The Utah Population Database. A Model for Linking Medical and Genealogical Records for Population Health Research

By Ken R. Smith, Alison Fraser, Diana Lane Reed, Jahn Barlow, Heidi A. Hanson, Jennifer West, Stacey Knight, Navina Forsythe and Geraldine P. Mineau

To cite this article: Smith, K. R., Fraser, A., Reed, D. L., Barlow, J., Hanson, H. A., West, J., Knight, S., Forsythe, N., & Mineau, G. P. (2022). The Utah Population Database. A Model for Linking Medical and Genealogical Records for Population Health Research. *Historical Life Course Studies, 12*, 58–77. https://doi.org/10.51964/hlcs11681

HISTORICAL LIFE COURSE STUDIES

Content, Design and Structure of Major Databases with Historical Longitudinal Population Data

VOLUME 12, SPECIAL ISSUE 5, 2020

GUEST EDITORS
George Alter
Kees Mandemakers
Hélène Vézina

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at hlcs.nl.

Co-Editors-In-Chief:
Paul Puschmann (Radboud University) & Luciana Quaranta (Lund University)
hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

HISTORICAL LIFE COURSE STUDIES
VOLUME 12 (2022), published 03-05-2022

The Utah Population Database

A Model for Linking Medical and Genealogical Records for Population Health Research

Ken R. Smith	University of Utah
Alison Fraser	University of Utah
Diana Lane Reed	University of Utah
Jahn Barlow	University of Utah
Heidi A. Hanson	University of Utah
Jennifer West	University of Utah
Stacey Knight	Intermountain Healthcare, Salt Lake City, Utah
Navina Forsythe	Utah Department of Health
Geraldine P. Mineau	University of Utah

ABSTRACT

Improving our understanding of the socio-environmental and genetic bases of disease and health outcomes among individuals, families, and populations over time requires extensive longitudinal data on multiple attributes for entire communities, states or nations. This requirement can be difficult to achieve. In this paper we describe a successful example of a database that meets these needs. The Utah Population Database (UPDB) is a unique and powerful database rarely found in the world that has been addressing these data requirements for over 40 years. The UPDB at the University of Utah is one of the world's richest sources of in-depth information that supports research on genetics, epidemiology, demography, history, and public health. Genetic researchers have used UPDB to identify and study individuals and families that have higher than normal incidence of diseases or other traits, to analyze patterns of genetic inheritance, and to identify specific genetic mutations. Demographers and other social scientists are increasingly using the UPDB to study issues such as trends in fertility transitions and shifts in mortality patterns for both infants and adults. A central component of the UPDB is an extensive set of Utah family histories, in which family members are linked to demographic and medical information. The UPDB includes medical information about cancer, causes of death, and medical details associated with births. It also includes diagnostic records from statewide insurance claims data and healthcare facilities (hospital discharge, ambulatory surgery, emergency department encounters). UPDB is also linked to Medicare claims data, a federal health insurance program generally for persons age 65 or older. The UPDB provides access to information on more than 11 million individuals and supports nearly 400 research projects. We describe in detail the data components of the UPDB, how it can be accessed, issues related to its development, record linkage, governance and privacy protections, as well as plans for future developments.

Keywords: Historical demography, Demography of Utah, Record linking, Administrative records, Data privacy, Genetics

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs11681

1 INTRODUCTION

A strategy for understanding human health and well-being is to collect and curate extensive data on large, well-defined populations and all of its members over time with the appropriate data and privacy safeguards. Several successful (some longstanding) examples of these databases exist, many of which have been fundamental contributors to key medical and social science discoveries. The Framingham Heart Study in the US and the National Survey of Health and Development in the UK are exemplary in this respect.

In this paper, we describe a unique resource, the Utah Population Database (UPDB), which offers exceptional and unique data and research opportunities for population scientists, demographers, epidemiologists, historians, geneticists, health services researchers, and behavioral scientists, among others, all of whom work on population health and medical research. A distinctive quality of the UPDB is that it is based on links at the individual-level of administrative and medical records derived from a range of sources spanning decades for some sources and centuries for others (Casey, Schwartz, Stewart, & Adler, 2016; Hurdle, Smith, & Mineau, 2013) with many sources updated up to the present. The individuals with linked records comprising life histories are in turn linked to their family members, a feature of UPDB that allows analysts to study families, shared and unshared environments, and genetic associations over many generations. Moreover, these linkages create up to 17 generations and the concept of family can be expanded such that many individuals are frequently connected to tens of thousands of relatives by blood or marriage. Members of these multi-generational pedigrees have extensive event and date information but also spatial attributes at varying levels of geographic detail. These latter data elements allow projects to link geo-coded data to the UPDB in order to investigate environmental exposures as well as factors related to propinquity such as the geographic distance separating relatives or travel time needed to access a hospital.

This article is structured around central features and characteristics of UPDB's history and its structure and management. The paper starts with the specific components that comprise it today and the historical circumstances that led to their inclusion in UPDB. We then describe our conceptual data model and how and why UPDB has been able to thrive and grow for so many decades, in short, due to consistent institutional commitments. This is followed by a section devoted to details about the record linking methods used to create UPDB. Given the sensitive nature of the data in UPDB and the need to maintain the highest level of data security, we describe the regulatory protections of the data and how research access to the data can be obtained. Confidentiality and privacy issues are discussed in the context of UPDB as well as how these matters relate to UPDB's relationship to the many agencies which provide data. The paper ends with final thoughts and directions for the future.

2 DATA

2.1 SOURCES

UPDB was established over 40 years ago and has been a premiere research resource that had the early vision to integrate genetics and the social sciences. Selected key dates representing important developments in the evolution of UPDB are shown in Figure 1. Its beginnings can be dated to the years 1973–1974 when several researchers at the University of Utah realized the research opportunities that could be gained by first obtaining extensive genealogy records and constructing a population-based resource that would link these genealogical data to high quality medical records in order to investigate the genetic basis of a number of important diseases. Central to the launch of UPDB was geneticist Mark Skolnick, who was recruited to the University of Utah to lead a computerization of family history records with links to medical records. He then created an initial consortium of two additional key scientists, cardiologist Roger Williams and demographer Lee L. Bean.

Figure 1 *Utah Population Database (UPDB) - Selected events in the history of UPDB*

