**Project Definition and Research Proposal**

15th April 2019

John H Booth (13133420)

BBK\_BUCI058D7\_1819 - MSc Data Science Project

MSc in Data Science

Birkbeck, University of London, UK

Contents

[1 Introduction 3](#_Toc5045002)

[2 Literature Review 4](#_Toc5045003)

[2.1 Background 4](#_Toc5045004)

[2.2 Data Extraction 5](#_Toc5045005)

[2.3 Data Wrangling 5](#_Toc5045006)

[2.4 Analytics 5](#_Toc5045007)

[3 Project Aims 7](#_Toc5045008)

[4 Project Description 8](#_Toc5045009)

[4.1 Data Extraction and Cleaning 8](#_Toc5045010)

[4.2 Exploratory Data Analysis and Data Wrangling 8](#_Toc5045011)

[4.3 Feature Selection 9](#_Toc5045012)

[5 System Architecture 10](#_Toc5045013)

[5.1 Originating System 10](#_Toc5045014)

[5.2 Analytic Platform 10](#_Toc5045015)

[6 Software Languages and Tools 11](#_Toc5045016)

[6.1 Python 11](#_Toc5045017)

[6.2 SQL 11](#_Toc5045018)

[6.3 R 11](#_Toc5045019)

[7 Methodology and Work Plan 12](#_Toc5045020)

[8 Ethics Approval and Data Governance 14](#_Toc5045021)

[References 15](#_Toc5045022)

[Bibliography 16](#_Toc5045023)

[Appendix A – Microsoft Access Structure of PM Research Database 17](#_Toc5045024)

[Appendix B – Event Attribute Value Model 18](#_Toc5045025)

[Appendix C – Summary of output from initial Sprints 20](#_Toc5045026)

# 1 Introduction

Great Ormond Street Hospital for Children NHS Trust (GOSH) is the country’s leading centre for treating sick children. With the UCL Great Ormond Street Institute of Child Health, GOSH is the largest centre for paediatric research outside the US.

Specialist Paediatric Pathologists perform perinatal, infant and childhood post-mortems including hospital referrals, forensic cases and those on behalf of Her Majesty’s Coroner.

The Pathology Department has established a research database containing details of all post-mortems performed between 1996 and 2017.  The database was originally used specifically for research into Sudden Unexpected Death in Infancy (SUDI). Since then it has been utilised for a number of other projects investigating SUDI, stillbirths and various aspects of paediatric autopsy procedure.

Currently the database holds 7000 records, each record representing an individual post-mortem. Up to 300 items of data can be defined for each post-mortem.

The purpose of this project is to use data science analytic techniques to develop operational strategies that can be applied to paediatric post-mortems to prioritise which data is required to achieve the target of specifying the cause of death.

This document consists of a literature review that covers the background to the project, the schema used to extract data from the research database, the data wrangling required to prepare the data for analytics and finally a deep dive on the chosen analytical methods. Following on from the review the documents details the projects aims, a detailed description of the project including the system architecture and software languages and tools to be employed on the project and a methodology and work plan.

# 2 Literature Review

In the literature review I intend to consider the following aspects of the project:

* Background – what is a post-mortem, what are the specific aspects of a paediatric post-mortem, how a paediatric post-mortems are categorised and the role GOSH plays as a national and international centre for paediatric post-mortem research.
* Data Extraction – how this project will handle the specific nature of data extraction and the use of the Entity Attribute Value schema for structuring data for the move from the research gathering database to the datasets ready for analytics.
* Data Wrangling – The handling of numeric variables in a categorical dataset, with the added complications of organ measurements in cases that cover the range from foetus to young children.
* Analytics – Why Decision Trees have been selected as the primary analytic method for this project, why ensemble techniques will be utilised such as Random Forests and Gradient Boosting.

## 2.1 Background

From the NHS web site [*NHS Contributors (2018)*] we get the following information:

“A post-mortem examination will be carried out if it's been requested by:

* a coroner – because the cause of death is unknown, or following a sudden, violent or unexpected death
* a hospital doctor – to find out more about an illness or the cause of death, or to further medical research and understanding.”

