American Journal of Epidemiology
Copyright © 2006 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved; printed in U.S.A.

Vol. 163, No. 5
DOI: 10.1093/aje/kwj057
Advance Access publication January 18, 2006

Practice of Epidemiology

Reliability of Self-reported Ancestry among Siblings: Implications for Genetic Association Studies

Melinda S. Burnett[1], Kari J. Strain[2], Timothy G. Lesnick[2], Mariza de Andrade[2], Walter A. Rocca[1,2], and Demetrius M. Maraganore[1]

[1] Department of Neurology, Mayo Clinic College of Medicine, Rochester, MN.
[2] Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN.

Received for publication February 3, 2005; accepted for publication October 13, 2005.

This study investigated the reliability of self-reported ancestry by comparing the interview responses of probands and their siblings. A total of 546 sibling pairs were ascertained in a family-based study of susceptibility genes for Parkinson's disease and asked to identify maternal and paternal countries of origin. Probands were recruited prospectively from the Department of Neurology of the Mayo Clinic in Rochester, Minnesota, from June 1, 1996, through May 31, 2005. Probands resided within Minnesota or one of the four surrounding states (Wisconsin, Iowa, South Dakota, North Dakota). Only 49 percent of these sibling pairs, primarily Caucasian, agreed completely on the countries of origin of both parents. The agreement increased to 68 percent when named countries were postcoded into six population genetic clusters (as previously defined by microsatellite markers). Self-reported ancestry may not be a reliable method to reduce the possible impact of population stratification in genetic association studies of outbred populations, such as in the United States.

alleles; case-control studies; ethnic groups; population groups; reproducibility of results

Genetic association studies investigate susceptibility genes for complex diseases; however, population stratification may confound their results in outbred populations (1). Unequal distributions of ethnic strata in cases and controls may lead to spurious allelic associations, because the frequency of genetic polymorphisms often varies according to ancestral origins. Genetic association studies use various methods to reduce the impact of population stratification (2). Matching of cases and controls according to self-reported ancestry is the simplest and most economical approach. However, few studies have evaluated the reliability of self-reported ancestry. This study investigated the reliability of self-reported ancestry by comparing the interview responses of a series of probands and siblings (sharing the same parents) ascertained as part of a family-based study of Parkinson's disease.

MATERIALS AND METHODS

Sibling pairs

We included probands and their siblings from an ongoing family-based study of susceptibility genes for Parkinson's disease. Probands were recruited prospectively and sequentially from the Department of Neurology of the Mayo Clinic in Rochester, Minnesota, from June 1, 1996, through May 31, 2005. Probands resided within Minnesota or one of the four surrounding states (Wisconsin, Iowa, South Dakota, North Dakota). Recruitment of potential probands was through the Neurology Appointment Center. The recruitment was monitored for completeness through a computerized tracking system. All probands underwent a clinical assessment at enrollment to determine eligibility. Enrolled probands provided

Correspondence to Dr. Demetrius M. Maraganore, Department of Neurology, Mayo Clinic College of Medicine, 200 First Street S.W., Rochester, MN 55905 (e-mail: dmaraganore@mayo.edu).

genealogic information and permission to contact their siblings. We recruited all living siblings aged 40 years or older. Siblings were screened for Parkinson's disease by use of a validated telephone screening instrument (3). In addition, both probands and siblings underwent a series of interviews and assessments that documented demographic, clinical, and risk factor information.

Because Parkinson's disease is often accompanied by cognitive impairment that can limit the quality of self-reported ancestry information, we excluded from this study subjects with cognitive impairment. For examined subjects (cases, siblings who screened positive for Parkinson's disease), we defined cognitive impairment as a Mini-Mental State Examination score of less than 24. For subjects who were only interviewed (siblings who screened negative for Parkinson's disease), we defined a cognitive impairment as a Telephone Interview for Cognitive Status score of less than 28 (4, 5).

