COMPUTER IDENTIFICATION OF SPANISH SURNAMES[1]

Edwin W. Jackson, M.D.[2] and Robert Buechley, Ph.D.[3]

Knowledge about a special group's health needs often requires statistical identification of the group in question. A new computer method of identifying Spanish surname groups is described here. It provides an effective alternative to the laborious hand methods previously used for this purpose.

Introduction

Nearly ten per cent of all Californians are of Mexican-American descent. Almost 2,000,000 strong, they constitute the largest cultural sub-group in the State. Yet, in spite of their numbers, their health status has been difficult to describe. Birth and death certificates and most medical records fail to identify Mexican-Americans as members of a distinct ethnic or racial group. For this reason, except for a few special studies, there has been very little information available that could help in planning and evaluating health programs for Mexican-Americans in California.

In general, the people of Mexican descent have maintained their cultural integrity. They have distinct social characteristics ranging from continued use of the Spanish language and traditional food habits to particular patterns of family formation and religious practice. Some health facilities serving this group have shown they recognize its distinctness by recruiting Spanish-speaking nurses and social workers, by preparing Spanish language manuals, and by providing Spanish training for other health workers. However, very little has been done to accurately assess the health problems having a particularly important or unique impact on the Mexican-American, or to provide basic data for establishing some order of priorities (*1*). The assumption has been that because Mexican-Americans represent a "minority" culture and tend to be in a lower economic class they must have high morbidity and mortality rates. However, little evidence has been presented to document this view.

This study describes a method of developing health indices by applying a computer program to vital statistics. Resulting preliminary data on infant, fetal, and perinatal mortality rates are presented and discussed.

Methods

Historical Background

Since 1930 the U.S. Census Bureau has used various approaches in attempting to enumerate and describe the Mexican population of the southwestern United States. Methods employed in 1930 and 1940 have been shown inadequate. In 1950 and 1960 a different procedure was used; data on persons of Spanish-American and Mexican-American origin were obtained by identifying Spanish surnames during the census coding procedure. This approach has proved satisfactory, and will be used again in the 1970 Census. The procedure requires regular coders to classify a name as Spanish only if it appears on a list of about 7,000 names. Other names of apparent Spanish origin are referred to specialists, who differentiate them from those pertaining to other Romance languages.

Americanization of Spanish surnames, and use of non-Spanish surnames that are spelled the same as Spanish ones, impose some limitations on this method. In addition, proper interpretation of Spanish surname data often requires knowledge about the subjects' country of birth or their parents' country of birth in order to properly define the group in question.

---

[1]Published in Spanish in the *Boletín de la Oficina Sanitaria Panamericana,* Vol. LXIX No. 5 (November 1970), pp. 436-441.

[2]California Department of Public Health, Bureau of Maternal and Child Health, Berkeley, California, United States of America.

[3]Ecological Research Branch, National Air Pollution Control Administration, Durham, North Carolina, United States of America.

In most areas of California, Spanish surname data refers generally to a population of Mexican origin or descent. In other locales, people with Spanish surnames may be predominantly from Cuba (Florida) or Puerto Rico (New York); or they may come from a wide array of Spanish-speaking countries with no particular one predominating (Washington, D.C.).

Computer Program

The major obstacle to using census data for health planning is that it requires the user to obtain additional information. For instance, the census publication tells how many Californian males 20-29 years old have Spanish surnames (*2*). If a health planner wants to know their death rate, he must determine the number of deaths in this group before the census-provided "denominator" can be used. To determine this, some way of finding Spanish surnames must be applied to individual health records such as vital data, disease reports, or medical survey questionnaires.

If the census coding method is used, special training, supervision, and additional clerical personnel are usually required. It thus becomes a slow and costly procedure, particularly when large numbers of records are involved. These difficulties can be overcome if the surnames to be entered in the vital or medical records are key punched for other purposes (to make registry listings and billings, for instance), with a computer program being employed to make the Spanish surname determination. The problems of the hand coding method, combined with the lack of basic health data on California's Spanish surname population, led to the decision to develop and apply a computer program to provide new health information for this group.

The computer program was developed in the California Department of Public Health by Robert W. Buechley. The program, as now designed, uses a set of general rules supplemented by a few short name lists. The general rules are based on examination of each surname's terminal letters, to a depth of four or sometimes five letters. These rules are as follows:

Rule O: If this surname is the same as the previous surname, make the same decision. This is just a "money saver", as it requires only one decision tree to be climbed for each surname. Whether the names are listed alphabetically or by families, the number of decisions is cut drastically.

Rule 1: No Spanish names end in B,C,F,H,J,K,P,Q,V, or W. All names ending in these letters are rejected.

Rule 2: Only very exceptional Spanish names end in D,G,M,T,X, or Y. Names ending in these letters are checked against a single very short list of surnames. If found they are accepted, otherwise they are rejected.

Rule 3: Only a few Spanish names end in L,N, or R. These names can be listed individually; each name ending in one of these letters is then checked against its proper list of surnames. Only those names found are accepted.

