Creating American Indian and Asian American Life Tables

1. Introduction

This grant request seeks funds to support the US portion of an international effort to produce life tables for Indigenous peoples in Australia, Canada, New Zealand and the United States.¹ All four countries have a similar history of being invaded by Europeans who treated the new lands they “discovered” as their own, pushing aside the first inhabitants or worse.² The creation of life tables for the American Indians/Alaska Natives [AI/AN], Native Hawaiians and Asian Americans [aka Asian] has not been carried out in the United States.³ This is so despite their creation regularly for other US population subgroups after each decennial census.⁴ Developing the Life tables will improve Vital Health Statistics for American Indian’s in numerous federal and private agencies. If funded this project will prototype the needed AI/AN and Asian life tables based on data from the following proposed states, Arizona, California, Montana, New Mexico, and Washington.⁵ The methods of using record linkages, while done in other settings [e.g., Jaro 1995, Winkler 1995 and Arias 2011] to create life tables are untested for these populations and difficult. That linkage will be our main challenge in this project⁶. While we will produce point estimates, our focus will be on measuring uncertainties and data gaps, hence on range estimates. Armed with these insights we will seek funding to scale up the effort nationwide.

¹ We are also seeking funds for Asian American life tables, as these Americans have many points in common with American Indians and are in fact ethnically and racially similar [ Gibbons 2012 ]. At a later time in , perhaps in a separate grant, we will work to produce life tables for Native Hawaiians. This is because following, Census Standards Native Hawaiians and AI/AN are two separate race categories.

² In Australia the date of first European contact is called “Invasion Day.” There are many tellings of this history. A recent popular, but broadly accurate, version is that by Diamond [ 1998]. Perhaps the biggest initial disaster that befell Indigenous people was the (initially largely inadvertent) introduction of smallpox, which decimated the native peoples. The just sainted Kateri Tekakwitha, a Mohawk Indian woman, was badly disfigured as a young girl by smallpox but supposedly one of the miracles cited at her canonization was that at death her disfigurement is said to have disappeared [Greer 2012]. Her shrine is found in a small reservation (REZ) on an island just outside Montreal in Canada. St. Kateri, also known as “Lily of the Mohaks”, is the first North American Indian to be sainted.

³ The use of these two designations is clearly to offer a label that oversimplifies groups whose members have different backgrounds [Lott 1998]. The use of the acronym AI/AN for American Indian and Alaskan Native. Native Hawaiians will be a separate category keeping with Census standards. It is important to note that the AI/AN designation consist of 566 Federally Recognized and many other States recognized Tribes. We will be focusing on Federally recognized Tribes [Federal Register / Vol. 77, No. 155 / Friday, August 10, 2012 / Notices. Page 47868]. Asians can be divided into west Asians (e.g., from Turkey, Iran, Armenia, Saudi Arabia), South Asians (primarily Pakistan and India) and East Asians (primarily China, Japan, and Korea). Asians are defined to a person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent, including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam. It includes people who indicated their race(s) as “Asian” or reported entries such as “Asian Indian,” “Chinese,” “Filipino,” “Korean,” “Japanese,” “Vietnamese,” and “Other Asian” or provided other detailed Asian responses. AI/AN Incidentally, the team assembled to do this work is primarily American Indian, with contributions from those of Asian and European heritage, as well.

⁴ The US National Center for Health Statistics (NCHS) life tables based on the 2010 Decennial year, can be found in separate reports for NonHispanic Whites [Anderson 1999], NonHispanic Blacks[Anderson 1999], and most recently for Hispanics [Arias 2010 ]. Currently, as noted, there are no successful attempts creating life tables for AI/AN and Asian Americans.

⁵ We need California to do our Asian Pilot with at least two other States. We intend to use Montana and Washington for the AI/AN life tables and have team members from Tribes in those States. We will also use Hawaii at a later point to create Native Hawaiian life tables.

⁶ Also look at Statistics New Zealand 2006.
Unquestionably, we expect, from work done elsewhere [ABS 2005; Ajwani et al. 2003, Health Canada 2003] that there may be significant differences in the mortality experience of American Indians, say, versus persons of European stock. The life table data and methodology proposed have the potential to identify and, ultimately, alleviate health disparities among the AI/AN people, saving lives, and increasing equality of opportunity.

After this brief introduction (Section 1) background on the AI/AN and Asian subpopulation measurement issues are found in Section 2. In Section 3, to illustrate the issues we focus on the largely solved problem of producing differential US infant mortality statistics by race and ethnicity [e.g., CDC WONDER]. In Section 4 we look at the issue of reliably coding American Indian and Asian American ethnicity. The typical, commonly used, linkage software developed in a cultural context different from that of the AI/AN and Asian populations. Even though this is just a proposal to do research a great deal of research done by others has already been reviewed. The bibliography available separate has over 200 references.

