Published in final edited form as:
Ethn Health. 2010 December ; 15(6): 639–647. doi:10.1080/13557858.2010.505979.

Validation of an Arab names algorithm in the determination of Arab ancestry for use in health research

Abdulrahman M. El-Sayed[1,2,3], **Diane S. Lauderdale**[4], and **Sandro Galea**[5]

[1]Center for Social Epidemiology and Population Health, University of Michigan, Ann Arbor, USA

[2]University of Michigan Medical School, Ann Arbor, USA

[3]Department of Public Health, University of Oxford, Oxford, UK

[4]Department of Health Studies, University of Chicago, Chicago, USA

[5]Department of Epidemiology, Columbia University, New York, USA

Abstract

Objective—Data about Arab-Americans, a growing ethnic minority, is not routinely collected in vital statistics, registry, or administrative data in the US. The difficulty in identifying Arab-Americans using publicly available data sources is a barrier to health research about this group. Here, we validate an empirically-based, probabilistic Arab name algorithm (ANA) for identifying Arab-Americans in health research.

Design—We used data from all Michigan birth certificates between 2000-2005. Fathers' surnames and mothers' maiden names were coded as Arab or non-Arab according to the ANA. We calculated sensitivity, specificity, and positive (PPV) and negative predictive values (NPV) of Arab ethnicity inferred using the ANA as compared to self-reported Arab ancestry.

Results—State-wide, the ANA had a specificity of 98.9%, a sensitivity of 50.3%, a PPV of 57.0%, and a NPV of 98.6%. Both the false positive and false negative rates were higher among men than among women. As the concentration of Arab-Americans in a study locality increased, the ANA false positive rate increased and false-negative rate decreased.

Conclusion—The ANA is highly specific but only moderately sensitive as a means of detecting Arab ancestry. Future research should compare health characteristics among Arab-American populations defined by Arab ancestry and those defined by the ANA.

Keywords

Name Algorithm; Arab-American; health; identification

Introduction

Arab-Americans are United States residents who trace their ancestry, cultural or linguistic heritage, or identity back to one of 22 Arab countries (El-Sayed and Galea 2009). Approximately 3.5 million Arab-Americans currently live in the US, and it is a growing minority group (de la Cruz and Brittingham 2003). Despite substantial focus on this group in recent years, in part due to the September 11, 2001 attacks and the subsequent wars in

Corresponding Author Abdulrahman M. El-Sayed Department of Epidemiology 1415 Washington Heights Ann Arbor, MI 48109 elabdul@umich.edu Phone: 248.930.0522 Fax: 734.763.5706 .

Afghanistan and Iraq—there is very limited research about this group's health and well-being.

A recent review about the health of Arab-Americans in the United States highlighted several reasons for continued research about this ethnic group (El-Sayed and Galea, 2009). First, substantial disparities in health metrics, such as cancer incidence, risk factors for cardiovascular disease, and smoking behavior, between Arab-Americans and the general population found in several of the reviewed studies suggest that the health of Arab-Americans may differ in meaningful ways from the general population, and is often worse (Schwartz *et al*. 2004, Grekin and Ayna 2008, Kridli *et al*. 2006, Hatahet *et al*. 2002, Jamil *et al*. 2008, Jaber *et al*. 2003, Jaber *et al*. 2004). Second, findings showed that exposures such as discrimination-associated stress and acculturation which are specific to the ethnic minority experience in the US, and among Arab-Americans in particular, were important contributors to adverse health and wellbeing among this ethnic group (El-Sayed and Galea, 2009). Two studies, for example, demonstrated links between discrimination-associated stress, and higher risk for metrics of poor health, including preterm birth, low birth weight, psychological distress and low self-esteem (Lauderdale 2006, Ajrouch 2007). In other studies, acculturation was demonstrated to be an important correlate of diabetes, asthma, low birth weight, and self-rated health and activity limitations among Arab-Americans (Ammer and Hovey 2007, Jaber *et al*. 2003, Read *et al*. 2005, Johnson *et al*. 2005, El-Sayed and Galea 2009). Third, despite evidence demonstrating significant health disparities between Arab-Americans and majority groups in the US (Schwartz *et al*. 2004, Grekin and Ayna 2008, Kridli *et al*. 2006, Hatahet *et al*. 2002, Jamil *et al*. 2008, Jaber *et al*. 2003, Jaber *et al*. 2004), there is little consensus regarding the burden of important diseases, such as diabetes, cancer, and/or heart disease among Arab-Americans (El-Sayed and Galea 2009).

