Linkable administrative files: Family information and existing data

Leslie L. Roos University of Manitoba, Canada
Leslie_Roos@cpe.umanitoba.ca
Randy Walld University of Manitoba, Canada
Charles Burchill University of Manitoba, Canada
Noralou P. Roos University of Manitoba, Canada
Nathan Nickel University of Manitoba, Canada

(Received February 2016 Revised March 2017) http://dx.doi.org/10.14301/llcs.v8i3.406

Abstract

Linkable administrative data have facilitated research incorporating files from various government departments. Examples from Canada, Australia, and the United Kingdom highlight the possibilities for improving such work. After expanding on comparisons of linkable administrative data with several famous studies, we forward suggestions on improving research design and expanding use of family data. Certain characteristics of administrative data: large numbers of cases, many variables for each individual, and information on parents and their children, provide building blocks for implementing these proposals.

Traditional longitudinal epidemiological approaches can be modified to facilitate a quasi-experimental perspective. Incorporating multiple research designs within the same project handles threats to validity more easily. Family data provide a number of opportunities for both same-generation and intergenerational research. Risk factors associated with a number of conditions can be studied. Bad events can affect all family members, and cross-sectoral information can extend analyses beyond health to include educational outcomes. Parent/child linkages suggest several lines of research exploring within-family relationships.

Complicated data call for family identification systems to estimate project practicality. Manitoba administrative data are presented to illustrate one such system. Problems in maintaining core data element – such as marital status – have been highlighted. The productivity and potential of cross-sectional, longitudinal, and life course research using existing information have emphasised the value of investments in such work.

Keywords

Linkable data; research strategies; information rich settings; international; primary data; longitudinal studies; family studies; small areas

Introduction

Increasingly, scholars around the developed world use record linkage and data routinely collected by administrative bodies for service rather than research purposes to create information-rich environments. These are environments where researchers can leverage linkable, individual-level information on the health, socioeconomic, social, and biological characteristics of virtually the entire population of a jurisdiction to support a wide variety of studies (Brook, Rosman, & Holman, 2008; Jutte, Roos, & Brownell, 2011). Information-rich environments have made it possible to study regional differences in health and health care delivery, specific diseases and interventions, child development, and aging. Policy-relevant issues associated with education, poverty, children taken into out-of-home-care by protective services, and social housing have also been explored (Public-Academic Research Colloquium Leveraging Administrative Data for Social Policy, 2016). Epidemiology and pharmacoepidemiology have been emphasised in pioneering studies using linked registers from Sweden, Denmark, Norway, and Finland (Mortensen, 2013). These information-rich environments are spreading. A recent listing noted 267 data linkage centers in 34 countries (International Population Data Linkage Network, 2016). After supporting earlier studies using linked databases, United Kingdom funders have invested £100 million in four Centers of Excellence in health informatics, four Administrative Data Research Centers focusing on social and economic datasets, and a Clinical Practice Research Datalink (Lyons, Ford, Moore, & Rodgers, 2014). Australia has made major efforts to develop appropriate linkage infrastructure for health research, policy and planning (Smith et al., 2011). Analysts in various states and provinces have stressed the desirability of merging information across different government departments (Cowan, 2015; Williams, McClellan, & Rivlin, 2010). State-level activity in Florida has led to new research on neonatal health and cognitive development (Figlio, Guryan, Karbownik, & Roth, 2014). Recent work with American de-identified tax records has examined neighbourhood effects both on intergenerational mobility and on income and life expectancy (Chetty, Hendren, Katz, & Lawrence, 2016a; Chetty et al., 2016b).

This expansion of research based on administrative data has provided analytic files which, for some applications, equal or outperform the best known primary data collection approaches (Johnson & Schoeni, 2011; Levy & Brink, 2005; Power, Kuh, & Morten, 2013). Record linkage provides a cost-effective means for creating 'wide' data files with many variables capturing information on significant aspects of individual lives across the life course. In addition, researchers using administrative data on the entire population have access to information on individuals difficult to include in routine survey research – e.g., those living in rural/remote regions, those residing in poor neighborhoods, etc. Because an entire population is typically available, a very large number of cases can be analysed, often over long periods of time. For example, by linking birth registers with a wide variety of administrative data, scholars can conduct sibling, twin, and family analyses, sometimes across generations, to study a wide range of health and social issues. Such characteristics of administrative data create possibilities for going considerably beyond the usual observational before/after comparisons.

Many strengths of administrative data are largely underutilised and not well understood. For example, several data linkage conferences highlighting a diverse set of projects included only six clinical trials, five natural and quasi-experiments, and just three sibling analyses out of approximately 700 presentations (Exploiting Existing Data for Health Research International Conference, 2013; International Data Linkage Conference, 2012; International Health Data Linkage Conference, 2014). The number of (primarily observational) projects is encouraging but many additional opportunities could be explored. Following Hand (2016), "sharing and linking behavioral data, both public and private, holds tremendous promise for improved public policy" and advancing medical and social science research.

This paper builds on extensive experience with databases organised in the Canadian province of Manitoba to illustrate the potential of administrative data. Comparisons are made to deal with the following questions: What features of linkable administrative data expand the opportunities for research? What characteristics of these data can facilitate better analysis? How can family information not only improve design but expand the number of opportunities? What are common limitations of administrative data?

Table 1 compares common characteristics of information-rich environments built from linkable administrative data with those of several famous long-term studies based on primary data collection – the Panel Study on Income Dynamics (PSID), the British Birth Cohorts, and the Framingham Study (Johnson & Schoeni, 2011; Levy & Brink, 2005; Power et. al., 2013). As can be seen, administrative data have several strengths relative to longitudinal primary data collection. A central advantage is the ability to cost-effectively update files on a regular (often annual) basis. This automatically facilitates construction of birth cohorts, longitudinal follow-up, growth in the number of individuals available for study, and new information on life events and family relationships. New research opportunities follow such expanded information.

Table 1. Comparing Linkable Administrative Data and Longitudinal Primary Data

Relevant characteristics	Linkable administrative data (Specific sites in Canada, United Kingdom, Australia, Scandinavia, United States)	Longitudinal primary data (Panel Study on Income Dynamics, British Birth Cohorts, Framingham Study)
Number of cases	Often more than one million	Several thousand or smaller
Populations	Often built on registry of an entire population or a specific group (e.g., Medicare enrollees)	Subjects sampled and tracked
Record linkage	Critical to expand scope of information and check data quality	Very useful to expand scope of information
Files typically linked	Health, geographic, housing, child protection, justice, interventions, and other files	Health utilisation, genomic data, and other files
New data and updating	Routine arrival of updates (often annually) Access to new files is negotiated	New data must be collected and merged with existing data
'Nonusers'	Individuals in population, but not in substantive file, often of interest	Such analyses not relevant
Loss to follow-up	Attrition by out of area migration	Attrition by nonresponse
Design and analysis		
Time	Information provided at relatively short intervals (from daily to annually)	Information must be collected (typically annually or at longer intervals)
Longitudinal	Number of years will vary with site; tracking from birth possible	Number of years will vary with site; tracking for many years from birth possible
Place	Often specified to postal code level	Often specified to address
Life events	Information may be available from registry or other sources for different ages	Information collected if part of design; often available for different ages
Interventions and evaluations	Longitudinal data allow double pretest designs and long follow-up	Constrained by *N*; detailed information may be available for specific conditions (Framingham).
Statistical methods	Modeling and family fixed effects analyses can partially compensate for omitted variables	Analyses often use standard approaches; complicated samples sometimes integrated
Sibling identification	Sibling and twin studies facilitated	Sibling studies facilitated if part of design
Quantity of information/individual	Many variables/individual ('wide data') but defined for administrative purposes	Generally fewer variables/individual but may be better targeted
Variables and indices	Using nonstandard variables to create meaningful indices is time-consuming	Defined by researchers; scaling may be relatively easy
Diagnoses	Testing for data quality necessary	Vary considerably with study
Health providers	Provider information to aid measurement	Provider information may be available
Individual follow-up	Before and after an event	Before and after an event
Limitations	Important information may be omitted or available only for a subpopulation	Important information likely to be collectable for entire sample if part of design
Data collection and quality	Data collection and access out of investigators' control; frequent quality checks needed	Investigators control data collection, access, and quality
Family Information		
Spousal identification	Can help to understand determinants of need/demand for aging services	Many analyses constrained by *N*
Constructing families over time	Family data may allow ethnic group identification to facilitate community and geographic comparisons	Many analyses constrained by *N*; special samples can be drawn
Risk factors	Wide range of factors can be analysed	Specific risk factors may be relevant
Bad events	Effects on family members can be studied	Analyses often constrained by N
Child development	Analyses of dyads and triads within families possible	Some dyadic and triadic comparisons possible

Part of Table 1 was presented in Roos et al. (2008).

