Published in final edited form as:

Genet Epidemiol. 2014 December ; 38(8): 692–698. doi:10.1002/gepi.21855.

Use of an Electronic Medical Record to Create the Marshfield Clinic Twin/Multiple Birth Cohort

John Mayer¹, Terrie Kitchner², Zhan Ye¹, Zhiyi Zhou³, Min He^{1,2}, Steven J. Schrodi², and Scott J. Hebring^{2,*}

¹Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI 54449, USA

²Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI 54449, USA

³Parkland Center for Clinical Innovation, Parkland Health & Hospital System, Dallas, TX 75235, USA

Abstract

Population-based genetic analyses, such as the Genome-Wide Association Study (GWAS), have proven powerful for describing the genetic complexities of common disease in epidemiologic research. However, the significant challenges faced by population-based study designs have resulted in a revitalization of family-based approaches, including twin studies. Twin studies are unique in their ability to ascertain both heritable and environmental contributions to human disease. Several regional and national twin registries have been constructed using a variety of methods to identify potential twins. A significant challenge in constructing these large twin registries includes the substantial resources required to recruit participants, collect phenotypic data, and update the registries as time progresses. Here we describe use of the Marshfield Clinic electronic medical record (EMR) to identify a cohort of 19,226 patients enriched for twins or multiples. This cohort defines the Marshfield Clinic Twin/Multiple Birth Cohort (MCTC). An EMR system provides both a mechanism to identify potential twins and a source of detailed phenotypic data in near real time without the need for patient contact outside of standard medical care. To demonstrate that the MCTC can be used for genetic-based epidemiologic research, concordance rates for muscular dystrophy (MD) and fragile-X syndrome, two highly heritable diseases, were assessed. Observations indicate that both MD and fragile-X syndrome are highly correlated among affected twins in the MCTC ($P \approx 3.7E-6$ and $1.1E-4$, respectively). These findings suggest that EMR systems may not only be an effective resource for predicting families of twins, but can also be rapidly applied to epidemiologic research.

Keywords

twins; heritability; muscular dystrophy; fragile-X syndrome

*Correspondence to: SJ Hebring, Center for Human Genetics, Marshfield Clinic Research Foundation, 1000 N Oak Avenue, Marshfield, WI 54449, USA, hebring.scott@mcrflmfdclin.edu, Phone: 1-715-389-3122.

Introduction

Population-based genetic strategies, including the Genome-Wide Association Study (GWAS) design, have been efficient for demonstrating the genetic complexities of common phenotypes. With over 1,000 GWASs published, nearly 4,000 genetic variants across the human genome have been associated with hundreds of human conditions [Hindorff et al., 2013]. Although the GWAS study design has been a powerful approach for the study of human genetics, significant challenges remain resulting in a recent resurgence of family-based approaches, including twin studies.

Twins represent a unique familial relationship. Monozygotic twins are nearly identical in their genetic makeup whereas dizygotic twins share, on average, half their genetic information. Unique to many twins is a shared exposure to social and environmental factors, starting at conception. As such, populations of twins are used as a gold standard when parsing the genetic and environmental influences of human disease. Multiple national twin registries exist including the United Kingdom Adult [Moayyeri et al., 2013; Spector and Williams, 2006], Australian [Hopper et al., 2013; Martin et al., 1981], Sri Lankan [Sumathipala et al., 2000; Sumathipala et al., 2013], Chinese National [Li et al., 2013; Yang et al., 2002], Danish [Spector and Williams, 2006; Kyvik et al., 1996], and Swedish [Cederlof et al., 1970; Magnusson et al., 2013] Twin Registries. The Swedish and Danish Twin Registries are the largest, containing nearly 200,000 participants, whereas the United Kingdom Adult, Australian, Sri Lankan, and Chinese National Twin Registries range from approximately 12,000 to 70,000 participants. A summary of these registries, along with other twin registries, has been published previously [van Dongen et al., 2012].