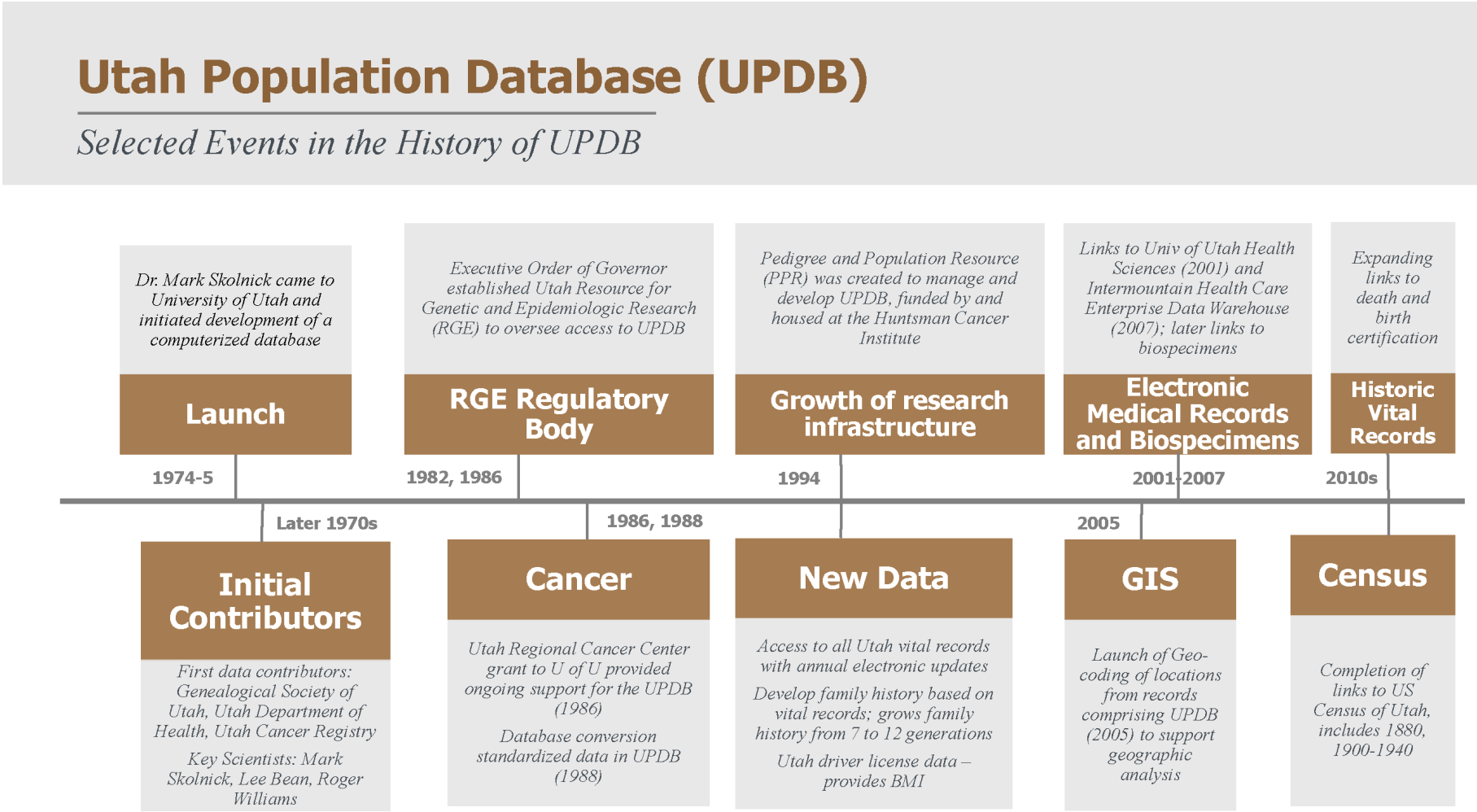


Table 1 *Population of Utah between 1850 and 2020*

Year	Population
1850*	11,380**
1860	40,273**
1870	86,786
1880	143,963
1890	210,779
1900#	276,749
1910	373,351
1920	449,396
1930	507,847
1940	550,310
1950	688,862
1960	890,627
1970	1,059,273
1980	1,461,037
1990	1,722,850
2000	2,233,169
2010	2,763,885
2020	3,249,879

* *Members of the Church of Jesus Christ of Latter-day Saints arrive in Utah July 24, 1847.*

** *Population of the Territory of Utah which included parts of present-day states of Colorado, Nevada, and Wyoming.*

\# *Utah granted statehood January 4, 1896.*

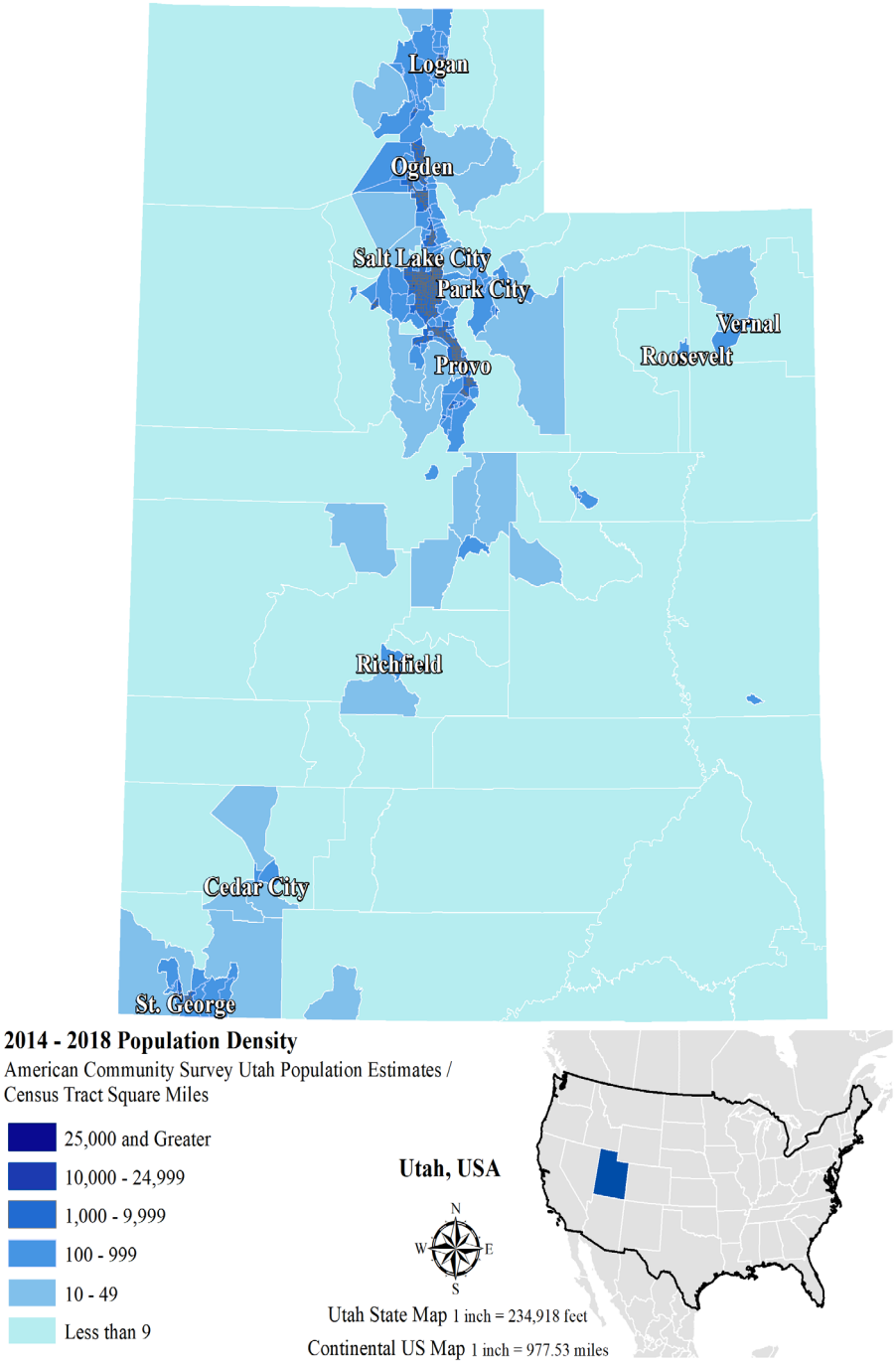
Sources: U.S. Census Bureau and Utah History Encyclopedia

UPDB is largely derived from records pertaining to events in Utah although connections to events outside the state are included when available, as described later. Utah is the 11th largest state in the US and has a median household income of $71,414 (2018). Between 2020 and 2021 the population experienced a 1.8% increase. In Table 1 we show the growth of the population of Utah since its settlement in 1847, through the time Utah was admitted to statehood in 1896, and up to the present. Based on the 2020 US Census of Utah, Utah is comprised of 90.6% white (of which 77.8% are non-Hispanic), 1.5% African American/Black, 1.6% American Indian/Alaska Native, 2.7% Asian, 1.1% Native Hawaiian/Pacific Islander, and 2.6% two or more races; 14.4% are Hispanic or Latino. While Utah has 33.6 residents per square mile, it has the 8th highest percentage of people living in urban areas (2010 US Census) among the 50 states. Figure 2 illustrates how Utah has a low population density and high urbanization levels. Utah has a period life expectancy at birth of 79.9, which is above the US figure of 78.7 (2018).

The original set of genealogy records used when the UPDB was being developed comprised approximately 185,000 documents representing, on each form, three generations: a husband-wife pair, their four parents, and the couple's offspring and their respective spouses. These initial documents were selected to represent approximately 1.9 million individuals. Linking these across generations (e.g., a child in one group sheet is a parent on another) creates thousands of multi-generational pedigrees, providing astonishing insights regarding the population (Song & Campbell, 2017).

These early genealogical records comprise the original backbone of UPDB. The founding research team secured access to the Utah Cancer Registry (UCR, a Surveillance, Epidemiology and End Result (SEER) Registry) and Utah death certificates (from the Utah Department of Health) as the basis for medical outcomes to be linked to the genealogies at the individual level. Accordingly, many of the early studies focused on cancer based on these cancer records (Skolnick et al., 1981), as well as cardiovascular mortality (Williams et al., 1979) and demographic studies (Bean, May, & Skolnick, 1978; Skolnick et al., 1978) based on death and genealogy records (Skolnick, Bean, Dintelman, & Mineau, 1979).

Figure 2 *Population density and map of Utah*



The UPDB is a research resource that has been expanded extensively in its 40 years of existence. At this time, UPDB includes information on approximately 11 million individuals who have basic demographic information and is a data source for nearly 400 research projects. The time period within UPDB covers birth cohorts from the 1700s but are more extensive starting in the mid-1800s and run through the present. Using the UPDB Query tool (https://uofuhealth.utah.edu/huntsman/utah-population-database/services/query.php; requires registration) in December 2021, we show there are 304,104 individuals in UPDB born before 1847 (the year members of the Church of Jesus Christ of Latter-day Saints first arrived in Utah), 975,081 born between 1847–1899, 2,491,970 born between 1900–1949, 2,494,871 born between 1950–1974, 2,814,316 born between 1975–1999, and 1,734,962 born between 2000–2020, the latest update.

