Development and Representativeness of a Large Population-Based Cohort of Native Californian Twins

Myles G. Cockburn, Ann S. Hamilton, John Zadnick, Wendy Cozen, and Thomas M. Mack

University of Southern California/Keck School of Medicine

We have established a large cohort of twins to facilitate studies of the role of genetics and environment in the development of disease. The cohort has been derived from all multiple births occurring in California between 1908–82 (256,616 in total). We report here on our efforts to contact these twins and their completion of a detailed 16 page risk factor questionnaire. Addresses of the individuals were obtained by linking the birth records with the California Department of Motor Vehicles (DMV) roster of licensees. To date this has been completed for twins born between 1908 and 1972 (200,589 individuals). The linkage has revealed 112,468 matches and, because of less complete DMV records in some years, was less successful in older females than in younger females and all males. Over 41,000 twins have participated by completing the questionnaire. Based on estimates of numbers of individuals receiving a questionnaire, we estimate our crude response rate to be between 42.2% and 49.6%, highest among females in their 40s (62.8%). We describe the representativeness of the twins in the original birth cohort, those identified by the linkage, and those completing the questionnaire. Compared to the 1990 resident population of California-born resident singletons, the respondents were of similar age, sex, race and residential distribution (for although we were able to locate fewer older females, they had a higher response rate), but were less likely to have been educated for more than 12 years. We provide a brief synopsis of studies nested within this cohort. We also elucidate our plans for expanding the cohort in the near future.

Twins offer great advantages as subjects for the study of disease (Martin et al., 1997). They are fully or partially matched on genetic determinants, share a common childhood environment, and can often describe the relative (twin vs. co-twin) differences in their past experience (Hamilton & Mack, 2000).

We have previously established a roster of volunteer twins with cancer and other chronic diseases from North America, and have made use of subsets for purposes of chronic disease epidemiology (Mack et al., 2000). The advantages of that roster were the large number of pairs affected with pertinent serious conditions, and the high level of compliance offered by self-selected participants. The disadvantages include the unrepresentative nature of the subjects with less severe conditions, and the absence of exposure criteria by which to choose cohorts for prospective studies (Mack et al., 2000). To overcome these liabilities requires a different resource, one with a large sample representative of the population. The difficult task is to identify a roster of twin subjects that is large enough, unbiased in terms of both common exposure and common outcome, and representative of the source population (Hawkes, 1997).

Here we describe the establishment of such a twin resource, a cohort of twins based on the 256,616 twin individuals born in California between 1908 and 1982. Included are all twin subjects with and without pertinent environmental exposures, healthy and diseased, paired and surviving, monozygotic and dizygotic, and like-sex and unlike-sex. We describe the formation and enrollment of members of the cohort, discuss the degree to which its members are representative of California twins and Californians generally, and describe the potential for using it to address pertinent scientific questions.

Materials and Methods

Cohort Establishment and Recruitment of Participants

Records of live multiple births in the state of California occurring between 1908 and 1982 were obtained from the California Department of Vital Statistics. This set was linked to the records of the California Department of Motor Vehicles (DMV) in 1989, 1998, 1999 and in 2000 using first and last name (linked to first, last and "a.k.a." name of DMV record) and date of birth, and returned a current address and new married name, when available. DMV records were computerized in the 1960s, with entry of names of then-current driver's licenses. Thus the maiden names of women then holding licenses under a married name were not entered into the record. Individuals are required to update their address information with the DMV, in order that they receive vehicle registration papers. If they have no vehicle, they are still required to update address information on their license/ID every 4 years, if they are still residing in California.

The resulting file contained all matches linked to a driver's license or California Identification card, with a current address (regardless of place) and date of last information update. This report includes results of our efforts to

Address for correspondence: Myles G. Cockburn, 1441 Eastlake Ave MC9175, Los Angeles CA90089-9175, USA. Email: cockburn@usc.edu

https://doi.org/10.1375/twin.4.4.242 Published online by Cambridge University Press

recruit twins born between 1908 and 1972. Recruitment of twins born 1973 to 1982 is ongoing.

We refer to those twins identified from birth records as the 'birth cohort' of California native twins, as opposed to those for whom we have subsequently received questionnaires, whom we describe as the 'respondents'. Future studies will be able to use the birth cohort for linkage with cancer registries and death records, whereas more detailed studies of exposure will employ the respondent set.

We carried out recruitment in 3 'waves', one in each of 1991, 1998 and 1999. Each 'wave' was conducted in a similar manner. After comparing the DMV-linked file to the National Change of Address Index (NCOA) to remove or update addresses (due to typographical errors, or out-of-date street names or zip codes), we sent letters of invitation to the twins with valid addresses. The letter contained a reply-paid postcard for the twin to update their information, inform us of the location of their twin, and space to inquire about the study. At this stage, and at each subsequent mailing, Address Service and Forwarding was requested, so that we could update our records, or remove twins with incorrect addresses from the subsequent questionnaire mailing.