In the United States, several regional twin registries exist, including the Minnesota [Krueger and Johnson, 2002; Lykken et al., 1990], University of Washington [Afari et al., 2006; Strachen et al., 2013], Michigan State University [Burt and Klump, 2013; Klump and Burt, 2006] and Mid-Atlantic [Anderson et al., 2002; Lilley and Silberg, 2013] Twin Registries, ranging in size from nearly 14,000 to 50,000 participants. Nation-wide twin registries also exist in the United States, including the Vietnam Era [Eisen et al., 1987; Tsai et al., 2013] and National Academy of Sciences-National Research Council [Page, 2006; Jablon et al., 1967] Twin Registries, but are biased toward male veterans from the Vietnam War and World War II, respectively. Recruitment methods for twin registries vary widely across studies. For example, the Washington Twin Registry utilized Washington State driver's license data and the Sri Lankan Twin Registry leveraged census data. The United Kingdom Twin Registry took advantage of multimedia to recruit twins, whereas other twin registries have leveraged birth records to identify twins. A commonality for many of the twin registries is the use of surveys, questionnaires, and/or interviews to collect phenotypic data. As such, the expense associated with recruiting twins and collecting phenotypic data can be significant. This fact is exacerbated when phenotypic and participant information is updated over time. Alternative approaches that allow for rapid generation of large cross-sectional populations of twins with phenotypic data available in real-time would be ideal and may be achievable with the use of an extensive electronic medical record (EMR) system.

An EMR can serve as a rich data source for demographic and phenotypic information and usually contains longitudinal health histories, including prescription records, family histories, laboratory and imaging test results, physician notes, biometric measurements, procedure codes, and diagnostic codes. In addition, an EMR may include sex, ethnicity, education, occupational exposures, postal address, and account numbers. In Europe, EMR systems have been widely accepted, driven in part by nationalized medicine. However, in 2006 fewer than 10% of American healthcare institutions had fully integrated systems [Smaltz and Berner, 2007]. As a result of the “Affordable Care Act” passed by United States Congress in 2010, efforts to implement EMR systems across American healthcare institutions are underway. Importantly, medical institutions such as the Marshfield Clinic in Wisconsin have been pioneers in the development and application of EMR systems. Marshfield Clinic has used an internally developed EMR system in clinical practice since 1984 with phenotypic data dating back to 1979.

Use of an EMR system for research purposes allows for extraction of rich phenotypic and demographic information related to complex topics. EMR systems across the United States, including the Marshfield Clinic’s EMR, have been applied to genetic epidemiological studies [McCarty et al., 2011]. These EMR-driven genetic studies take advantage of biobanks of presumed unrelated patients. A limiting factor for these genetic studies is the ability to recruit patients and the expense necessary to acquire genetic data. Whereas population-based epidemiological studies have the capacity to identify genetic risk factors, family-based studies, including twin studies, often offer more statistical power. Identification of large populations of twins in an EMR system for genetic research has not yet been demonstrated. Here, we describe the generation of the Marshfield Clinic Twin/Multiple Birth Cohort (MCTC) using data available in the Marshfield Clinic EMR. The MCTC is one of the largest cross sectional populations enriched for twins in the United States and allows for collection of phenotypic information in near real-time without the need for birth records, surveys, or patient contact through a research protocol. As proof-of-principle that MCTC can be used for genetic-based epidemiologic research, disease concordance rates of muscular dystrophy (MD) and fragile-X syndrome, two highly heritable diseases, were quantified using nothing more than EMR data.

Methods

The Marshfield Clinic health system consists of 57 physical clinic locations and two hospitals in Western, Central, and Northern Wisconsin as well as Michigan’s Upper Peninsula. In 2011, Marshfield Clinic treated approximately 400,000 unique patients with nearly 3.8 million patient encounters. As of January 1, 2013, Marshfield Clinic has records from nearly 2.6 million unique patients. Due to Marshfield Clinic’s geographic coverage, serving a very stable population with strong agricultural ties, a significant proportion of Marshfield Clinic patients receive the majority of their healthcare in the Marshfield Clinic system. The Marshfield Clinic was one of the first health care institutions to develop and apply an EMR system. A variety of structured and unstructured health data are available in Marshfield Clinic’s EMR system dating back to 1979. In Marshfield Clinic’s EMR, billing codes, procedure codes, lab values, and physician notes are readily available. For patients last seen prior to 1979, basic demographic information is available in the EMR while paper

charts are also accessible. This EMR system was used to identify families of twins/multiples.

At the basic level, predicted twin/multiple births were identified by those patients who share a last name and date of birth (DOB). These selection criteria are similar to those applied to generate the Vietnam Era Twin Registry [Eisen et al., 1987]. To account for potential changes of last name, especially for women who may have changed last name due to marriage, the first recorded last name was used. A total of 100,284 patients in Marshfield Clinic's EMR system were identified who shared a last name and DOB.

