomplete Info"
"6" = "Ever Smoker, Unclassified"
"7" = "Bad Data"
"8" = "Ever Pipe/Cigar Smoker"

Data Network



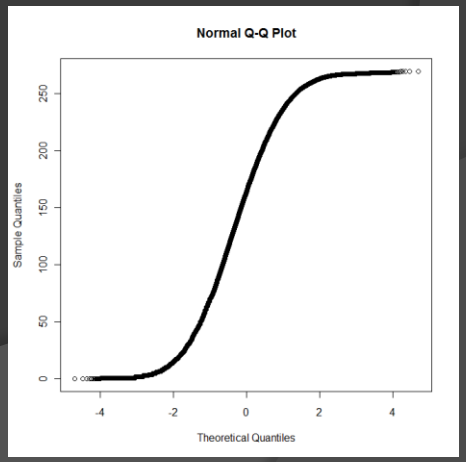
Data Network



Hypothesis testing for survival time (male vs female)

$$H_0 : \mu_{male} = \mu_{female}$$
$$H_1 : \mu_{male} \neq \mu_{female}$$

- P-value<2.2e-16
- Mean of male = 156(month)
- Mean of female = 142(month)
- Statistically female has shorter survival time than male.



Finial data after raw data manipulation

X	BDAYDATE	DATEDD	RACE.O	RACE	SMK82	CPD82	XCPD82	CPDTMP	CPD	DUR82	XDUR82	DURTMP	DUR	DTINT82	AGEQT82	DEAD	DEATHDT	LUNGDTH	SEX	survtimeold	time	status
1	08/15/1894	6/15/1987	1	1	1	NA	NA	0	NA	NA	NA	0	NA	9/15/1982	M	1	6/15/1987	0	2	57.00822	57.01	0
2	3/15/1907	2/15/1985	1	1	2	20	NA	20	20	61	NA	61	61	9/15/1982	M	1	2/15/1985	0	1	29.06301	29.06	0
3	1/15/1907	12/4/2003	1	1	3	NA	10	10	10	NA	22	22	22	9/15/1982	40	1	12/4/2003	0	2	254.79452	254.79	0
6	5/15/1934	5/10/2004	1	1	2	20	NA	20	20	30	NA	30	30	9/15/1982	M	1	5/10/2004	0	2	259.98904	259.99	0
7	9/15/1915	9/13/1999	1	1	1	NA	NA	0	NA	NA	NA	0	NA	9/15/1982	M	1	9/13/1999	0	2	204.06575	204.07	0
10	8/15/1921	5/26/1994	1	1	1	NA	NA	0	NA	NA	NA	0	NA	9/15/1982	M	1	5/26/1994	0	2	140.41644	140.42	0
11	7/15/1918	12/11/2000	1	1	3	NA	20	20	20	NA	26	26	26	9/15/1982	44	1	12/11/2000	0	1	219.02466	219.02	0
15	11/15/1929	4/15/1988	1	1	2	20	NA	20	20	35	NA	35	35	9/15/1982	M	1	4/15/1988	0	1	67.03562	67.04	0
17	11/15/1922	3/6/1995	1	1	2	20	NA	20	20	43	NA	43	43	9/15/1982	M	1	3/6/1995	0	1	149.75342	149.75	0
19	9/15/1927	3/26/1995	3	3	2	20	NA	20	20	25	NA	25	25	9/15/1982	M	1	3/26/1995	0	2	150.41096	150.41	0
20	9/15/1919	6/15/1987	3	3	2	15	NA	15	15	44	NA	44	44	9/15/1982	M	1	6/15/1987	0	1	57.00822	57.01	0
21	8/15/1928	11/15/1986	3	3	2	20	NA	20	20	29	NA	29	29	9/15/1982	M	1	11/15/1986	0	2	50.03836	50.04	0
32	1/15/1941	5/2/1999	1	1	3	NA	20	20	20	NA	10	10	10	9/15/1982	25	1	5/2/1999	0	2	199.66027	199.66	0
33	2/15/1905	9/15/1986	1	1	3	NA	60	60	60	NA	12	12	12	9/15/1982	33	1	9/15/1986	0	1	48.03288	48.03	0
34	12/15/1908	4/28/1997	1	1	3	NA	20	20	20	NA	35	35	35	9/15/1982	52	1	4/28/1997	0	2	175.52877	175.53	0
35	8/15/1922	2/17/2003	1	1	2	20	NA	20	20	45	NA	45	45	9/15/1982	M	1	2/17/2003	0	2	245.26027	245.26	0
38	10/15/1906	1/17/1997	1	1	3	NA	1	1	1	NA	2	2	2	9/15/1982	21	1	1/17/1997	0	2	172.20822	172.21	0
40	1/25/1920	8/13/1991	1	1	2	40	NA	40	40	41	NA	41	41	9/15/1982	M	1	8/13/1991	0	2	106.98082	106.98	0
43	12/15/1905	2/2/1993	1	1	1	NA	NA	0	NA	NA	NA	0	NA	9/15/1982	M	1	2/2/1993	0	2	124.70137	124.70	0
44	10/15/1940	1/15/1990	1	1	2	40	NA	40	40	27	NA	27	27	9/15/1982	M	1	1/15/1990	0	1	88.07671	88.08	0

PARAMETRIC WAY TO ANALYSIS OF LUNG CANCER DATA

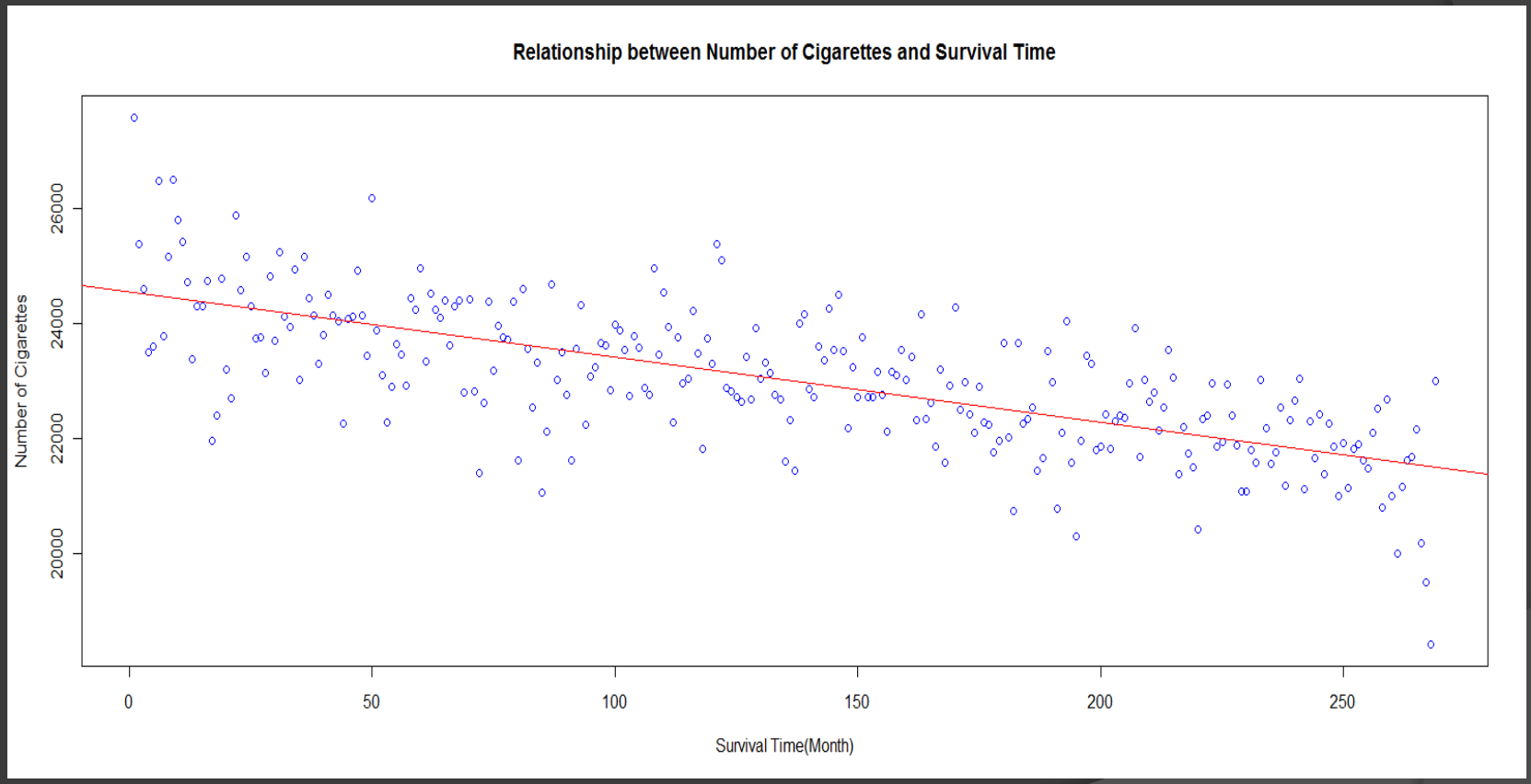
USF CANCER RESEARCH TEAM

Why parametric ?

