Administrative Record Linkage as a Tool for Public Health Research

Douglas P. Jutte, Leslie L. Roos, and Marni D. Brownell

1School of Public Health, University of California, Berkeley, California 94720-1190; email: dpjutte@berkeley.edu
2Department of Community Health Sciences, Faculty of Medicine, University of Manitoba, Winnipeg, Manitoba R3E 3P5, Canada; email: Leslie_Roos@cpu.umanitoba.ca, Marni_Brownell@cpu.umanitoba.ca

Keywords
population health, social determinants of health, life course, socioeconomic status, longitudinal data

Abstract
Linked administrative databases offer a powerful resource for studying important public health issues. Methods developed and implemented in several jurisdictions across the globe have achieved high-quality linkages for conducting health and social research without compromising confidentiality. Key data available for linkage include health services utilization, population registries, place of residence, family ties, educational outcomes, and use of social services. Linking events for large populations of individuals across disparate sources and over time permits a range of research possibilities, including the capacity to study low-prevalence exposure-disease associations, multiple outcome domains within the same cohort of individuals, service utilization and chronic disease patterns, and life course and transgenerational transmission of health. Limited information on variables such as individual-level socioeconomic status (SES) and social supports is outweighed by strengths that include comprehensive follow-up, continuous data collection, objective measures, and relatively low expense. Ever advancing methodologies and data holdings guarantee that research using linked administrative databases will make increasingly important contributions to public health research.
INTRODUCTION

Insights into the complex web of factors influential in health (12), the importance of determining pathways and mechanisms of causation (48), and the role of life course health trajectories (5, 43) highlight the need for data capable of supporting the next stages of public health and population health research. Traditionally, the most powerful tools for unraveling these complicated questions have been large-scale, longitudinal studies. However, longitudinal collection of data is difficult to coordinate, time-consuming, and expensive. These studies are complicated by issues of loss-to-follow-up, underrepresentation of ethnic and racial minorities, reduced participation by individuals in the tails of the socioeconomic distribution, and frequent use of self-report for disease and risk exposure, as well as limitations resulting from exclusion criteria and patient consent requirements that reduce generalizability to the overall population (4, 91). We explore a different approach to data collection and research that could circumvent some of the thorniest of these problems and may shed new light on the complex interplay of risk and protective factors on health over time.

Linked administrative data comprise information already widely and diligently collected on large populations for other purposes, then merged at the individual level using unique, anonymized identifiers and made available for academic research. More extensive than routine vital statistics collected on birth, death, or disease and more inclusive than disease-specific registries, administrative health databases are increasingly linked to population-wide institutional data from social service agencies, educational institutions and census reports and utilized to identify interpersonal family connections within and across generations. These information-rich environments can, relatively inexpensively, supply the very large samples and long-term observations that primary data collection often cannot. And critically for epidemiologic research, in locations where population registries have been created, it is possible to determine the disposition of an entire population (the denominator) rather than that of only those interacting with a health or educational institution (73, 75). Although important issues remain regarding database governance, privacy protection, research access, and distribution of findings, these obstacles are not insurmountable (82).

Some researchers have described linked administrative information as an opportunity to create new data from existing sources. The merging of data derived from disparate authorities to include events occurring to individuals and families over time and intergenerationally, combined with the ability to do this for large populations, allows for a wide range of important and often unique public health investigations. It also provides a relatively low-cost supplement to the too few longitudinal studies critical for future public health research.

LOCALES OF MAJOR DATA LINKAGE EFFORTS

The published literature has identified several centers noted for highly productive research utilizing linked administrative data while maintaining privacy and confidentiality (36, 72, 73). These include the Oxford Record Linkage Study (ORLS), the Scottish Record Linkage System, the MigMed2 Database (Sweden), Statistics Norway, the Rochester Epidemiology Project (Minnesota), the Western Australia Data Linkage System (WADLS), and several Canadian centers, including the Manitoba Center for Health Policy (MCHP), the Center for Health Services and Policy Research (CHSPR) in British Columbia, and the Institute for Clinical and Evaluative Sciences (ICES) in Ontario. Detailed descriptions of these research centers and their investigator-initiated research and government-funded policy work have been published elsewhere (10, 29, 35, 36, 40, 45, 61, 72, 86).

More recently, New Zealand and Wales have begun data linkage operations, and a nationwide Australian effort, the Public Health
Research Network (PHRN), will soon provide that country of 21 million the world’s largest population health database (24, 35). The International Health Data Linkage Network (IHDLN), comprising a substantial subset of existing health data linkage centers, was formed in 2008 to foster collaboration and exchange, describe best practices, and record output from research and programs based at its participating centers.

**DATA LINKAGE METHODOLOGY**

Data linkage is defined as “the bringing together of information from two records that are believed to relate to the same individual or family” (9). Linkage is achieved by using a limited set of basic sociodemographic factors, “linkage variables,” to identify uniquely and reliably an individual across two or more datasets (35, 72).

