On the Distribution of Underlying Causes of Death

ALAN M. GITTELSOHN

Abstract: The feasibility of applying surveillance techniques to large health data sets is being explored through study of a national mortality data base encompassing 21 million United States death records for the period 1968–1978. Through the development of efficient file structures and information recovery techniques, it is possible to pose a series of questions and follow-up questions of the entire data set within budgetary constraints. Initial screening of the mortality data base reveals that major changes have occurred over the 11 years with marked declines for diseases of cardiovascular, respiratory, digestive and renal systems, and maternal and perinatal mortality. There is a tendency for increased usage of non-specific terminology. The occurrence of unlikely and unusual causes in the data set is documented and reasons for their inclusion discussed in terms of underlying cause of death logic. Problems in the study of geographic distributions of cause specific mortality are outlined with illustrations of the dispersion of standardized mortality ratios for major causes of death over areas of the country. Clusters of high mortality areas require interpretation in terms of underlying dispersion and possible reporting artifacts arising out of geographic differentials in diagnostic labeling practice. (*Am J Public Health* 1982; 72:133–140.)

Introduction

Surveillance, meaning oversight or close supervision, is based on the acquisition of raw data which are evaluated and organized so that coherent patterns can be perceived. The application of surveillance techniques to health data systems implies monitoring for the detection of changes in levels of incidence and the occurrence of sentinel health events. The requisites are a current flow of reliable information into an accessible data base and an efficient method of data recovery. Presently, we are exploring the feasibility of applying such methodology to the death reporting system for the United States.

Mortality reporting has attained virtual completeness for several decades with the expansion of the death registration area to the entire nation. The occasional deaths missed through disappearance are few in number, the only significant problem in underreporting occurring in certain sections of the country in relation to fetal deaths and newborns dying shortly after birth.[1] Beyond completeness of ascertainment, the quality of the information on the death record tends to be high for characteristics such as age, sex and date of death, intermediate for residence, and uncertain and variable for causes of death. A review of the literature on the reliability of statements of cause of death by the author[2] suggested that major problems exist. These are related to the current state of medical knowledge, to incomplete information available at the time of death, and to the variable practice of physicians in completing the medical certification on the death record. The matter is compounded by classification procedures which require assignment of a single underlying "cause" of death, implying presence or absence of the condition without allowance for uncertainty. For this reason, changes in styles of labeling diagnoses may be confused with trends in mortality rates, and differential labeling practice between communities may be confounded with differential levels of mortality. Clearly, a surveillance system cannot improve on the quality of information contained in the data base. Rather, it must deal with the material incorporated therein, taking cognizance of the possibilities of reporting artifacts.

The present report is concerned with initial findings on cause of death distributions based on total US mortality coded according to the Eighth Revision of the International Classification of Diseases (ICD8).[3] The approach is one of exploring the mortality data base for the country as a test and demonstration of the application of surveillance methodology.

Material and Methods

The mortality data set presently comprises 21 million United States death records for the period 1968–1978. The information, derived from the Public Use Tapes of the National Center for Health Statistics (NCHS),[4] has been condensed into a minimum record including the items of age, race, sex, date of death, county of residence, and underlying cause of death. Population data are based on the 1970 Census and annual county estimates by age, race, and sex prepared by the Bureau of the Census for the period 1970

Address reprint requests to Alan M. Gittelsohn, Professor, Department of Biostatistics, Johns Hopkins School of Hygiene and Public Health, 615 North Wolfe Street, Baltimore, MD 21205. This paper, submitted to the Journal December 22, 1980, was revised and accepted for publication October 2, 1981.

Editor's Note: See also related editorial p 125 this issue.

through 1977.[5] Estimates for the years preceding and following this period have been developed by linear extrapolation. As additional information becomes available, the material is incorporated in the files in order to maintain currency. The two files, one for deaths and one for population, constitute a national mortality data base which can be drawn upon to develop mortality rates for specified groupings of cause by time and place. Details of processing methods are contained in Appendix A.

General Observations

Examination of the frequency distribution of ICD8 for the 21 million deaths reveals the diversity of underlying cause assignments in the data base. Over the 11 years, 2,358 distinct causes occurred at least once with a mean of 9,051 deaths per four-digit cause and a standard deviation of 1,683. The maximum of the distribution was the 3.4 million deaths ascribed to ICD 410.9 "acute myocardial infarction without mention of hypertension." Causes with fewer than 500 deaths represent 80 per cent of the cause assignments and comprise under 5 per cent of all deaths. Several admissible causes had no occurrences, notably the infectious diseases which have been controlled through preventive and therapeutic measures. Females had a wider variety of cause assignments than males because of the more detailed classification of gynecological as contrasted with male genital causes, and because of the Group XI causes relating to pregnancy, childbirth, and the puerperium.

