CITS3401 Data Exploration and Mining
Project 1

Medicare Australia Data Warehouse

Mitchell Pomery
21130887

April 28, 2014

Introduction

Medicare Australia wishes to use data from it's previous years to assist in making decisions to improve their services, analyze expenditure and detect individuals who are abusing their system. Each center stores information about visits in an Online Transaction Processing (OLTP) database, these are then collated at a state and country wide level. The patient, doctor, treatment and prescriptions for each visit are stored. This document outlines the data cube designed facilitate in the decision making processes of Medicare.

Requirements

The authors' interpretation of the requirements are listed below.

Object	**Restrictions**
Location	State or Territory in Australia
Center	3 Centers in each State/Territory
Patient	
Tests	Only one test will occur per visit.
Diseases	Only one disease will be diagnosed per visit maximum.
Referrals	Occur when a disease has been diagnosed.
Date	2006 to 2011, broken down into quarters.

Table 1: Requirements

Assumptions

Assumptions were made where the requirements were incomplete or insufficient, to simplify the schema and keep it manageable, and to make the scenario as realistic as possible.

1. Only a small number of patients, diseases, physicians, hospitals, specialists and pathology clinics exist.
2. Doctors are irrelevant, only the name of the clinic matters.
3. Patients will always visit a General Physician before seeing a specialist.
4. The cost of treatment, as well as the person or company who pays for the treatment is irrelevant.
5. People only visit medical centers in their own state.
6. All data is complete and easily available in the desired format.

Warehouse Schema

A star schema was designed to make the data cube simpler, and the queries faster than a snowflake schema or fact constellation. The date of the visit was broken into two dimensions, Year and Quarter, to allow for comparison between different years. This allows us to see which diseases reoccur at high rates each year, and when they occur.

The requirements also state that Medicare is interested in using this data for analyzing several areas of their business. They would like to be able to determine which patients frequently return for diagnosis and prescriptions and which doctors frequently refer patients to the same specialists. The discovery of trends in diseases and referrals, and the discovery of outliers in patients, medical centers and treatment times would be beneficial for Medicare.

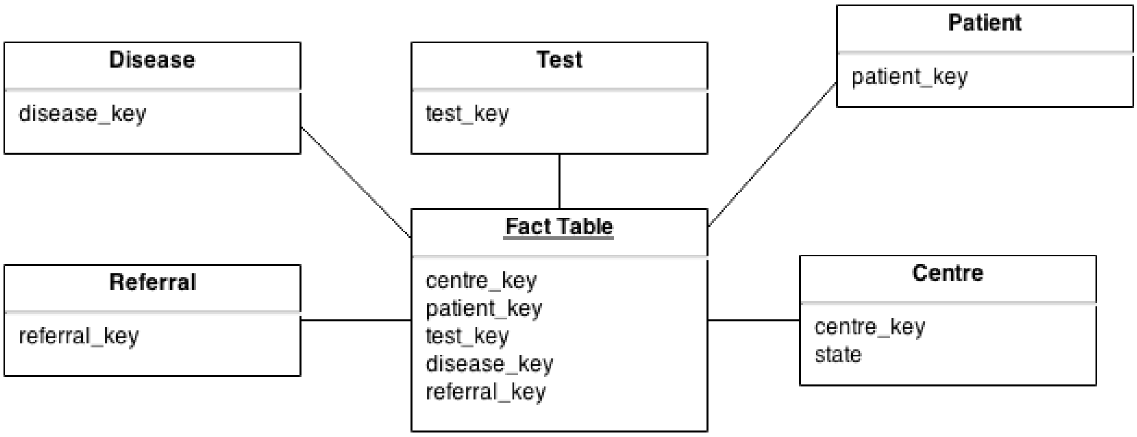


Figure 1: Fact Table in Star Schema

Prototype Warehouse

Data Generation

Prototype data was generated using the Python script `gendata.py`. It takes sets of words for diseases, clinics, names, medical tests and states, creates random people and outputs information about their visits from 2006 to 2011. An attempt has been made to make it reflect reality, by restricting the average persons visits to a couple of times a year, charging based on the location visited and making a small percentage of users abuse the system each year.