Privacy and the maintenance of secure information are always of utmost concern. To achieve high-quality linkages without compromising confidentiality, multistage deidentification processes have been developed by the different research centers to address legal and political concerns without reducing the value of the data (44, 72, 93). Although the details vary across sites, to enhance privacy in most cases an outside agency serves as a linkage center. The merging of information is completed through a series of steps. In the first step, the linkage center receives files from data trustees (e.g., health or educational agencies) containing only linkage variables but no program data. In the second step, these linkage variables are used to identify individuals with information present across agency datasets. Third, a unique record number is assigned to the identified individuals. Finally, this unique record number is sent back to the data trustees, who use it to provide the data repository or researcher with the requested administrative content data but without the associated identifying information. Thus, confidentiality is maintained because neither the linkage center, researchers, or repository are ever simultaneously in possession of both the individually identifiable linkage data and programmatic content associated with that individual. Only the original agency maintains those complete records (13, 36, 45, 72).

Despite the potential for larger scale data sharing, this approach can actually enhance individual confidentiality over the traditional, investigator-initiated use of private health data (73). For example, Trutwein and colleagues (90) found that research projects utilizing name-identified health data dropped dramatically after the introduction of a record linkage system in Western Australia (Figure 1).

**DATA CONTAINED WITHIN LINKED ADMINISTRATIVE RESOURCES**

The information available for linkage across databases varies by locale. Variation results from the length of time a center has been in existence, the type of health care financing (e.g., single- versus multipayer systems), local politics, financial resources, and often the leadership of participating agencies and data linkage centers. Using Manitoba Centre for Health Policy as an example, Figure 2 demonstrates the range of data sources that can be linked at the individual level within a population.

**Health Data**

The backbone for linked administrative health systems is hospital discharge diagnoses and vital statistics on birth, death, and disease. However, a wide range of additional health information has been linked at the individual level, including pharmaceutical and immunization records, outpatient clinical diagnoses, emergency room treatment, home care, nursing records, maternal care and birth records, and cancer or other disease registries. Existing administrative data have also been successfully linked to case-control survey populations with collected biological specimens (46).

**Residential Information**

A second major category of data is place of residence, which serves several important purposes.
Length of time spent in high- or low-income areas provides valuable contextual information. Census-based information tied to region, neighborhood, or urban block face can supplement, or even replace, often limited individual-level socioeconomic data (47, 65, 84). Residential information over time provides evidence of mobility, a variable known to influence health outcomes for both children and adults (22, 53). Such mobility may be a disruptive life event if occurring during childhood (2, 33, 70). Change in residence across neighborhoods of differing socioeconomic status (SES) over time also provides information on upward or downward socioeconomic trajectories for individuals over the life course (31).

**Family-Related Variables**

Number of children in the family, birth order, parental age and marital status, and number of years living in a single-parent household have all been identified as factors important in health (10, 14, 27, 32). Changes in family structure over time can also be ascertained and are predictive of child well-being (83). Although methods for identifying family relationships vary across centers, near-universal mother-child specification is possible (enabling sibling links) and fathers have been identified in upwards of 85%
of cases (72). In some settings, more extended family links can be made to first-degree cousins and grandchildren as well, the latter allowing for transgenerational inferences (7, 10, 41).

**Educational and Social Variables**

The most recent categories of data available for linkage, and potentially the most important for future public health research (60, 79, 88), are markers for social and educational predictors and outcomes. Detailed, population-wide information on educational outcomes is available in a limited number of locations for children (15, 56, 58) and in even fewer locations for adults (10, 87). Educational data provide, for health studies in children, the opportunity to examine more than one domain of well-being simultaneously, e.g., Jutte et al. 2010 (41) and, for adults, the ability to account for differences in prior educational success when examining health outcomes. Similarly, population-wide data on interactions with the justice system, the child protection system, or receipt of government financial assistance can be used to identify high-risk individuals better or serve as additional examples of negative individual outcomes. Data on income assistance have been used to supplement census-based socioeconomic information with individual-level information on low-income status (41, 76). Linkages to routine screening data can provide detailed social information at the individual level. For example, the Families First program in Manitoba, Canada, screens all newborns for risks such as maternal history of prenatal smoking, alcohol use, and family stress (17). Population-wide screening of child development at kindergarten entry using the Early Development Instrument (EDI) has been incorporated into administrative data in several jurisdictions (39, 56).

**Population Registry**

Another important variation across centers is the existence, or not, of a population registry. Jurisdictions with identification numbers provided for every resident, such as the provincial single-payer health care systems in Canada, allow for creation of a population registry. For example, in Manitoba each resident is provided with a personal health identification number (PHIN). Because a high proportion of the residents interact with the health system annually [e.g., in childhood, greater than 98% over any four-year age period (20)], Manitoba Health is able to maintain a registry of active PHINs covering the entire population. This population registry is ideal for providing a denominator for all population-level analyses. Jurisdictions without an individually linked population registry must rely on less precise regional census information or similar data.

**RESEARCH UTILIZING LINKED ADMINISTRATIVE DATA**

To illustrate the potential for public health investigations, we explore some examples of research using linked administrative data.

**Population-Based Cohort and Case-Control Studies**

One strength of population-wide administrative data is the capacity to study low prevalence disease-disease and procedure-disease associations, associations unlikely to be funded for study using randomized controlled trials (30). For example, using a series of nested, case controls in the ORLS, Goldacre and colleagues examined links between gall bladder disease and colon cancer, appendectomy and inflammatory bowel disease, vasectomy and prostate disease, and mumps immunization and aseptic meningitis (30, 49, 62). Similarly, using Swedish MigMed data, Sundquist and colleagues (80) have examined Graves’ disease and risks of several cancers.