Table 1 summarizes changes in mortality rates by age and by major cause over the decade of ICD8. The total number of deaths remained almost constant at 1.9 million per year while the population increased at a rate of .8 per cent per year averaging 210 million between 1968 and 1978. Significant declines in total mortality occurred during the period for all race, sex, and age groups, the single exception being White males aged 15–24 for whom the death rate remained unchanged. For most age groups, rates for non-Whites declined more rapidly than rates for Whites, resulting in an important decrease in the mortality differentials between the two groups. Sex differentials persisted largely unchanged over the 11 years. Mortality, measured by age adjusted rates, declined between 10 per cent and 40 per cent for the four race-sex-groups for each major cause group, the exception being malignant neoplasms with a 3 per cent increase for females and an 8–13 per cent for males, largely due to the continuing rise in lung cancer death rates. Infant mortality and mortality from cerebrovascular diseases decreased by nearly one-third, diseases of the digestive and genito-urinary systems by about one-fifth, and diseases of the heart by over one-sixth. Mortality from respiratory causes and external causes in the aggregate decreased by over 10 per cent, particularly among non-Whites. These marked downward trends over the decade of ICD8 present a challenge to the mortality analyst in their interpretation. It is against them that changes in death rates for specific causes should be measured.

Trends toward Nonspecificity

Examination of reporting trends for particular causes of death suggests a growing tendency towards use of non-

TABLE 1—Annual Deaths per 10,000 by Age, Race, and Sex and Major Cause Group, United States, 1968–1969 and Per Cent Change to 1977–1978

	1968–1969 Deaths Per 10,000 Per Year				Per Cent Change to 1977–1978			
	White		Non-White		White		Non-White	
Age	F	M	F	M	F	M	F	M
< 1	159	213	335	418	−28%	−31%	−32%	−33%
1–4	7	9	14	17	−24	−16	−36	−36
5–9	3	5	4	7	−19	−15	−30	−27
15–24	6	17	11	29	− 5	− 1	−28	−31
25–34	9	18	24	53	−20	−10	−40	−28
35–44	20	35	53	91	−25	−21	−40	−28
45–54	46	89	103	171	−14	−17	−28	−22
55–64	101	223	200	319	−10	−18	−23	−14
65–74	254	498	387	584	−18	−15	−24	−14
75+	913	1,201	806	1,040	−14	− 8	− 8	− 8
Cause of Death								
All Causes*	74	123	104	158	−15	−13	−22	−17
Malignant neoplasms	13	20	15	24	+ 3	+ 8	+ 3	+13
Heart	29	51	36	51	−19	−17	−22	−19
Cerebrovascular	10	11	15	17	−27	−30	−36	−35
Respiratory system	4	9	5	11	−14	−11	−43	−31
Digestive system	3	5	3	6	−21	−20	−14	−11
Genito-urinary system	1	2	3	4	−21	−26	−22	−28
External causes	4	11	6	21	−14	−11	−22	−23

*Rates adjusted by direct method using combined US age distribution, 1968–1978.

specific terminology. Trends for selected causes during the ICD8 period are exhibited in Table 2. Unspecified septicemia (ICD 38.9) more than doubled while most other infectious and parasitic diseases declined. Among neoplasms, unspecified site (ICD 199) increased by 38 per cent. For non-Hodgkins lymphoma, the one-third decline in lymphosarcoma and reticulum cell sarcoma (ICD 200) has been compensated by a doubling of unspecified primary malignancies of lymphoid tissue. While most diseases of the heart have declined, other forms of heart disease increased by 39 per cent, particularly cardiomyopathy (ICD 425) and symptomatic heart disease (ICD 427). The 29 per cent decrease in mortality from cerebrovascular diseases is noteworthy for the rapid decreases in cerebral hemorrhage and thrombosis and the almost constant trend for generalized, unspecified, and ill-defined categories. All pneumonias declined by 31 per cent while unspecified pneumonia (ICD 486) increased 9 per cent. The 25 per cent decline for all diseases of the genito-urinary system was accompanied by a doubling in the rate for unspecified diseases of the kidney and urinary tract. Senility and ill-defined death rates have increased by 14 per cent, the major component being the six-fold increase in sudden infant death.