7 In the case of Asians we expect the life tables to show a mortality experience that for Americans of European stock. This is so, despite the fact that racially the ancestors of American Indians may all be from Asia [e.g., Gibbons 2012; Wallace 1992]

8 Given major health differentials and disparities between American Indians and the general population, it seems unconscionable that there has been no successful effort to attempt to create life tables. This lack of information has proven to be destructive [e.g., Langwell, 2007]. American Indian Tribal Governments, Indian Health Services, and other organizations who deal with American Indians do not have the kind of information contained in life tables to respond effectively to these disparities. The information resulting from AI/AN life tables may potentially save many lives.
Sections 5 and 6 cover two conventional concerns, how age misreporting and net coverage error could potentially affect AI/AN and Asian American life tables. In Section 7 we discuss one of the main tools we will use, record linkage [Scheuren et al 2007]. The record linkage literature, while extensive, has some gaps that we hope to fill in with this project. As laid out in Section 8, we emphasize the calculation of measures of uncertainty. The tool of choice here is Multiple Imputation or MI [Rubin 1987]. While cited in the record linkage literature [e.g., Larsen 1999] no completely worked out approach yet exists. MI is a tool natural for handling the inherent uncertainty that arises when carrying out probabilistic record linkages. Section 8, also, spells out our analysis plans, once the AI/AN and Asian American life tables are in hand.

There are two remaining sections (Section 9 and 10). Section 9 describes some of the early stages of work we will be partnering with tribal nations. Section 10 as required by the National Science Foundation, gives a sketch of our semination/dissemination plans. Whether we succeed or not we will be providing full details on what is done. If successful those details will help others extend our work. If unsuccessful, we will provide enough details so our mistakes can be avoided. Either way the science will be advanced.

The afterword deals with some of the organizational complexities that exist in carrying out this project. After all, we have to deal with several sets of gatekeepers and stakeholders various federal and state government agencies and programs, as well as tribal governments. Limited funding for this support has been requested but we have no illusions regarding the long and arduous task of this undertaking. To see the vast portions of Indian Country please refer to map 19. The Tribal governments, however, may be the key to our success. Difficulties stemming from the residual anger of past injustices, institutional traditions, and other antipathies have made it difficult to facilitate full cooperation in the past and we expect that work be the case again despite the project staffing.

2. Background
What we are envisioning builds on the work done by others, most recently [Arias 2010] for Hispanic Americans. We face two of the challenges Arias encountered in her report for the National Center of Health Statistics (NCHS).¹⁰

1. Most notably is the degree to which the population of interest was stably coded and reliably identified over a person’s entire lifetime. This, if anything [e.g., Bertolli 2007], poses a bigger problem for AI/AN than it did for Hispanics.

2. Another issue Arias successfully dealt with and which we have to address is the misreporting of date of birth and hence age. Of course, errors in the reporting of age directly impact on life tables and must be “corrected.”

There are still more concerns:

¹⁰ Several collaborators at NCHS, led by Sam Notzon, are currently partnering with us. Other international collaborators from Australia, Canada and New Zealand are involved and their past work and anticipated roles in this project will be discussed later.
3. The 2010 Decennial Census counts of AI/AN and Asian Americans are at best suspect. AI/AN, particularly, may be seriously undercounted at least in some age groups [e.g., Alexander 2010] and Asians potentially overcounted [e.g., Waksberg 2000], albeit perhaps only slightly. Our approach, thus, will be to rely on administrative (i.e., birth) record data to the extent possible for the denominator of the life table death rates, employing the reporting of the mother to determine ethnicity.

4. The numerator of the death rates, the National Death Index (NDI), also has problems. Coding of ethnicity is an issue; and, potentially, from our own experience there may be an undercount of AI/AN deaths on some reservations.

There is one further point of difference to mention before moving toward the successful creation of AI/AN and Asian life tables, relative to those of the larger American subpopulations (Europeans, African Americans and Hispanics). The AI/AN and Asian populations are smaller, by a factor of perhaps 10 or more. This means that relationships are inherently less stable and can be affected by “edge” or misclassification issues. The whole idea of using annual life tables comes into question here too. A time series approach, if feasible, would be better for “sampling” reasons [Deming 1950] and may be affordable within the budget we are requesting.

Central to addressing several of these concerns is the use of computerized record linkage techniques. For this project, we hope to make small advances in existing linkage methods. Naturally, as developed below, we will build on the iconic work of Fellegi and Sunter [1969] but also on many recent papers.

To review, in the Fellegi-Sunter world (and still today), there are three linkage decisions: True Links, True Nonlinks and a third category, Indeterminate, where more data or better data are needed to make a decision. The Fellegi-Sunter approach, under a model, presets error bands for the decisions on which cases are treated as True Links or True Nonlinks. If the estimated true link/nonlink probabilities, under the model chosen, fall outside these bands no computerized decision is made and further review, presumably with more data, has to be done before an acceptably low error rate can be arrived at.

This is a sound theory if three assumptions hold: (1) The probability model is roughly right (sadly, seldom true in practice), (2) The errors that remain are small enough to be de minimus (so that the linked cases can be analyzed without further work, also an approach unlikely to be adequate) and (3) there are enough additional data to manually determine with the same or a lower error rate the true status of the Indeterminate cases (again seldom possible, even in the days (no more), when extensive manual reviews were affordable [Scheuren et al 1972-1980].

