Supplement – Appendix A.

Description of the Utah Population Database with References

Overview. The Utah Population Database (UPDB) is a research resource housed at the University of Utah that contains computerized data records for over 11 million individuals who have lived in Utah or currently reside in the state(1). As previously described in detail(2), the UPDB was created in the 1970's from genealogic data of pioneer founders and their descendants (1.63 million individuals born from the early 1800's to the middle of the 1900's) who were primarily of Northern European descent(3) and who had low levels of inbreeding(4). Computerized genealogic data were linked to statewide cancer records in the Utah Cancer Registry (UCR), a Surveillance Epidemiology and End Results (SEER) registry, and to individual-level Utah vital records (deaths, births, marriages, and divorces) from the early 1900's. Other high-quality, statewide data sets were incorporated into the UPDB during the 1990's including driver's licenses and medical records from inpatient and ambulatory facilities in Utah. These data are updated annually, and probabilistic record-linking is performed with individuals in UPDB as described (5). In conjunction with data from the United States Census of Utah (1880-1940), these records are used to extend family histories to the present day or create new family histories as previously unlinked individuals have events recorded in the state. Because of its size and the varied sources of its information, most families living in Utah are represented(6); for example, of all individuals born in Utah in 1950, 84% have grandparent information and 67% have five or more previous generations documented in the UPDB(1). In the case of children and adolescents, the UPDB is able to establish follow-up from probabilistic linked records of both a child and those of her or his mother. As such, it is the only database of its kind in the U.S., and one of only a few in the world.

Links to healthcare systems data. Intermountain Healthcare is the largest and University of Utah Healthcare (UUHC) the second largest healthcare system in Utah. Both not-for-profit healthcare delivery

systems operate multiple hospitals, outpatient clinics, ambulatory surgery centers, laboratories, and health insurance plans covering Utah and southeastern Idaho. In addition to tertiary-level teaching and research facilities, Intermountain and UUHC also have several small hospitals and clinics that are the only source of care in some rural Utah communities. The Intermountain and UUHC enterprise data warehouses (EDWs) were created to bring together comprehensive health and administrative data beginning in 1996 from all facilities to allow researchers to study patient data. Patients in the EDWs are identified by an enterprise master patient index, which is used to link data resulting from all patient encounters(8). There are regular audits of enterprise master patient index numbers that look for duplicates and inaccuracy across systems. There are more than five million patients listed in the Intermountain EDW and 2 million in the UUHC EDW connected to more than 35 billion data points such as laboratory results, discharge summaries, pharmacy orders and diagnosis codes. Our methodology allows the record linking activity to be completed using patient demographic information without exposing any medical information. After the linking is complete, a master subject index (MSI) is created, the identifying information of each respective EDW is deleted, and a copy of the MSI is held by each institution to facilitate future projects. The MSI allows each institution to maintain control of their information and protects the confidentiality of their patients. When research projects request use of these data, the investigator is required to obtain approval from Resource for Genetic and Epidemiologic Research and the institutional review boards from each institution, and it is only at this point that information from both institutions' EDWs are accessed and combined(8).

Follow up procedures in UPDB. In population-based cohort studies, an accurate assessment of follow-up times in both cases and unaffected controls merits careful consideration. Valid exposed-unexposed comparisons using retrospective cohort data depend on appropriate matching of exposure periods and longitudinal tracking of individuals. The UPDB records the date that each person in the database was last known to be residing in Utah, as described in a detailed appendix(9). To better capture follow-up periods for healthy adults and to avoid any potential bias in selecting controls for recently diagnosed cases, we developed a secondary procedure to modify the "last residence in Utah" date. If an

individual's “last residence” date is based on a driver's license renewal, we assume that an individual who renews a Utah driver's license as required (once every 5 years) is still living in Utah until she or he becomes licensed outside of Utah, surrenders a license or dies (9). Assessing length of follow up for potential population-based controls can be problematic, if the first date an individual appears in Utah is unknown, particularly for those born outside of Utah. We developed an algorithm for determining a ‘first residence in Utah’ date to reduce potential bias from selecting controls new to Utah that may have been exposed elsewhere as described in a detailed appendix(10). By creating a statewide pool of individuals with adequate follow up based on time between an index event in Utah and the most recent event recorded in the UPDB, we ensure appropriate matching of exposure periods between cases and controls in our study.

References

1. The University of Utah Pedigree and Population Resource: Utah Population Database, Overview. http://healthcare.utah.edu/huntsmancancerinstitute/research/updb/.

2. Feldkamp ML, Carey JC, Pimentel R, Krikov S, Botto LD. Is gastroschisis truly a sporadic defect?: familial cases of gastroschisis in Utah, 1997 to 2008. Birth Defects Res A Clin Mol Teratol. 2011;91(10):873-878.

3. McLellan T, Jorde LB, Skolnick MH. Genetic distances between the Utah Mormons and related populations. Am J Hum Genet. 1984;36(4):836-857.

4. Jorde LB. Inbreeding in the Utah Mormons: an evaluation of estimates based on pedigrees, isonymy, and migration matrices. Ann Hum Genet. 1989;53(Pt 4):339-355.

5. Prahalad S, O'Brien E, Fraser AM, et al. Familial aggregation of juvenile idiopathic arthritis. Arthritis Rheum. 2004;50(12):4022-4027.

6. Kerber RA, O'Brien E. A cohort study of cancer risk in relation to family histories of cancer in the Utah population database. Cancer. 2005;103(9):1906-1915.

7. Duvall SL, Fraser AM, Kerber RA, Mineau GP, Thomas A. The impact of a growing minority population on identification of duplicate records in an enterprise data warehouse. Stud Health Technol Inform. 2010;160(Pt 2):1122-1126.

8. DuVall SL, Fraser AM, Rowe K, Thomas A, Mineau, GP. Evaluation of record linkage between a large healthcare provider and the Utah Population Database, Journal of the American Medical Informatics Association, Volume 19, Issue e1, 1 June 2012, Pages e54–e59, https://doi.org/10.1136/amiajnl-2011-000335.

9. Curtin K, Smith KR, Fraser A, Pimentel R, Kohlmann W, Schiffman, JD. Familial risk of childhood cancer and tumors in the li-fraumeni spectrum in the utah population database: Implications for genetic evaluation in pediatric practice. Int. J. Cancer 2013;133: 2444–2453.

10. Curtin K, Fleckenstein AE, Robison RJ, Crookston MJ, Smith KR, Hanson GR. Methamphetamine/amphetamine abuse and risk of Parkinson's disease in Utah: a population-based assessment. Drug and alcohol dependence. 2015 Jan 1;146:30-8.