To further refine and assess the MCTC, additional criteria were applied including available health histories, family size, last name filtering, home address, billing accounts, and evidence of relationships through unstructured text data, specifically key word searches for "twin" and "triplet." A pictorial representation of the methods applied can be seen in Figure 1. EMR data extraction was conducted through SAS programming. To validate findings and further refine the methodologies, manual chart reviews were conducted by a trained study coordinator on 600 predicted families. This included assessment of paper charts predating Marshfield Clinic's EMR system. This project has been approved by the Marshfield Clinic Institutional Review Board (HEB10413PM-C).

As proof-of-principle that EMR data can be used to identify families of twins/multiples, and that the MCTC can be used for genetic studies, concordance rates for predicted families with MD or fragile-X syndrome were measured. Patients with evidence of MD or fragile-X syndrome were identified by the International Classification of Disease, version 9 (ICD9) codes that define MD (ICD9 359.1) or fragile-X syndrome (ICD9 759.83). A probability-based test was used to determine if the number of concordant twins observed is more frequent than expected by chance assuming all individuals were unrelated. The P-value was calculated as detailed below.

$$P \cong \sum_{k=x}^N \binom{N}{k} \left(\frac{n}{2N}\right)^{2k} \left[1 - \left(\frac{n}{2N}\right)^2\right]^{N-k}$$

Where x denotes the number of twin pairs concordant for having the disease, N denotes the total number of twin pairs examined, and n is the number of individuals within the 2N total number of individuals that have the disease. A higher than expected concordance rate within a population may indicate a shared genetic component; alternatively, it could be influenced by shared environmental factors.

Results

Twin/Multiple Birth Cohort

From the population of 2.6 million unique Marshfield Clinic patients with records in the EMR, 100,284 patients were identified who shared both a last name and DOB. The largest predicted family of multiples consisted of 16 individuals. These patients all shared a DOB of January 1st in the early 1900s. In fact, the vast majority of unrealistic sized families had

January 1st birthdays at the turn of the 20th century. It is expected that a default DOB may have been assigned to these patients if DOB was unknown. A property shared amongst these large predicted families is the availability of demographic information, but lack of health histories in the EMR. Assuming that many of these large “multiple birth” families were false positives and that only those families with clinical data in the EMR were of interest, the cohort was filtered to include only those with available medical histories in the EMR. This reduced the group to 39,246 patients with no obvious bias towards January 1st DOBs. Under this scenario, the largest predicted family consisted of eight individuals. These individuals all shared a Laotian surname. A significant population of Laotian refugees migrated to central Wisconsin following the Vietnam War in the 1970s and 1980s. Again, it was assumed that DOB was unknown for these patients and was artificially assigned in the EMR. To further reduce the potential for false positives, the maximum number of multiples was restricted to four per family. This restricted the population to 19,062 families consisting of 39,144 patients.

To characterize the group of 39,144 patients, additional selection conditions were applied. Based on Marshfield Clinic’s EMR system, 8,648 individuals shared a home address, 5,904 individuals shared a billing account, and 8,841 individuals, or their predicted family member had the key-word “twin” or “triplet” in unstructured physician notes totaling 14,102 unique patients. This 14,102 patient set is referred to as Subset A. Subset B represents the 25,042 remaining patients without additional evidence of relationships. Home address and billing accounts are captured in the EMR, but longitudinal histories for these data are unavailable due to data replacement. Not surprisingly, many of the patients who shared a billing account overlapped with those that shared a home address (Figure 2). Also not surprisingly, patients who shared a home address or billing account tended to represent a younger proportion of the population while the keyword search identified older patients (Figure 3).

To assess the positive predictive value of using same last name and same DOB, while restricting the patients to those with available health histories and families no larger than four, manual validation of 600 predicted families was performed. Three-hundred predicted families were randomly selected from Subset A. Of the 300 families, 2.7% had insufficient clinical histories to assess. Of the remaining 292 families, 84% were validated as twins or multiples. As expected, individuals who shared two or more attributes including a billing account, home address, and/or key word, increased the accuracy of the prediction (Figure 2). An additional 300 predicted families in Subset B were also randomly selected for manual validation. Of the 300 predicted families in Subset B, 53 (18%) had insufficient clinical histories to assess manually. Of the 247 remaining families, 14 (6%) were validated as twins.

