Design and Implementation of a Twin-Family Database for Behavior Genetics and Genomics Studies

Dorret I. Boomsma,[1] Gonneke Willemsen,[1] Jacqueline M. Vink,[1] Meike Bartels,[1] Paul Groot,[1] Jouke Jan Hottenga,[1] C. E. M. Toos van Beijsterveldt,[1] Therese Stroet,[1] Rob van Dijk,[2] Rien Wertheim,[2] Marco Visser,[2] and Frank van der Kleij[2]

[1] *Biological Psychology, VU University Amsterdam, Amsterdam, the Netherlands*

[2] *Furore Amsterdam, Amsterdam, the Netherlands*

In this article we describe the design and implementation of a database for extended twin families. The database does not focus on probands or on index twins, as this approach becomes problematic when larger multigenerational families are included, when more than one set of multiples is present within a family, or when families turn out to be part of a larger pedigree. Instead, we present an alternative approach that uses a highly flexible notion of persons and relations. The relations among the subjects in the database have a one-to-many structure, are user-definable and extendible and support arbitrarily complicated pedigrees. Some additional characteristics of the database are highlighted, such as the storage of historical data, predefined expressions for advanced queries, output facilities for individuals and relations among individuals and an easy-to-use multi-step wizard for contacting participants. This solution presents a flexible approach to accommodate pedigrees of arbitrary size, multiple biological and nonbiological relationships among participants and dynamic changes in these relations that occur over time, which can be implemented for any type of multigenerational family study.

Advances in behavioral and epidemiological genetic research, the increasing size and complexity of genetic studies, both in terms of increasing sample size and high dimensionality of phenotypic and genotypic data, and the need to integrate data in large scale collaborative studies, necessitate the introduction of new database systems. In this article we focus on the design and implementation of a database system that contains the pedigree information of biologically related and unrelated individuals. The system, called Panter (Person Administration of the Netherlands Twin Register) was designed for studies of extended twin families which are registered with the Netherlands Twin Register (NTR; Boomsma et al., 2006), but is applicable to any type of data collection that consists of clustered observations.

Many large scale twin registers that started out with collections of twin pairs and higher-order multiples are rapidly expanding into registers that include the parents of twins, their siblings, their spouses and offspring. Parents and siblings of twins add value to molecular genetic projects (Martin et al., 1997), and provide information on processes of social interaction and cultural transmission among family members (e.g., Eaves, 1977; van Leeuwen et al., 2008). Spouses of twins give information with respect to processes of assortment (e.g., Reynolds et al., 1996; van Grootheest et al., 2008). The offspring of twins design is very powerful in studying intergenerational associations between environmental variables and outcomes in children (e.g., Magnus et al., 1985). Different participants in a (twin) register may take part in research projects with different 'roles'. For example, parents of young twins may take part as informants on the behavior and development of their offspring, or they may be phenotyped for traits of interest themselves and their data included in the analyses of genotype-phenotype relations. In addition, data on twins or other family members may be collected from informants who are not related to the individuals in the study. For example, projects that assess behavior in young twins often make use of teacher ratings (e.g., Derks et al., 2006; Simonoff et al., 1998). Many studies are not limited to collecting cross-sectional phenotypic, or genotypic, data but employ prospective designs. Longitudinal studies should keep track of changes in the addresses of participants, for example, to study how neighborhood characteristics influence mental and physical health, of changing relations among

Received 24 March 2008 ; accepted 1 April, 2008.

Address for correspondence: Dorret Boomsma, Biological Psychology, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, the Netherlands. E-mail : dorret@psy.vu.nl

participants in a pedigree and of changes in willingness to participate in particular research projects.

Relationships among participants in a register may be dynamic and can change over time, as do other characteristics of participants. Biological relations do not change, but social relationships can change quite frequently. In the case of relationships such as teacher-student, the relationship simply ceases to exist without any further complications for additional relationships. In the case of a separation between two spouses and a new marriage for one or both individuals, additional family relations may be introduced, such as stepparents, unrelated siblings who grow up together, half-siblings, or stepchildren. Clear definitions of these relationships and the changes that may occur over time is essential for genetic studies, but is also of importance in studying the effects of altering family situations on health and behavior.

