Identification of Twin Pairs From Large Population-Based Samples

Dinand Webbink,[1] Jaap Roeleveld,[2] and Peter M. Visscher[3]

[1] CPB Netherlands Bureau for Economic Policy Analysis, The Hague, the Netherlands
[2] SCO-Kohnstamm Instituut, University of Amsterdam, Amsterdam, the Netherlands
[3] Queensland Institute of Medical Research, Brisbane, Australia

The basis of most twin studies is the ascertainment of twins, often through twin registries, and determination of zygosity. The current rate of twin births in many industrialized countries implies that in the near future around 3% or more of individuals will be a twin. Hence, there are and will be a lot of twins around and many of those will not participate in twin studies. However, if large population-based samples are available that include appropriate identifiers, then twins can be detected and twin studies performed, even in the absence of zygosity information. We quantified the number of twin pairs that could be detected from a longitudinal survey in the Netherlands, which aims to answer questions about educational strategies and performance in primary education in the Netherlands. We detected 2865 twin pairs if we used a coded name identifier, date of birth, school, grade and year of survey, which is 2.01% of 284,945 pupils in five cohorts. Relaxing our selection criteria increased the number of apparent twin pairs identified, most of which are false positives due to chance matching of identification criteria. We show that the intraclass correlation on measured phenotypes can be used as a quality control measure for twin identification, and quantify the proportion of false negatives (true twin pairs not identified) due to missing data and data coding errors. We compared our estimated rate of twins in the sample to census data and estimate that with our most stringent selection criteria we detect more than 80% of all twin pairs in the sample. We conclude that the identification of twin pairs from large population-based samples is feasible, rapid and accurate if the appropriate identifiers are available, and that twin pairs from such sources are a valuable resource for studies to answer scientific question about twins versus nontwins and about genetic and environmental factors of twin resemblance.

Ever since Francis Galton (1822–1911) recognized their potential for research in human biology, twins have been used in behavioral, psychological and medical studies. Apart from studies that contrast characteristics of twin pairs against singleton-born individuals, that is, those that are interested in differences between being a twin versus a nontwin, most other research is concerned with the separation of genetic and environmental factors that determine observed family resemblance. As we now know (and Galton guessed), the nature–nurture separation relies on the existence of monozygotic (MZ) and dizygotic (DZ) twin pairs and an accurate assignment of individual pairs into the two zygosity classes. With the establishment of twin registries in many countries in the second half of the 20th century, the standard procedure for twin research is by recruitment and ascertainment of twin pairs through such registries. Zygosity is determined either through a set of specific questions about how alike the twins are or, increasingly so, through the use of DNA markers. Maintaining twin registries and, in particular, obtaining good quality phenotypes and tissue samples for genotyping, is expensive.

However, there are additional resources available in many countries that allow twin studies, namely large population-based samples conducted in education research, epidemiology and the social sciences. If twin pairs can be identified from such samples, as well as the sex of the individuals, then genetic studies can be carried out even in the absence of zygosity information (Benyamin et al., 2005; Scarr-Salapatek, 1971). The information to separate genetic and environmental causes of similarity in such samples comes from the knowledge that opposite-sex (OS) pairs are per definition DZ, whereas same-sex (SS) pairs are a known mixture of MZ and DZ pairs. Genetic studies using population-based samples have the potential advantages of large samples sizes and no ascertainment bias with respect to being a twin. The disadvantages are the uncertainty that an identified pair of individuals are twins, the absence of zygosity information for SS pairs, the absence of blood samples (and therefore the impossibility of DNA marker-based genetics research), and the lack of (quality) control of the phenotypes. Depending on the scientific question asked, in some

Received 24 March, 2006; accepted 4 May, 2006.

Address for correspondence: Dr Dinand Webbink, CPB Netherlands Bureau for Economic Policy Analysis, PO Box 80510, 2508 GM, The Hague, the Netherlands. E-mail: H.D.Webbink@cpb.nl

cases the advantages, in particular with respect to sample size, may outweigh the disadvantages.