Organising information

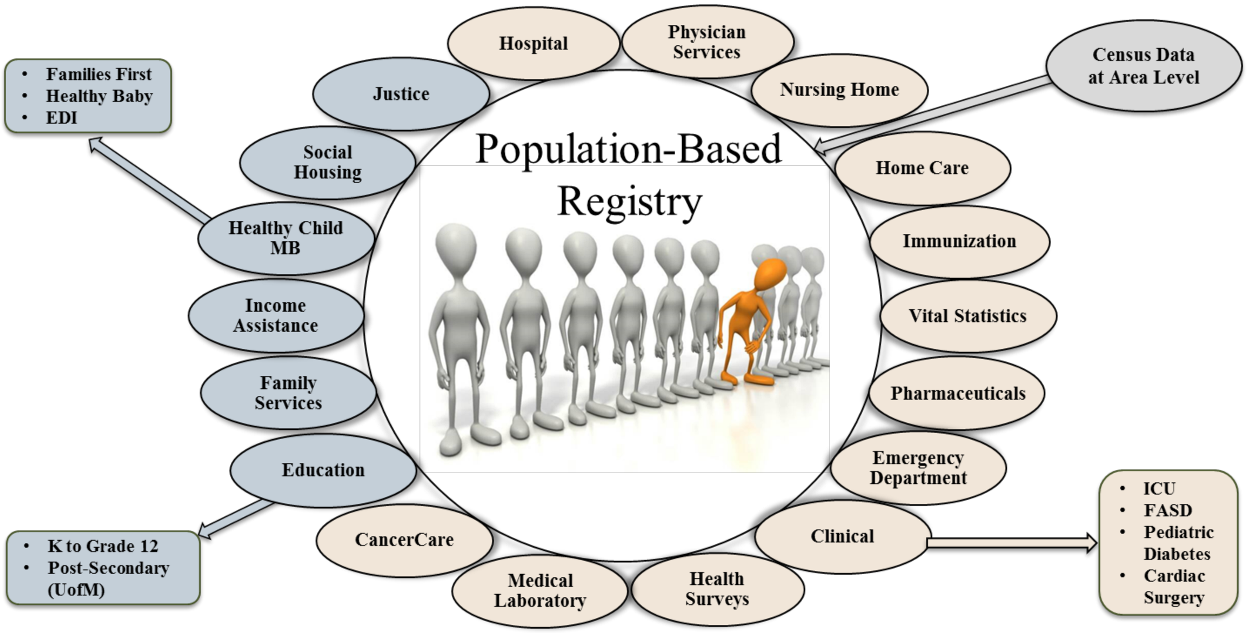
The Whole Population

Because government departments typically collect data on everyone using their systems, researchers working with administrative data can benefit from having registries comprising the entire population (i.e., population-based), whether they are readily available (Canada, Wales) or have to be built more laboriously from multiple sites (Australia) (Ford et al., 2009; Roos & Roos, 2011; Stanley, Glaubert, McKenzie, & O'Donnell, 2011). Information on population-based births, deaths, marital status, migration, and place of residence allows building such registries. A few groups, such as individuals in federal prison or in the military, may not be included. Registries also enable population-based analyses of non-users, providing otherwise unavailable information about who does not receive preventive measures (cervical cancer screening, childhood immunisations, at-risk family screening) or who is not enrolled in school (Brownell et al., 2006; Gupta, Roos, Walld, Traverse, & Dahl, 2003). With registries typically based on a whole (or good approximation to a whole) population, bias associated with dealing with unrepresentative population subsets is minimised.

Figure 1 shows several types of data files linkable with a population registry. This Canadian provincial repository has been used for hundreds of published studies over the past forty years.

Figure 1

A deidentified population research data repository (Manitoba)



Ties with CancerCare are relatively recent. Cadham refers to the Cadham Provincial Laboratory. Healthy Child Manitoba encompasses several provincial programs, including testing involving the EDI (Early Development Instrument).

Wide data files created through record linkage

Repositories such as Manitoba's have typically been started with health information and expanded as other data sets became available. Having many descriptors for each individual ('wide' data files) facilitates work in several ways (National Research Council, 2013). Population-based indices can be created by combining files using techniques developed by Mosteller & Tukey (1977) (Roos et al., 2013). Data cleaning is a significant issue that has been treated elsewhere, and some general tools have been developed (Smith et al., 2015). Close relations with data providers facilitate ensuring quality.

Merging with clinical databases containing details on diseases increases opportunities to build good measures and provide long-term follow-up (Bernstein et al., 2016). Both sites based on linkable administrative data and those anchored in primary data have been continually working to add files and to expand the range of topics covered (Morris, 2015; Smith et al., 2011). Bringing in new files from other agencies or clinical investigators increases research opportunities but, often takes years of negotiations ('smiling persistence') (Borghol et al., 2012).

Large N

The large numbers of cases and variables available using administrative data facilitate both observational and interventional research. An investigator can focus on low prevalence conditions or rare events and still find enough cases for study.

Longitudinal nature

Both observational studies and quasi-experiments benefit from longitudinal data generated by the periodic updating of administrative files. Evaluating the long-term impact of health and educational interventions is an important application of such data (Murnane & Willett, 2011). The large N and ability to track individuals over long periods can be most useful in studies of the developmental origins of adult disease.

Locating individuals

Information as to where each individual in a population resides is very valuable in judging exposure to many variables. A highly influential American research program has been built on geographic variation in medical care (Wennberg, 2010), while Shadish (2013) has stressed the importance of control groups based on local conditions.

Sibling, twin, family analyses.

Parents, particularly mothers, can generally be specified through linkage of registries and hospital data to support more detailed work on families. Information on mothers and children allows specification of half-siblings. As discussed later, the more detailed the family information, the more possibilities for better casual inference.

Particular unexploited characteristics of administrative data can strengthen analytic approaches. Murnane and Willett (2011) have emphasised using multiple perspectives in worrying about selection bias. There are many types of selection bias. In the causal inference literature, "selection bias" is concerned with threats to internal validity and refers to systematic differences (both observed and unobserved) between those who select into the exposure state and those who select into the unexposed state. Murname and Willett's (2011, p. 330) concern is that "selection into treatments is as likely to be based on unobserved variables as on observed variables, and this source of bias remains despite the best efforts at statistical adjustment". Better research depends on "pattern-matching", on probing hypotheses using designs with different limitations to deal with various threats to internal validity (Jaffee, Strait, & Odgers, 2012; Steiner, Cook, Shadish, & Clark, 2010). Observational studies and quasi-experiments use design features to rule out many plausible, alternative explanations for an association but ensuring equivalence of the groups being compared is difficult, if not impossible. A pattern based on consistent results, a triangulation of evidence using different methods, allows more robust casual inference (Gage et al., 2016).