A primary methodological obstacle to quality health research about Arab-Americans remains the effective identification and differentiation of members of this group from the general population (El-Sayed and Galea 2009). In the US, Arab-Americans are considered "white" according to the United States Office of Management and the Budget. This federal body, which collects federal statistics and oversees public policy, defined the race category "white" as "persons originating in Europe, the Middle East, and North Africa (Office of Management and the Budget 1978, Executive Office of the President 1997). Therefore, self-identified Arab ethnicity is not collected in federal vital registry data in the US. However, the state of Michigan, the state with the largest per capita Arab-American population in the country (Arab American Institute Foundation 2003), is the only state of which we are aware that does collect self-reported Arab ethnicity in state vital registry statistics.

One possible solution is the use of an Arab name algorithm (ANA) that can identify Arab ethnicity from databases where last names are available. A similar algorithm has been used to identify Hispanics (Passel and Word 1980, Word and Perkins 1996). One such algorithm has been developed by Morrison and colleagues to identify Arab-Americans (Morrison *et al*. 2003). This algorithm, and a similar one, have been used to identify Arab-Americans for health research purposes in several studies (Schwartz *et al*. 2004, Lauderdale 2006, El-Sayed *et al*. 2008). A recent report from the Institute of Medicine of the US National Academies of Science highlighted the importance of name algorithms for the indirect identification of ethnic minorities where self-report data is not routinely collected- it also stressed the importance of qualifying the accuracy of such tools (Ulmer *et al*., 2009). In this study, we were interested in assessing the performance of Morrison's ANA using birth registry data between 2000-2005 from the state of Michigan. We also wanted to test the ANA across gender, across a specific geographic stratum known to have a high concentration of Arab-Americans, and across birthplace, in order to assess the effects of high and low Arab-American concentration on the utility of the ANA.

Methods

The methods used by Morrison and colleagues to create the ANA drew from those used in the creation of algorithms for identifying Asian and Hispanic names (Morrison *et al*. 2003, Lauderdale and Kestenbaum 2002, Elo *et al*. 2004). Detailed steps regarding the construction of the version of the ANA validated here are discussed by Morrison and colleagues (Morrison *et al*. 2003) and Lauderdale (Lauderdale 2006).

We worked with the Michigan Department of Community Health (MDCH), which created a special file for us that lacked individual identifiers but included special ethnic identifiers created for this project. The MDCH coded the maiden names of all mothers and the surnames of all fathers of children born in Michigan between 2000-2005 as Arab or non-Arab according to the ANA. Then, mothers and fathers were coded as Arab or non-Arab according to self-reported ancestry, which MDCH forms specifically query. Finally, using self-reported ancestry as the "gold standard," we calculated sensitivity (the proportion of those who reported Arab ancestry who were correctly identified as Arab by the ANA), specificity (the proportion of those who did not report Arab ancestry who were correctly labeled as non-Arab by the ANA), and positive (PPV- the proportion of those who were labeled as Arab by the ANA who reported Arab ancestry) and negative predictive values (NPV- the proportion of those who were labeled as non-Arab by the ANA who did not report Arab ancestry) for the ANA across the entire study population.

We ran one statewide analysis followed by three stratified analyses. The first stratification was by gender with analyses run among men and women separately. The second stratification was by community context. Because 82% of Michigan's Arab-American population live in one of three contiguous counties: Oakland, Macomb, or Wayne counties (Arab American Institute Foundation 2003), parents were coded for residence within or outside of one of these three counties (Arab-American dense counties) using zip codes, and analyzed separately. The third stratification was by maternal nativity status (US-born versus foreign-born).

Results

Table 1 shows two-by-two cross-tabulations of Arab ancestry and Arab name among all parents in Michigan between 2000-2005, as well as among each strata.