Post-mortems are carried out by Pathologists, doctors who specialise in the causes of diseases. Paediatric post-mortems have their own specific issues as explained on the Royal College of Pathologists [*RCPath Contributors (2018)*]:

“Paediatric and perinatal pathology is concerned with identification of disease in the fetus, infant and child. It is age-specific rather than organ-specific and includes investigation of that organ unique to the fetus, the placenta. The spectrum of disease in this age range is very different from that seen in adults and the interaction of congenital malformation and growth of the child interact to produce unique pathology.”

The Lullaby Trust, a charity that supports parents who have suffered the sudden loss of a child support research in this field gives a detailed breakdown of the different categories or presentations of post-mortems [*Lullaby Trust (2018)*]:

* TOP: Termination of pregnancy so the patient has not reach full term less than 24 weeks
* Still birth: 24 weeks to full term
* SUDI: Patients less than one year old
* SUDC: Patients over 1 year.

Great Ormond Street Hospital, Histopathology Department is both a national and international centre for paediatric post-mortem research and I have cited a number of research papers produced on data gathered at GOSH: [Ben‐Sasi, et al, 2013], [Weber MA. Et al. 2008], [Weber MA. Et al. 2009], [*Weber, MA., Klein, NJ.et al 2008*]

## 2.2 Data Extraction

The GOSH post-mortem research database is held in a conventional set of hierarchical tables that are linked to a large number of individual lookup tables. All though this model works well for data capture it will not be optimised for data analytics.

The Entity Attribute Value model would be a good schema to use to extract the data for analytics. The advantages of using the EAV model for healthcare data are outlined by [*Löper, D., et al, 2013*] but the two main advantages in respect to this project are:

* Flexible and extensible model: This model is highly flexible, because different types of data objects can be included, even if the underlying format is different.
* Simple restoring capabilities: No data transformation is required in order to store the incoming information, the entities, attributes and values remain the same. No data is lost. This is especially crucial in medical applications.

One of the additional reasons for using the EAV model is the efficiency of storing data as described in [*Dinu, et al. (2007)]*

* Data are sparse and have a large number of applicable attributes, but only a small fraction will apply to a given entity
* Numerous classes of data need to be represented, each class has a limited number of attributes, but the number of instances of each class is very small.

A clear example of the flexibility of the EAV model for health care data is given in [*Borodin, et al. (2015)]*

## 2.3 Data Wrangling

Although the majority of data held on a post-mortem is categorical a significant number of data is numeric data, lengths and weights. The importance of these values in determining cause of death is detailed by [*Horn,et.al., 2004*. ].

Even within the main presentations of post-mortems described above these values can vary considerably; two main methods will be considered:

* Scaling – Each value i.e. heart weight could be expressed as a percentage of body weight.
* Growth charts - Individual analysis of each measurement by age of patient. Given the number of data points for each age group a distribution for each value could be determined and then each measurement could be classified into low, normal or high classification.

The approach of using growth charts in post-mortem analysis is described in [*Pryce, J.W., et.al. 2014]*

## 2.4 Analytics

The base analytic technique for this project will be Decision Trees with cross validation. Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable [*Song, et al.2015*]. The key advantages of the Decision Tree technique are:

* Simplifies complex relationships between input variables and target variables by dividing original input variables into significant subgroups.
* Easy to understand and interpret.
* Non-parametric approach without distributional assumptions.
* Easy to handle missing values without needing to resort to imputation.
* Easy to handle heavy skewed data without needing to resort to data transformation.
* Robust to outliers

[*Song, et al.2015*]

The main disadvantage of the technique is that using a single tree a model will suffer from low variance and high bias [*Analytics Vidhya Contributors (2016)*]. To combat this situation the project will consider ensemble methods which look to combine different techniques to better balance variance versus bias *[Abolfazl R, 2018*.].