We have one more motivating perspective to add. We conjecture that AI/AN and Asian life tables would look different, when finished, from those of the majority European subpopulation and also from major African American and Hispanic subpopulations (referenced earlier). And, of course, some of these differences may be a challenge to us for action. Our analysis of the resulting life tables will, therefore, be comparative. Particularly telling may be the analysis

---

11 See [Haldane 1956] for a delightful treatment of this issue.
12 Scheuren and Winkler 1993, 1997, among many others found in the extensive bibliography that accompanies this project description.
13 There are historical and cultural differences between American Indians and other Americans. Health differences and disparities are also wide. American Indians tend to have greater prevalence rates for diabetes, alcohol and drug abuse, injuries from car accidents, and teen suicide.
between American Indians and Asians, each racially, albeit not culturally, roughly a counterfactual proxy for the other (See Section 8.) In any case we will employ these as working hypotheses to be validated or refuted as part of our work under this project.

3. Infant Mortality Illustration

The purpose of this project, as already noted, is ultimately the creation and analysis of at first two sets of life tables, one for the American Indians/Alaska Natives, and the other for Asian Americans in the United States\(^\text{14}\). As a start on accomplishing this goal, we are looking at linkage statistics compiled between infant birth and death certificates by race.\(^\text{15}\).

![Figure 1. Neonatal deaths per 1,000 live births](image1)

![Figure 2. Postneonatal deaths per 1,000 live births](image2)

Figure 1 is an illustration of trends in neonatal deaths, which are deaths of infants in the first month of life and while figure 2 indicates trends in postneonatal infant deaths or deaths occurring in the next 11 months of life. These infant mortality rates were obtained through a linkage of State birth certificates to death certificates for the years 1983-2007.

Figure 1 shows neonatal deaths between/among different ethnicities. Neonatal is defined as 0-28 days of age. As shown in figure 1, AI/AN and Asian rates fall in line with the other ethnicities, with the exception of African Americans where low birth weight babies are common than elsewhere. Figure 2 depicts postnatal deaths across race and ethnicity. Postnatal is defined as from 29-days to a year old. This graph shows how AI/AN population’s postnatal deaths are the highest among all race/ethnicities until 1990 when there is a sharp drop between the years 1990 to 2000, paralleling a similar drop in African American infant death rates. At the end of the period African American and AI/AN infant death rates are about the same, both, though, remain higher than for the other ethnic groups graphed.

The CDC compiles the data it receives from each State. It is up to each State to provide the data in a consistent fashion. But there are missing data problems for AI/AN and Asian infants, potential problems of inconsistent coding by ethnicity, potential inconsistencies in data linkage determinations over time and/or over geography, Thus, we believe more work is needed.\(^\text{16}\)

---

\(^\text{14}\) At a later point we will attempt to produce life tables for Native Hawaiians. This, as mentioned, is because Native Hawaiians are categorized differently using Census standards from the rest of the AIAN group.

\(^\text{15}\) The statistics are by year and ethnicity. The data from these graphs were compiled from the CDC WONDER system (http://wonder.cdc.gov).

\(^\text{16}\) There has been an attempt by NCHS to develop some consistency in the linkage process by providing back to the States for their consideration a trial linkage of birth and death records. But the development of full paradata and metadata remain unfinished. One of the ways we intend to examine the consistency of coding over time and space is to cross-tabulate the final State codes against the provisional ones assigned by NCHS analysts. We will use the paradata as part of a sensitivity analysis of any disparate results obtained.
State has its own confidentiality rules and problems of aggregating across geography, say, by Census Division cannot be done with the publicly available datasets because of different suppression rules[http://wonder.cdc.gov]. Thus, even in a best case scenario there is a lot of work to do in this project. Population counts by State and by Tribe may be worth exploring [e.g., Norris 2012] despite small area estimation issues [Scheuren 2012] and the confidentiality concerns already mentioned.

4. American Indian Identity

American Indian identity is based on Tribal affiliation. Individuals within tribes can be enrolled members or descendants whose ancestors were enrolled members. Tribes maintain records of enrolled members and descendants. Enrollment criteria are set by each Tribe. Currently, there are 566 federally recognized Tribes; many States recognized Tribes and still other Tribes, seeking Federal and/or State recognition. The recognition process is complex and lengthy.

The complex and legalistic recognition process evolved since the time of first encounter with Europeans in this country. Most early American Indian identity was through kinship. Historically lineage was always the dominant way to show ones Tribal descent. After European contact a colonial practice of blood quantum started that began to evolve an alternative concept of an American Indian identity. Over generations a shift occurred from an identify coming from indigenous social-cultural-territorial-based definitions to legal race-based definitions arising through Congressional laws, administrative regulations and Supreme Court Cases. The Treaty Period (1817-1871), Reservation Period (1871-1887), and Allotment Period (1887-1934) each in its own way led to permanent shifts in the measurement of American Indian identity:

1. The Treaty Period set up references to blood quantum to define entitlement to specific properties or lands, not Tribally-defined membership.

2. The Reservation Period setup legislative and federal regulations to encourage, even at times coherence, assimilation of Indians into American society.17

3. The allotment period created by the General Allotment Act (also known as the Dawes Act) was an attempt by the federal government to set up an assimilation policy to disband tribal entities by encouraging individual family framing, thus ridding the US federal government of their responsibility to tribal members as “guardians and protectors.” 18

Thus, American Indian identity is a significant historical issue and remains contentious. There is, even, a lot of turmoil when it comes to American Indian identity classifications, and among American Indians themselves. This problem has been forced upon the American Indian culture through these historical events. The identity problem can directly affect the linkage of the individuals involved who are American Indian. There will remain, as we will see later, a degree

---

17 As oversight by the Bureau of Indian Affairs (BIA) intensified a question of blood quantum came into play more and more. The BIA wanted to establish blood quantum’s for memberships of individual tribes. In Elk v Wilkins [112.U.S.94, 99 [1884]]the Supreme Court ruled that Indian Tribes have the ability for self-determination of citizenship in the tribe.