The increasing trends in mortality from non-specific and ill-defined categories are evident for each of the major groupings of the International Classification. Septicemia and bacterial disease not elsewhere classified now constitute half of all infectious and parasitic disease deaths. Secondary and site unspecified malignancies are the fourth leading type of cancer among males after lung, colon, and prostate. Generalized, unspecified, and ill-defined cerebrovascular diseases constitute 58 per cent of all stroke deaths. Unspecified pneumonia (ICD 486) has increased to encompass 55 per cent of all pneumonia deaths while ill-defined disease of the kidney and urinary tract have doubled to 42 per cent of all diseases of the genito-urinary system. An analogous phenomenon may have occurred with ischemic heart disease (ISHD), acute myocardial infarction having declined 30 per cent while chronic ISHD declined 9 per cent.

Unusual and Unlikely Causes

Underlying cause of death means the initiating step in the chain of circumstances leading to death. Application of this principle in certain situations can lead to aberrations in

TABLE 2—Annual Death Rates and Per Cent Change for Selected Causes of Death, United States 1968–69 and 1977–78

Cause of Death	ICD8 Codes	Deaths Per Million Per Year		Per Cent Change
		1968–69	1977–78	
All Causes		9,979	8,481	− 15%
Infectious and parasitic diseases	000–136	87	79	− 9
Septicemia, unspecified	038.9	17	38	+123
Malignant neoplasms	140–209	1,655	1,757	+ 6
Secondary, site unspecified	195–199	106	122	+ 14
Site unspecified	199	51	70	+ 38
Non-Hodgkin's lymphoma	200,202	48	50	+ 4
Lymphoma, unspecified	202.2	13	27	+101
Chronic rheumatic heart disease	393–398	80	58	− 27
Hypertensive disease	400–404	135	70	− 48
Ischemic heart disease	410–414	3,522	2,813	− 20
Acute myocardial infarction	410	1,904	1,344	− 30
Other ischemic	411–414	1,618	1,469	− 9
Other forms of heart disease	420–429	191	267	+ 39
Cardiomyopathy	425	6	23	+311
Symptomatic heart disease	427	93	181	+ 95
Cerebrovascular disease	430–438	1,104	781	− 29
Cerebral hemorrhage	431	236	103	− 56
Cerebral thrombosis	433	316	176	− 44
Generalized, unspecified CVD	436–438	482	456	− 5
Pneumonia	480–486	331	229	− 31
Specified	480–484	61	36	− 40
Bronchopneumonia, unspecified	485	155	67	− 57
Pneumonia, unspecified	486	116	126	+ 9
Diseases of genito-urinary system	580–629	154	116	− 25
Other diseases of kidney	593	15	31	+102
Other diseases of urinary tract	599	10	18	+ 87
Senility and ill-defined causes	790–796	111	126	+ 14
Sudden death—all ages	795	5	26	+469
infants	795	227	1,609	+610
Other ill-defined	796	95	93	− 2

Rates adjusted to total US age distribution—1968–1978.

ICD assignments. The International Classification is designed for indexing both mortality and morbidity; for the latter many entries are unlikely as causes of death in terms of current understanding of disease processes. For the most part, these are conditions affecting external organs generally regarded as being of a non-life threatening nature. A complete listing of such occurrences from the ICD8 decade is too extensive for inclusion. Examples are the 773 deaths assigned to diseases and conditions of the eye including two refractive errors, 42 strabismus cases, and 428 cataracts. Respiratory diseases include 282 deaths due to the common cold, 368 to hypertrophy of the tonsils, eight to deflected nasal septum, and 20 to nasal polyp. Dental caries, carbuncles of the extremities, bunion, synovitis of the toe, and polydactyly are additional examples of unlikely causes of death which occur throughout the data set.

There are several possible explanations for admissible causes of death such as dental caries, strabismus, and bunion. The first is that of miscoding which is bound to arise in all data systems. The number of these instances appears to be extremely small when measured against the 21 million death records in the file. In any event, no estimate of coding error can be derived by simple tests for internal consistency without reference to external data sources. The second explanation is to be found with the medical certifier and a breakdown of the querying process at the local and state levels. The third lies with the rules for assigning underlying cause of death which explicitly state that deaths associated with medical interventions are to be assigned to the condition for which the intervention was initiated.[6] Thus, the logic of ICD classification is that a child dying from aspiration of blood following tonsillectomy be counted as tonsilar hypertrophy and that a death due to anesthesia during cholecystectomy be counted under cholelithiasis. The information about enlarged tonsils and gall stones is of small value since both conditions are ubiquitous and generally not threatening to life. Such an explanation can account for many of the deaths due to conditions for which opthalmologic, cardiovascular, gynecological, neurological, and orthopedic procedures are initiated. The mortality reporting system, by capturing only a single cause and using an arbitrary assignment rule, is precluded from providing information about potentially avoidable and preventable problems resulting in death.