In order to provide sufficient time to receive returned mail and update our records, on average three weeks after mailing the introductory letter we mailed a 16 page questionnaire (available for viewing at http://twins.usc.edu/questionnaire) with its reply-paid envelope to all those individuals whose addresses were thought to be valid. After a further 5 weeks, allowing time for questionnaire responses and return addressing updates, we sent a reminder post-card again urging twins to take part in the study.

After a further 3 weeks (on average), we sent a second copy of the questionnaire to those twins not yet responding, refusing, or whose package was returned with an unknown address. This mailing included twins whose addresses were added after DMV linkage by virtue of information from Post Office address update, or from twins notifying us of the whereabouts of their co-twin ("late additions").

Evaluation of Non-Response

In order to evaluate possible reasons for non-response related to incorrect address information, we conducted a study of 206 respondent and 204 non-respondent individuals, randomly chosen within blocks stratified by geographical region within California. We conducted independent searches for address information and telephone numbers using a local reverse directory look-up, and Internet search engines. Those non-respondents for whom telephone numbers were found were interviewed to verify receipt of the questionnaire and determine reasons for non-response. Along with a sample of respondent twins, they were asked about recent address changes that might affect delivery of the questionnaire. We employed a private investigator to independently trace a sample of both groups using records from a credit-reporting agency that contains regular address information updates on most individuals resident in California. The proportion of 'found' respondents was compared to the proportion of 'found' non-respondents to estimate the proportion of non-respondents who did not receive the questionnaire.

Estimation of the True Denominator of Twins Able to Receive a Questionnaire and Subsequent Response Rates

We made estimates of the true denominator for our response rates based on a number of assumptions. We eliminated those whose addresses were rejected by the NCOA linkage and for whom we did not subsequently receive a new address in future DMV linkages. From those who were sent a questionnaire, we eliminated those returned by the Post Office, due to lack of a forwarding address, expiration of the mail forwarding notice, or insufficient address. Second, we reduced the population of potentially-respondent twins by the excess proportion of non-respondent twins (over and above the proportion of respondent twins) for whom no address could be found in our sub-study (above), on the assumption that they had either moved out of state or were deceased. We assessed the variation in response rates, by age, sex and geographical location (county) to determine potential sources of bias.

Finally, we considered the breakdown of respondents by zygosity, derived from self-report, and compared the pairwise distribution to that of the original cohort to assess their representativeness by sex and zygosity.

The Questionnaire

The 16 page questionnaire asked about basic demographic characteristics (age, sex, education, occupation, marital status), perceived zygosity (Kasriel & Eaves, 1976), growth and development, reproductive history, use of medical services, dietary preference, disease experience (including cancer occurrence), and lifestyle choices (smoking, alcohol consumption, exercise, sun exposure). For some questions, participants were asked to compare themselves to their co-twin (eg "How much taller or shorter than you is your twin" and "How much earlier or later did your twin have her first menstrual period?").

Source of Comparison Population(s) for Determining Respondent Representativeness

We compared the birth cohort and the questionnaire respondents to the 1990 California resident population of California natives. These comparison data were obtained from the Integrated Public Use Microdata Series (IPUMS — http://www.ipums.umn.edu; Ruggles & Sobek, 1997). We used the 5% California State sample of the 1990 census, obtaining data on age, sex, race/ethnicity, education and 1990 county of residence for individuals in that sample. We used 8 categories when comparing race in the Census data set to race derived from questionnaire responses (White, Latino, Black, American Indian, Japanese/Chinese, Filipino/Thai, other and missing) derived from identical questions. The census data was available only for those over the age of 33 years, so age comparisons are available for only 5 age groups. For comparisons of census data to data regarding non-respondents, we use only 4 race/ethnicity categories, because birth records only distinguish White/Latino, Black, Asian and Other race. For additional education comparisons for the California-born population no longer resident in California, we used the 1% sample of the entire US 1990 census population, also from IPUMS.

The number of twin pairs of each zygosity-gender type (identical or fraternal, male, female or mixed) was

https://doi.org/10.1375/twin.4.4.242 Published online by Cambridge University Press

compared to the number of expected live births in the Pacific States as reported previously (Mack et al., 2000).

Results

There were 200,590 multiple births registered in the State of California between 1908 and 1972, slightly more females (50.2%) than males. The entire 1908–72 cohort was aged 16 years or greater in 1990, and was therefore eligible to have an active record in the DMV. Their distribution by age (and year of birth) and sex is given in Table 1.

Linkage with the Department of Motor Vehicles

There were 112,468 successful results from linkage of this cohort with the records of the DMV. Female names were less successfully linked, especially those over age 43, with the disparity between the sexes increasing with increasing age (Table 1). While more than 60% of men under the age of 53 years and women under the age of 43 years were found, this proportion was almost halved for age-groups of males over 73 years. Less than 20% of females in any age group over 53 years were found.