Rule 4: Spanish names ending in E, I and U are also rare. So far, all RRE, GUI and ZU names investigated have turned out to be Basque. If the ending is not one of these, a single list of E's, I's, and U's is consulted. It contains PONCE, ARCE, ALOU, CANTU, etc. If found they are accepted, otherwise they are rejected.

The four rules leave only names ending in A,O,S, and Z. The Z rule is:

Rule 5: All names ending in Z preceded by a vowel are Spanish. Of the "Z" names not preceded by a vowel, only SAENZ is Spanish. It is put in the "exception" list used in applying Rule 2.

Rule 6: Names ending in S may be Spanish.

Rule 6a: If the name ends in ES or IS the whole name is compared with a specific list and accepted only if found on that list.

Rule 6b: If the name ends in AS or OS, the final four letters before the "S" are treated like the last four letters in words covered by rule 7 below.

Rule 7: Spanish names ending in A and O are the most common and most troublesome. Many characteristic Spanish endings are not found in Italian or Portuguese, but many others, such as GRECO, are common to all three languages. All A and O names are checked against A and O endings to a depth of four letters. If not found there, they are checked against A and O name lists. If not found, they are rejected.

Rule 8: Names accepted because of characteristic endings are checked against a short

list of non-Spanish names with Spanish endings. These include BARRE, CICERO, GROZA, SOUZA and MAYO.

This computer approach provides quick acceptance or rejection without requiring computer reference to long lists. The set of rules using just the final letter of each name permits about 80 per cent of the names to avoid lists entirely, leaving 20 per cent to be tested against a series of short lists. The present cost of doing this for the State of California's birth tapes is $65 per 100,000 names, amounting to a total cost of $225 for Spanish surname identification of all 350,000 California births in 1969.

The computer approach also offers complete reproducibility, and gives results which are comparable to the Census Bureau's hand coding technique. The computer calls over 95 per cent of all Spanish surnames Spanish, and calls only 2 per cent of non-Spanish names Spanish. It identifies about 8 per cent more Spanish surnames than does the 1960 Census hand coding method, so that only a simple adjustment is required for use of census figures to calculate rates.

Results and Discussion

The program was run using taped records of births, together with those of fetal and infant deaths, to make a traditional calculation of mortality rates in infancy. The results, shown in the table below, did not indicate any marked differences between Spanish surname and non-Spanish surname neonatal and postneonatal mortality rates. The Spanish surname fetal mortality rates were slightly higher, but not significantly so.

Further analysis considering the father's occupational level was also performed. The Spanish surname perinatal mortality rate was found to be approximately that of the non-Spanish surname group at all occupational levels, excepting farm laborers. For the latter, the respective rates of the Spanish and non-Spanish surname groups was 37.9 and 30.0 per 1,000 live births.

The Mexican-American population in California is largely urban, with less than 15 per cent living in a rural setting. The high perinatal mortality rates in the farm labor groups point to an area of continued concern for maternal and child health programs in farming regions, and suggest that other high risk subgroups should be sought out.

Further work is needed and will be carried out. Specifically, county mortality rates will be calculated to seek out possible high risk areas within the State, and analysis of infant mortality for rural and urban areas is indicated. Finally, as 1970 Census figures become available, the Spanish surname population's age and sex mortality rates for selected diseases will be derived. It is believed that this approach will provide new information needed for health

TABLE 1—Births and fetal, neonatal and postneonatal deaths by Spanish surname and race. (California 1966, preliminary figures).

Race or group	Live births	Fetal death rate[a]	Neonatal death rate[b]	Postneonatal death rate[b]
White				
Spanish surname	61,946	14.1	13.5	5.5
Non-Spanish surname	232,234	10.6	14.6	4.5
Negro	31,564	19.7	23.5	8.8
Other non-white	12,092	6.8	16.8	5.8
Total	337,836	11.9	15.3	5.2

[a]Rate per 1,000 total births.
[b]Rate per 1,000 live births.
Source: State of California, Department of Public Health, Birth and Death Records.

planning and extension of health services. It should also prove a useful tool for medical and epidemiological studies.

Summary

Vital health statistics on the United States of America's Hispanic-American population have been lacking because of the difficulty of identifying this group. Now computer methods offer a rapid and efficient alternative to the manual name classification techniques previously employed. A computer program devised for this purpose permits Spanish surname classification of most names according to a few simple rules; those that remain are checked against special name lists. Applied to the State of California's computerized birth records, this method cost $225 for Spanish surname identification of 350,000 names. The program is easily reproducible and quite sensitive. More than 95 per cent of the true Spanish surnames are so classified, while only 2 percent of non-Spanish names are read as Spanish. When this program was applied to California's 1966 births and infant deaths it showed an infant mortality rate in the Spanish surname population comparable to those of the white non-Spanish surname group. Analysis by fathers' occupational levels showed the perinatal mortality rate of Spanish surname farm workers to be higher than that of the total Spanish surname population.

REFERENCES

(*1*) Morton, W. E. "Demographic Redefinition of Hispanos." *Public Health Rep 85*:617-623, 1970.

(*2*) United States Bureau of the Census. *The U.S. Census of Population, 1960; Subject Reports; Persons of Spanish Surname; Final Report PC (2)* - 1 *B*. U.S. Government Printing Office, Washington, D.C.