Multiple perspectives are particularly important because biases posed by unmeasured variables affect almost all observational studies. Without special circumstances (such as autocorrelated variables), typically less than 30% of the variance in social outcomes is handled by measured predictors (Roos et al., 2013). Measured relationships between risk factors and outcomes of interest tend to be spuriously high because statistical adjustment to deal with potential confounders "will usually be incomplete. Not only must all confounders be measured, but confounders must suffer from no measurement error" (Gage et al., 2016, p. 569); moreover, confounders must be included in the model with the correct functional form (e.g., linear vs. higher-ordered terms or as an interaction term). The many possible predictors, family linkages, and 'wide' data files often available in administrative data increase the variables and approaches available to attempt "to control for selection bias in causal research based on observational data" (Murnane & Willett, 2011). The different approaches to selecting comparison groups presented below are often relatively easy to implement with administrative data.

Comparison groups using geographical information

Several studies comparing randomised trials with different quasi-experiments have shown careful selection of covariates to often lead to better adjustment than using more covariates or relying on a particular form of data analysis (such as propensity scores or analysis of covariance) (Pearl, 2009; Shadish, Cook, & Campbell, 2002). Such 'careful selection' may well involve choosing the closest available neighbours (Cook, Shadish, & Wong, 2008; Lyons et al., 2014). Some jurisdictions will have geographic information that facilitates using such sophisticated methods as linear programming to select appropriate neighbours (Roos, Walld, & Witt, 2014).

Comparison groups using propensity scoring

Propensity score methods use analytic techniques to compare individuals exposed to an event of interest to those unexposed but demonstrably comparable on a wide range of observed factors. These methods often "group individuals on a range of characteristics that pre-date their exposure to a given risk factor of 'treatment'" (Jaffee et al., 2012, p. 7). One Manitoba project linking with Statistics Canada surveys used over 200 covariates to construct propensity scores which were then used to identify a group of smokers comparable to a group of non-smokers on these covariates; these two groups were then compared vis-à-vis their health service use to identify those differences attributable to smoking (Martens et al., 2015).

Propensity score methods have been steadily improving and provide many advantages. However, available covariates must well describe selection processes; Shadish (2013) also advocates the "use of comparison groups that are from the same location with very similar focal characteristics." Results from clinical trials often are found to diverge from those generated using propensity scores (Murnane & Willett, 2011; Sturmer, Glynn, Rothman, Avorn, & Schneeweiss, 2007). For example, meta analyses of randomised clinical trials on invasive cardiac management found 8-21% improvements in survival after AMI; two propensity score methods using American Medicare data showed a substantially greater effect (a 50% improvement) (Stukel et al., 2007).

Comparison groups using instrumental variables

Geographic information is also particularly useful for instrumental variable analyses, analyses "which identify an unconfounded proxy (an 'instrument') for the exposure of interest and assess the association between that and the outcomes to remove the biases of unmeasured confounding" (Gage et al., 2016, p. 580). Geographic variation in rates often provides the instrument. Robust instrumental variable analysis has been shown to provide unbiased estimates of causal effects. For example, in the Stukel et al. (2007) research, instrumental variable analysis, with geographical variation in the rates of invasive cardiac management providing the instrument, produced results close to those of the clinical trials.

Comparison groups using family data

Often underutilised family data can facilitate incorporating multiple comparisons within a single project. Observational studies based on comparing family members represent stronger designs for causal inference than traditional epidemiological studies. These family fixed-effects approaches are able to adjust for unobserved family-level factors that may confound the relationship between exposure and outcome. "Confounds due to passive gene-environment correlations" occur because "the same gene variants that influence how parents behave with their children may be transmitted to children and influence children's behaviour or abilities" (Jaffee et al., 2012, p. 274). Since children typically grow up within the same family, sibling designs help correct for error from omitted parental variables (such as income, which may affect the outcomes) and substantially reduce confounding in tests of casual hypotheses (Lahey & D'Onofrio, 2010). D'Onofrio, Lahey, Turkheimer, and Lichtenstein (2013) argue against traditional comparison of unrelated individuals in observational studies, explaining how details on family genetic relationships can help deal with different threats to interpretation.

Siblings share many experiences (and varying degrees of genetic similarity), but these experiences may occur at different times in the developmental process (Turkheimer & Waldron, 2000). Such "natural experiments" based on family relationships rely on enough sibling pairs differing from each other on important, measured predictors. Analyses of human capital in Canada, Scandinavia, and Florida focusing on 'sibship' (defined as having the same mother) have used administrative data, information on siblings and twins, long periods of follow-up, and very large numbers of cases (Black, Devereaux, & Salvanes, 2007; Figlio et al., 2014; Oreopoulos, Stabile, Walld, & Roos, 2008). As discussed later, specifying fathers is difficult in many administrative databases. This paper builds on information on mothers and children for the discussion of family relationships; thus, half-siblings form the basis for much of the discussion. The Scandinavian data have sometimes been able to study relatives "differing in both their genetic connectedness and the extent to which they were reared together" (Bjorklund et al., 2005). The power of these family-based designs to rigorously examine casual inferences will vary with the amount of information available (D'Onofrio et al., 2013).

For example, although smoking during pregnancy is a risk factor for offspring conduct issues, normal multivariable approaches appear insufficient. Several sibling fixed-effects analyses have shown that siblings "differing in their exposure to tobacco smoke in utero showed no differences in externalising behaviour" up through adolescence and into adulthood (Jaffee et al., 2012, p.10). In another analysis, Mortensen (2013) compared birth weights of children of Danish women differing in education with birth weights of children of their sisters (and cousins) also differing in education. Education appeared to be a much less important influence than would have been thought without the "children of sisters and cousins" comparisons.

Families and generations

Several information-rich environments and two of the well-known primary data collection efforts (the Framingham Study and the Panel Study of Income Dynamics) (D'Agostino et al., 2008; Johnson & Schoeni, 2011) facilitate making comparisons across generations within the same family. The Family Connections Genealogical Project in Western Australia has actively utilised this capacity and the large *N* for focused genetic epidemiological research (Brameld et al., 2014). The Danish Family Relations Database has been used to study how a number of diseases cluster in families (Boyd et al., 2009; Oyen et al., 2012).

Information across multiple generations may be gathered in at least two ways in administrative repositories. At the start of a government system (such as health insurance coverage) a defined population is typically entered into the database. If a registry organises individuals into families, older and younger family members will be noted. Wives and husbands are likely to be specified; informed assumptions about mothers and fathers can be made. Secondly, as a system continues over the years, children are born, individuals die, and people move in and out of the relevant jurisdiction. This presents the opportunity to develop birth cohorts and, in repositories combining registries and birth records, to specify mothers accurately.

Figure 2 diagrams the three generations of available Manitoba data. Much of the family research using these data has analysed individuals labeled as Generation 2. However, recent hip fracture studies have gone back to Generation 1 (Yang et al., 2016) while ongoing analyses incorporate Generations 2 and 3. With fewer divorces and more parents being married, information on fathers (at least married male parents) is easier to obtain for work based on Generation 1. With each birth cohort averaging between 12 and 16 thousand individuals, and loss-to-follow-up from birth to age 65 ranging from 1-1.5% annually, large numbers of cases are readily accumulated for multi-generational research. New individuals entering the province increase the N available for shorter-term observational or interventional studies.

Figure 2. Simplified overview of Manitoba generational information in repository

Generation 1	Parents of those born in 1970 and later (Health info from 1970 on; information from birth not available for parents)	
Generation 2	Born in 1970-1978 period in Manitoba; born in 1979 and subsequently (Health info from 1970 on; Education info for those born 1979 and after)	**New individuals entering Manitoba**
Generation 3	Children of those born from 1970 on in Manitoba (Health info from birth; Education info for those born 1979 and after)	

Each information-rich environment has characteristics facilitating certain types of research. Various registries are likely to have different strengths and weaknesses vis-a-vis identifying family relationships. Intergenerational analyses and large numbers of cases can be especially useful for creatively identifying subsamples, highlighting risk factors, studying bad events, and understanding child development.

