Doll-Hill Progress Report No. 1

1 September 1969

The following report covers the period from the inception of the study until 1 September 1969. In order to fully understand the progress of the study, it seems worthwhile to describe in this first report what constitutes the data set and some of its limitations.

I. Sample Description

The study is concerned with the "pure" smoking experience of 15,938 British doctors during the period 1951-1961. "Pure" smoking is defined as cigarette smoking uncomplicated by pipes, cigars, or mixtures of the three types of smoking. Non-smokers as well as well as ex-smokers are included in this sample but again, these subjects either never smoked cigarettes or smoked only cigarettes at least one year prior to 1951.

The distribution of this sample by age and smoking groups at the beginning of the study is shown in Table 1. A natural question arises: how well does this sample of pure cigarette smokers, ex-smokers, and non-smokers compare to a) the total Doll-Hill population and b) the census population of British doctors. The latter group represents another sampling procedure of the same population. Doll and Hill sampled the British physician population with questionnaires while the Registrar General sampled via census methods.

Table 1

Percentage of British Male Doctors by Age and Smoking Habits in 1951

Age Group	Non Smokers	Ex-Pure Cigarette Smokers	Current Pure Cigarette Smokers: 1-14/Day	15-24/Day	25+/Day	Total
35-44	9.108%	5.827%	7.600%	11.730%	7.191%	41.457
45-54	5.205	5.519	5.475	7.921	6.676	30.796
55-64	2.282	3.420	2.810	3.921	2.653	14.886
65-74	1.446	2.081	2.307	1.537	.993	8.423
75-84	.999	1.220	1.139	.427	.138	3.935
85+	.176	.195	.107	.013	.013	.503
Total	19.316%	18.261%	19.449%	25.308%	17.664%	

Table 2 compares our present sample with the above. It can be readily seen that our sample has a greater concentration of younger physicians than either Doll-Hill's original sample or the Census.

Table 2

Comparison of Three Samples of British Male Physicians by Age Group

	Present Sample		Doll-Hill* Sample		Census** Population	
Age Group	No.	%	No.	%	No.	%
<35	-	-	10,140	29.4	11,204	28.4
35-44	6,595	41.5	8,886	25.6	9,747	24.6
45-54	4,899	30.8	7,117	20.6	8,290	21.0
55-64	2,368	14.9	4,094	11.8	4,879	12.4
65-74	1,340	8.4	2,694	7.8	3,279	8.3
75-84 85+	706	4.4	1,563	4.5	2,108	5.3
Total	15,908		34,494		39,507	

* 1956 Doll-Hill Article (Reference 1)
** Registrar General's Decennial Supplement for England and Wales (Reference 2)

This is probably due to the lack of pipe and cigar smokers in the present sample since these people tend to be older. Dr. Doll did not include any smoking information on the less than 35 year old subjects (113 subjects) in the starting population that he sent us. He did, however, include these subjects in the 2,705 deaths attributed to the initial population. Due to the omission we will be unable to calculate mortality rates for the less than 35 year old group since the population at risk is unknown. We will be able to use these younger subjects when they enter the 35-44 age group and contribute person-years to this cohort. Table 2 presents excellent agreement between the Doll-Hill and the Census populations in regard to the age distribution. The discrepancy noted between the two totals (34,494 and 39,507) may be due to the inclusion of radiologists in the Census population. If this is not the reason, we can offer no other at this time and this question probably should be brought to Dr. Doll's attention.

II. Progress to Date

The data from Doll-Hill consists of the following items:

1) Identification number
2) Age of subject at the start of the study (1951)
3) Number of years subject was in study before his death
4) International Classification of Death code (ICD)
5) Smoking classification of subject where

NON = Non-smoker

EX = Ex-smoker; smoker who stopped smoking cigarettes at least one year prior to the study

C01 = Smoker of 1 to 14 cigarettes/day

C15 = Smoker of 15 to 24 cigarettes/day

C25 = Smoker of 25 or more cigarettes/day

There are 2,705 deaths total over the ten year span (1951-1961) from all causes. Table 3 indicates a breakdown of this total by age and smoking groups.