To draw inference about differences between twins and nontwins or differences in similarity between OS and SS pairs from large population-based samples it is important that twin pairs are identified accurately. The aim of this study was to use a large longitudinal sample of school children with educational performance in the Netherlands to quantify the parameters that determine the identification of twin pairs. In particular we aim to estimate the proportion of all twin pairs that are not identified (the false negative rate) and the proportion of assigned twin pairs that are not twins (the false positive rate).

Material and Methods

Data

Data were available from the longitudinal PRIMA survey in the Netherlands, which aims to answer questions about educational strategies and performance in primary education in the Netherlands (Driessen et al., 1994; Driessen et al., 2004). We used the first five waves of the PRIMA survey including data on pupils, parents, teachers and schools. Each wave had approximately 60,000 pupils. The waves are: 1994, 1996, 1998, 2000 and 2002. The PRIMA project targets a panel of schools (approximately 600 schools yearly). Within each school, pupils in Grades 2, 4, 6 and 8 (average age: 6, 8, 10, 12 years) are tested in languages and maths. In addition, information on the social background is collected and teachers are asked about the behavior of the child in school. The subsequent waves of the project focus on the same sample of schools. If schools no longer wish to participate, they are replaced by schools with a comparable composition of pupils. With this setup the project intends to collect data on the same children on their way through primary education. However, pupils drop out from the project if they repeat a grade, move to another school or if the school stops participating. The first five waves contained 192,102 unique pupils.

Identification of Twins

Twins were identified by matching on family name, date of birth, school and year of the survey. Because of privacy reasons we did not obtain the family name but a coded name identifier. This variable replaced names with identifiers separately for each school, grade and year of the survey. Hence, if two pupils within a certain grade of a certain school and year have the same family name then they are assigned the same identifier. If there are two pupils with exactly the same values on these matching variables they are considered to be twins. In the analysis we focused on twins and did not consider matches of three or more pupils.

As other population-based datasets may not have name identifiers, we investigated the results of a matching procedure without family name. We used all available background variables for the matching procedure. These variables were (i) a weight factor used for the financing of primary schools based on the education and ethnicity of the parents, (ii) the education and country of birth of mother and father, (iii) the household composition, (iv) whether the father or mother spoke Dutch at home and (v) the proportion of time that the pupil lived in the Netherlands.

We calculated intraclass correlations for test in languages, arithmetic and IQ for twin pairs based on two matching procedures. The IQ test has two components, 'composition of figures' and 'exclusion'. For the psychometric properties of the IQ test, see Driessen et al. (1994). The first procedure, which we call 'correct', uses name, date of birth, school and year of the survey. The second procedure, which we call 'incorrect', uses all variables without name. For the second procedure we calculated intraclass correlations for all identified pairs and for pairs that were only identified in the second procedure. We expect that the latter pairs were not really twins.

We compared and contrasted the estimated proportion of twins in the sample with census data on twin births for the same age groups (CBS, 2005). For this comparison we do not use the total sample of PRIMA because of an oversampling of disadvantaged students but focus on the so-called reference sample which is representative of the population. The reference sample is approximately two thirds of the total sample and has 35,000 to 40,000 pupils in each survey.

The longitudinal nature of the data was used to estimate the proportion of data-coding errors. In addition, for a small subset of the data, 1127 pupils of 10 schools participating in the survey, we obtained the full (uncoded) names of the individuals, and used these for validation purposes.

Results

Table 1 compares the results of a number of matching procedures with the results of the matching which uses the name identifier. We find 5730 twins (2865 pairs) if we use name, date of birth, school, grade and year of survey, which is 2.01% of 284,945 pupils in all five waves. This is shown in the first row of Table 1. The second row shows the share and number of twin pairs if we only use date of birth, school and year of the survey as matching variables. The third row shows the proportion and number of twins if we add weight factor to the matching procedure. The last rows show the results if we exploit the longitudinal character of the data. The columns on the right of Table 1 show some measures of the accuracy of the procedure. Column 3 shows which proportion of the identified pairs are really twins. Column 4 shows the proportion of twin pairs in the sample that is identified with this procedure.

