Automating the Integration of Clinical Studies into Medical Ontologies

Mark Roantree, Jim O' Donoghue, Noel O'Kelly
School of Computing
Dublin City University, Ireland.
{mark,jodonoghue,nokelly}@computing.dcu.ie

Martin van Boxtel, Sebastian Köhler
School for Mental Health & Neuroscience
Maastricht University, Netherlands.
{martin.vanboxtel,s.koehler}@maastrichtuniversity.nl

Abstract

A popular approach to knowledge extraction from clinical databases is to first define an ontology of the concepts one wishes to model and subsequently, use these concepts to test various hypotheses and make predictions about a person's future health and wellbeing. The challenge for medical experts is in the time taken to map between their concepts/hypotheses and information contained within clinical studies. Presently, most of this work is performed manually. We have developed a method to generate links between Risk Factors in a medical ontology and the questions and result data in longitudinal studies. This can then be exploited to express complex queries based on domain concepts, to extract knowledge from external studies.

1. Introduction

As part of medical research, many specialists conduct clinical studies over long periods of time, in order to observe human demographics, lifestyles, behavior and choices, to try to understand how illnesses can occur and more importantly, devise means of preventing or delaying various forms of ill health. These clinical studies often contain significant knowledge but one must understand how to search or data mine for it correctly. Often this requires advance (and considerable) knowledge of the clinical study and the extraction is usually a manual process, using spreadsheets or statistical software. When a fresh approach to studying a particular domain is initiated, medical researchers may wish to view existing clinical studies from different perspectives and these perspectives may differ from the focus or structure of the original studies. This provides the primary motivation for the research tackled in this paper. We are seeking to automatically link a set of medical concepts or requirements to existing clinical studies, in order that knowledge relating to these concepts of interest can be extracted without the need for the specialist to select the right sections or question(s) in the study which are relevant to the query. This is a significant development as many of these clinical studies have thousands of questions where the answers are used to generate metrics or provide a platform for the analytics used to predict likely outcomes.

This research takes place as part of a project to investigate means to decrease dementia risk and/or delay the onset of dementia by combining areas of social innovation, multi-factorial modeling and clinical expertise [9]. One of the aims is to quantify dementia risk and deliver personalized strategies and support to enable individuals to reduce their risk of dementia in later life. The strategy is for specialists in dementia to work together with information management researchers to devise algorithmic methods for reusing existing clinical studies and quantifying dementia risk and possible reduction.

1.1. Background and Motivation

Clinical research using large datasets generally uses a sensor based approach to gather data such as [22] or reuses data from large studies such as [10]. Previous research efforts on constructing medical ontologies have provided frameworks for incorporating data from operational medical systems [18] or tackling the general issue of interoperability across medical applications [21,1]. One of the requirements of our project is to develop and test a number of different hypotheses. Currently, testing must be performed manually using spreadsheets or statistical software but the role of data management researchers on the project is to automate this process.

When describing ontologies, they are seen as *a formal specification of the terms in a target domain and a description of relationships between concepts* [19]. For this project, the domain is dementia and concepts are the Risk Factors associated with dementia. One of the advantages of adopting an ontological approach to any specific problem is that it defines a common vocabulary for researchers wishing to share information in a domain. At its core, it comprises of machine-interpretable definitions of the major concepts

and their relations and thus, provides the ability for the system to interpret problems and assist in decision making. The ontological approach has been shown to be effective in areas such as intensive care [2] and even in broader healthcare like the Lifeline project [4]. In terms of research into dementia, one of the earliest approaches that involved ontology construction was in [15] where they sought to formally describe concepts and the relationships between them. Each of these projects demonstrated the impact of a formal approach to classifying terms and relationships and how the ontologies can be exploited for a greater understanding of data in different domains.

The main goals in the development of the initial Ontology can be identified as: Identify the main Risk Factors for dementia (the core concepts in the ontology); Model the Risk Factors by identifying the properties that best describe them; Model the Question/Answer Database from a suitable study; Create links between Risk Factors and Question/Answer databases appropriate to the ontology; Develop a series of protocols for testing various hypotheses.

1.2. Problem Description

This paper describes the efforts at the fourth step in ontology construction: linking knowledge from existing longitudinal studies to risk factors identified by health specialists for the particular medical domain. What may appear as a relatively straightforward task (and currently performed manually) is in reality quite a difficult problem. When specialists devise a series of hypotheses to be tested using one or more longitudinal studies it requires the interaction and manipulation of potentially thousands of questions. The problem is in the quick identification of those areas of the study that best test the hypothesis. Given the manual nature of this approach, it is often difficult to ensure that all relevant questions and answers are used and thus, the accuracy of the results can be difficult to measure. Our goal is to avoid human matching of ontology concepts with sections of studies and instead provide an automated approach to constructing the queries necessary to express each hypothesis and generate the appropriate result set. However, simple keyword matching between risk factors and the terms used in question-based studies results in a poor level of matching. While ideas such as ontologies for managing healthcare surveys as proposed in [19] could greatly assist in matching new concepts to older datasets, the reality is that this type of structured approach to medical studies does not exist. Thus, a more innovative approach is required to exploit older studies when creating new ontologies.

