Mapping Longitudinal Studies to Risk Factors in an Ontology for Dementia *

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

Maria Pierce, Kate Irving
School of Nursing, Dublin City University, Ireland.
{maria.pierce, kate.irving}@dcu.ie

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

Abstract

A common activity carried out by healthcare professionals is to test various hypotheses on longitudinal study data in an effort to develop new and more reliable algorithms that might determine the possibility of developing certain illnesses. The In-MINDD project provides input from a number of European dementia experts to identify the most accurate model of inter-related risk factors which can yield a personalised dementia risk quotient and profile. This model is then validated against the large population-based prospective Maastricht Aging Study (MAAS) dataset. As part of this overall goal, the research presented in this paper demonstrates how we can automate the process of mapping modifiable risk factors against large sections of the aging study and thus, use information technology to provide more powerful query interfaces.

Keywords **Dementia, Modifiable Risk Factors, Ontology, Word Matching**

1. Introduction

Dementia is a serious loss of cognitive ability beyond what might be expected from normal ageing. Worldwide, the number of people with dementia is currently estimated to be 44 million and expected to reach approximately 76 million by 2030 and 135 million by 2050 [25]. While dementia is one of the most feared age-related conditions and a chronic and progressive illness with no cure, there is now strong evidence that dementia can potentially be delayed by adopting lifestyle changes in midlife aimed at improving cardiovascular health, addressing low mood and poor diet and increasing physical and cognitive activity. Given the huge social and economic costs of dementia, even a delay of one year would make such interventions cost-effective [2,4]. Increasingly, the importance of dementia prevention and the need to take prevention measures based on existing knowledge are being highlighted at an international level.

The In-MINDD project [11], funded by the European Union (In-MINDD FP7/2009-2013), is taking place against this backdrop. It has three main objectives. First, it seeks to create a multi-factorial model for dementia risk, taking into account a broad spectrum of factors including cardiovascular risks, mood, physical and cognitive activity. A key element in creating this model is to test its validity against the Maastricht Aging Study dataset, which is a large population-based prospective dataset. Second, it aims to develop a state-of-the-art online profiler for use in primary care to assess the risk that individuals in mid-life (40-60 years) have of developing dementia in later life. The project will also develop personalised strategies to reduce risks to participants' future cognitive health via a supportive online environment which has access to the best available on-line strategies and locally sourced options for delaying the onset of dementia. Third, it will evaluate the use of the In-MINDD profiler and online support environment with practitioners and patients. The article relates to the first of these three objectives.

In some clinical research projects, sensors can be used to harvest data [27] with the effect of generating very large datasets. In other approaches large studies are compiled over significant periods of time [12] using a question and answer type system to compile knowledge on individuals with different demographics, lifestyles etc. In both cases, the next step is often to build the ontology to provide context and generate new knowledge from the existing dataset. Previous research efforts on constructing medical ontologies have provided frameworks for incorporating data from operational medical systems [22] or tackling the general issue of

*The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2012) under grant agreement no. 304979.

interoperability across medical applications [26,1]. One of the requirements of our project is to develop and test a number of different dementia-risk 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. The ontological approach has been shown to be effective in areas such as intensive care [3] and even in broader healthcare like the Lifeline project [6]. In terms of research into dementia, one of the earliest approaches that involved ontology construction was in [18] 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.