- We have known that parametric way will for sure be the best.

- Next step is using K-D, K-M and Cox ph

Linear Relationship between Mean Number of Cigarettes and Survival Time



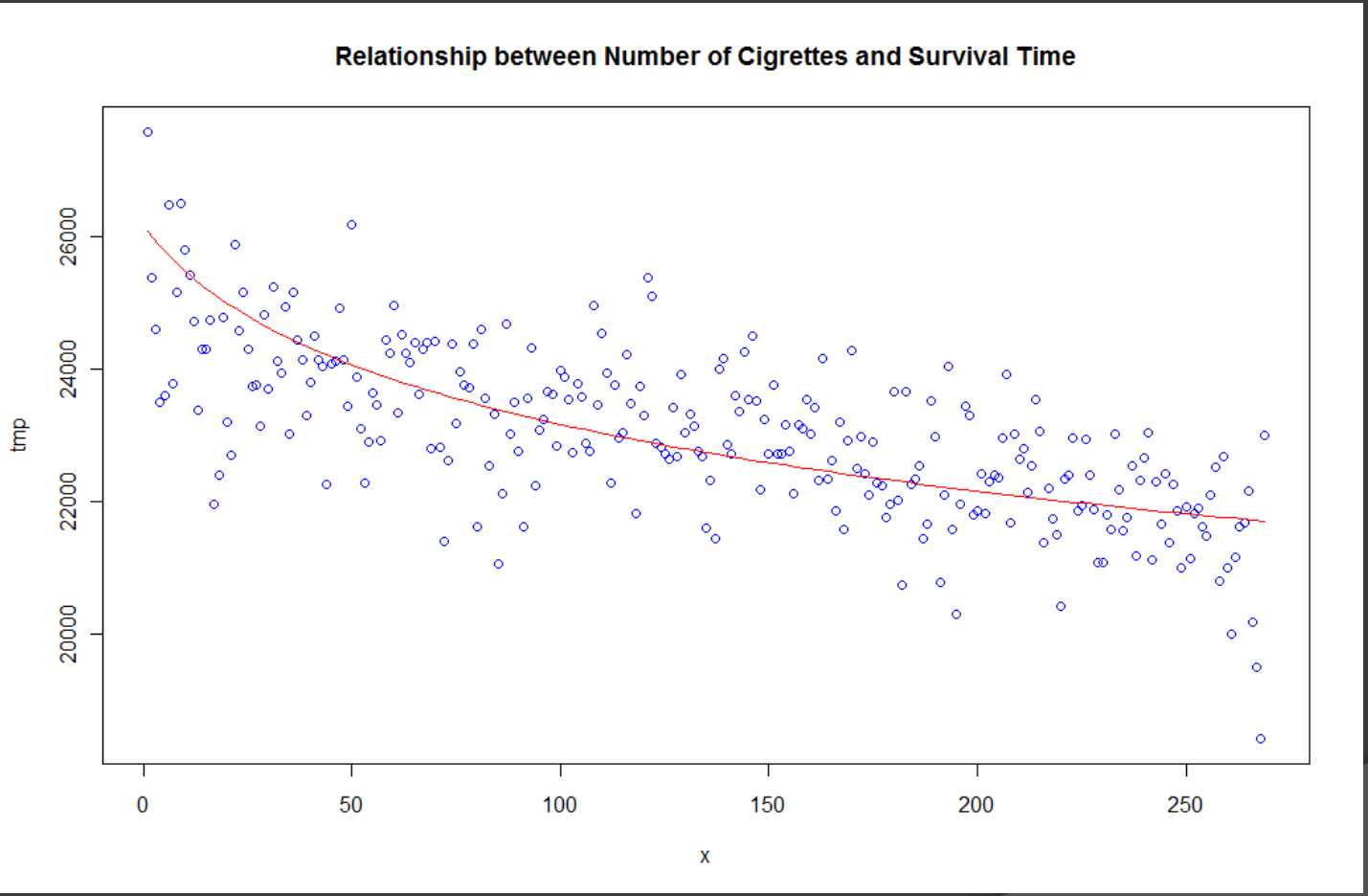
Linear Relationship between Mean Number of Cigarettes and Survival Time

The result of regression analysis give us the regression line as

 # cig=24561.92-11.38time

P-value < 2e-16 and Radj=0.488 which means the negative effect of between mean number of cigarettes and survival time is significant.

Exponential Relationship between Mean Number of Cigarettes and Survival Time



Exponential Relationship between Mean Number of Cigarettes and Survival Time

The result of regression analysis give us the regression line as

 # Cig = -ln(time)/0.0005+32248.4

=> time=exp(0.0005(32248.4-#cig))

=1.0005 exp(32248.4-#cig)

P-value < 2e-16 and Radj=0.429

which means the exponentially decaying effect between mean number of cigarettes and survival time is significant.

DIE WITHIN 5 YEARS

Probability Density Functions

P.D.F (Survival time) **Johnson SU Distribution**	$f(t) = \frac{\delta}{(\lambda\sqrt{2\pi}\, z(1-z))} \exp(-\frac{1}{2}(\gamma + \delta \ln(\frac{z}{1-z}))^2)$ where $z = \frac{t-\xi}{\lambda}$, and $\gamma = 0.1722, \delta = 0.7424, \lambda = 63.122$, and $\xi = -0.2701$.
P.D.F (number of cigarettes) **Generalized Logistic Distribution**	$f(x) = \frac{(1+kz)^{-1-1/k}}{\alpha(1+(1+kz)^{-1/k})^2}$ where $z = \frac{x-\mu}{\sigma}$, and $k = -0.03337, \sigma = 3840.3$, and $\mu = 24339.0$.
Survival function	$S(x) = 1 - \Phi(\gamma + \delta \ln(\frac{z}{1-z}))$ where Φ is the Laplace Integral.

DIED WITHIN 10 YEARS

Probability Density Functions

P.D.F (Survival time) **Power Distribution**	$f(t) = \frac{\alpha(t-a)^{\alpha-1}}{(b-a)^{\alpha}}$, where $\alpha = 1.2604$, $a = 0.85801$, and $b = 120.99$.
P.D.F (number of cigarettes) **Generalized Logistic Distribution**	$f(x) = \frac{(1+kz)^{-1-1/k}}{\alpha(1+(1+kz)^{-1/k})^2}$ where $z = \frac{x-\mu}{\sigma}$, and $k = -0.04162$, $\sigma = 3848.6$, and $\mu = 23975.0$.
Survival function	$S(t) = 1-\left(\frac{t-a}{b-a}\right)^{\alpha}$

DIED WITHIN 15 YEARS

Probability Density Functions(15 Years)

P.D.F (Survival time) **Johnson SU Distribution**	$f(t) = \frac{\delta}{(\lambda\sqrt{2\pi}\, z(1-z))} \exp(-\frac{1}{2}(\gamma + \delta \ln(\frac{z}{1-z}))^2)$ where $z = \frac{t-\xi}{\lambda}$, and $\gamma = 0.2162$, $\delta = 0.6924$, $\lambda = 187.62$, and $\xi = -2.7478$.
P.D.F (number of cigarettes) **Generalized Logistic Distribution**	$f(x) = \frac{(1+kz)^{-1-1/k}}{\alpha(1+(1+kz)^{-1/k})^2}$ where $z = \frac{x-\mu}{\sigma}$, and $k = -0.03349$, $\sigma = 3806.5$, and $\mu = 23614.0$.
Survival function	$S(t) = 1 - \Phi(\gamma + \delta \ln(\frac{z}{1-z}))$ where Φ is the Laplace Integral.