Table 3

Percentage and Number of Deaths Due to All Causes
By Age and Smoking Groups in the Present Doll-Hill Sample

	Smoking Group									
	Non		C01		C15		C25		EX	
Age Group	No.	%	No.	%	No.	%	No.	%	No.	%
23-34	30	6.8	20	3.4	39	6.4	20	3.8	4	.01
35-44	33	7.5	34	5.7	53	8.7	71	13.6	32	.06
45-54	60	13.6	80	13.5	162	26.7	177	33.3	66	12.2
55-64	73	16.5	126	21.2	162	26.7	139	26.6	122	22.6
65-74	99	22.4	175	29.5	136	22.4	98	18.7	143	26.5
75-84	121	27.4	143	24.1	52	8.6	17	3.3	144	26.7
85+	26	5.9	16	2.7	2	.3	1	.2	29	5.4
Total	442	100.1	594	100.1	606	99.8	523	100.0	540	99.1

Cross-classification frequency distributions have been generated from magnetic tape for the four variables: age at entry, ICD code, number of years in study, and smoking habits. These are shown in the Appendix.

Frequency tabulations and the plotting routines for the major causes of death are being implemented by means of computer programs and will be ready for the next progress report. Computer programs completed as of this date are:

(1) Mantel-Haenszel Summary Chi-Square Procedure

(2) Program to compute the crude mortality rates by the migration and "standard" methods for the major causes of death.

The adjustment procedures for the crude mortality rates have been considered. Two methods, the direct and indirect, are available for use. The direct method, the one Doll used for his adjustment, relates the age-specific death rates of a given community (namely the British male doctors) to some population taken as a standard (for example, the total population of England and Wales). It finds for each age group what the expected number of deaths in the standard population would be if the age-specific mortality of the community were applied. This is accomplished by multiplying the specific rate for each age group by the population for the corresponding age group in the standard population. The adjusted death rate, direct method, is formed by adding the expected deaths for each age group and dividing the sum by the total population that was taken as a standard.

The indirect method is computed by multiplying the crude death rate of the community by an adjustment factor that is designed to take account of the peculiarities of the age composition of

the community. Since the direct method has been used by Doll and the results of the two methods do not differ appreciably, our analysis should utilize the direct method as well. Mention of the other method is only for information purposes.

A more serious question should be considered concerning the adjustment procedure. This relates to the basic problem of converting the physician's death rates to those of the general population, that is, the value of the adjustment process itself. The physicians represent a homogeneous group of people, probably selected for their availability and intelligence. Can the rates of such a specialized group justifiably be related to the heterogenous general population? The purpose of adjustment is for comparability of one community or one time period with another. If smoking is a deterrant to health, why shouldn't this fact show up within the physician population. It seems to us that the adjustment procedure is unnecessary and without merit for our purposes in this study. Of course, we have an obligation to reproduce the Doll-Hill procedure using the "pure" smoking groups and to adjust using the male British physician population but these reasons are empirical and unimportant in the establishment of smoking and health relationships.

The crux of the adjustment problem may lie in statistical nomenclature and in the historical concepts of Vital Statistics. The latter branch of statistics is always concerned with time and

place comparisons and therefore adjustments are required to enable these comparisons to be conducted. In fact, the mortality rate estimates would be meaningless and useless unless they could be compared with some other set. However, the mortality rates of the male British physicians are calculated to determine the existence of a link between smoking and causes of death for this specific occupation. Statisticians consider this a fixed sample rather than a random one. Almost the entire British medical population is included as evidenced by the Census figures. This is hardly a random sample and therefore inferences concerning this and only this population are entitled to be made. Adjustments cannot alter this restriction since the general population of British males was not sampled. It is suffice to say that the physicians and the general population differ in many ways and to sample one and make inferences concerning the other seems quite erroneous and misleading.