While the original development of UPDB was derived from three sources (genealogy, cancer record and death records), the UPDB now includes substantially more records and from diverse sources (see Table 2). These are fully described at https://uofuhealth.utah.edu/huntsman/utah-population-database/.

Table 2 *Records available in the Utah Population Database*

Record Type	Years Available	Notes	Records
Original Family History Records	1700's–1975	The original genealogical portion of UPDB holds Utah family histories organized into pedigrees based on Genealogical Society of Utah documents that hold demographic/kinship data.	1,917,111
UTAH VITAL RECORDS			
Birth Certificates	1915–1921, 1926–2020	Data on parents and children and their demographic medical information; volume of data varies by year.	3,162,090
Death Certificates	1904–2021	Causes of death are coded using International Classification of Diseases (ICD) revisions 6–10.	982,662
Marriage Certificates	1978–2010	Husband and wife name and age, marriage date, and county of marriage; (1988+): birth date and birth place, education, number of marriages, and type of marriage (civil/religious).	692,838
Divorce Records	1978–2010	Husband and wife name, marriage and divorce dates, and county where the divorce was issued; (1988+): birth dates and education of the husband and wife, number of marriages, number of children, and number of children under age 18.	298,928
Fetal Deaths	1978–2020	Stillbirths/fetal deaths of 20 weeks or greater gestation as calculated from the mother's last normal menses period to the date of delivery.	11,933
MEDICAL RECORDS			
Ambulatory Surgery Utah	1996–2020	Diagnosis and procedure codes and external injury E-codes.	12,342,203
Inpatient Hospital Claims Utah	1996–2020	Diagnosis and procedure codes and external injury E-codes.	6,696,825
Emergency Department	1996–2020 (older records forthcoming)	Diagnosis and procedure codes and external injury E-codes.	16,167,073
All Payer Claims Data	2013–2020	The APCD data captures medical and financial information for nearly all encounters involving 3rd party payers.	>200,000,000
Utah Cancer Registry	1966–2019	UCR is a statewide cancer registry that monitors cancer incidence & mortality. It participates in the NCI Surveillance, Epidemiology, and End Results (SEER) Program.	420,185
Birth Defect Network	1995–2018		23,910
ADDITIONAL RECORDS			
U.S. Census of Utah	1880,1900–1940	Individual-level records provide a range of data including SES, household composition, migration, literacy, and neighborhoods.	2,300,084
Social Security Death Index	Last updated 2011	Date and state of deaths regardless of their place of death.	581,373
Utah Driver License Division	Last updated 2021	DLD has residential data for all Utah drivers. DLD is also a good source for height and weight, or BMI.	4,175,080
Utah Voter Registration	Last updated 2020	Variables include residential information; updated during presidential election years.	2,251,922
TOTAL			>252,024,217

Externally Linked Records — Demographic records of external records that are linked to UPDB but substantive variables are held by data provider until investigators obtain IRB Approval			
Record Type	**Years Available**	**Notes**	**Individuals**
University of Utah Health Sciences	1992–Current	All University of Utah inpatient and out-patient clinics. Demographic records and medical and pharmacy information.	3,232,154
Intermountain Healthcare (IH)	1992–Current	All IH hospitals inpatient and out-patient clinics. Demographic records and medical and pharmacy information.	8,141,654
Centers for Medicare & Medicaid Services	1992–2012	Diagnostic, procedure and other risk factor data. 18 years of data linked to UPDB; new links are underway for more recent records.	700,000
Utah Department of Human Services (DHS)	1995–2020	DHS comprises 12 divisions including the child-serving Divisions of Child and Family Services (DCFS), Juvenile Justice Services (DJJS) and Services for People with Disabilities (DSPD). Data include records related care as supervision, wraparound services while in custody, therapeutic services, and in-home services.	616,894

Briefly, in addition to the original genealogy records, UPDB includes:

1. All electronically available Utah vital records (births, deaths, marriages, divorces and fetal deaths) from the Utah Department of Health from 1904 at the earliest onwards depending on the certificate.
2. Statewide health data from the Utah Department of Health including:
    a. Ambulatory Surgery records which contain medical, financial and diagnostic information regarding visits occurring at designated surgical out-patient units;
    b. Inpatient Hospital Discharge records which contain medical, financial and diagnostic data upon discharge from a hospital as an inpatient;
    c. Emergency Department records which describe the medical and diagnostic information about the emergency visit;
    d. All Payer Claims which hold data on medical, financial, diagnostic and pharmacy data that involve claims to a third-party health insurance provider;
    e. Utah Cancer Registry data from the Utah Department of Health which hold statewide medical data on all incident cancer diagnoses except non-melanoma skin cancer;
    f. Utah Birth Defect Network is a statewide, population-based surveillance system that identifies birth defects in children born in Utah since 1994; UPDB has data up to 2018.
3. Social Security Death Index records which provide place and date of death for persons who have ever been enrolled in the Social Security system.
4. Utah Voter Registration which provides information about whether still living in Utah.
5. Utah driver license records which contain data on spatial information on the place of residence as well as height and weight.
6. The 1880 and 1900–1940 Utah individual-level censuses.

UPDB is noteworthy with respect to linkages to other large federated medical data sets (i.e., links to UPDB but the information does not reside within UPDB). First, a "master subject index" or MSI has been created that links the UPDB with the demographic records from the Enterprise Data Warehouses (EDWs) of the two largest health providers in Utah: University of Utah Health Sciences and Intermountain Healthcare (DuVall, Fraser, Rowe, Thomas, & Mineau, 2012). These linkages are based on demographic information only and do not involve in any way medical, treatment or diagnostic information for the purposes of record linking. These two health care providers represent

inpatient and outpatient electronic medical information for approximately 85% of the state's medical encounters starting from the mid-1990s. These medical data are not held within the UPDB but are securely maintained by the enterprise data warehouses of these health care providers. Medical data are joined with the demographic and genealogical data in UPDB after the research project receives the necessary approvals from appropriate Institutional Review Boards (IRB) and the Utah Resource for Genetic and Epidemiologic Research (RGE), which oversees research access to the UPDB as described below.

A third and related medical data linked to the UPDB are those derived from Medicare claims, a federal health insurance program generally for persons age 65 or older. The Medicare data are available due to funding from National Institutes of Health (NIH) grants that were originally designed to facilitate the study of healthy aging and health expectancy among the Medicare-eligible (age 65 or older) population. These data relate to claims from 1992–2015 and more recent years are being added, if an individual had a claim at some point in Utah. These data are available to researchers beyond its original purposes using the UPDB but they must not only obtain IRB approval for their use but also approval from the federal Centers for Medicaid and Medicare Services (CMS).

Another data set linked to the UPDB stems from the Department of Human Services (DHS). The DHS data represent nearly all persons in Utah identified by the DHS including those using Aging and Adult Services, Child and Family Services, and Juvenile Justice Services. The linking also uses the "master subject index" methodology. Demographic data of included persons are provided for linking without any indication of the reason regarding their inclusion in the dataset. With DHS, RGE, and IRB approval, all requested information in each DHS purview is provided to the researcher.

2.2 RESEARCH OPPORTUNITIES

Linking these records within UPDB creates diverse types of datasets with unique research opportunities including:

1. Creation of reproductive histories

Using data from the Utah Department of Health that includes Utah birth certificates from 1915 to the present, we have extended the genealogical holdings of UPDB considerably. Information for the same mother and/or the same father on multiple birth certificates are linked, a technique similar to family reconstitution. This allows us to see that specific individuals share common parents and are therefore siblings. The children named on these birth certificates (the second generation) are then linked to the birth certificates of their children (the third generation, that is the grandchildren of the first generation). This provides an efficient and non-biased approach for representing the current Utah population as these families propagate the next generation. For the many families that remain in the state over their reproductive years, a complete history is possible. Moreover, this strategy creates broader genealogies connecting individuals more distantly related. Because birth certificates provide gestational age and birth weight as well as other features such as adverse obstetric events and birth complications, this strategy has provided a valuable source for analysis of preterm births, cesarean sections and preeclampsia in families and across generations (Hammad et al., 2020; Theilen et al., 2016, 2018). It is noteworthy that many of the genealogies derived from vital records also link into the legacy genealogies that are part of the UPDB. Note that this strategy is restricted to births visible on Utah birth certificates. Certainly, instances exist where a woman or a couple will bear children in Utah and others who were born elsewhere. Some data about past fertility patterns are represented on each birth certificate such as the number of previous pregnancies and live births (the availability of these data varies by birth year). This type of "retrospective" information from birth certificates is captured in UPDB but is not useful for constructing and expanding genealogies since the identities of these previous offspring born elsewhere are not known via birth certificates. Note that other sources of data, such as from the Genealogical Society of Utah or death certificates in UPDB that identify other offspring are used whenever possible.