DIED WITHIN 20 YEARS

Probability Density Functions

P.D.F (Survival time) **Beta Distribution**	$f(t) = \frac{1}{B(\alpha_1, \alpha_2)} \frac{(t-a)^{\alpha_1 - 1}(b-t)^{\alpha_2 - 1}}{(b-a)^{\alpha_1 + \alpha_2 - 1}}$ where B is the Beta Function, and $\alpha_1 = 1.2885$, $\alpha_2 = 1.0385$, a = 0.79202, and b = 240.99.
P.D.F (number of cigarettes) **Generalized Logistic Distribution**	$f(x) = \frac{(1+kz)^{-1-1/k}}{\alpha(1+(1+kz)^{-1/k})^2}$ where $z = \frac{x - \mu}{\sigma}$, and k = -0.02596, σ = 3782.0, and μ = 23219.0.
Survival function	$S(t) = 1 - I_z(\alpha_1, \alpha_2)$ where $z = \frac{t - a}{b - a}$, and I is the regularized Incomplete Beta Function.

DIED WITHIN 22.3(268 MONTH) YEARS

Probability Density Functions

P.D.F (Survival time) **Beta Distribution**	$f(t) = \frac{1}{B(\alpha_1, \alpha_2)} \frac{(t-a)^{\alpha_1 - 1}(b-t)^{\alpha_2 - 1}}{(b-a)^{\alpha_1 + \alpha_2 - 1}}$ where B is the Beta Function, and $\alpha_1 = 1.2944$, $\alpha_2 = 1.0638$, a = 0.79201, and b = 268.5.
P.D.F (number of cigarettes) **Generalized Logistic Distribution**	$f(x) = \frac{(1+kz)^{-1-1/k}}{\alpha(1+(1+kz)^{-1/k})^2}$ where $z = \frac{x - \mu}{\sigma}$, and k = -0.04162, σ = 3848.6, and μ = 23975.0.
Survival function	$S(t) = 1 - I_z(\alpha_1, \alpha_2)$ where $z = \frac{t - a}{b - a}$, and I is the regularized Incomplete Beta Function.

FINDING ATTRIBUTABLE VARIABLES

USF CANCER RESEARCH TEAM

n=162455 (174150 observations deleted due to missingness)

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
factor(SEX)2	1.390e-01	1.149e+00	1.478e-01	0.941	0.346759	
factor(SMK)2	1.094e+00	2.985e+00	1.528e-01	7.159	8.14e-13	***
factor(SMK)3	NA	NA	0.000e+00	NA	NA	
factor(SMK)6	NA	NA	0.000e+00	NA	NA	
factor(SMK)8	NA	NA	0.000e+00	NA	NA	
CPD	9.977e-03	1.010e+00	3.304e-03	3.020	0.002531	**
DUR	4.512e-02	1.046e+00	2.899e-03	15.563	< 2e-16	***
factor(SEX)2:factor(SMK)2	3.099e-01	1.363e+00	2.112e-01	1.467	0.142322	
factor(SEX)2:factor(SMK)3	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:factor(SMK)6	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:factor(SMK)8	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:CPD	1.497e-02	1.015e+00	5.663e-03	2.644	0.008184	**
factor(SMK)2:CPD	9.654e-03	1.010e+00	4.955e-03	1.948	0.051362	.
factor(SMK)3:CPD	NA	NA	0.000e+00	NA	NA	
factor(SMK)6:CPD	NA	NA	0.000e+00	NA	NA	
factor(SMK)8:CPD	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:DUR	-1.755e-02	9.826e-01	4.517e-03	-3.885	0.000102	***
factor(SMK)2:DUR	-1.535e-02	9.848e-01	3.947e-03	-3.889	0.000100	***
factor(SMK)3:DUR	NA	NA	0.000e+00	NA	NA	
factor(SMK)6:DUR	NA	NA	0.000e+00	NA	NA	
factor(SMK)8:DUR	NA	NA	0.000e+00	NA	NA	
CPD:DUR	4.682e-05	1.000e+00	9.139e-05	0.512	0.608386	
factor(SEX)2:factor(SMK)2:CPD	-1.103e-02	9.890e-01	7.846e-03	-1.405	0.159885	
factor(SEX)2:factor(SMK)3:CPD	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:factor(SMK)6:CPD	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:factor(SMK)8:CPD	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:factor(SMK)2:DUR	-2.562e-03	9.974e-01	5.883e-03	-0.435	0.663235	
factor(SEX)2:factor(SMK)3:DUR	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:factor(SMK)6:DUR	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:factor(SMK)8:DUR	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:CPD:DUR	-5.832e-05	9.999e-01	1.667e-04	-0.350	0.726446	
factor(SMK)2:CPD:DUR	-1.722e-04	9.998e-01	1.282e-04	-1.344	0.179086	
factor(SMK)3:CPD:DUR	NA	NA	0.000e+00	NA	NA	
factor(SMK)6:CPD:DUR	NA	NA	0.000e+00	NA	NA	
factor(SMK)8:CPD:DUR	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:factor(SMK)2:CPD:DUR	2.129e-04	1.000e+00	2.176e-04	0.978	0.327847	
factor(SEX)2:factor(SMK)3:CPD:DUR	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:factor(SMK)6:CPD:DUR	NA	NA	0.000e+00	NA	NA	
factor(SEX)2:factor(SMK)8:CPD:DUR	NA	NA	0.000e+00	NA	NA	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