Table 2 shows sensitivities, specificities, PPVs, NPVs and percentage of parents reporting Arab ancestry in each stratum. Statewide, 2.8% of all parents reported Arab ancestry. In the statewide analysis, sensitivity of the ANA in determining self-reported Arab ancestry was 50.3%, specificity was 98.9%, PPV was 57.0% and NPV was 98.6%.

Our first stratification was by gender. The proportion of both men and women reporting Arab ancestry was 2.8%. Among men, sensitivity was 36.1%, specificity was 98.4%, PPV was 39.7% and NPV was 98.2%. All metrics were higher among women; sensitivity was 64.3%, specificity was 99.4%, PPV was 75.2% and NPV was 99.0%.

Our second stratification was by residence in three Arab-American dense counties versus all other counties. The percentage of parents reporting Arab ancestry in the three Arab-American dense counties was 5.5% compared to 1.0% outside of these counties. We found that specificity and NPV were lower in Arab-American dense counties as compared to outside of these counties, while sensitivity and PPV were higher in Arab-American dense counties as compared to outside of these counties.

Our third stratification was by maternal nativity. Among the foreign-born, 17.8% reported Arab ancestry, compared to 1.2% in the native-born group. We found that among the foreign-born group, specificity and NPV were lower and sensitivity and PPV were higher as compared to among the native-born group.

Discussion

Using birth registry data from the state of Michigan, we found that the ANA is highly specific but only moderately sensitive as a means of detecting Arab ancestry when compared to self-reported ancestry, which, to our knowledge, the state of Michigan uniquely includes in vital statistics forms. As the concentration of Arab-Americans in a study locality increased, the sensitivity and the PPV of the ANA increased and the specificity and the NPV of the algorithm decreased. The NPV was high: 93.5% or higher in all stratifications regardless of the prevalence of Arab Ancestry.

To the best of our knowledge, this is the first attempt to validate an Arab surname algorithm for use in health research. A Middle East Surname List (MESL) that includes surnames from Arab countries, as well as Middle Eastern countries where Arab ethnicity is uncommon, has been developed and validated (Nasseri 2007). The MESL was derived using name data from the Social Security Administration, the California Cancer Registry, and expert opinion, using country of birth as the standard for determining Middle Eastern status; it was reported to be relatively accurate in determining birth in a country of the Middle East (Nasseri 2007). The sensitivity of the MESL in determining Middle Eastern birthplace among patients in the California Cancer Registry was 88.62%; specificity was 99.46%; PPV was 68.54%; and NPV was 99.85%. The MESL methodology differs from our ANA in three fundamental ways. First the MESL includes names from “Middle Eastern” countries, including Iran, Afghanistan, Pakistan, and Armenia. Second, multiple sources were used in the compilation of the final MESL. Third, the MESL was tested using Middle Eastern country of birth as the “gold standard,” rather than self-reported ancestry, which systematically omits native-born Middle-Easterners (Stronks *et al*. 2008).

We found substantially higher false positive and false negative rates associated the ANA among men compared to women. Using the ANA, more Arab-American men were likely to be identified as non-Arab-American then Arab-American women, and more non-Arab-American men were likely to be identified as Arab-American then non-Arab-American women. Our findings may reflect systematic differences in ancestral self-reporting between men and women.

In evaluating screening tests in clinical situations, one expects the sensitivity and specificity of a test to be constant in all populations, and the PPV and NPV to vary with prevalence of the condition, with PPV higher in groups with higher prevalence of the condition. While we found the PPV to be higher in strata with higher proportions of Arab-Americans, we found that the sensitivity of the ANA was much higher in areas with greater Arab-American concentration and among women, although not as high as other surname algorithms (Nasseri 2007). This points to differences in the Arab-American population between high and low density areas. Arab-Americans who live in high Arab-American concentration localities, or enclaves, are likely to have lower socioeconomic status (SES) than Arab-American in lower Arab-American concentration localities and are also more likely to be Muslim (Logan *et al*. 2002, Abudabbeh 1996, Abu-Laban and Suleiman 1989, El-Badry 1994, Naff 1985, Naff 1985, Amer and Hovey 2007). We hypothesize that our finding may be partially explained by the ratio of Christian Arab-Americans to Muslim Arab-Americans in areas with high Arab-American concentrations compared to areas with low Arab-American concentrations. Because biblical names are common among Christian Arabs, they are less likely to have

