BrainIT 2018 – Datathon

Introduction

The BrainIT group (www.brainit.org) has collected clinical and minute by minute physiological data from 261 adult patients with traumatic brain injury. The data definition is described in the data definition document downloadable from [1] and further described in the paper also downloadable from [2].

New analysis methods and models continue to be developed, many of them frequently known only to specific engineering, statistical, mathematical or computer science domains. The BrainIT group comprise predominately clinical researchers who recognize the usefulness of opening up the dataset to researchers from non-clinical specialties.

Datathon Agenda

To foster developing and assessing fresh approaches to analysis of this unique dataset and to demonstrate the proof of concept of this approach for establishing new collaborations, the BrainIT group opened up their datasets to a 2-day datathon as part of the BrainIT group annual meeting on Thursday 6th to Friday 7th December 2018 hosted at the Usher Institute for Population Health Sciences and Informatics at 9 Bioquarter [4].

Five data scientists participated in this pilot datathon. Three were PhD students from Edinburgh, one was an experienced data scientist from industry (Aridhia) and one was a graduate statistician on a training scholarship from the Ivory Coast.

Datathon Stages

1. Two days before the start of the meeting, the BrainIT data definition document was emailed to all participants. It is expected that the participants studied the schema to familiarize themselves with the structure, content, data types and element names. Importantly, as this is a domain specific dataset, datathon members were expected to compile a list of questions about the data that they can pose to BrainIT clinicians who will be available to answer any domain specific questions on the first day of the BrainIT meeting.

2. On the morning of the first day of the meeting (Thursday 6th December) and at the start of the meeting there was an initial two-hour session where two clinicians(Prof Per Enblad and Dr Chris Hawthorne) posed two specific clinical questions which formed the data challenge for the group. There then followed a 90 minute period where questions arising from the data scientists about the posed research questions or the data structure/content/clinical content could be addressed at the opening session. At the termination of this opening session, datathon members broke away into a nearby room (Ness Room) at the Usher Institute and spent the rest of the afternoon downloading

the data and becoming familiar with the datasets. BrainIT data scientists were available over lunch to answer any questions.

3. The second day of the datathon challenge was hosted in the main meeting room and after working for a further couple of hours, the scientists made informal presentations of their interim analysis approaches to available clinicians. Analysis continued until 4 pm when the data scientists made final presentations of their analyses approaches and interim results to addressing the two research questions.

Data Description Overview

In summary, the data consists of three classes of data: a) "One-off" demographic and clinical data elements collected only once per patient, b) "Episodic" clinical data items collected more than once per patient at un-predictable times and c) "Periodic" time-series data items collected at a frequency of typically once per minute. The latter data type is predominately physiological monitoring data acquired from bedside monitors recording vital signs summary measure data from patients during their acute management in neuro-intensive care following their traumatic brain injury. This physiological data is the largest component of the dataset with approximately 2 million records.

The data is stored in 9 tables archived as separate R data repositories (.RDS). Each table contains a common patient study ID identifier. The tables are stored in an FTP accessible data repository. The table names are:

1. demographic.rds
2. daily_observations.rds
3. icu_monitoring.rds
4. lab_results.rds
5. neurological.rds
6. other_clinical_events.rds
7. physiological.rds
8. surgery.rds
9. target_therapies.rds

Importantly, the BrainIT dataset contains standardized clinical outcome data (Glasgow Outcome Scale) obtained at 6 months post injury for the majority of patients thus enabling prognostic models to be developed and tested.

The BrainIT website contains a list of peer-reviewed papers of analyses already conducted on this dataset [3].

Questions Posed

Q1. Has the incidence of abnormal blood pressure, ICP or CPP across BrainIT centres changed between the project data collection periods ending 2005 (EU Grant 1) and 2010 (AvertIT project)?

A) What are the distributions of the physiology between the two datasets?
    Are there:
                - Centre Effects?
                - Age?
                - Diffuse Vs Focal?
                - First 24 hour Vs First 72 Hours?

B) Using the time-series signals, can one use a machine learning/pattern recognition approach to identify possible differences within or between the two datasets?
        - supervised approaches – prediction of future events/outcome
        - unsupervised approaches – eg: clustering

Q2. Intracranial Pressure (ICPm) is a physiological mechanism that also is subject to circadian rhythms and as such might be expected to change periodically in value over a 24 hour (or longer/shorter?) period. Are there detectable time-dependent periods to ICP?