Such matched set analysis can be done for any comparison of choice (23). For example, to examine the relationship between residential mobility and schizophrenia, Lix and colleagues (53) created three cohorts matched on age, sex, and urban/rural residence. One group
had diagnosed schizophrenia. The other two served as comparisons with severe physical illness (inflammatory bowel disease) or no substantial mental/physical problems. With this approach, they showed consistent determinants of residential mobility across all groups (marital status, income, use of physician services) and demonstrated increased residential mobility for patients with schizophrenia.

**Multiple and Overlapping Outcome Domains**

Data linked from disparate sources provide the capacity to examine outcomes from different domains within the same cohort of individuals. For example, using a regional birth cohort of children, Jutte et al. (41) examined both medical and educational outcomes in relation to several biological and social risk predictors present at birth. Figure 3 shows rates of hospitalization and scholastic failure by gestational age and maternal age (complete data online, Supplemental Figure 1). Follow the Supplemental Material link from the Annual Reviews home page at http://www.annualreviews.org).

Population-Based Prediction of Disease

Using health survey data linked to administrative health services data, ICES researchers in Ontario, Canada, developed and validated an

**Figure 3**

Proportion of poor health and educational outcomes associated with early-life clinical and social predictors: gestational age and maternal age. Used with permission from Reference 41 (complete study comparison results available as Supplemental Figure 1; follow the Supplemental Material link from the Annual Reviews home page at http://www.annualreviews.org).
algorithm for population-based prediction of diabetes (78). The Diabetes Population Risk Tool (DPoRT) accurately predicts diabetes risk in a population using self-reported measures available in routine population health surveys rather than collating only the detailed clinical data used for individual diabetes risk assessment. This population-level approach to estimating disease incidence allows for improved, lower-cost population health planning and enhanced assessment of the impact of illness prevention strategies.

The Inclusion of Individuals Not Receiving Services

In public health research, knowledge of the denominator—the total population—can be critical. In locales with a population registry (e.g., Sweden, Norway, and Manitoba, Canada) or locations utilizing multiple data sources capable of capturing nearly the entire populace (e.g., Western Australia), the population denominator is available. A study of school children completed by MCHP researchers vividly illustrates the importance of the denominator (15). In Figure 5, the left graph illustrates the passing rate for all twelfth-grade students who took the compulsory language arts exam. A clear gradient is present across neighborhood SES quartiles; 75% of the lowest-SES children passed versus 92% of highest-SES children. In contrast, the right set of columns shows the disposition of all 18-year-olds from the 1984 birth cohort who should have taken the exam, not just those in school the day of the exam. When accounting for the children who had missed the exam, or had fallen behind or withdrawn from school, the actual passing rate among the lowest-SES group dropped to 27%, just one-third of the 77% passing rate in the highest-SES neighborhoods.

Likewise, linked data can help investigators determine the extent to which a program or screening tool reaches the entire target population. For example, the province of Manitoba screens all newborn infants for health and social risk using the Families First screen (originally called BabyFirst) to stratify children into no-risk and at-risk categories. However, the population registry also identifies a critical third category of children: those born but not screened. Brownell et al. (17) showed that although there existed a threefold variation across health districts in percentage of infants (born 2002) meeting at-risk criteria (8%–24%), in some jurisdictions nearly three-quarters of infants had never been screened (see Supplemental Figure 2 for details). Subsequent analysis, again using data linkage, found that among children requiring foster care by age two, 42% had been among those not screened. Combined, these findings provided powerful evidence for policy makers that screening efforts were, in some cases, inadequate and that these missed children represented a particularly high-risk group meriting additional resources.

Chronic Disease Surveillance and Improved Prevalence Estimates

Case definition using a single data source alone (e.g., physician office claims or hospital discharge diagnoses) is generally inadequate for surveillance purposes (55). However, using data from more than a single source improves the necessary sensitivity and specificity for cost-effective chronic disease surveillance...
For example, links to prescription data can improve the capture of conditions with disease-specific medications such as diabetes and asthma. The eventual linkage of gold-standard clinical data or laboratory evidence will further improve case identification at the population level (54).

Longitudinal administrative data linkage can also improve estimates of disease prevalence and health service utilization. For example, in any given year ~7% of children in Manitoba are involved with Child and Family Services. However, over a 10-year period, more than 17% of children required intervention (77). Information such as this has been used successfully to create a sense of policy urgency and to provide a perspective on the true proportion of health disparities and population health risks (77).

**Life-Course and Transgenerational Investigations**

Another feature of administrative data that is difficult to achieve in other settings is the enhanced capability to evaluate exposure effects over the life span because data are collected continuously over time. For example, to assess the impact of low-income housing on later life chances, Oreopoulos (67) linked home address data identifying children in public housing to information on long-run labor market outcomes (eventual earnings, unemployment likelihood, and welfare participation) extending into their thirties. Similarly, Jutte et al. (42) used existing data to follow children of teen mothers into early adulthood and assessed outcomes at different points over the life course.

The addition of parent-to-child family links also allows for transgenerational analyses. For example, using nationwide population data in Sweden, Sundquist, Li, and colleagues examined the differential impact of maternal and paternal factors on the transmission of coronary heart disease (85), risk of small-for-gestational age status (52), and risk of schizophrenia and other psychotic disorders (51).