Other Forms of Ascertainment

We found a further 8,214 twins by other means, the majority by referral from respondent co-twins. A small number of others heard about our study and contacted us (via website: http://twins.usc.edu, or by telephone). Females were more likely than males to be ascertained in this manner (not shown).

Address Refinement and Postal Service Returns

The process of 'cleaning' addresses obtained from the DMV resulted in the alteration or deletion of 12,542 addresses, mostly because of out of date zip codes. Address correction service resulted in the return of 5,882 letters, and subsequently we mailed a total of 102,258 questionnaires. The distribution of 18,424 individuals who were not sent a questionnaire for any reason did not differ by gender, but younger people were more likely to be excluded by incorrect address information (not shown). Where DMV linkage was successful for females, the address we obtained was less likely to result in a postal return or NCOA deletion. For both males and females, the youngest age groups (under 33 years) suffered the greatest loss from invalid addressing. Females aged 33 to 52 years had less than 10% loss due to postal returns or NCOA deletions.

Postal returns from each 'wave' varied, and the following describes only the final outcome for each member of the cohort — for example, if no match was found in an early DMV linkage, but subsequently a match was found and a questionnaire received, the individual was counted as a successful DMV linkage and questionnaire return.

Response and Response Rates

Of 102,258 questionnaires sent to individuals, only 98,105 could possibly have received the questionnaire — the remainder were either deceased or without a known address. Of these, 41,367 were returned for a crude overall response rate of 42.2% (41,367/98,105). Above the age of 52 years, males were slightly more likely than females to respond, but under the age of 53 years, females were far more likely to respond, with almost twice the proportion of females than males responding in the youngest twins, those aged 18 to 23 years (Table 2). While the majority of respondents were White, response rates were similar in Whites and Asians, and substantially lower among the small number of African-American twins and those reporting "other" race. These differences did not vary markedly by age (not shown) or sex (Table 2).

Sub-Study Evaluating Non-Response

Of the 206 respondents and 204 non-respondents selected for our sub-study, we were able to independently locate 77 (37.4%) and 58 (28.4%) respectively, using reverse directory and Internet searches. The remaining 146 non-respondents and a 25% sample of respondents were traced using a credit reporting agency. Using all available means, 75% of the respondent sample could be located, whereas 55% of the non-respondent sample could be located. We therefore estimated that 20% of the non-

Table 1

Demographic Comparison of Twins Identified in the California Twin Program and the Original Birth Cohort from Whom They Were Drawn, by Age and Sex

		California birth record of multiple births, 1908–1972 (n = 200,589)				Twins found by any method (n = 120,682)					
		Males		Females		Males			Females		
Birth year	Age (1990)	Number	% distribution	Number	% distribution	Number	% distribution	% found	Number	% distribution	% found
1908–17	73–82	2,596	2.6%	2,645	2.6%	687	1.0%	26.5%	116	0.2%	4.4%
1918–27	63–72	4,853	4.9%	5,121	5.1%	2,037	3.1%	42.0%	554	1.0%	10.8%
1928–37	53–62	5,956	6.0%	6,164	6.1%	3,295	5.0%	55.3%	942	1.7%	15.3%
1938–47	43–52	12,381	12.4%	12,801	12.7%	7,912	12.1%	63.9%	5,130	9.3%	40.1%
1948–57	33–42	24,958	25.0%	25,174	25.0%	18,119	27.7%	72.6%	16,508	29.9%	65.6%
1958–67	23–32	34,021	34.0%	33,575	33.4%	23,085	35.2%	67.9%	21,737	39.4%	64.7%
1968–72	18–22	15,207	15.2%	15,137	15.0%	10,362	15.8%	68.1%	10,198	18.5%	67.4%
Total		99,972	100%	100,617	100%	65,497	100%	65.5%	55,185	100%	54.8%

https://doi.org/10.1375/twin.4.4.242 Published online by Cambridge University Press

Table 2

Response Rates Among All Twins Identified in the California Twin Program (n = 41,367 Respondents) by Age, Sex and Race

		Male	Response rates			Female	Response rates		
			1	2	3		1	2	3
Birth year	Age (1990)	Responses	%	%		Responses	%	%	
1908–17	73–82	282	49.1%	55.7%	*	48	49.5%	51.1%	*
1918–27	63–72	1,026	56.4%	59.7%	*	265	55.1%	57.9%	*
1928–37	53–62	1,687	58.1%	60.5%	*	446	54.8%	56.8%	*
1938–47	43–52	3,501	50.8%	52.3%	*	2,849	61.2%	62.8%	*
1948–57	33–42	6,555	41.3%	42.9%	*	8,492	56.1%	57.7%	*
1958–67	23–32	4,673	25.7%	27.0%	*	7,135	40.5%	42.1%	*
1968–72	18–22	1,609	18.8%	19.8%	*	2,799	32.3%	34.2%	*
Race									
White/Latino		18,399	37.6%	39.3%	†	20,530	49.1%	50.9%	†
Black		516	11.7%	12.5%	†	928	21.2%	22.4%	†
Asian		335	35.4%	36.5%	†	397	49.3%	50.4%	†
Other		83	16.6%	17.2%	†	178	35.7%	36.9%	†
Overall response and response rates		19,333	35.3%	36.9%	41.5%	22,034	46.4%	48.2%	54.6%

