

Tuskegee Study – Statistical Analysis

I. Summary Information

The dataset consisting of 624 records. Of these, 427 (68.4%) were Subjects, 185 (29.6%) were Controls, and 12 (1.9%) were Control to Subject(CtoS) conversions. Unless otherwise specified, analysis was performed after adding the 12 CtoS conversions to the Subjects, giving 185 (29.6%) controls and 439 (70.4%) subject “cases”. Cause of death information was provided for 394 (63.14%) of the 624 records; of those records with cause of death, 285 (72.3%) were subjects and 109 (27.7%) were controls. In all cases, cause of death analysis used the first cause mentioned for the record. For each analysis below, number of observations is determined by information available. For example, 538 records (383 subjects and 155 controls) have information for “age at first exam.”

II. Age when study started

Information on Entry into Study was not available. “Age at first exam” was used as a proxy for age at study start. Table 1 below summarizes descriptive information on age at first exam by case status.

Table 1. Age at first exam, by case status

Group	Age at exam	Mean	Standard Deviation	Median	Minimum	Maximum
Subjects	383	43.46	13.33	43	21	90
Control	155	43.86	13.29	45	16	87

The difference (between controls and subjects) in the mean age at first exam is not statistically significant (t Statistic = 0.3143, p-value = 0.7534).

III. Date of lesion

Data are available on date of first lesion for 147 men. Of these, 143 also have information on data of first exam (these include one C to S crossover). The mean length of time from first lesion to first exam for

these 143 men was 19.7 years (median 19, standard deviation 12.2, range 1-47 with one outlier at 72) (Table 2).

Table 2. Distribution of years from first lesion to first exam (143 men).

<u>Years</u>	<u># Men</u>
1-5	19
6-10	25
11-15	13
16-20	19
21-25	20
26-30	22
31-35	12
36-40	9
41-45	2
46-50	1
72	1

IV. Age/Date at death, by case status

Although the data show no difference between age at first exam (Table 1), case status does have an impact on age at death. Subjects died, on average, almost 5 years earlier than controls (Table 3).

Table 3. Age at first exam by case status

Group	Age at death	Mean	Standard Deviation	Median	Minimum	Maximum
Subjects	320	65.17	14.15	57	22	93
Control	120	69.91	13.26	62	26	97

The difference in the mean age at death is highly statistically significant (t Statistic = 3.18, p-value = 0.0016).

To investigate the effect of the development of new knowledge in medical care, we investigated the effect of “decade of death” by case status (Table 4).

Table 4. Decade of death by case status

Case Status	1930-39	1940-49	1950-59	1960-69	1970-79	Total
Subject (Row%) [Col%]	65 (19.8) [81.3]	79 (24.1) [75.2]	75 (22.9) [71.4]	78 (23.8) [71.6]	31 (9.5) [59.6]	328 (100.0) [72.7]
Control (Row%) [Col%]	15 (12.2) [18.8]	26 (21.1) [24.8]	30 (24.4) [28.6]	31 (25.2) [28.4]	21 (17.1) [40.4]	123 (100.0) [27.3]
Total	80 (17.7) [100.0]	105 (23.3) [100.0]	105 (23.3) [100.0]	109 (24.2) [100.0]	451 (100.0) [100.0]	451

The differences in death decade are statistically significant (Chi-squared statistic = 7.93, p-value = 0.09); men who were subjects tended to die in earlier decades. Specifically, for each decade prior to 1950, proportionately more subjects died than for the controls (see row %), even though the groups were the same age at study entry.

Death rates before 1950 are of interest since this date marks the start of availability of penicillin. Table 5 shows death rates by case status and death by 1950.

Table 5. Death by 1950 by Case Status

Case Status	Dead by 1950	Alive in 1950	Total
Subjects (Row%) [Col%]	144 (44.6) [77.8]	179 (55.4) [69.6]	323 (100.0) [73.1]
Control (Row%) [Col%]	41 (34.5) [22.2]	78 (65.5) [30.4]	119 (100.0) [26.9]
Total	185 (41.9)	257 (58.1)	442 (100.0)

The relative risk of death prior to 1950 is 44.6% (144/323) for subjects compared to 34.5% (41/119) for controls; that is, subjects were about 10% more likely to die in that period (p = 0.03).