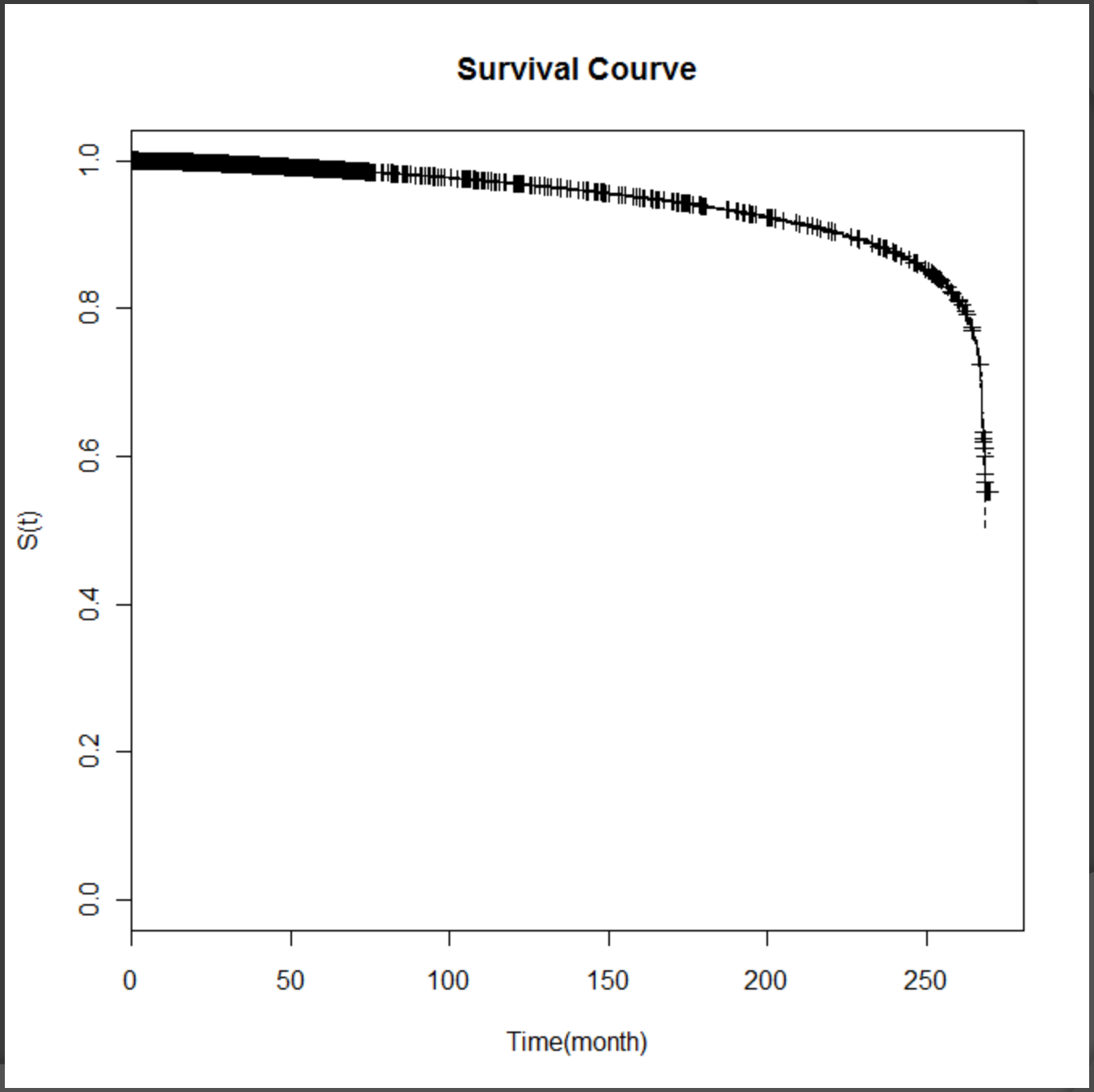
Attributable Variables

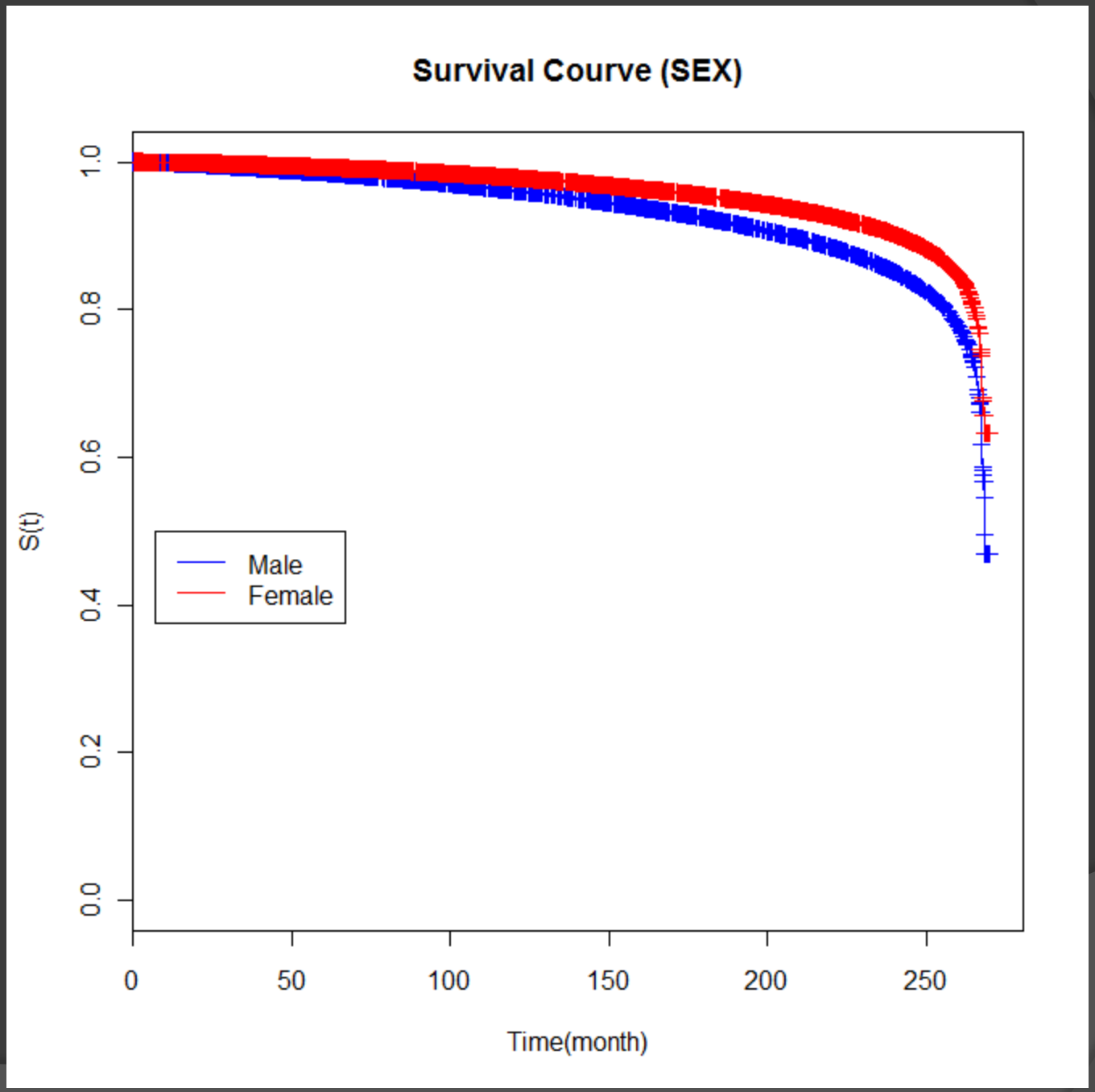
- **The following variables have significant effect on survival time**

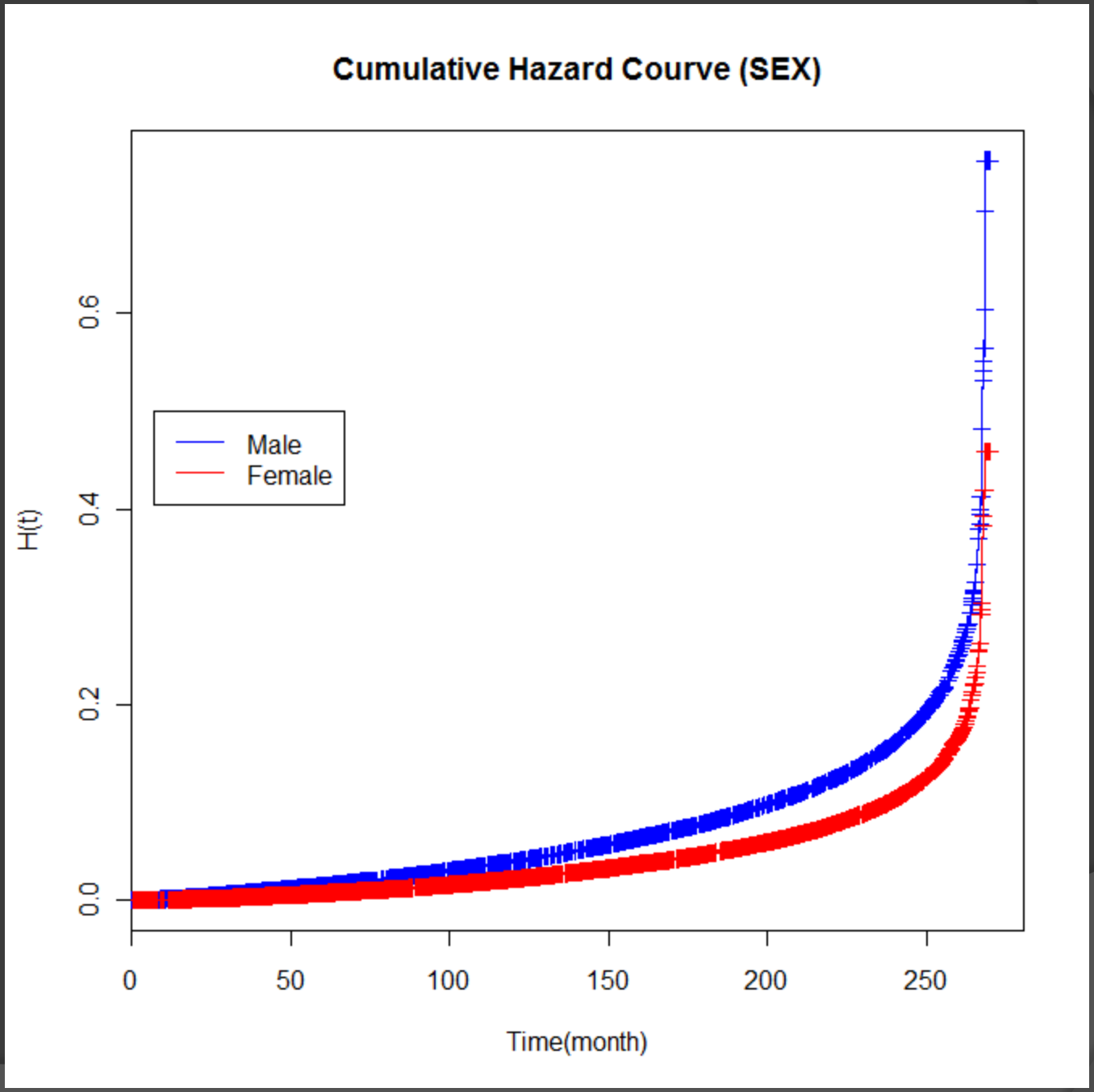
- SMK=2 (Current Smoker)
- CPD (Cigarettes per day)
- DUR (Duration time)
- SEX=2 (Female)
- SEX=2:DUR (Female : Duration time)
- SMK=2:DUR (Current Smoker : Duration time)