The first technique to be considered will be Random Forest where the training data is split into a number of different sets and a tree is calculated for each set and the results combined *[Abolfazl R, 2018.*].

Gradient boosting is another technique that looks to decrease bias. Gradient boosting is a technique that looks to combine parameters that give a low prediction accuracy to produce a higher prediction accuracy [*Prashant G 2017.*]

A further extension of the base decision tree methodology would be to consider a neural network approach [*Yang, Y. et al, 2018*]. One of the key aspects of a decision tree methodology is the transparency of the process. Trying to maintain this key aspect while using the opaque approach of neural networks would be interesting to explore.

# 3 Project Aims

The aims of this project are:

1. To develop a routine to extract data from the existing Post-mortem Research Database into an entity attribute value schema that will make the data more readily available for data analytics.
2. To apply the Decision Tree Analytical method to the extracted data to develop operational strategies that can be applied to paediatric post-mortems to prioritise which data is required to achieve the target of specifying the cause of death.
3. To investigate ensemble strategies, specifically Random Forests and Gradient Boosting to see how these techniques can improve on the basic Decision Tree method.
4. If time permits a final aim would be to consider if the approach of Neural Networks could be employed to enhance the results.

# 4 Project Description

The project will be split into a three main of sections:

## 4.1 Data Extraction and Cleaning

The current post-mortem research database is held in Microsoft Access in a conventional set of hierarchical structure based around the model in which the data is collected – see simple schematic in Appendix A. The 30 main tables are linked to 150 lookup tables which hold the various lists of data that define the possible inputs for the categorical columns. All though this model works well for data capture it will not be optimised for data analytics.

The data will be transformed during the extraction process into an Entity, Attribute, Value model (EAV) see the literature Review for more details and the pros and cons of using this model. Each post-mortem is represented as a single entity or event and all the data captured on the post-mortem are stored as attributes for that event. The definitions of the categorical data will be stored in a single concepts table. Patient details will be held in a patients table and any patient specific data will be held as patient attributes.

One of the key benefits of using the EAV model is that only attributes with defined values need to be store and that all values have to be held in one of 4 columns: text, numeric, datetime and concept id. In this way the limited number of columns ensures that the data must conform to a basic standard at this early stage.

This section of the project ends with the data being in its revised EAV format is transferred to the data analytics platform. The details of the transferring process and the data analytics platform are given in the Section on System Architecture.

## 4.2 Exploratory Data Analysis and Data Wrangling

The 7000 post-mortems are split into 4 main presentations: SUDI, Sudden Unexpected Death in childhood (SUDC), Still birth and Termination of pregnancy (TOP). How are the 400 data items that could be collected for each post-mortem split between the different presentations?

Basic data distributions for the data items recorded for each post-mortem across the different presentations.

Each post-mortem is broken down into 3 main sections:

* External Examination – basic external measurements and the recording of external features.
* Internal Examination – Part 1 – Organ weights
* Internal Examination – Part 2 – Individual investigation of the 7 main bodily systems.

Additional information is also gathered on the patient’s history, circumstances of their death and other information relevant to each case.

Additional to the data recorded during the post-mortem samples are taken and various microbiology and virology tests are undertaken and the results recorded with the post-mortem record.

Even within the 4 main presentations the amount of data recorded for each post-mortem can vary significantly so during this section a base data set for each presentation will be defined such that each post-mortem cases needs to have to be included in the analysis.

Particularly attention will be given to the numeric data, lengths and weights, as even within the main presentations these values can vary considerably; two main methods will be considered:

* Scaling – Each value i.e. heart weight could be expressed as a percentage of body weight.
* Growth charts - Individual analysis of each measurement by age of patient. Given the number of data points for each age group a distribution for each value could be determined and then each measurement could be classified into low, normal or high classification.

## 4.3 Feature Selection

Of the 300 potential items of information that can be recorded during a post-mortem which are the most significant items for predicting the cause of death?