Table 1 shows that the most extended matching procedure identifies 25% more twins than the procedure with the name identifier. This procedure makes two types of mistakes: not identifying twins, and identifying twins who are not twins. The share of not identified

Table 1

Share of Twins Using Different Matching Variables for Identification

Matching variables	(1) % identified as twin	(2) Number of pairs	(3) Correct twin/ Identified as twin	(4) Identified correct twins/ Total twins in sample
Name, date of birth, school, year	2.01	2865	100%	100%
Date of birth, school, year	7.10	10,109	26.8%	94.5%
+ weight factor	5.20	7413	36.3%	94.0%
+ socioeconomic status	3.78	5383	49.2%	92.5%
+ country of birth father/mother	3.36	4787	54.8%	91.5%
+ education father/mother	2.71	3855	66.8%	90.0%
+ household composition	2.65	3781	67.9%	89.6%
+ father/mother speaks Dutch at home	2.58	3676	68.2%	87.6%
+ time in the Netherlands	2.55	3634	68.6%	87.1%
Using longitudinal data				
Name, date of birth, school, grade, year	2.10	2995	100%	100%

Table 2

Intraclass Correlations of Twin Pairs Using Two Matching Procedures

	Identifier	Grade 2	Grade 4	Grade 6	Grade 8
Language	Correct	.578	.496	.554	.527
	Incorrect (1)	.540	.495	.450	.476
	Incorrect (2)	.461	.482	.209	.334
Arithmetic	Correct	.641	.617	.569	.656
	Incorrect (1)	.581	.548	.472	.560
	Incorrect (2)	.470	.404	.217	.350
IQ	Correct		.392	.497	.518
	Incorrect (1)		.365	.354	.418
	Incorrect (2)		.281	.044	.171

Note: Incorrect (1) using all matching variables except family name. Incorrect (2) only uses the pairs that are not identified with the correct matching procedure.

twins (Column 4) increases with adding more matching variables. This is most likely explained by coding errors and missing values. Twins will not be identified if variables are coded incorrectly or have missing values for one of the twins of a twin pair.

The longitudinal aspect of the data can be used in the construction of the twin variable. Pupils identified as twins in a certain wave and not identified as twins in another wave can be considered to be twins. Applying this correction increases the share of twins from the basis procedure (using family name) to 2.1% (5990 twins, 2995 pairs). If we apply this correction to the procedure without name, we have to make an assumption about which twin identification is correct.

Table 2 shows the intraclass correlations for twin pairs for test in languages, arithmetic and IQ after applying different matching procedures. We note that this correlation reflects a mixture of underlying correlations for MZ and DZ twin pairs and correlations between unrelated pairs of individuals that are incorrectly assigned as being a twin pair. The IQ test was not taken in Grade 2. The results show that the correlations are always lower for the incorrect matching procedure. The lowest correlations are found for pairs identified with the incorrect procedure and not identified with the correct procedure. This finding is in line with our expectation that the latter group are nonrelated individuals. The substantial intraclass correlation of paired nontwins presumably reflects the effects of common environmental factors, in particular school, education level of parents and socioeconomic status.

In Table 3 we compare twin rates from birth statistics with estimates from our data. Twinning rates in the Netherlands in previous decades have been reported previously (e.g., Orlebeke et al., 1991). The second column shows the population statistics, the third column shows the comparable estimates from our data. We focus on pupils in Grade 2 because, unlike in higher grades, grade repeating is not an issue from Grade 1 to Grade 2. Pupils in Grade 2 of the 1994/1995 survey were born in 1988/1989. Our estimate of the proportion of individuals that are a twin from the PRIMA data is lower than the population estimate. Assuming that the census data are correct it appears that we identified between 83% to 100% of the twins in the population.

Table 3

Twin Rates from Birth Statistics and From Identification in PRIMA Grade 2

Born in	% twin in population	% twin in PRIMA	Difference	PRIMA/ Population
1988/1989	2.42	2.01	0.41	83.1 %
1990/1991	2.74	2.36	0.38	86.1 %
1992/1993	2.81	2.80	0.01	99.6 %
1994/1995	2.94	2.48	0.46	84.4 %

Note: Source: Statistics Netherlands and PRIMA (Driessen et al., 2004; CBS, 2005)

The proportion of missing values was 0.022 on date of birth and 0.023 on family name (results not shown in tables). If the proportion of twins in the population is 2.5% and missing values occur randomly in our dataset, then this translates into 2.5*(0.022 + 0.023)*2 = 0.225% fewer twins identified. Hence, missing values could account for an underestimation of the proportion of individuals that are a twin by 0.225/2.5 = 9%.