Besides the challenge of keeping track of the social and biological family ties, researchers may be confronted with additional challenges. For instance, a father may be willing to complete questionnaires on the behavior of his children, but may not want to provide information on his own behavior. This implies that participants should be able to indicate for which type of relationship they want to be approached and that selection on the level of relationships is possible. Database systems need to be able to cope with these new developments. Systems must link the longitudinal phenotypic information and keep track of a large number and wide range of social and biological relationships between participants that can change over time.

Database activities can be distinguished into administrative processes and scientific applications. Administrative processes include importing new individuals and families or adding new family members, address management (tracing twins and families who have moved), documenting the participation status of individuals (moved, not willing to participate, ill, deceased), and storing information on questionnaire mailings, responses to mailings and reminders, approaching nonresponders and approaching subjects for other than survey projects, such as biobank studies.

Importantly, any system that keeps track of personal information needs to adhere to guidelines concerning privacy. Identifying information, such as name, date of birth and address, needs to be kept separately from the phenotypic information collected. While this may sound simple, the actual implementation may be more complex. For instance, when sending parents the questionnaires which they need to complete about more than one child (e.g., their twin children and additional siblings of the twins), it needs to be clear to the parents which questionnaire they need to complete for which child.

We choose to support administrative and scientific work by different database-systems (see Figure 1), which each work with separate anonymous IDs. The Administrative database with person and family relation data includes names, date of birth, addresses, relation information among participants, information on willingness to participate in different types of research projects and information on administration of research projects, questionnaires and surveys (e.g., which subjects have been approached, which subjects responded, and which subjects

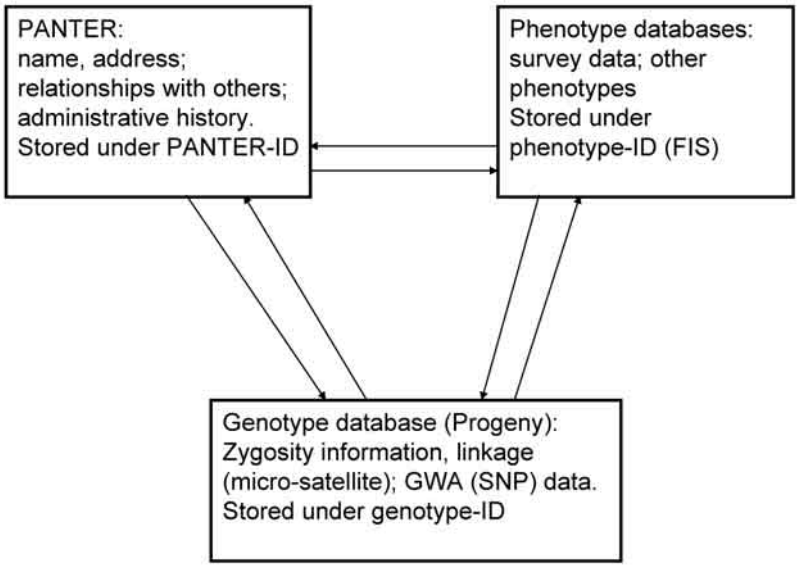

Figure 1

Overview of the databases used for administrative and scientific work. Please note that the arrows indicate data exchange patterns, but that the databases are not linked directly.

https://doi.org/10.1375/twin.11.3.342 Published online by Cambridge University Press

received reminders). The Phenotype database has information regarding longitudinal phenotyping and the Genotype database contains marker information (microsatellite markers, SNPs). In this article we describe the Administrative database.