18 The Dawes Commission was established to set up Tribal Rolls to establish allotments. The question of blood quantum was established again. Many, now called “mixed-bloods,” were denied allotments because they were US citizens. Until the late 1960s, in order for them to be called US Citizens, they had to assimilate into American Society, denouncing all of their American Indian identity. (See also Graham and Scheuren 2012)
of indeterminacy that is not resolvable at a reasonable price, given the current vital record systems available for linkage. It is here that we intend to employ multiple imputation; not to resolve the issues, rather to characterize their impacts.

5. American Indian Issues in Self Re-naming

American Indians have numerous issues when it comes to reliable data. American Indian Tribes are sovereign nations which means they do not have to share their tribal membership data with any federal, state, or local entity. Because of this federal and state agencies, often, do not have the needed data. This causes coverage error issues when attempting to understand the dynamics of American Indian populations.

Probabilistic record linkage does not account well for these and other “renamings” of American Indians or even Asians. The system for record linkage, as, say, set out the Fellegi-Sunter model has Eurocentric cultural aspects; thus, it cannot work well without significant modifications in “Indian Country.” It has been set up for a European culture, originally a northern European culture. It can be made to work in a Hispanic setting and with more work, potentially, in an Asian setting. It does not account well, though, when confronted with the challenge of American Indian identity issues. See Section 7 below for more on linkage.

6. Asian American Identity and Issues

The history of Asians in the United States started with discrimination. When Asians first immigrated to America, like many immigrants before them, they were welcomed as low-cost, albeit, also, low-skilled laborers. Over time, Asians were to face racial prejudices, in the form of the Chinese Exclusion Act of 1882, the Immigration Act of 1917, and the National Origins Act of 1924. After the Immigration and Nationality Act of 1965 was passed, waves of new immigrants started to arrive from Asia (and Africa too).

If you fast forward to 2012; according to Pew Research Center’s report, “Asian Americans are the highest-income, best-educated and fastest-growing racial group in the United States.” The report also states that Asians have surpassed Hispanics as the largest wave of new (legal) immigrants to the United States, which makes them the fastest-growing racial group in the country. However, that does not give a complete picture.

While Asian American immigrants are considered well off today, immigrants from Asia were not always so lucky. True, many now come with increased opportunities, thanks to Asian immigrant’s ever expanding U.S. and worldwide economic presence. Many 19th Century Chinese
immigrants who come to America come with hopes of achieving the “American Dream”, just like the Jews, Italians and other later 19th century European immigrants. It was only, for the most part, their great grandchildren who were to “make it.” Alas, this has yet to occur for American Indians, despite the fact that they got here first. This could be a principle reason to carry out this research with AI/AN a companion study.

In the mindset of Americans, Asians are considered the model minority group which other immigrants groups should be patterned after. This ‘model minority” status, while earned, causes many problems. Asian American problems have to be pushed off to the sides in favor of other minority groups. Who knows? That may be one of the reasons life tables for American Asians were not considered a high enough priority to be produced.

The current problems that are affecting Asian Americans are whether there is an accurate coverage of their populations. Most Asian immigrants prefer to settle in urban areas, where there are more job opportunities. The coverage problem is common in rural areas where there are a fewer Asians living there and many, in the western U.S., whose families arrived in the 19th to work on the transcontinental railroad [Ambrose 2000].

Some Asians’ rename themselves in order to assimilate by taking American names (e.g., as by our NORC colleague Dr. Yanghe Yang, who goes professionally by the name Michael Yang [e.g., Yang 2012]. But on his passport Yanghu still appears, not Michael. This causes linkage issues if their formal and informal identities cannot be linked together. Often names are not legally changed.

A big problem that there is not ample insight on the translation of names listed on official documents. In Chinese there is no direct translation of a Chinese character-based name to an American one, written in the Latin alphabet. For example, the common last name “Zhang” could be written as “Chang” [Tian 2012]. The same can apply to the surname “Lee” (Le, Ly, Li, etc.). In China there are few last names relative to people, so first names are key identifiers. Thus a common Americanized name like “Daniel Lee” could be hard to link uniquely, especially when middle names, as is the case of most Chinese, are not common.

7. Record Linkage

Neither American Indian nor Asian life tables have been successfully completed in the United States for 2010. Because of cultural differences, as noted, the ability to create life tables for either group is difficult.

American Indians (even today), along with Asian Americans (albeit mainly historically) have been discriminated against and marginalized. Both have been forcibly and, sometimes frequently relocated from place to place. Witness the repeated “Trails of Tears” in American history and the encouraged and sometimes forced Bureau of Indian Affairs relocation policy of

---

28 The book: Unraveling The" Model Minority" Stereotype: Listening to Asian American Youth
29 Many European Americans also rename themselves using a nickname [Newcombe et al. 1992], so arguably the issue is a matter of degree not difference.
30 The existing vital records system may not be flexible enough to handle to emerging multicultural society which we are become. There are good reason as imply above to add a second name field.
American Indians to inner cities in the 50’s, 60’s, and 70’s. For the Japanese, the Internment camps still are in the living memory of many.