        Is there any influence of other factors:
                ·   Age?
                ·   Injury? (GCS, Pupils, Diffuse/Focal?)

        Are there any differences between the two datasets?

Participants

There were the following five data scientist's participating in this Datathon event:

1. Evangelos Kafantaris, Edinburgh University Engineering PhD Student.

2. Maixent ASSI: Aberdeen University - Visiting Statistician

3. Harry Peaker: Aridhia Ltd, Edinburgh. Data Scientist

4. Elsie Home: Usher Institute, Edinburgh - PhD Student

5. Michael Camilleri - Edinburgh Informatics Institute, PhD Student.

Results

Below are excerpts from two of the presenting groups on their interim analysis approaches.

The full reports can be downloaded from the site: http://www.brain-it.eu/icuprog/

Group 1.

This group focused chiefly upon Q2 (circadian rhythms).

The plot below shows the distribution across all patients of a frequency based analysis of the three "dominant frequencies" present over the duration of their monitoring period.

It can be seen that there appears to be a predominant periodic event at around 2-4 hours and also a peak at 24 hours. On discussion with the clinical staff, the former may be related to standard nursing manouvers such as "turning the patient" which occurs on a regular basis in patients with a frequency of once ever 2-3 hours. The peak at 24 hours could similarly be part of a daily intervention by nursing or medical staff. No clear circadian effect could be identified, nor was there any obvious correlation with clinical outcome (based on GOSe at 6 months post injury). Upon discussion, it was felt that other approaches beyond that of Fourier analysis might be considered which are less dependent upon the "stationarity" assumption about the time-series data. Wavelet analysis is one approach which could be considered.
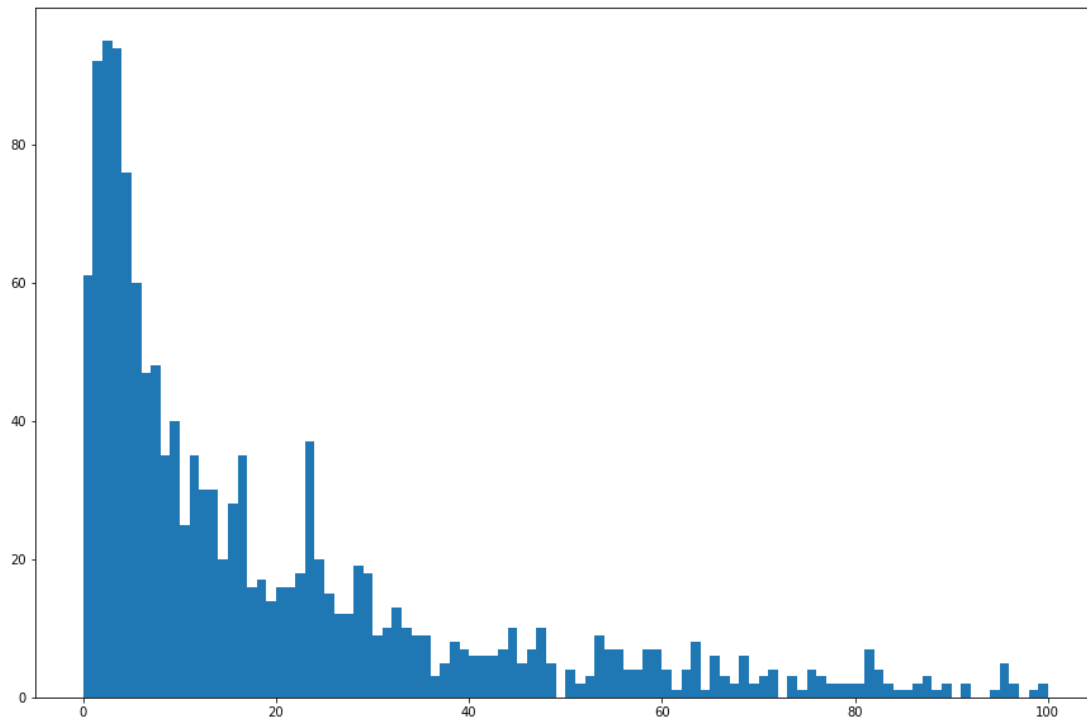
Group 2.