Data Analysis

`Palo.xlsx` was created to import the data into the OLAP database for analysis, however due to issues with PALO, it was soon discarded. `search.py` was written to do data analysis on the data generated by `gendata.py`, and the output was manually entered into `Python.xlsx` . `search.py` enumerates all the possible elements for each dimension into several arrays, then counts the number of transactions that satisfy certain conditions, such as state, and places it into a table.

Medicare Australia is then able to view the output data and determine what action, if any, needs to occur. By looking at the visits per patient per year, Medicare will be able to determine if any individuals are using their services abnormally, they will then be able to investigate if the visits are legitimate. By looking at the yearly breakdown of clinics, it is possible to see trends in patient numbers for each year. This is useful as it will allow Medicare to determine what areas are in need of upgraded infrastructure.

Scenarios

Abuse of Services

By analyzing how many times patients visits medical centers, and clustering them, it is possible to identify outliers. These people visit the medical centers in a pattern that is dissimilar to the general population, and so looks suspicious. By then investigating these people, Medicare will be able to determine if they are abusing the services provided and take action.

Insufficient Infrastructure

Analyzing the number of visits to each clinic, it is possible for Medicare to see what clinics are under heavy throughput. By further analyzing the amount of time spent with patients, and comparing it to other clinics, it is possible to determine if the clinic needs more staff and better training, or if a new clinic should be opened nearby to reduce the throughput.

Virus Epidemics

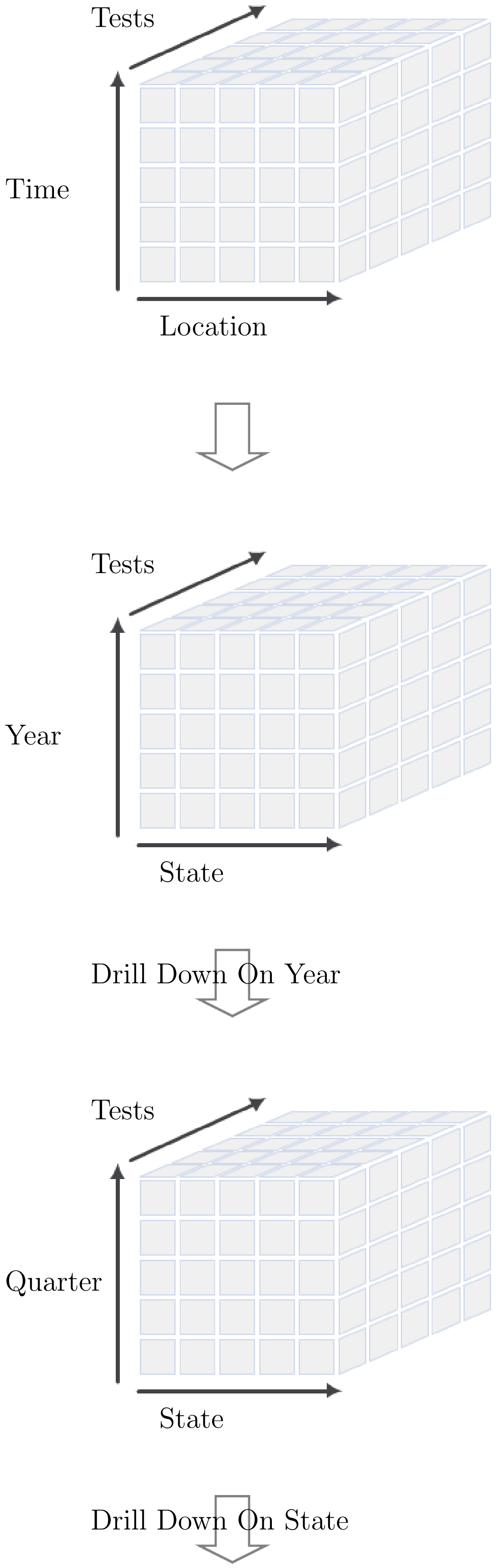
Looking at the number of outbreaks of diseases over time and analyzing the trends will allow Medicare to see what areas of Australia the disease travels through. This will allow Medicare to implement plans to stop the spread of the disease and to help eradicate it.

