Data sharing in the Dunedin Study

Terrie E. Moffitt and Richie Poulton
tem11@duke.edu, richie.poulton@otago.ac.nz

Introduction:

We are enthusiastic about the open-science movement to enhance reproducibility (http://www.sciencemag.org/content/348/6242/1422.full?ijkey=ha1o5D9wvW4ZQ&keytype=ref&siteid=sci). The Dunedin Study has had a data-sharing policy in place for over 20 years, and we have registered all data-analysis plans on a public google site for the last five years. This document is intended to give interested parties a full explanation of our data sharing policy, and the longstanding rationales behind it. The document ends by describing requirements for accessing Dunedin Study data.

We seek a careful balance between the benefits of data-sharing in research and any potential risks to study participants. Seeking this balance is aided by consultation of research and policy on data sharing (http://www.nature.com/news/researchers-wrestle-with-a-privacy-problem-1.18396). For example, the 2015 version of the **Helsinki Declaration** addresses the balance between the aims of open-access data and the aims of human-subjects protection: "Principle 8. While the primary purpose of medical research is to generate new knowledge, this goal can never take precedence over the rights and interests of individual research subjects" (http://jama.jamanetwork.com/article.aspx?articleid=1760318). Similarly, our policy conforms to the revised Common Rule for Human Subjects Protection, which requires advanced informed consent for data sharing, even of de-identified data http://www.nejm.org/doi/full/10.1056/NEJMp1512205?query=TOC. As another example, in August 2014, NIH issued its Genomic Data Sharing (GDS) Policy. This policy was intended for investigators who generate large-scale genomic and phenotypic data from federally funded new collections. The Dunedin Study is not among that group. Nevertheless, we think the basic principles of the GDS Policy apply well to data sharing in general, across many types of data. To explain our own policy here, we drew on published materials from the NIH GDS Policy (http://osp.od.nih.gov/under-the-poliscope/2015/08/genomic-data-sharing-two-part-series).

The 2014 NIH GDS Policy establishes that data-sharing can only occur with the advance consent of research participants, even if the datasets generated have been de-identified. NIH has taken this new approach to informed consent because formal research into participants' preferences has revealed that participants expect to be asked for permission before scientists use and share their de-identified data for research (for a special issue on this research see The End of Privacy, *Science* 22 Feb 2015, www.sciencemag.org). Moreover, as has been well-documented, the risk of re-identification of data, particularly genomic data but also other types of data, is no longer a theoretical possibility. Re-identified data could potentially be used to discriminate against or stigmatize participants, their families, or groups. As such, it is no longer tenable for scientists to hold that anonymization is still achievable or to allow unrestricted sharing of "de-identified" datasets without consent on the premise that de-identified use is without risk to the donor. The GDS Policy, and the

2015 revision of the Common Rule, both urge that the research enterprise must begin to respect the wishes of participants in relation to data access.

The Dunedin Study has not sought informed consent for unrestricted data sharing because data from the Dunedin study have historically been deemed by the Duke and Otago IRBs as being in a **high-risk** category that precludes making the data set available for unrestricted, unsupervised open-access data sharing. Consent documents for the study used over the past 40 years have informed each study member that "…all the information obtained by the researchers at the Dunedin Multidisciplinary Health and Development Research Unit will be treated as STRICTLY CONFIDENTIAL to members of the research team," and "Only approved Dunedin Study researchers will have access to your data." These consent documents were last signed by Study members at the age-38 assessment, which ended in 2012. This means that the Dunedin Study participants have not at this point given their informed consent for unrestricted data sharing, and therefore data deriving from their participation cannot be made available for unrestricted use.

There are 5 main reasons for the Dunedin Study's approach to informed consent and data sharing; each derives directly from the special circumstances of an ongoing 4-decade longitudinal multi-generational study of a birth cohort of human participants and their families. We note each below.