**LIMITATIONS AND STRENGTHS OF ADMINISTRATIVE DATA**

Roos et al. (72) published a detailed list of the attributes associated with linked administrative data compared with those of a well-respected example of longitudinal primary data.
collection, the Panel Study on Income Dynamics (PSID). See Table 1. However, a few limitations and strengths of administrative data merit specific attention.

One important limitation of administrative data is the frequent lack of individual-level SES information. Traditional SES measures of income, wealth, achieved educational level, or job description often are not recorded at the population level, Scandinavian databases being the notable exception (8, 11). However, several researchers have shown that census data can provide a fair approximation of family or household levels of income (47, 65, 84). Others have successfully linked survey data on a subset of the population to partially replicate analyses with individual-level data and, thus, correlate the findings (74). Composite indices of census-derived data have also shown promise in approximating household income

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Longitudinal primary data (Panel Study on Income Dynamics)</th>
<th>Population-based administrative data (Manitoba and other sites)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cases</td>
<td>Several thousand or smaller</td>
<td>Often more than one million</td>
</tr>
<tr>
<td>Cost</td>
<td>High on a per-person basis</td>
<td>Very low on a per-person basis</td>
</tr>
<tr>
<td>Representativeness</td>
<td>Often national</td>
<td>Often from a province or state</td>
</tr>
<tr>
<td>Population studied</td>
<td>Subjects sampled and tracked</td>
<td>Built on a registry of an entire population</td>
</tr>
<tr>
<td>Research design</td>
<td>Often complex designs needed to increase power and control costs</td>
<td>Given a population, complex designs can be imposed retrospectively as needed</td>
</tr>
<tr>
<td>Record linkage</td>
<td>Useful in some contexts</td>
<td>Critical to check data quality and expand scope of information sources</td>
</tr>
<tr>
<td>Individual follow-up</td>
<td>Before and after an event</td>
<td>Before and after an event</td>
</tr>
<tr>
<td>Coverage and loss to follow-up</td>
<td>Nonresponse and differential attrition possible</td>
<td>Differential attrition possible</td>
</tr>
<tr>
<td>Updating</td>
<td>New data must be collected and merged with existing data</td>
<td>Multifile information must be cleaned and merged with existing data. Cleaning relies on record linkage</td>
</tr>
<tr>
<td>Time</td>
<td>Information must be collected (typically annually or at longer intervals)</td>
<td>Information provided at relatively short intervals (from daily to annually)</td>
</tr>
<tr>
<td>Place</td>
<td>Information at time of study (historical reconstruction possible)</td>
<td>Detailed information usually provided close to date of move</td>
</tr>
<tr>
<td>Longitudinal</td>
<td>Goes back many years</td>
<td>Goes back many years</td>
</tr>
<tr>
<td>Neighborhoods</td>
<td>Flexible construction from postal code or census area</td>
<td>Flexible construction from postal code or census area; large N may permit flexible assignment to generate nearest neighbors</td>
</tr>
<tr>
<td>Life events</td>
<td>Collected as part of design</td>
<td>Possibly available from registry or other sources</td>
</tr>
<tr>
<td>Family and intergenerational data</td>
<td>Collected as part of design; sibling and intergenerational studies facilitated</td>
<td>Assessing a family composition at any point in time possible; sibling and twin studies facilitated</td>
</tr>
<tr>
<td>Limits</td>
<td>Important information likely to be collectable for entire sample</td>
<td>Important information may be missing or available only for a subpopulation</td>
</tr>
<tr>
<td>Variables</td>
<td>Defined by researchers; scaling possible</td>
<td>Defined by others for administrative purposes; creating meaningful variables may be very time-consuming; scaling possible</td>
</tr>
<tr>
<td>Intellectual history</td>
<td>Scope of data collection often expanded to provide a rich data set</td>
<td>Scope expanded beyond initial health care data by receiving files from other agencies</td>
</tr>
</tbody>
</table>
levels. For example, the socioeconomic factor index (SEFI) is a validated measure derived from principal components analysis that utilizes neighborhood-level unemployment rate, rate of high-school completion, percentage of lone parent households, and female workforce participation to provide a summary score for any residential address (59). Similar efforts in Western Australia resulted in the design and use of the socio-economic indices for areas (SEIFA) (58, 89). Such indices can also be used as a contextual community measure. For example, Swedish researchers have incorporated a neighborhood deprivation index into multilevel analyses of coronary heart disease (87, 94).

Data on social supports and interpersonal relationships are another type of information not often available in population-wide administrative data. Studies have shown these to be important components of health and well-being (19, 37). Race/ethnicity is also often unavailable in administrative data. However, at least in children, published epidemiologic work has shown that minority status and ethnicity are inconsistently related to poor health outcomes, particularly after accounting for SES factors (66, 69). Combined, these limitations remain important issues to overcome for research attempting to disentangle the roles of SES, discrimination, social supports, and neighborhood effects on population health.

