Mining electronic medical records

ICD 10 Code
Certain infectious and parasitic diseases (A00
Intestinal infectious diseases (A00-A09)
A00 Cholera
A01 Typhoid and paratyphoid fevers
A02 Other salmonella infections
A03 Shigellosis
A04 Other bacterial intestinal infections
A05 Other bacterial foodborne intoxications
A06 Amoebiasis
A07 Other protozoal intestinal diseases
Tuberculosis (A15-A19)
A17† Tuberculosis of nervous system
A18 Tuberculosis of other organs
A19 Miliary tuberculosis
Certain zoonotic bacterial diseases (A20-A28)
A20 Plague
A21 Tularaemia
A22 Anthrax
A23 Brucellosis
A24 Glanders and melioidosis
A25 Rat-bite fevers
A26 Erysipeloid
A27 Leptospirosis
Other bacterial diseases (A30-A49)
A30 Leprosy [Hansen's disease]
A31 Infection due to other mycobacteria
A32 Listeriosis
Electronic Medical Record System
Home
Show
Print
Send
Preview
Edit
Save
Quit
Personal Information
Social Information
Insurance
Diagnosis
Treatment
Medical History
Healthcare Calendar
Schedule
Appointment
HN:
Patient Name
Doctor :
Operative Note :
Personal Information
Search
HN
Name
Doctor
Search

Source: This is an entirely fictional data set created by R.W. Oldford.

Problem

Imagine a disease epidemic has broken out in the population of some country. It is thought that adults under the age of 60 appear to be particularly vulnerable. Both men and women contract the disease and need to be treated. Those who go untreated die within 5 days of contracting the disease.

The medical community has tried two quite different approaches to treat patients having the disease – call these `Treatment A` and `Treatment B`. For the health of the country, it is important to determine which of these two treatments is more effective.

Plan

To investigate which is the better treatment, it was decided to mine the medical records from another country of those who had contracted the disease and had been treated with one of the two treatments. Patients treated with either A or B survive the disease and recover fully; some however still die.

Electronic medical records available from several of the more populous districts are accessible. These can be searched to provide records from patients that have received treatment. It was decided that there should be the same number of records drawn for each treatment.

Moreover, concern was raised that the investigation have gender balance (i.e. equal numbers of males and females). So, to make sure that both sexes were equally represented, it was also decided that the number of female patients would be the same as the number of male patients.

Finally, it was desirable to detect even small differences in success rates of the two treatments since small differences could mean many more lives being saved. A sample size of about $n = 3000$ was decided on. Records would be collected until 3,000 were found, 1500 of which were treated with `A`, 1500 with `B`, and there were equal numbers of males and females in the study.

Data

In this stage, the plan is executed. Instead of 1500 records of treatment `A` and `B`, 1600 of each were found. The number of males and females was kept equal (now 1600 of each sex).

The process was to search the records in order, selecting those first encountered to get 1600 for each treatment and 1600 of each sex. Many records might be discarded whenever one quota was met and the search continued to meet the other quotas. It was also noticed that the patient's age was available for each record, so that the effect of treatment on younger and older adults might also be considered.

The counts which fell into the various categories were assembled into the `medicalRecords` data set in the loon.data package.

```
data("medicalRecords", package = "loon.data")
```

The first few rows of the data set are

```
head(medicalRecords, 4)
```

```
##     Age    Sex Treatment   Outcome Freq
## 1 20-39   Male         A Recovered   60
## 2 40-59   Male         A Recovered   90
## 3 20-39 Female         A Recovered  270
## 4 40-59 Female         A Recovered  540
```

out of 16 rows, where the frequency (`Freq`) of each combination has been compiled from the 3200 selected records.