Having reviewed the literature this project will use decision trees as the basic statistical method in determining the significance of individual items of data, the main reasons being that the data recorded for a post-mortem lends itself to a decision tree based approach combined with the clarity in which the results of a decision tree analysis can be communicated back to the users.

One of the main problems with a basic decision tree analysis is that it is subject to a high variance based on the specific set of data used for a particular analysis. The recognised method of reducing variance is to use an ensemble analysis where multiple approaches are combined in a single analysis. Two common ensemble methods used with decision tree analysis are:

* Random Forest
* Gradient boosting

Both of these methods have their pros and cons and their individual benefits can only be realised by a detailed analysis on the specific data to be used. This comparison will form the analytic basis of this project.

# 5 System Architecture

The system architecture used during the project is best described in two parts:

## 5.1 Originating System

The current GOSH Post-mortem Research database is held in a Microsoft Access database. It contains the records of over 7000 post-mortems performed at GOSH since 1996.

The data is held in a hierarchical data structure that has been optimised for data entry, see Appendix A. As part of the project the data will be transformed into an EAV data structure, see Appendix B, written out to CSV files, one for each of the originating tables and then imported into the GOSH Analytic Platform.

## 5.2 Analytic Platform

GOSH provide a cloud based Digital Research Environment (DRE). The DRE provides a collaborative environment for the management, visualisation and analysis of routinely collected de-identified clinical and other data. The platform incorporates both technical and process components that offer strong data governance by design.

Analysis is undertaken for individual projects using a pre-defined ‘workspace’. Data is imported into the project workspace using CSV files that are presented back to the researchers as SQL tables. As part of the import process the system uses a pre-defined table definition file to pseudonymise any keys and removes all identifiable columns.

Within the workspace the researcher has access to develop both SQL scripts for data manipulation and R scripts for visualisation and analytics.

# 6 Software Languages and Tools

The software languages that will be used on this project are:

## 6.1 Python

Python will be used to extract the data from the PM Research database and transform the data from the hierarchical model to the EAV model. The primary Python package involved will be ‘PyODBC’ which enables Python to connect to any ODBC defined source and use SQL based commands to extract and manipulate data. The Python coded process will end with the production of the CSV files required to import into the GOSH analytic platform.

To ensure robust code and an efficient development process both unit testing and profiling will be used.

## 6.2 SQL

In the process to move the data from the PM Research database GOSH analytic platform SQL will be used to extract the data from the hierarchical data structure, create the EAV tables and insert the transformed data into the new table structure.

Within the GOSH analytic platform SQL will be used to list and cast data from the EAV model into datasets optimised for the chosen analytical process.

## 6.3 R

Once the data has been structured within the GOSH analytic platform all data visualisation and statistical analysis will be done using the R software language. A number of packages will be used:

* Dplyr – data manipulation
* Ggplot2 – data visualisation
* Tree – decision trees
* RandomForest – Random Forests
* Xgboost – Gradient boosting

# 7 Methodology and Work Plan

The project will use an agile approach with the work being broken down into a number of sprints. Currently the sprints are envisioned to be:

* Convert data to EAV format
  + Create tables (Python, pyODBC and SQL)
  + Extract data (Python and pyODBC)
* Upload data to GOSH DRE workspace
  + Create CSV files (python)
  + Import CSV tables into EAV table structure (SQL)
* Initial data summary and visualisations
  + Create list then cast datasets (SQL)
  + Initial data summary and visualisations (R)
* Data wrangling
  + Numeric parameters
    - Scaling
    - Growth charts
  + Summarise parameters
    - Body system abnormalities
* Initial Decision Tree Analysis
  + Create datasets based on (SQL):
    - Post-mortem presentation
    - Post-mortem stages
  + Execute analysis (R)
  + Model evaluation
* Apply ensemble methods – Please note that as time for this project is limited not all these ensemble methods may be attempted.
  + Random Forests
    - Create datasets
    - Execute analysis
    - Model Evaluation
  + Gradient boosting
    - Create datasets
    - Execute analysis
    - Model Evaluation
  + Neural Networks
    - Create datasets
    - Execute analysis
    - Model Evaluation
* Write up project report

Due to the reduced workload in the autumn term the first three sprints have been completed. The results are summarised in Appendix C.