Administrative data do have a number of important strengths, however. For example, such data are free from many of the measurement and loss-to-follow-up problems associated with longitudinal surveys (4, 72). The flexibility of continuously collected administrative data presents some advantages over surveys dependent on expensive, repeated interviews of the same respondents at fixed intervals (72). Researchers are able to study the health or health care utilization of a population for any length of time and for any year or set of years (72). This flexibility can be useful for natural experiments such as abrupt economic changes or the implementation of new policies. Roos et al. (73) highlight a number of capabilities for administrative data pertinent to health policy analysis. These include the ability to compare regions, areas, and hospitals; study policy interventions longitudinally; combine information related to physicians and their patients; sum expenditures across different services within the health care system; and examine social determinants of health using education and family services data in conjunction with health-related information. The latter was demonstrated in several of the examples provided above.

**IMPORTANT CHALLENGES FACING ADMINISTRATIVE DATA LINKAGE**

Issues of privacy and confidentiality continue to be among the most prominent concerns facing the collection and use of administrative health data. In a 2006 issue of *BMJ* dedicated to this topic, Davies & Collins (21) noted that “interpretations of legislation [related to topics of health data use] seem to have been driven less by careful consideration of the likelihood of real harm for individuals than by the desire to minimize the risk of criticism for organisations.” They argued that, in fact, the overinterpretation of protective legislation could represent a risk to the public in its own right. The British Academy of Medical Sciences (1) noted there is little evidence that the use of confidential records in medical research has caused serious harm, and they further stated that advances in the fields of public health and population health have been increasingly inhibited by inappropriate constraints laid down by confusing legislation as a result of an undue emphasis on privacy and autonomy.

A related controversial issue is whether informed consent is required from patients prior to the use of personal data within administrative databases. Although deliberate care is required to create standards that protect privacy (81, 82), the rigid use of a “consent or anonymize” policy can be detrimental to research in terms of scientific opportunity, time, and financial resources (1). For example, Al Shahi & Warlow (3) argue that an “authorization bias” is introduced by the consent/authorization
process, which limits potential uses of the data and makes generalizability uncertain. Tu et al. (91) reported in the New England Journal of Medicine that in a clinical stroke registry requiring consent, a substantial bias in costs and death rate (a threefold lower in-hospital death rate for consented patients) occurred, even in a setting with specialized assistance available for the consent process. These authors argued for the need for legislation that would permit waivers of informed consent for minimal-risk observational research. Others believe that in the case of publicly funded health care, a social obligation exists for patients to permit use of their deidentified information without consent to allow the health care system to be monitored and improved for the benefit of all (3, 92).

The form in which administrative data is maintained is another decision that confronts groups creating or managing linked information and researchers planning to use the data. Some locales maintain a data repository of information collected periodically from a number of contributing agencies, e.g., the MigMed2 database in Sweden or the Manitoba Population Health Research Data Repository. However, the degree of data integration maintained on an individual varies. For example, the Manitoba repository does not maintain longitudinal, complete data files on each individual. Rather, the repository partitions sociodemographic and content data until requested for a specific research project. One impact of the repository approach is the need, in most cases, for the investigator to work with a dedicated and specially trained programmer or to be on-site to access the data.

Alternatively, some centers use a virtual repository, where primary data custodians (government departments or ministries) maintain direct control of their data and do not provide copies to a third party for storage and access. For example, Western Australia and PHRN (35) maintain only demographic data with which they use a linkage algorithm to request specific data from custodians only upon request from researchers with approved projects. Although this approach is seen by some to be at lower risk for breach of confidentiality, it requires greater computing power to recreate the linkages from scratch with each new project.

NEXT STEPS, NEW RESEARCH

Information-rich environments based on longitudinal administrative data provide an opportunity for both methodological improvements and new research directions.

Design and Analysis

Linked data can support better research design for both academic and policy research. Work by Mosteller & Tukey (64) has proven particularly helpful in allowing messy categorical data to be combined so as to generate reasonably normalized distributions. In Manitoba, this has facilitated building and validating several indices of educational performance (based on the type of information in Figure 5). Such population-based indices support multivariate analyses of a variety of outcomes.

Multilevel modeling seems especially well-suited for administrative data with different types of variables measured at different levels. For example, individuals can be nested within families nested within particular neighborhoods (26). Such modeling can account for (a) the dilemma of unit of analysis, (b) the lack of independence among observations (e.g., children in the same family), and (c) unobserved, confounding variables (57). Using this approach with administrative data, the influence of social environmental factors can be compared with that of individual-level factors (eg., 58, 87, 94), although the relationships among variables at different levels also suggest comparing sibling with neighbor correlations (67).

Sibling and Twin Analysis

With birth cohorts constructed from multiple years of administrative data, sibling and twin designs are possible (7, 11). Statistical analyses (e.g., family fixed-effects regression and
multilevel modeling) controlling for unobserved factors that lead to sibling correlations are particularly important when widely used measures such as individual household income and parental education are not generally available on a population-wide basis, as is the case with most administrative data (7, 63, 68). Although not true experiments, such analyses of sibling data represent one of the most powerful nonexperimental approaches (63). The future should see more sibling/parent designs linking parental and child histories to assess the effects of such conditions as attention-deficit/hyperactivity disorder on subsequent offspring (71).

**Life Course Research**

Life course research typically examines aspects of family background that may affect health and other outcomes into adulthood. Strohschein et al. (83) used the longitudinal nature of the Manitoba population registry to characterize family circumstances supportive of (or detrimental to) well-being over time. Building on this study, ongoing work is directed toward construction of a different history for each child using significant events (parental death, parental mental health issues, divorce, residential mobility, change of school) and the age at which each occurred. The ability to link such experiences to subsequent outcomes while accounting for interfamily differences is a powerful resource.