Typically, AI/AN and Asian Americans rename themselves culturally and not just personally (as by taking a nickname). This renaming has made it difficult to complete the linkage of birth certificates and death certificates of these populations, as infants become adults. For example, in many Tribes when an AI/AN has a ceremony in their culture, as a mark of adulthood, they are given an Indian name that stays with them for the rest of their life (and, it is believed, into their afterlife). Another example of this renaming is when Asians immigrate to the United States. Many take on an American name which is completely different from the name listed in their passport or visa. They may even use this “Americanized” name for a drivers’ license or other paperwork.

A reminder on how record linkage is typically done may be in order. For linkages of individuals, the key variables are surnames or in America last names, first names, dates of birth, gender, and potentially ethnicities (usually too inconsistently reported, to be useful), and other dates, including date of death, if applicable. Suppose that one component of the records associated with each of the two populations to be linked, say A and B is the surname. The comparison of surnames on two records will result in a component of the linkage comparison vector. This component may be a simple comparison component such as “name agrees” or “name disagrees” or “name missing on one or both records”; or it may be a more complicated vector component such as for example “records agree on Soundex code, the Soundex code is B650; the first 5 characters of the name agree; the second 5 characters of the name agree; the surname is BROWNING.”

At the Social Security Administration (SSA), incidentally, the SSNs assigned and identifying data, names, dates of birth and gender are recorded on what is called the NUMIDENT file. Now the NUMIDENT has a crude version of the birth certificate ethnicity. But it is too crude or incomplete to be used to classify whether a child is indigenous (American Indian) or not. It has, in fact, now being discontinued. That is why we need access to the State copy of the SSN stamped version of the birth certificate. This is an expensive step, if we require the data for all 20 years, as seems desirable.

In what follows we look graphically at the Fellegi-Sunter linkage space. In Figure 3 we start with a world in which the linkage variables are always present, unique and without errors of any kind (see figure 3 below). This was the system that was to suppose to have occurred when Social Security Numbers were introduced. Of course, it did not and, subsequently, fixes, adding part of the family name and, still other changes were made to “robustize” the system against

33 Just like I have used “Fritz,” since I was 17, even though my birth certificate says Frederick.,
34 In Chinese tradition, the surname is the first name.
35 The Social Security Administration for many years produced a file with social security number and date of death, with a full name. Social Security recently stopped this because of concern of identity fraud. This file is still useful for earlier years.
36 Browning example from Fellegi-Sunter opt. cit.
37 We do intend to use the historical NUMIDENT, probably at the Census Bureau, since the ethnicity shown there can help in the (multiple) imputation we will have to rely on outside the period where we will have birth certificates with SSNs attached
error and incompleteness. Anyway for our purposes should protect against the failure to link caused by cultural differences between the various subgroups in the US that may arise from renaming and subsequently lower linkage rates for the non-European groups.

Despite these problems the successful linkage of birth certificates and death certificates of infants by the States, facilitated by the National Center for Health Statistics, can be extended by using another birth certificate feature. With the assignment of Social Security Numbers (SSNs) at birth, completed in all States by 1993 the SSN was added to the birth certificate. So it may be available, at a fee for further linkage. We intend to employ this feature now that the Social Security Number is available on the birth certificate. Allowing for uncertainty in the linkage we can describe the matching as an overlay of two distributions (see figure 4 below). There are true links and true nonlinks but because of the overlap we have to add a third indeterminate category of potential links.

In this idealized, but more realistic world than that depicted in figure 3, two types of errors are made: cases where we treat a true link as a nonlink (This is done with probability $\alpha$.) Conversely, cases where we treat a true nonlink as a link (This is done with probability $\beta$.) The desired probabilities $\alpha$ and $\beta$ are set as low as the system and available resources will allow. Now, as we have said already, relative to the SSA/IRS systems, ad hoc fixes were needed because the original systems were not intended to be employed the way they are now. This is also true of the vital record systems. And that is why we need to use the social security numbers (SSNs) placed on the birth certificates. How to meld the pre-1990s data, where SSNs were on the certificates with those prior to 1993, where they are not, is an open question. At this point we will employ Multiple Imputation here. We believe that Multiple Imputation will afford us another benefit in that we will not have to rely on Fellegi-Sunter type linkage models. Our experience in previous NSF Grants is that, even if a successful award is made, not all the requested funds sought may be provided.

We have two visuals to introduce that make, again, our observations relative to linkage, especially for American Indians. These are figures 5 and 6. Figure 5 displays a moderate number

---

38 The idea of building in an error detection/correction code did not occur to the founders of the US Social security System in the US in time for use as has been done in the Nordic countries, for example.