The Administrative database was designed specifically to support research involving many different relationships among persons. It consists of a graphical user interface (GUI) to facilitate input, output and management of person data, and a relational database that can be accessed directly from external programs such as SPSS. The software was written in .NET 2.0 and based on modern n-tier architecture (Fowler, 2003) with separate Presentation, Domain, Service and Datasource Layers. Such an n-tier architecture increases maintainability, supports scalability of performance and users and is perfectly suited for easy adaptation to new situations (for example changing database vendor or adding a web interface). Design patterns (Fowler, 1996) are used wherever applicable to further enhance the maintainability. To ensure the long-term stability, unit tests have been implemented which will trap most of the errors caused by software changes in later years (see www.nunit.org). For this particular project, a high code coverage of the unit tests of 70% was reached. The database used is Microsoft SQL Server 2005, but can be replaced by a different database type as nHibernate (Bauer & King, 2005) to access the data.

Persons and Relations

Previous database systems have often been proband or twin-centric, in the sense that twins were placed at the heart of the web of possible relations. Although this makes it relatively easy to define first- and second-degree relationships, it becomes more difficult when pedigrees broaden, when multiple pedigrees are linked, or when subjects become part of the register for more than one reason (e.g., they are the teachers of a twin but are also a twin themselves). In the Administrative database discussed here, the concept of