2. Creation of residential exposures and histories.

Residential location information is derived from several sources in UPDB including Driver License Division (DLD) data, voter registrations, and vital records, while other records provide location information at a higher level of geographic aggregation such as the ZIP code. One use of DLD is to provide current residence status for individuals in UPDB. In this way a researcher is able to determine if an individual

is currently under observation, while residence information on death records verify if an individual was under observation until their death. This helps with generating population denominators. Every four years after major federal elections, Voter Registration records are obtained and linked to UPDB which give geographic information at a particular point in time. Additionally, DLD data hold information on height and weight from which we have derived the Body Mass Index (BMI) for each individual (Chernenko, Meeks, & Smith, 2019; Smith et al., 2008; Smith et al., 2011; Zick et al., 2009). Addresses from any of its sources used to derive residential histories within UPDB have been geo-coded when sufficient address information is available in the source records. This creates the opportunity for linking any geo-referenced data set (e.g., census block, air quality monitors) with individual-level data. These residential histories can capture important points in the life history of an individual from mother's residence at birth (own birth certificates), residence in childhood (birth certificates of latter born siblings), place of residence of offspring (children's birth/fetal death certificates), adult locations (census records, DLD, voter registration, health facilities data), and death (own or spouse: death certificates). Finally, residential histories described here refer to places within the state of Utah. Some records, including data from the Genealogical Society of Utah and the U.S. Census of Utah, contain information about locations for individuals outside the state of Utah. These may refer to places that precede or follow a period of time when an individual was living in Utah. In addition to that, to deal with potential selective migration into and out of Utah, UPDB staff have created date variables that mark the point in time when we first saw individuals and when we last saw them in the state of Utah, subject to the data availability within UPDB. For minors who do not yet vote or drive, mother's information is used. When possible, prior or subsequent specific locations are available in UPDB but even when they are unavailable analysts may use our entry and exit dates in order to adjust for possible selection bias. For analyses that span historic periods covered by U.S. Census records linked to UPDB, decennial sightings of locations are available whether or not they occur within Utah's boundaries.

3. Creation of Links with Individual-Level Census Records

The addition of the micro level census records from 1880 and 1900–1940 (and those to come) to UPDB now allows for several types of studies. First, it is now possible to observe mobility, both geographic and socioeconomic, and its causes and consequences. Seeing the population before the censuses and decades after the last one in 1940 enable investigators to see how personal fortunes (or penury) during these early years as reflected in the Census are associated with later life health and well-being. Second, given the manner in which census enumerators were assigned to districts to conduct the full count of the population, the data can be used to cluster individuals into neighborhoods. Accordingly, individuals identified in the census can be characterized by the quality of their 'neighborhoods' and how these spatial attributes may alter later life outcomes. These census records provide valuable independent information about family composition, co-residence, and genealogical data that may not be possible from other sources of data in the UPDB. Again, in terms of residential history, the censuses add value since they provide information about birthplace (important for the 19th century since Utah was greatly affected by international in-migrants) and in some cases (1910 Census) the year of entry to the US.

4. Creation of a Life Course Dataset to measure adversity and opportunity over time

With administrative data linked over many decades, the possibility of conducting life course analysis at the population level grows substantially. Since UPDB holds data from its earliest years in the 18th and 19th centuries up to the present, it is possible to see entire life spans within individuals and across generations. Apart from linking basic demographic and genealogical connections, UPDB annotates these records with information from vital records starting in the early 20th century, adds micro-level census information from 1880–1940, introduces cancer incidence information in the mid-1960s and then grows to include more medical data from the mid-1990s to the present. Family connections and geographic information exist throughout these years, though the spatial data vary in terms of their geographic resolution given the type of records available in a given period. UPDB has been the basis for the Demographic Child Adversity Exposure (DECADE) scale (Hollingshaus, 2015) which measures how challenges early in life may be associated with serious health outcomes, such as suicide, later in life. For contemporary years where data-rich birth certificates are available, other indicators of socioeconomic status include education (after 1968) and occupation (all years through 2008) of the parents are available in UPDB, along with marital status. This enables investigators to see children born into single-parent households or to find same-sex unions (for the latter, this has only been possible in recent years and is under development). These variables can be used to examine the effects of early life adversity on life courses for modern decades (Stroup et al., 2017).

5. Creation of Datasets with Links to External Datasets

UPDB also has the capacity to link its data to ongoing projects that have arisen independent of UPDB. For example, the Cache County Memory and Health Study was launched in 1995 to study factors related to dementia and Alzheimer's disease risk. Participants were 65 and older at enrollment and were from a single county in Northern Utah. With the appropriate approvals, all were linked to UPDB. This linkage provided an opportunity to open up new life course studies of dementia and Alzheimer's disease (Norton et al., 2010, 2011, 2016).

In the end, the diversity of data sources and the annual updates of many of the data sources has created a resource in UPDB that includes nearly all of the residents of Utah. An assessment of the number of people alive and living in Utah in 2010 based on US census estimates shows close agreement with those represented in the UPDB.

3 CONCEPTUAL MODEL

A fundamental goal of the UPDB is to preserve the integrity of the data in the form in which it was received and yet create a set of unique individuals which can easily be used for record linking, statistical analysis and pedigree construction. Therefore, each dataset added to UPDB and all individuals listed on the records from those distinct datasets are assigned a unique dataset-specific identification number while relationships are created between individuals, such as a husband and wife on a marriage record or parent and child on a birth certificate. Information that is unique to each data source and time period is stored together in a separate dataset such as the manner of death on a death certificate or birth weight on a birth certificate. Major format changes to vital records with different information collected result in separate datasets. Personal information that is common across many data sources and is used in the matching process, is stored together in other datasets, including, but not limited to, demographic information, names, places, addresses, and relationships.

Initially, as each individual is loaded into UPDB, they exist with all their original information (archival information) and they also exist as a "composite" person, but the composite person only reflects the data originally received. After at least two or more persons are determined to be the same individual via the linking process, to facilitate further record linking and analysis of the data, the unique "composite" person record is re-created for that individual. The composite person is created by using a rules-based program which evaluates discrepant information. So, this rules-based program is only used after a link between two persons has been established to determine the most accurate and current name, demographic and relationship information of an individual based on the frequency and source of information. The objective is to create a person-oriented data structure where the person-specific information is selected in order to construct as complete as possible the life history of the individual from the many streams of data representing that individual. There are instances where the source data comprising the individual's life seems correct based on information at a given time but are deemed in error (or at least some portion) based on new information that subsequently comes to light as new records are added and linked. In this way, the person-oriented model is dynamic as new data are added to UPDB.

The durability, sustainability, and success of UPDB can be attributed in large measure to several factors outside the UPDB structure. First, complex and large linked databases such as UPDB are understandably expensive to build and maintain. In this instance, the Huntsman Cancer Institute has provided support to the Pedigree and Population Shared Resource (PPR) since the mid-1990s. This institutional foundation has given the PPR staff the stability it needs to engage in rational planning and support growth. This institutional basis has also been supplemented by the University of Utah beyond that provided by the Huntsman Cancer Institute. This funding model has been essential for the growth and the quality of UPDB. It also permits individual investigators to propose studies using UPDB data where the infrastructure costs have been largely paid by the institution (through philanthropic giving from the Huntsman Cancer Foundation and returned overhead to the University of Utah from extramural grants). Accordingly, research grants can accommodate the project-specific costs associated with PPR expenses through support from federal agencies. This has been a successful model given the large number of extramurally funded grants awarded to investigators using UPDB data.

A more subtle but important aspect of UPDB's success relates to the relatively small size of Utah's population and institutions. Utah's small population at the outset in the mid-1970s likely contributed to its inauguration. The volume of data was more manageable and the ability of the principal institutions to interact was conducive to creating a collaborative atmosphere between the key institutions (the Genealogical Society of Utah, the University of Utah, and the Utah Department of Health). The geographic proximity of these institutions contributed to negotiations and agreements that would likely have been more problematic in much larger states.

The growth and evolution of investigators and topics reliant on UPDB can in part be attributed to the catalyzing effects of big data on team science (Sellers et al., 2006; Shah, Pico, & Freedman, 2016; Stokols, Misra, Moser, Hall, & Taylor, 2008). The diversity and quality of UPDB data that is curated and made available has served to induce large and ambitious projects that require investigators from multiple disciplines. This has created teams that often combine medical, population and social sciences. Such multidisciplinary efforts generally serve to make the science stronger and have served to make UPDB essential to the larger research mission of the University of Utah.

4 RECORD LINKING

4.1 OVERVIEW

The linking process is fundamental to the core purposes of the UPDB, its utility, the representation of the diverse data sets it comprises, and the structure and scope of the pedigrees it contains. The objective is to identify efficiently the same individuals across millions of records historically as well as with each scheduled update of new records. The "composite" person is created using available identifying information including (when available) full name, birth date, death date, addresses, phone numbers, place of birth or death, encrypted Social Security Number, and names and specific relationships of family members.