Table 3 displays a selected group of deaths occurring in the United States during the ICD8 decade assigned to conditions and diseases for which surgery is commonly performed, and ICD 930–932, deaths due to complications and misadventures in operative, therapeutic, and diagnostic procedures. Estimates of the total number of surgical procedures performed in the country during the period are based on the Hospital Discharge Survey of NCHS.[7] Combining these with conservative estimates of short-term surgical case fatality rates[8,9] provides a general notion of the magnitude of surgical deaths occurring in the nation. For example, the estimate of 4.5 million cholecystectomies performed and a case fatality rate of one per cent yield 45,000 surgical deaths associated with the one procedure. The 2.8 million prostatectomies, most often performed in older men for benign hypertrophy of the prostate, are estimated to have case fatalities ranging between 1 and 4 per cent. Using the lower figure yields an estimated 28,000 deaths following the surgery. These are to be contrasted with the 23,518 deaths assigned to ICD 930 "surgical complications" which is necessarily an understatement of the true incidence.

Geographic Variations

Investigation of geographic variations in mortality rates presents a complex series of problems, particularly with diagnoses which are subject to variable interpretation. The first of these pertains to the possibility of the confounding of the decedent's place of residence with the practice location of the cause of death certifier. A tendency to favor a particular set of diagnostic labels in one community may well account for an excessive death rate. A hypertensive diabetic suffering an acute myocardial infarction while hospitalized with a broken hip resulting from a fall can be described in widely varying terminology. Residence as a decedent characteristic may lose its meaning for long-term diseases since the disease itself can result in selection of residence. The

TABLE 3—Deaths due to Conditions for Which Surgery is Commonly Performed and Estimated Number of Surgical Deaths United States 1968–1978

ICD Code	Cause of Death	Number Reported	Expected Surgical Deaths	Estimated Case Fatality Rate	Estimated Number of Surgeries (in millions)	
374	Cataract	428	4,450	.0013	3.4	lens extraction
454	Varicose veins	2,806	1,000	.0010	1.0	ligation
455	Hemorrhoids	310	4,180	.0018	2.3	hemorrhoidectomy
500	Hypertrophy of tonsils	368	1,860	.0002	9.3	tonsillectomy
540–543	Appendicitis	11,774	11,400	.0030	3.8	appendectomy
550–553	Abdominal hernia	27,542	16,100	.0028	5.7	herniorrhaphy
574–576	Cholelithiasis	53,601	45,000	.0100	4.5	cholecystectomy
600–602	Diseases of prostate	23,268	28,000	.0100	2.8	prostatectomy
930	Operative procedures	23,518				
931	Therapeutic procedures	7,879				
932	Diagnostic procedures	2,518				

high rate of chronic obstructive pulmonary disease in Arizona has been explained on the basis of asthma and emphysema patients seeking a desert climate. Retirees moving to Florida may account for the high rate of lung cancer observed in several communities in that state.

The 3,075 counties identified in the data system range in size from several hundreds to over 7 million. In order to achieve stability of rates, the geographic units we selected for study are the 509 state economic areas (SEAs) of the country defined by the Census Bureau with larger counties treated separately to form 600 area groupings. Age has been subdivided into ten-year groups. The time period was restricted to the most recent five-year period available, the full 11 years of data encompassing causes which have undergone marked changes between 1968 and 1978. In order to simplify matters, the issue of race and sex differentials has been avoided by restricting attention to White males.

The basic calculations involve production of a three-dimensional age by area by cause frequency table $\{D_{ijk}\}$ and a corresponding two-dimensional person-year count table $\{N_{ij}\}$ for deaths occurring between 1974 and 1978. Causes have been selected with an occurrence rate of at least 10 per million per year among White males (see Appendix B). Measures of geographic variability are based on direct adjusted rates, (DARs), standardized mortality ratios (SMRs), and their variances.