1.3. Research Focus and Contribution

Our method is to use word distance algorithms and the Wordnet approach [16,25] to match ontology concepts (risk factors) to questions expressed in selected studies. Using this approach, we compare ontology keywords with all questions in the study using a *similarity threshold*, so that terms do not require an exact match but can be *semantically close* to the ontology keyword. There are four steps in our method to determine where the highest levels of matching to questions in clinical trials were achieved and also, to measure how much of the study could be mapped to concepts in our ontology. The automatic generation of these links will provide a query based system that links directly from ontological concepts through to datasets from clinical studies.

It should be made clear that we not addressing the problem of inference as with many research contributions using ontologies. Our ontology is constructed from scratch, using one or more clinical studies but our focus is on matching user needs to appropriate parts of clinical studies and in the provision of a query interface to interrogate the dataset. As we focus on the matching aspects in this paper, our contribution is at three levels:

- We provide a framework in which ontologies and clinical studies can be mapped or integrated;
- Using existing word matching technologies, we provide a hybrid matching method which uses both word similarities and the structure of the clinical studies to map ontological concepts (Dementia Risk Factors) to questions/data in clinical studies;
- By storing all matching instances inside a data warehouse, we can easily extract the analysis used to adapt the parameters for subsequent matching experiments.

The paper is organized as follows: in Section 2, we provide an overview of our system and the process for linking to existing clinical studies; in Section 3, we describe the keyword matching used to create the links between ontology concepts and specific aspects of clinical studies; in Section 4, we show how the structure of clinical studies can also be used in the mapping process; in Section 5, we present our evaluation; in Section 6, we discuss related work on the topic of medical ontologies; and finally, in Section 7, we present our conclusions and outline future work in this area.

2. Ontology Components

In this section, we provide an outline view of the system and briefly describe the main components.

While there are different approaches to constructing ontologies [3], the six major steps outlined in [19] are broadly present in each approach. In the first three steps, the basic concepts are identified, their properties described, and in some ontologies, a vocabulary is created containing all of the allowable terms for the ontology. We place each of these steps into a single phase which we call *Ontology Initialization*. In our project, these concepts are Risk Factors (associated with dementia) and the properties are those characteristics used to describe or measure a particular Risk Factor. In Figure 1, the basic ontology is shown as a classification of Risk Factors, the system vocabulary, the associations between these components, and the process for initializing the ontology.

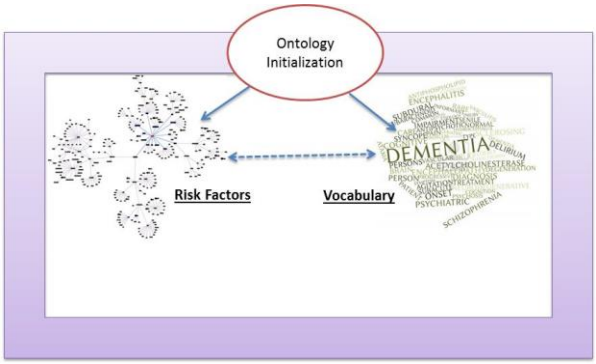

Figure 1: Initializing the Ontology.

Figure 1 also shows the vocabulary, specified by specialists in the area of dementia and the links between those vocabulary keywords and the Risk Factors. In many cases, many-to-many links are formed between vocabulary keywords and Risk Factors.

When a clinical study has been identified as a candidate for knowledge extraction or any form of query processing, it is first necessary to import all of the questions presented in the study into the system. In effect, this is a process of generating metadata. Most studies will have some form of structure where questions are asked in a specific order, or the study is sub-divided into clearly labeled sections. This process is known as the "Model Clinical Study" phase as all questions are imported and are then sectioned into clusters as determined by the study.

Python is used for this process because it is a natural fit for this type of exercise (scripting language) and can be extended with the very popular Natural Language Toolkit (NLTK) [10]. It is regarded as one of the leading platforms for building Python programs to work with human language data, containing a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

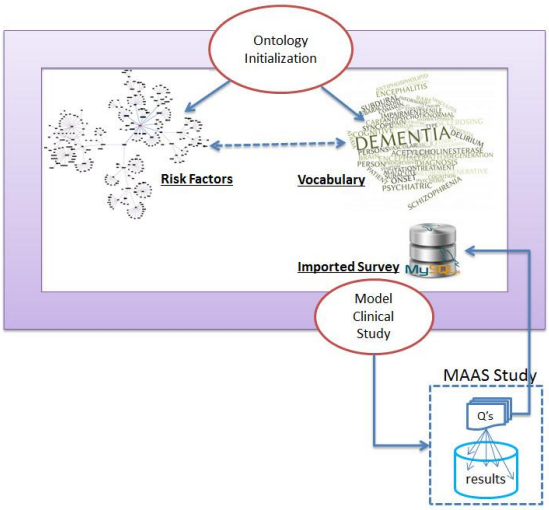

Figure 2: Modeling the Study's Structure.

The result of this process is that all questions are given unique identifiers (many studies will already contain this information), and clusters are also given unique identifiers. In the case of clusters, many studies will already have these labels (e.g. *Family History* or *Details of Activity/Exercise*) although it is not necessary for the system to have meaningful labels. In other words, the system need not understand labels as they are merely used to classify questions into clusters. This process of importing questions and results is shown in figure 2.