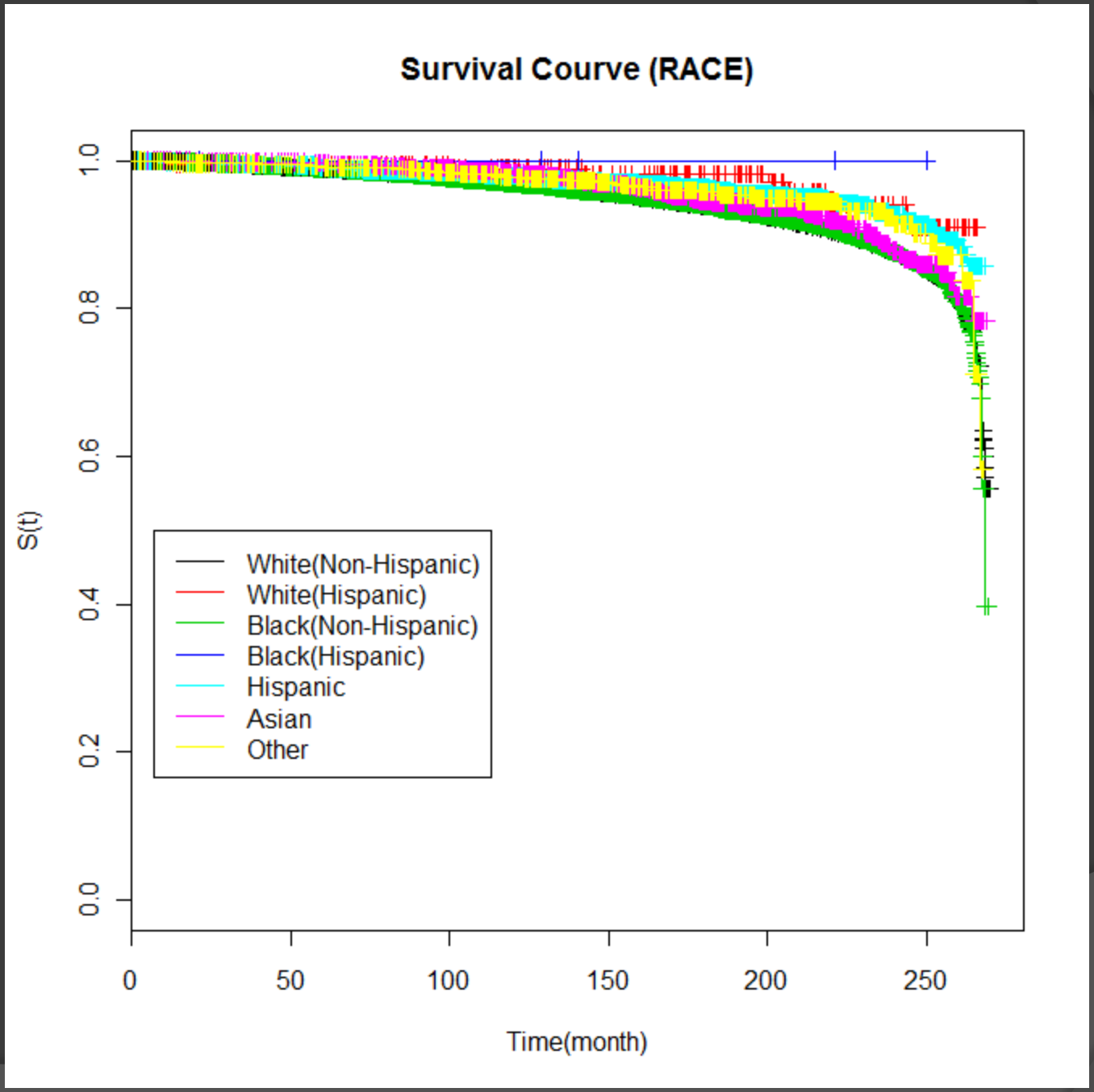
SEMI PARAMETRIC WAY TO ANALYSIS OF LUNG CANCER DATA