The project will be continued in the summer term after the exams have been completed and will be completed by the 16th September

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Activity | June | July | August | September |
| Continue project on exam completion | \* |  |  |  |
| Data wrangling | \*\* |  |  |  |
| Decision Tree Analysis | \*\* |  |  |  |
| Apply Ensemble Methods as time allows:   1. Random Forest 2. Gradient Boosting 3. Neural Network |  | \*\*\*\* |  |  |
| Discussion, iteration and report  framework |  |  | \*\*\*\* |  |
| Complete project report |  |  |  | \*\* |
| Project Completion -  16th September |  |  |  | \* |
|  |  |  |  |  |

# 8 Ethics Approval and Data Governance

Use of the de-identified PM Research database has Health Research Authority (HRA) Approval

Use of the GOSH Digital Research Environment (DRE) has HRA Approval for carrying out de-identified research using GOSH clinical data.

HRA Approval was granted after having obtained Research Ethics Committee (REC) approval. Additionally all research carried out by GOSH is managed by the Joint Research and Development (R&D) Office.

Within a DRE analytic workspace de-identified data is stored and processed in a secure, cloud-enabled, dedicated GOSH secondary use data store which has all usual NHS security arrangements. This allows secured researcher access from multiple computers on the GOSH network and external collaborators.

# References

*Abolfazl Ravanshad, 2018. Ensemble Methods [online] Medium. Available at:* [*https://medium.com/@aravanshad/ensemble-methods-95533944783f*](https://medium.com/@aravanshad/ensemble-methods-95533944783f) *[Accesses 2019-03-24]*

*Abolfazl Ravanshad, 2018. Gradient Boosting vs Random Forest. [online] Medium. Available at:* [*https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80*](https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80) *[Accesses 2019-03-24]*

*Analytics Vidhya Contributors (2016). A Complete Tutorial on Tree Based Modeling from Scratch. [online] Analytics Vidhya Avaiable at:* [*https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/*](https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/) *[Accesses 2019-03-24]*

*Ben‐Sasi, K., Chitty, L.S., Franck, L.S., Thayyil, S., Judge‐Kronis, L., Taylor, A.M. and Sebire, N.J., 2013. Acceptability of a minimally invasive perinatal/paediatric autopsy: healthcare professionals' views and implications for practice. Prenatal diagnosis, 33(4), pp.307-312.*

*Borodin, A. and Zavyalova, Y., 2015. On an EAV based approach to designing of medical data model for mobile healthcare service. UBICOMM 2015, p.33.*

*Dinu, Valentin; Nadkarni, Prakash (2007), "Guidelines for the effective use of entity-attribute-value modeling for biomedical databases", International Journal of Medical Informatics,****76****(11–12): 769–79*

*Horn, L.C., Langner, A., Stiehl, P., Wittekind, C. and Faber, R., 2004. Identification of the causes of intrauterine death during 310 consecutive autopsies. European Journal of Obstetrics & Gynecology and Reproductive Biology, 113(2), pp.134-138.*

*Löper, D., Klettke, M., Bruder, I. and Heuer, A., 2013. Enabling flexible integration of healthcare information using the entity-attribute-value storage model. Health information science and systems, 1(1), p.9.*

*Lullaby Trust (2018). Facts-and-Figures-for-2015-released-2017. Updated 2018. London: Lullaby Trust.*

*NHS Contributors (2018). Post Mortem. [online] NHS. Available at: https://www.nhs.uk/conditions/post-mortem [Accessed 2019-03-24].*

*Prashant Gupta 2017. Decision Trees in Machine Learning. [online] Medium. Avaiable at:* [*https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052*](https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052) *[Accessed 2019-03-24]*