Linking is accomplished primarily using probabilistic techniques supplemented by deterministic linking and manual linking as a result of manual review. The probabilistic linking software used for UPDB has evolved over time, from a command line program called Automatch using probabilistic linkage techniques based on Howard Newcombe's seminal work (Fair, Lalonde, & Newcombe, 1991; Newcombe, 1969; Newcombe, Kennedy, Axford, & James, 1959) to the current linking software called QualityStage, IBM's Websphere Information Integration Solution™ family of tools and applications (IBM, Armonk, NY, USA). QualityStage draws on information theory and advanced pattern recognition features to provide the highest level of automation for standardization and matching (Duvall et al., 2012).

For some data sources, the information is insufficient to use with probabilistic linking techniques. For example, Ambulatory surgery records may not provide names but only contain encrypted Social Security number, birth date, gender and ZIP code; in this situation, deterministic linking is used. Also, if a child on a birth certificate is linked to his or her own death certificate using probabilistic methods, then the parents listed can be linked with assurance using deterministic methods with only their names. The principles of using a combination of probabilistic and deterministic linking techniques with systematic validation supplemented with hand edits as needed has remained constant for UPDB record linkage throughout the database's existence. Validation processes external to QualityStage are used which may result in manual review of potential links.

The process of UPDB record linkage begins with the receipt of new records that are scheduled to be transferred to the UPDB team every six months or every year from the data contributors. The fact that UPDB adds contemporary data (perhaps with a one to one and a half year lag from the date of the record, a lag induced by internal data processing required by the data contributors to achieve their original mandate) is a central strength of the UPDB for studying contemporary outcomes which may be linked to more distant historical circumstances. This tempo of creating up-to-date information on living individuals within UPDB and building their life histories is attractive to researchers from a range of disciplines who are often focused on past causal events that may affect current responses (e.g., diagnosed with COVID-19).

4.2 INFORMATION STANDARDIZATION

All source records are securely transferred and loaded onto UPDB servers. Individuals in these data sets are represented by records created in UPDB, assigned an ID specific to the source (which is distinct from the person-level ID for the composite person), and the variables in these records are standardized according to UPDB protocol. For example, Social Security numbers are encrypted, punctuation and spaces are removed from within names, street addresses are standardized to match the US Postal Service standards (e.g., Street is abbreviated to ST) while cities, counties, states and countries are matched against a UPDB-specific dictionary to remove common spelling errors and abbreviations. When a record is received with multiple individuals identified with their family relationships, such as on a birth certificate (child, mother, father), a distinct record for each individual on this record is created in UPDB, assigned an ID number and these genetic or marital relationships between the individuals are indicated. Each individual record exists with all the information that was received initially in the source (archive) record; some of this information is also maintained in the record of the "composite" individual or Person Record.

The information that is used for record linking is selected from the tables of standardized information for the "composite" person and a single file is created. This file contains sex, names (original and Soundex), birth dates, death dates, birth place, death place, whether the individual is a twin/triplet, encrypted Social Security numbers, current address, and phone number of the individual. Also included are parent's name and their encrypted Social Security number as well as spouse's name, birth date, encrypted Social Security number, marriage date and birth date. Multiple records will exist for an individual with multiple spouses. For linking census records, information on the four eldest children are also included. All of these fields are available for use in the QualityStage program. Additional information can be used for validation, such as previous addresses for one person matching the current address of a potential link.

As additional records are loaded and linked, the validity and quantity of the information on a given person increases. In addition, social (e.g., marriage) and genetic (offspring) relationships are added and verified as multiple records containing relationship information are added and linked. An example of this arises in the case of names appearing on birth and death certificates. Parents who are listed on a death certificate can also be listed as parents on the decedent's birth certificates. When children on birth certificates are linked to their own death certificate, the two sightings of the parents (once on each vital record) are evaluated and linked if possible. In this example, a relationship may change from being identified as a birth parent relationship on a death certificate to an adoptive parent.

The resulting data are stored in tables in a relational database that cover a range of concepts or domains. These domains are numerous and include relationships (e.g., ego-mother-father), demographic features, medical diagnoses, insurance claims, birth/death details, residential history, and follow-up information. This domain-oriented data structure may draw on information from a variety of sources or they may be based on a single source. Tailored datasets approved for analysis will be created from multiple domains.

4.3 GENERAL LINKING STRATEGY

Potential links are initially created based on exact matching on one or more fields. These fields are called blocking variables or fields which create a subset of records that satisfy this exact matching. For example, the combination of encrypted Social Security number and birth year or the combination of first name, last name and birth year that exactly match across two records would be assessed and appear in the subset. Many different blocking combinations are used to account for instances where only one character may differ in a name or a birth year may differ by a single year.

Within sets of blocked records, statistical weight (affecting soundex) are then calculated for fields with sufficient variability (e.g., last name, first name, birth place, mother's maiden name) and used to measure the contribution of these fields to the probability of matching two records accurately. These weights are an extension of the Fellegi and Sunter algorithm (Fellegi & Sunter, 1969) developed by Jaro (Jaro, 1995; see also DuVall, Kerber, & Thomas, 2010) and derived from probabilities that utilize the frequency of the distinct values in the field which are generated by QualityStage. These field-specific agreement weights are based on the probability that the field agrees given that the records are true matches (m-probability) and the probability that the field agrees given that the records are

not true matches (u-probability). The combination of the m- and u-probabilities form the basis of the weight assigned to the matching for a given field. If the fields do not match, a disagreement weight is assigned dependent on parameters that define the likelihood of a mismatch given that the two records are true matches; therefore, knowledge of the data and their quality can be incorporated into the weights. A positive weight for comparison of mismatched variables may be assigned by evaluating the similarity of the two strings using an algorithm that is based on information theory principles. A composite weight is computed by summing the distinct (dis)agreement component weights of each variable comparison. Threshold values are used to classify a "good" link if the composite weight is above a threshold value, a nonmatch if it is below lower threshold, and undecided otherwise (these undergo manual review or further validation with other variables). A series of passes over the data are performed using different combinations of blocking and matching fields. Relationships between individuals can also be utilized for blocking. Several illustrative examples include:

- The first name on one record may be blocked with the middle name from another record; the last name on one record of a woman may be blocked with the spouse's last name from another record where the woman is recorded with only her maiden name.
- To address potential keying errors for date fields, birth day on record A is blocked with birth month on record B and birth month on record A is blocked with birth day on record B.
- Relationships that exist between parent and children as well as spouses allow for blocking on their names.
- The US Census of Utah records have been linked to genealogy records using blocking on parent's name and four of their eldest children's names or with just a single relationship between child – mother, child – father or husband – wife.

Each set of candidate links are processed through a set of validation checks before being incorporated into the UPDB. These validation programs identify links that may need additional attention and manual review. The set of links to be reviewed indicate the check that generated the need for additional attention and the source of the inconsistency. These checks include general logical inconsistencies, such as children born before parents, born to implausibly young parents, having two birth/death certificates, or different birth places. Often information from family members may be used to help validate a questionable link. For instance, if encrypted Social Security numbers do not match for an individual, they are compared with those of relatives (parents, spouse, children) for the case where the encrypted Social Security numbers is used on a record but is not that of the individual in question. The same process may be performed that involve mismatched addresses and phone numbers.

Finally, records identified as valid matches, compiled from QualityStage based on composite scores, validation programs, and human review, are then processed through the "composite" person creation program and incorporated into UPDB. On a monthly basis, additional validation programs are run and assigned a priority value with the highest priority given to the most egregious logical inconsistencies and resolved. To provide an example, due to timing of the processing of links, imagine Individual A who has minimal information (example, name and birth date only) and who may link to two different individuals (B and C). The validation program which assesses A and B as well as A and C separately may not identify any problems. However, after the composite person is created from information from Individual A, B and C, the logical inconsistencies may arise such that the composite person now has multiple parents with different birth dates indicating an incorrect link. There are additional procedures in place to assess this problem of multiple records linked during a single linking run, however with multiple linkers and the lag time of creating the "composite" person, some of these links can only be caught during the monthly validation checks.

When invalid links are discovered, these links are broken and then new composite person records are created with a re-assessment of the best information to retain. There is also a process to permanently reject links so that they do not happen again with a new linking process by adding the IDs to a table that is checked every time a new link is processed. If new information comes to light, the rejected link can be removed from the table. This is a process that always involves human intervention.

5 PRIVACY AND CONFIDENTIALITY

Our principal concern regarding use of linked datasets in UPDB is protecting identities of individuals in these data. To establish some basic definitions, privacy refers to an individual's ability to control information about him/herself while confidentiality is the obligation of a second party to not reveal private information about an individual to a third party without the permission of the person concerned (Wylie & Mineau, 2003).