For this reliability study, we compared the self-reported race and ancestry responses of subjects with Parkinson's disease (probands) with those of a single full sibling (sharing both parents). Probands or siblings who provided no information regarding ancestry were excluded. If a proband had multiple siblings, only one sibling was selected for inclusion. We first matched siblings to probands on gender (when possible) and then by closest age at study. Additional details regarding proband and sibling methods, including the diagnostic criteria for Parkinson's disease, the clinical assessments performed in examined subjects, the screening interview for Parkinson's disease in siblings, and the risk factors interview used for both probands and siblings, were previously reported (6).

Assessment of ancestry

The probands and their siblings were asked to list their father's countries of origin (up to four free-form answers) and their mother's countries of origin (up to four free-form answers). Our script read verbatim: ''*What is the country of origin of your father's ancestors? For example, were they from Germany, Ireland, or China?*'' Multiple answers were allowed. The same question was asked for mother's ancestors. They were also asked to define their race by selecting from the following categories: Caucasian, African American, American Indian/Alaskan Native, Asian, Native Hawaiian/Pacific Islander, and Hispanic. Only information obtained directly from probands and siblings was included in this reliability study. Information obtained via proxy interview was excluded. The Mayo Clinic Investigational Review Board approved all study methods.

To determine whether agreement improved when ancestry was more broadly defined, we postcoded self-reported countries of origin according to population genetic clusters (as previously defined by studies using microsatellite markers) (7). The clusters were Eurasia, East Asia, Oceania, America, Africa, and the Kalash group of Pakistan. As the majority of our subjects reported European ancestry, we further subdivided the European continent into geographically defined subcontinents (i.e., Northern, Central, and Southern Europe). ''Northern European'' included Scandinavian, Swedish, Norwegian, Finnish, Danish, Irish, or British origins. ''Central European'' included French, Belgian, Dutch, Swiss, Luxembourgian, German, Austrian, Hungarian, Polish, Czechoslovakian, or Russian origins. ''Southern European'' included Italian, Spanish, Portuguese, Greek, or Yugoslavian origins.

Statistical analyses

We evaluated overall agreement on countries of origin. In addition, we stratified analyses according to the maximum number of countries listed by any member of a sibling pair, in order to determine whether agreement would decline when at least one member reported multiple countries of origin. We also stratified analyses by the year of birth of the proband, in order to determine whether older generations were more knowledgeable of their ancestral origins. Finally, we restricted analyses to female-female and male-male sibling pairs, in order to determine whether there were gender-specific differences in the reliability of self-reported ancestry.

For all proband-sibling comparisons of ancestry, we used two different measures of intersibling reliability. Using the ''strict'' measure, sibling pairs were considered in agreement if their answers matched completely. Using the ''liberal'' measure, sibling pairs were considered to be in agreement if their answers matched for at least one item (8).

We also performed sensitivity analyses, excluding probands or siblings with Parkinson's disease and comparing self-reported ancestry for other sibling pairs (one pair per family, selected for the same gender when possible and then for the closest age at study). We used a paired t test to compare age at study between probands and siblings. We used McNemar's test for comparisons of gender and years of education between probands and siblings.

RESULTS

Sibling pairs

More than 90 percent of eligible probands and more than 90 percent of eligible siblings participated in the study. We identified 546 proband-sibling pairs who were informative regarding their parents' countries of origin (table 1). The probands and siblings did not differ significantly in age ($p = 0.07$) or educational level ($p = 0.80$). They did, however, differ slightly in gender, with more probands being male and more siblings being female ($p = 0.02$). The subjects were well educated, with only 9.5 percent of probands ($n = 52$) and 9.0 percent of siblings ($n = 49$) having less than 12 years of education.