Several Australian states and Canadian provinces are collecting population-wide data using the Early Development Instrument, a holistic measure of children’s developmental status at kindergarten (25, 37, 56). An increasing interest in early life suggests the need for empirical estimates of the effects of various factors at different stages in child development (34); as linkages are performed and these younger cohorts age, the analytical possibilities will multiply.

Life course research merges into intergenerational research, raising many questions of interest across disciplines. In Western Australia, the Family Connections Genealogical Project has utilized the linkage of births, midwives, and marriage data to the core system. More than 10 research projects are now under way, analyzing the genealogy of up to three generations (28).

**CONCLUSION**

Investigators are engaged in many efforts to increase the scale and scope of administrative data links. As mentioned previously, efforts are now under way to expand the extensive work done in Western Australia to incorporate the remaining Australian states into the PHRN, thus encompassing the 21 million inhabitants of the continent. The United Kingdom is moving forward in similar efforts to merge hospital to hospital and hospital to death records for the entire population of 51 million (M.J. Goldacre, personal communication).

In addition to enlarging the scale of administrative data, work is now ongoing in multiple locales to increase the scope of information. In Manitoba, for example, databases on subsidized housing and from the Justice Ministry are being added to the Population Health Research Data Repository. Many sites are interested in links to clinical, radiological, and laboratory results. The incorporation of routine blood screens on newborns, pregnant women, and adults could also present new research opportunities. Genetic data—requiring collection only once over the lifetime—could also theoretically be incorporated, although an even higher level of scrutiny for privacy/confidentiality protections would be required.

The increase in scale and scope has allowed for expanded interdisciplinary research. Population-based studies seem especially suitable for combining the perspectives of economics with its “search for the closest approximation to an experiment” and those of social epidemiology with a focus on the “accretion and consistency of a broad range of research and evidence” (38, p. 12). The potential for increasing links among the work of researchers in public health and population health, education,
clinical medicine, housing, community development, and justice holds great promise.

The significance of these new insights has also influenced policy development. A review of projects using Western Australia’s linked data showed consequential policy and clinical practice reforms (13). An assessment of the impact of the Manitoba Center for Health Policy led to similar conclusions (50). Influences on projects as diverse as scaling back rural hospital construction plans, moving to a policy on the generic substitution of drugs, and focusing health planners on the mismatch between population need and services delivered were all recognized as positive outcomes from research utilizing the population data repository. Advancing research methodologies and increasing holdings of linked health and social data in ever more locales guarantee continued important contributions to investigations in public health and the social determinants of health and disease.

**SUMMARY POINTS**

1. Linked administrative databases are powerful resources that provide longitudinal health and social data on large populations for flexible and relatively low-cost investigation of pressing public health concerns.

2. Well-established research centers in several jurisdictions around the world have developed and implemented methods that achieve high-quality data linkages for conducting health and social research while maintaining individual privacy and confidentiality.

3. Information-rich repositories link individual-level longitudinal data from a variety of sources, including data on health services, population and disease registries, place of residence information, family-related social and structural variables, educational outcomes, and use of social services.

4. Linking events for large populations of individuals across disparate sources and over time permits a range of research possibilities including the capacity to study low-prevalence exposure-disease associations, different domains within the same cohort of individuals, individuals within a population who do not use services, chronic disease surveillance, population-based disease prediction, and life course and transgenerational patterns in health.

5. Common limitations of administrative data include a lack of individual-level measures of socioeconomic status, social supports, and nonfamilial interpersonal relationships.

6. Strengths of administrative data include comprehensive follow-up, inclusion of under-represented racial/ethnic and socioeconomic groups, reduced reliance on self-report, flexibility of study period due to continuously collected data, and relatively low expense because information is routinely collected for other purposes.

7. Concerns regarding privacy, confidentiality, and informed consent continue to challenge the use of linked administrative data despite rigorous security procedures and demonstrated public benefit.

8. Emerging methodologies and expanding data holdings guarantee that research using linked administrative databases will continue to make important contributions to the public health arena.
**FUTURE ISSUES**

1. The increasing use of sibling and twin analyses will allow for greater accounting of contextual variables unmeasured in population-wide data.

2. Multilevel modeling merits increased use in future studies to better utilize the strengths of administrative data derived from different levels (individual, family, school, community).

3. Sibling/parent research designs should be expanded to support the incorporation of multigenerational data into outcome analyses.

4. Further work is needed to improve the statistical manipulation of multisource, categorical data to generate population distributions more amenable to analysis and sensitivity testing.

5. Future efforts must continue toward creating links to existing population-wide clinical data such as radiological findings, laboratory results, and routine screening tests (e.g., newborn and prenatal blood screens).