When individuals agree to participate in research studies or when UPDB data are provided to researchers, it is with the understanding that the information will only be used to advance research and will be kept confidential. Only the minimum data necessary to conduct the research is provided. Strategies such as removing explicit identifiers, e.g. name, full birth date, street address, and Social Security number, have been used to ensure confidentiality before releasing information to researchers.

Even when these measures have been implemented, potential re-identification methods could be used, such as matching to other databases or by looking at unique characteristics found in the fields of the database itself resulting in possible deductive disclosure of the identities of the individuals represented. Even when current methods of protecting identifying information from researchers are employed, there is still some risk to privacy and confidentiality when linking and sharing health information for research (Gymrek, McGuire, Golan, Halperin, & Erlich, 2013).

Using identifiers in UPDB is designed to optimize matching individuals across data sets, whether linking is being conducted with historical or contemporary data, so other approaches to protect confidentiality need to be employed. Because state regulations regarding individually identifying information may differ and because federal regulations and requirements will vary according to type of information and its use, the protections for privacy and confidentiality have to be tailored in different areas to comply with those regulations.

5.1 RESOURCE FOR GENETIC AND EPIDEMIOLOGIC RESEARCH (RGE)

Access to UPDB data is regulated by the Utah Resource for Genetic and Epidemiologic Research (RGE). The RGE was created by an Executive Order of the Governor of Utah on July 14, 1982. Relying on enabling statutes in state health code, the RGE was established as a "data resource for the collection, storage, study, and dissemination of medical and related information" to operate "for the purpose of reducing morbidity or mortality, or for the purpose of evaluating and improving the quality of hospital and medical care". Originally administered under the direction and supervision of the Utah Department of Health, the RGE was transferred to the University of Utah by a second Executive Order in 1986. RGE is the legal custodian for the data contained within the UPDB and is responsible for developing and maintaining contractual agreements with organizations that contribute data to the UPDB or that links records to the UPDB.

Each project requesting access to data from the UPDB or linked electronic medical records applies to RGE for review. Applications are reviewed by the RGE Committee, which includes representatives from the university faculty with expertise in several disciplines including demography, genetics, public health and epidemiology, as well as representatives from each of the data contributors. Each data contributor has the right to veto the use of its own data if the representative determines the proposal describes an inappropriate use of its data. In practice, representatives of the data contributors rarely exercise their veto power because most applications can be revised to address concerns. All projects are also required to obtain approval by the appropriate Institutional Review Board(s) and Privacy Board(s) before access to data is granted.

RGE has the responsibility to protect the sensitive confidential information in UPDB. The RGE requires that users with access to data sign the RGE Confidentiality and Data Use Agreement. Each user on a project is also required to disclose any relationship with a for-profit company that might have an interest in the research being conducted with UPDB data. Relationships with for-profit companies are not prohibited, but require an assurance that data will be protected. The RGE Committee evaluates the data security for each location in which UPDB data will be stored. Finally, before any research is published, investigators must submit manuscripts to RGE for review, which includes scrutiny of any potentially identifying data presented for publication.

5.2 DATA SHARING AND RELATIONSHIPS WITH DATA CONTRIBUTORS

Important issues arise when collaborating with agencies who contribute data to the UPDB. For UPDB to exist, it requires the full participation of numerous organizations interested in advancing research and willing to share data for the purposes of research. The University of Utah is the steward of the data comprising UPDB but these data are not owned by the University of Utah. Formal agreements with the contributing agencies to facilitate long-term sustainability of using linked datasets for research in UPDB have been addressed and involve addressing the following issues:

1. Most datasets in UPDB were not collected specifically for research but are allowed for research use with consideration of privacy and confidentiality concerns by the data contributor.
2. Data collected by investigators for administrative purposes, research (including biospecimen data), and from high-risk clinics are linked to UPDB, but any diagnostic, relationship or residential information cannot be released until permission is given by the investigator who provided the data to UPDB to be used by other investigators with appropriate IRB approval.
3. RGE works with data contributors to carefully describe to data users the data contributors' authority for allowing research use of the contributors' data linked to other datasets.
4. RGE and PPR staff negotiate agreements with data contributors regarding data security measures and the methods used to protect the confidentiality of these data.
5. When scientific discoveries are made, intellectual property needs to be clarified by formal agreements between institutions since these scientific advancements are based on links between data sets which represent new and synergistic information.

We illustrate these principles with an example that relates to an NIH grant directed by Dr. Ken Smith. In that study, Medicare claims data (Principle #1) were requested to allow age-eligible individuals in UPDB to be matched to their Medicare records (Principle #2). The University of Utah owns the links but not the Medicare data themselves which RGE explains to users (Principle #3). In this sense, researchers may function in ways that are similar to data contributors since they may (1) contribute the links outright to the research resource or (2) maintain control over future use, as institutional data contributors do, while establishing the necessary data security and privacy protections. For the latter, Medicare records must remain on a single secure server in a manner compliant with the requirements of the Centers for Medicaid and Medicare Services, the federal agency that collects and allows approved release of Medicare files (Principle #4). Several studies have used these Medicare data linked to UPDB (Hanson, Horn, Rasmussen, Hoffman, & Smith, 2017; Hanson, Smith, & Zimmer, 2015; Hollingshaus et al., 2016; Wirostko et al., 2016), publications that acknowledge the value and approved access to the Medicare data (Principle #5).

5.3 UPDB AS A RESEARCH RESOURCE

Certain kinds of research infrastructures, secure data centers, or statistical coordinating centers often link data sets for research use such as is done with UPDB. These research entities generally hold identifying data, link records and then provide approved de-identified data or limited data sets to investigators. Such entities have policies associated with the release of these linked data information. Like other such research entities, UPDB relies on the following RGE policies and procedures:

1. To create "minimum-necessary" (often de-identified or limited) data sets to be released for research use.
2. To preclude researchers from linking to other data resources (without approval) to prevent disclosure of individual information outside the scope of the original research agreement (Kohane & Altman, 2005; Winickoff, 2006).
3. To develop confidentiality agreements with users/investigators that require users not attempt to re-identify individuals and will disclose any breeches of confidentiality to RGE.
4. To develop methods for contacting and recruiting individuals for participation in a research protocol (Wylie & Mineau, 2003).

There are several models for creating research resources. The concept of the "Charitable Trust" has been suggested from the field of genomics biobanks (Winickoff & Winickoff, 2003). This means that academic medical centers might (and often) elect to transfer blood, tissue, and medical data to private biobanks in an exchange for access for research and equity. With this Charitable Trust approach,

when a person agrees to donate tissue, the trust is the steward of the tissue and is obligated to ensure protection of the donated tissue. Maintaining the viability and sustainability of the Trust is a challenge with this strategy, for participants, universities and for-profit organizations (Master, Campo-Engelstein, & Caulfield, 2015; Turner, Dallaire-Fortier, & Murtagh, 2013).

Others have proposed disease registries (e.g., statewide cancer registries) that could use a national system that separates information under the control of three groups, including the Disease Registry, a Population Registry (a trusted agency that maintains the personal identifying information) and an Identifier Translation Agency (another trusted third party that has the key to translate the unique identifier assigned by the Population Registry and by the disease registry) (Churches, 2003). This strategy would ensure safety but makes linking data more cumbersome, reduces the effectiveness of identity matching, increases costs, and imposes added costs for conducting research.

Some agencies, institutes and centers may provide infrastructure for record linking activities. Major data providers allow access to confidential data at secure data centers for approved users who have achieved security clearances. Gaining access to confidential micro-data, such as those held by the US Census and National Center for Health Statistics through the Federal Statistical Research Data Centers, represents an important example of this strategy. There are also university data centers that operate secure computing systems that have policies that allow researchers to acquire, maintain, and analyze restricted-use data.

In Utah, RGE is a resource that has addressed these issues for decades. It is a dynamic institution whose policies and procedures address increasing complexities related to managing and linking individual-identifying records for research use, and embodies elements of all three approaches: RGE controls access to the data as they are provided to UPDB through agreements with data contributors. UPDB encrypts Social Security Numbers when working with the statewide healthcare facilities and claims records as required by the Utah Department of Health. Medical record numbers are replaced with random unique IDs by University of Utah Health and Intermountain Healthcare before linking their electronic health records to the UPDB.

6 FEATURES OF UPDB THAT FACILITATE THE INTEGRATION OF GENETICS AND DEMOGRAPHY

There are substantial opportunities afforded to researchers using UPDB. Every year, the number of generations represented increases such that the detection of familial aggregation of diseases and outcomes improves. UPDB provides the capacity to identify multigenerational pedigrees with significant excess prevalence of specific conditions and to then enroll them for genetic, outcomes or other population-based studies. Indeed, deeper and larger comprehensive genealogies enhance the likelihood of gene discoveries. Because records from many data sets are linked together, UPDB is able to combine, confirm, or improve information at the individual level. Creating and maintaining a database similar to the UPDB would require resources beyond the scope of any single research project. While the genealogy records in UPDB may appear to be similar to those available through web-based genealogical databases, they are not since those sources generally only represent primary source data for use by individuals doing their own genealogies. With the use of vital records and driver license records, UPDB is a statewide resource and individuals born and living in Utah are more comprehensively represented. UPDB supports the larger goal of treating the entire state as a platform for genetic, population and outcomes research.