*Pryce, J.W., Bamber, A.R., Ashworth, M.T., Kiho, L., Malone, M. and Sebire, N.J., 2014. Reference ranges for organ weights of infants at autopsy: results of> 1,000 consecutive cases from a single centre. BMC clinical pathology, 14(1), p.18.*

*RCPath Contributors (2018). Paediatric Pathoogy. [online] Royal College of Pathologists. Available at:* [*https://www.rcpath.org/trainees/examinations/examinations-by-specialty/paediatric-pathology.html*](https://www.rcpath.org/trainees/examinations/examinations-by-specialty/paediatric-pathology.html) *[Accessed 2019-03-24].*

*Song, Yan-Yan and Ying Lu.2015. “Decision tree methods: applications for classification and prediction” Shanghai archives of psychiatry vol. 27,2: p130-5.*

*Weber, M.A., Ashworth, M.T., Risdon, R.A., Hartley, J.C., Malone, M.A.R.I.A.N. and Sebire, N.J., 2008. The role of post-mortem investigations in determining the cause of sudden unexpected death in infancy. Archives of disease in childhood, 93(12), pp.1048-1053.*

*Weber, M.A., Ashworth, M.T., Risdon, R.A., Malone, M., Burch, M. and Sebire, N.J., 2008. Clinicopathological features of paediatric deaths due to myocarditis: an autopsy series. Archives of disease in childhood, 93(7), pp.594-598.*

*Weber, M.A., Ashworth, M.T., Risdon, R.A., Brooke, I., Malone, M. and Sebire, N.J., 2009. Sudden unexpected neonatal death in the first week of life: autopsy findings from a specialist centre. The Journal of Maternal-Fetal & Neonatal Medicine, 22(5), pp.398-404.*

*Weber, M.A., Klein, N.J., Hartley, J.C., Lock, P.E., Malone, M. and Sebire, N.J., 2008. Infection and sudden unexpected death in infancy: a systematic retrospective case review. The Lancet, 371(9627), pp.1848-1853.*

*Wikipedia Contributors (2019). Entity Attribute Model. [online] Wikipedia. Available at:* [*https://en.wikipedia.org/wiki/Entity-attribute-value\_model*](https://en.wikipedia.org/wiki/Entity-attribute-value_model) *[Accessed 2019-03-24].*

*Yang, Y., Morillo, I.G. and Hospedales, T.M., 2018. Deep neural decision trees. arXiv preprint arXiv:1806.06988.*

