Multimorbidity Cluster Analysis Toolkit

January 2016

Contents

Background

In examining the issue of multimorbidity, previous literature has focused on the descriptive counting of individual diseases or the simplistic link between co-occurring pairs of diseases. However, the analysis of cumulative interactions and non-random associations between disease diagnoses will lead to a deeper understanding of the multidimensional burden of multimorbidity (Garin et al., 2014; Sinnige et al., 2013; Prados-Torres et al., 2012). This burden and impact of multimorbidity can be assessed from the patient, caregiver, health care provider and health system perspective. A computational cluster analysis can explore the distinct clinical profiles that exist within a sample of individuals living with multimorbidity.

This toolkit is designed to allow researchers to identify the distinct clusters or clinical profiles that exist within a sample of patients or individuals with multimorbidity. This toolkit can be adapted for use with varying diagnostic systems, multimorbidity definitions, sample sizes, target populations and settings. Its intent is to create a consistent approach to identify subgroups of patients or individuals with multimorbidity, based on co-existing conditions or diseases. This information is driven by the data, and the results should be assessed carefully. In fact, it is most ideal to incorporate clinical and contextual insight for interpretation. This information can be a helpful resource for research, clinical care and health policy decisions.

Development of This Toolkit

This toolkit was developed by a research team at Western University from the Departments of Epidemiology & Biostatistics and Computer Science. It was developed using data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), which is based at Queen's University and is funded by the Public Health Agency of Canada under a contribution agreement with the College of Family Physicians of Canada. The views expressed herein do not necessarily represent the views of the Public Health Agency of Canada. This toolkit is available to all academic researchers interested in exploring multimorbidity. When utilized in research projects, it is requested that acknowledgement (below) is made in any publications.

Bauer M & Nicholson K. Multimorbidity Cluster Analysis Toolkit. 2016.

Who Should Use This Toolkit?

This toolkit is intended for use by researchers interested in examining multimorbidity in a set of data (particularly in large, secondary datasets such as the CPCSSN electronic medical record dataset), and who are interested in determining the most frequently occurring permutations (ordered) and combinations (unordered) clusters of conditions/diseases in their sample. An example of an ordered cluster or permutation of multiple conditions/diseases is when a group of patients have received diagnoses in the same specific sequence: first diagnosed with obesity, then with hypertension, then with cancer. An example of an unordered cluster or combination of multiple conditions/diseases is when a group of patients have received the same three diagnoses (obesity, hypertension and cancer), but in varying order and without a specific sequence. For example, some patients may have been diagnosed with hypertension, then cancer, then obesity. This toolkit will conduct a patient-level computational cluster analysis to determine the frequency and type of mutually exclusive clusters of conditions/diseases among a sample of individuals with multimorbidity. This analysis could also be tailored (by tailoring the input data file) to explore the burden of multimorbidity among a specific subset of conditions/disease, such as patients living with diabetes mellitus or cancer. While these results could be used to inform health policies and resource allocation for complex patients, it is important to note that this analysis cannot assess or detect a causal link between condition/disease diagnoses.

What Does This Toolkit Contain?

As a companion to the Multimorbidity Cluster Analysis Tool, this toolkit contains the following items: summary of the Multimorbidity Cluster Analysis Tool; summary of input and output files; and Frequently Asked Questions. The Multimorbidity Cluster Analysis Tool (which consists of JAVA code and an executable file) has been developed and tested to support up to 150,000 individual records and up to 100 condition/disease categories. The simple setup of the input tables will allow for varying diagnostic systems to be used (e.g., ICD-9, ICD-10, SNOMED CT, Read Codes), as well as definitions of multimorbidity. The time (in days) elapsing between diagnoses can also be explored, if the date of first diagnoses is available within the data.