Reliability of self-reported ancestry

Overall agreement for race was excellent, with 99 percent agreement by use of the strict method and 100 percent agreement by use of the liberal method (table 2). Nearly all subjects (99 percent) were Caucasian. Among non-Caucasians, agreement for self-reported race was perfect, except for the few pairs who claimed American Indian/Alaskan Native

TABLE 1. Demographic characteristics of study subjects recruited at the Mayo Clinic, Rochester, Minnesota, between 1996 and 2005

Characteristic	Probands (n = 546)		Siblings (n = 546)		p value
	Median	Range	Median	Range	
Age (years) at study	67	36–90	66	35–89	0.07*
	No.	%	No.	%	
Gender					
Male	336	61.5	270	49.5	0.02†
Female	210	38.5	276	50.5	
Education (years)					
<12	52	9.5	49	9.0	0.80†
12	198	36.3	192	35.2	
13–15	116	21.3	124	22.7	
>15	179	32.8	181	33.2	
Unknown	1	0.2	0	0.0	

* Paired *t* test.

† McNemar's test.

ancestry (one of three pairs agreed, 33 percent). Because of regional demographics and referral patterns in the upper Midwest, no African Americans were included in this study.

By use of the strict method of comparison, 62 percent of pairs ($n = 339$) agreed on their mother's countries of origin and 67 percent of pairs ($n = 366$) agreed on their father's countries of origin (table 3). When agreement for both parents' origins combined was analyzed, only 49 percent of sibling pairs ($n = 265$) were in complete agreement by use of the strict method. Results improved with the liberal method of comparison, with 81 percent of pairs ($n = 441$) agreeing partially on their mother's countries of origin, 82 percent of pairs ($n = 449$) agreeing partially on their father's countries of origin, and 91 percent of pairs ($n = 495$) agreeing partially regarding either parent's countries of origin.

Female-female pairs had greater agreement for either the strict or liberal method than did male-male pairs. Specifically, for female-female pairs, the agreement was 52 percent (strict method) or 93 percent (liberal method). For male-male pairs, the agreement was 48 percent (strict method) or 90 percent (liberal method). The percentage of agreement decreased with increasing number of countries listed by either member of the pairs. For example, by the strict method, 80 percent of pairs agreed on their mother's country of origin when only one answer was given, but only 19 percent agreed when two countries were given. The liberal method revealed a less dramatic decrease in agreement with multiple answers; in fact, agreement increased with the number of answers for father's countries of origin. The level of agreement did not vary noticeably by year of birth of the proband for either the strict or liberal methods.

Surprisingly, the postcoding of self-reported countries of origin according to broader geographic definitions resulted in only modest improvements in agreement (table 4). When postcoding ancestry according to population genetic clusters (as previously defined by microsatellite markers) (7), we found that only 68 percent of pairs ($n = 373$) agreed completely regarding both parents using the strict method, while 100 percent of pairs ($n = 546$) agreed using the liberal method. For the mother's clusters of origin, we found that 77 percent of pairs ($n = 423$) agreed using the strict method and 100 percent of pairs ($n = 544$) agreed using the liberal method. For the father's clusters of origin, we found that 83 percent of pairs ($n = 453$) agreed using the strict method and 100 percent of pairs ($n = 546$) agreed using the liberal method.

TABLE 2. Frequency of agreement for self-reported race among study subjects recruited at the Mayo Clinic, Rochester, Minnesota, between 1996 and 2005

Stratum	Sibling pairs (n = 546)			
	Strict method* agreement		Liberal method† agreement	
	No.	%	No.	%
Overall	543/546	99	545/546	100
By subgroup‡				
Caucasian	540/541	100	540/541	100
Caucasian/American Indian/ Alaskan Native	1/3	33	3/3	100
Asian	1/1	100	1/1	100
Native Hawaiian/Pacific Islander	1/1	100	1/1	100

* The strict method defines a sibling pair to be in agreement if they agree completely on self-reported ancestry.

† The liberal method defines a sibling pair to be in agreement if they agree at least partially on self-reported ancestry.

‡ Because of regional demographics and patterns of referral in the upper Midwest, no African Americans were included in this study.