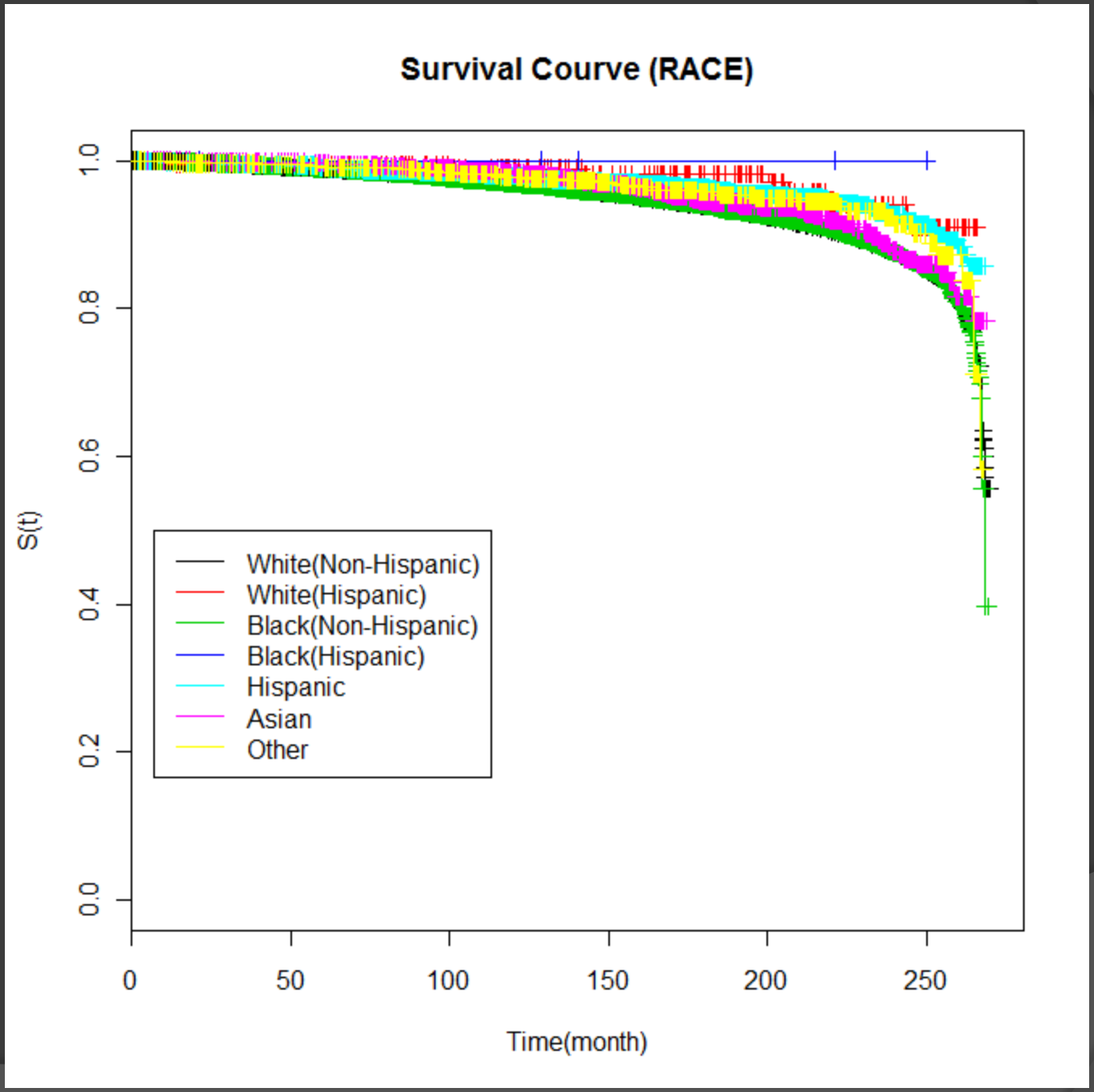
USF CANCER RESEARCH TEAM

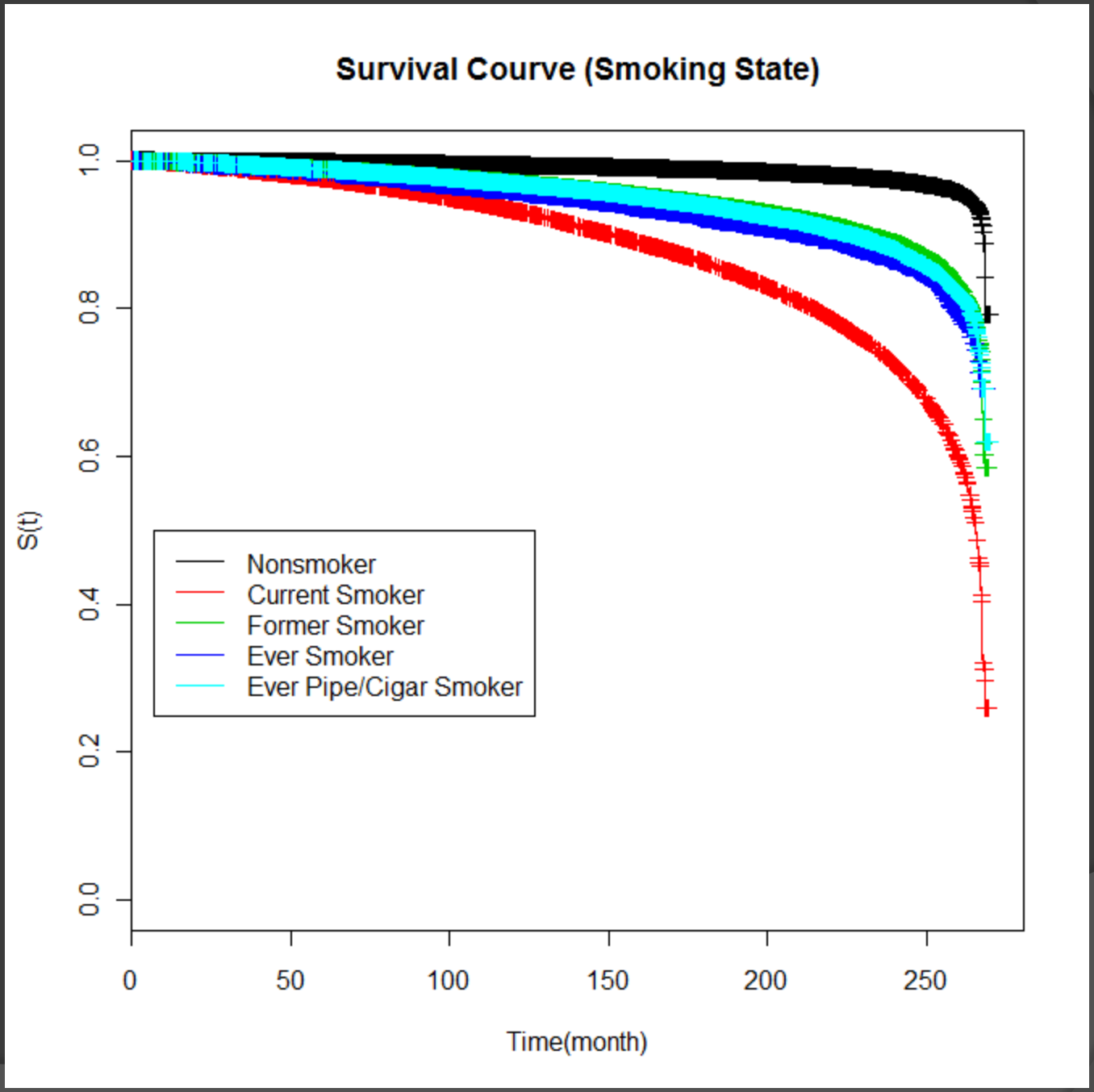

Survival Courve
S(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
Time(month)