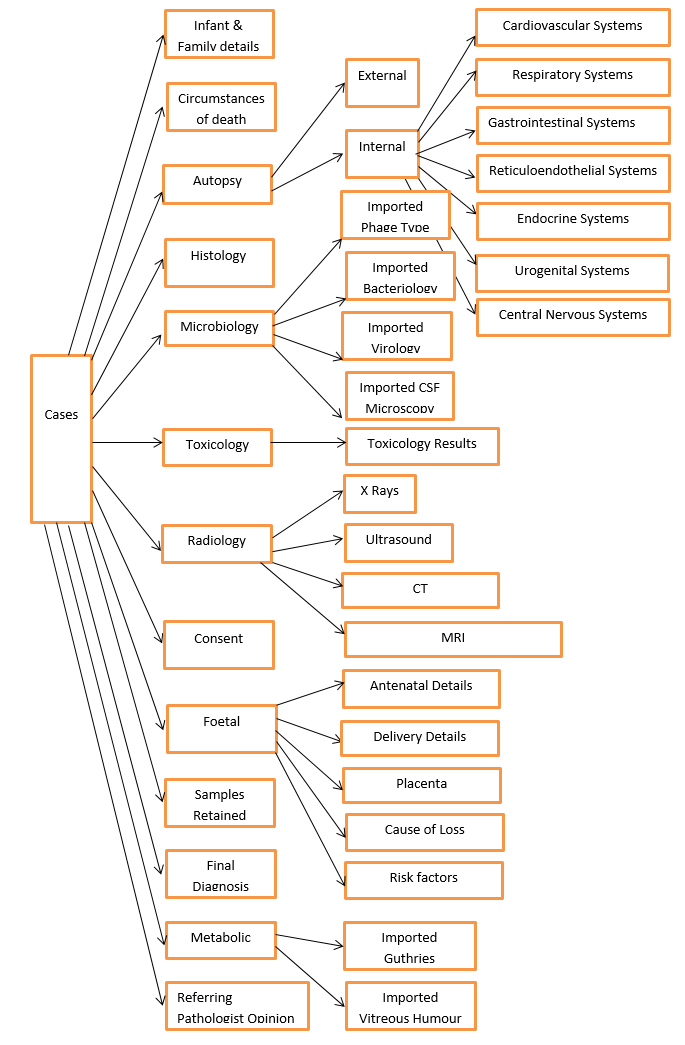
# Bibliography

*Caruana, R. and Niculescu-Mizil, A., 2006, June. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM.*

*James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112, p. 18). New York: springer.*

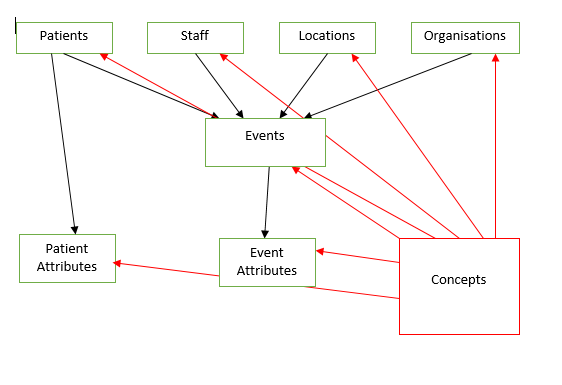
*Quinlan, J.R., 1986. Induction of decision trees. Machine learning, 1(1), pp.81-106.*

# Appendix A – Microsoft Access Structure of PM Research Database



The above simple schematic shows the basic set of tables and how they cascade down from the master Cases table. Each arrow represents a foreign key link from the level above. For clarity I have not included the over 400 individual lookup tables.

# Appendix B – Event Attribute Value (EAV) Model



The above is a simple representation of the EAV model that can represent the total data stored in the database shown in Appendix A including the contents of the look-up tables Please note for this project the Staff, Locations and Organisation tables are not required.

The arrows represent foreign keys back to their originating tables.

An example of the SQL used in the original table to extract a piece of data:

MS Access:

SELECT tblcases.caseid,   
       tblautopsies.dateautopsy,   
       tlkhema\_heartmacroabn.heartmacroabntext   
FROM   (((tblcases   
          INNER JOIN tblautopsies   
                  ON tblcases.caseid = tblautopsies.caseid)   
         INNER JOIN tblinternalexams   
                 ON tblautopsies.autopsyid = tblinternalexams.autopsyid)   
        INNER JOIN tblcardiovascularsystems   
                ON tblinternalexams.internalexamid =   
                   tblcardiovascularsystems.internalexamid)   
       INNER JOIN tlkhema\_heartmacroabn   
               ON tblcardiovascularsystems.heartmacroabn\_hemaid =   
                  tlkhema\_heartmacroabn.heartmacroabnid;

EAV Schema:

SELECT

ha\_events.start\_date,   
  HC\_1.concept\_type  AS EA\_concept\_type,   
  HC\_1.label         AS EA\_label,   
  ha\_event\_attributes.value\_numeric,   
  HC \_2.label

FROM

(ha\_concepts   
     INNER JOIN ha\_events   
      ON ha\_concepts.concept\_id = ha\_events.event\_type\_id)   
     INNER JOIN ((ha\_event\_attributes   
       LEFT JOIN ha\_concepts AS HC\_1   
         ON ha\_event\_attributes.event\_attribute\_type\_id = HC\_1.concept\_id)  
       LEFT JOIN ha\_concepts AS HC\_2   
         ON ha\_event\_attributes.value\_id = HC\_2.concept\_id)   
       ON ha\_events.event\_id = ha\_event\_attributes.event\_id