The final phase, matching the ontology to the target clinical study is the primary focus of this paper and is described in depth in the following section. In brief, the goal is to link each ontological concept (Dementia Risk Factor) with all relevant questions in the clinical study. In Figure 3, a process for comparing Risk Factors with questions from the clinical study results in the generation of mappings or links between them. This removes the need for human preprocessing as required querying or data mining operations can now exploit the links to auto-generate query expressions. Python is again used to store all matches in a MySQL data warehouse, as well as XML which is being used in a separate research project.

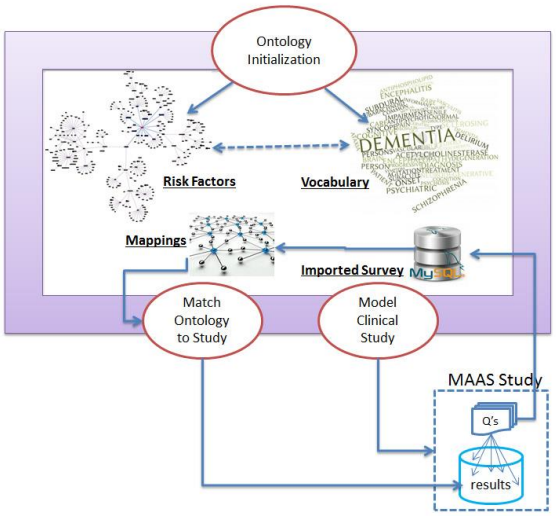

Figure 3: Mapping Risk Factors to Study Questions

3. Keyword Matching Method

The first three steps focus on matching ontology keywords to terms used in the questions in clinical studies. In essence, each step adopts the same approach but uses different keywords for matching with the questions. For step 1, we use the Risk Factor names (see Table 4); for step 2, we use the properties that describe the Risk Factor; and for step 3, we use all keywords that are contained in the vocabulary and are associated with the Risk Factor. The fourth and final step is described in Section 4.

3.1. Method

For each iteration of keyword matching, we begin with a set of terms that represent the Risk Factor:

$$RF_i = \{RF_{t1}, RF_{t2}, \ldots, RF_{tn}\}$$

Each term RF_{ti} is passed to the WordNet system [25] together with a threshold T_{ti} which represents the level of synonym match to be used. Wordnet then returns set(s) of word matches, which are combined so that for each term RF_{ti}, there is now a set of terms. In some cases, this will be a singleton set where only the term itself is returned.

$$RF_{t1} = \{RF_{t11}, RF_{t12}, \ldots, RF_{t1n}\}$$
$$RF_{t2} = \{RF_{t21}, RF_{t22}, \ldots, RF_{t2m}\}$$
…
$$RF_{tp} = \{RF_{tp1}, RF_{tp2}, \ldots, RF_{tpq}\}$$

The goal at this point is to reduce multiple sets of synonyms to a single set for each Risk Factor as argument for the comparison algorithm. For Risk Factor RFi, we refer to the set of all possible synonyms as RFTi. A union operation is used to create a single set so that for Risk Factor RF_i, all synonyms are present in RFT_i:

$$RFT_i = RF_{t1} \cup RF_{t2} \cup RF_{t3} .. \cup RF_{tn}$$

At this point, we have a single set of terms to represent each Risk Factor RF_i.

As described in the previous section, each clinical study is imported into the system as a series of *clusters* (representing sub-sections) of questions. Each cluster has an identifier C_m and within any cluster each question is identified by Q_n. Thus, every question in the clinical study has a unique identifier provided by $\{C_m,Q_n\}$ where C_m represents the cluster identifier and Q_n the question identifier. Each word in each question can then be addressed by the triple $\{C_m,Q_n,W_o\}$.

Every term RF_{ij} is matched against each word W_o (excluding stop words) in each question $\{C_m,Q_n\}$ contained in the clinical study. If any terms $\{RF_{ij}, C_mQ_nW_o\}$ match, then that question Q_n is linked to the Risk Factor RF_i in the ontology.

For all 3 steps in the matching process, a Fact Table FT_{wm} with details of each comparison $\{RF_{ij}, C_mQ_nW_o\}$ is created inside the data warehouse with the structure shown in Definition 1. The first three attributes uniquely identify a question compared with a specified Risk Factor; this is followed by the particular step in which this comparison took place; the next two attributes are the words compared, followed by the threshold used and the result of the comparison.

Definition 1. **Fact Table Structure**

FT_{wm} = {CID, QID, RFID, Step, RF_{ti}, $C_mQ_nW_o$, Tti, Result}

Definition 1 shows the relation for the Fact table containing all comparison tests. CID is cluster identifier; QID the question identifier; RFID the Risk Factor identifier; Step has a value of 1,2,3 or 4 depending on which step the comparison occurs; RF_{ti} is the Risk Factor; $C_mQ_nW_o$ is the word identifier; Tti, the threshold; and finally, Result is Boolean and indicates if the comparison was true or false. As the structure suggests, we adopt a purely relational approach to data mining queries for performance reasons and as we are dealing with a single relational *style* dataset. However, we can adopt an XML-based approach where it is necessary to combine ontologies as we have shown in [7] that similar levels of performance can be achieved.

