**Proposal for MSc (Research)**

**Supervisor**: Thomas Lumley t.lumley@auckland.ac.nz

**Research Student:** Junhuang Xue jxue533@aucklanduni.ac.nz

**Proposed study:** Anomaly detection in hierarchical datasets

**Background:**

Large datasets often implement a hierarchical structure, i.e. data are organised into a tree-like structure, where a higher hierarchy category may disaggregate into finer categories with lower hierarchy. For example, Geographic locations of New Zealand may first divide into North and South Island, and then into different regions, and so on. Such datasets are abundant in industry and government agencies and require attention. However such structure often focus on organisation and structure, and often create complications in statistical analysis.

We are interested in whether modelling the full hierarchy gives better detection of anomalies or than just modelling one category at a time. When you model too broad a category, you may miss detailed variation; when you model too narrow a category, you may have too little data with little significance. There may come a time when Analysts have to decide on which hierarchy may be best suited for attention. Finding how levels of hierarchies affect the rate of anomaly detection could provide insight into how hierarchical structure data behave and could have real-life implications.

Simulation data will be implemented for theoretical evaluations; Emergency department visits data will be implemented for evaluation of anomaly detection rate between hierarchies in a real-life setting. The project will provide some insights on how levels of hierarchies will affect anomaly detection rate and inform analysts or policymakers about how hierarchically structured data will behave.

**Suitable journal:**

For the assessment of Anomaly detection in hierarchical datasets, *Journal of the American Statistical Association* and *Australian & New Zealand Journal of Statistics* could be suitable. For the implementation, potentially *JMIR Public Health and Surveillance* and *The Journal of Emergency Medicine.*

**Resources required:**

We plan to use hospital emergency department visits over time data, from U.S (MMIC) or New Zealand (NMD). We will require permission to obtain these data and the application process are expected to take over a month. If we could not obtain our preferred datasets, we will try to obtain some relatively large, publically available datasets that has a variable with hierarchical structure, such as NZ Census or Thousand Genomes project. Usage of real life data is preferred, we will use simulation but it may not be adequate.