BRIEF WORKSHOP REPORT

Expanding the Data Repository: New Technology and Resources for the 21st Century

Accroître le registre de données : nouvelles technologies et ressources pour le XXI^{e} siècle

MARK SMITH, MSC
Associate Director, Repository
Manitoba Centre for Health Policy (MCHP), University of Manitoba
Winnipeg, MB

LESLIE L. ROOS, PHD
Distinguished Professor, Senior Research Scientist
MCHP, Department of Community Health Sciences, University of Manitoba
Winnipeg, MB

CHARLES BURCHILL
Associate Director, Data Use and Access
MCHP, Department of Community Health Sciences, University of Manitoba
Winnipeg, MB

These are exciting times to be at the Manitoba Centre for Health Policy (MCHP). The size and diversity of databases at MCHP provide an unprecedented opportunity to conduct population health and health services research and research into the determinants of health. Current holdings now represent over 90 databases of health, education, social services, survey and clinical data, as well as instruments such as the Early Development Indicator (EDI) now routinely administered to all primary school children in Manitoba. In the near future we expect to add public housing, infectious disease reporting and justice. Individual records in all these data sets can be linked, as approvals from research ethics boards, health information privacy committees and individual data trustees permit, through the use of an encrypted individual identifier.

Because of previous requirements, only individuals who were physically on-site could analyze data. For off-site investigators wanting to do their own data analysis, this meant that working with MCHP data was cumbersome. To address this limitation, several new and emerging technologies have been combined and will be piloted at MCHP over the next four

years using funds provided by a recent Canadian Foundation for Innovation (CFI) award. Chief among these are: two-level authentication, the SASTM Scalable Performance Data Server (SPDS) and thin clients running virtual desktops. (A thin client is one that relies on the server to do most of its processing.) SPDS provides finer control over data access (to the row and column level in data sets) and better user tracking than previous implementations. The final requirement is that all thin clients be available only in controlled-access areas (using swipe card technology) and that users log use of all thin clients. These requirements are part of a planned auditing process that will occur both routinely as well as randomly on short notice. By using these technologies and meeting these requirements researchers and analysts will be able to analyze MCHP data from remote locations. Users who breach confidentiality agreements are subject to penalties imposed by their respective institutions, as well as losing rights of access to the Repository.

We also undertook a review of other policies and procedures and noted that the better our online documentation is, the fewer are the demands placed on MCHP staff. As a consequence, a major reorganization of documentation and its accessibility is currently taking place, some of which will be available only on internal portal sites (for confidentiality reasons). This reorganization will include integrated access, for every data set, to trustee-provided documents, data schemas, data dictionaries, codebooks, history and file revisions, common problems and data quality reports.

Perhaps one of the largest undertakings was the development of an accreditation process, a half-day workshop that must be completed by all investigators who sign a data access agreement. This workshop introduces researchers to MCHP policies and procedures concerning access to confidential information, the expectations we have of them and those that they can have of us, and the wide range of supports (many online) for learning about and using administrative data. The entire accreditation process is available online at www.umanitoba.ca/medicine/units/mchp/.

Combining data from new and different domains does not come without analytical and statistical pitfalls. First, individual household-level data are available for only a subset of the population (based on information from the Manitoba prescription drug insurance plans); families receiving income assistance can also be specified. For everyone else, only ecological data on income are available. Second, scores on standardized tests provide information on student achievement, but many children in the appropriate age groups will not have scores recorded. Other files on enrolment, school grades, year in school and residency in the province without school attendance must be combined to provide a fuller picture for a given birth cohort. Techniques pioneered by Mosteller and Tukey (1977) have proven to be particularly useful in showing how messy data from multiple files can be put together to generate "normalized" distributions relevant for an entire population. For index creation, a standardized score for each individual can be computed by assuming an underlying logit distribution, divided into pieces according to the percentage of cohort members in each category (Roos et al. 2008; Willms 1986). Such distributions facilitate the use of powerful statistics.

MCHP research has highlighted the development of several kinds of family information

(Strohschein et al. 2009). Particularly useful may be the construction of family histories to assess the effects of critical life events – e.g., parental divorce or death – at different stages of childhood on educational achievement, health status and labour force participation. Second, longitudinal information on place of residence should permit the study of residential mobility as both a correlate and a predictor of health and achievement (Lix et al. 2006). Third, sibling–parent designs that link parental histories to child histories can be used to assess the contribution of such conditions as maternal depression to a child's mental health and subsequent outcomes. Multi-level modelling is a particularly useful statistical tool that allows the comparison of individual, family and neighbourhood factors within a single analysis (Gelman and Hill 2007).

Finally, the socio-economic gradient can be studied from a variety of perspectives over a rich array of outcomes. Preliminary analysis using educational achievement shows that siblings living in lower-income families have lower, but more highly correlated, scores than their higher-income counterparts. This finding ties in with research in behavioural genetics suggesting that a more restricted environment leads to higher within-family correlations in IQ and achievement scores (Loehlin et al. 2007). The possibilities for interdisciplinary collaboration are many.

REFERENCES

Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Lix, L., A. Hinds, G. DeVerteuil, J.R. Robinson, J. Walker and L.L. Roos. 2006. "Residential Mobility and Severe Mental Illness: A Population-Based Analysis." *Administration and Policy in Mental Health* 33(2): 160–71.

Loehlin, J.C., K.P. Harden and E. Turkheimer. 2007. "The Effect of Assumptions about Parental Assortative Mating and Genotype–Income Correlation on Estimates of Genotype–Environment Interaction in the National Merit Twin Study." *Behavior Genetics* 39(2): 165–69.

Mosteller, F. and J.W. Tukey. 1977. *Data Analysis and Regression. A Second Course in Statistics*. Reading, MA: Addison-Wesley.

Roos, L.L., M. Brownell, L. Lix, N.P. Roos, R. Walld and L. MacWilliam. 2008. "From Health Research to Social Research: Privacy, Methods, Approaches." *Social Science and Medicine* 66(1): 117–29.

Strohschein, L., N. Roos and M. Brownell. 2009. "Family Structure Histories and High School Completion: Evidence from a Population-Based Registry." *Canadian Journal of Sociology* 34(1): 83–103.

Willms, J.D. 1986. "Social Class Segregation and Its Relationship to Pupils' Examination Results in Scotland." *American Sociological Review* 51(2): 224–41.