Querying this fact table is used as part of the validation process that determines both the accuracy of the links created between Risk Factors and Clinical Studies, and in cases where false hits occurred, to quickly drill down and determine the process which resulted in the false hit.

4. Using Structure to Map Questions

Due to the nature of the questions in clinical studies, there remain many unmatched questions after the first three rounds of word matching as discussed in the previous section. Example 1 shows a sample question (a) and statement (b) for participants to provide input. However, there is no context with which to associate either with a particular risk factor. Our approach is to associate this type of question with other questions that are richer in context or have clear keywords and an existing match to a Risk Factor. The second phase in the matching process uses the inherent structure in clinical studies to attempt to match remaining questions.

Example 1. No-Context Questions

a. Did you ever feel that it is all a bit too much? Choose option 1/2/3/9 as described in item 1
b. For most people it is easier to remember interesting facts than uninteresting facts. Answer from 1-9.

This stage begins with the creation of a matrix of Clusters by Risk Factor. Recall that we use Clusters to group sets of questions. The matrix is populated with the *percentage of questions matched so far*, for each cluster against each Risk Factor. For example, if cluster C_i has a total of 10 questions of which 5 are matched for Risk Factor RF_1 and 8 for RF_2, then:

$RF_1C_i = 0.5$

$RF_2C_i = 0.8$

The algorithm is simple in approach. If any pairing {Ci,RFj} exceeds a set threshold T_s, then all of the remaining questions in cluster C_i are mapped to RF_j. For the purpose of this analysis, the setting was T_s = 0.3.

The results can be seen in Table 4 and are discussed in the section 5 but beforehand, we present an example of one cluster of questions passing through each of the 4 steps.

4.1. Illustrated Example

This section is used to illustrate how a cluster of questions `tol_test` is matched against the all risk factors. For reasons of simplicity and space, we chose a cluster with only 3 questions.

Step 1: This step uses the risk factor titles only. A single question is matched based on the term *cognitive* for Risk Factor *Cognition*.

cid	qid	RF	Term	Tt	Step	Result
tol_test	tolstat	Cognition	cognitive	0.3	1	1

Table 1: Step 1 Result.

Step 2: This step uses the attributes modelled for each Risk Factor. Two more questions (with id=tolscore and id=tolstat) are matched, for risk factor *Inactivity*, using the synonyms *task* and *physical* respectively. The question previously matched in step 1, is again matched but this time through the synonym *instruction*.

Qid	RF	Term	Tt	Step	Res
tolstat	Cognition	Cognitive	0.3	1	1
tolstat	Cognition	instruction	0.8	2	1
tolscore	Inactivity	Task	0.8	2	1
tolstat	Inactivity	Physical	0.8	2	1

Table 2: Step 2 Result
(`cid` omitted, all values = `tol_test`)

Step 3 uses terms from the vocabulary but for this cluster, no further questions were matched.

Step 4 seeks to add the remaining questions from the cluster into each of the Risk factors that exceed the selected threshold based on the percentage of questions (set at 30% for this experiment) already matched for that cluster. In other words, where any cluster has the number of matched questions ≥ 30% of the total questions in the cluster, the entire cluster of questions is mapped.

Qid	RF	Term	Tt	Step	Res
tolstat	Cognition	Cognitive	0.3	1	1
tolstat	Cognition	Instruction	0.8	2	1
tolscore	Inactivity	Task	0.8	2	1
tolstat	Inactivity	Physical	0.8	2	1
tolserie	Cognition		0.3	4	1
tolscore	Cognition		0.3	4	1
tolserie	Inactivity		0.3	4	1

Table 3: Step 4 Result for Tt = 0.3.

The final result sees all 3 questions mapped to both Risk Factors due to the threshold being reached.

5. Experiments and Evaluation

In order to evaluate our work, we used the MAAS epidemiological study into biological, medical and psychosocial aspects of normal and pathological cognitive aging [10]. It was a prospective cohort study using more than 1800 community-dwelling individuals aged between 24 and 81 years. In total, the dataset contained 2,372 questions spread across 79 clusters (or questionnaire sub-sections). Clusters had between 3 and 121 questions, with an average of 30 questions across each cluster.

Experiments were run in Intel Core 2 Duo processor CPU E8400 running at 3GHz on a 64 bit Ubuntu 12.04 LTS platform. The Natural Language Tool Kit (NLTK) [17] and Wordnet [25] technologies were incorporated into a Python 2.7 application. Most ontology-based research employs RDF to model data but we adopt a SQL/UML model as we use real world datasets that are (close to) relational in structure. Not discussed here is our usage of XML both for interoperability across clinical studies and for creating web extended ontologies. This is primarily due to the optimization based approach in our previous work [12,13] where issues with XML performance can affect the usage of large ontologies.

5.1. Evaluation Methodology

One of the benefits of this approach is the automatic generation of evaluation forms which are linked to appropriate parts and questions of the clinical study. A sample of questions matched to Risk Factors were presented to dementia experts and they were asked to indicate those which were true and false hits. Our Fact Table can easily be queried to determine at which of the 4 steps, the question was matched. For those matched correctly, we would like it to be matched as early as possible; for those matched incorrectly, we must determine which step provided the false hit.