39 If we were granted just, say, 60% of our request we would have to give up producing the Asian life tables, seeking funds for them elsewhere. Anyway, we do not intend to try to do everything by giving every subtask a “haircut.” The biggest single cost is acquiring and coding the State Birth certificates. Because we are working with NCHS, the Census Bureau and Indian Health Services we are expecting support from them; hence, not more than nominal additional expenses for their important contributions, especially in the calculation of trial life tables and in analyses by cause of death, but also in their application of multiple imputation to record linkage.
of indeterminate links, as might be true of a European American subpopulation, while figure 6 displays a setting where there are a many indeterminates, due to the weaker (multicultural) linkages that can be made with indigenous peoples, certainly in the US but perhaps everywhere in the world [Accattoli 1998].

Setting these two graphs, side by side, allows us to address still another issue. The \((\alpha, \beta)\) bounds are not culturally neutral. As we have tried to suggest, for the AI/AN and Asian subpopulations using common (European motivated) limits, as is done, say, at the US Census Bureau in its person linkages leads to differential coverages and different values of, say, \(\beta\) for each subpopulation. Here, again, a Multiple Imputation approach while not solving this problem can at least illuminate it and help us look at robustness issues. We have, in another setting, developed an adjustment for the differential coverage issue that we think, with a Multiple Imputation, content “correction” may work, although it has yet to be tried recently\(^{40}\).

\[\begin{align*}
\text{Figure 5.} & \quad \text{Linkage with Modest Uncertainty} \\
\text{Log Frequency} & \\
\text{True Mentions} & \quad \text{Indeterminate Potential Link} & \quad \text{True Units} \\
\text{Link Probability} & \\
\end{align*}\]

\[\begin{align*}
\text{Figure 6.} & \quad \text{Multicultural Linkage} \\
\text{Log Frequency} & \\
\text{True Mentions} & \quad \text{Indeterminate Potential Link} & \quad \text{True Links} \\
\text{Link Probability} & \\
\end{align*}\]

8. Analysis Plan

Looking at the differences between figures 5 and 6 led us to employ multiple imputation to address the wider uncertainty caused largely by the cultural differences between European Americans, AI/AN or Asians. Our analysis goal is similar to our colleagues in Australia\(^{41}\), Canada\(^{42}\), New Zealand\(^{43}\) and elsewhere in the U.S.\(^{44}\).

\(^{40}\) But in the Scheuren, et al 1980 work we were successful

\(^{41}\) Linked Perinatal, Birth and Death Project (Australia), The current linkage method for IM, child mortality: Data on births are not routinely linked with perinatal, infant and child deaths data at a national level in Australia. Data linkage will be carried out by the Data Linkage Unit of the Australian Institute of Health and Welfare. Linkage will be mostly probabilistic because of the absence of unique identifiers on the various data sets to be linked. Probabilistic linkage will require the following variables as both blocking and linking variables: first and last name (middle name where this is available), sex, full date of birth, full residential address, name and address of facility where the birth occurred, dates of episodes of care (hospitalisations) for children who were hospitalized, date of death (for children who died)

\(^{42}\) Perinatal Outcomes Study: Live Birth, Infant Death and Stillbirth Records Linked to 20% Sample Data from the 1996 and 2006 Censuses of Population (Canada), The purpose is to assess perinatal outcomes in Canada according to risk factors related to socioeconomic, ethno-cultural and environmental circumstances. It will examine preterm and small/large-for-gestational-age births, as well as birth weight-specific and gestational age-specific fetal and infant mortalities.


\(^{44}\) United States Linked Birth-Infant Death Data Set (Past Work), The linked birth and infant death data set is a valuable tool for monitoring and exploring the complex inter-relationships between infant death and risk factors present at birth. The purpose of the linkage is to use the many additional variables available from the birth certificate to conduct detailed analyses of infant mortality patterns. The linked files include information from the birth certificate such as age, race, and Hispanic origin of the parents, birth weight, period of gestation, plurality, prenatal care usage, maternal education, live birth order, marital status, and
Even in the days, now largely long gone, when manual researches were undertaken to determine the link/nonlink status of members of the indeterminate (middle) linkage group, a large middle group was thought to be a weakness. It was obviously expensive, time consuming, and as it turned out different determinations were made by different researcher clerks (scheuren et al. 1980). Thus, we were adding to the (typically) unmeasured degree of uncertainty and, typically, without a way to measure that added uncertainty.

The structural response by “Linkers” was to augment the linkage software by computerizing more of the determinations. This had the effect of shifting costs to an earlier stage of the matching/linkage where variables were computerized. While there may have been cost savings at the same quality of the linkage, no scientific studies on this point that we know of exist, so we are silent here, except to speculate that the effect was adverse in that the linking work was further removed from the analysts control and, still, no quantification was done of the now hidden uncertainty.

- Multiple Imputation to measure linkage uncertainties. Truth to tell, there are many options available to handle linkage uncertainties. We also use multiple imputation, mainly, because we are familiar with it and have software to use it already developed.
- Multiple imputation [Rubin 1987] replaces each missing value with a set of plausible values. The distribution of these “plausibles” gives us a way to represent the uncertainty about the right value to impute.
- Each completed multiply imputed data set can be analyzed using standard procedures for complete data, and the results across these analyses combined, so that all the uncertainty components in the analysis—model uncertainty and missing value uncertainty are accounted for in the analysis.
- Our most recently use of multiple imputation was when we were testifying on an Indian class court suit action, now before the Supreme Court that began in 1996. This was a lawsuit filed in the U.S. District Court for the District of Columbia seeking to compel a historical accounting of Individual Indian Money (IIM) accounts. Below we can sketch the mathematics underlying our approach, borrowed from that case.