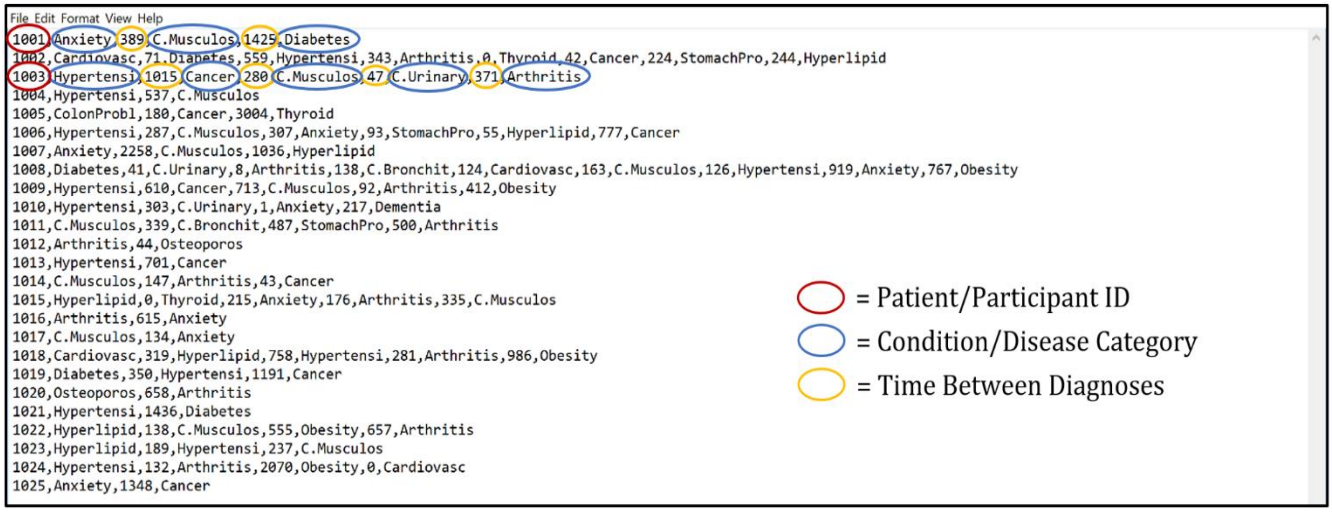
How Should This Toolkit Be Used?

Step One

1. Patient/Participant Identifier

- A unique identifier (ID) should be created for each patient/participant. This ID can be maintained from the original study ID (e.g., 00012345, 00012346) or can be created as a new unique identifier in the input file (e.g., 1, 2, 3).
- The patient/participant ID should begin each line, followed by condition/disease diagnoses that the patient/participant has received and the time occurring between each condition/disease diagnosis. Only one patient/participant ID should be included on each line, and the diagnoses and time variables (included on the same line) should correspond to each ID.
- This file can be created in a data management program (e.g., SAS, STATA, EXCEL) and saved as or exported as a ".txt (comma separated values)" type file.
- After the file has been saved, it is encouraged that the input data file is opened and assessed for appropriate structure and layout (displayed below).

Example Input Data File:



The input data file should consist of the following information:

Patient/Participant ID, Condition/Disease Category, Time Between Diagnoses

This data can be prepared in a SAS/STATA/EXCEL program and saved as a .txt file.

2. Condition/Disease Category

- The condition/disease categories included in the definition of multimorbidity should be created and finalized prior to creating the input data file. For example, all relevant ICD-9 or ICD-10 codes should be assigned to the category "Anxiety" for anxiety diagnoses. The condition/disease category names can be maintained, up to a maximum of ten characters (e.g., Anxiety, Cancer, Hypertensi). It is important that these condition/disease categories are maintained throughout the entirety of the .txt input file. These category names will also be used to create the output files.

3. Time Between Diagnoses

- The time occurring between each date of condition/disease diagnosis should be calculated (in days) and included in the input data file.
- The first date of diagnosis for each condition/disease should be used. The accuracy of these date of diagnosis should be assessed, and biases acknowledged if necessary.
- The time elapsing between condition/disease diagnoses can be calculated in the original dataset by the following equation: [Date of Diagnosis 2] – [Date of Diagnosis 1]. This calculation should be used to determine time elapsing between all condition/disease diagnoses, measured in days and the resulting values should be rounded to the nearest whole number.
- It is important to structure the input data file as follows:

  ID, DiseaseCategory1, Time1, DiseaseCategory2, Time2, DiseaseCategory3

  Where Time1 = Time (in days) elapsing between DiseaseCategory2 and DiseaseCategory1

  Time2 = Time (in days) elapsing between DiseaseCategory3 and DiseaseCategory2

The final input data file should be saved as a .txt file and named appropriately to be easily identifiable for use in the program (e.g., mmpatients.txt).
A new folder should be created that holds the .txt file and the Multimorbidity Cluster Analysis Tool.