Our approach was to set initial threshold low in order that Risk Factors could be linked to as many questions in the clinical study as possible. Clearly, this has the risk of a high number of questions incorrectly linked to Risk Factors but our analytical tools allow us to quickly identify the step at which the hit occurred and even the keyword. The purpose was to empirically determine the optimum thresholds for all four steps in matching links. The goal is to maximize matched questions to Risk Factors while minimizing the number of false hits.

5.2. Results and Analysis

We present a number of tables that form the basis of our analysis and discussion. Table 4 shows the breakdown of Risk Factors by each step with the numbers of questions and clusters matched. The figures are cumulative showing most Risk Factors increasing their quantities of matched questions at each step. Both the number of questions and clusters matched are provided to illustrate if questions tend to be matched in a low number of clusters or across many.

In general, matched cluster numbers are low (to be expected), although in the case of *Low Cognitive Activity* it matched questions across a very high number of clusters. This would generally imply that there are many false hits as it is unlikely that a single Risk Factor can be the focus of so many segments of the clinical study, conversely it could be that cognitive activity is a focus of the study, as its focus is ageing and dementia. However, this is limited to just this risk factor – linked to 52 clusters in step 1, and 65 by the end of the process. The table is sorted by the overall number of questions matched to illustrate the range of numbers of matched questions.

Risk Factor	Step 1		Step 2		Step 3		Step 4	
	Q's	C's	Q's	C's	Q's	C's	Q's	C's
Low Cognitive Activity	72	52	141	60	290	65	633	65
Mid life obesity	0	0	92	16	272	70	625	70
Physical Inactivity	0	0	201	63	234	68	358	68
Depression	21	4	102	30	160	35	309	35
Diet	0	0	96	22	105	22	150	22
Alcohol	5	3	37	11	83	17	91	17
Diabetes	5	4	18	7	26	9	74	9
Hypertension	2	2	28	8	41	14	45	14
Cholesterol	3	2	3	2	39	16	43	16
Coronary	25	11	25	11	34	14	38	14
Smoking	3	2	16	11	23	13	30	13
Renal	0	0	0	0	23	8	27	8
Functional imparment	0	0	0	0	0	0	0	0

Table 4: $Tt_1 = 0.3$; $Tt_2 = 0.3$; $Tt_3 = 0.3$; $Tt_4 = 0.2$

The purpose of Table 5 is to illustrate the *degree* of matching by questions. It can be seen that 1143 of 2372 (or 48%) were unmatched. However, it also highlights those questions which appear to be relevant to more than one risk factor. Interestingly, 102 questions are matched to five or more risk factors. Our approach at this point was to determine which steps produced the most false hits and increase thresholds to eliminate as many false hits as possible. As can be seen from the Table 4 and 5 captions, thresholds were both very low and with little variation.

Degree	Questions	Clusters
0	1143	5
1	548	3
2	319	1
3	165	3
4	92	9
5+	102	63

Table 5: $Tt_1 = 0.3$; $Tt_2 = 0.3$; $Tt_3 = 0.3$; $Tt_4 = 0.2$

The next step was to randomly extract between 4 and 6 matched questions for each risk factor and automatically create a validation template to identify false and correct hits. This template was passed to dementia experts for decision making. The template was modified with every match being classed as A (Applicable), PA (Partially Applicable) and NA (Not Applicable). As every word-to-word comparison and every decision for the structural matching step were recorded in a data warehouse style fact table, a number of simple queries could generate statistics which informed us of which steps provided the best and worst hit rates per risk factor. In Definition 2, our query template for retrieving different aspects of the matching operation is shown.

Definition 2. **Result Analytics Sub-expression**
select <*Query Type*>
from Match_Fact_Table
where RF = <*Risk Factor*> **and**
(QID = <*Query ID*> | CID = <*Cluster ID*>)

The expression in Definition 2 is a standard SQL expression with three variables automatically extracted from the validation results, depending on the type of analytics required. *Query Type* can be one of *Step*, *Risk_Factor_Term*, *Question_Term* or *Threshold*. For example if we wish to determine at which *step* a question was linked to a risk factor. The *Risk Factor*, *Query ID*, and *Cluster ID* variables are extracted from the report for those matches that are marked as "Not Appropriate". The clause with QID provides more detailed analysis while the clause with CID provides a more abstract analysis. Example 2 shows a query expression generated by the system.

Example 2. **Result Analytics Sub-expression**
select Step
from Match_Fact_Table
where RF = *'diabetes'* **and** QID = '*loa_u'*;

The system was used to run query expressions for all false hits found in the validation report in order to conduct a high level analysis. In all, the number of false hits from the initial set of thresholds came to just over 70%. The analysis of false hits can be summarized as follows: Step 1 had 11%; Step 2 had 42.5%; Step 3 had 46%; and Step 4 had no false hits. As a result of this process we modified the thresholds for 3 of the 4 steps as shown in the captions for tables 6 and 7. The threshold for step 1 remained the same; thresholds for steps 2 and 3 were significantly higher; while the threshold for step 4 was lowered.