**Some details of Approach.** A good robust multivariate model for use with multiple imputation is the multivariate normal model with a noninformative prior [Schafer (1997)]. The complete-data posteriors, which are used to generate imputations, are

\[
\begin{align*}
\Sigma | Y & \sim W^{-1}(n - 1, (n - 1) S), \\
\mu|\Sigma, Y & \sim N(\bar{Y}, \frac{1}{n} \Sigma), \\
Y_i'|\mu, \Sigma & \sim N(\mu, \Sigma) \quad \forall i = 1, \ldots, n,
\end{align*}
\]

maternal smoking, linked to information from the death certificate such as age at death and underlying and multiple cause of death.

where \( W^{-1} \) denotes the inverse Wishart distribution, \( Y \) is the completed data matrix (which is composed of the observed values, \( Y_{obs} \), and the filled-in missing values, \( Y_{mis} \)), \( Y_i \) is a row in the data matrix, \( n \) is the number of years in the data matrix to be completed, \( S \) is the sample covariance matrix, \( y \) is the sample mean vector, and \( N(\cdot) \) denotes the multivariate normal distribution.

- **Software.** The imputation of the missing values can be completed using the SAS MI procedure, An MCMC algorithm that would generate observations from the posterior distribution. An example of a SAS program that implements this can be found at [http://www.norc.org/iim](http://www.norc.org/iim). The imputation of each missing value in our earlier work was done 10,000 times, more than we may need here. The multiple imputation literature, written over 30 years ago during an era of expensive computing, generally suggests that 3-to-5 imputations would be sufficient for assessing the contribution to an estimated value’s uncertainty due to missing information. The theory behind this relies on the use of a multivariate normal distribution, which is also the basis for our imputation modeling. Since we live in an era of less expensive computing, we would choose to use a larger number of imputations, although not necessarily 10,000.

Bottom line, our computing power will be sufficient for many imputations.

- For infants (0-1), the existing record linkages are pretty complete deaths are matched back to the birth certificates. We believe but will check that the amount of uncertainty in the linkage is small, perhaps under 5%. That’s about as good as it gets however.
- For infants/children born in 1993 or later there are SSNs available on the birth certificate.
- For young children (2-10) there has been demographic analysis done by Census, but there is not correct data on migration, it is hard to keep track of them. No administrative records exist for this population, since the Census Bureau or NCHS do not collect data for this age group. One suggestion to track them is to look at the movement of the parents, particularly the mother. Since there is a section for parent’s social security number in the application of social security number, that helps identify mobility of the child. Another option to track young children’s mobility is to look at school records, but that is impossible to get hold of the State’s data, alas that is a “bridge too far”. A common theme of infants, young children, and older children is the overwhelming theme of children being placed into the various state child protective services. These children may be adopted into non-Indian families and have a change in name after the adoption. If the tribe wants to assert jurisdiction over such cases they can along with other children.\(^{46}\)
- For older children (11-17) and young adults (18-30), we have available since 1993, the birth certificate with Social Security Numbers, so that covers the age group. However problems arises, if the population is like that of AIAN and under a “rebranding”. When the child grows up they change their name, but they still have their social security number. That is ok, if they still stay within the same geographic area. But we must be careful because the linkage might get thrown out because it is considered a nonlink. So it is suggested that if the records contain a different name but same social security number, we look other traits that match (e.g., gender, date and place of birth, mother’s name and other parental details.). About right so we can say it is a potential link. There is no good data on young adults. We have some identification on the NUMIDENT file. We can get an idea from that as to how many people are identified as indigenous. What we want to do is compare the information on the NUMIDENT file from

Census (which is not good) to the ethnicities listed on the birth certificate (once permission is granted). We would have to use multiple imputation for the later years after 1993 for ethnicity.

For adults (31-64), social security numbers are available for the NUMIDENT for everyone. We will try to match them. The uncertainty will be greater, but we will impute for that. Also, there is a need to find the Social Security Numbers of the individuals with the name changes. That process can be easier if they are working adult in the system (e.g., W-2 forms).

[The IRS produces an estimate of an “underground economy”]. To have a consistent form of data for American Indians we can use the data from the Indian Health Service, even though coverage limited. Even if a person changes their name in order to receive services from Indian Health they must fill out paper work and prove enrolment in a federal recognized Tribe. This can sometimes solve the problem of rebranding issues if the person keeps receiving services from Indian Health.

For retired adults (65+) they have considerable incentives to be linked because of Social Security Earnings. This different from Social Security Income the only requirement is age. Social Security Earnings requires contributions from work to qualify. The quality of the records and results vary enormously depending on the age of the individual. The younger individuals usually have better records, thanks to Social Security Numbers on the birth certificate. We want to make the birth records electronic from 1940 to now. Using Census to do so is a possibility, but the results are not as good.

9. Partnering with Tribal Nations

In order to accomplish this grant an integral part will be to partner with Tribal Nations. Underreporting of AI/AN populations and vital statistics can be attributed to constant mistrust of the Federal Government to Tribal Nations. This mistrust can be an uphill battle to overcome. To overcome this battle will be partner with Tribal Leaders to bring the idea of, the creation of life tables to them. These meeting with be conducted on a government-to-government consultation following Presidential Executive Order 13175 that states the unique legal and political relationship with Indian Tribal governments, established through and confirmed by the Constitution of the United States, treaties, statutes, executive orders, and judicial decisions.