The value of the UPDB comes, in large measure, to the synergies arising from the number, time coverage and diversity of records added to the resource. A number of new sources of records and infrastructure developments exist that will expand the data collection needed to improve the utility of the UPDB.

1. Data Coverage Expansion. In addition to the annual updates of data sources described previously, UPDB continues to add other records such as historic birth certificate data and historic Census records (1950) as they become available.

2. Environmental and Geo-Spatial Capabilities. Expansion of the linkage of georeferenced environmental-geographic-socioeconomic data to UPDB facilitates environmental epidemiology, health services research, gene-environment interaction research and studies of social disparities.

3. Utah Genome Project and the Center for Genomic Medicine at the University of Utah. The Utah Genome Project (UGP) is a large-scale, multi-year initiative to advance better disease prevention, diagnosis, and treatment methods through discovery of genetic signatures for human diseases and response to drug therapies. UGP supports projects that collect biologic samples and sequence DNA identified through UPDB families with excess burden of disease; diseases targeted for support from UGP have significant public health impact and are deemed to be scientifically feasible targets for analysis. The UGP is a component of the Center for Genomic Medicine (CGM) which supports additional analytic and translation objectives along with data access to the UPDB.

4. Visualization and a New Pedigree construction. Important enhancements to UPDB's Kinship Analysis Tools (KAT) (Kerber, 1995; Kerber, O'Brien, Smith, & Cawthon, 2001) are being made. The use of visualization and network-based tools takes a set of individuals of interest, provided by a researcher, and traverses the full extent of genealogies in the UPDB, connecting these individuals through all existing family relationships, whether close or distant relatives, and creating a comprehensive multi-lineage pedigree. These programs are fully scalable and will fill a current gap in describing family structure, social networks and provide depictions of complex pedigrees necessary for sophisticated genetic analyses. Additional tools are being developed for the analysis of genetic heterogeneity and gene–environment interactions. (Hanson et al., 2020)

5. UPDB Limited (UPDB-L) Query Tool. Potential investigators may access large subsets of data from the UPDB through the online UPDB Limited Query Tool. Version 1.0 was released in 2009 and provided access to all death and birth certificates, Inpatient Hospital Claims and Ambulatory Surgery Claims, statewide cancer diagnoses, geographic and demographic information, along with data on familial relationships and pedigrees. Plans are in place to expand the tool to include emergency department claims, All Payer Claims Database and University of Utah health data.

6. UPDB Linkages to Biospecimens and Clinical Measures. The UPDB has benefitted from linkages to biospecimens. In working with the Huntsman Cancer Institute's Research Informatics Shared Resource and the Biospecimen and Molecular Pathology, UPDB is linked to clinically annotated biobanking, histology services, and molecular diagnostics. Since UPDB is also linked to the records of Intermountain Healthcare and the University of Utah Hospitals and Clinics, these sources also hold clinical data that arise as a matter of patient care; these clinical data are only transferred to researchers from the respective clinical enterprise data warehouses when projects are approved.

7 CONCLUSION

The UPDB offers demographers, historians, geneticists, oncologists, physicians, epidemiologists, and other social scientists an unparalleled data resource from which to launch the next generation of studies that rely on data heretofore unavailable, including the prospect of novel data types such as full genome and exome sequencing. Access to these novel data joined with the depth of information from the UPDB make it an extremely attractive research resource for both investigators and their trainees and students and have contributed to the success and popularity of the UPDB. Moreover, given the sensitivity of the data (spanning family relationships, linkages to DNA and biobanks, geospatial markers), users of the UPDB receive the protections and oversight of the Utah Resource for Genetic and Epidemiologic Research (RGE) that have been in place for decades and permit responsible and ethical use of the data while protecting the identities of the individuals whose data are the basis for the research undertaken. Over time, the UPDB has grown in terms of the number of individuals and families represented as well as the diversity of data sources. This growth, with the proper privacy protections, portend continued use of UPDB across a range of topics and disciplines. Nonetheless, the stakes remain high when managing such large volumes of data. The structure of the UPDB requires that it continues to earn the trust and confidence of the public, state government representatives, and

the data contributors. In the end, UPDB represents a valuable data resource which scientists can use to test next-generation hypotheses.

REFERENCES

Adams, J., Lam, D. A., Hermalin, A. I., & Smouse P. E. (Eds.) (1990). *Convergent issues in genetics and demography*. New York: Oxford University Press.

Bean, L. L., May, D. L., & Skolnick, M. (1978). The Mormon historical demography project. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 11*(1), 45–53. doi: 10.1080/01615440.1978.9955216

Casey, J. A., Schwartz, B. S., Stewart, W. F., & Adler, N. E. (2016). Using electronic health records for population health research: A review of methods and applications. *Annual Review of Public Health, 37*, 61–81. doi: 10.1146/annurev-publhealth-032315-021353

Cawthon, R. M., Smith, K. R., O'Brien, E., Sivatchenko, A., & Kerber, R. A. (2003). Association between telomere length in blood and mortality in people aged 60 years or older. *The Lancet, 361*(9355), 393–395. doi: 10.1016/S0140-6736(03)12384-7

Chernenko, A., Meeks, H., & Smith, K. R. (2019). Examining validity of body mass index calculated using height and weight data from the US driver license. *BMC Public Health, 19*, 100. doi: 10.1186/s12889-019-6391-3

Churches, T. (2003). A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BMC Medical Research Methodology, 3*, 1. doi: 10.1186/1471-2288-3-1

DuVall, S. L., Fraser, A. M., Rowe, K., Thomas, A., & Mineau, G. P. (2012). Evaluation of record linkage between a large healthcare provider and the Utah Population Database. *Journal of the American Medical Informatics Association, 19*(e1), e54–59. doi: 10.1136/amiajnl-2011-000335

DuVall, S. L., Kerber, R. A., & Thomas, A. (2010). Extending the Fellegi–Sunter probabilistic record linkage method for approximate field comparators. *Journal of Biomedical Informatics, 43*(1), 24–30. doi: 10.1016/j.jbi.2009.08.004

Fair, M. E., Lalonde, P., & Newcombe, H. B. (1991). Application of exact ODDS for partial agreements of names in record linkage. *Computers and Biomedical Research, 24*(1), 58–71. doi: 10.1016/0010-4809(91)90013-m

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210. doi: 10.1080/01621459.1969.10501049

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science, 339*(6117), 321–324. doi: 10.1126/science.1229566

Hammad, I. A., Meeks, H., Fraser, A., Theilen, L. H., Esplin, M. S., Smith, K. R., & Varner, M. W. (2020). Risks of cause-specific mortality in offspring of pregnancies complicated by hypertensive disease of pregnancy. *American Journal of Obstetrics and Gynecology, 222*(1), 75.e1–75.e9. doi: 10.1016/j.ajog.2019.07.024

Hanson, H. A., Horn, K. P., Rasmussen, K. M., Hoffman, J. M., & Smith, K. R. (2017). Is cancer protective for subsequent Alzheimer's disease risk? Evidence from the Utah Population Database. *The Journals of Gerontology: Series B, 72*(6), 1032–1043. doi: 10.1093/geronb/gbw040

Hanson, H. A., Leiser, C. L., Madsen, M. J., Gardner, J., Knight, S., Cessna, M., . . . Camp, N. J. (2020). Family study designs informed by tumor heterogeneity and multi-cancer pleiotropies: The power of the Utah Population Database. *Cancer Epidemiology, Biomarkers & Prevention, 29*(4), 807–815. doi: 10.1158/1055-9965.EPI-19-0912

Hanson, H. A., Smith, K. R., & Zimmer, Z. (2015). Reproductive history and later-life comorbidity trajectories: A medicare-linked cohort study from the Utah Population Database. *Demography, 52*(6), 2021–2049. doi: 10.1007/s13524-015-0439-5

Hollingshaus, M. S. (2015). *Seeds of sorrow: A life-course approach to early-life parental death and later-life suicide and behavioral health risk* (Doctoral dissertation). Utah: University of Utah. Retrieved from https://collections.lib.utah.edu/ark:/87278/s6060q8v

Hollingshaus, M. S., Coon, H., Crowell, S. E., Gray, D. D., Hanson, H. A., Pimentel, R., & Smith, K. R. (2016). Differential vulnerability to early-life parental death: The moderating effects of family suicide history on risks for major depression and substance abuse in later life. *Biodemography and Social Biology, 62*(1), 105–125. doi: 10.1080/19485565.2016.1138395