1.2 Research Focus and Contribution

This paper describes our approach to linking knowledge from existing longitudinal studies to risk factors identified by health specialists for the area of dementia. 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 hundreds of questions from the original study. 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 the relevant questions and answers are used and thus, the accuracy of the results can be difficult to measure. Our method matches sections of the clinical study to defined ontology Risk Factors, in a four step process of *Concept Name Match*, *Concept Property Match*, *Vocabulary Match* and *Structural Match*. For the In-MINDD project focusing on dementia, these risk factors are: Low Cognitive Activity, Physical Inactivity, Depression, Mid-life Obesity, Cholesterol, Alcohol, Hypertension, Coronary Disease, Renal Impairment, Diabetes, Diet, Smoking and Functional Impairment (which was later dropped). However, simple keyword matching between these risk factors and the keywords used in question-based studies results in a very low level of matching. While ideas such as ontologies for managing healthcare surveys as proposed in [23] 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. In earlier work [28], we used WordNet [20,31] to generate synonyms at three descriptive levels for Risk Factors: risk factor name, risk factor properties, vocabulary associated with a risk factor. These synonyms together with the original terms were matched against all questions in the clinical study. While there were a number of desirable outcomes that emerged from this approach, it resulted in a high number of false positives where inappropriate questions were matched to some risk factors. As a result a new approach was necessary where the goal was to reduce the false positives as much as possible, while still using the same methodology of matching in four iterations, our previous paper [28] can be referenced for what terms are used in each step.

This focus on the creating links between any longitudinal studies focused on dementia and our ontology risk factors, has three broad contributions.

- We provide a framework in which ontologies can be populated with data from clinical studies;
- We introduce a new matching method which uses word stemming and comparison at the phrase level as opposed to single word matches;
- We audit every comparison between the risk factors and questions in the clinical study so that we can record at which of the four steps, matching is found. This has the benefit that simple queries can detect the precise point at which a false positive occurred. This in turn, provides input for the next iteration of the algorithm to determine selection thresholds for links to the clinical study. The results are shown in the summary statistics of Tables 1, 2 and 3.

1.3 Article Structure

The paper is organized as follows: in Section 2, we discuss related work in this area; in Section 3, the In-MINDD architecture and ontology are presented; in Section 4, two mapping strategies are described; in Section 5, we discuss how the process can provide useful metrics for fine-tuning the process; in Section 6, we present an evaluation of both approaches and demonstrate why our current method outperforms the previous version; and finally, in Section 7, we offer some conclusions.

2. Related Research

Automating ontology construction [13,16,23], population [8,30], and reuse [7] have long been topics of research. There have been many different applications but none seem to have the overall aim of the creation of a system to quickly understand the scope of a longitudinal study and identify relevant areas of interest to a clinical specialist.

The MedLingMap tool [19] is built with a similar goal in mind to ours – finding sections of interest in a corpora – but instead of clinical trial data it is the published literature of a niche research area.

There are research streams [16,24] that overlap with the goals of this research, but instead aim to construct and extend ontologies rather than semantically enrich and query them. The authors of the first project [16] take pre-existing ontologies and use natural language processing (NLP) techniques, specifically semantic analysis and subject indexing, to completely overhaul existing ontologies and give them greater semantic depth. They process natural language (NL) descriptions of attributes and turn them into subject term descriptions for concept attributes. Through this semantic analysis and subject indexing they extract concepts from the NL description and create non-taxonomical links between concepts instead of only having inheritance links, we do not adopt this approach as we are enriching the ontology with external knowledge rather than enriching the links between concepts. Other work [24] links to external sources of information where they mine domain texts, glossaries and dictionaries to extract feature and glossary groups found through the input of a seed-ontology. This is a semi-automated process as these groups aid an ontology creator in extending the ontology by presenting possible additions instead of automatically creating links to an external source. Again, this project is different to the work we have carried out in that it is mining domain corpora and linking them to an ontology rather than linking them to a clinical trial. They extract feature groups by stripping the stop words existent in the domain text as well as words irrelevant to the domain, and only after this do they extract features (words) depending on their lexical co-occurrence within similar contexts

Researchers in [7] focus on automated ontology reuse instead of construction and enrichment, also using NLP to complete this task. Natural language web pages are evaluated to figure which bests fit their criteria through matching elements in the web-pages i.e. concept names and concept values, to those that are in the ontology. The process here is the reverse of the work that we have carried out. They use NLP to establish the link between concepts, relationships and attributes in documents with existing ontology sub-trees whereas we use NLP to link an ontology and a sub-section of the natural language elements in a longitudinal study to determine how the study can best be queried in order to test dementia risk hypotheses.