Table 4 exhibits the variability of mortality over the 600 areas of the country for malignant neoplasms of the digestive and hematopoietic systems. Space does not permit inclusion of all sites. Mortality rates at the extremes of the distribution for each cause are presented when the normalized rate for an area has a Z-value of at least 3. By large sample criteria, a normalized rate is significantly different from the national average $\bar{R}$ when $|Z| > 1.95$ at the 5 per cent level. For each of the causes listed, standard tests for the equality of the SMRs $\{\theta_j\}$ are rejected leading to the conclusion that variation in the rates is statistically significant. Primary cancer of the colon, the leading digestive system site and the second most common neoplasm after lung cancer with $\bar{R} = 194$ per million per year, exhibits nearly a fourfold range from 84 in Ocone, South Carolina to 293 in Camden, New Jersey.

The DARs $\{R_j^*\}$ and the SMRs $\{\theta_j\}$ are subject to random variation which contributes to the range of variation. The systematic variance of the SMRs for a cause, $V(\theta)$ has

TABLE 4—Variability in Mortality from Malignant Neoplasms of the Digestive and Hematopoietic Systems, US White Males 1974–1978

		Extreme Areas						
Site	U.S. Rate $\bar{R}$	DAR R_j^*	SMR θ_j	z_j	Years 1000S N_j	Area(j)	State	SMR Variance $V(\theta)$
Colon	194	84	.41	−4	259	Ocone	SC	.042
		89	.46	−5	535	Alexander	NC	
		293	1.51	+5	504	Campbell	KY	
		293	1.51	+7	985	Camden	NJ	
Pancreas	104	50	.48	−3	418	Utah	UT	.010
		55	.51	−3	467	Kalamazoo	MI	
		173	1.67	+4	316	Acadia	LA	
		179	1.74	+3	167	Bullock	AL	
Stomach	81	16	.23	−4	119	Chester	SC	.062
		27	.32	−3	178	Surry	VA	
		139	1.72	+9	1,773	Providence	RI	
		146	1.81	+9	1,306	Suffolk	MA	
Rectum	52	6	.15	−7	399	Beaufort	SC	.135
		8	.26	−7	475	Anoka	MN	
		108	2.07	+5	277	Fulton	NY	
		137	2.51	+5	151	Franklin	MA	
Liver	15	1	.10	−4	539	Barton	MO	.133
		2	.15	−3	463	Atcheson	KS	
		51	3.16	+6	573	Nueces	TX	
		62	4.27	+5	212	Crockett	TX	
Leukemia	86	41	.40	−3	824	Alaska	AK	.010
		42	.49	−3	384	Hancock	ME	
		163	1.92	+3	167	Bullock	AL	
Lymphoma	58	16	.27	−3	551	Arapaho	CO	.028
		17	.33	−4	312	Colfax	NM	
		84	1.42	+4	1,266	San Mateo	CA	
		91	1.56	+3	505	Kanawha	WV	
Hodgkins	14	1	.11	−5	637	Cambria	PA	.052
Disease		2	.14	−4	568	Bay	FL	
		34	2.41	+3	331	Buncombe	NC	
		36	2.64	+3	201	Brookings	SD	

DAR R_j^* adjusted by total US White Male Population.
$z_j = (R_j^* - R\cdot)/S.E.(R_j^*)$ normalized rate.

the random component removed and therefore provides an additional insight into variability of rates. In terms of SMR variance, primary malignant neoplasms of the rectum and the liver are the most variable of the sites displayed with $V(\theta) = .135$ and .133 respectively. The cancer death record reliability study of Percy, *et al,*[10] based on the Third National Cancer Survey, reports a false negative rate of 44 per cent for rectal cancer and 50 per cent for liver cancer. The corresponding false positive rates of 14 per cent and 23 per cent suggest that both tumors are significantly underreported on the average. No estimates of area differences in reliability are presented by the authors. It can be inferred that such differences are bound to exist perhaps explaining in part the wide range in mortality rates for the two tumors. By contrast, the two sites exhibiting the least geographic variation, are leukemia and pancreatic cancer, both with $V(\theta) = .01$. False positive and false negative rates for the latter two cancers are reported as being about 10 per cent for pancreas and 4 per cent for leukemia.[10] The extent to which differential physician reporting practice contributes to the observed variation in cancer mortality for each site cannot be estimated without mounting a national study aimed at validating death certificate diagnoses with the best available clinicopathology information.