For Subjects (including C to S conversions), the mean age at death is 65.17 (std dev 14.15), median age at death 66.

V. Autopsy

Autopsy information was available for 353 records: 256 subjects and 97 controls. For these 353 records, table 6 shows the association between case status and autopsy likelihood.

Table 6. Association of Autopsy with Case Status

Case Status	No Autopsy	Autopsy	Total
Subjects (Row%) [Col%]	94 (36.7) 75.2	162 (63.3) 71.1	256 (100.0) 72.5
Control (Row%) [Col%]	31 (32.0) 24.8	66 (68.0) 28.9	97 (100.0) 27.5
Total	125 (35.4)	228 (64.6)	353 (100.0)

The "risk" (rate) of autopsy was $162/256=63.3\%$ for subjects and $66/97=68\%$ for controls. Thus subjects were about 87% as likely to be autopsied as controls. This difference is not statistically significant ($p\sim 0.20$) but is suggestive of a difference (see section on cause of death below).

If we consider only men who were dead by 1973 (the end of the study), the rate of autopsy was 64.1% for subjects versus 71.9% for controls (p -value = 0.09), a result significant at $p < 0.10$ and based on large enough numbers to be unlikely to be due to chance (Table 7).

Table 7. Association of Autopsy with Case Status, Death by 1973

Case Status	No Autopsy	Autopsy	Total
Subjects (Row%) [Col%]	88 (35.9) [77.9]	157 (64.1) [71.0]	245 (100.0) [73.4]
Control (Row%) [Col%]	25 (28.1) [22.1]	64 (71.9) [29.0]	89 (100.0) [26.6]
Total	113 (33.8)	221 (66.2)	334 (100.0)

VI. Cause of Death

To investigate causes of death (COD), narrative cause of death information was grouped into causes corresponding to ICD5 categories (the ICD coding system used in the 1940s). For each record, the first cause of death recorded was used. Table 8 lists the COD groups.

Table 8. Causes of death.

- 1.1 aortitis
- 1.2 aneurysim
- 1.3 Heart disease (other than 1.1 or 1.2)
- 2 Malignant neoplasm (Cancer)
- 3 Cerebrovascular disease (Stroke, cerebral hemorrhage); CVD
- 4 Chronic lower respiratory disease; CLRD
- 5 Emphysema
- 5.1 Tuberculosis (TB)
- 5.2 Lung (other)
- 6 Unintentional injures (falls, car crash)
- 7 Influenza and pneumonia (P&I)
- 8.1 Suicide
- 8.2 Homicide
- 9 diabetes mellitus (Diab)
- 10 Artherosclerosis
- 11 Chronic liver disease and cirrhosis
- 12 Alzheimer's disease and other dementia
- 13 Nephtiris, Nephrosis, Nephrotic Syndrome (Neph)
- 14 Syphilis
- 15 Other
- 16 Unknown

Next is information on cause of death and infection status (Table 9). Here, C to S conversions are grouped with subjects (the results are not very different if the conversions are excluded). These data are not appropriate for assessing statistical significance (too many categories) but give a descriptive picture of potential associations. For example, aortitis, aneurism, and other heart disease are more prevalent causes of death among subjects than for controls.