TABLE 3. Frequency of agreement for self-reported countries of origin among study subjects recruited at the Mayo Clinic, Rochester, Minnesota, between 1996 and 2005

	Sibling pairs ($n = 546$)			
Stratum	Strict method* agreement		Liberal method† agreement	
	No.	%	No.	%
Parents' countries of origin				
Overall	265/546	49	495/546	91
Female-female pairs	87/168	52	156/168	93
Male-male pairs	109/228	48	205/228	90
Mother's countries of origin				
Overall	339/546	62	441/546	81
By maximum no. of countries listed				
1	315/396	80	315/396	80
2	23/122	19	105/122	86
3	1/23	4	18/23	78
4	0/5	0	3/5	60
By year of birth of the proband				
1951–1975	27/40	68	33/40	83
1926–1950	239/388	62	312/388	80
1900–1925	73/118	62	96/118	81
Father's countries of origin				
Overall	366/546	67	449/546	82
By maximum no. of countries listed				
1	353/435	81	353/435	81
2	13/98	13	83/98	85
3	0/11	0	11/11	100
4	0/2	0	2/2	100
By year of birth of the proband				
1951–1975	29/40	73	33/40	83
1926–1950	264/388	68	324/388	84
1900–1925	73/118	62	92/118	78

* The strict method defines a sibling pair to be in agreement if they agree completely on self-reported ancestry.

† The liberal method defines a sibling pair to be in agreement if they agree at least partially on self-reported ancestry.

Within the Eurasian cluster, the majority of subjects reported Central or Northern European subcontinent origins (table 5). Only 57 percent of pairs ($n = 312$) agreed regarding both parents' European subcontinents of origin by use of the strict method, and 93 percent of pairs ($n = 506$) agreed by use of the liberal method. For the mother's subcontinents of origin, 71 percent of pairs ($n = 361$) agreed by use of the strict method and 91 percent of pairs ($n = 463$) agreed by use of the liberal method. For the father's subcontinents of origin, 76 percent of pairs ($n = 390$) agreed by use of the strict method and 91 percent of pairs ($n = 468$) agreed by use of the liberal method.

When we performed sensitivity analyses restricted to unaffected sibling pairs ($n = 357$), the results were similar to those for the proband-sibling pairs (data not shown). For the unaffected sibling pairs, 52 percent agreed completely on the countries of origin of both parents (as compared with 49 percent for proband-sibling pairs). For the unaffected sibling pairs, 68 percent agreed completely on the countries of origin of both parents when the named countries were post-coded into six population genetic clusters (as compared with 68 percent for proband-sibling pairs).

DISCUSSION

This study of proband-sibling pairs demonstrates that self-reported ancestry has limited reliability. The degree of disagreement within sibling pairs regarding the countries of origin of their parents was striking, particularly when a mother or father had more than one country of origin.

TABLE 4. Frequency of agreement for population genetic clusters of origin* among study subjects recruited at the Mayo Clinic, Rochester, Minnesota, between 1996 and 2005

Stratum	Sibling pairs (n = 546)			
	Strict method† agreement		Liberal method‡ agreement	
	No.	%	No.	%
Parents' clusters of origin	373/546	68	546/546	100
Mother's clusters of origin				
Overall	423/546	77	544/546	100
Eurasia	421/540	78	540/540	100
East Asia	1/1	100	1/1	100
Oceania	0/0	0	0/0	0
America	1/5	20	3/5	60
Africa	0/0	0	0/0	0
Father's clusters of origin				
Overall	453/546	83	546/546	100
Eurasia	452/543	83	543/543	100
East Asia	1/1	100	1/1	100
Oceania	0/0	0	0/0	0
America	0/2	0	2/2	100
Africa	0/0	0	0/0	0

* Genetic cluster of origin as previously reported by Rosenberg et al. (Science 2002;298: 2381–5) (7).

† The strict method defines a sibling pair to be in agreement if they agree completely on self-reported ancestry.

‡ The liberal method defines a sibling pair to be in agreement if they agree at least partially on self-reported ancestry.