There are two other bodies of work [8,30] whose aim is to populate ontologies with the use of NLP, but they differ to this research in that they are not linking the ontologies to external sources in order to find the best areas for domain specialists to query. Ontology population consists of instance identification and maintenance (adding new concepts to the ontology that were not previously present). The first [30], uses Hidden Markov Models (HMMs) to recognise instances of a particular ontological concept. The HMMs are trained on sparsely and semantically annotated corpora and the algorithm is used at runtime to identify matches. The second [8] combines two different techniques that are NLP - as previously mentioned – and information extraction (IE). These techniques are combined in the fashion that NLP identifies instance candidates and IE is used to construct a classifier and then classify the instances.

In work most similar to that presented here, researchers [13] use existing ontologies (AMT – Australian Medical Terminologies, SNOMED CT) and link them to the Australian Imaging, Biomarker and Lifestyle (AIBL) study of ageing. Our goal was to enrich the *ontology* from an existing dementia study rather than enriching the *clinical trial data* itself. They identified instances of a class in the ontology, like that of the drug paracetamol via the OpenClinica data standard meta-data (used in the creation of the study to give meaning to the trial data) and a two phase mapping process. Also suggested in the paper was a Linked Clinical Data Cube for more exhaustive querying of the clinical trial data through its links with ontologies. Our approach used NLP to identify inexact matches as opposed to pre-existing meta-data inherent in the trial. Longitudinal studies are not always constructed in a framework with such exhaustive meta-data therefore we think the natural language approach is very pertinent. Also we do not link clinical trial concepts with other class instances to see how they interact, but instead to allow clinical specialists to explore if the clinical, and in this case dementia risk factors that were proposed in the literature are corroborated in the dataset. We also use a data cube in our validations, but we chose to develop in SQL rather than RDF, it contains the results of the matching instead of results in the trial for efficient exploration and validation.

3. The In-MINDD Architecture and Ontology

In this section, we provide an outline view of the In-MINDD ecosystem and briefly describe the main components. While there are different approaches to constructing ontologies [5], a common approach is to identify the basic (dementia) concepts, describe the properties of these concepts, and generate a vocabulary of concept-related keywords. This process here is referred to as *Ontology Initialization*, see Figure 1, and the terms

in each stage of initialisation map directly to those used in each of our three stages of matching. As the In-MINDD project is focused on dementia, these concepts are the associated (dementia) Risk Factors and the properties are those characteristics used to describe or measure a particular Risk Factor. 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 identifiable sub-sections. This process is known as the *Model Clinical Study* phase as all questions are imported and are then sectioned into *clusters* as pre-determined by the study.

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.