Risk Factor	Step 1		Step 2		Step 3		Step 4	
	Q's	C's	Q's	C's	Q's	C's	Q's	C's
Low Cognitive Activity	72	52	97	58	277	64	807	64
Physical Inactivity	0	0	162	62	183	67	537	67
Depression	21	4	95	27	151	34	508	34
Mid life obesity	0	0	15	3	57	17	230	17
Cholesterol	3	2	3	2	27	8	117	18
Alcohol	5	3	12	7	52	11	86	11
Hypertension	2	2	28	8	41	14	41	14
Coronary	0	0	0	0	27	9	31	9
Renal	0	0	0	0	23	8	27	8
Diabetes	5	4	5	4	13	6	17	6
Diet	0	0	2	1	11	3	11	3
Smoking	3	2	3	2	4	3	8	3
Functional impairment	0	0	0	0	0	0	0	0

Table 6: $Tt_1 = 0.3$; $Tt_2 = 0.8$; $Tt_3 = 0.8$; $Tt_4 = 0.1$

In Table 6, we can see a significant drop in matches across all risk factors by step 3 due to the increased thresholds. As will be explained shortly, this led to a significant decrease in false hits. However by step 4, the final numbers of matched questions were very different from the initial experiment. Risk factors such as *Low Cognitive Activity*, *Physical Inactivity* and *Depression* all showed a significant increase but the majority now matched to less clusters and thus, had their overall numbers of matched questions reduced.

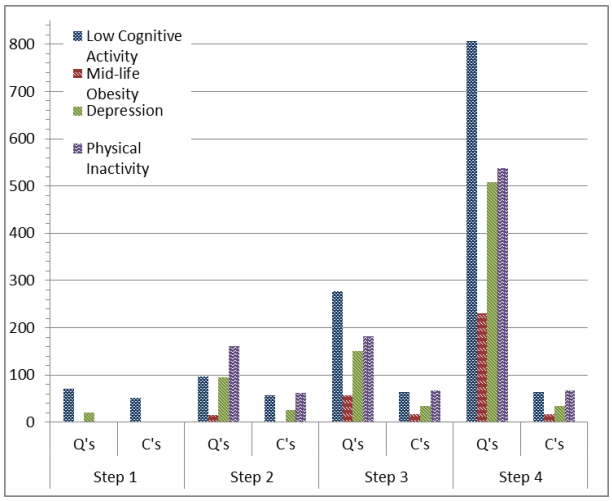

Figure 4a: $Tt_1 = 0.3$; $Tt_2 = 0.8$; $Tt_3 = 0.8$; $Tt_4 = 0.1$

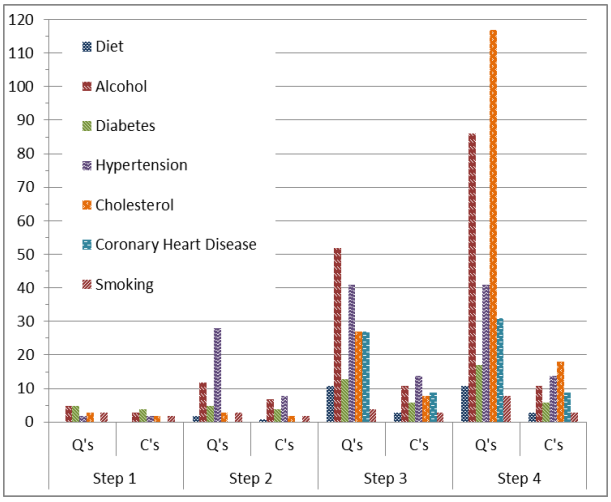

Figure 4b: $Tt_1 = 0.3$; $Tt_2 = 0.8$; $Tt_3 = 0.8$; $Tt_4 = 0.1$

Some risk factors continued to be matched to the same number of clusters but more questions, as previously unmatched questions were added due to the lower cluster-match threshold. The goal is to ensure that the word matching steps are optimized to take advantage of the higher impact of structural matching.

Degree	Questions	Clusters
0	**932**	**9**
1	688	1
2	504	2
3	141	14
4	77	2
5+	28	55

Table 7: $Tt_1 = 0.3$; $Tt_2 = 0.8$; $Tt_3 = 0.8$; $Tt_4 = 0.1$

Table 7 shows the degree of Risk Factor matching for each question. The first row shows the expected decrease in unmatched questions at 932 (39%) which is an improvement in terms of matching on the initial round of experiments. However, when we re-examined the validation template, the number of false hits was now at just under 20%. From this evidence, we were able to make a number of assumptions. Firstly, once provided with initial matching data, the structural matching provides a high degree of accuracy and allows for a high number of matches. Secondly, where properties used to describe Risk Factors were abstract or generic, this led to a high number of false hits. Finally, we detected the fact that matching individual words from Risk Factor *attributes* is where a lot of the false hits occurred. For example, matching 'week' provides false hits whereas matching 'units of alcohol per week' does not.