1 — Response rate as % of twins to whom we sent a questionnaire

2 — Response rate as % of twins likely to have received the questionnaire (removed those known to be deceased and Post Office returns)

3 — Response rate as % of twins we believe actually received the questionnaire (removed 15% of non-respondents) on the basis of our sub-study

* No age-specific data available † No race-specific data available

Table 3

Refusal Rates Among All Twins Approached in the California Twin Program, by Age and Sex, as a Percentage of Those Receiving a Questionnaire

		Male				Female			
		Hard[1]		All[2]		Hard[1]		All[2]	
Birth year	Age (1990)	Responses	%	Responses	%	Responses	%	Responses	%
1908–17	73–82	32	5.6%	41	7.1%	7	7.2%	7	7.2%
1918–27	63–72	87	4.8%	117	6.4%	33	6.9%	41	8.5%
1928–37	53–62	89	3.1%	119	4.1%	33	4.1%	48	5.9%
1938–47	43–52	133	1.9%	179	2.6%	99	2.1%	119	2.6%
1948–57	33–42	213	1.3%	322	2.0%	225	1.5%	318	2.1%
1958–67	23–32	188	1.0%	526	2.9%	158	0.9%	420	2.4%
1968–72	18–22	38	0.4%	170	2.0%	47	0.5%	171	2.0%
Race									
White/Latino		724	1.5%	1,349	2.9%	546	1.4%	997	2.5%
Black		37	0.9%	83	2.0%	35	0.9%	79	1.9%
Asian		14	1.5%	25	2.7%	15	1.9%	31	3.9%
Other		5	1.0%	17	3.5%	6	1.2%	17	3.5%
Totals		780	1.4%	1,474	2.7%	602	1.3%	1,124	2.4%

[1] 'Hard' — age-specific firm refusals from any source (e-mail, mail in card, telephone call)

[2] 'All' — age-specific refusals also considering a blank returned questionnaire to be a refusal

respondent sample did not respond because we were unable to reach them (mail forwarding and address correction being unsuccessful in these individuals). From 28 completed telephone interviews of non-respondent twins, most declared their intent to return the questionnaire, and there were no predominant reasons identified for refusals.

We concluded that, over-and-above the 5% who would have been removed by NCOA or mail forwarding in this group, a further 15% of those sent a questionnaire may not have received it, and would not have been discovered in subsequent postal returns. After adjusting the number actually receiving the questionnaire according to these estimates, we

https://doi.org/10.1375/twin.4.4.242 Published online by Cambridge University Press

arrived at revised response rates of 41.5% and 54.6% for males and females respectively, and 49.6% overall. We have not presented age- or sex-specific estimates of this response calculation because numbers are insufficient to make accurate estimates of age-specific questionnaire loss.

Refusals

Firm refusals were received by telephone or e-mail (n = 967), by comments on returned questionnaires or on returned reminder post-cards (n = 533), for a total of 1,382 refusals (excluding duplicate forms of refusal). The proportion of refusals was higher in females than males over the age of 42, and increased with age in that group, but was negligible in both males and females younger than 43 (Table 3). In addition 1,216 twins sent back a blank questionnaire. Combining these with the direct refusals produced a total number of 2,598 refusals, but did not alter the age or sex distribution of overall refusals (Table 3). Refusal did not differ substantially by race.

Comparison of Census Data, Twin Cohort, and Respondents

Male respondents were on the whole slightly younger than the birth cohort of male twins (Table 4), and due to our inability to locate older female twins, the female respondents were very much younger on average than both the cohort of female twins, and the California-born 1990 resident population of females. Although the respondents were predominantly white, Latino respondents were slightly over-represented, and accordingly whites and African-Americans were slightly under-represented. This difference was more striking among females (Table 4). Around 3% of respondents did not give a race, so misclassification could have accounted for any of these observations.

Male respondents were substantially more likely to be married than the male census population, but in older females (over age 53 and under 72 years) the marriage rate among respondents was lower than in the census (not shown).