**DISCLOSURE STATEMENT**

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

**LITERATURE CITED**


18. Deleted in proof


82. Souhami R. 2006. Governance of research that uses identifiable personal data will improve if the public and researchers collaborate to raise standards. BMJ 333:315–16
Contents

Symposium: Determinants of Changes in Cardiovascular Disease

Cardiovascular Disease: Rise, Fall, and Future Prospects
Russell V. Luepker ........................................................................ 1

Proportion of the Decline in Cardiovascular Mortality Disease due to Prevention Versus Treatment: Public Health Versus Clinical Care
Earl S. Ford and Simon Capewell ................................................ 5

Prospects for a Cardiovascular Disease Prevention Polypill
Kaustubh C. Dabhadkar, Ambar Kulshreshtha, Mohammed K. Ali, and K.M. Venkat Narayan ...................................................... 23

Social Determinants and the Decline of Cardiovascular Diseases: Understanding the Links
Sam Harper, John Lynch, and George Davey Smith ......................... 39

Sodium Intake and Cardiovascular Disease
Alanna C. Morrison and Roberta B. Ness ...................................... 71

Epidemiology and Biostatistics

Administrative Record Linkage as a Tool for Public Health Research
Douglas P. Jutte, Leslie L. Roos, and Marni D. Brownell ................. 91

Cardiovascular Disease: Rise, Fall, and Future Prospects
Russell V. Luepker ........................................................................ 1

Proportion of the Decline in Cardiovascular Mortality Disease due to Prevention Versus Treatment: Public Health Versus Clinical Care
Earl S. Ford and Simon Capewell ................................................ 5

Social Determinants and the Decline of Cardiovascular Diseases: Understanding the Links
Sam Harper, John Lynch, and George Davey Smith ......................... 39

Sodium Intake and Cardiovascular Disease
Alanna C. Morrison and Roberta B. Ness ...................................... 71
Prenatal Famine and Adult Health

L.H. Lumey, Aryeh D. Stein, and Ezra Susser ........................................... 237

Environmental and Occupational Health

Advances and Current Themes in Occupational Health and
Environmental Public Health Surveillance
Jeffrey D. Shire, Gary M. Marsh, Evelyn O. Talbott, and Ravi K. Sharma ........... 109

Climate Change, Noncommunicable Diseases, and Development:
The Relationships and Common Policy Opportunities
S. Friel, K. Bowen, D. Campbell-Lendrum, H. Frumkin, A.J. McMichael,
and K. Rasanathan .............................................................................. 133

Genetic Susceptibility and the Setting of Occupational
Health Standards
Paul Schulte and John Howard ............................................................. 149

New Directions in Toxicity Testing
Daniel Krewski, Margit Westphal, Mustafa Al-Zoughool, Maxine C. Croteau,
and Melvin E. Andersen .................................................................. 161

Promoting Global Population Health While Constraining the
Environmental Footprint
A.J. McMichael and C.D. Butler ......................................................... 179

Prenatal Famine and Adult Health
L.H. Lumey, Aryeh D. Stein, and Ezra Susser ........................................... 237

Public Health Practice

Accelerating Evidence Reviews and Broadening Evidence Standards to
Identify Effective, Promising, and Emerging Policy and
Environmental Strategies for Prevention of Childhood Obesity
Laura Brennan, Sarah Castro, Ross C. Brownson, Julie Claus,
and C. Tracy Orleans ........................................................................ 199

Action on the Social Determinants of Health and Health Inequities
Goes Global
Sharon Friel and Michael G. Marmot .................................................... 225

Prenatal Famine and Adult Health
L.H. Lumey, Aryeh D. Stein, and Ezra Susser ........................................... 237

The Growing Impact of Globalization for Health and Public
Health Practice
Ronald Labonté, Katia Mobindra, and Ted Schrecker ................................ 263
Using Marketing Muscle to Sell Fat: The Rise of Obesity in the Modern Economy  
*Frederick J. Zimmerman* ................................................................. 285

Cardiovascular Disease: Rise, Fall, and Future Prospects  
*Russell V. Luepker* ................................. ................................. 1

New Directions in Toxicity Testing  
*Daniel Krewski, Margit Westphal, Mustafa Al-Zoughool, Maxine C. Croteau,  
and Melvin E. Andersen* ................................. ................................. 161

Prematurity: An Overview and Public Health Implications  
*Marie C. McCormick, Jonathan S. Litt, Vincent C. Smith,  
and John A.F. Zupancic* ................................. ................................. 367

Proportion of the Decline in Cardiovascular Mortality Disease due to Prevention Versus Treatment: Public Health Versus Clinical Care  
*Earl S. Ford and Simon Capewell* ................................. ................................. 5

The U.S. Healthy People Initiative: Its Genesis and Its Sustainability  
*Lawrence W. Green and Jonathan Fielding* ................................. ................................. 451

**Social Environment and Behavior**

Ecological Models Revisited: Their Uses and Evolution in Health Promotion Over Two Decades  
*Lucie Richard, Lise Gauvin, and Kim Raine* ................................. ................................. 307

Environmental Risk Conditions and Pathways to Cardiometabolic Diseases in Indigenous Populations  
*Mark Daniel, Peter Lekkas, Margaret Cargo, Ivana Stankov,  
and Alex Brown* ................................. ................................. 327

*Kenneth E. Powell, Amanda E. Paluch, and Steven N. Blair* ................................. ................................. 349

Prematurity: An Overview and Public Health Implications  
*Marie C. McCormick, Jonathan S. Litt, Vincent C. Smith,  
and John A.F. Zupancic* ................................. ................................. 367

The Social Determinants of Health: Coming of Age  
*Paula Braverman, Susan Egerter, and David R. Williams* ................................. ................................. 381