Creatively identifying samples

Family data generated from a population registry can help with ethnic identification; Manitoba researchers have used such information to specify members of First Nations (North American Indian) communities, Metis (descendants of First Nations and Whites), and French Canadians. For example, registration of children in Status Verification files (to be considered First Nations benefitting from 'Treaty' status) often takes several years. Once a single family member has been specified from these files and registry data, parents and children could be noted as belonging to a particular minority group. In a 'mixed marriage', this would lead to errors but there has been no other way to proceed (Martens et al., 2011). Another project used enrollment in Francophone schools (run by the Franco-Manitoban School Division) for identifying French Canadian children who could then be linked to their families.

Highlighting risk factors

Risk factors for many medical conditions have a family component. Linkable administrative data can advance such research by: 1) noting conditions for which parental history can help indicate risk of the condition in their children, and 2) including siblings in addition to parents. Administrative and survey data are generally comparable in predicting hip fracture risk among the adult children of parents checked for their fracture history (Lix et al., 2017). Suggestions have been generated to help improve a widely adopted clinical tool (the World Health Organization's FRAX measure) to quantify patient fracture risk. Framingham researchers have generated a number of cardiovascular risk algorithms, but administrative data allow looking at a wider range of conditions. Scandinavian investigators have been particularly active in studying atrial fibrillation, mental health, and cancer (D'Agostino et al., 2008; Mortensen, 2013). Finally, risk might be incorporated in evaluation of screening programs, such as those for breast and prostate cancer. Administrative data may allow examining population coverage of such programs; a cost-benefit perspective suggests looking at how well screening works among those at higher risk (Gupta et al., 2003).

Studying adverse events

Most families will experience adverse events over the years; linked files from various agencies can help define events and specify outcomes. Such events will impact the mental health, life chances, and financial prospects of other family members. The importance of such shocks seems likely to vary with a family's socioeconomic status, geographic location, and so forth. For example, although the effects of accidents (such as concussions) can be studied just using diagnostic information from individuals in one or more databases, adverse events probably have wider impacts. Death of one family member (by suicide or automobile) has been explored in terms of its impact on other family members (Bolton et al., 2013; Bolton et al., 2014). Children's injuries may (or may not) affect parental health (Enns et al., 2016). Extending the research to examine longer-term effects would be very feasible. Additional worthwhile projects include looking at maternal and child outcomes when a child is taken into care by a government agency and examining family outcomes when one family member has had a serious medical diagnosis. Cross-sectoral information is particularly valuable; adverse events may affect the educational achievement (in addition to mental health) of younger family members.

Understanding child development

Population-based linkages directed toward studying parent-child interaction suggest many opportunities. In Manitoba, developmental patterns of mother and child can be studied from the birth of the mother and that of the child. Beginning in 1979, both health and educational information are present and accessible over 37 years. Various combinations of mother and child dyads can be constructed to assure adequate numbers of cases. For example, approximately 9,000 mother-child pairs can currently be followed up to age 15. This number will increase substantially over the coming years. Facilitated by the large N and many years of follow up, age comparisons can be coordinated. For example, a mother's health in her first ten years of life can be compared to that of her children during their first decade, expanding on the family history literature (Cunliffe, 2015; Dhiman, Kai, Horsfall, Walters, & Qureshi, 2014).

More generally, the analysis of dyadic and triadic relationships within large numbers of families is becoming increasingly practical. Recent Manitoba work examines the associations among mother—sibling1—sibling2; the influence of a mother's teen pregnancy is compared with that of an older sister's on the probability of a younger sister's teen pregnancy (Wall-Wieler, Roos, & Nickel, 2016a). Propensity scoring created control groups as equivalent as possible. A second paper looks at the effects of an older sister's experience with pregnancy (terminated versus continued to child birth) on the outcomes of a younger sister's pregnancy.

Additional parental variables can be incorporated into longitudinal observational studies using multiple birth cohorts (Currie, Stabile, Manivong, & Roos, 2010; Oreopoulos et al., 2008). For example, maternal mental health at various developmental stages may well influence long-term outcomes, even after controlling for a number of other measures. Maternal participation in various perinatal programs can also be used in the analysis. Three major outcomes include: infant health at birth, one-year mortality, and health and educational achievement at age 15. As noted below, study design is sensitive to the availability of different data sets.

Limitations

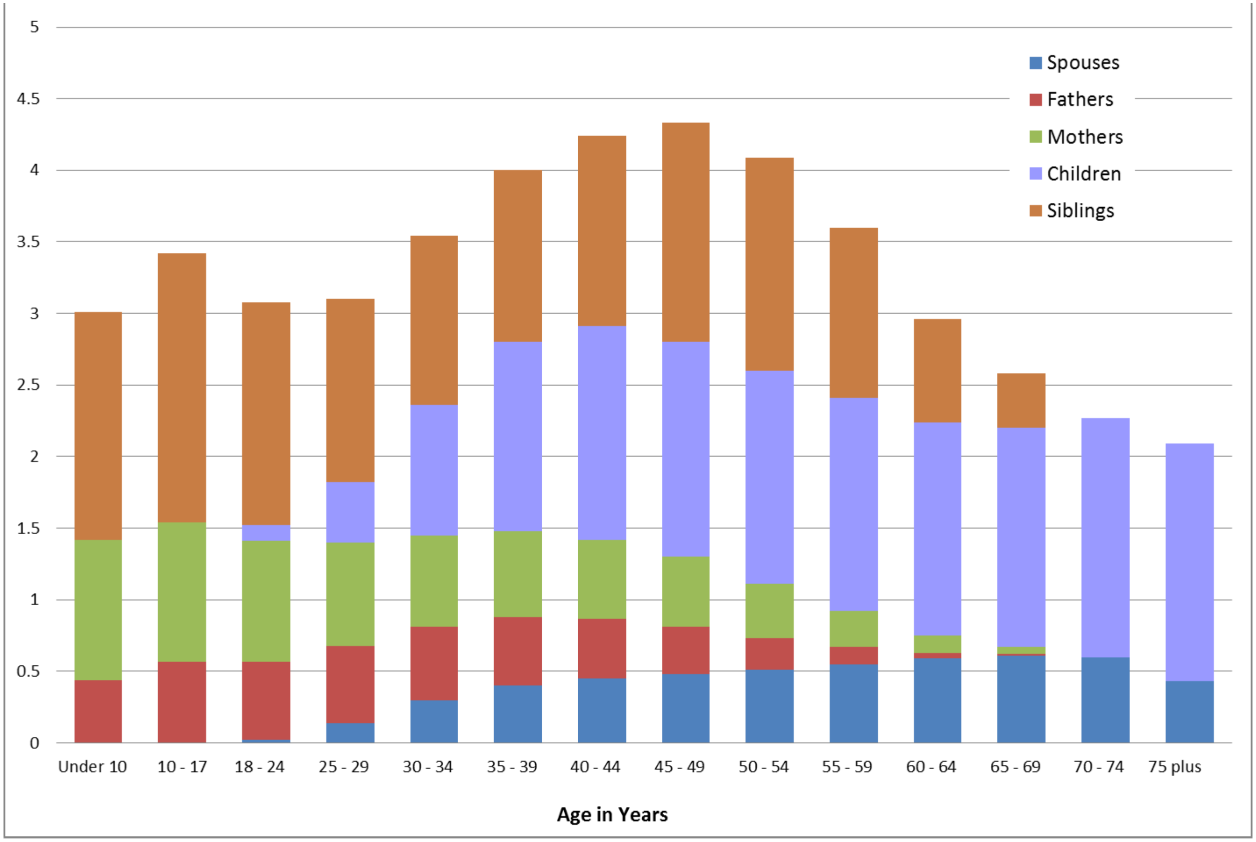
Tradeoffs

New data sets are always welcome but they typically provide new variables for only more recent years. Thus, in Manitoba, population-based information on Employment Income Assistance (similar to welfare) and on Child and Family Services (regarding serious family issues) is only available after 1995. Incorporating such information affects the birth cohorts chosen, the number of individuals studied, and the number of years of follow-up. Work examining the influence of an older sister's teenage pregnancy on a younger sister focused on younger sisters from age 16 to 19. Including adolescents aged 14 and 15 in the main analyses would have meant foregoing a grade nine educational achievement measure (girls were too young) and the Child and Family Services information (present for only a limited number of years). This decision did reduce the number of younger sisters available. However, sensitivity testing using only those covariates available in both samples showed few differences.