This group focused chiefly upon Q1 Data distribution specifically on the demographics/injury variables and utilised a type of cluster/principal components analysis called "t-SNE".

The demographic features initially selected for modeling were the presence of additional extra-cranial injury into groupings of "Chest", "Facial", "Limb", "Pelvic", " Abdominal", "Spinal" and "Other".
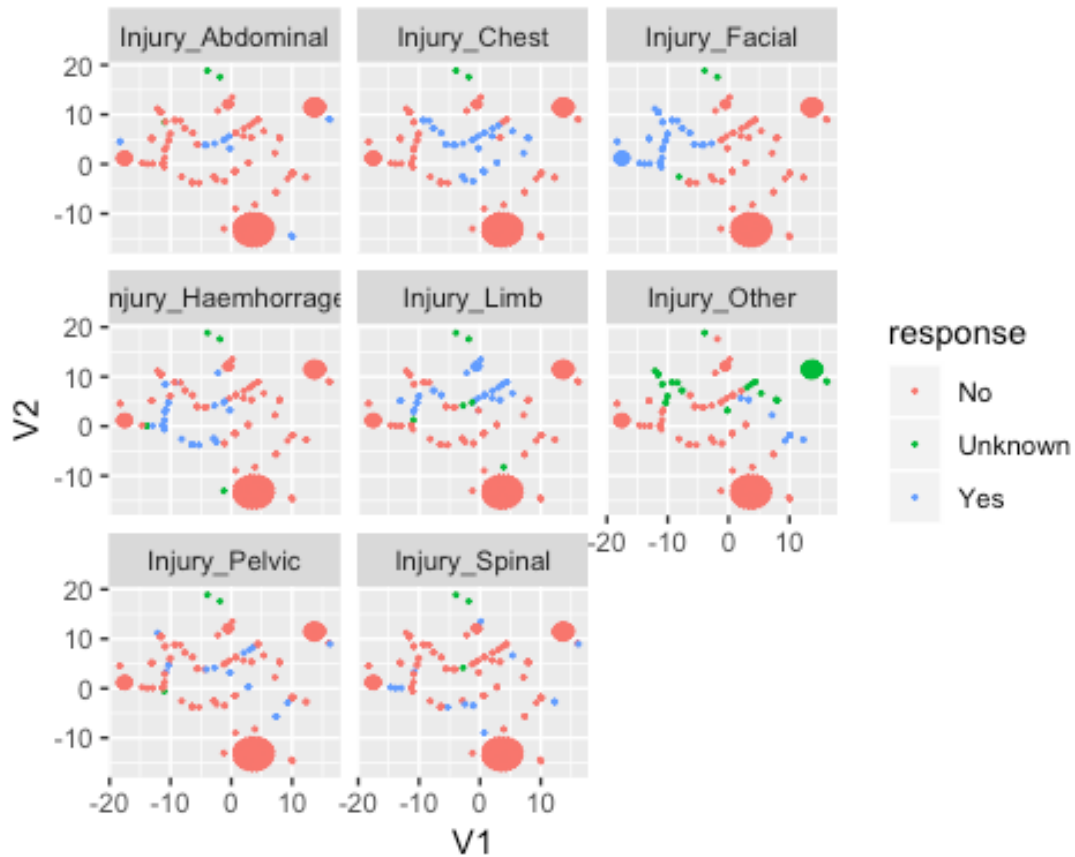
One can see from this type of visualization approach that patients in the large cluster at the bottom have No across all injury variables. The small cluster of patients on the left all have facial injuries, the small cluster of patients at the top all have limb injuries etc. It looks as though most of the patients scattered in the middle have Yes to at least one of the injury variables.