Table 5 displays variability of adjusted heart disease mortality rates in the range ICD400–429 for entities with US rates exceeding 10 per million per year. The extreme area rates are significantly different from the US rate with $|Z| > 3$ and in the lower or uppermost percentiles of the distribution of the 600 mortality rates. Acute and chronic ischemic heart disease (ISHD), the leading causes of death, vary by a factor of more than threefold between extremes. The possibility of differential physician labeling practice between acute and chronic forms of the disease is present with the per cent acute of all ISHD ranging from under one-third in Alameda, California to nearly 80 per cent in areas of Georgia, Tennessee, and Kentucky.

The systematic variance of SMRs for heart disease entities, $V(\theta)$, is high when contrasted with the malignant neoplasms, particularly for the symptomatic and ill-defined diagnostic labels. A hypertensive heart and renal disease cluster (ICD404) is evident in Franklin, Ohio with an SMR of 19.45, representing a rate about 20 times that recorded in the rest of the country. More extreme is ICD 411 "other acute and subacute forms of ISHD" which includes such diagnoses as coronary failure, coronary insufficiency, and intermediate coronary syndrome. Fifteen per cent of all ICD 411 deaths in the country occurred to residents of Oklahoma and Maricopa, Arizona with SMRs of 24 and 15 respectively. High intrastate variability occurred with the assignment of congestive heart failure (ICD 427.0) as cause of death in Connecticut, rates varying from 34 per million in Hartford to 589 in New Haven. Other myocardial insufficiency (ICD 428), which includes fatty degeneration, myocardial disease, and myocarditis NOS, is highly localized in Montgomery, Maryland with an SMR of 36. If the 519 deaths in the latter county are deducted from the US total of 6,726, the Montgomery rate of 524 per million is to be contrasted with 13 per million for the remainder of the country.

Ill-defined heart disease (ICD 429) comprises generic terminology including cardiac decompensation, organic heart disease, cardiac hypertrophy, and ventricular dilata-

TABLE 5—Variability in Mortality from Diseases of the Heart, US White Males 1974–1978 (600 Areas)

ICD Code	Heart Diseases	US Rate $\bar{R}$	DAR R_j^*	SMR θ_j	Years 1000S N_j	Area	State	SMR Variance $V(\theta)$
402	Hypertensive heart	23	132	5.64	259	Marathon	WI	.229
			0	.00	370	Benton	AR	
403	Hypertensive renal	19	89	4.64	208	Wood	WV	.178
			1	.08	1,496	Fairfax	VA	
404	Hypertensive heart and renal	16	309	19.45	1,794	Franklin	OH	.465
			0	.00	754	Richmond	NY	
410	Acute ischemic	2,008	3,289	1.62	126	Boyd	KY	.052
			970	.45	859	Bernalillo	NM	
411	Subacute ischemic HD	25	603	23.89	1,136	Oklahoma	OK	.897
			387	14.79	2,867	Maricopa	AZ	
			0	.00	594	Pulaski	AR	
412	Chronic ischemic HD	1,588	2,986	1.86	1,176	Hudson	NJ	.065
			680	.39	295	Gwinnett	GA	
425	Cardiomyopathy	22	57	2.48	387	Clark	WA	.156
			2	.11	378	Alcorn	MS	
427.0	Congestive heart failure	73	589	8.01	1,669	New Haven	CT	.366
			34	.46	1,826	Hartford	CT	
427.2	Cardiac arrest	75	455	6.23	576	Hancock	MS	.365
			11	.15	627	Kane	IL	
428	Myocardial insufficiency	15	524	35.69	1,261	Montgomery	MD	.424
			2	.15	1,826	Hartford	CT	
429	Ill-defined heart disease	45	864	19.46	320	McLennan	TX	.987
			809	18.27	417	Brown	WI	
			.656	15.02	820	Colbert	AL	
			0	.00	786	Burlington	NJ	

tion. As such, it is a highly variable label, constituting a leading cause of death in several Texas, Alabama, and Wisconsin areas. Rates above 800 per million were recorded in McClennan, Texas and Brown, Wisconsin. Of the 25 rates above 300 per million, 13 occurred in Texas communities. The lowest decile, with rates below six per million, included large metropolitan counties such as Los Angeles, San Diego, and Orange (California) and Queens and Nassau (New York), each with more than three million person years. Six of the lowest 25 areas were New Jersey counties including Burlington with 0 deaths to be contrasted with an expected number of 28. The high geographic variability of the uncertain cause ICD429 would appear to be related to geographic differences in certification practice. Ill-defined heart disease mortality is of such a magnitude that it may account for the low ISHD mortality observed in several areas of the South East and South West Central Regions of the country.