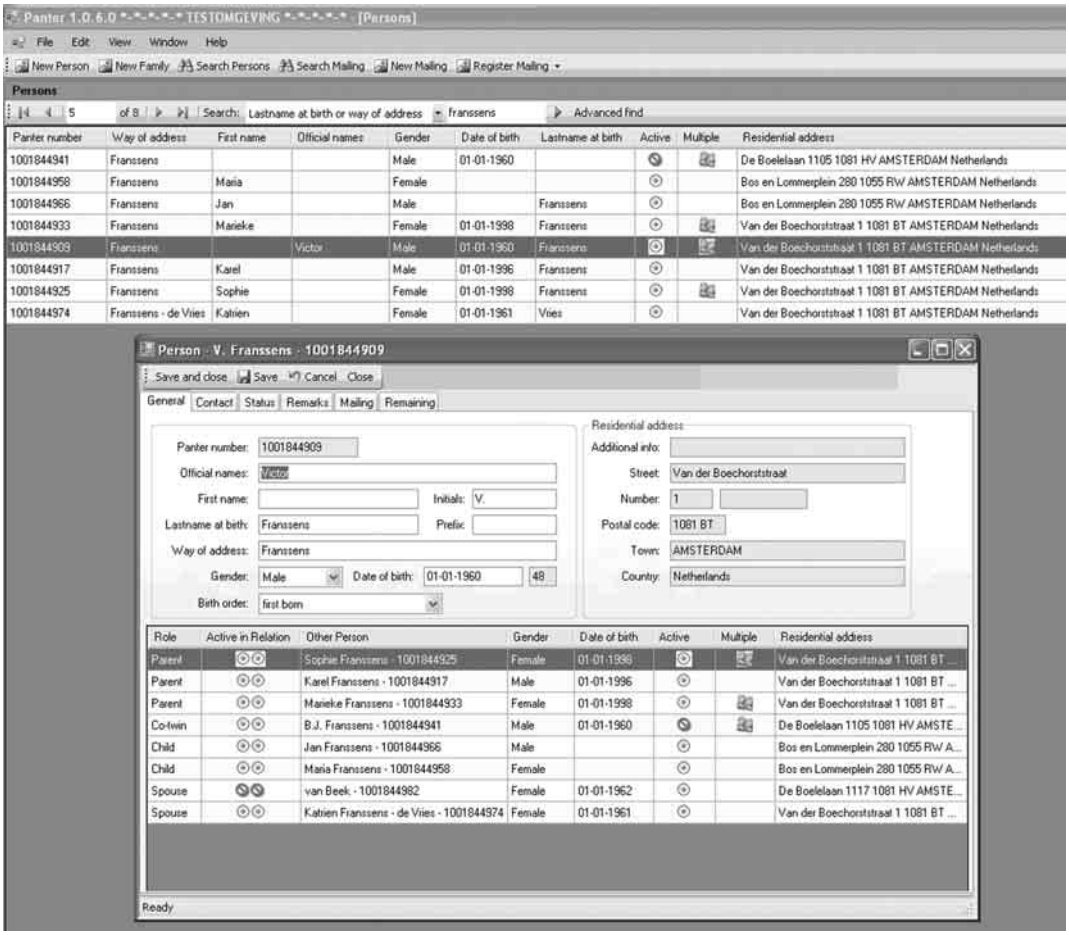

Figure 2

The man selected in the top screen (Victor) is an active participant in the study and is part of a twin pair. The bottom screen indicates he has several roles; he is also a father of twins (Sophie and Marieke) and has a singleton son (Karel). He is the child of Maria and Jan who are both active participants, but his co-twin is at present not active in the study. His current spouse is Katrien and he also has a former spouse (van Beek) who is still active in the study.

Note: Data presented here are fictional.

https://doi.org/10.1375/twin.11.3.342 Published online by Cambridge University Press

a proband/twin focus has been abandoned and replaced with the generic notion of persons and relations. Persons are stored in the database using common fields such as name and date of birth. A separate table contains the relations between persons, which is a one-to-many structure. The relations themselves have one or more characteristics that define the type of relation: twin/co-twin, student/teacher, child/parent, and so forth. These characteristics are user-definable and extendible. Each relationship between two individuals is thus identified. Figure 2 illustrates this setup. In a table the relationship between Victor and his parents is identified (Victor is a child of Maria; Victor is a child of Jan). Victor is the spouse of Katrien and Victor is co-twin to Berend Jan and brother to Karel. These relationships are all stored in one table that is addressed when creating pedigree output. In this table the relationship between Victor's parents is also defined (Maria is the spouse of Jan). In this way, all siblings connected through the same father and mother may be identified but two pedigrees may also be linked through a spousal relationship. Siblings may also be linked directly, without the need to add nonparticipating parents to the database, which may be preferable in an administrative database. For genetic analytic purposes, 'dummy' parents can be created when transferring the relationship data to a standard pedigree file. Similarly, in this way teacher/child relations may be defined, which would not be possible in a traditional pedigree set-up consisting of individual ID plus father and mother ID.

This setup is completely flexible and applicable to any research project that involves clustered observations. Figure 3 shows how these concepts are related using UML notation (see www.uml.org). Pedigree data can be accumulated using the defined relations and exported as a comma-separated file that can be imported into other applications. Users may employ different definitions of a pedigree (e.g., include or exclude nonbiological relatives or spouses), and the current program can output both smaller and more extended pedigrees. A small pedigree structure may be defined through all first-degree relationships of an individual, consisting of the person's parents, children, and possible co-twin and siblings. A larger, extended pedigree may exist of two pedigrees linked by marriage. If a user wants to identify these larger pedigrees, the current programming within the database clusters all individuals who have a biological or nonbiological relationship with each other. For example, in the case of an extended pedigree, the brother of the parent of a twin will be included in the output pedigree, as will be the brother of the spouse of an individual. Other relationships such as teacher-pupil can also be