6. Related Research

Ontology research has attracted great deal of interest from many sectors of the research community since it came into the fore as both a means for knowledge representation and sharing, and a valuable tool for the semantic web. There has been a number of research projects in the area of automating ontology construction [11,14,20], population [6,24], and reuse [5]. Tools have been developed for the problems mentioned using Natural Language Processing (NLP) and Information Extraction (IE) techniques, as well dynamic programming, data standard meta-data and linked data structures. Ontology construction and maintenance is both time-consuming and expensive as it requires a domain expert to perform the task [20,5,6]. Therefore, it is apt that we focus our discussion on state of the art methods that focus on partial or full automation of ontology construction, population, enrichment and reuse.

A project with an aim similar to ours was presented in [11]. Their goal was to enrich clinical trial data with linked data sources by linking existing ontologies - like AMT (Australian Medical Terminologies) and SNOMED CT - to the AIBL (Australian Imaging, Biomarker and Lifestyle) Study of Ageing. Their approach identified instances of a class in the ontology – such as the drug paracetamol – via OpenClinica data standard meta-data (that was used to structure and give meaning to the trial data) and a two-phase mapping process. They then proposed a Linked Clinical Data Cube for more efficient and exhaustive querying of the clinical trial data through its links with the ontologies. Our approach differs in that we are building the ontology from scratch and need to enrich or populate the ontology rather than enriching the clinical trial data. We also use NLP for inexact matching as opposed to standard meta-data as this is not always present in clinical trials. Our aim is also not to see how to link clinical trial concepts with other class instances to see how they interact, but instead to test if risk factors identified in the literature are corroborated in the dataset. In [11], their goal was to determine how drugs mentioned in the clinical trial would interact with other drugs and to identify the chemical name of a drug if given the brand name through the ontological information. In our validation, we use an SQL data cube instead of an RDF cube and this contains results of the matching instead of results from the trial for efficient validation.

The research in [14] and [20] focuses on ontology construction and extension rather than semantic enrichment and querying. In [14], they take pre-

existing ontologies, reformat them and build more exhaustive ontologies in their place. They use NLP techniques – semantic analysis and subject indexing – to change NL attribute descriptions into subject term descriptions for concept attributes. This extracts concepts from the NL description and creates non-taxonomical links between concepts instead of a direct inheritance between concepts and their natural language attribute descriptions. We cannot adopt this approach as clinical studies represent external knowledge sources. In [20] however, the authors do mine external sources but not to link to clinical trials. They instead mine domain texts and glossaries/dictionaries to come up with what they call feature groups and glossary groups based on a seed-ontology. These groups then aid the ontology creator in extending the ontology by presenting possible additions or updates instead of automatically creating links to the external source. A feature group is extracted from a domain text by stripping the stop words and terms irrelevant to the domain and then extracting features (words) depending on their lexical co-occurrence within similar contexts. As with our research, they search for words that have a similar meaning to ontology terms but instead of using a synonym finder in the Python NLTK they mine a number of domain glossaries and extract any terms that exist in two thirds of the definitions.

In [5], the authors focused on automated ontology reuse instead of construction or enrichment. They, like us, use NLP to aid this process. They analyze natural language web-pages to determine which best fit their scope by matching concept names and concept values to those in the ontology. Although they use NLP to link the concepts, relationships and attributes in the documents, it is only to establish what sub-tree of an existing ontology to use. We instead use this to create a link between an ontology and a sub-section of a natural language document to see how it can be best queried in order to test risk hypotheses.

In [6,24], the authors focus on ontology population i.e. instance identification and maintenance (adding newly found concepts to the ontology that are not previously present). In [24], they employ Hidden Markov Models (for each set of instances that belong to an ontological concept) trained on sparsely and semantically annotated corpora. They then use an algorithm to identify matches at runtime. In [6], they use both Natural Language Processing and Information Extraction techniques to populate their ontology. They use NLP to identify instance candidates, IE to construct a classifier and then classify the instances. In both cases, they are populating the ontology whereas we are semantically enriching ours by linking it to a clinical trial and identifying the best areas for domain specialists to query.

Finally in [26], they attempt to intelligently predict the intent of a user's query. This is similar to what we are trying to do in that we are linking one possible user query (a hypothesis about a risk factor or their interactions) [5] with queries that have been tried and tested in clinical trials to yield the best results. Where we differ is that we are linking an ontology to a clinical trial whereas they are mining user query logs and building query trees based on the output.

7. Conclusions

Long term clinical studies provide rich sources of knowledge for researchers in different medical domains. However, the extraction of knowledge has generally been a manual process and in this paper, we presented a means of automating this process. Our approach was to automatically link information from clinical studies to the concepts of interest to medical researchers. We do this by matching those concepts to the questions in clinical studies; and by doing so create direct links to the actual data. Our work was validated through a series of experiments where we sought to match high numbers of questions (where possible) but to ensure that the number of false hits were as low as possible. Our experimental output demonstrated the effectiveness of this approach and our ability to optimize matching levels across different stages of the process.

While our evaluation shows that we can link appropriate segments of the clinical study to risk factors, it is the development of the query interface that will demonstrate the significant reduction in manual effort that this work requires, and will greatly widen its impact. Our current focus is twofold. Firstly, we are working on running our system with multiple clinical studies, both as a means of further testing but also as a mechanism for integration across clinical studies. Secondly, we are building the query interface for which this research was designed. This allows the medical expert to present hypotheses to a clinical study and have our system detect the appropriate parts of the study necessary to compute the results.