Practicality

Family identification mapping systems (developed by both the Panel Study of income Dynamics and the Manitoba Centre for Health Policy) can help assess project practicality by specifying the mean number of family members associated with any individual. On average, each 35-39 year old in Manitoba (approximately 80,000 in 2014) has almost five family members who can be linked to in the database—this includes one sibling, 1.5 children, a mother, and a father and approximately 40% to a spouse. Given the initiation of the registry in 1970, siblings over 19 could not be specified; this led to the 'disappearance' of siblings in the older age groups. Figure 3 suggests another problem: fathers of younger children are becoming increasingly difficult to identify.

Figure 3. Manitoba registry – Mean number of specifiable family members by age relationships to persons living as of June 1, 2014



Nonstandard and omitted variables

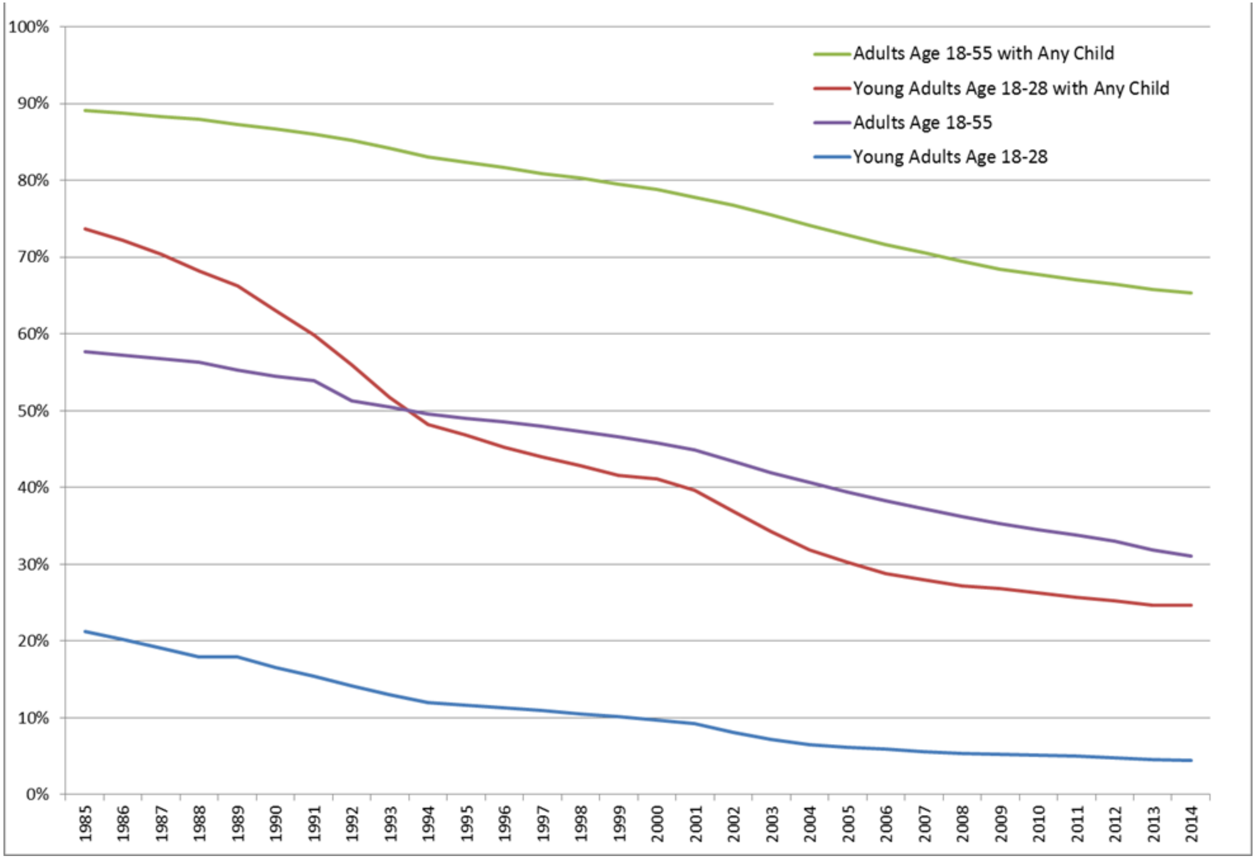
Certain types of measures, including an individual's socioeconomic status, are often not present in administrative data; several interrelated techniques can partially deal with such omitted information. As noted earlier, sibling analyses play a major role in handling omitted variables and better controlling for unmeasured risk factors. Multi-level modeling is generally helpful, while statistical tests for omitted variables are possible. Finally, the variance explained in Manitoba research compares favorably with that explained in the Framingham Study and the Panel Study of Income Dynamics (D'Agostino et al., 2008; Roos et al., 2013; Wall-Wieler, Roos, Chateau, & Rosella, 2016b).

Researchers have long discussed the substitution of small area data on socioeconomic status for (often unavailable) variables describing parental education and household income in considerable detail (Krieger, Waterman, Chen, Rehkopf, & Subramanian, 2016). Area-based socioeconomic measures are widely available but their appropriateness depends on the topic under study, the size of the areas providing information, the heterogeneity of the area's population, and other factors. Canadian analyses may be based on census areas or six-digit postal codes (considerably smaller than American census tracts). In at least some provinces small area data have been shown to approximate individual level measurement. (Mustard, Derksen, Berthelot, & Wolfson, 1999; Pampalon, Hamel, & Gamache, 2009; Schuurman, Bell, Dunn, & Oliver, 2007; Roos et al., 2013; Subramanian, Chen, Rehkopf, Waterman, & Krieger, 2006). One set of American researchers has used a group of individual measures plus neighbourhood income to create an index of family socioeconomic status that performs well (Figlio et al., 2014).

Social change, government policy, and data quality

Recent changes in society and in government data collection have combined to present significant challenges to the Manitoba family data (Roos et al., 2013). With more couples having children outside of marriage, the meaning of such important measures as marital status may be changing. Moreover, since 1996 the role of marital status and family number has become deemphasised in managing the provincial Pharmacare reimbursements. The ability to readily identify fathers has steadily deteriorated (Figure 4). Over 31% of recent mothers appear as Not Married on the registry but Not Single Parent on provincial Families First surveys. Statistics Canada tabulations suggest many of the mothers not responding to the surveys headed single parent families. Here, other linkable files such as various Manitoba surveys containing information on marital status or partners—plus better access to Vital Statistics files—may allow improvements in these data. The maintenance of such 'core data elements'– elements used across studies and across disciplines – has been a critical feature of both the Panel Study of Income Dynamics and the Framingham Study. The inability to mandate such elements as part of the core information represents a real weakness of administrative data.

Figure 4. Manitoba registry – Recording of per cent married by year



Discussion

Administrative data provide a wide range of opportunities. The capacity to organise family structure, health, and residential mobility information for large numbers of cases across various intervals can expand ‘life course epidemiology’. Tracking the occurrence of family events (deaths, divorces), diagnoses, and childhood moves can help assess their importance on later health and social outcomes (Currie et al., 2010). Life course epidemiology pioneered in the United Kingdom has traditionally used survey information to evaluate the possible effects of changes in one or two variables (Viner et al., 2015). Five time-varying measures have been incorporated in ongoing Manitoba work using administrative data to study high school graduation and externalising mental health conditions.