Table 9. Cause of death (COD) and infection status, all years

Infection Status			
COD	C	S	TOTAL
aortitis	2	3	5
Row %	40.0	60.0	100.0
Col %	1.8	1.1	1.3
aneurysim	1	4	5
Row %	20.0	80.0	100.0
Col %	0.9	1.4	1.3
Heart, other	33	85	118
Row %	28.0	72.0	100.0
Col %	30.3	29.8	29.9
Cancer	13	31	44
Row %	29.5	70.5	100.0
Col %	11.9	10.9	11.2
CVD	13	25	38
Row %	34.2	65.8	100.0
Col %	11.9	8.8	9.6
CLRD	0	2	2
Row %	0.0	100.0	100.0
Col %	0.0	0.7	0.5
TB	1	7	8
Row %	12.5	87.5	100.0
Col %	0.9	2.5	2.0
Lung, other	5	6	11
Row %	45.5	54.5	100.0
Col %	4.6	2.1	2.8
Unintent Inj	2	10	12
Row %	16.7	83.3	100.0
Col %	1.8	3.5	3.0
P&I	10	23	33
Row %	30.3	69.7	100.0
Col %	9.2	8.1	8.4
Suicide	0	1	1
Row %	0.0	100.0	100.0
Col %	0.0	0.4	0.3
Homocide	0	1	1
Row %	0.0	100.0	100.0
Col %	0.0	0.4	0.3

Diab	1	0	1
Row %	100.0	0.0	100.0
Col %	0.9	0.0	0.3
Artherosclerosis	7	22	29
Row %	24.1	75.9	100.0
Col %	6.4	7.7	7.4
Liver	0	4	4
Row %	0.0	100.0	100.0
Col %	0.0	1.4	1.0
Alzheimers	1	0	1
Row %	100.0	0.0	100.0
Col %	0.9	0.0	0.3
Neph	10	25	35
Row %	28.6	71.4	100.0
Col %	9.2	8.8	8.9
Syphillis	0	16	16
Row %	0.0	100.0	100.0
Col %	0.0	5.6	4.1
Other	6	14	20
Row %	30.0	70.0	100.0
Col %	5.5	4.9	5.1
Unknown	4	6	10
Row %	40.0	60.0	100.0
Col %	3.7	2.1	2.5
TOTAL	109	285	394
Row %	27.7	72.3	100.0
Col %	100.0	100.0	100.0

Again, to investigate the effect of development of medical knowledge, cause of death by decade is given(1930-39, 1940-49, 1950-59, 1960-69, 1970-79). Table 10 presents the number of deaths for each cause in each decade by infection status.

Table 10. Deaths by decade and infection status

Cause of Death: 1930-1939	Controls	Subjects
Aortitis		
Aneurysim		XX
Heart disease (other)	XXX	XXXXXXXXXX
Malignant neoplasm	XX	XXX
Cerebrovascular dis		X
Chronic lower resp dis		
Emphysema		
Tuberculosis (TB)		XXXX
Lung (other)		
Unintentional injures		XX
Influenza, pneumonia	XX	XX
Suicide		
Homicide		
Diabetes mellitus		
Artherosclerosis		XXX
Chronic liver disease		X
Alzheimer's /dementia		
Nephtiris	XX	XXXXXXXXXXXXXXXX
Syphilis		XXXXXXXXXX
Other	XXX	XXXXX
Unknown	X	
Total	13	54

In the first decade, Subjects were more likely than controls to die of heart disease, nephritis, or syphilis.

Cause of Death: 1940-1949	Controls	Subjects
Aortitis		XX
Aneurysim	X	
Heart disease (other)	XXXXXXXXX	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
Malignant neoplasm	XXX	XXXXX
Cerebrovascular dis	X	XXXXXXX
Chronic lower resp dis		
Emphysema		X
Tuberculosis (TB)	X	XXX
Lung (other)		X
Unintentional injures	X	X
Influenza, pneumonia	X	XXXXXX
Suicide		
Homicide		X
Diabetes mellitus	X	
Artherosclerosis	XX	XXX
Chronic liver disease		
Alzheimer's /dementia		
Nephtiris	XXXX	XXX
Syphilis		XXXX
Other	XX	XX
Unknown		X
Total	25	68

In the second decade, heart disease continued to dominate deaths in both groups, disproportionately subjects. Cerebrovascular disease and influenza were also higher in subjects.