For the coding errors we first compared the date of birth of Grade 2 pupils with the date of birth of Grade 4 pupils in the next wave. For 0.5% of the pupils that participated in Grade 2 and Grade 4 of two subsequent waves we find different dates of birth. A second step for assessing coding errors is based on an investigation of a subsample of pupils (1127 individuals) for which we obtained the full names. Matching on name, date of birth, school and grade resulted in 22 twins, that is, 1.95% of the sample. Matching on name, school and grade resulted in 92 twins, or 8.2%. In a check on the 35 additional pairs in the procedure without date of birth we find four pairs with the same name, school and grade that differ by only one digit in their date of birth. If one of these four pairs really is a twin then this would increase the proportion of twins by 0.18%. In addition, we looked at coding errors in names. Matching on school, grade and date of birth but without name resulted in 88 twins, that is 7.81%. In a check on the 33 pairs additionally identified in the second procedure we found no coding errors. We found in only one case comparable names that differed by two letters. Table 4 summarizes the findings on missing values and coding errors. For the coding errors we separately show the findings for the two approaches.

A third reason for not identifying twins is that they do not attend the same school (van Leeuwen et al., 2005). These authors reported that 857 out of a total of 7595 twin pairs (11.3%) had at least one of the twins at age 7 with a disease or handicap or receiving special education, but it was not reported if these twins were at separate schools. A more accurate estimate of the proportion of twins that may be at different schools cannot be obtained from the data presented in that study on the remaining 6738 pairs, because no distinction was made between twins being in a different class at the same school and being in different schools. Our data do not allow us to investigate the importance of not identifying twins because they attend different schools.

For the pupils in Grade 2 of the reference sample we estimated the proportion OS twins as 0.76% and the share of SS twins as 1.67%. Hence, the ratio of OS pairs is 0.76/2.43 = 0.31, consistent with a proportion of 62% of DZ and 38% of MZ twins. The standard error of the estimate of the proportion of DZ (or MZ) twins is approximately $2\sqrt{(0.38*0.62/477)} = 0.044$. The proportion of twins that were DZ born in the period 1988 to 1995 in the Netherlands was 0.6645 (*SE* = 0.003; Imaizumi 2003). Hence, the proportion of MZs in our sample appears to be slightly larger than that in the population but the difference is not statistically significant.

Table 4

Errors in Matching Procedure and Their Effect on the Proportion of Identified Twins

	Date of birth	Name	Total	% twins not identified
Missing values	2.2	2.3	4.5	0.225
Coding error (1)	0.5	0.0	0.5	0.045
Coding error (2)				0.18

Discussion

We have quantified the accuracy of twin identification from a large population-based sample. Extracting twin pairs from a database using matching criteria is not new (Deary et al., 2005; Goldberg et al., 1993; Scarr-Salapatek, 1971). In the Nordic countries, twin registries have been built using record linkage methods keyed to a unique personal identification number issued at birth (see references in Goldberg et al., 1993). But, to our knowledge there are no previous reports of extracting and validating school-aged twins from longitudinal studies. Including a name identifier we estimate that more than 83% of twin pairs in the population are identified correctly. When we exploit the longitudinal nature of the data, a total of 5990 individuals, that is, 2995 twin pairs, are identified, for whom we have longitudinal data on educational performance and IQ scores. This is a formidable resource for twin studies, including research into the effect of being a twin on cognitive ability and research to estimate genetic and nongenetic components of variance. A twin study on the genetics of IQ in children and educational performance of 12-year-olds from the same (Dutch) population consisted of 691 twin pairs with educational achievement and 209 pairs with detailed IQ phenotypes (Bartels et al., 2002).

We have shown that a large part of the difference between our estimate of the proportion of twins and the census data can be attributed to missing values and coding errors on the matching variables name and date of birth. The accuracy of data entry will vary from study to study, so that the difference in ascertainment proportion that we observe cannot be extrapolated to other samples. With an increasing amount of computerization of administration databases at schools, one would predict that the discrepancies that we observe would diminish over time.