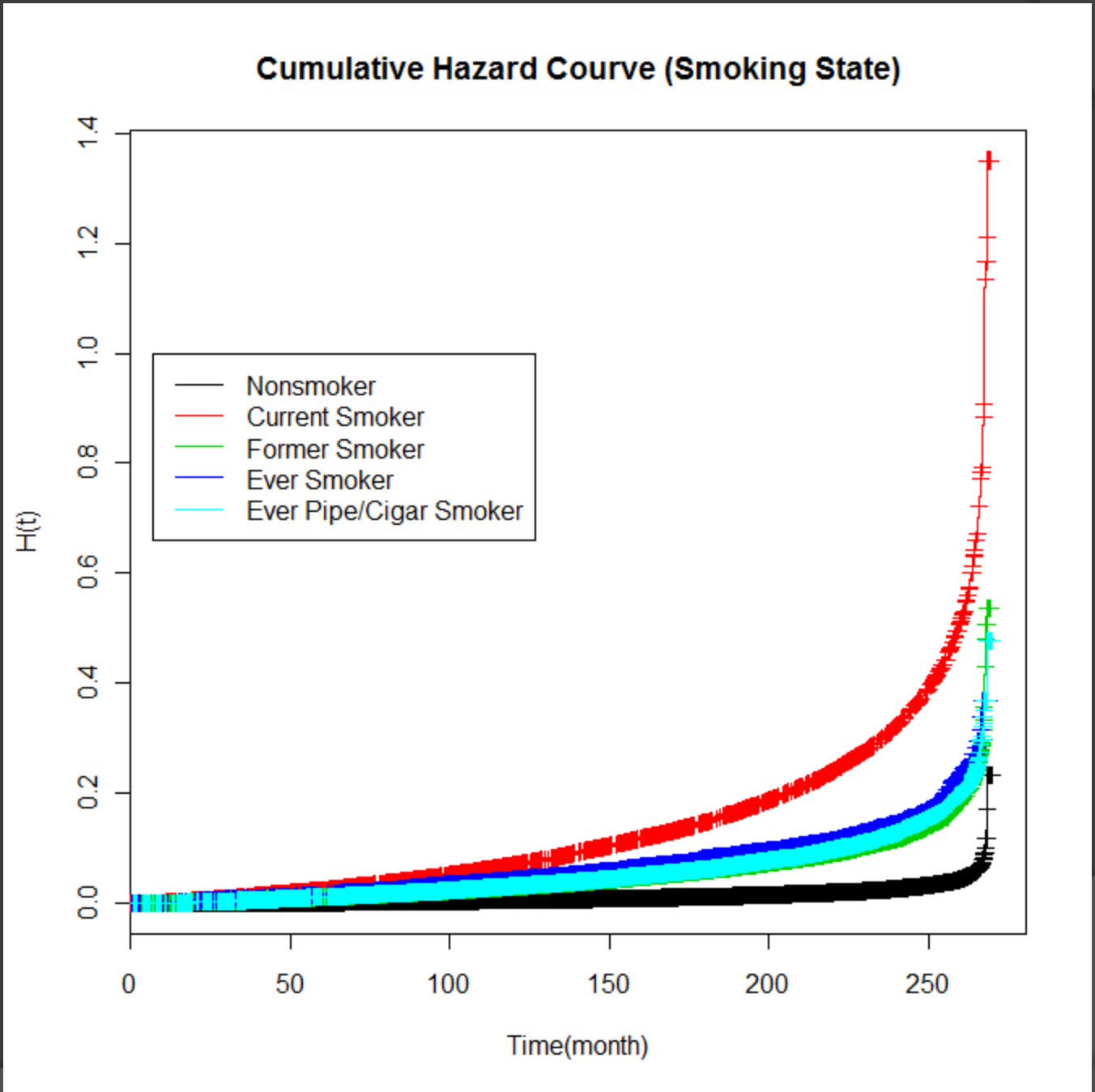

Survival Courve (SEX)
S(t)
1.0
0.8
0.6
0.4
0.2
0.0
Male
Female
0
50
100
150
200
250
Time(month)

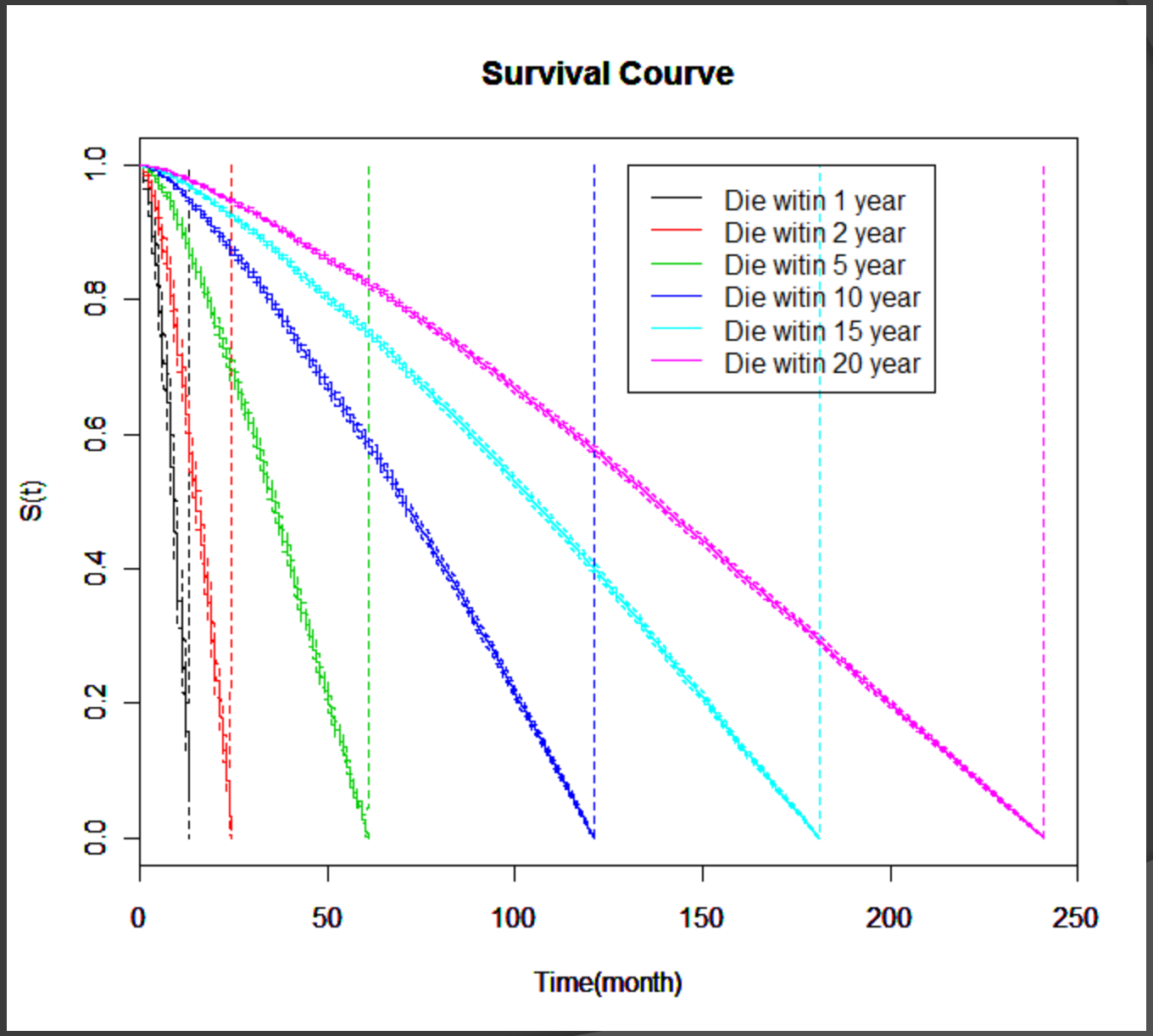

Cumulative Hazard Courve (SEX)
H(t)
0.0
0.2
0.4
0.6
Male
Female
0
50
100
150
200
250
Time(month)