Table 4

Demographic Comparison of Census-derived Figures for California-born 1990 Resident Population, the California Twin Birth Cohort, and Respondents to Our Study, (Percentages Are the Percentage in Each Age/Sex Group) by Age, Race, Education and Occupational Groups

		Male Census	Male Twin birth cohort	Male response	Female Census	Female Twin birth cohort	Female response
Birth year	Age (1990)						
1908–17	73–82	4.6%	5.1%	2.2%	6.0%	5.1%	0.4%
1918–27	63–72	10.9%	9.6%	7.9%	11.9%	9.9%	2.2%
1928–37	53–62	13.9%	11.7%	12.9%	14.3%	11.9%	3.7%
1938–47	43–52	24.4%	24.4%	26.8%	23.4%	24.7%	23.6%
1948–57	33–42	46.1%	49.2%	50.2%	44.5%	48.5%	70.2%
Race			N/A			N/A	
White		85.7%		84.6%	85.2%		82.1%
Latino		5.7%		6.4%	5.8%		7.4%
Black		3.7%		2.1%	4.2%		3.4%
American Indian		1.3%		1.3%	1.5%		1.1%
Japanese/Chinese		2.9%		1.6%	2.7%		1.7%
Filipino/Thai		0.5%		0.2%	0.5%		0.3%
Other		0.1%		0.8%	0.1%		0.7%
Missing		0.0%		3.0%	0.0%		3.3%
Education			N/A			N/A	
Under 12 yrs		13.3%		24.6%	14.3%		21.5%
12 yrs		22.9%		31.2%	28.1%		35.1%
over 12 yrs		63.8%		44.2%	57.6%		43.5%
Occupation			N/A			N/A	
Managerial, professional and specialty		28.3%		34.9%	25.3%		34.8%
Technical, sales and admin. Support		19.0%		12.6%	33.4%		25.2%
Service		6.8%		4.0%	9.3%		5.5%
Farming, forestry and fishing		3.4%		1.4%	0.8%		0.3%
Precision product, craft and repair		17.3%		11.2%	1.8%		0.4%
Operators, fabricators and laborers		14.0%		13.2%	4.1%		2.7%
Other and unspecified		11.2%		22.7%	25.4%		31.2%

https://doi.org/10.1375/twin.4.4.242 Published online by Cambridge University Press

Respondents were less likely to have completed more than 12 years of education than the comparative population of California-born 1990 residents, more so in males than in females (Table 4) and the disparity was greater in those aged 45 years or older (not shown). However, respondents were substantially better educated than the 1% sample of all US residents (for less than 12 years education, 12 years education and greater than 12 years education respectively, males: 38.9%, 23.0%, 38.0%; females: 37.3%, 26.7%, 36.0%), and better educated than the California-born US residents (for less than 12 years education, 12 years education and greater than 12 years education respectively, males: 41.3%, 18.2%, 40.5%; females: 39.0%, 19.9%, 41.1%).

All counties each with more than 5% of the population residing in them had nearly identical distributions of respondents and California-born 1990 resident population (Table 5). In only 2 counties (Sutter and Napa) did the response account for less than half the expected number. These 2 counties account for only 1.1% of the California population.

Table 5

Geographical Comparison of the Respondents and Census-derived Figures for California-born 1990 Resident Population

Geographical location (county)	Census population		California twin cohort	
	Number	Percent	Number	Percent
Alameda	10,163	5.1%	1,792	5.3%
Sacramento	9,014	4.5%	1,536	4.5%
San Diego	12,071	6.0%	2,224	6.5%
Santa Clara	10,539	5.2%	1,611	4.7%
Fresno	6,052	3.0%	837	2.5%
Orange	14,164	7.1%	2,861	8.4%
Los Angeles	43,548	21.7%	7,184	21.1%
Ventura	4,354	2.2%	938	2.8%
Riverside	6,648	3.3%	1,158	3.4%
San Bernardino	8,341	4.2%	1,534	4.5%
All others	76,150	37.7%	12,413	36.3%
Total	201,044	100%	34,088	100%

Differences Between Double- and Single-Respondent Twin Pairs

For 53.3% (n = 10,296 individuals) of male respondents and 59.4% of females (n = 13,084 individuals), both members of the twin pair returned the questionnaire. A substantially larger proportion of females than males belonged to double-respondent pairs in each age group, but for both males and females, older twins, except the very oldest, were more likely to respond in tandem than younger twins were (not shown).

Zygosity and Sex of Respondents Compared to the Original Cohort

Female-female pairs were slightly more prevalent among respondents than in the birth cohort of California twins (Table 6). Of 13,090 individual respondents from MZ twin pairs, slightly more than half (52.4%) were female. While we have no way of knowing if the respondents' zygosity is representative of that of all native born Californian twins, the distribution of pairs by gender is similar to that of the birth cohort (Table 6). The proportion of MZ males and females in the respondent cohort was lower than the proportion of estimated live births, and subsequently the proportion of DZ twins of both genders, and DZ twin pairs of mixed gender, were higher than the proportion of estimated live births (Table 6). In total we received responses from at least one member

Table 7

Twin Pairs Represented by Current Respondents

Zygosity	Double respondents[1]	Single respondents[1]	Total pairs
MZ (male)	1,765	2,752	4,518
MZ (female)	2,351	2,220	4,570
DZ (male-male)	1,779	3,857	5,636
DZ (female-female)	2,301	2,968	5,269
DZ (male-female)	3,035	6,406	9,441
Unknown	185	332	517
Total	11,416	18,534	29,951

[1] 'double' respondents are pairs where both twins have sent back a questionnaire, 'single' respondents are pairs where only one twin has sent back a questionnaire.