Expected Deaths

Most smoking and health studies will attempt to assess the total mortality experience of the population under study. Doll and Hill neglect to do this and explain the "figures for total mortality should not, therefore, be interpreted until the mortality of each of its principal disease components has been separately studied" (p. 1402 Reference 3). This premise can be easily argued

especially when other investigators (Hammond) have elected to perform such an analysis.

It should be noted that the calculation of expected deaths which the authors perform in the 1954 and 1956 papers (References 1 and 4) is not performed in the 1964 paper (Reference 3). The expected deaths are utilized in the chi-square test. They state, "The numbers of deaths in most of these categories are so large that tests of statistical significance are hardly necessary." (P. 1401 Reference 3). This presents a definite inconsistency because later on in the 1964 paper (Reference 3) probability values are quoted for the various causes of death. Where were these probability levels derived from if not as the results of statistical significance tests. We would assume then that significance tests were performed and in the same manner as in the 1954 and 1956 papers. It would seem in order to examine the total mortality experience in this report even though Doll and Hill do not.

The 1954 paper (p. 1452 Reference 4) details the method the authors used to conduct these significance tests. We have similarly performed the same test but using our present sample which contains only "pure" smokers. Table 4 is identical to Table I of Reference 4.

Table 4

Percentage of Male Smokers Surviving by Age at Commencement and Years of Death

Cigarette Smokers

Age in Years	Non Smokers	C01	C15	C25	Ex Smokers	Total
35-44	22.0	18.3	28.3	17.3	14.1	100.0
45-54	16.9	17.8	25.7	21.7	17.9	100.0
55-64	16.0	18.9	24.3	17.8	23.0	100.0
65-74	17.2	27.4	19.0	11.8	24.7	100.1
75-84	25.4	29.2	10.9	3.5	31.0	100.0
85+	35.4	21.5	2.5	2.5	38.0	99.9

The statistical significance of the difference between death rates can be more easily assessed from the actual numbers of deaths recorded; that is, by comparing them with the numbers which would have been expected to occur in each smoking category if smoking were unrelated to the chance of dying.

Table 5

Total Number of Deaths Due to All Causes by Age Group

Age Group	Number
23-34	113
35-44	223
45-54	545
55-64	622
65-74	651
75-84	477
85+	74
Total	2705

To calculate the expected number of deaths, multiply the percentages for a particular age group from Table 4 by the total

number of deaths for that age group in Table 5. If mortality is unrelated to smoking, then the total deaths from Table 5 for each age group should be distributed in the same proportion as the smoking groups of Table 4. This then is the expected number of deaths as shown in Table 6.

Table 6

Expected and Observed Number of Deaths Due to All Causes by Age and Smoking Group

Age Group		Non Smokers	Cigarette Smokers C01	C15	C25	Total
35-44	Expected	43.0	40.9	63.1	38.7	31.3
	Observed	33	34	53	71	32
45-54	Expected	92.1	96.9	140.2	118.1	97.7
	Observed	60	80	162	177	66
55-64	Expected	99.6	117.4	151.3	110.8	142.9
	Observed	73	126	162	139	122
65-74	Expected	111.7	178.3	123.4	76.8	160.8
	Observed	93	175	136	96	143
75-84	Expected	171.2	139.4	51.8	16.8	147.8
	Observed	121	143	52	17	144
85+	Expected	25.9	15.7	1.8	1.8	27.7
	Observed	26	16	2	1	29
Total	Expected	499.5	588.6	531.6	363.0	608.2
	Observed	412	574	567	503	536

The observed numbers are the actual deaths due to all causes by age and smoking groups at the end of the study. The totals represent the total expected deaths in each smoking group and the

total observed number. These are the values subjected to the chi-square test shown in Table 7.