Without a name identifier, the number of identified pairs appears to be overestimated. This is not really surprising, since the probability of matching all criteria by chance is not small. For large schools, for example, those having 100 pupils in a particular grade, the

probability of one or more pairs of unrelated individuals sharing a date of birth is substantial. If we assume that dates of birth are uniformly distributed throughout the year, then the expected number of pairs of individuals that share the same birth date is (99 + 98 + … + 1)/365 = 13.56. Hence, by chance we would 'detect' a proportion of 27.12% twins in the sample even if there were none. With a group size of 50, the expected number of pairs is 3.36, or 6.71% being an apparent twin (which is comparable to the share in row 4 of Table 1).

In conclusion, it is relatively easy to identify OS and SS twin pairs from large cohort studies, as long as name, date of birth and sex identifiers are available. In combination with measured phenotypes, such samples, which are unbiased with respect to being a twin, can be used to answer scientific questions about twins versus nontwins, and about the differences in similarity between OS and SS twin pairs. The proportion of identified twins can be compared to census data as a quality control measure. Correlations of phenotypes for (apparent) twin pairs can also be used to calibrate the stringency of twin assignment. If there are too few selection criteria then too many apparent twin pairs will be identified.

Acknowledgments

We thank Dorret Boomsma for many useful comments and suggestions and the Dutch Organization for Scientific Research for allowing us to use a subsample with the full names. We thank the Boys (Theo Arentze, Han Kamphuis, Gert de Lange, Fred Reinders, Joan Wassink, Jan Webbink and Jan Zomer) for inspiration and many thought-provoking discussions.

References

Bartels, M., Rietveld, M. J., Van Baal, G. C., & Boomsma, D. I. (2002). Heritability of educational achievement in 12-year-olds and the overlap with cognitive ability. *Twin Research, 5,* 544–553.

Benyamin, B., Wilson, V., Whalley, L. J., Visscher, P. M., & Deary, I. J. (2005). Large, consistent estimates of the heritability of cognitive ability in two entire populations of 11-year-old twins from Scottish mental surveys of 1932 and 1947. *Behavior Genetics, 35,* 525–534.

CBS. (2005) Statline, Birth Statistics. Retrieved March, 2006, from statline.cbs.nl

Deary, I. J., Pattie, A., Wilson, V., & Whalley, L. J. (2005). The cognitive cost of being a twin: Two whole-population surveys. *Twin Research and Human Genetics, 8,* 376–383.

Driessen, G., Van Langen, A., & Oudenhoven, X. (1994). *De toetsen voor de cohort Primair onderwijs, verantwoording* [The tests for the Cohort Primary Education, justification]. Nijmegen: ITS.

Driessen, G., Van Langen, A., & Vierke, H. (2004). *Basisrapportage PRIMA-cohortonderzoek, Vijfde meting 2002-2003* [Report on PRIMA-longitudinal research project, Survey 2002-2003]. Nijmegen: ITS.

Goldberg, J., Henderson, W. G., Eisen, S. A., True, W., Ramakrishnan, V., Lyons, M. J., & Tsuang, M. T. (1993). A strategy for assembling samples of adult twin pairs in the United States. *Statistics in Medicine, 12,* 1693–1702.

Imaizumi, Y. (2003). A comparative study of zygotic twinning and triplet rates in eight countries, 1972–1999. *Journal of Biosocial Science, 35,* 287–302.

Orlebeke, J. F., Eriksson, A. W., Boomsma, D. I., Vlietinck, R., Tas, F. J., & de Geus, E. C. (1991). Changes in the DZ unlike/like sex ratio in The Netherlands. *Acta Geneticae Medicae et Gemellologiae, 40,* 319–323.

Scarr-Salapatek, S. (1971). Race, social class, and IQ. *Science, 174,* 1285–1295.

van Leeuwen, M., van den Berg, S. M., van Beijsterveldt, T. C., & Boomsma, D. I. (2005). Effects of twin separation in primary school. *Twin Research and Human Genetics, 8*, 384–391.