Residual disagreement was substantial even when ancestral regions of origin were more broadly defined. Only 68 percent of sibling pairs agreed completely on the population genetic clusters of origin of both parents (strict method). It was surprising that older generations, temporally more proximate to their ancestral migrations, did not provide more reliable information regarding countries or regions of origin than did younger generations. It was also noteworthy that agreement was better for the father's countries or regions of origin than for the mother's countries or regions of origin. This perhaps reflects cultural traditions emphasizing the father's heritage over the mother's, such as taking the father's surname.

One possible limitation of our study was that the probands and siblings included were often of different genders. It is possible that the reliability of self-reported ethnicity is gender specific. We explored this possibility by restricting analyses to male-male and female-female pairs. Even among same-gender pairs, disagreement was substantial. Another limitation is that the wording of the ancestry questions that we used may have contributed to disagreement. Although we allowed for multiple countries of origin per parent, we did not state this verbatim. Unfortunately, our study was limited primarily to Caucasian subjects. We were unable to provide reliability information for other racial groups.

There are only limited studies on the reliability of self-reported ancestry. One compared the race data reported by the parents of children in the third and fourth grades who participated in the Cardiovascular Health in Children Study (9). When independently asked to assign only one race to their child, 2.6 percent of 2,164 parent pairs disagreed on the race of their child. Of the 1,048 children asked to report a race, 5 percent disagreed with the race assigned by their parents. Other studies reported only limited reliability for race. One study found that self-reported race varied over time, with only 58.3 percent of 5,991 Americans in the National Health and Nutrition Examination Survey reporting the same race after 10 years (8). In another survey of US households, only 65 percent of respondents were classified with the same ancestry after 1 year (10). Poor reliability may be caused by changing social attitudes and definitions of race, by the challenge of placing simple racial labels on complex ethnic heritages, or by limited commitment of respondents to participation in the surveys. Self-reported race may be perceived as a marker of cultural identity rather than of genetic ancestry (11), further limiting the value of this information for genetic association studies.

There is considerable debate as to whether, and when and to what extent, population stratification is a confounder of genetic association studies, and on how to remedy the

TABLE 5. Frequency of agreement for European subcontinents of origin among study subjects recruited at the Mayo Clinic, Rochester, Minnesota, between 1996 and 2005

Stratum	Sibling pairs ($n = 546$)			
	Strict method* agreement		Liberal method† agreement	
	No.	%	No.	%
Parents' subcontinents of origin	312/546	57	506/546	93
Mother's subcontinents of origin				
European	361/511	71	463/511	91
Northern European‡	151/213	71	192/213	90
Central European§	195/258	76	233/258	90
Southern European¶	5/6	83	5/6	83
European, mixed regions	10/34	29	33/34	97
Father's subcontinents of origin				
European	390/515	76	468/515	91
Northern European‡	160/208	77	187/208	90
Central European§	217/279	78	255/279	91
Southern European¶	7/7	100	7/7	100
European, mixed regions	6/21	29	19/21	90

* The strict method defines a sibling pair to be in agreement if they agree completely on self-reported ancestry.

† The liberal method defines a sibling pair to be in agreement if they agree at least partially on self-reported ancestry.

‡ "Northern European" includes Scandinavian, Swedish, Norwegian, Finnish, Danish, Irish, or British origins.

§ "Central European" includes French, Belgian, Dutch, Swiss, Luxembourgian, German, Austrian, Hungarian, Polish, Czechoslovakian, or Russian origins.