Table 6

Paired Gender Comparisons Between Birth Record of California Twins and Respondents, Including Distribution of Zygosity Among Respondents

	California record of multiple births	Estimated proportion of twins live born in Pacific US[1]		Respondents							
Gender of pair	%	MZ %	DZ %	MZ No.	%	DZ No.	%	unknown No.	%	Total	%
Male	34.5%	18.0%	15.1%	6,230	15.1%	7,353	17.8%	358	0.9%	13,940	33.7%
Female	34.9%	18.5%	15.8%	6,861	16.6%	7,505	18.1%	446	1.1%	14,812	35.8%
Mixed	30.6%	—	32.6%	—	—	12,369	29.9%	—	—	12,369	29.9%
Unknown	—	—	—	—	—	—	—	246	0.6%	246	0.6%
Total	100.0%	36.5%	63.5%	13,090	31.6%	27,227	65.8%	1,050	2.5%	41,367	100.0%

[1] Taken from Mack et al., 2000.

https://doi.org/10.1375/twin.4.4.242 Published online by Cambridge University Press

of 29,951 pairs of twins, almost a third of whom are MZ pairs (Table 7).

Discussion

We present here the development of a large population-based study of twins recruited in a novel manner that continues to be expanded and followed up. Such a resource can be of use in three ways. First, it is a population-based cohort of native Californian residents that can be used for cross-sectional analyses of multiple health-related conditions and exposures — it should be noted that the respondent cohort is currently the largest available population-based cohort of any kind in California. For these analyses to be valid, the respondents must be representative of native born Californians. Secondly, we can consider twins as paired individuals who in some cases share a common genome (MZ twins), or share on average half their genome (DZ twins), and at the very least share a majority of childhood exposures potentially pertinent to the etiology of disease. For these twins we already have a comprehensive set of exposure histories, and self-reports of disease that can later be verified by planned linkage to cancer registries and death indices. Used in this case for traditional twin studies comparing the disease or death concordance in MZ versus DZ twins, the key is whether or not MZ and DZ twins are selected in an unbiased (i.e., representative) fashion and with equal probability. Finally, the respondent cohort is a useful source for nested exposure-control or case-control studies selecting twins efficiently on the basis of self-reported exposure discordance. Again the key is in the non-selective ascertainment of the individuals in the cohort, but in contrast to such studies among singletons, the comparability of the control is assured.

Many twin cohorts are based on heavily biased samples of volunteers (Easton et al., 1992) or special enrollment populations (Friedman & Lewis, 1978; Henderson et al., 1990; Kendler et al., 1992). However, large twin cohorts in Scandinavia are arguably as population-based as one could hope for, developed by the linkage of comprehensive population registers of birth and socialized medical coverage (Cederlof & Lorich, 1978; Hauge et al., 1968; Kaprio, 1994; Kaprio et al., 1990; Kaprio et al., 1978; Kringlen, 1978; Kyvik et al., 1996). Our analyses demonstrate that this twin cohort is representative of the California born population with a few exceptions. This representativeness makes the cohort a useful one for cross-sectional analyses, but our ability to characterize the respondents in comparison to the population from which they were drawn also allows us to describe potential sources of selection bias in nested case-control studies. The DMV linkage process generated a young, non-migrating group of potential questionnaire respondents who, if older and female, were less likely to be married. Out-of-date addressing information from the DMV, however, favored the exclusion of younger twins who did not necessarily migrate out-of-state, but who were mobile enough to have invalid DMV addressing despite recent DMV contact (the youngest group must have supplied address information to the DMV in the past few years). These features of the respondent cohort are relevant to the generalizability of cross-sectional analyses, such as prevalence of health care utilization or cigarette smoking. However, we were able to well characterize the differences between respondents and the population from which they were drawn, so we will likewise be able to adjust such prevalence estimates.

The opportunity for selection bias to affect any study nested within the cohort can be crudely assessed by the overall response rate, and then characterized by describing the response among demographic sub-groups of the respondents. Among those people successfully located, we experienced a response rate in excess of that reported for most other population-based studies, at least 42.2%, and probably as high as 49.6% based on our sub-study of likely non-receipt of the questionnaire. This may be because we appealed to subjects as twins, and twins recognize their value as research subjects. We also employed multiple mailings and postcard reminders to increase response. The refusals we received were rare and their distribution tended to indicate that most non-response was due to apathy, particularly among males, rather than to an objection to our methods. The higher response rates among older women compensated somewhat for our inability to find contact details for them by DMV linkage. The respondent females may therefore represent a non-random (unmarried or more liberal, for example) segment of the population — certainly females in the older groups were the most active in their firm refusal. Our respondents could thus be argued to be more likely a motivated group, which will be important when assessing selection bias for exposures such as physical activity, health behaviors (screening etc) and attitudes and knowledge about health.