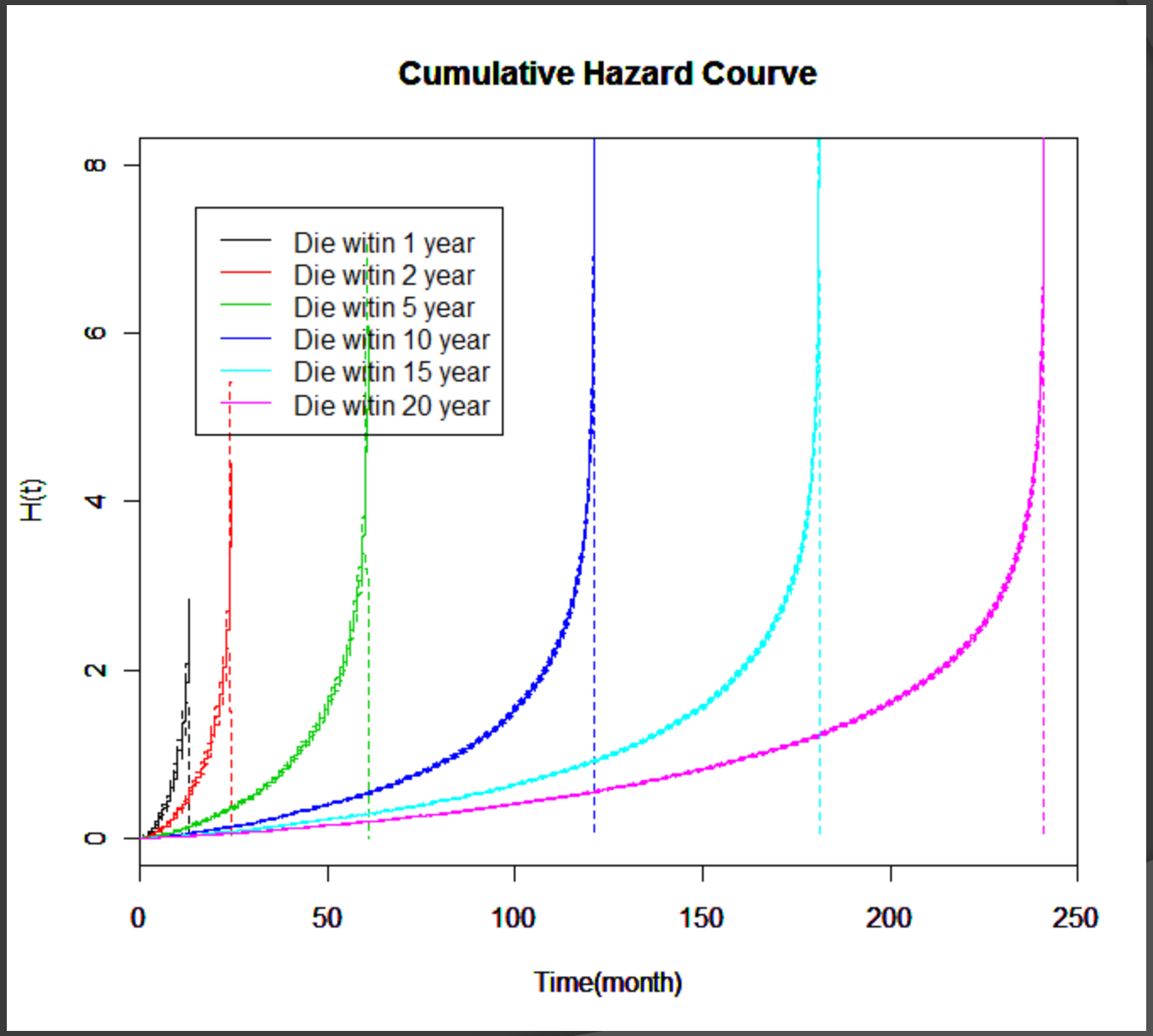

Survival Courve (RACE)
S(t)
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
Time(month)
White(Non-Hispanic)
White(Hispanic)
Black(Non-Hispanic)
Black(Hispanic)
Hispanic
Asian
Other

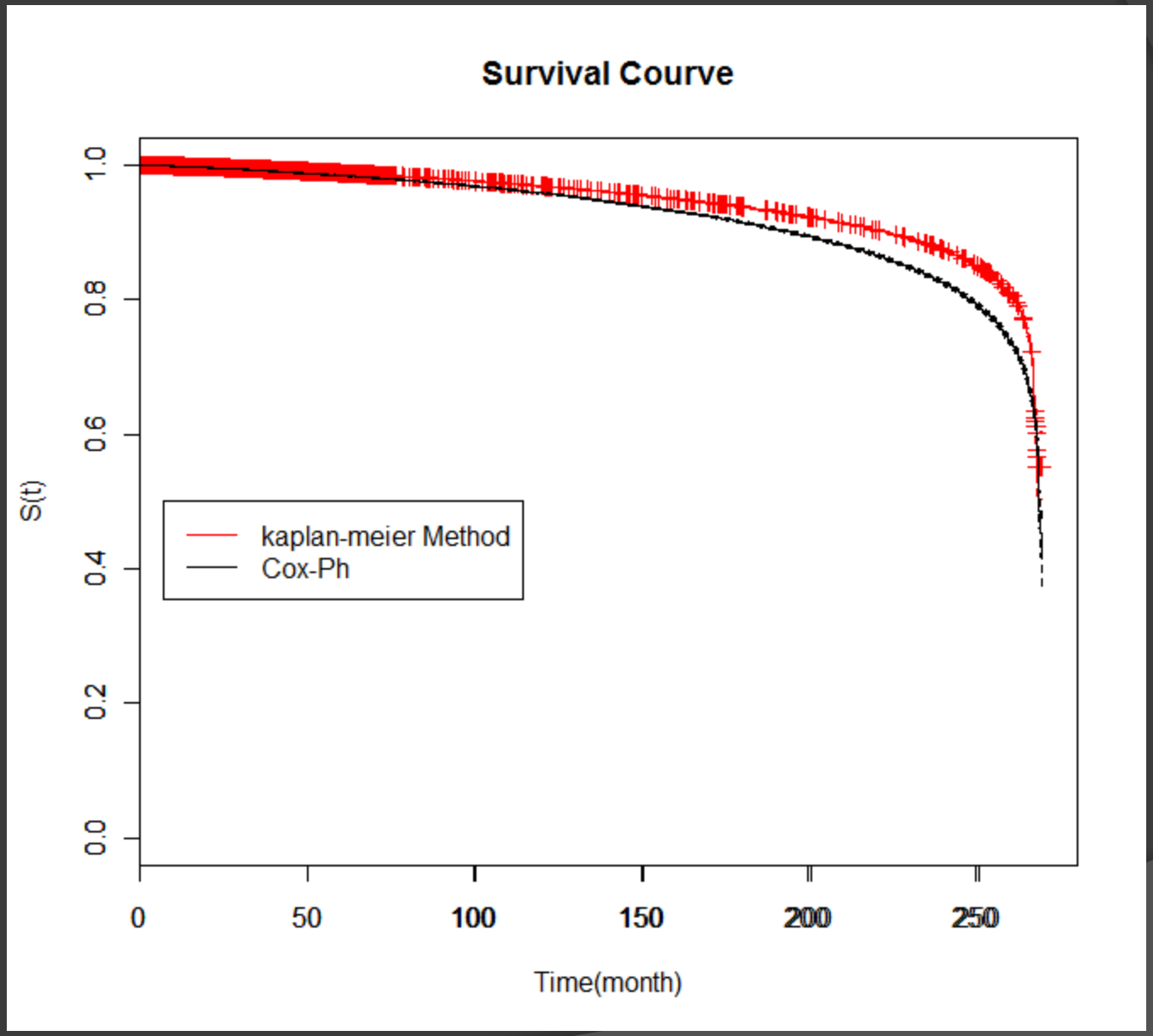

Survival Courve (RACE)
S(t)
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
Time(month)
White(Non-Hispanic)
White(Hispanic)
Black(Non-Hispanic)
Black(Hispanic)
Hispanic
Asian
Other


Survival Courve (Smoking State)
S(t)
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
Time(month)
Nonsmoker
Current Smoker
Former Smoker
Ever Smoker
Ever Pipe/Cigar Smoker


Cumulative Hazard Courve (Smoking State)
H(t)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
50
100
150
200
250
Time(month)
Nonsmoker
Current Smoker
Former Smoker
Ever Smoker
Ever Pipe/Cigar Smoker


Survival Courve
S(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
Time(month)
Die witin 1 year
Die witin 2 year
Die witin 5 year
Die witin 10 year
Die witin 15 year
Die witin 20 year


Cumulative Hazard Courve
Die witin 1 year
Die witin 2 year
Die witin 5 year
Die witin 10 year
Die witin 15 year
Die witin 20 year
H(t)
0
2
4
6
8
0
50
100
150
200
250
Time(month)


Survival Courve
S(t)
1.0
0.8
0.6
0.4
0.2
0.0
kaplan-meier Method
Cox-Ph
0
50
100
150
200
250
Time(month)

Reference

- American Cancer Society http://www.cancer.org
- A. W. Fyles, D. R. McCready, L. A Manchul., M. E. Trudeau, P. Merante, M. Pintile, L. M. Weir, and I. A. Olivotto, Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer, *New England Journal of Medicine*, 351 (2004) 963-970.
- B. Abraham and J. Ledolter, Introduction to regression modeling, 2006
- C. A. McGilchrist and C.W. Aisbett. Regression with Frailty in Survival Analysis. Biometrics, 47(2):461-466, 1991.
- C. A. McGilchrist, REML Estimation for Survival Models with Frailty. Biometrics, 49(1):221-225,1993
- D. Collett, Modeling survival data in medical research (Chapman & Hall/CRC) , 2003.
- D. P. Harrington, T.R. Fleming, A Class of Rank Test Procedures for Censored Survival Data. *Biometrika*, 69(3):553-566, 1982.
- D. R. Cox, Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society Series B*, 34: 187–220, 1972.
- D. R. Cox and D. Oakes, Analysis of survival data (London: Chapman & Hall), 1984.
- E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. 53:457-448, 1958.
- J. P. Klein. Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm. Biometrics, 48(3)795-806, 1992.

Thank You !