Toward a Fourth Generation of Disparities Research to Achieve Health Equity  
*Stephen B. Thomas, Sandra Crouse Quinn, James Butler, Craig S. Fryer,  
and Mary A. Garza* ................................. ................................. 399
Action on the Social Determinants of Health and Health Inequities
Goes Global
Sharon Friel and Michael G. Marmot ............................................. 225

Social Determinants and the Decline of Cardiovascular Diseases:
Understanding the Links
Sam Harper, John Lynch, and George Davey Smith ................................ 39

Using Marketing Muscle to Sell Fat: The Rise of Obesity in the
Modern Economy
Frederick J. Zimmerman ................................................................. 285

Health Services

Prospects for a Cardiovascular Disease Prevention Polypill
Kaustubh C. Dabhadkar, Abnar Kulshreshtha, Mohammed K. Ali,
and K.M. Venkat Narayan ............................................................... 23

The Health Care Workforce: Will It Be Ready as the Boomers Age?
A Review of How We Can Know (or Not Know) the Answer
Thomas C. Ricketts ................................................................. 417

The Health Effects of Economic Decline
Ralph Catalanito, Sieda Goldman-Mellor, Katherine Saxton,
Claire Margerison-Zilko, Meenakshi Subbaraman, Kaja LeWinn,
and Elizabeth Anderson .............................................................. 431

The U.S. Healthy People Initiative: Its Genesis and Its Sustainability
Lawrence W. Green and Jonathan Fielding ........................................ 451

Underinsurance in the United States: An Interaction of Costs to
Consumers, Benefit Design, and Access to Care
Shana Alex Lavarreda, E. Richard Brown, and Claudie Dandurand Bolduc .......... 471

Administrative Record Linkage as a Tool for Public Health Research
Douglas P. Jutte, Leslie L. Roos, and Marni D. Brownell ...................... 91

Indexes

Cumulative Index of Contributing Authors, Volumes 23–32 ....................... 483
Cumulative Index of Chapter Titles, Volumes 23–32 ................................ 488

Errata

An online log of corrections to Annual Review of Public Health articles may be found at
http://publhealth.annualreviews.org/
New From Annual Reviews:

**Annual Review of Statistics and Its Application**

Volume 1 • Online January 2014 • http://statistics.annualreviews.org

Editor: **Stephen E. Fienberg**, Carnegie Mellon University

Associate Editors: **Nancy Reid**, University of Toronto  
**Stephen M. Stigler**, University of Chicago

The *Annual Review of Statistics and Its Application* aims to inform statisticians and quantitative methodologists, as well as all scientists and users of statistics about major methodological advances and the computational tools that allow for their implementation. It will include developments in the field of statistics, including theoretical statistical underpinnings of new methodology, as well as developments in specific application domains such as biostatistics and bioinformatics, economics, machine learning, psychology, sociology, and aspects of the physical sciences.

**Complimentary online access to the first volume will be available until January 2015.**

**TABLE OF CONTENTS:**

- A Systematic Statistical Approach to Evaluating Evidence from Observational Studies, David Madigan, Paul E. Stang, Jesse A. Berlin, Martijn Schuemie, J. Marc Overhage, Marc A. Suchard, Bill Dumouchel, Abraham G. Hartzema, Patrick B. Ryan
- The Role of Statistics in the Discovery of a Higgs Boson, David A. van Dyk
- Brain Imaging Analysis, F. DuBois Bowman
- Statistics and Climate, Peter Guttorp
- Climate Simulators and Climate Projections, Jonathan Rougier, Michael Goldstein
- Probabilistic Forecasting, Tilmann Gneiting, Matthias Katzfuss
- Bayesian Computational Tools, Christian P. Robert
- Bayesian Computation Via Markov Chain Monte Carlo, Radu V. Craiu, Jeffrey S. Rosenthal
- Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models, David M. Blei
- Structured Regularizers for High-Dimensional Problems: Statistical and Computational Issues, Martin J. Wainwright
- High-Dimensional Statistics with a View Toward Applications in Biology, Peter Bühlmann, Markus Kalisch, Lukas Meier
- Next-Generation Statistical Genetics: Modeling, Penalization, and Optimization in High-Dimensional Data, Kenneth Lange, Jeannette C. Papp, Janet S. Sinsheimer, Eric M. Sobel
- Breaking Bad: Two Decades of Life-Course Data Analysis in Criminology, Developmental Psychology, and Beyond, Elena A. Erosheva, Ross L. Matsueda, Donatello Telesca
- Event History Analysis, Niels Keiding
- Statistical Evaluation of Forensic DNA Profile Evidence, Christopher D. Steele, David J. Balding
- Using League Table Rankings in Public Policy Formation: Statistical Issues, Harvey Goldstein
- Statistical Ecology, Ruth King
- Estimating the Number of Species in Microbial Diversity Studies, John Bunge, Amy Willis, Fiona Walsh
- Dynamic Treatment Regimes, Bibhas Chakraborty, Susan A. Murphy
- Statistics and Related Topics in Single-Molecule Biophysics, Hong Qian, S.C. Kou
- Statistics and Quantitative Risk Management for Banking and Insurance, Paul Embrechts, Marius Hofert

Access this and all other Annual Reviews journals via your institution at [www.annualreviews.org](http://www.annualreviews.org).