While the respondent group may have been motivated, they certainly did not appear better educated than the general population. The disparity between the educational status of the respondents and the census population of California native current residents is notable. Both sources defined education as the completion of various grade levels, and while it is possible that "completion" could be misconstrued as 'attending' versus 'completing' a particular grade level (eg attended 12th grade versus obtaining a high school diploma), such misclassification would not generate the observed disparity confined to higher level education. Nor would the small proportion of respondents failing to provide a response to the education question — even if they were all in the most educated group, there would still be a disparity. Better-educated individuals may be among those migrating out-of-state or migrating frequently within the state and therefore not locatable, or they may have less time to complete questionnaires. This is contrary to expectation — we would have predicted that the better educated were more likely to respond, but perhaps they simply do not take the time to.

There was a higher response among Latinos, which might explain part of the educational disparity if their educational experience differed from the remainder of the cohort. When we looked at race-specific educational attainment (not shown), we found that the Latino respondents were less educated than both the White respondents and the census Latino population. However, among Whites there was still a substantial difference in educational attain-

https://doi.org/10.1375/twin.4.4.242 Published online by Cambridge University Press

ment (also not shown). Our respondents had education levels intermediate between those of California native current residents, and those California natives living elsewhere in the United States. The educational attainment of the respondent cohort is more similar to that of all California natives (twins and singletons) than to those currently resident in California.

The roster obtained by successful DMV linkage, and the respondent cohort were both remarkably uniformly distributed over all counties of California, giving us no reason to assume that there were local disparities in response rates, or that we preferentially selected any geographically discrete groups in our study. Given that the distribution of race, socioeconomic status and exposure to environmental insult is determined largely by geographical location in California, this is further evidence of the population-based nature of the respondent group, and of its value for both prevalence surveys and nested studies, although the sample sizes for racial groups other than White are small.

The key concern regarding the usefulness of any twin population for classical studies of the role of genetics in disease is the extent to which MZ and DZ twins are equally representative, and if not, how they differ with respect to concordance under the null hypothesis (Hawkes, 1997; Martin et al., 1997). While respondents were more likely to be female, the paired gender of twin respondents was not substantially different from that of the birth cohort, tending to indicate that we did not substantially preferentially recruit pairs in which both twins were female. Our comparison of the respondent cohort with the proportions of zygosity-gender pairs estimated to occur among live births in the Pacific States provides a more accurate estimate of the ascertainment of zygosity types — we have slightly under-ascertained MZ pairs. Using Weinberg's rule (Lykken et al., 1987) we would estimate that the number of DZ unlike sex twin pairs is equal to the total number of like-sex DZ pairs, and the number of MZ twins is thus the number of like-sex pairs minus the number of unlike-sex pairs, and we would have estimated that 38% of this respondent cohort should be MZ. By either count, the respondent cohort is slightly under-represented by MZ twins. Possible reasons for this under-representation, beyond possibility of detection in this study, are a higher mortality rate among MZ pairs, and a higher rate of out-of-state migration at an early age of both members of the pair, resulting in our inability to identify them using DMV records.

Overall however, this group is a representative sample of California twins with respect to zygosity, and future comparisons of the rate of disease among MZ versus DZ twins in this group seem unlikely to be biased by enrollment. Contrast this with our International Twin Study, a cohort of twins recruited from newspaper advertisements soliciting twins with chronic disease, in which MZ twins with disease were substantially over-ascertained (Mack et al., 2000). Concordance, in addition to zygosity, may determine compliance (Mack et al., 2000), and it remains to be seen whether twins concordant for particular exposures or conditions were preferentially recruited here.

In summary, we have developed a large representative sample of the population of twins born in California. The ways in which this respondent cohort varies from the population with respect to age, sex, race, education, occupation, are now known and can be used to adjust the results of studies choosing to use this population simply as a cohort of native Californians. Likewise we can accurately estimate the role of selection bias in studies using these twins as subjects, and determine the extent to which cohort members followed for disease outcomes in future are likely to have been differentially ascertained with respect to zygosity. We are currently preparing papers using the respondent cohort to investigate population-based risk factors for smoking uptake and cessation, population-based estimates of physical activity levels and characteristics, and risk factors for mole size and frequency in California. We are conducting classic twin analyses of the risks for mammographic density among female twins, and a case-control study of the role of cigarette smoking in the development of cytokines in identical twins discordant for smoking.

We can identify subsets of twins for further study requiring re-contact (such as the collection of DNA samples), identified on the basis of either discordance for exposure (again derived from questionnaire data already at hand) or subsequent disease. This group will clearly provide an excellent resource for studies of the genetic basis of cancer etiology after we have linked it to the records of the California Cancer Registry. In addition to being able to contrast risk factors between almost 30,000 MZ and DZ twin pairs, we will be able to determine the role of under-ascertainment of respondents by disease status, since we can compare the original birth record to Cancer Registry and mortality records. We can then comment on the effect of ascertainment on disease concordance and subsequent MZ/DZ comparisons as we have done elsewhere (Mack et al., 2000).