Step Two

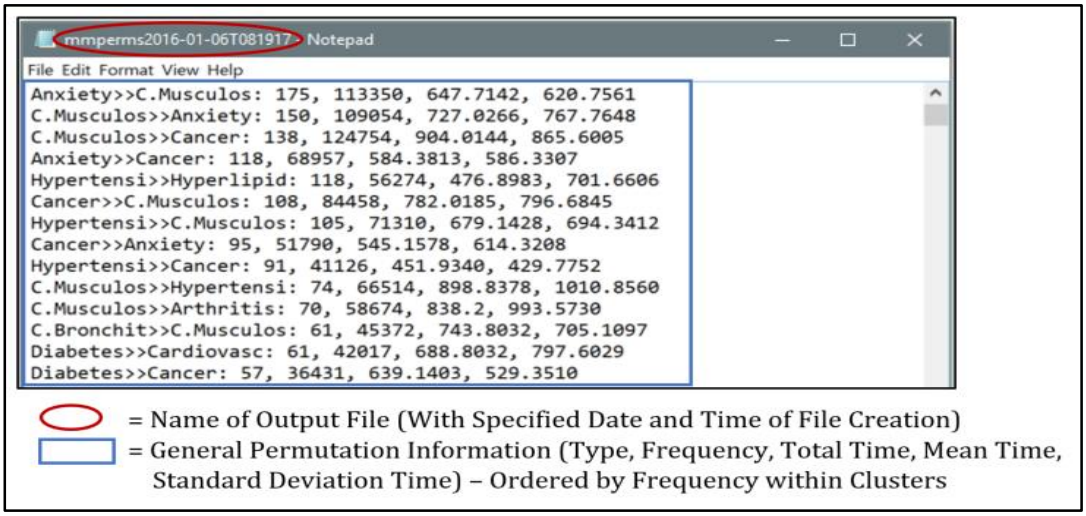
1. Running Program

- Once the final input data file has been prepared and saved as a .txt file, the Multimorbidity Cluster Analysis Tool can be run. To do so, the tool is accessible [from www.cpcssn.ca/]* or from www.csd.uwo.ca/faculty/bauer/ under "Multimorbidity Tool". **N.B. Placeholder, not yet available on CPCSSN website**
- To download the program, click on "mm cluster tool". When asked to save the program, select "Yes". The program will download onto the computer system and is labelled as "mm cluster tool.jar". The .jar program should be saved and moved into the same folder as your input data file with patient/participant data.
  - A JAVA runtime environment is required on the system. If the "mm cluster tool.jar" does not run, a JAVA runtime environment is needed. A JAVA runtime environment can be downloaded for free from: http://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html. To download, select the version that is required for the system (e.g., Windows x86). Install this JAVA runtime environment.
- To run the cluster analysis program, double click on the saved "mm cluster tool.jar" file. The program will first prompt for the input data file using an "Open" box. Select the appropriate input data file and select "Open".
- The program will produce a sequence of display messages, indicating completed steps of the program (e.g., Reading from file; Number of records processed; Number of permutations/combinations found; Writing permutations to file; Completed writing permutations; Writing combinations to file; Completed writing combinations; Processing completed). Select "OK" for each step, as the program pauses and waits for a user response.
- Each output file name indicates if it holds the permutations (mmperms.txt) or combinations (mmcombs.txt) and whether it holds detailed results (mmpermsDetails.txt or mmcombsDetails.txt). Each output file name also contains the date and time of file creation (see below). This means that consecutive runs of the program will produce uniquely named files and previous files will not be overwritten.

2. Output Files

- The program will automatically save the data output files (as .txt files) in the same folder as the .jar program and input data file.
- A total of four output files will be created and each are described further below. Example output data files are also included below.
  - The "mmpermsDATETIME.txt" output file contains all permutations (ordered clusters) of condition/disease diagnoses. The output is a sort list of permutations, which is presented in order from most frequent to least frequent for each group of patients/participants with the same number of diagnoses (e.g., 2 diagnoses, 3 diagnoses). These permutations are represented using the ">>" character, which indicates an additional diagnoses (in that specific sequence). The format of this output file is: Condition/Disease Permutation, Number of Occurrences (Number of Patients/Participants), Total Time (Cumulative from First to Last Diagnosis in Days), Mean Time (Days), Standard Deviation Time (Days).