names that are ethnically distinctive and included in the ANA. Therefore, some of the reason for the false-negative rate of the ANA may occur because Christian Arab-Americans with non-ethnically distinctive names are not recognizably Arab-American by surname. Another possibility is that there may be relatively more persons with partial rather than full Arab ancestry outside the ethnic enclaves and although they endorse the Arab ancestry question, such ancestry may be confined to the maternal side and not reflected in their surnames. For reasons of confidentiality the actual surnames on the birth registry records were not available for our use. We were therefore unable to test our hypotheses about Christian Arab names and partial ancestry.

Specificity was lower in the higher Arab-American concentration areas, meaning a higher proportion of non-Arab-Americans were incorrectly identified as Arab-American. As multi-ethnic congregations are common among Muslim communities in the US (Haniff 2003), and Arab-Americans in the US are disproportionately Muslim (Arab American Institute Foundation 2008), there may be higher proportions of Muslims of other ethnicities in contexts with high Arab-American density. One source of false positives observed may be that some names in the ANA belong to Muslims with other ancestries, such as South Asian or African-American converts to Islam. Although the derivation of the ANA list at the Social Security Administration was designed to identify names that were distinctive for Arab countries, some names are included that are much more common in Arab countries but nonetheless appear in other populations with a high proportion of Muslims. These names would identify some persons with other ancestries.

When interpreting the results of this study it is important to recognize that our analysis includes only parents, who range in age from older adolescents to middle-aged adults, which may not be representative of the entire Arab-American population. Using data about parents may over-represent first generation Arab-Americans, who may be more or less likely to have ethnically distinctive names than the general population. Another limitation to consider is that the standard of measurement to which the ANA was compared is self-identified Arab ancestry, rather than ethnicity. Because the concept of ancestry may be interpreted differently than ethnicity, our findings may not generalize to the algorithm's ability to determine ethnicity. There has not, to our knowledge, been any validation of the self-reported Arab ancestry question. Finally, it is important to consider the purpose of a names algorithm in health research about ethnic minority populations. The most important use of the ANA may not be its ability to detect the largest proportion of Arab-Americans, but in its ability to identify a sample of that population that is very likely to be Arab-American and that does not differ systematically from the general Arab-American population. For example, Shin and Yu suggested that because 22% of the Korean population shared the surname "Kim", metrics among only those with the Kim surname were generalizable to the Korean-American population (Shin and Yu 1984). If the ANA is capable of defining a representative sample of Arab-Americans, it may be well-equipped for use in health research despite only moderate sensitivity. The NPV of the ANA, which was found to be uniformly high, may therefore be more important than its sensitivity for use in assessing health metrics among the Arab-American population.

The ANA is a highly specific, but only moderately sensitive potential tool for investigators interested in identifying Arab-Americans for the purposes of health research. The performance of the ANA in identifying Arab ancestry is dependant upon the proportion of Arab-Americans in the study population; the ANA was most sensitive in study populations with high proportions of Arab-Americans but actually had slightly higher specificity in areas with lower Arab-American concentrations. Because of its high specificity, the ANA may be useful as a means of assessing health metrics among the Arab-American population;

however, research is needed to compare the actual health characteristics among Arab-American populations defined by Arab ancestry and those defined by the ANA.

Key Messages

The difficulty in identifying Arab-Americans using vital statistics, registry, or administrative data is a barrier to health research about this group. The Arab Names Algorithm is a highly specific, but only moderately sensitive means of detecting Arab ancestry for use in health research about Arab-Americans. As the concentration of Arab-Americans in a study locality increased, the ANA false positive rate increased and false-negative rate decreased. Future research should compare health characteristics among Arab-American populations defined by Arab ancestry and those defined by the ANA.