Cause of Death: 1950-1959	Controls	Subjects
Aortitis	XX	X
Aneurysm		
Heart disease (other)	XXXXXXXX	XXXXXXXXXXXXXXXXXXXX
Malignant neoplasm	XXXX	XXXXXXXXXXXX
Cerebrovascular dis	XXXXXXXX	XXXXXXXX
Chronic lower resp dis		
Emphysema		
Tuberculosis (TB)		
Lung (other)		XXX
Unintentional injures		XXXX
Influenza, pneumonia	XX	XXXXXXXXXXXX
Suicide		
Homicide		
Diabetes mellitus		
Artherosclerosis		XXXX
Chronic liver disease		XXX
Alzheimer's /dementia		
Nephtiris	XX	XXXXX
Syphilis		
Other		XXX
Unknown	X	XX
Total	26	70

In the third decade, cancer, heart disease and influenza are disproportionately represented among subjects.

Cause of Death: 1960-1969	Controls	Subjects
Aortitis		
Aneurysm		XX
Heart disease (other)	XXXXXXXXXX	XXXXXXXXXXXXXXXXXXXXXXXXXXXX
Malignant neoplasm	XXX	XXXXXXXXXX
Cerebrovascular dis	XXXX	XXXXXXX
Chronic lower resp dis		
Emphysema		
Tuberculosis (TB)		
Lung (other)	XXX	
Unintentional injures		X
Influenza, pneumonia	XX	XXXX
Suicide		
Homicide		
Diabetes mellitus		
Artherosclerosis	XXXX	XXXXXXXXXXXX
Chronic liver disease		
Alzheimer's /dementia		
Nephtiris	X	XXXX
Syphilis		XX
Other		X
Unknown	X	XXX
Total	28	67

Heart disease, cancer, and artherosclerosis are predominant in the 4th decade.

Cause of Death: 1970-1979	Controls	Subjects
Aortitis		
Aneurysim		
Heart disease (other)	XXXX	XXXXX
Malignant neoplasm	X	
Cerebrovascular dis		XXX
Chronic lower resp dis		
Emphysema		X
Tuberculosis (TB)		
Lung (other)	XX	XX
Unintentional injures	X	X
Influenza, pneumonia	XX	
Suicide		X
Homicide		
Diabetes mellitus		
Artherosclerosis	X	X
Chronic liver disease		
Alzheimer's /dementia	X	
Nephtiris	X	
Syphilis		X
Other		XXX
Unknown	X	
Total	14	18

Causes of death for those syphilitic men who had received at least 10 shots during the study (Table 11). Although numbers are small (there are 39 such men; 25 of these have listed cause of death), there is statistical evidence of variation in causes of death over time (Chi-square = 33.8, $p=0.09$).

Table 11. Cause of Death for Men receiving Shots

Decade of Death					
COD	1940-49	1950-59	1960-69	1970-79	TOTAL
Heart	2	3	2	0	7
Row %	28.6	42.9	28.6	0.0	100.0
Col %	66.7	25.0	25.0	0.0	28.0
Cancer	0	4	2	0	6
Row %	0.0	66.7	33.3	0.0	100.0
Col %	0.0	33.3	25.0	0.0	24.0
CVD	0	0	1	1	2
Row %	0.0	0.0	50.0	50.0	100.0
Col %	0.0	0.0	12.5	50.0	8.0
CLRD	1	0	0	0	1
Row %	100.0	0.0	0.0	0.0	100.0
Col %	33.3	0.0	0.0	0.0	4.0
Lung	0	1	0	0	1
Row %	0.0	100.0	0.0	0.0	100.0
Col %	0.0	8.3	0.0	0.0	4.0
P&I	0	3	1	0	4
Row %	0.0	75.0	25.0	0.0	100.0
Col %	0.0	25.0	12.5	0.0	16.0
Artherosclerosis	0	0	1	0	1
Row %	0.0	0.0	100.0	0.0	100.0
Col %	0.0	0.0	12.5	0.0	4.0
Neph	0	1	1	0	2
Row %	0.0	50.0	50.0	0.0	100.0
Col %	0.0	8.3	12.5	0.0	8.0
Syphlis	0	0	0	1	1
Row %	0.0	0.0	0.0	100.0	100.0
Col %	0.0	0.0	0.0	50.0	4.0
TOTAL	3	12	8	2	25
Row %	12.0	48.0	32.0	8.0	100.0
Col %	100.0	100.0	100.0	100.0	100.0