Hurdle, J. F., Smith, K. R., & Mineau, G. P. (2013). Mining electronic health records: An additional perspective. *Nature Reviews Genetics, 14*, 75. doi: 10.1038/nrg3208-c1

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine, 14*(5–7), 491–498. doi: 10.1002/sim.4780140510

Kerber, R. A. (1995). Method for calculating risk associated with family history of a disease. *Genetic Epidemiology, 12*(3), 291–301. doi: 10.1002/gepi.1370120306

Kerber, R. A., O'Brien, E., Smith, K. R., & Cawthon, R. M. (2001). Familial excess longevity in Utah genealogies. *The Journals of Gerontology: Series A, 56*(3), B130–139. doi: 10.1093/gerona/56.3.b130

Kohane, I. S., & Altman, R. B. (2005). Health-information altruists — A potentially critical resource. *The New England Journal of Medicine, 353*(19), 2074–2077. doi: 10.1056/NEJMsb051220

Master, Z., Campo-Engelstein, L., & Caulfield, T. (2015). Scientists' perspectives on consent in the context of biobanking research. *European Journal of Human Genetics, 23*(5), 569–574. doi: 10.1038/ejhg.2014.143

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., . . . Skolnick, M. H. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science, 266*(5182), 66–71. doi: 10.1126/science.7545954

Neklason, D. W., Stevens, J., Boucher, K. M., Kerber, R. A., Matsunami, N., Barlow, J., . . . Burt, R. W. (2008). American founder mutation for attenuated familial adenomatous polyposis. *Clinical Gastroenterology and Hepatology, 6*(1), 46–52. doi: 10.1016/j.cgh.2007.09.017

Newcombe, H. B. (1969). The use of medical record linkage for population and genetic studies. *Methods of Information in Medicine, 8*(1), 7–11.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science, 130*(3381), 954–959. doi: 10.1126/science.130.3381.954

Norton, M. C., Fauth, E., Clark, C. J., Hatch, D., Greene, D., Pfister, R., . . . Smith, K. R. (2016). Family member deaths across adulthood predict Alzheimer's disease risk: The Cache County Study. *International Journal of Geriatric Psychiatry, 31*(3), 256–263. doi: 10.1002/gps.4319

Norton, M. C., Smith, K. R., Østbye, T., Tschanz, J. T., Corcoran, C., Schwartz, S., . . . Welsh-Bohmer, K. A. (2010). Greater risk of dementia when spouse has dementia? The Cache County Study. *Journal of the American Geriatrics Society, 58*(5), 895–900. doi: 10.1111/j.1532-5415.2010.02806.x

Norton, M. C., Smith, K. R., Østbye, T., Tschanz, J. T., Schwartz, S., Corcoran, C., . . . Welsh-Bohmer, K. A. (2011). Early parental death and remarriage of widowed parents as risk factors for Alzheimer disease: The Cache County study. *The American Journal of Geriatric Psychiatry, 19*(9), 814–824. doi: 10.1097/JGP.0b013e3182011b38

Sellers, T. A., Caporaso, N., Lapidus, S., Petersen, G. M., & Trent, J. (2006). Opportunities and barriers in the age of team science: Strategies for success. *Cancer Causes & Control, 17*, 229–237. doi: 10.1007/s10552-005-0546-5

Shah, R., Pico, A. R., & Freedman, J. E. (2016). Translational epidemiology: Entering a brave new world of team science. *Circulation Research, 119*(10), 1060–1062. doi: 10.1161/CIRCRESAHA.116.309881

Skolnick, M., Bean, L. L., Dintelman, S. M., & Mineau, G. (1979). A computerized family history data base system. *Sociology and Social Research, 63*(3), 506–523.

Skolnick, M., Bean, L., May, D., Arbon, V., De Nevers, K., & Cartwright, P. (1978). Mormon demographic history I. Nuptiality and fertility of once-married couples. *Population Studies, 32*(1), 5–19.

Skolnick, M., Bishop, D., Carmelli, D., Gardner, E., Hadley, R., Hasstedt, S., . . . Smart, C. (1981). A population-based assessment of familial cancer risk in Utah Mormon genealogies. In F. E. Arrighi, P. N. Rao, & E. Stubblefield (Eds.), *Genes, chromosomes, and neoplasia* (477–500). New York: Raven Press

Smith, K. R., Brown, B. B., Yamada, I., Kowaleski-Jones, L., Zick, C. D., & Fan, J. X. (2008). Walkability and body mass index: Density, design, and new diversity measures. *American Journal of Preventive Medicine, 35*(3), 237–244. doi: 10.1016/j.amepre.2008.05.028

Smith, K. R., Hanson, H. A., Mineau, G. P., & Buys, S. S. (2012). Effects of *BRCA1* and *BRCA2* mutations on female fertility. *Proceedings of the Royal Society B, 279*(1732), 1389–1395. doi: 10.1098/rspb.2011.1697

Smith, K. R., Zick, C. D., Kowaleski-Jones, L., Brown, B. B., Fan, J. X., & Yamada, I. (2011). Effects of neighborhood walkability on healthy weight: Assessing selection and causal influences. *Social Science Research, 40*(5), 1445–1455. doi: 10.1016/j.ssresearch.2011.04.009

Song, X., & Campbell, C. D. (2017). Genealogical microdata and their significance for social science. *Annual Review of Sociology, 43*(1), 75–99. doi: 10.1146/annurev-soc-073014-112157

Stokols, D., Misra, S., Moser, R. P., Hall, K. L., & Taylor, B. K. (2008). The ecology of team science: Understanding contextual influences on transdisciplinary collaboration. *American Journal of Preventive Medicine, 35*(2 Suppl.), S96–S115. doi: 10.1016/j.amepre.2008.05.003

Stroup, A. M., Herget, K. A., Hanson, H. A., Reed, D. L., Butler, J. T., Henry, K. A., . . . Smith, K. R. (2017). Baby Boomers and birth certificates: Early-life socioeconomic status and cancer risk in adulthood. *Cancer Epidemiology, Biomarkers & Prevention, 26*(1), 75–84. doi: 10.1158/1055-9965.EPI-16-0371

Theilen, L. H., Fraser, A., Hollingshaus, M. S., Schliep, K. C., Varner, M. W., Smith, K. R., & Esplin, M. S. (2016). All-cause and cause-specific mortality after hypertensive disease of pregnancy. *Obstetrics & Gynecology, 128*(2), 238–244. doi: 10.1097/AOG.0000000000001534

Theilen, L. H., Meeks, H., Fraser, A., Esplin, M. S., Smith, K. R., & Varner, M. W. (2018). Long-term mortality risk and life expectancy following recurrent hypertensive disease of pregnancy. *American Journal of Obstetrics and Gynecology, 219*(1), 107.e1–107.e6. doi: 10.1016/j.ajog.2018.04.002

Turner, A., Dallaire-Fortier, C., & Murtagh, M. J. (2013). Biobank economics and the "Commercialization Problem". *Spontaneous Generations: A Journal for the History and Philosophy of Science, 7*(1), 69–80. doi: 10.4245/sponge.v7i1.19555

Williams, R. R., Skolnick, M., Carmelli, D., Maness, A. T., Hunt, S. C., Hasstedt, S., . . . Jones, R. K. (1979). Utah pedigree studies: Design and preliminary data for premature male CHD deaths. *Progress in Cinical and Biological Research, 32*, 711–729. Retrieved from PMID: 523491

Winickoff, D. E. (2006). Health-information altruists. *The New England Journal of Medicine, 354*(5), 530–531. doi: 10.1056/NEJMc053390

Winickoff, D. E., & Winickoff, R. N. (2003). The charitable trust as a model for genomic biobanks. *The New England Journal of Medicine, 349*(12), 1180–1184. doi: 10.1056/NEJMsb030036

Wirostko, B. M., Curtin, K., Ritch, R., Thomas, S., Allen-Brady, K., Smith, K. R., . . . Allingham, R. R. (2016). Risk for exfoliation syndrome in women with pelvic organ prolapse : A Utah project on Exfoliation Syndrome (UPEXS) study. *JAMA Ophthalmology, 134*(11), 1255–1262. doi: 10.1001/jamaophthalmol.2016.3411

Wylie, J. E., & Mineau, G. P. (2003). Biomedical databases: Protecting privacy and promoting research. *Trends in Biotechnology, 21*(3), 113–116 doi: 10.1016/S0167-7799(02)00039-2

Zick, C. D., Smith, K. R., Fan, J. X., Brown, B. B., Yamada, I., & Kowaleski-Jones, L. (2009). Running to the store? The relationship between neighborhood environments and the risk of obesity. *Social Science & Medicine, 69*(10), 1493–1500. doi: 10.1016/j.socscimed.2009.08.032