Once a relationship has been established we will work with Tribal Leaders to determine the community involved with the project, stakeholders (people with a direct or indirect interest in the research partnership or otherwise affected), capacity of community (ability to identify, mobilize, and address social and public health problems or additional resources). After stakeholders are identified within the community we will work with Tribal Leaders to set up a Tribal Internal Review Board (IRB) if not all ready established. The reason for establishing a Tribal IRB is to ensure the community, Tribal Leaders and other stakeholders the protection of data in research and have the ability to review projects being conducted. We will use examples such as the California Rural Indian Health Board, Chickasaw Nation Research Review Committee, Oglala Sioux Tribal Research Review Board, and Navajo Nation Human Research Review Board.

A second partner we will utilize in Indian Country will be the Indian Health Services (IHS). Indian Health Services, an agency within the Department of Health and Human Services, is responsible for providing federal health services to American Indians and Alaska Native peoples. Indian Health Services are located on or near reservations. Indian Health Services also partners with The Urban Indian Health Programs (UIHP) a non-profit program also providing health services to American Indians and Alaskan Natives.
Indian Health Services can be a partner in sharing of information. Because IHS provides services exclusively to members of federally recognized tribes they are a resource for data on AI/AN. Unfortunately, a gap emerges with AI/AN for those not registered with IHS.

States proposed for this study are Montana and Washington at first. At a later time Hawaii will be included to create life tables for AI/AN and Native Hawaiians. Twenty-two facilities are located in Montana on reservations or near various reservations. Washington has thirty-six facilities on or near reservations. The facilities are of IHS direct service, Urban Indian Health Programs, and tribal consortiums.

Studies have indicated that death certificates tend to be underreported for the AI/AN population. [Bertolli, Espey, Hahn]. Tribal membership can be attributed to underreporting for AI/AN. American Indian Tribes have the ability to set their own membership status. Historically American Indians used kinship through various systems of lineal descent. This system varied from tribe to tribe. After European contact a system of blood quantum was created by the federal government. This system was set up to assimilate Indian people. Tribal Governments maintain records for enrolled members and descendants. Administrative (i.e. enrollment) Tribal records could do abundant to alleviate misclassification for births and deaths.

Tribal records are administered by individual Tribal Governments. Tribal Governments do not release Tribal enrollment records to outside agencies. Manual matches may have to be done in some instances. We want to highlight the fact that results from life tables have the potential to help Tribes understand the impact of health conditions and impairments resulting in death on Tribal citizens now and in the future. The life tables for Asian Americans are expected also to be actionable. The contrast between the life tables between Asian Americans and American Indians may turn out to be most useful. While the final report will be public, in addition, we want to provide Tribes with the knowledge needed to identify relevant statistics, as well as provide Tribes with detailed reports of our findings.

10. Dissemination and Publication of Results

The scientific results will be presented in three ways:

■ First, the linkage application of multiple imputation will be described and elaborated on in a second edition of my Springer book with Tom Herzog and Bill Winkler on *Data Quality and Record Linkage Techniques*. The new edition is scheduled to be published next year (2013).

■ Second, when the Study is all done (2015) we will host another Record Linkage Conference like those I organized in the 80’s and 90’s.

■ Third, the life table results will be disseminated in refereed Public Health Journals and NCHS publications. Every effort will be made to expand the pilot work to cover more States and Tribes, so results may be forthcoming later.

The contrast between AI/AN and Asian American mortality experience and its causes is also expected to lead to new science outside the fields of statistics and demography.

---

Supporting Bibliography for NSF Grant

AIAN Identity


Asian American Identity and Issues


Health Issues


NATSIHC (National Aboriginal and Torres Strait Islander Health Council) 2003. National Strategic Framework for Aboriginal and Torres Strait Islander Health: framework for action by governments. Canberra: NATSIHC.

National Cancer Institute, Surveillance Epidemiology and End Results, County attributes. 2008. www.seer.cancer.gov/seerstat/variables/countyattrs/.

National Cancer Institute, Surveillance Epidemiology and End Results. Accessing datasets and tools: SEER behavior recode for analysis. seer. cancer.gov/behavrecode/index.html.


Seattle Indian Health Board-Urban Indian Health Institute. The Health Status of Urban American Indians and Alaska Natives: An Analysis of Select Vital Records and Census Data Sources, 2004. Available at URL:


USDHHS, Indian Health Service. Indian Health Service Fact Sheet. 2006. www.ihs.gov/PublicInfo/PublicAffairs/Welcome_Info/ThisFacts.asp.


**Misclassification**


Indian Health Service. IHS Fact Sheets – IHS Year 2011 Profile. Available at: http://www.ihs.gov/PublicAffairs/ IHSBrochure/Profile2010.asp.

Indian Health Service, U.S. Department of Health and Human Services, Adjusting for Miscoding of Indian Race on State Death Certificates. Rockville, MD: 1996


Record Linkage


Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: a comparison of Link Plus, the Link King, and a ‘basic’ deterministic algorithm. Health Informatics J 2008;14:5–15.


**Life Tables**


International Work