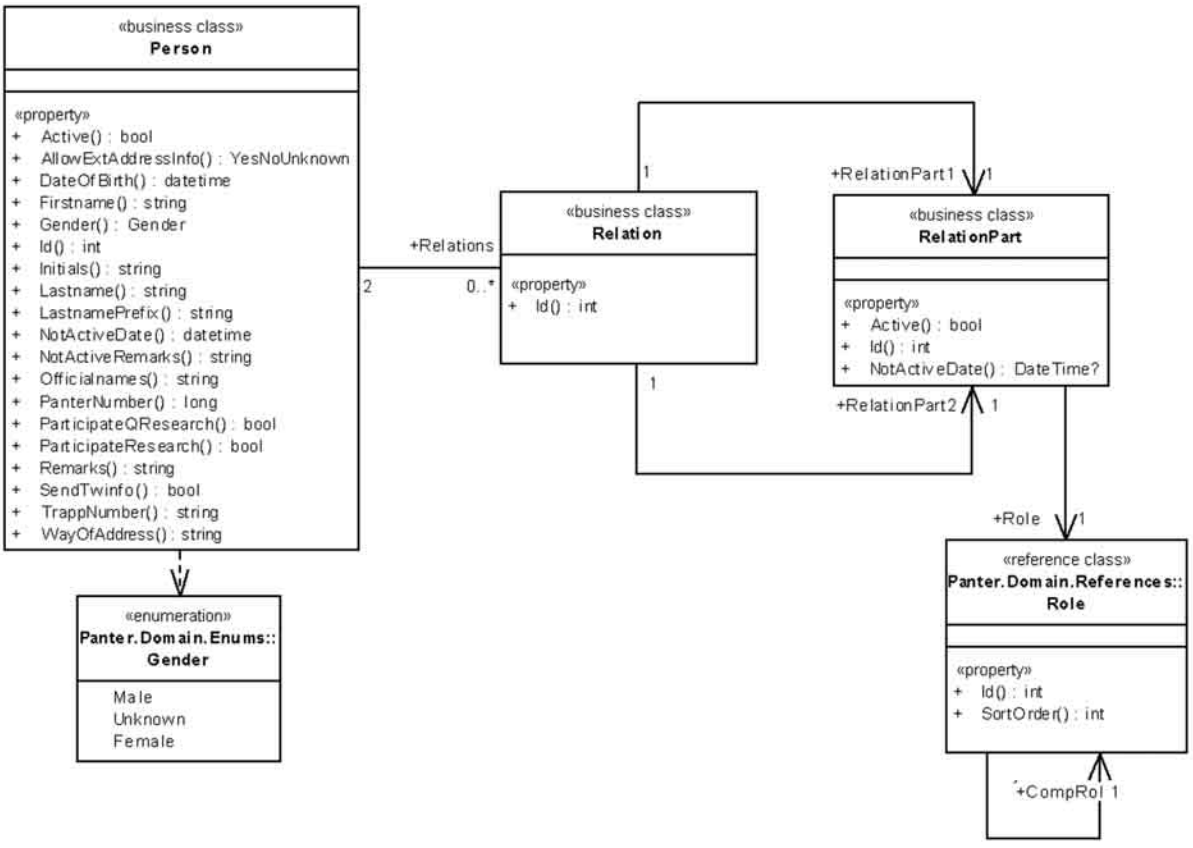

Figure 3

Class diagram of the relationship between persons, relations, and roles (UML notation).

https://doi.org/10.1375/twin.11.3.342 Published online by Cambridge University Press

incorporated, but these relationships generally are not relevant for genetic studies.

Data Entry, Alterations and Integrity

The user interface can be used for adding persons to the database who may have relationships with persons already existing in the database, or for adding 'standard' twin families consisting of a father, mother and a twin pair. In case of the addition of a large number of individuals, a bulk import can take place through comma-separated files.

Demographic data can be easily entered and managed, and include names (first names, given names, surnames), address, phone number/email, participation status and the reason for (temporary) nonparticipation, mailing history and additional notes. Addresses can be postal as well as residential, and can also be defined at the relationship level. A teacher who is part of a twin pair himself can thus receive personal correspondence at home, while any mailings concerning the twins in his class could be sent to the school address. In addition, foreign addresses are also supported. When adding new data such as persons, families or addresses, it is important to avoid creating duplicate records that differ only by a small amount of data but actually refer to the same entity. For this purpose the fields of the new records are compared with existing records and if the difference does not reach a threshold, the user is warned that a duplicate record may be inserted. When adding address data, a table containing national zip codes can be used to ensure integrity of the names of streets entered. Also, in the case of altering the address of a person, the program checks whether more individuals live at the old address and asks whether the other persons found at the address should also be moved.

Historical Data

When a move of address is received and subsequently processed, a new record is inserted in the database with a date of alteration. The old address is retained in the database, together with the same alteration date. In this way, the history of addresses a person has lived at is stored in the database and can be used in longitudinal studies. Historical data are also retained in the case of altered relationships. In case a spousal relationship is ended, an end date is included in the database and the relation is set to inactive, indicating that the relationship has become that of ex-spouses. In this way, both spouses and ex-spouses may still be included in future studies.