Acknowledgments

The authors thank Glenn Copeland and Glenn Radford from the Michigan Department of Community Health for their help acquiring the data. Funded in part by the Rhodes Trust and NIH Grants GM07863, DA022720 and DA017642.

References

Abudabbeh, N. Arab families. In: McGoldrick, M.; Giordano, J.; Pearce, JK., editors. Ethnicity and family therapy. 2nd edition. NY: Guilford Press; New York: 1996. p. 333-46.

Abu-Laban, B.; Suleiman, MW. Arab-Americans, continuity and change. Association of Arab-American University Graduates; Belmont, MA: 1989.

Ajrouch KJ. Resources and well-being among Arab-American elders. Journal of Cross Cultural Gerontology 2007;22(2):167–182. [PubMed: 17226096]

Amer MM, Hovey JD. Socio-demographic differences in acculturation and mental health for a sample of 2nd generation/early immigrant Arab Americans. Journal of Immigrant and Minority Health 2007;9(4):335–347. [PubMed: 17340173]

Arab American Institute Foundation. State profile: Michigan. 2003. Available at << http://www.google.co.uk/url?sa=t&source=web&ct=res&cd=1&ved=0CAcQFjAA& url=http%3A%2F%2Fwww.aaiusa.org%2Fpage%2Ffile%2Ff6bf1bfae54f0224af_ 3dtmvyj4h.pdf%2FMIdemographics.pdf&ei=AEoES-aPBsLP-QbftditCA&usg=AFQjCNGwRDd8kRsO2KOX8WXn7RovUZAMow&sig2=0itwmq kaWNHwb6e4XDW91A >>

Arab American Institute Foundation. Arab Americans; demographics. 2008. Available at <<http://www.aaiusa.org/arab-americans/22/demographics>>

de la Cruz, G.; Brittingham, PA. The Arab population: 2000. U.S. Census Bureau; 2003. Available at << www.census.gov/prod/2003pubs/c2kbr-23.pdf>>

El-Badry S. The Arab-American market. American Demography 1994;16:22–30.

Elo IT, et al. Mortality among elderly Hispanics in the United States: Past evidence and new results. Demography 2004;41(1):109–128. [PubMed: 15074127]

El-Sayed AM, Hadley C, Galea S. Birth outcomes among Arab-American before and after the terrorist attacks of September 11, 2001. Ethnicity & Disease 2008;18(3):348–356. [PubMed: 18785451]

El-Sayed AM, Galea S. Community context, acculturation, and low birth weight risk among Arab-Americans: Evidence from the Arab-American Birth Outcomes Study. Journal of Epidemiology and Community Health. 2009 published Online First.

El-Sayed AM, Galea S. The health of Arab-Americans living in the United States: a systematic review of the literature. BMC Public Health 2009;9:272. [PubMed: 19643005]

Executive Office of the President. Office of Management and Budget. Office of Information and Regulatory Affairs. Washington D.C.: Office of Federal Statistical Policy and Standards, U.S.

Department of Commerce; 1997. Revisions to the standards for the classification of federal data on race and ethnicity.

Grekin ER, Ayna D. Argileh use among college students in the United States: An emerging trend. Journal of Studies on Alcohol and Drugs 2008;69(3):472–475. [PubMed: 18432392]

Haniff GM. The muslim community in America: A brief profile. Journal of Muslim Minority Affairs 2003;23(2):303–311.

Hatahet W, Khosla P, Fungwe TV. Prevalence of risk factors to coronary heart disease in an Arab-American population in southeast Michigan. International Journal of Food Sciences and Nutrtion 2002;53(4):325–335.

Jaber, et al. Epidemiology of diabetes among Arab Americans. Diabetes Care 2003;26(2):308–313. [PubMed: 12547854]

Jaber LA, et al. The prevalence of the metabolic syndrome among Arab Americans. Diabetes Care 2004;27(1):234–238. [PubMed: 14693995]

Jamil H, et al. Self-reported heart disease among Arab and Chaldean American women residing in southeast Michigan. Ethnicity & Disease 2008;14(1):19–25. [PubMed: 18447094]

Johnson M, et al. Asthma prevalence and severity in Arab American communities in the Detroit area, Michigan. Journal of Immigrant and Minority Health 2005;7(3):165–178.