¶ "Southern European" includes Italian, Spanish, Portuguese, Greek, or Yugoslavian origins.

problem (12–16). One study used empirical data (for one gene and one disease from one US study of non-Hispanic Caucasians of European origin) and a statistical estimation of the confounding risk ratio to stimulate a bias of less than 1 percent from population stratification. Evaluation of a wide range of allele frequencies and representative disease rates that exist across European populations predicted that the risk ratio might be biased by less than 10 percent in US studies that ignore ethnicity (15). By contrast, another study examined approximately 15,000 genome-wide single nucleotide polymorphisms typed in three population groups and noted that the consequences of population structure increased markedly with sample size. The authors concluded, "For the size of study needed to detect typical genetic effects in common diseases, even the modest levels of population structure within population groups cannot safely be ignored" (16, p. 512). In addition, it was recently shown that even within a relatively homogenous genetic isolate, there was substantial substructure (17). Matching of cases and unrelated controls for self-reported ancestry is a simple and economical approach, and a recent study suggested that self-reported race is strongly correlated with a cluster of genetic markers (18). However, our findings suggest that self-reported ancestry is not a reliable method to prevent confounding. For studies of cases and unrelated controls, the genotyping of hundreds of additional markers to define population substructure ("genomic control") is presently costly and beyond the scope of many academic laboratories. Furthermore, genomic control may not correct for structure if too few genomic markers are used and, in other settings, may result in overcorrections with a loss of statistical power (16). For populations with low non-paternity rates, genetic association studies of discordant sibling pairs may avoid the pitfall of population stratification. However, despite several strengths, family-based case-control studies may have reduced statistical power because of extensive matching of pairs, are prone to selection biases when siblings are unavailable, and are prone to confounding when pairs are of different genders and ages (19). In conclusion, while population stratification and its confounding of genetic association studies remain controversial, none of the available remedies (including matching for self-reported ancestry) is entirely satisfactory.

ACKNOWLEDGMENTS

This study was supported by grants R01 ES10751 and R01 NS33978 from the National Institutes of Health.

Conflict of interest: none declared.

REFERENCES

1. Knowler WC, Williams RC, Pettitt DJ, et al. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 1988;43:520–6.
2. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet 2003;361:598–604.
3. Rocca WA, Maraganore DM, McDonnell SK, et al. Validation of a telephone questionnaire for Parkinson's disease. J Clin Epidemiol 1998;51:517–23.
4. Folstein MF, Folstein SE, McHugh PR. ''Mini-mental state.'' A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–98.
5. Welsh K, Breitner JCS, Magruder-Habib K. Detection of dementia in the elderly using telephone screening of cognitive status. Neuropsychiatry Neuropsychol Behav Neurol 1993; 6:103–10.
6. Rocca WA, Peterson BJ, McDonnell SK, et al. The Mayo Clinic family study of Parkinson's disease: study design, instruments, and sample characteristics. Neuroepidemiology 2005;24:151–67.
7. Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. Science 2002;298:2381–5.
8. Hahn RA, Truman BI, Barker ND. Identifying ancestry: the reliability of ancestral identification in the United States by self, proxy, interviewer, and funeral director. Epidemiology 1996;7:75–80.
9. Bomar PJ, Harrell JS, Webb JP. Family member discrepancies in report of a child's race. J Cult Divers 1997;4:104–9.
10. Polednak AP. Racial and ethnic differences in disease. New York, NY: Oxford University Press, 1989.
11. Blustein J. The reliability of racial classifications in hospital discharge abstract data. Am J Public Health 1994;84: 1018–21.
12. Wacholder S, Rothman N, Caporaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 2002;11:513–20.
13. Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002;11: 505–12.
14. Risch N, Burchard E, Ziv E, et al. Categorization of humans in biomedical research: genes, race and disease. Genome Biol 2002;3:1–12.
15. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiological studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 2000; 92:1151–8.
16. Marchini J, Cardon LR, Phillips MS, et al. The effects of human population structure on large genetic association studies. Nat Genet 2004;36:512–17.
17. Helgason A, Yngvadottir B, Hrafnkelsson B, et al. An Icelandic example of the impact of population structure on association studies. Nat Genet 2005;37:90–5.
18. Tang H, Quertermous T, Rodriguez B, et al. Genetic structure, self-identified race/ethnicity, and confounding in case-control studies. Am J Hum Genet 2005;76:268–75.
19. Maraganore DM. Blood is thicker than water: the strengths of family-based case-control studies. Neurology 2005;64: 408–9.