Table 7

Difference Between Expected and Observed Deaths Due to All Causes for Chi-Square Determination

Smoking Group	None	C01	C15	C25	EX
Difference	87.5	14.6	-35.4	140.0	72.2
Chi Square	15.33	0.36	2.36	53.99	8.57
Significance	SS	NSS	NSS	SS	SS

SS = Statistical Significance at $p<.05$
NSS = Not statistically significant $p>.05$

The results are very interesting and show the lack of any smoking gradient. The total chi square is highly significant but this is due to the large contribution of the C25 group. The non-smokers represent the second largest contribution which certainly is contrary to previous findings reported in the literature.

The machinery has been set up to perform this analysis on each of the major causes of death listed in the contract.

III. Interesting Avenues for Investigation

It seemed interesting to compare the Doll-Hill deaths with that of the Registrar General's (Reference 2) Census for male British doctors. The Doll-Hill study covered a ten-year (1951-1961) period while the Census data is for the five-year period (1949-1953). We chose malignant neoplasms, all sites, ICD Codes 140-205

to examine this comparability more thoroughly the nine population. The deaths [illegible] when converted to the number occurring in one year to equalize the time span. Since the time span overlaps but fails to coincide, the Census deaths for the doctors should reflect better diagnostic procedures which should lead to more treatment successes.

Table 8

Age Group	Doll-Hill Sample (Regardless of Smoking Category) No.	Per Yr (÷10)	Registrar General's Census Data (No Smoking Information) No.	Per Yr (÷5)
<35	24	2.4	14	2.8
35-44	41	4.1	18	3.6
45-54	100	10.	53	10.6
55-64	129	12.9	92	18.4
65-74	131	13.1	149	29.8
75+	58	5.8	193	38.6
	483		519	
Population at Risk:	15,908		39,507	

This problem is far from resolved in our mind but note the wide discrepancy between the deaths/year for the last two age groups. This sort of exploration, that is, using material other than the Doll-Hill sample for comparative purposes seems worthy of further investigation and may shed light on some sampling problems in the Doll-Hill report.

Another item that should be looked into is the mortality rate calculation using the individual ages rather than age groups.

Our experience indicates that the average of an age group is rarely the midpoint but some other value usually close to one of the bounds. It should not be difficult to calculate rates for each age with the use of a computer program.

We have also become aware of a paper by Sprigett[5] which cautions against using long study intervals in the study of lung cancer, due to rapid rate of change of mortality from this cause. We intend to split the Doll-Hill sample into two five-year groups to see whether deaths from ICD cause 162 differ in the two cohorts.

Summary

The progress to date has been the production of frequency distributions of the cross-classified variables, implementation of computer programs to produce the required plots and distributions by major causes of death, and development of algorithms to calculate person/years by the migration and "standard" methods. The computer printouts are available for your inspection at your request.

REFERENCES

(1) R. Doll and A. Bradford Hill, "Lung Cancer and Other Causes of Death in Relation to Smoking", British Medical Journal, 11-10-1956, pp. 1071-1081.

(2) The Registrar General's Decennial Supplement England and Wales, 1951, Occupational Mortality, Part II, Vol. 2, p. 138, Table 3A(i).

(3) R. Doll and A. Bradford Hill, "Mortality in Relation to Smoking: Ten Years' Observations of British Doctors", British Medical Journal, 1964, 1, 1399-1410.

(4) R. Doll and A. Bradford Hill, "The Mortality of Doctors in Relation to Their Smoking Habits", a preliminary report, British Medical Journal, June 26, 1954, pp. 1451-1455.

(5) V. H. Springett, "The Beginning of the End of the Increase in Mortality From Carcinoma of the Lung", Thorax, 21:132-138, March 1966.