Kridli SA, Herman WH, Brown MB. The epidemiology of diabetes and its risk factors among Chaldean Americans. Ethnicity & Disease 2006;14(1):351–365. [PubMed: 17682235]

Lauderdale DS. Birth outcomes for Arabic-named women in California before and after September 11. Demography 2006;43(1):185–201. [PubMed: 16579214]

Lauderdale DS, Kestenbaum B. Mortality rates of elderly Asian American populations based on medicare and social security data. Demography 2002;39(3):529–540. [PubMed: 12205756]

Logan JR, Alba RD, Zhang W. Immigrant enclaves and ethnic enclaves in New York and Los Angeles. American Sociological Review 2002;67(2):299–322.

Morrison, PA., et al. Developing an Arab American surname list: Potential demographic and health research applications. Presented at the Annual Meeting of the Southern Demographic Association; Alexandria, Va.. October; 2003. p. 23-25.

Naff, A. The early Arab immigrant experience. Southern Illinois University Press; Carbondale, IL: 1985.

Naff, A. The early Arab immigrant experience. In: McCarus, E., editor. The development of Arab-American identity. University of Michigan Press; Ann Arbor, MI: 1985.

Nasseri K. Construction and validation of a list of common Middle Eastern surnames for epidemiological research. Cancer Detection and Prevention 2007;31(5):424–429. [PubMed: 18023539]

Office of Management and Budget. Statistical Policy Handbook. Washington D.C.: Office of Federal Statistical Policy and Standards, U.S. Department of Commerce; 1978. Directive no. 15: Race and ethnic standards for federal statistics and administrative reporting.

Passel, JS.; Word, DL. Constructing the list of Spanish surnames for the 1980 census: An application of Bayes' theorem. 1980. Presented at the Population Association of America Annual Meeting; Denver, CL. Apr; 1980. p. 10-12.

Read JG, Amick B, Donato KM. Arab immigrants: A new case for ethnicity and health? Social Science & Medicine 2005;61(1):77–82. [PubMed: 15847963]

Schwartz KL, et al. Cancer among Arab Americans in the metropolitan Detroit area. Ethnicity & Disease 2004;14(1):141–146. [PubMed: 15002934]

Shin EH, Yu EY. Use of surnames in ethnic research: The case of Kims in the Korean-American population. Demography 1984;21(3):347–360. [PubMed: 6479394]

Stronks K, Kulu-Glasgow I, Agyemang C. The utility of 'country of birth' for the classification of ethnic groups in health research: the Dutch experience. Ethnicity & Health 2008;14:3. First published on: 03 December 2008 (iFirst).

Ulmer, C., et al. Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. National Academies Press; Washington, D.C.: 2009.

Word, DL.; Perkins, RC. Building a Spanish surname list for the 1990's- A new approach to an old problem. Population Division, U.S. Bureau of the Census; Washington D.C.: 1996. Technical Working Paper No. 13

Table 1

Two-by-two cross-tabulations of Arab ancestry and Arab name statewide and in other strata among all parents of newborns in Michigan between 2000-2005

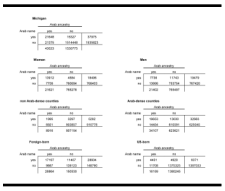

Table 2

Sensitivity, specificity, positive predictive value, negative predictive value, and Arab ancestry analyses statewide and in other strata among all parents of newborns in Michigan between 2000-2005

	Sensitivity	Specificity	PPV	NPV	Arab Ancestry	N
Michigan	50.3%	98.9%	57.0%	98.6%	2.8%	1,573,798
Men	36.1%	98.4%	39.7%	98.2%	2.8%	786,899
Women	64.3%	99.4%	75.2%	99.0%	2.8%	786,899
Arab-dense counties	57.6%	97.9%	60.1%	7.7%	5.5%	657,728
non Arab-dense counties	22.4%	99.6%	37.7%	99.2%	1.0%	916,070
Foreign born	64.0%	92.4%	60.1%	93.5%	17.8%	177,394
US-born	27.5%	99.6%	47.5%	99.2%	1.2%	1,396,404