One analysis of the impact of early life events based on administrative data can be directly compared with work relying on primary data. Manitoba short- and long-term outcomes have been tracked over 18 years incorporating several measures of health at birth (birth weight, Apgar scores, and gestational length) and different outcomes (education, health, employment assistance) (Oreopoulos et al., 2008). Investigators using the Panel Study of Income Dynamics to study early life benefit from survey information on pregnancy intentions, family income data over the life course, and the ability to follow birth cohorts well into adulthood (Johnson & Schoeni, 2011). PSID outcomes include health (both childhood and adult) educational achievement (total years of schooling), and several labor market measures (annual income, hours worked, and hourly wages). Full siblings (particularly in recent years) can be better distinguished by primary data collection. Administrative data research (in Manitoba and Scandinavia) has been able to incorporate twin analyses and information on child health at various life stages (Black et al., 2007; Currie et al., 2010; Oreopoulos et al., 2008). Both administrative and primary data have the potential to measure important within-family differences (but using different variables).

Possibilities and missing pieces

Many linkable data sets are neglected or only partially utilised. Considerable research potential has been lost because of an inability to receive permission to link, or funding limitations. Expensive, large-scale surveys can be underused by researchers wanting to employ data linkage. For example, Statistics Canada has incorporated 'permission to link' items in many of its community health surveys; typically, over 90% of the responses are affirmative. Surveys (often including questions on respondent and parental education, obesity, and activity level) have been available for linkage to administrative data for many provincially approved projects for several years. Even in small provinces, combining these periodic surveys can provide over 50 thousand respondents. Although access for investigator-initiated research building on these high-quality data sets has been problematic, changes at the federal level may facilitate future studies.

Political constraints can limit important research efforts. In Australia, a Population Health Research Network across states/territories has been funded "to facilitate linkage between jurisdictional data sets, and between these data sets and research data sets, using demographic data" (Smith et al., 2011). However, Australian states do not automatically hold population registries and obtaining partial substitutes (such as voter registries) from the Dominion Bureau of Statistics has often been difficult. Under these constraints, much Australian work has focused on hospital information held by the states. This may change with recent efforts to make welfare data more accessible (Cowan, 2015).

Information-rich environments excel in the breadth of data available for work across many disciplines. The demand for studies using these environments seems likely to grow. However, in Canada only three provinces have developed large, relatively accessible population databases. Several others have cooperated on specific projects such as CNODES, the Canadian Network for Observational Drug Effect Studies (Suissa et al., 2012). Start-up costs are not inconsiderable; high levels of cooperation among ministries and university researchers are essential to provide both timely data and necessary economies of scale.

Linkable administrative data have generated widespread interest; international linkage networks continue to expand across fields (International Population Data Linkage Network, 2016). In just one example noted by Gage et al. (2016), if access to large cohort datasets continues to increase and such information can be linked to genomic research, existing information becomes more valuable. The breadth of possible work multiplies the possibilities.

Acknowledgements

Support was provided by the Canadian Institute for Health Research and the Canadian Institute for Advanced Research. The funding sources had no involvement in study design, analysis and interpretation of data, in writing the article, and in the decision to submit for publication.

The results and conclusions are those of the authors and no official endorsement by the Manitoba Centre for Health Policy, Manitoba Health, or other data providers is intended or should be inferred. The work was approved by the Research Ethics Board, Faculty of Health Sciences, University of Manitoba and the Health Information Privacy Committee of Manitoba Health, Healthy Living and Seniors. Data used in this study are from the Population Research Data Repository housed at the Manitoba Centre for Health Policy, University of Manitoba and were derived from information provided by Manitoba Health, Healthy Living and Seniors and Manitoba Education and Training under projects #2013/2014-04 and 2014/2015-20. All data management, programming and analyses were performed using SAS® version 9.3. We would like to thank Theresa Daniuk for her work on the manuscript.

References

Bernstein, C.M., Banerjee, A., Targownik, L., Singh, H., Burchill, C., Chateau, D., & Roos, L.L. (2016). Caesarean delivery is not a risk factor for the development of inflammatory bowel disease: A population-based analysis. Clinical Gastroenterology and Hepatology, *14*(1), 50-57. https://doi.org/10.1016/j.cgh.2015.08.005

Bjorklund, A., Jantti, M., & Solon, G. (2005). Influences of nature and nurture on earnings variation: A report on a study of various sibling types in Sweden. In: S. Bowles, H. Gintis, M. Osborne Groves (Eds). Unequal Chances: Family Background and Economic Success. Princeton, NJ: Princeton University Press (pp. 145-164).

Black, S.E., Devereux, P.J., & Salvanes, K.G. (2007). From the cradle to the labor market? The effect of birth weight on adult outcomes. Quarterly Journal of Economics, *122*(1), 409-439. https://doi.org/10.1162/qjec.122.1.409

Bolton, J.M., Au, W., Leslie, W.D., Martens, P.J., Enns, M.W., Roos, L.L., Katz, L.Y., Wilcox, H.C., Erlangsen, A., Chateau, D., Walld, R., Spiwak, R., Seguin, M., Shear, K., & Sareen, J. (2013). Parents bereaved by offspring suicide: a population-based longitudinal case-control study. JAMA Psychiatry, *70*(2), 158-167. https://doi.org/10.1001/jamapsychiatry.2013.275

Bolton, J.M., Au, W., Walld, R., Chateau, D., Martens, P.J., Leslie, W.D., Enns, M.W., & Sareen, J. (2014). Parental bereavement after the death of an offspring in a motor vehicle collision: A population-based study. American Journal of Epidemiology, *179*(2), 177-185. https://doi.org/10.1093/aje/kwt247

Borghol, N., Suderman, M., McArdle, W., Racine, A., Hallett, M., Pembrey, M., Hertzman, C., Power, C., & Szyf, M. (2012). Associations with early-life socio-economic position in adult DNA methylation. International Journal of Epidemiology, *41*(1), 62-74. https://doi.org/10.1093/ije/dyr147

Boyd, H.A., Poulsen, G., Wohlfahrt, J., Murray, J.C., Feenstra, B., & Melbye, M. (2009). Maternal contributions to preterm delivery. American Journal of Epidemiology, *170*(11), 1358-1364. https://doi.org/10.1093/aje/kwp324

Brameld, K.J., Dye, D.E., Maxwell, S., Brisbane, J.M., Glasson, E.J., Goldblatt, J., & O'Leary, P. (2014). The Western Australian family connections genealogical project: Detection of familial occurrences of single gene and chromosomal disorders. Genetic Testing and Molecular Biomarkers, *18*(2), 77-82. https://doi.org/10.1089/gtmb.2013.0254

Brook, E.L., Rosman, D.L., & Holman, C.D.J. (2008). Public good through data linkage: Measuring research outputs from the Western Australian data linkage system. Australian Journal of Public Health, *32*(1), 19-23. https://doi.org/10.1111/j.1753-6405.2008.00160.x

Brownell, M. Roos, N.P., Fransoo, R., Roos, L.L., Guevremont, A., MacWilliam, L., Yallop, L., & Levin, B. (2006). Is the class half empty? Socioeconomic status and educational achievement from a population-based perspective. IRPP Choices, *12*(5), 1-30.

Chetty, R., Hendren, N., Katz, L.F., & Lawrence, F. (2016a). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. American Economic Review, *106*(4):855-902. https://doi.org/10.1257/aer.20150572

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A., & Cutler, D. (2016b). The association between income and life expectancy in the United States, 2001-2014. Journal of the American Medical Association, *315*(6), 1750-1766. https://doi.org/10.1001/jama.2016.4226

Cook, T.D., Shadish, W.R., & Wong, V.C. (2008). Three conditions under which experiments and observational studies produce comparable casual estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management, *27*(4), 724-750. https://doi.org/10.1002/pam.20375

Cowan, P. (2015). House of Reps approves release of de-identified welfare data. Itnews. Available at: http://www.itnews.com.au/news/house-of-reps-approves-release-of-de-identified-welfare-data-410375.