Options for Searching

Various capabilities for searching assist the user to find particular persons. The most simple search interface allows searching for persons based on simple characteristics such as name, zip code and date of birth. For more complicated searches, any database field can be used with expressions such as 'greater than' or 'starts with'. The most advanced search option uses predefined expressions, which support queries to find for example 'spouses of twins' or 'sisters who both have multiples'. Additional predefined expressions that define search criteria may be programmed. It is also possible to base a search on IDs contained in an input file. All these search options can be combined to narrow down the list of persons returned. The results are shown onscreen and can optionally be written to a file.

Wizard Mailings

A mailing wizard is available that provides a step-by-step user interface to create and register mailings that can be sent by traditional mail (using MS Word to create address labels) or by e-mail through transfer to a different database system. Using the wizard, a selection of subjects can be made. It is possible to indicate whether these subjects are the addressees (who complete the questionnaire for themselves) or whether other persons, for example the parents, are the addressees (who complete a questionnaire about their child). In the final step of the mailing wizard, a mailing output is created including information on both the addressee and the subject, allowing the user to create labels for envelopes (containing the name and address of the addressee) and questionnaires (containing the phenotype IDs of the subject(s)). When a certain type of mailing is sent frequently a template can be made, ensuring that each time all settings for the mailing are entered in the same way. When a questionnaire is returned, receipt can be done manually, by using an import file or a barcode-scanner. To optimize response rates, each mailing has the options 'resend' and 'reminder'. The first option is used to resend the mailing to participants for whom the mail was undeliverable but who subsequently have been located. The second option allows for sending a reminder to a selection of participants based on mailing status and date range.

Output Options

Panter has several built-in options for data export to be used in other applications. Files with demographic information can be generated for selected samples for mailings or to check whether surveys were completed by the intended person. More importantly, Panter can output the relation information. All possible relationships among individuals can be exported to a file, for example, person 1 is father of person 3, person 1 is co-twin of person 5. A pedigree number is assigned to all family members within one pedigree. As stated previously, the definition of the pedigree is flexible. Users may specify small nuclear families (one person with parents, additional co-twin/siblings, and offspring) or more extended pedigrees (all individuals linked through biological or nonbiological relationships). Using a separate SQL module, the pedigree file can further be made suitable for use in genetic linkage analyses. This algorithm is able to deal with family loops and ensures that all individuals are included only once. In

https://doi.org/10.1375/twin.11.3.342 Published online by Cambridge University Press

this module the selected relations are limited to parents, twins, siblings and spouses. In case of a sibling relation without parents, dummy parents are created by the module just for the output. Spouses are included, but can easily be removed from the file. With the addition of information on sex, the whole output is reorganized into a comma-delimited file that can subsequently be used for the pedigree import module of the Progeny software. This import format is a derivate of the standard linkage format (Ott, 1999) and can easily be converted to this standard format. Pedigree numbers are generated and assigned at the moment of export, and are based on grouping the individuals that are linked to each other with the given relations. Typed genetic markers are later used to confirm these relations such as the zygosity of twins.

Privacy and Security

Several measures have been taken to ensure the privacy of individuals and the security of the system as a whole. Users must first log on to the system before they can perform the actions they are authorized to, according to their login credentials. The authorization model is currently based on the Windows Active Directory Group the user belongs to, but can easily be replaced by a database-driven authorization or even a custom implementation. There are different levels of authorization, varying from view-only permissions up to administrator level with full permissions including deletion of persons from the database. The system runs on a separate network that is not physically connected with any of the other university networks, and is therefore safe from intruder attempts from the Internet. Any data exchange required has to take place using memory sticks or similar media. For the sake of authorization, the separate network runs its own Domain Controller that contains the user information.

To further ensure the privacy of individuals, the administrative database works with different IDs to those used in the databases containing the phenotype and genotype data. The different IDs for a particular individual are never stored in the same database, but can be calculated from the others using a (pluggable) encryption scheme (requiring the highest level of authorization in Panter). Part of the ID is used as a checksum to ensure data integrity.