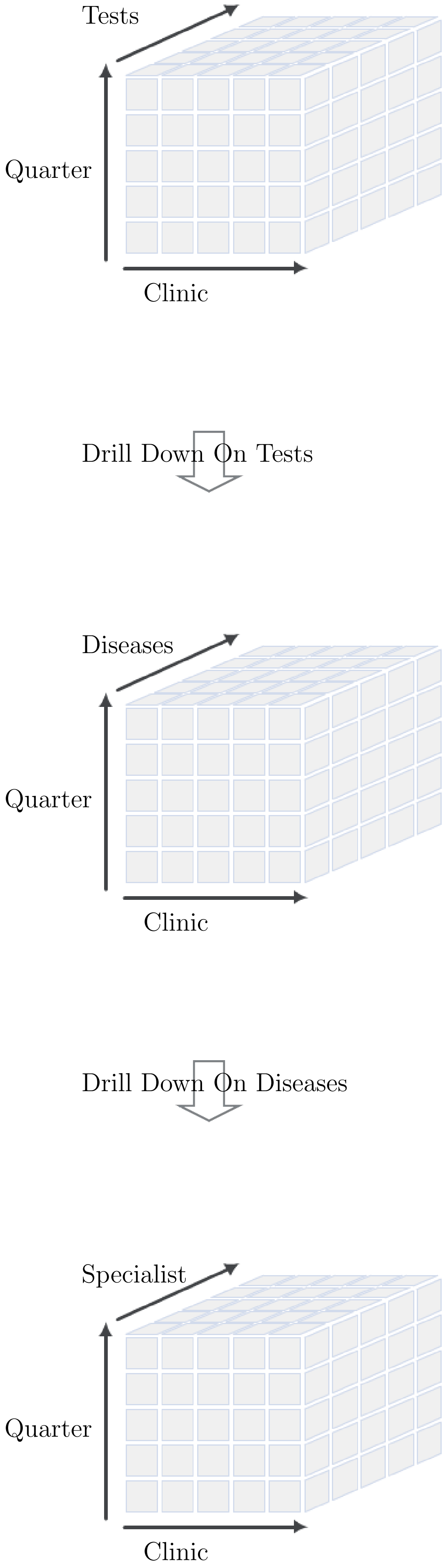
Patient Care

The time spent with patients is important to Medicare, and by analyzing it, clinics that do not spend enough time with their patients can be determined. Further analysis on these clinics can help determine if it is due to them understaffed or the doctors being unwilling to provide the best service possible.

Data Cube

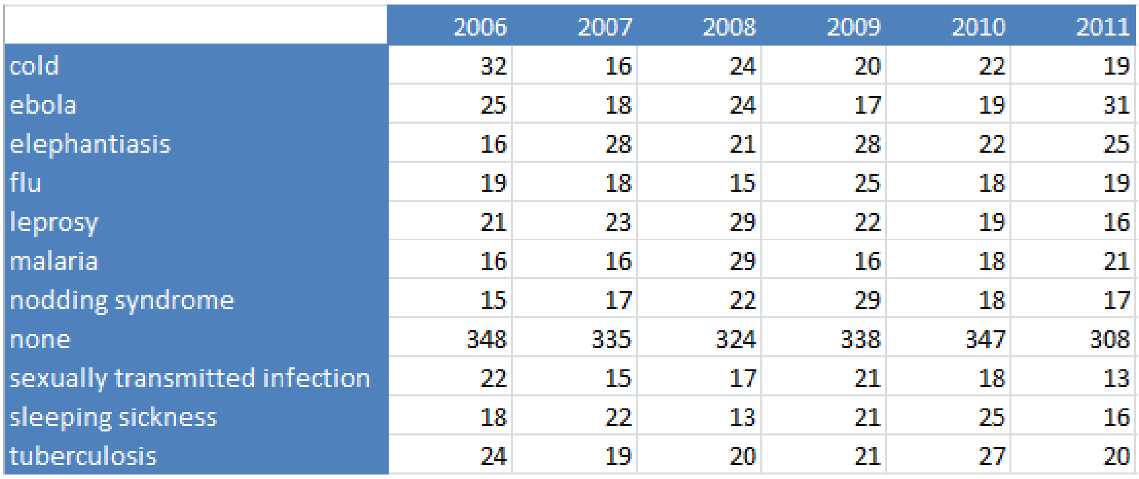
A visualization of the data cube is provided below:




Tests
Quarter
Clinic
Drill Down On Tests
Diseases
Quarter
Clinic
Drill Down On Diseases
Specialist
Quarter
Clinic

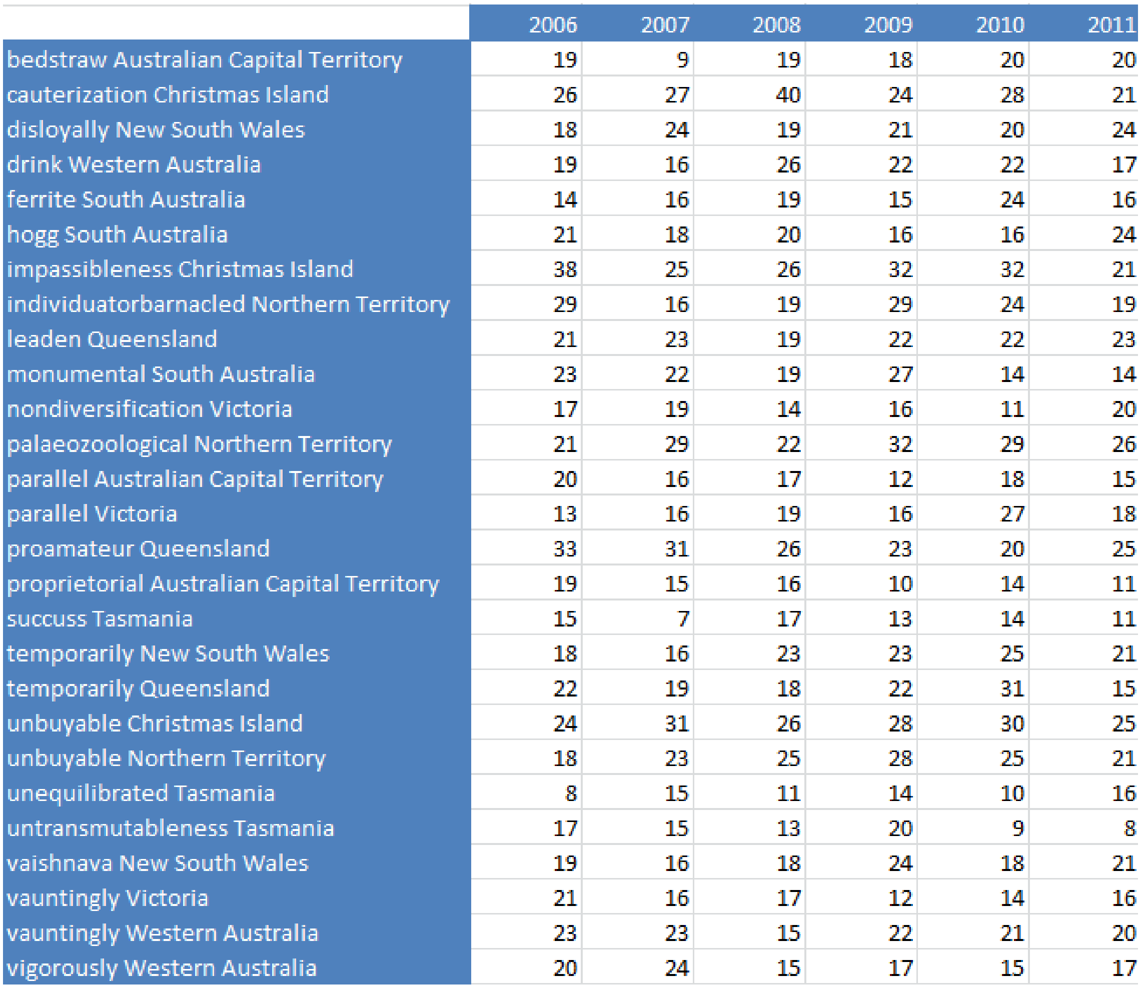
Example Output

A visualization of the data cube is provided below:



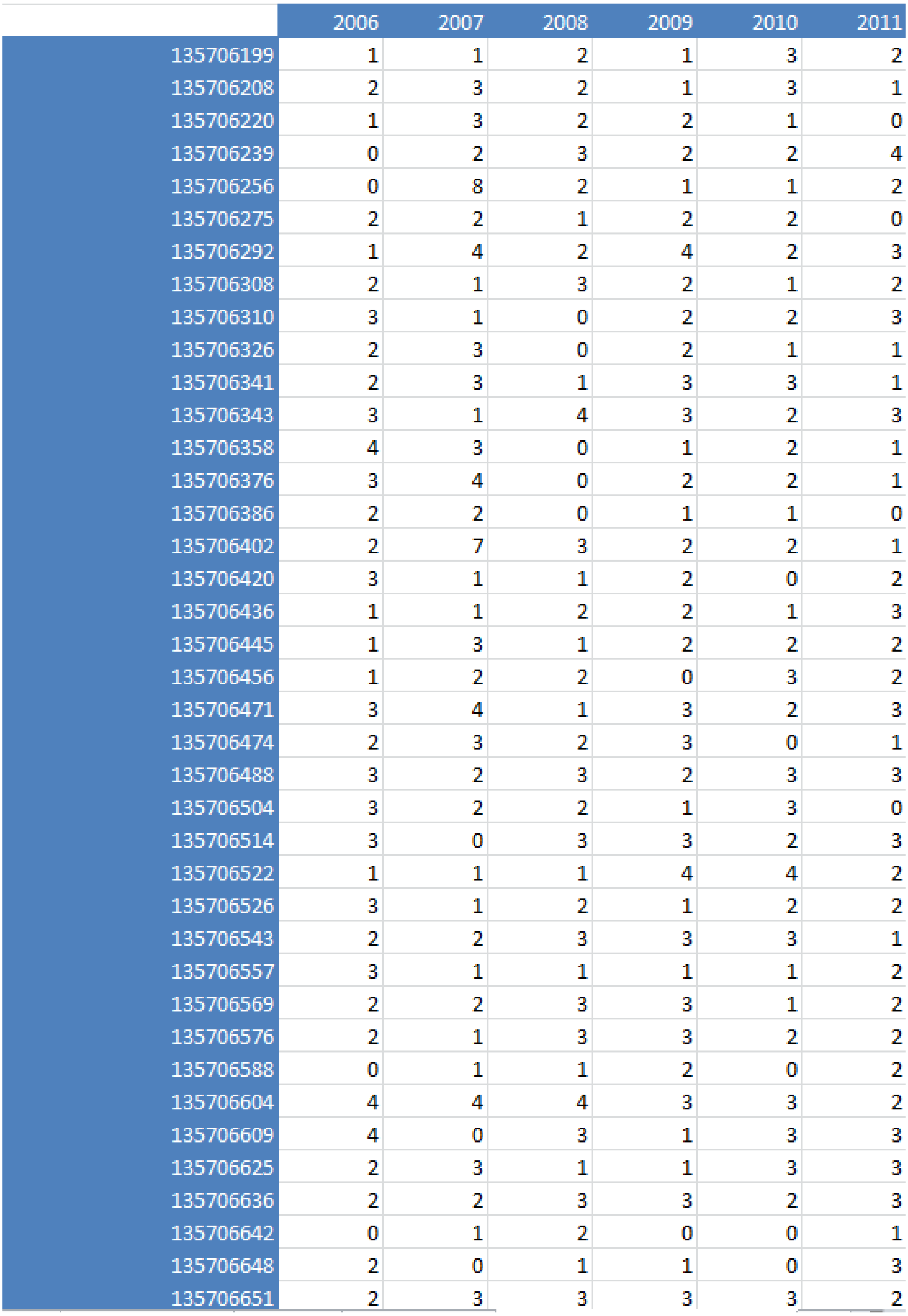
	2006	2007	2008	2009	2010	2011
cold	32	16	24	20	22	19
ebola	25	18	24	17	19	31
elephantiasis	16	28	21	28	22	25
flu	19	18	15	25	18	19
leprosy	21	23	29	22	19	16
malaria	16	16	29	16	18	21
nodding syndrome	15	17	22	29	18	17
none	348	335	324	338	347	308
sexually transmitted infection	22	15	17	21	18	13
sleeping sickness	18	22	13	21	25	16
tuberculosis	24	19	20	21	27	20