Cunliffe, V.T. (2015). Experience-sensitive epigenetic mechanisms, developmental plasticity, and the biological embedding of chronic disease risk. Wiley Interdiscip Rev Syst Biol Med, *7*(2), 53-71. https://doi.org/10.1002/wsbm.1291

Currie, J., Stabile, M., Manivong, P., & Roos, L.L. (2010). Child health and young adult outcomes. Journal of Human Resources, *45*(3), 517-548. https://doi.org/10.1353/jhr.2010.0013

D'Agostino, R.B. Sr., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., & Kannel, W.B. (2008) General cardiovascular risk profile for use in primary care: The Framingham Heart Study. Circulation, *117*(6), 43-53. https://doi.org/10.1161/circulationaha.107.699579

D'Onofrio, B.M., Lahey, B.B., Turkheimer, E., & Lichtenstein, P. (2013). Critical need for family-based quasi-experimental designs in integrating genetic and social science research. American Journal of Public Health;*103*(Suppl 1), S46-S55. https://doi.org/10.2105/AJPH.2013.301252

Dhiman, P., Kai, J., Horsfall, L., Walters, K., & Qureshi, N. (2014). Availability and quality of coronary heart disease family history in primary care medical records: implications for cardiovascular risk assessment. PLoS One, *9*1, e81998. https://doi.org/10.1371/journal.pone.0081998

Enns, J., Gawaziuk, J.P., Khan, S., Chateau, D., Bolton, J.M., Sareen, J., Stone, J., Doupe, M., & Logsetty, S. (2016). Mental and physical health outcomes in parents of children with burn injuries as compared with matched controls. Journal of Burn Care and Research, *37*(1), e18-26. https://doi.org/10.1097/BCR.0000000000000309

Exploiting Existing Data for Health Research International Conference. (2013). St. Andrews, Scotland, August 28-30.

Figlio, D.N., Guryan, J., Karbownik, K., & Roth, J. (2014). The effects of poor neonatal health on children's cognitive development. American Economic Review *104*(12), 3921-3955. https://doi.org/10.1257/aer.104.12.3921

Ford, D.V., Jones, K.H., Verplancke, J-P., Lyons, R.A., John, G., Brown, G., Brooks, C.J., Thompson, S., Bodger, O., Couch, T., & Leake, K. (2009). The SAIL databank: Building a national architecture for e-health research and evaluation. BMC Health Services Research, *9*, 157. https://doi.org/10.1186/1472-6963-9-157

Gage, S.H., Munafo, M.R., & Davey Smith, G. (2016). Causal inference in developmental origins of health and disease (DOHaD) research. Annual Review of Psychology *67*, 567-580. https://doi.org/10.1146/annurev-psych-122414-033352

Gupta, S., Roos, L.L., Walld, R., Traverse, D., & Dahl, M. (2003). Delivering equitable care: Comparing preventive services in Manitoba, Canada. American Journal of Public Health, *93*(12), 2086-2092. https://doi.org/10.2105/AJPH.93.12.2086

Hand, D.J. (2016). Editorial: 'Big data' and data sharing. Journal of the Royal Statistical Society: Series A (Statistics in Society), *179*, 629-631. https://doi.org/10.1111/rssa.12185

International Data Linkage Conference. (2012). Advancing Knowledge for Better Health and Social Outcomes. Perth, Western Australia, May 2-4.

International Health Data Linkage Conference. (2014). Data Linkage for Better Public Policy, Vancouver, British Columbia, April 28-30.

International Population Data Linkage Network. (2016). Wales, United Kingdom, Available at: https://www.ipdln.org/.

Jaffee, S.R., Strait, L.B., & Odgers, C.L. (2012). From correlates to causes: Can quasi-experimental studies and statistical innovations bring us closer to identifying the causes of antisocial behavior? Psychological Bulletin, *138*(2), 272-295. https://doi.org/10.1037/a0026020

Johnson, R.C., & Schoeni, R.F. (2011). The influence of early-life events on human capital, health status, and labor market outcomes over the life course. B.E. Journal of Economic Analysis and Policy, *11*(3), 2521. https://doi.org/10.2202/1935-1682.2521

Jutte, D.P., Roos, L.L., & Brownell, M. (2011). Administrative record linkage as a tool for public health research. Annual Review of Public Health, *32*, 91-108. https://doi.org/10.1146/annurev-publhealth-031210-100700

Krieger, N., Waterman, P.D., Chen, J.T., Rehkopf, D.H., & Subramanian, S.V. (2016). Geocoding and monitoring US socioeconomic inequalities in health: An introduction to using area-based socioeconomic measures. The Public Health Disparities Geocoding Project monograph. Boston, MA: Harvard School of Public Health. Available at: https://www.hsph.harvard.edu/thegeocodingproject/.

Lahey, B.B., & D'Onofrio, B.M. (2010). All in the family: Comparing siblings to test casual hypotheses regarding environmental influences on behavior. Current Directions in Psychological Science, *19*(5), 319-323. https://doi.org/10.1177/0963721410383977

Levy, D., & Brink, S. (2005). Change of Heart: Unraveling the Mysteries of Cardiovascular Disease. New York, NY: Vintage Books.

Lix, L.M., Leslie, W.D., Yang, S., Yan, L., Walld, R., Morin, S.N., Majumdar, S.R., & Roos, L.L. (2017). Accuracy of offspring-reported parental hip fractures: a novel population-based parent-offspring record linkage study. American Journal of Epidemiology, 185 <needs to be italicized> (10), 974-981. https://doi.org/10.1093/aje/kww197

Lyons, R.A., Ford, D.V., Moore, L., & Rodgers, S.E. (2014). Use of data linkage to measure the population health effect on non-health-care interventions. Lancet, *383*(9927), 1517-1519. https://doi.org/10.1016/S0140-6736(13)61750-X

Martens, P., Nickel, N., Forget, E., Lix, L., Turner, D., Prior, H., Walld, R., Soodeen, R.A., Rajotte, L., & Ekuma, O. (2015). The Cost of Smoking: A Manitoba Study. Winnipeg, MB: Manitoba Centre for Health Policy.

Martens, P.J., Bartlett, J.G., Prior, H.J., Sanguins, J., Burchill, C., Burland, E., & Carter, S. (2011). What is the comparative health status and associated risk factors for the Metis? A population-based study in Manitoba, Canada. BMC Public Health, *11*, 814. https://doi.org/10.1186/1471-2458-11-814

Morris, A. The Farr Institute. (2015). Who are we and how to engage. Presentation at: Digital Health Assembly: Open Innovation Conference, Cardiff, UK, February 11, 2015.

Mortensen, L.H. (2013). Socioeconomic inequality in birth weight and gestational age in Denmark 1996-2007: Using a family-based approach to explore alternative explanations. Social Science and Medicine, *76*(1), 1-7. https://doi.org/10.1016/j.socscimed.2012.08.021

Mosteller, F., & Tukey, J.W. (1977). Data Analysis and Regression: a Second Course in Statistics. Reading, MA: Addison-Wesley.

Murabito, J.M., Pencina, M.J., Nam, B.H., D'Agostino, R.B. Sr., Wang, T.J., Lloyd-Jones, D., Wilson, P.W., & O'Donnell, C.J. (2005). Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. Journal of the American Medical Association, *294*(24), 3117-3123. https://doi.org/10.1001/jama.294.24.3117

Murnane, R.J., & Willett, J.B. (2011). Methods Matter: Improving Casual Inference in Educational and Social Science Research. New York, NY: Oxford University Press.

Mustard, C.A., Derksen, S., Berthelot, J-M, & Wolfson, M.C. (1999). Assessing ecologic proxies for household income: A comparison of household and neighbourhood-level income measures in the study of population health status. Health & Place, *5*(2), 157-171. https://doi.org/10.1016/S1353-8292(99)00008-8

National Research Council. (2013). Frontiers in Massive Data Analysis, Washington, DC: The National Academies Press.