The present methods of scanning the mortality file for cause of death clusters have been illustrated for neoplasms and diseases of the heart. Space does not permit similar presentation for other groups of diseases. An initial aspect of mortality surveillance should include the detection of outliers and measurement of variability as a surrogate for estimating the statistical reliability of a cause. The existence of high frequency indeterminate causes throughout the ICD list introduces major uncertainty into rates for specific causes particularly when the former are highly variable. Examples include the ill-defined and secondary malignant neoplasms, ill-defined and symptomatic heart disease, and the Group XVI causes of symptoms and ill-defined conditions.

Summary

We have outlined current approaches to scanning large health data bases, using US mortality for the 11-year period 1968 through 1978. An essential first step is a file design whereby subsets can be accessed easily and inexpensively. This entails designing efficient code structures and data storage for rapid recovery of information. A second requirement is the design of flexible and reasonably efficient programs with control specification of variables, subsets, and tables. The program generator is dictated by the size of the mortality file which comprises 21 million records for the study period. Thereby, the control specifications generate the specific machine language code for the particular application at hand.

Application of these techniques to screening the US death file reveals that major shifts in mortality have occurred over the decade and that major mortality differentials exist between areas and by race and sex. A pronounced feature of the data set is the tendency toward increased usage of nonspecific terminology. Unlikely causes of death in the data set have been described and it has been speculated that their usage is derived from the underlying cause logic of ICD. Dental caries or strabismus as cause of death is not relevant information from a public health standpoint when the death was probably iatrogenic. The matter could be improved immediately by maintaining multiple cause information in the file. Several problems in the interpretation of spatial distributions of mortality have been outlined.

The first objective of developing a surveillance methodology for health data bases is to define reporting problems in order to assist in the design of programs for improving data quality. The second is to screen the files to detect unusual occurrences and changes over time in order to help specify potential health problems and to evaluate the impact of health programs. Tools for approaching both objectives have been described as including an efficiently organized data base, an efficient and flexible system of information recovery, and a strategy for search. The latter requires historical public health perspective and wide contact with current developments, implying the need for collaborative inputs from public health and clinical medicine.

REFERENCES

1. McCarthy B, Terry J, Rochat R, Quave S, Tyler C: The underregistration of neonatal deaths. Am J Public Health 1980; 70:977–982.
2. Gittelsohn A: Cause of Death Validation: Review of Literature and Annotated Bibliography. Final Progress Report NCHS HRA 230-77-0032, Washington, DC, September 1978.
3. Eighth Revision of International Classification of Diseases Adapted for Use in the United States, Washington, DC: US Dept. HEW, Public Health Service Pub. No. 1693.
4. Standardized Micro-data Tape Transcripts. Washington, DC: US DHEW, NCHS, DHEW Pub. No. (PHS) 78-1213.
5. US resident population by age and sex, Series P-25, No. 721, Washington, DC: US Bureau of Census.
6. Vital Statistics: Instructions for Classifying Underlying Cause of Death, 1976–1977. Washington, DC: DHEW, NCHS, 1975.
7. Surgical Operations in Short Stay Hospitals, US 1973, Washington, DC: DHEW Pub. No. (HRA) 76-1775, 1976.
8. Surgery in the United States: A Summary Report of SOSSUS American College of Surgeons and the American Surgical Society, 1975.
9. Bunker J: Comparison of hospitals with regard to outcomes of surgery. Health Services Research, Summer 1976, 112–127.
10. Percy C, Stanek E, Gloeckler L: Accuracy of cancer death certificates and its effect on cancer mortality statistics. Am J Public Health 1981; 71:242–250.
11. Gail M: The analysis of heterogeneity for indirect standardized mortality ratios. J R Statist Soc A 1978: 141, 224–234.

ACKNOWLEDGMENTS

The work has been supported by the National Center for Health Statistics, RFP No. 233-78-2090. Processing the three billion bytes of mortality data received from NCHS would have been impossible without the guidance of Dr. James Tonascia of the Department of Biostatistics, Johns Hopkins School of Hygiene and Public Health. Marie Diener, Lucy Mead and William Huster, graduate students in the Department, have made significant contributions to this project.

APPENDIX A

The programming system used to process the mortality data set is termed CHOMPS for "comprehensive health management programming system." It has evolved over the past decade in dealing with population based health data sets including hospitalization, ambulatory care, morbidity, natality, and mortality under the current National Center for Health Statistics (NCHS) mortality surveillance project. Written in extended Fortran, the program has been adapted

for use in several computing environments including the IBM 3031 at Johns Hopkins Hospital. It is a general tabulator and event rate maker driven by control specifications developed with the notion that the public health analyst be enabled to process data sets without the requirement for professional programmer intervention. The code produced by Fortran has been compared with Cobol and PL/I and found to be reasonably efficient. A CHOMPS manual is available from the author. CHOMPS has been bench-marked against general program packages and found to be faster by an order of magnitude.