1. **Risk of mental pain and suffering.** The research team and IRBs recognize the risk to study members of mental pain and suffering from worry about the security of their lifetime of data. Due to the depth, multidisciplinary breadth and duration of the Dunedin Study, this dataset differs markedly from data sets created when research participants take part in a one-time limited data–collection session. The Dunedin data set contains sensitive information regarding topics concerning participants' IQ, income, health behaviors, credit ratings, conviction records, social welfare records, and medical records. Unusual in research, the data set includes information divulged by study participants in confidential interviews about, for example, their lifetime history of mental disorders, sexual preference, suicidality, physical and sexual abuse victimisation, substance use, high-risk sexual behavior, domestic violence, stressful life events such as an abortion, parenting of children, and crimes committed. Also unusual, the data set contains information about the medical and psychiatric histories of four generations of the study members' families, from their grandparents to their offspring. Since the 1990's the data set contains genetic and genomic data and biomarker data, which have special ethical status because they allow re-identification, and because researchers are in the position to know information about study members' genes and health that they themselves do not know. As reported in all publications from the study, the cohort members are born during an identifiable year in an identifiable city and they are from a small-population country. An ill-intentioned user could very easily misuse the data of the longitudinal study to illicitly identify individual study

members and their families, and to expose confidential and potentially destructive details of their lives. The likelihood of any scientist doing this is immaterial. *What is material is the study members' perceptions of the potential for data-security risk, and their concerns about it.*

2. **At-risk participants.** Substantial proportions of the cohort belong to at-risk groups. It is standard IRB policy that such groups require a simple-to-understand, unconditional guarantee that all of their data are held in strict confidence by the research team. These groups include incarcerated prisoners, patients with chronic mental illnesses (such as schizophrenia or autism), and individuals whose tested cognitive abilities are in the diagnoseable range of mental retardation or mild cognitive impairment. For these groups, trust is achieved by putting a face on who will use their data, and this is inconsistent with seeking consent for unrestricted data-sharing.

3. **Multiple suppliers of data.** Because this study has been underway for four decades, much of the data were collected from individuals and agencies who gave consent under the condition that data would be kept strictly confidential and used only by the Dunedin Study research team. These data sources include mothers, fathers, teachers, peer informants, partners, doctors, schools, government agencies, and private companies who provided administrative data. These various Dunedin Study data sources are not now accessible to us for re-consent and therefore data derived from them cannot be shared for unrestricted use. (Government agencies and private companies sometimes place restrictions on data access that the research team must respect; for example, the US Social Security Administration provides data for the Health and Retirement Survey, but stipulates that these data may only be shared with scientists who hold a federally funded grant.)

4. **Risk of cohort attrition caused by participants' concerns about data security.** The Dunedin Study will be actively ongoing for years into the future. The scientific value of the longitudinal design relies on future follow-ups of the cohort, and high participation rates at those future follow-ups. As such, the benefits of data-sharing for a single paper project in the short run must always be weighed against the greater benefit of preserving the cohort intact for the multi-decade longitudinal study as a whole, in the longer run. Our surveys of our cohort members indicate that their continued participation is contingent on the consent forms' stating that "Your data are held in strict confidence" and "Only staff members of the Research Unit will have access to your data."

5. **Growing public concern about data security stimulated by media coverage.** The Dunedin Study cohort families were first enrolled in the study many years ago, in a kinder, gentler era. Their first two decades of participation were marked by growing trust in the research team, which was based on personal contact between researchers and families, and on our

proven track record for preserving participants' confidentiality. In the early days, the Dunedin Study was not internationally visible, data were not kept in electronic format, genomic and biomarker data were not collected, and the "open-science" movement for unrestricted open-access data-sharing had yet to emerge on the scientific scene. Requests for data-sharing were few and could be handled by a welcoming stance toward collaboration. Study families were comfortable with our collaborative data-sharing approach.