In the plot below, one can see an attempt to assess if "Type of Trauma" can explain in part some of the clustering of the data.
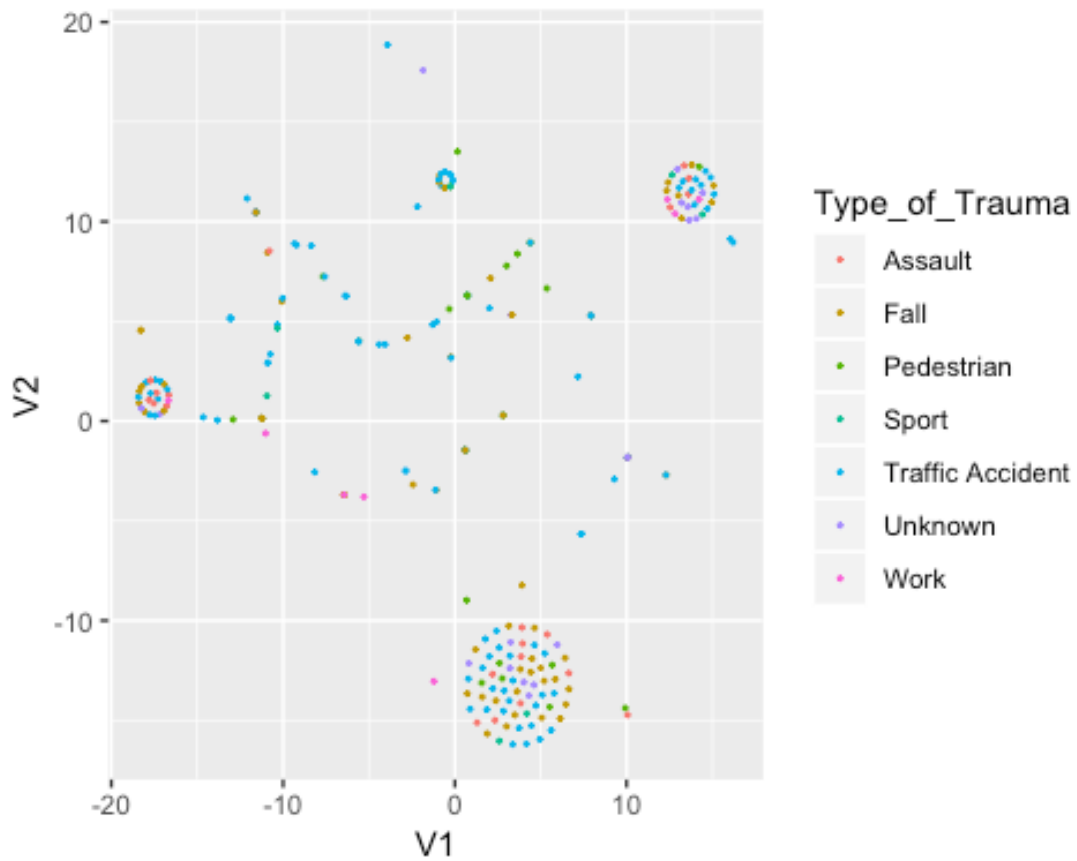
Discussion of this approach with the clinical staff highlighted how interesting this type of visualization approach can be and as with all clinical datasets, dependent on domain knowledge for assessing explanatory variables. It was felt this type of approach merits further investigation with other demographic factors and also with derived summary measures from the time-series data from the BrainIT datasets.

Discussion

The participants (both Data Scientists and Clinicians) were delighted with how stimulating this initial experiment with collaboration between external data scientists and the BrainIT clinicians had been. In particular, the clinicians were impressed with how quickly the data scientists managed to get to grips with the data and its clinical context.

This exercise has shown the feasibility of the approach of bringing in and collaborating with external data scientists who may not have worked within the specific critical care domain before. Even though only a few hours notice of the research question was given and less than 24 hours of analysis time provided, some potentially interesting and novel approaches to analyses of this dataset were identified.

Whats Next?

As a result, we propose to hold regular datathons at our future BrainIT meetings, but identify key questions well in advance and give interested data scientists sufficient time to make more substantive analyses which can be discussed more fully during a datathon session. Three of the data-scientists who participated in this datathon are planning on continuing with their analyses and already one formal BrainIT database analysis has been proposed and added to our web-site project list.

References

[1] http://www.brain-it.eu/files/Brain-IT_Core_Dataset_Operations_Manual_22_Sept_2010.pdf

[2] http://www.brain-it.eu/files/bit_piper_coredata_concept_2003.pdf

[3] http://www.brain-it.eu/publications

[4] https://www.ed.ac.uk/usher/

[5] http://www.brain-it.eu/criteria

Photos of Meeting