Most of the processing costs are engendered by reading the file and computing table subscripts. Recognition of this fact has led to the development of a program generator whereby the control specifications generate specific machine language code for input into the tables. This process is transparent to the user. The input routine is assembled and then linked with a fixed control specification editor and a fixed output subroutine to form an operating load module.

File design is an important aspect of efficient event rate generation. Several designs have been explored for the mortality project. The following binary files are presently in use:

File Type	Item	Byte	Bits	
Both	Cause of death	1–2		ICD code
	Year of death	3		68, 69, . . ., 79
	Age	4	1–6	Ten year groups
	Race	4	7	White, other
	Sex	4	8	
	County	5–6		MSP county code
SORT8	Month of death	7	1–4	
	Day of death	7–8	5–9	
	Age in years	8	2–8	0 to 126+ years
MULT8	Multiplicity	7–8		

The multiplicity file MULT8 stores the data as a six-dimensional frequency tabulation by cause, year, age, race, sex, and county of residence. It represents a 60 per cent data reduction over the unit record file SORT8 with a corresponding gain in processing times. The entire US mortality file for the period 1968–1978 occupies one-third of a tape reel recorded at 6250 BPI. Compared with the NCHS public use tapes, the MSP MULT8 represents a 50 to 1 data reduction. An additional 5 to 1 reduction is achieved by binary rather than character storage. Overall, the gain in processing times is more than two orders of magnitude. A similar approach has been used with the county population file.

APPENDIX B

For $\{D_{ijk}\}$ deaths due to cause k in age group (i) and area (j) and $\{N_{ij}\}$ person-years, the age specific rate is given by

$$R_{ijk} = D_{ij}/N_{ij} \tag{1}$$

The direct adjusted rate (DAR) is the weighted average

$$R^{*}_{jk} = \sum_i W_i R_{ijk} \tag{2}$$

where the weights $\{W_i\}$ are based on the internal standard

$$W_i = N_{i\cdot}/N_{\cdot\cdot} \tag{3}$$

The variance of the DAR R^{*}_{jk} is approximated by

$$V(R^{*}_{jk}) \doteq \sum_i W_i^2 R_{ijk}/N_{ij} \tag{4}$$

using standard Poisson assumptions. The normalized rate is

$$Z_{jk} = (R^{*}_{jk} - \bar{R}_{\cdot k})/SE(R^{*}_{jk}). \tag{5}$$

Large sample theory permits treatment of the $\{Z_{jk}\}$ as normal (0, 1). The standardized mortality ratio (SMR) is based on the expectancy

$$E_{jk} = \sum_i R_{i\cdot k} N_{ij} \tag{6}$$

with the SMR

$$\theta_{jk} = D_{jk}/E_{jk}. \tag{7}$$

A test for age homogeneity is given by Gail[11] as

$$SSW = \Sigma\Sigma(D_{ijk} - E_{ijk})^2/E_{ijk} \tag{8}$$

which is approximately χ^2 with $(I - 1) \times (J - 1)$ degrees of freedom. When dealing with national samples, the null hypothesis is rejected for virtually all causes of death except those low occurrence rates suggesting that the criterion is too sensitive to minor perturbations in the data. Gail provides a test for the equality of SMRs based on

$$SSB = \Sigma(D_{jk} - E_{jk})^2/E_{jk} \tag{9}$$

which is approximately χ^2 with $(J - 1)$ degrees of freedom. The test is rejected for almost every cause of death {k}, again suggesting that the chi-square criterion is far too sensitive when analyzing national mortality samples over a decade. An approximate measure of the variability of the $\{\theta_{jk}\}$ is given by MacPherson* with

Total $$V_T \doteq \sum \log_e(\theta_{jk})^2$$

Random $$V_R \doteq \frac{1}{J - 1} \sum 1/E_{jk}$$

Systematic $$V(\theta_{jk}) = V_T - V_R$$

As a descriptor of variability in mortality rates, the systematic variance $V(\theta_{jk})$ provides a means for characterizing the distribution of a cause over time or geographic areas. Neoplasms, particularly lymphoma, tend to have low variances while the less certain causes have high variances.

*MacPherson K: Personal communication, 1981.