Discussion

The database system developed, Panter, has at its heart an extremely flexible handling of persons and relations, which is independent of proband or twin status. Panter has no intrinsic notion of types of relations, for example, biological parents versus stepparents, so a child may have an unlimited number of parents in the database. New relations among participants are easy to add and old relations can be made inactive, still retaining the information on the past existence of this relationship. Selection of subjects can be based on individual characteristics or on relations among participants. Panter was set up to make it easily adaptable for other environments. The solutions built into Panter present a flexible approach to accommodate pedigrees of arbitrary size, multiple biological and nonbiological relationships among participants and dynamic changes in these relations that occur over time, which can be implemented for any type of multigenerational family study. Information on procuring Panter may be obtained from the first author.

Panter has been designed and built to accommodate new developments in genetic epidemiological research. Typical examples of such developments are increasing data collections, both in sample size and dimensionality, and the need to integrate data in collaborative studies (e.g., Muilu et al., 2007). Panter facilitates research by offering a smooth handling of administrative processes and unrestricted options for relations among participants.

Acknowledgment

The work described in this article was made possible by NWO-MagW Investeringsaanvraag 480-04-004.

References

Boomsma, D. I., de Geus, E. J., Vink, J. M., Stubbe, J. H., Distel, M. A., Hottenga, J. J., Posthuma, D., van Beijsterveldt, T. C., Hudziak, J. J., Bartels, M., & Willemsen G. (2006). Netherlands Twin Register: From twins to twin families. *Twin Research and Human Genetics, 9*, 849–857.

Bauer, C., & King G. (2005). *Hibernate in action*. Greenwich, CT: Manning.

Derks, E. M., Hudziak, J. J., Beijsterveldt, C. E. M., Dolan, C.V., & Boomsma, D.I. (2006). Genetic analyses of maternal and teacher ratings on Attention Problems in seven-year-old Dutch twins. *Behavior Genetics, 36*, 833–844.

Eaves, L. J. (1977). Inferring the causes of human variation. *Journal of the Royal Statistical Society, A., 140,* 324–355

Fowler, M. (1996). *Analysis patterns reusable object models*. Reading, MA: Addison Wesley.

Fowler, M. (2003). *Patterns of enterprise application architecture*. Boston, MA:Addison-Wesley.

Magnus, P., Berg, K., & Bjerkedal, T. (1985). No significant difference in birth weight for offspring of weight discordant monozygotic female twins. *Early Human Development*, *12, 55–59*.

Martin, N., Boomsma, D. I., & Machin G. (1997). A twin-pronged attack on complex traits, *Nature Genetics*, *17,* 387–391.

Muilu, J., Peltonen, L., & Litton, J. E. (2007). The federated database: A basis for biobank-based post-genome studies, integrating phenome and genome data from

https://doi.org/10.1375/twin.11.3.342 Published online by Cambridge University Press

600,000 twin pairs in Europe. *European Journal of Human Genetics, 15*, 718–723.

Ott, J. (1999). *Analysis of human genetic linkage* (3rd ed.). Baltimore, MD: The Johns Hopkins University Press.

Reynolds, C. A., Baker, L. A., & Pedersen, N. L. (1996). Models of spouse similarity: Applications to fluid ability measured in twins and their spouses. *Behavior Genetics, 26,* 73–88.

Simonoff, E., Pickles, A., Hervas, A., Silberg, J. L., Rutter, M., & Eaves L. (1998). Genetic influences on childhood hyperactivity: Contrast effects imply parental rating bias, not sibling interaction. *Psychological Medicine, 28,* 825–837.

van Grootheest, D. S., van den Berg, S. M., Cath, D. C., Willemsen, G., & Boomsma, D. I. (2008). Marital resemblance for obsessive-compulsive, anxious and depressive symptoms in a population-based sample. *Psychological Medicine* [Epub ahead of print].

van Leeuwen, M., van den Berg, S., & Boomsma, D. I. (2008). A twin-family study of general IQ. *Learning and Individual Differences*, *18,* 76–88.

https://doi.org/10.1375/twin.11.3.342 Published online by Cambridge University Press