WHERE

(HC\_1.concept\_type = "\event attribute type\post mortem\tblcases"  
     AND ha\_concepts\_1.label = "case id")  
   OR (HC\_1.concept\_type = "\event attribute type\post mortem\tblcardiovascularsystems"   
         AND HC\_1.label = "heartmacroabn\_hemaid") 

# Appendix C – Summary of output from initial Sprints

What I have achieved in the initial sprints:

* Python code using pyodbc package:
  + 1750 lines of code
  + Unit testing on functions as I converted them
  + Problem with initial code which would have taken 5 days to extract and re-model data.
  + Profiling found the time sink was opening new selects. NB inserts took very little time but don’t return any values
  + Also significant impact using linked tables not required in ODBC model.
  + Reduced time from 60 seconds per event to just under 10 seconds
  + Process then took 20 hours (7000 events)
* Created CSV files and uploaded to workspace
  + Got initial breakdown
  + Needed to create some new attributes for reporting based on existing attributes
    - Cause of death (COD) summary into Known and unknown
    - Number of attributes per event
* SQL in workspace to create basic visualisation datasets.
  + List and pivot (pivot in R as no pivot function implemented in Greenplum postgrSQL in platform).
* Added Microbiology and Virology events
  + 4 events with small number of attributes
    - Specimen Taken
      * Specimen type
      * Site
    - Lab Episode
      * Lab number
    - Lab Test Set
      * Test Set code and description
    - Lab Test
      * Test code
      * Result
  + More complex reporting
    - Linked events by using patient\_id
* Added more attributes for reporting
  + External/Internal/Microbiology/Virology examinations undertaken – True/False
  + Number of samples taken
* Finalised numbers for report – see below.
* Began filling out structure of project proposal

Initial breakdown of all post-mortem cases broken down by presentation (REF) and cause of death (COD2\_SUMM):

|  | **COD2\_SUMM** |  |  |  |
| --- | --- | --- | --- | --- |
| **REF** | **001:Unknown** | **002:known** | **003:Other** | **<NA>** | **Total** |
| **001:Sudden Death < 12 months** | 836 | 877 | 23 | 3 | 1739 |
| **002:Sudden Death > 12 months** | 190 | 504 | 125 | 4 | 823 |
| **006:Hospital PM NOS** | 28 | 451 | 12 | 1 | 492 |
| **007:Miscarriage < 24 weeks** | 380 | 378 | 135 | 68 | 961 |
| **008:Stillbirth > 24 weeks** | 585 | 242 | 118 | 49 | 994 |
| **009:Neonatal Death** | 8 | 32 | NA | 6 | 46 |
| **010:TOP** | 28 | 694 | 57 | 12 | 791 |
| **011:Forensic PM** | 217 | 154 | 13 | NA | 384 |
| **012:Coroner’s PM NOS** | 5 | 17 | 9 | NA | 31 |
| **014:Insufficient Clinical Details** | 684 | 6 | 4 | 1 | 695 |
| **Total** | 2961 | 3355 | 496 | 144 | 6956 |

NOS (Not otherwise specified)

Filtering the data on required presentations and cause of deaths the following number of post-mortem cases ca be used for analysis:

|  | **COD2\_SUMM** |  |  |
| --- | --- | --- | --- |
| **REF** | **001:Unknown** | **002:known** | **Total** |
| **001:Sudden Death < 12 months** | 811 | 813 | 1624 |
| **002:Sudden Death > 12 months** | 167 | 427 | 594 |
| **007:Miscarriage < 24 weeks** | 321 | 275 | 596 |
| **008:Stillbirth > 24 weeks** | 514 | 193 | 707 |
| **009:Neonatal Death** | 7 | 26 | 33 |
| **010:TOP** | 6 | 414 | 420 |
| **Total** | 1826 | 2148 | 3974 |

The total number of attributes used across all post-mortem cases was 307. The breakdown of number of attributes used by referral type:

