

# International Journal of Population Data Science



Swansea University  
Prifysgol Abertawe

Journal Website: [www.ijpds.org](http://www.ijpds.org)

## Making Sense of a Hot Mess: Cleaning and Validating Messy Administrative Data to study Supportive Housing in Winnipeg, Manitoba

Yogendran, M<sup>1</sup>, Doupe, M<sup>1</sup>, Schultz, J<sup>1</sup>, and McDougall, C<sup>2</sup>

<sup>1</sup>Manitoba Centre for Health Policy

<sup>2</sup>Manitoba Centre for Health Policy, University of Manitoba

### Introduction

While supportive housing (SH) is an important alternate to nursing home (NH) use, these data have never been linked to administrative records in Manitoba. By conducting linkages to other administrative records, we describe a process for cleaning and validating SH data, in preparation to conduct policy-relevant research.

### Objectives and Approach

SH data (N=516 units) from Winnipeg were received at the Manitoba Centre for Health Policy (MCHP) in three different files. File 1 (2004-2008; 1005 records) contained monthly client snapshots. File 2 (2008-2010; 1336 records) contained application, move-in, cancellation, and move-out dates. File 3 (2010-2011; 729 records) contained one line of text for each record showing the application, processing, and move-in/cancellation date. We used overlapping data from these files plus linkages to other data sources (Manitoba Population Registry, nursing home data, and Vital Statistics) to clean and assess the accuracy of SH data.

### Results

The original files contained 2039 people with 3070 records. From this we excluded: i) 215 records with unusable Personal Health Identification Numbers; ii) 949 records with missing SH move-in dates; iii) 691 records that did not match to the Manitoba Health Registry; and iv) 25 records where data did not match to the NH, hospital, or Vital Statistics files. The result was 1190 people each with one record. SH move-out dates were often missing from these records. This field was imputed from other data sources (NH, Vital Statistics). Some people transferred between SH sites, and these data were retained in the same record. Aside from the first year of operation when capacity was low, most SH dwellings operated at 80-100% occupancy annually.

### Conclusion/Implications

Using several verification methods including linkages to other data sources, we successfully cleaned and verified the accuracy of the SH data for use at MCHP. High annual SH occupancy rates suggest that the file contains the vast majority of SH users, and can now be used in follow-up research.