Oreopoulos, P., Stabile, M., Walld, R., & Roos, L.L. (2008). Short, medium, and long term consequences of poor infant health: An analysis using siblings and twins. Journal of Human Resources, *43*(1), 88-138. https://doi.org/10.1353/jhr.2008.0003

Oyen, N., Ranthe, M.F., Carstensen, L., Boyd, H.A., Olesen, S.C., Olesen, S.P., Wolfahrt, J., & Melybe, M. (2012). Familial aggregation of lone atrial fibrillation in young persons. Journal of the American College of Cardiology, *60*(10), 917-921. https://doi.org/10.1016/j.jacc.2012.03.046

Pampalon, R., Hamel, D., & Gamache, P. (2009). A comparison of individual and area-based socio-economic data for monitoring social inequalities in health. Statistics Canada. [accessed 23 Sept 2013]. Available at: http://www.statcan.gc.ca/pub/82-003-x/2009004/article/11035-eng.htm.

Pearl J. (2009). Causality: Models, Reasoning and Inference. 2nd ed. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Power, C., Kuh, D., & Morton, S. (2013). From developmental origins of adult disease to life course research on adult disease and aging: Insights from birth cohort studies. Annual Review of Public Health, *34*, 7-28. https://doi.org/10.1146/annurev-publhealth-031912-114423

Public-Academic Research Colloquium: Leveraging Administrative Data for Social Policy. November 29 & 30, 2016. Washington, District of Columbia.

Roos, L.L., Walld, R., & Witt, J. (2014). Adolescent outcomes and opportunities in a Canadian province: Looking at siblings and neighbors. BMC Public Health, *14*(1), 506. https://doi.org/10.1186/1471-2458-14-506

Roos, L.L., Hiebert, B., Manivong, P., Edgerton, J., Walld, R., MacWilliam, L., & de Rocquigny, J. (2013). What is most important: Social factors, health selection, and adolescent educational achievement. Social Indicators Research, *110*(1), 385-414. https://doi.org/10.1007/s11205-011-9936-0

Roos, L.L., Brownell, M., Lix, L., Roos, N.P., Walld, R., & MacWilliam, L. (2008). From health research to social research: Privacy, methods, approaches. Social Science & Medicine, *66*(1), 117-129. https://doi.org/10.1016/j.socscimed.2007.08.017

Roos, N.P., & Roos, L.L. (2011). Administrative data and the Manitoba Centre for Health Policy: Some reflections. Healthcare Policy *6*(Special Issue), 16-28. https://doi.org/10.12927/hcpol.2011.22116

Schuurman, N., Bell, N., Dunn, J.R., & Oliver, L. (2007). Deprivation indices, population health and geography: An evidence of the spatial effectiveness of indices at multiple scales. Journal of Urban Health, *84*(4), 591-603. https://doi.org/10.1007/s11524-007-9193-3

Shadish, W.R. Propensity score analysis: Promise, reality and irrational exuberance. (2013). Journal of Experimental Criminology, *9*(2), 129-144. https://doi.org/10.1007/s11292-012-9166-8

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA: Houghton Mifflin.

Smith, M., Roos, L.L., Burchill, C., Turner, K., Ostapyk, T., Towns, D.G., Hong, S.P., Jarmasz, J.S., Ginter, J., Martens, P.J., Roos, N.P., Lix, L.M., Brownell, M., Azimaee, M., Soodeen, R.A., Nicol, J.P. (2015). Health services data: Managing the data warehouse: 25 years of experience at the Manitoba Centre for Health Policy. In: B. Sobloev, A. Levy, S. Goring (Eds). Data and Measures in Health Services Research. New York, NY: Springer Science+Business Media.

Smith, M., Semmens, J., Rosman, D., Ford, J., Storey, C., Holman, C.D.J., Fuller, E., & Gray, V. (2011). International health data linkage network. Healthcare Policy, *6*(Special Issue), 94-96. https://doi.org/10.12927/hcpol.2011.22127

Stanley, F., Glaubert, R., McKenzie, A., & O'Donnell, M. (2011). Can joined-up data lead to joined-up thinking? The Western Australian Developmental Pathways Project. Healthcare Policy, *6*(Special Issue), 63-73. https://doi.org/10.12927/hcpol.2011.22120

Steiner, P.M., Cook, T.D., Shadish, W.R., & Clark, M.H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. Psychological Methods, *15*(3), 250-267. https://doi.org/10.1037/a0018719

Stukel, T.A., Fisher, E.S., Wennberg, D.E., Alter, D.A., Gottlieb, D.J., & Vermeulan, M.J. (2007). Analysis of observational studies in the presence of treatment selection bias: Effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. Journal of the American Medical Association, *297*(3), 278-285. https://doi.org/10.1001/jama.297.3.278

Sturmer, T., Glynn, R.J., Rothman, K.J., Avorn, J., & Schneeweiss, S. (2007). Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. Medical Care, *45*(Suppl 2), S158-165. https://doi.org/10.1097/MLR.0b013e318070c045

Subramanian, S.V., Chen, J.J., Rehkopf, D.H., Waterman, P.D., & Krieger, N. (2006). Comparing individual- and area-based socioeconomic measures for the surveillance of health disparities: A multilevel analysis of Massachusetts births, 1998-1991. American Journal of Epidemiology, *164*(9), 823-834. https://doi.org/10.1093/aje/kwj313

Suissa, S., Henry, D., Caetano, P., Dormuth, C.R., Ernst, P., Hemmelgarn, B., Lelorier, J., Levy, A., Martens, P.J., Paterson, J.M., Platt, R.W., Sketris, I., Teare, G., & Canadian Network for Observational Drug Effect Studies. (2012). CNODES: The Canadian Network for Observational Drug Effect Studies. Open Medicine, *6*(4), e134-140.

Turkheimer, E., & Waldron, M. (2000). Nonshared environment: A theoretical, methodological, and quantitative review. Psychological Bulletin, *126*(1), 78-108. https://doi.org/10.1037/0033-2909.126.1.78

Viner, R.M., Ross, D., Hardy, R., Kuh, D., Power, C., Johnson, A., Wellings, K., McCambridge, J., Cole, T.J., Kelly, Y., & Batty, G.D. (2015). Life course epidemiology: Recognising the importance of adolescence. Journal of Epidemiology & Community Health, *69*(8), 32-37. https://doi.org/10.1136/jech-2014-205300

Wall-Wieler E, Roos LL, & Nickel N. (2016a). Teenage pregnancy: The impact of maternal adolescent childbearing and older sister’s teenage pregnancy on a younger sister. BMC Pregnancy and Childbirth, *16(*1), 120. https://doi.org/10.1186/s12884-016-0911-2

Wall-Wieler, E., Roos, L.L., Chateau, D.G., & Rosella, L.C. (2016b). What predictors matter: risk factors for late adolescent outcomes. Canadian Journal of Public Health, *107*(1), e16-22. https://doi.org/10.17269/cjph.107.5156

Wennberg, J.E. (2010). Tracking Medicine: A Researcher’s Quest to Understand Health Care. New York, NY: Oxford University Press.

Williams, D.R., McClellan, M.B., & Rivlin, A.M. (2010). Beyond the affordable care act: Achieving real improvements in Americans' health. Health Affairs, *29*(8), 1481-1488. https://doi.org/10.1377/hlthaff.2010.0071

Yang, S., Leslie, W.D., Yan, L., Walld, R., Roos, L.L., Morin, S.N., Majumdar, S.R., & Lix, L.M. (2016) Objectively verified parental hip fracture is an independent risk factor for fracture: A linkage analysis of 478, 792 parents and 261, 705 offspring. Journal of Bone and Mineral Research. *31*(9), 1753-1759. https://doi.org/10.1002/jbmr.2849