During manual validation, it was noticed that the majority of false positives in Subset B came from common surnames such as Anderson, Smith, and Jones. The distribution of surname frequencies in Subset A and B were highly divergent (data not shown). To assess the effect of common surnames on the accuracy rate, the most common surnames were incrementally removed from Subset B and then reassessed for accuracy based on the remaining validation data. Twin prediction rates were maximized to 25% after the top 286 most common surnames were removed from Subset B leaving 5,124 patients (Figure 4). The

14,102 patients in Subset A and the 5,124 remaining patients in Subset B represent 19,226 patients from 9,473 predicted families; the sum of these subgroups make up the MCTC. When considering the accuracy rates of the different populations, taken together with the proportions those populations represent in the MCTC, it is estimated that 68% of the MCTC are true families of twins and multiples.

The final MCTC contained 253 predicted multiples and 9220 predicted twins including 3179 members of opposite sex pairs, 3247 members of same sex male pairs, and 2794 members of same sex female pairs. The average patient age for all birth types was between 33.2 and 36.2 with comparable years of available EMR data (Table I). The number of predicted families decreased as minimum years of available EMR data increased (Figure 5A). Likewise, as age increased, years of EMR data also increased. It is not until age 18, an age where families tend to separate, where a leveling off of available EMR data was noticeable (Figure 5B).

Twin Concordance Rates

To demonstrate that the MCTC can be used to conduct genetic-based epidemiologic studies, concordance rates for those patients diagnosed with MD (ICD9 359.1) or fragile-X syndrome (ICD9 759.83) were determined. MD is a highly heritable disease with multiple modes of inheritance depending on the genetic etiology. The two most common forms of MD include Duchenne and myotonic MD. Duchenne MD is caused by mutations in *DMD*, typically affects children, and is an X-linked recessive condition. Myotonic MD results from mutations in *DMPK* or *ZNF9*, affects primarily adults, and is an autosomal dominant condition. Fragile-X syndrome is an X-linked recessive condition resulting in an expansion of a trinucleotide repeat upstream of *FMRI*. Fragile-X syndrome is also one of the most common forms of cognitive impairment in males. In the MCTC, 10 presumed twins were diagnosed with MD and two were diagnosed with fragile-X syndrome. The majority of MD patients (60%) and both fragile-X syndrome patients were males, demonstrating that both conditions can be X-linked, as expected. Of importance, four of the 10 MD patients and two of the two fragile-X syndrome patients were predicted twin pairs. All concordant MD and fragile-X patients were manually validated, for diagnosis and relationship. According to the frequency of MD and fragile-X syndrome in the cohort and according to the equation presented above, it is highly improbable that affected patients would be pairs of twins unless there is a strong shared factor ($P \approx 3.7E-6$ and $1.1E-4$, respectively). These significant P -values may be driven by a shared environmental exposure, but based on *a priori* knowledge, they are driven undoubtedly by shared genetic factors.

Discussion

Studies designed to make use of twin registries offer a powerful approach for ascertaining both genetic and environmental contributions to human disease in epidemiologic research [Goodin, 2012]. As such, many twin registries have been developed in the United States and across the globe [van Dongen et al., 2012]. A variety of methods have been applied when recruiting patients into these registries. Regardless of methodology, significant time and resources are required. Resources may be further strained if participant information is updated over time. To address these challenges, the MCTC was developed to identify a

clinic patient population enriched for twins and multiples using an extensive EMR system. The advantages of this approach are significant. An EMR system offers an extremely cost-effective and efficient mechanism to identify patients that may be related. Although others have applied large databases to efficiently identify presumed twins [van Dongen et al., 2012], use of an EMR system provides direct access to patient health information. As such, no patient contact is necessary to collect phenotypic data other than that collected during routine medical care. This is unique because predicted twins can be identified continuously while detailed phenotypic data can be obtained in near real time including disease status, biometric measurements, and lab values. Furthermore, using basic data in an EMR system (Figure 1) may allow this methodology to be translatable to other EMR-based healthcare systems. Although the Marshfield Clinic healthcare system has served nearly 2.6 million patients, far larger healthcare institutions linked to an EMR system exist offering the potential for larger cohorts. Furthermore, with many data types standardized in EMR systems, combining cohorts across regions or countries may be possible.

A significant advantage of current twin registries is the well-documented familial relationships. Without participant contact, identifying true relationships through questionnaires and/or genetic testing can be difficult. Using an EMR system may increase error rates, especially for families containing females where last name may change as a result of marriage prior to entry into an EMR system. Female name changes did not appear to contribute significantly for predicted male-female families; based on manual chart reviews, less than 1% of predicted families were identified as married couples.