However, times have changed, and we can no longer take Study members' approval of data security and data sharing for granted. Dunedin cohort members now often contact us in reaction to media coverage reporting that research participants and their families can easily be identified using only their DNA, age, and city (for example: http://www.nytimes.com/2013/06/18/science/poking-holes-in-the-privacy-of- DNA) (as another example: https://datafloq.com/read/re-identifying-anonymous-people-with-big-data/228) (as another example: http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/). Such media coverage is changing the way that cohort members think about their lifetime repository of data in the Dunedin Study. Media coverage of "The Surveillance Society," portrayal of forensic science in television and film (such as CSI), and news stories of criminal hackers accessing supposedly secure government and industry data bases appear daily. Today, efforts to recruit research participants routinely fail, so much so that the National Academy of Sciences convened a panel to address the problem, which is in part due to public lack of confidence in data security (Massey DS, Tourangeau R. New Challenges to Social Measurement. Annals of the American Academy of Political and Social Science. 2013;645:6-22). Thus, we and our IRBs recognize that an ironclad guarantee of confidentiality is essential to make Dunedin Study members feel safe, prevent mental pain and suffering caused by worries about data security, and to prevent cohort attrition.

Our data-sharing policy provides for researchers outside the Study to access data used in a published paper by becoming "honorary" staff members of the Dunedin Unit, so they can access the data via collaboration (policy on the Dunedin Study website [http://dunedinstudy.otago.ac.nz]). Applicant investigators are invited to submit a concept paper describing the data analysis project they wish to carry out. The applicant investigator is then nominated for a non-salaried appointment as an "Associated Investigator," under the sponsorship of a study PI, for a limited term corresponding to the duration of the project. We provide all such investigators with clean, well-documented data files and electronic data-set dictionaries. To ensure effective data sharing, the PI-sponsor discusses detailed data-analysis plans with each investigator in advance and stays actively involved throughout each project. Our involvement is required because our IRBs have long required that consent forms must include "The name and contact information of an individual who is affiliated with the institution and familiar with the research and will be available to address participant questions." (NIH GDS Policy requires this as of January 2015.)

We provide participants with the names and contact details of PI's involved in the Dunedin Study at the time of consent, and we list these individuals on the Study website. One of them must be the sponsor responsible for each data-analysis project.

Access requirements in a nutshell. Proposed data-analysis projects from qualified scientists must have a concept paper describing the purpose of data access, IRB approval at the applicants' university, and provision for secure data access. We offer secure access on the Duke and Otago campuses. These access requirements parallel those used by dbGap and the Health and Retirement Study.

We register all concept papers on a publically accessible site before analysis begins: http://www.moffittcaspi.com/content/proposed-project-concepts (or https://sites.google.com/site/dunedineriskconceptpapers/)

All scripts and analysis files for Dunedin Study published papers are available.

Voluntary data-sharing. It is useful to keep in mind that the Dunedin Study's data sharing has always been voluntary, not compelled. This data-sharing policy has been in place and operating effectively for over 20 years. Unlike dbGaP and the HRS, the Dunedin Study has never been funded as a data provider. In addition, much of the data, including the genomic data, have not been funded by US taxpayers. Like the large Scandinavian register data bases that require travel to Scandinavia to access their data, the Dunedin Study contains data on non-US citizens only but the Dunedin Study makes data available in the USA at Duke University, without travel to New Zealand. Our data-sharing policy was last approved in 2015 by NIA as part of review of Dunedin Study competing-renewal funding.

Future plans:

1. When we next assess Dunedin Study members, we plan to seek IRB approval to offer two stage informed consent. At the first stage, Study members can consent to take part in the research with the longstanding guarantee: "Your data are held in strict confidence and only members of the Dunedin Study Research Unit will have access to your data." We will provide, as usual, an explanation of the risks and benefits to the Study member. At the second stage, Study members will be asked to consent to unrestricted data sharing with other qualified scientists. We will provide an explanation of the risks and benefits of unrestricted data-sharing, including the possibility of re-identification.