

# Causal Relationship Mining

*“Is it possible to identify and define mechanism to generate useful causal inter-relationships across data collections”*

## PhD Proposal

Supervisors: Frans Coenen<sup>1</sup>, Marwan Bukhari<sup>2</sup>, Nicky Goodson<sup>3</sup>, Marta García-Fiñana<sup>4</sup>

<sup>1</sup>Department of Computer Science, the University of Liverpool

<sup>2</sup>Department of Rheumatology, University Hospitals of Morecambe Bay NHS Trust

<sup>3</sup>University Hospital Aintree Academic Rheumatology Department, University of Liverpool

<sup>4</sup>Center for Medical Statistics and Health Evaluation, University of Liverpool.

## 1. Overview

Causal relationship mining, or cause and effect mining, is directed at the identification of relationships that link sets of fields, or specific field values, in data. At its simplest the data may comprise a single data table containing a few hundred records, at its most complex the data may comprise a distributed network of data tables containing thousands of records. The relationships may be expressed/modeled in a variety of ways. The most straightforward are simple implication rules such as  $X \Rightarrow Y$ , where X and Y are disjoint sets of conjuncted field values or expressions involving field values. Examples include:

Colour==green and Shape==spherical  $\Rightarrow$  Fruit=apple  
Age>65  $\Rightarrow$  Pensionable=yes

The first example assumes three database table fields “Colour”, “Shape” and “Fruit” and is read as “if colour value is green and shape value is spherical then fruit value is apple”. The second example assumes two data table fields “Age” and “Pensionable” and is read as “if age value is greater than 65 then pensionable value is yes”. We can of course get more complex by adding negations and disjunctions.

There are several techniques available for identifying rules/models of the above form. One simple approach is to use Association Rule Mining (ARM), this provides a good start point but is limited to binary valued fields which means data must be discretised or ranged. A second approach is to use some form of rule induction to induce rules from the data, or look at classification rule mining techniques. An alternative approach to identifying causal relationships is to build a model of the domain. For example multiple linear logistic regression may be used to construct such a model when the outcome variable of interest is the presence or absence of some condition X (e.g., presence/absence of a specific disease). The model can be used to answer specific questions such as the probability of a particular outcome (e.g., having the disease yes/no) in relation to several prognostic factors Y (e.g., age, gender, body mass index, smoker/non-smoker, etc.). The model also allows us to identify what factors are significantly associated with X and to create a prognostic index to distinguish those subjects likely or unlikely to have the outcome event.

We can also place restrictions on the rules we wish to identify, for example we may wish to find inter-relationships between two or more data sets and not intra-relationships within a single data collection.

The proposed research would consider the following question:

*“Is it possible to identify and define mechanism to generate useful causal inter-relationships across data collections, and if so what are the most appropriate mechanisms for this purpose”*

There are a number of issues associated with this research question:

1. To identify appropriate mechanisms a clear understanding must be obtained regarding the nature and representation of the desired relationships.
2. Having identified the nature of the desired relationships the mechanism to generate the relationships must be established. What these mechanisms might be is unclear at the time of writing; but, as indicated above, ARM and rule induction provide two start points. The field of classification, where classifiers are generated in terms of rule sets, may provide another fruitful avenue of research.
3. The third issue is how to determine whether a relationship is useful or not. Can we put some metric on a relationship? Can the generation process perhaps encapsulate some concept of usefulness?
4. Finally, can the generated model be used to predict outcomes, would it be possible to query the model?

To help focus the investigation, and provide answers to the above, the work will be directed at a particular medical application. This is discussed in the following Section.

## **2. The Application Domain**

It is conjectured that there is a relationship, in terms of body fat and bone density, between Osteoporosis and Cardiovascular disease (CVD), i.e that patients with osteoporosis have increased cardiovascular disease. Osteoporosis is a bone disease that increases the likelihood of bone fracture and is characterized by reduced bone density. Cardiovascular disease encompasses disease of the heart and or blood vessels. Atherosclerotic cardiovascular disease is the most common cardiovascular pathology and it is this condition that may be linked with adverse bone health. Atherosclerotic CVD often causes significant illness events that frequently result in admission to hospital. Two members of the supervisory team have considerable experience in this field and will be able to provide appropriate supporting expertise. More specifically the research team have access to data obtained from a Dual Energy X-Ray assessment (DEXA) scanner located within the osteoporosis service at the Royal Lancaster Infirmary (RLI). A DEXA scanner utilizes low energy X-ray absorption by bone and soft tissues to calculate hip and spine Bone Mineral Density (BMD) and % body fat. The osteoporosis service at Lancaster currently have some 8,000 patients referred to them each year, many from within the Lancaster area, some from outside. The DEXA scanner has been in operation since 2004 and consequently some 35,000 patient records are available to support the proposed inter-relationship mining research.

In addition it will be possible for the research team to get access to additional patient and risk factors data regarding cardiovascular disease and osteoporosis from RLI's main patient database. Cardiovascular events resulting in hospital admission can be identified from hospital admission statistics (HES). These are collected at both regional and national level.

The desire to determine the relationship between cardiovascular disease and osteoporosis in terms of body fat and bone density will therefore act as the exemplar application to drive the research.

There are, however, some complications associated with the application. The first is that the data is currently distributed and stored in a number of formats. The data would therefore have to be brought together in a single data warehouse. A second complication is that the anonymity of the patients must be insured, therefore all data must be anonymised. A third issue is that it is highly likely that the data will contain missing values, duplicate records and other anomalies. It is anticipated that construction of the application specific data warehouse will form a substantial part of the research. It is hoped that this part of the work will lead to some additional interesting insights into the mechanics for constructing warehouses that might merit publication. It is also anticipated that the researcher will spend a significant amount of time in Lancaster where the DEXA scanner is housed and the data is located.

### 3. Programme of Work

A three year programme of work is envisioned as describe below. Note that the work will include a substantial element of a data warehouse development. Once the data warehouse is “up and running” the programme of work is directed at investigating and comparing three separate techniques to achieve inter-relationship mining.

<b>Work Package</b>	<b>Start (month)</b>	<b>Duration (months)</b>	<b>Description</b>	<b>Deliverables</b>
Induction	0	0.5	Familiarisation with University procedures and general induction to the programme of work	None
Background reading	0.5	1.5	Familiarisation with application domain and current state of the art.	Technical report
Data warehousing	2	6	Construction of the data warehouse for the application domain. An essential component for the application.	A Data warehouse and a supporting technical report (note that is some interesting experiences are encountered and solved it may be possible to publish an “experience paper”.
Causal relationships	8	1	Investigation into causal relationships, their definition and how their appropriateness may be evaluated.	Technical report.
Relationship Mining Technique 1	9	6	A good stat point for the work will be Association Rule Mining. This is a well understood technology although it has limitations in that it typically works with binary valued data.	Causal relationship rule generator 1 software and (low level) research paper.
Relationship Mining Technique 2	15	6	The second investigation into generating causal relations will look at an alternative approach to technique 1. At time of writing rule induction or classification techniques seem to be an appropriate technology.	Causal relationship rule generator 2 software and (middle ranking) research paper that includes comparisons with previous technique.
Relationship Mining Technique 3	21	5	The third investigation would be directed at relationship modeling using multiple linear logistic regression approach.	Causal relationship model generator software and (significant) research paper that includes comparisons with previous techniques.
End Evaluation	26	2	Final evaluation of work.	Summary research paper.
Thesis Writing	28	6	Thesis writing	Thesis.