Manual chart reviews were conducted to assess error rates. Approximately 84% of the predicted families in MCTC who shared a home address, billing account, and/or were matched by a key word search (Subset A) were validated as twins or multiples (Figure 2). These patients tended to be younger and subsequently were easier to validate. Unfortunately, older patients, individuals who are more valuable due to deeper medical histories, are far more difficult to validate. Home address, billing account, and/or key word searches are less effective as individuals age and families separate causing the home address and account information to be replaced overtime. Regardless, when assessing those predicted families who did not share a home address, billing account, or did not have an indication of being a “twin” or “triplet” from text data (Subset B), 25% were identified as true twins/multiples after filtering to remove common last names. Although the prediction rate of Subset B is not optimal and caution should be made when applying this subset of patients for study, this nevertheless represents a significant enrichment of twins compared to a rate of approximately 3% in the general population [Martin et al., 2013]. When considering accuracy rates in the context of the populations assessed by manual chart review, MCTC as a whole can predict twins with 68% accuracy.

To increase prediction efficiency, other information may be leveraged such as birth records. As it relates to the MCTC, hospital birth records do not offer a significant benefit. Multiple births are not associated with the twins’ medical records, but are linked primarily to the mother’s. As such, only young twins that can be mapped back to maternal records can be assessed using hospital records. Unfortunately, a significant fraction of Marshfield Clinic patients are not born at a system hospital. Using criteria such as home address and billing

accounts serves a similar purpose without the requirement of patients being born at one of the two hospitals linked to the Marshfield Clinic's EMR system. Conversely, cross referencing government/state birth records with EMR systems would be a significant benefit, but this requires birth records to be structured electronically, a property that Wisconsin does not have prior to 1989. Regardless, even if sibling relationships can be ascertained, zygosity status will remain elusive without genetic testing. This limitation may be temporary as genomics becomes incorporated into standard medical practice [Rehm et al., 2013; van El et al., 2013]. Alternatively, zygosity status may be further enriched by capturing blood types in the EMR, which could rule out monozygosity in cases of differing blood types.

Another limitation of an EMR-based approach when developing twin cohorts may be incomplete health histories. This is especially relevant for patients whose birthdays pre-date the EMR system. Incomplete health histories may also pose a challenge for health systems covering transient patient populations or in locations where significant healthcare marketplace competition exists. In the Marshfield Clinic system, this limitation may be mitigated in part by a stable agricultural-based population and an extensive longitudinal EMR system (Figure 5B). This is exemplified by the MCTC containing nearly 4,000 families where all predicted family members had multiple years of clinical follow-up in the EMR; approximately 1,000 predicted families had greater than 15 years of clinical follow-up in the EMR (Figure 5A). Although biased toward younger patients (Figure 3), the MCTC may be a cross-sectional representation of the general population of Northern, Western, and Central Wisconsin and demonstrates that applying an EMR system to predict family relationships provides an efficient method to begin genetic-based epidemiologic studies.

As mentioned previously, twin registries can be extremely valuable for genetic studies, including heritability measurements [Goodin, 2012]. Although heritability cannot be directly measured in this cohort because relatedness at the genetic level is unknown, concordance rates may be measured as a proxy to heritability. The advantage of the MCTC is that the phenotypes that can be assessed are not limited to questionnaires or surveys, a limitation applicable to many of the large twin registries. EMR-based twin populations have the capacity to assess thousands of clinically relevant disease phenotypes [Hebbring, 2013; Hebbring et al., 2013]. As proof-of-principle that the MCTC can be applied to genetic research, this study demonstrates that MD and fragile-X syndrome are enriched for concordant pairs ($P \approx 3.7E-6$ and $1.1E-4$, respectively). Although it is possible that a shared environmental component may be driving these observations, the known genetic contribution to MD and fragile-X syndrome makes this possibility unlikely. Future studies assessing other diseases may compare concordance rates between same-sex and opposite-sex pairs to better separate genetic and environmental factors. As mentioned previously, incorporating blood types may also help to enrich for zygosity status.

In conclusion, this study demonstrates for the first time that an EMR system can offer an efficient mechanism to identify twin/multiple birth relationships and may be used in the study of genetic diseases. As EMR systems become standardized, widely applied, and incorporated with genomics, large and powerful family-based studies, including other family

structures beyond twin families, may be possible when studying the complex genetic and environmental contributions to human disease.

Acknowledgments

This study was funded in part by NCATS grant 9U54TR000021, NCRR 1UL1RR025011, and NLM 1K22LM011938. The authors gratefully acknowledge the support from the Marshfield Clinic Research Foundation. The authors would also like to thank Rachel Stankowski for her assistance in editing this manuscript.

References

- Afari N, Noonan C, Goldberg J, Edwards K, Gadepalli K, Osterman B, et al. University of Washington Twin Registry: construction and characteristics of a community-based twin registry. *Twin Res Hum Genet.* 2006; 9:1023–9. [PubMed: 17254446]
- Anderson LS, Beverly WT, Corey LA, Murelle L. The Mid-Atlantic Twin Registry. *Twin Res.* 2002; 5:449–55. [PubMed: 12537875]
- Burt SA, Klump KL. The Michigan State University Twin Registry (MSUTR): an update. *Twin Res Hum Genet.* 2013; 16:344–50. [PubMed: 23101567]
- Cederlof R, Floderus B, Friberg L. The Swedish twin registry--past and future use. *Acta Genet Med Gemellol (Roma).* 1970; 19:351–4. [PubMed: 5533790]
- Eisen S, True W, Goldberg J, Henderson W, Robinette CD. The Vietnam Era Twin (VET) Registry: method of construction. *Acta Geneticae Med Gemellol (Roma).* 1987; 36:61–6.
- Goodin DS. The genetic and environmental bases of complex human-disease: extending the utility of twin-studies. *PloS One.* 2012; 7:e47875. [PubMed: 23272039]
- Hebbiring SJ. The challenges, advantages and future of phenome-wide association studies. *Immunology.* 2013; 141:157–65. [PubMed: 24147732]
- Hebbiring SJ, Schrod SJ, Ye Z, Zhou Z, Page D, Brilliant MH. A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.* 2013; 14:187–91. [PubMed: 23392276]
- Hindorf LA, Wise A, Junkins HA, Hall PN, Klemm AK, Manolio TA. [accessed 1 February 2014] A catalog of published genome-wide association studies. 2012. www.genome.gov/gwastudies
- Hopper JL, Foley DL, White PA, Pollaers V. Australian Twin Registry: 30 years of progress. *Twin Res Hum Genet.* 2013; 16:34–42. [PubMed: 23200298]
- Jablon S, Neel JV, Gershowitz H, Atkinson GF. The NAS-NRC twin panel: methods of construction of the panel, zygosity diagnosis, and proposed use. *Am J Hum Genet.* 1967; 19:133–61. [PubMed: 6067093]
- Klump KL, Burt SA. The Michigan State University Twin Registry (MSUTR): genetic, environmental and neurobiological influences on behavior across development. *Twin Res Hum Genet.* 2006; 9:971–7. [PubMed: 17254439]
- Krueger RF, Johnson W. The Minnesota Twin Registry: current status and future directions. *Twin Res.* 2002; 5:488–92. [PubMed: 12537882]
- Kyvik KO, Green A, Beck-Nielsen H. The most recent part of the Danish Twin Registry. Establishment and analysis of zygote specific twinning rates. *Ugeskr Laeger.* 1996; 158:3456–60. [PubMed: 8650815]
- Li L, Gao W, Yu C, Lv J, Cao W, Zhan S, et al. The Chinese National Twin Registry: an update. *Twin Res Hum Genet.* 2013; 16:86–90. [PubMed: 23298737]
- Lilley EC, Silberg JL. The Mid-Atlantic Twin Registry, revisited. *Twin Res Hum Genet.* 2013; 16:424–8. [PubMed: 23218199]
- Lykken DT, Bouchard TJ Jr, McGue M, Tellegen A. The Minnesota Twin Family Registry: some initial findings. *Acta Genet Med Gemellol (Roma).* 1990; 39:35–70. [PubMed: 2392892]
- Magnusson PK, Almqvist C, Rahman I, Ganna A, Viktorin A, Walum H, et al. The Swedish Twin Registry: establishment of a biobank and other recent developments. *Twin Res Hum Genet.* 2013; 16:317–29. [PubMed: 23137839]

- Martin JA, Hamilton BE, Ventura SJ, Osterman MJK, Matthews TJ. Division of Vital Statistics. Births: final data for 2011. *Natl Vital Stat Rep*. 2013; 62:1–70.
- Martin NG, Gibson JB, Mathews JD. Australian NHMRC twin registry. *Med J Australia*. 1981; 2:518. [PubMed: 7198714]
- McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*. 2011; 4:13. [PubMed: 21269473]
- Moayyeri A, Hammond CJ, Hart DJ, Spector TD. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res Hum Genet*. 2013; 16:144–9. [PubMed: 23088889]
- Page WF. Update on the NAS-NRC Twin Registry. *Twin Res Hum Genet*. 2006; 9:985–7. [PubMed: 17254441]
- Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, et al. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med*. 2013; 15:733–47. [PubMed: 23887774]
- Smaltz, DA.; Berner, ES. *The Executive's Guide to Electronic Health Records*. Chicago: Health Administration Press; 2007.
- Spector TD, Williams FM. The UK Adult Twin Registry (TwinsUK). *Twin Res Hum Genet*. 2006; 9:899–906. [PubMed: 17254428]
- Strachan E, Hunt C, Afari N, Duncan G, Noonan C, Schur E, et al. University of Washington Twin Registry: poised for the next generation of twin research. *Twin Res Hum Genet*. 2013; 16:455–62. [PubMed: 23218177]
- Sumathipala A, Fernando DJ, Siribaddana SH, Abeysingha MR, Jayasekare RW, Dissanayake VH, et al. Establishing a twin register in Sri Lanka. *Twin Res Hum Genet*. 2000; 3:202–4.
- Sumathipala A, Siribaddana S, Hotopf M, McGuffin P, Glozier N, Ball H, et al. The Sri Lankan Twin Registry: 2012 update. *Twin Res Hum Genet*. 2013; 16:307–12. [PubMed: 23302519]
- Tsai M, Mori AM, Forsberg CW, Waiss N, Sporleder JL, Smith NL, et al. The Vietnam Era Twin Registry: a quarter century of progress. *Twin Res Hum Genet*. 2013; 16:429–36. [PubMed: 23177374]
- van Dongen J, Slagboom PE, Draisma HH, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. *Nat Rev Genet*. 2012; 13:640–53. [PubMed: 22847273]
- van El CG, Cornel MC, Borry P, Hastings RJ, Fellmann F, Hodgson SV, et al. Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. *Eur J Hum Genet*. 2013; 21:S1–5. [PubMed: 23819146]
- Yang H, Li X, Cao W, Lu J, Wang T, Zhan S, et al. Chinese National Twin Registry as a resource for genetic epidemiologic studies of common and complex diseases in China. *Twin Res*. 2002; 5:347–51. [PubMed: 12537857]

Figure 1.

Flow chart describing the filters applied in the construction of the Marshfield Clinic Twin/Multiple Birth Cohort (MCTC).

Figure 2.

Venn diagram illustrating overlap between 8,652 individuals who shared a home address (Add), 5,904 individuals who shared a billing account (Acc), and 8,856 individuals who had the key word “twin” or “triplet” in unstructured text data (KW) in the Marshfield Clinic Twin/Multiple Birth Cohort (MCTC). Also indicated are the estimated accuracy rates for predicting twins in the MCTC based on different criteria according to manual chart review.

Figure 3. Patient distribution by age for the entire Marshfield Clinic Twin/Multiple Birth Cohort (MCTC) (black) and those in MCTC who shared a home address (red), shared a billing account (green), and had the key word “twin” or “triplet” in unstructured text data (blue).

Figure 4.

Indicated are the number of patients (black distribution plot) and accuracy rates (grey line) when top surnames were removed from subset B.

Figure 5.

(A) Number of families that have at least one individual with indicated minimum years of electronic medical record (EMR) data in the Marshfield Clinic Twin/Multiple Birth Cohort (MCTC). (B) Years of EMR data by age for patients in the MCTC. Solid line indicates the mean; dotted lines represent the upper and lower 95% confidence intervals (CI).

Table 1

Demographics of the Marshfield Clinic Twin/Multiple Birth Cohort (MCTC)

Birth Type	Families	Mean Age (years)	Mean EMR History (years)
Twins M-F	3179	34.9	9.2
Twins M-M	3247	36.2	9.8
Twins F-F	2794	33.5	8.6
Triplets and Quadruplets	253	33.2	8.8

Abbreviations: M, male; F, female; EMR, electronic medical record.