ICD Code	Definition
2	Pulmonary tuberculosis
151	Malignant neoplasm (stomach)
153	Malignant neoplasm of large intestine, except rectum
154	Malignant neoplasm of rectum
157	Malignant neoplasm of pancreas
162	Malignant neoplasm of bronchus and trachae, and of lung specified as primary
177	Malignant neoplasm of prostate
181	Malignant neoplasm of bladder and other urinary organs
331	Cerebral hemorrhage
332	Cerebral embolism and thrombosis
*336	Cerebral hemorrhage with hypertension
*419	Coronary thrombosis with mention of hypertension
420	Arteriosclerotic hear disease, including coronary disease
422	Other myocardial degeneration
450	General arteriosclerosis
451	Aortic aneurysm
491	Bronchopneumonia
502	Chronic bronchitis
610	Hyperplasia of prostate
*999	Violent deaths

* Refers to classifications devised by Doll-Hill and therefore differs from ICD Code

TABLE II

Distribution by Smoking Group

ICD Code	Smoking Group					
	C01	C15	C25	EX	NON	Total
2	4	7	10	5	2	28
151	11	12	8	12	10	53
153	14	5	9	7	10	45
154	3	3	9	4	2	21
157	9	5	7	2	4	27
162	22	51	55	10	3	141
177	5	7	4	9	11	36
181	6	3	3	4	6	22
331	19	13	14	18	14	78
332	30	28	15	33	27	133
336	13	16	11	16	14	70
419	11	13	8	8	9	49
420	170	184	145	156	112	767
422	48	24	16	39	21	148
450	12	4	5	7	10	38
451	4	11	6	4	1	26
491	6	12	4	11	11	44
502	12	18	16	13	2	61
610	7	6	2	8	8	31
899	37	34	36	24	41	172
Total	534	606	523	540	442	

TABLE III

Distribution By Years In Grade

	Years in Grade										
[illegible] Code	1	2	3	4	5	6	7	8	9	10	Total
2	2	3	1	7	1	2	4	3	0	4	28
151	5	3	3	6	5	1	8	7	10	5	53
153	3	5	10	8	2	5	3	3	3	5	45
154	4	2	3	1	3	2	2	2	1	1	21
157	2	1	4	2	3	3	2	4	2	4	27
[illegible]	9	8	10	15	16	19	12	10	24	12	141
177	2	3	6	2	7	3	3	1	3	3	36
181	3	1	2	0	3	1	1	3	4	4	22
331	7	5	11	9	6	8	10	7	6	9	78
332	7	12	13	18	8	14	18	19	15	9	133
336	15	6	7	1	8	7	8	4	8	6	70
419	2	3	7	3	6	3	7	5	6	7	49
420	56	63	69	75	77	66	98	93	79	93	767
422	7	14	10	14	17	15	17	18	16	20	148
450	0	3	5	7	7	3	4	0	5	4	38
451	1	1	3	2	4	3	1	6	3	2	26
491	5	4	1	3	2	0	6	10	7	6	44
502	2	6	7	7	6	2	9	4	5	13	61
610	6	3	5	4	2	4	1	2	3	1	31
6[illegible]	16	15	16	13	19	17	22	19	17	16	172
Total	211	232	253	272	264	259	321	291	292	310	

5570 9455

TABLE [illegible]

Distribution by [illegible]

Years in Study	Smoking Group					
	C01	C15	C25	EX	NON	Total
1	58	43	31	44	35	211
2	52	52	34	47	47	232
3	67	47	54	46	39	253
4	55	60	54	51	42	272
5	63	58	52	47	43	264
6	55	63	51	46	44	259
7	62	72	60	78	49	321
8	66	54	65	62	44	291
9	58	76	58	52	48	292
10	58	81	63	57	51	310
Total	594	606	523	540	442	2705

TABLE V

Distribution by Age Group

Years in Study	Age Group							
	23-34	35-44	45-54	55-64	65-74	75-84	85-102	Total
1	9	10	30	39	51	56	16	211
2	8	13	38	52	45	63	13	232
3	9	15	47	63	67	47	7	253
4	11	25	41	62	67	54	11	272
5	11	21	56	52	61	53	10	264
6	12	14	43	71	70	35	4	259
7	13	27	67	75	88	48	3	321
8	11	28	72	65	63	48	4	291
9	13	33	76	70	62	35	3	292
10	16	34	70	72	77	38	3	310
Total	113	223	545	622	651	477	74	2705