Figure 2: Number of patients diagnosed with a disease



	2006	2007	2008	2009	2010	2011
bedstraw Australian Capital Territory	19	9	19	18	20	20
cauterization Christmas Island	26	27	40	24	28	21
disloyally New South Wales	18	24	19	21	20	24
drink Western Australia	19	16	26	22	22	17
ferrite South Australia	14	16	19	15	24	16
hogg South Australia	21	18	20	16	16	24
impassibleness Christmas Island	38	25	26	32	32	21
individuatorbarnacled Northern Territory	29	16	19	29	24	19
leaden Queensland	21	23	19	22	22	23
monumental South Australia	23	22	19	27	14	14
nondiversification Victoria	17	19	14	16	11	20
palaeozoological Northern Territory	21	29	22	32	29	26
parallel Australian Capital Territory	20	16	17	12	18	15
parallel Victoria	13	16	19	16	27	18
proamateur Queensland	33	31	26	23	20	25
proprietorial Australian Capital Territory	19	15	16	10	14	11
succuss Tasmania	15	7	17	13	14	11
temporarily New South Wales	18	16	23	23	25	21
temporarily Queensland	22	19	18	22	31	15
unbuyable Christmas Island	24	31	26	28	30	25
unbuyable Northern Territory	18	23	25	28	25	21
unequilibrated Tasmania	8	15	11	14	10	16
untransmutableness Tasmania	17	15	13	20	9	8
vaishnava New South Wales	19	16	18	24	18	21
vauntingly Victoria	21	16	17	12	14	16
vauntingly Western Australia	23	23	15	22	21	20
vigorously Western Australia	20	24	15	17	15	17

Figure 3: Number of visitors to each clinic



	2006	2007	2008	2009	2010	2011
135706199	1	1	2	1	3	2
135706208	2	3	2	1	3	1
135706220	1	3	2	2	1	0
135706239	0	2	3	2	2	4
135706256	0	8	2	1	1	2
135706275	2	2	1	2	2	0
135706292	1	4	2	4	2	3
135706308	2	1	3	2	1	2
135706310	3	1	0	2	2	3
135706326	2	3	0	2	1	1
135706341	2	3	1	3	3	1
135706343	3	1	4	3	2	3
135706358	4	3	0	1	2	1
135706376	3	4	0	2	2	1
135706386	2	2	0	1	1	0
135706402	2	7	3	2	2	1
135706420	3	1	1	2	0	2
135706436	1	1	2	2	1	3
135706445	1	3	1	2	2	2
135706456	1	2	2	0	3	2
135706471	3	4	1	3	2	3
135706474	2	3	2	3	0	1
135706488	3	2	3	2	3	3
135706504	3	2	2	1	3	0
135706514	3	0	3	3	2	3
135706522	1	1	1	4	4	2
135706526	3	1	2	1	2	2
135706543	2	2	3	3	3	1
135706557	3	1	1	1	1	2
135706569	2	2	3	3	1	2
135706576	2	1	3	3	2	2
135706588	0	1	1	2	0	2
135706604	4	4	4	3	3	2
135706609	4	0	3	1	3	3
135706625	2	3	1	1	3	3
135706636	2	2	3	3	2	3
135706642	0	1	2	0	0	1
135706648	2	0	1	1	0	3
135706651	2	3	3	3	3	2

Figure 4: Number of visits by an individual