8. References

[1] Anjum Ashiz et. al. The Requirements for Ontologies in Medical Data Integration: A Case Study. In Proceedings of the 11th International Database Engineering and Applications Symposium, 2007.

[2] Charlet J., Bachimont B., and Jaulent M.C. Building Medical Ontologies by Terminology Extraction from texts: An experiment for the intensive care units. Computers in Biology and Medicine, vol. 36, pp. 857-870, Elsevier, 2006.

[3] Cristani M. and Cuel R. A Survey on Ontology Creation Methodologies. International Journal on Semantic Web and Information Systems (IJSWIS), vol.1, no. 2, pp. 49-69, IGI Global, 2005.

[4] Dieng-Kuntz R. et al. Building and Using a Medical Ontology for Knowledge Management and Cooperative Work in a Healthcare Network. Computers in Biology and Medicine, vol. 36, pp. 871-892, Elsevier, 2006.

[5] Ding, Yihong, et al. "Generating ontologies via language components and ontology reuse." Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2007. 131-142.

[6] Faria, Carla, Rosario Girardi, and Paulo Novais. "Using domain specific generated rules for automatic ontology population." *Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on*. IEEE, 2012.

[7] Hao Gui, Mark Roantree: A Data Cube Model for Analysis of High Volumes of Ambient Data. Proceedings of the 3rd International Conference on Ambient Systems, Networks and Technologies, Procedia CS 10: 94-101, Elsevier, 2012.

[8] Huz S. and Karras B., A Proposed Ontology for Online Healthcare Surveys, Proceedings of AMIA Annual Symposium Proc., pp.304–308, 2003.

[9] In-MINDD - INnovative, Midlife INtervention for Dementia Deterrence, at: http://www.inmindd.eu/, 2012.

[10] Jolles J, Houx P.J., van Boxtel M.P.J. and Ponds R.W.H.M. (Eds.) Maastricht Aging Study: Determinants of cognitive aging. Maastricht: Neuropsych Publishers, 1995.

[11] Leroux Hugo, and Laurent Lefort. "Using CDISC ODM and the RDF Data Cube for the Semantic Enrichment of Longitudinal Clinical Trial Data." Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences, 2012.

[12] Liu J. and Roantree M. Precomputing queries for personal health sensor environments. ACM Conference on Management of Emergent Digital EcoSystems, ACM Press, pp. 49-56, 2009.

[13] Liu J., Roantree M. and Bellahsene Z. A Schema Guide for accelerating the view adaptation process. 29th International Conference on Conceptual Modeling, LNCS vol. 6412, Springer, pp. 160-173, 2010.

[14] Liu, Yao, et al. "Research on automatic construction of medical ontology." Biomedical Engineering and Computer Science (ICBECS), 2010 International Conference on. IEEE, 2010.

[15] Malhotra A. et. Al. ADO: A disease ontology representing the domain knowledge specific to alzheimer's disease. Alzheimer's & Dementia, article in press, Elsevier, 2013.

[16] Miller George et. al. WordNet: An on-line lexical database. Int. Journal of Lexicography, 3(4), pp. 235-244, 1990.

[17] Natural Language Toolkit V2.0. At: nltk.org, July, 2012.

[18] Novacek V., Laera L. and Handschuh S. Dynamic Integration of Medical Ontologies in Large Scale, In Proceedings of WWW2007/HCLSDI, ACM Press, 2007.

[19] Noy N. and McGuinness D. "Ontology Development 101: A Guide to Creating Your First Ontology." Available at: www.ksl.stanford.edu/people/dlm/papers/ontology101 /ontology101-noy-mcguinness.html, 2001.

[20] Parekh, Viral, Jack Gwo, and Tim Finin. "Mining domain specific texts and glossaries to evaluate and enrich domain ontologies." *International Conference of Information and Knowledge Engineering*. Vol. 37. 2004.

[21] Roantree M., Kennedy J., and Barclay P. Using a Metadata Software Layer in Information Systems Integration. Proceedings of 13th Int. Conf. on Advanced Information Systems Engineering (CAiSE), LNCS 2068, pp. 299-314, Springer, 2001.

[22] Roantree M., Shi J., Cappellari P., O'Connor M.F., Whelan M. and Moyna N. Data transformation and query management in personal health sensor networks. J. Network and Computer Applications 35(4): 1191-1202, 2012.

[23] Slegers, K., van Boxtel, M. P. J., and Jolles, J. Computer use in the Maastricht Aging Study (MAAS): Determinants and relationship with cognitive change. Computers in Human Behavior, 28(1), pp. 1-10, 2012.

[24]Valarakos, Alexandros G., et al. "Enhancing ontological knowledge through ontology population and enrichment." *Engineering knowledge in the age of the Semantic Web*. Springer Berlin Heidelberg, 2004. 144-156.Semantic Web. Springer Berlin Heidelberg, pp 144-156, 2004.

[25] wordnet.princeton.edu/wordnet/documentation/, 2012.

[26] Yin, Xiaoxin, and Sarthak Shah. Building taxonomy of web search intents for name entity queries. Proceedings of the 19th international conference on World Wide Web. ACM, 2010.