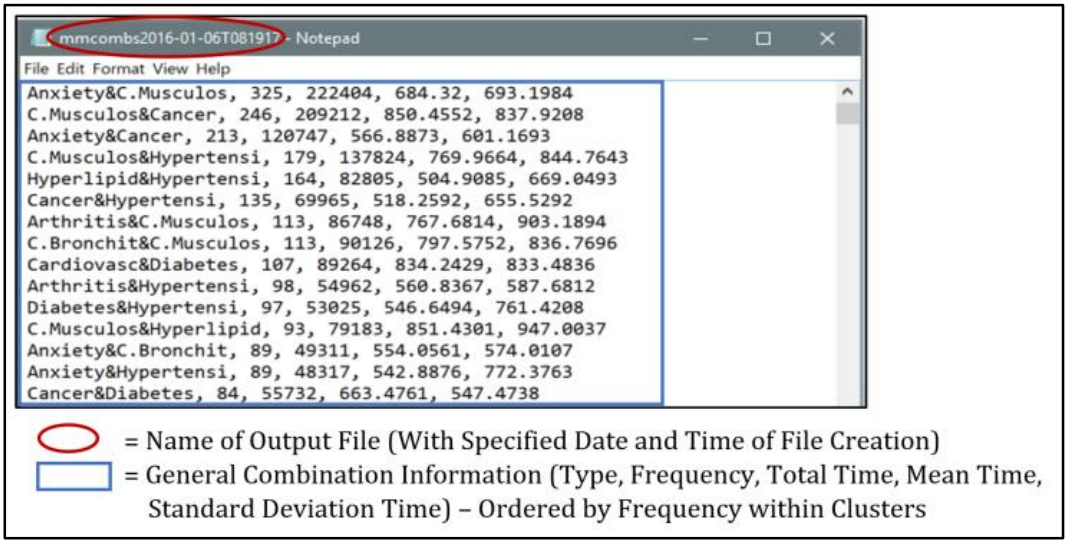
    Example Output Data File:



  - The "mmcombsDATETIME.txt" output file contains all combinations (unordered clusters) of condition/disease diagnoses. The output is a sort list of combinations, which is presented in order from most frequent to least frequent for each group of patients/participants with the same number of diagnoses (e.g., 2 diagnoses, 3 diagnoses). These combinations are represented using the "&" character, which indicates an additional diagnoses

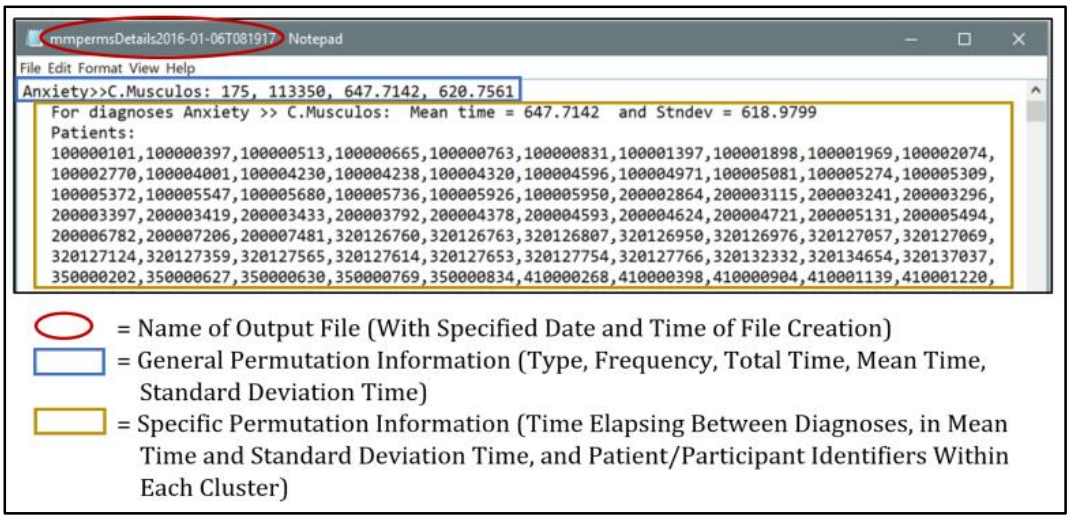
(regardless of sequence). The format of this output file is: Condition/Disease Combination, Number of Occurrences (Number of Patients/Participants), Total Time (Cumulative from First to Last Diagnosis in Days), Mean Time (Days) and Standard Deviation Time (Days).

Example Output Data File:



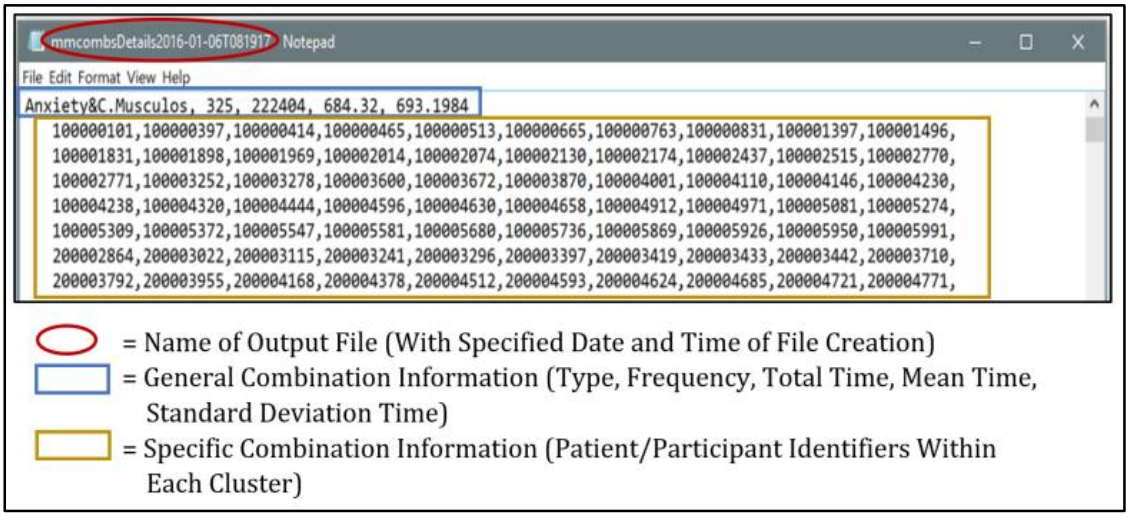
- The "mmpermsDetailsDATETIME.txt" output file contains the same permutations as the "mmpermsDATETIME.txt" output file. This includes the Condition/Disease Permutation, Number of Occurrences (Number of Patients/Participants), Total Time (Cumulative from First to Last Diagnosis in Days), Mean Time (Days) and Standard Deviation Time (Days). This output file also contains further details for each permutation. More specifically, the Mean Time (Days) and Standard Deviation Time (Days) is presented for each sequence of diagnoses within a permutation. The Patient/Participant IDs of individuals contained within these mutually exclusive clusters are also included in this output file.

  Example Output Data File:



- The "mmcombsDetailsDATETIME.txt" output file contains the same combinations as the "mmcombsDATETIME.txt" output file. This includes the Condition/Disease Combination, Number of Occurrences (Number of Patients/Participants), Total Time (Cumulative from First to Last Diagnosis in Days), Mean Time (Days) and Standard Deviation Time (Days). This output file also contains further details for each combination. More specifically, the Patient/Participant IDs of individuals contained within these mutually exclusive clusters are included in this output file.

  Example Output Data File:



- The output files can then be imported into data management programs (e.g., EXCEL) for further processing.

The four output data files will be saved as .txt files and consist of general and specific information of the permutations (ordered clusters) and combinations (unordered clusters) of condition/disease diagnoses. These output files can then be imported into data management programs (e.g., EXCEL) for further processing.

Frequently Asked Questions

Question 1. Do I have to abbreviate the condition/disease category names myself?

Answer: You may choose to shorten the condition/disease category names yourself or the program will automatically shorten the category names to ten characters.

Question 2. I cannot seem to find the "mm cluster tool.jar" file on my computer after accessing the tool online. Where is it located on my computer?

Answer: After accessing the tool online and downloading the file to your computer, the file may automatically be placed in your "Downloads" folder or to your "Desktop". You can also "Search" for the file on your computer. Once the file has been located, please relocate the file into the same folder that holds the input data file.

Question 3. The data that I will be using to create my input data file does not contain information on the date of diagnoses, so I cannot calculate the time between diagnoses. Can I still use this tool to determine the most frequently occurring clusters?

Answer: Yes, researchers who do not have data on the time between diagnoses can still use this tool. In order for the program to run properly, however, it is important to maintain a column for the time variable between each diagnoses (space holder = 0). For example, it is important to structure the input data file as follows:

ID, DiseaseCategory1, Time1, DiseaseCategory2, Time2, DiseaseCategory3

Where Time1 and Time2 = 0

Question 4. If I have a follow-up question or comment about the Multimorbidity Cluster Analysis Tool and/or Toolkit, is there a suitable way to submit these questions and/or comments?

Answer: Yes, further questions or comments about the Multimorbidity Cluster Analysis Tool and/or Toolkit can be directed to: mmclusteranalysis@gmail.com.

References

1. Garin N, Olaya B, Perales J, Moneta MV, Miret M, Ayuso-Mateos JL, et al. Multimorbidity patterns in a national representative sample of the Spanish adult population. PLoS ONE. 2014;9(1):e84794-803.

2. Prados-Torres A, Poblador-Plou B, Calderon-Larranaga A, Gimeno-Feliu LA, Gonzalez-Rubio F, Poncel-Falco A, et al. Multimorbidity patterns in primary care: Interactions among chronic diseases using factor analysis. PLoS ONE. 2012;7(2):e32190-202.

3. Sinnige J, Braspenning J, Schellevis F, Stirbu-Wagner I, Westert G, Korevaar J. The prevalence of disease clusters in older adults with multiple chronic diseases: a systematic literature review. PLoS ONE. 2013;8(11):e79641-53.