We will continue enrollment in the coming years, expanding the respondent cohort by contacting twins born between 1973 and 1982, previous non-respondents providing a new and valid address to the DMV at renewal time, and by re-contact of other previous non-respondents.

Acknowledgements

Initial project management was conducted by Rich Pinder, and the California Twin Program website was designed by Nick Fox. Jennifer Nedrud, Misha Birch, Saundra McGee and Travis Alexander conducted daily activities including receiving phone calls from twins and processing the enormous volume of mail required for this study. This study was funded by grants from the California Tobacco-related Disease Research Program (8RT-0107H and 6RT-0354H). We also thank the many twins who participated.

References

Cederlof, R., & Lorich, U. (1978). The Swedish twin register. In W. E. Nance, G. Allen & P. Parisi (Eds.), *Twin research* (pp. 189–195). New York: Alan R. Liss.

Easton, D. F., Cox, G. M., Macdonald, A. M., & Ponder, B. A. (1992). The study of nevi in British twins: Study design and description of the data set. *Cytogenetics & Cell Genetics, 59*(2–3), 165–166.

https://doi.org/10.1375/twin.4.4.242 Published online by Cambridge University Press

Friedman, G. D., & Lewis, A. M. (1978). The Kaiser-Permanente twin registry. In W. E. Nance, G. Allen & P. Parisi (Eds.), *Twin research: Part B, biology and epidemiology* (pp. 173–177). New York: Alan R. Liss.

Hamilton, A. S., & Mack, T. M. (2000). Use of twins as mutual proxy respondents for each other in a case-control study of breast cancer: Effect of item non-response and misclassification. *American Journal of Epidemiology, 152,* 1093–1103.

Hauge, M., Harvald, B., Fischer, M., Gotlieb-Jensen, K., Juel-Nielsen, N., Raebild, I., Shapiro, R., & Videbech, T. (1968). The Danish twin register. *Acta Geneticae Medicae et Gemellologiae, 17*(2), 315–332.

Hawkes, C. H. (1997). Twin studies in medicine — what do they tell us? *QJM, 90*(5), 311–321.

Henderson, W. G., Eisen, S., Goldberg, J., True, W. R., Barnes, J. E., & Vitek, M. E. (1990). The Vietnam era twin registry: A resource for medical research. *Public Health Reports, 105*(4), 368–373.

Kaprio, J. (1994). Lessons from twin studies in Finland. *Annals of Medicine, 26*(3), 135–139.

Kaprio, J., Koskenvuo, M., & Rose, R. J. (1990). Population-based twin registries: Illustrative applications in genetic epidemiology and behavioral genetics from the Finnish twin cohort study. *Acta Geneticae Medicae et Gemellologiae, 39*(4), 427–439.

Kaprio, J., Sarna, S., Koskenvuo, M., & Rantasalo, I. (1978). The Finnish twin registry: Formation and compilation, questionnaire study, zygosity determination procedures, and research program. *Progress in Clinical & Biological Research, 24*(B), 179–184.

Kasriel, J., & Eaves, L. (1976). The zygosity of twins: Further evidence on the agreement between diagnosis by blood groups and written questionnaires. *Journal of Biosocial Science, 8*(3), 263–266.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C., & Eaves, L. J. (1992). A population-based twin study of major depression in women. The impact of varying definitions of illness. *Archives of General Psychiatry, 49*(4), 257–266.

Kringlen, E. (1978). Norwegian twin registers. In W. E. Nance, G. Allen & P. Parisi (Eds.), *Twin research* (pp. 185–187). New York: Alan R. Liss.

Kyvik, K. O., Christensen, K., Skytthe, A., Harvald, B., & Holm, N. V. (1996). The Danish twin register. *Danish Medical Bulletin, 43*(5), 467–470.

Lykken, D. T., McGue, M., & Tellegen, A. (1987). Recruitment bias in twin research: The rule of two-thirds reconsidered. *Behavior Genetics, 17*(4), 343–362.

Mack, T. M., Deapen, D., & Hamilton, A. S. (2000). Representativeness of a roster of volunteer North American twins with chronic disease. *Twin Research, 3*(1), 33–42.

Martin, N., Boomsma, D., & Machin, G. (1997). A twin-pronged attack on complex traits. *Nature Genetics, 17*(4), 387–392.

Ruggles, S., & Sobek, M. (1997). *Integrated public use microdata series: Version 2.0.* Minneapolis: Historical Census Projects, University of Minnesota. Available: <http://www.ipums.umn.edu/usa/>.

https://doi.org/10.1375/twin.4.4.242 Published online by Cambridge University Press