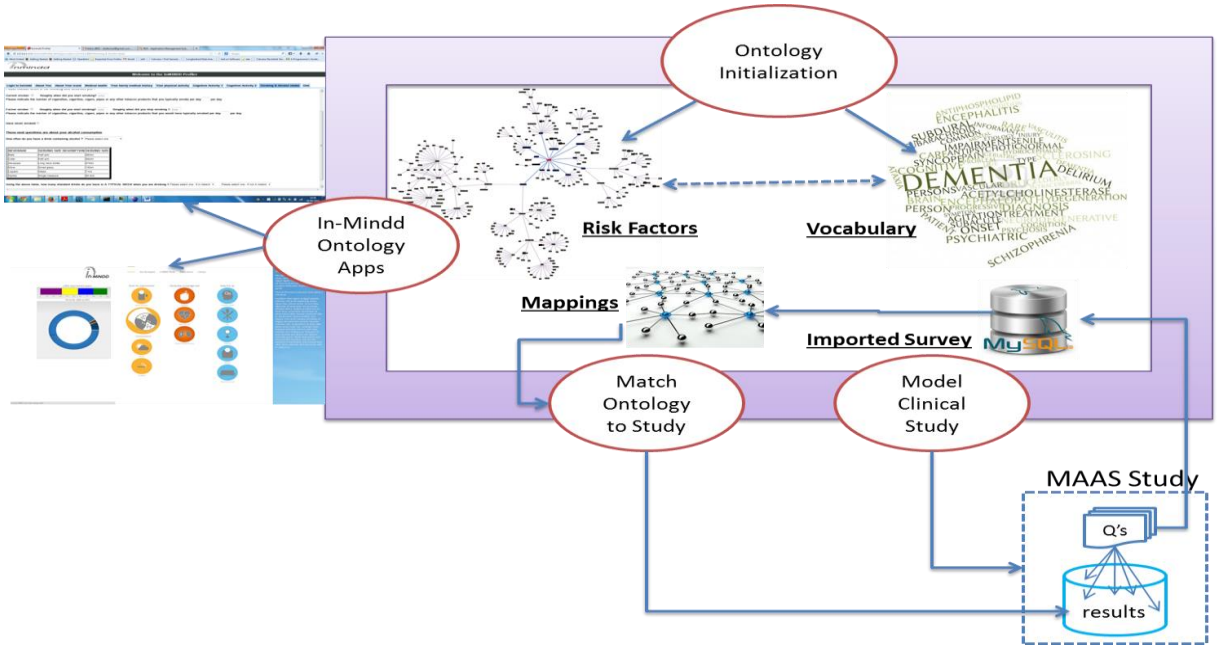

Figure 1: Ontology Population and Applications

The final phase of matching the ontology to the target clinical study is the primary focus of this paper, and is described in depth in the following section and builds on previous work [28]. In brief, the goal is to link each ontological concept (Dementia Risk Factor) with all relevant questions in the clinical study. In Figure 1, 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 pre-processing as required querying or data mining operations can now exploit the links to auto-generate query expressions. The In-Mindd apps pertain to an online profiler for dementia risk, a support environment for those with high risk, and the interface (not yet implemented) to the system described here to allow clinical researchers from the project to query dementia datasets in order to identify variables relevant to testing hypotheses.

The In-MINDD ontology contains 12 major concepts which are Risk Factors related to dementia: alcohol; coronaryHeartDisease; physicalInactivity; chronicKidneyDisease; diabetes; cholesterol; smoking; midlifeObesity; midlifeHypertension; diet; depression; and lowCognitiveActivity. Each concept has a series of properties (measures), a method to calculate a *score* for each risk factor; and an associated vocabulary which helps with matching questions from clinical studies. Due to space limitations, the properties and vocabulary cannot be described here. There have been two cloud-based applications developed using Google App Engine and the ontology: the Profiler which populates the properties of each risk factor and computes the scoring, and the Support forum which offers a means of reducing the *score* for selected risk factors.

4. Mapping Risk Factors

In this section, we present the two to match risk factors to clusters of questions in the clinical study. Method 1 uses NLTK[21] and method 2 uses Lucene[17]. Both use the same four steps as outlined in our previous paper [28]. These are *Concept Name Match*, *Concept Property Match*, *Vocabulary Match* and *Structural Match*. The first three steps focus on matching ontology keywords to terms used in the questions in clinical studies, each at a different level in the ontology, while the fourth uses the structure of the study to match further questions. 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 1); 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.

4.1 NLTK Method Using Synonyms

The NLTK approach was described in [28] and is briefly discussed again here for comparison with our more recent approach (discussed in Section 4.2). The method was to find synonyms for all keywords and use this larger keyword set to map to as much of the clinical study as possible. The underlying natural language technology used was WordNet[31] which allowed us to set thresholds for increasing or decreasing the matching proximity. 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 [31] 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.

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

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:

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

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

As 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.

4.2 Lucene Method Using Stemming and Phrases

Due to the high number of false hits [28] with NLTK, our second method did not use synonyms alone. Instead, we also adopted a stemming approach and used the Lucene library [17] to develop our algorithms. Lucene is an extremely rich and powerful full-text search library distributed by the Apache Software Foundation [17]. At Lucene's core is the Analyzer, which was used to create an inverted index for the MAAS dataset. The algorithms selected were part of the SnowBall analyser and facilitated the usage of word stems (a search on the word 'smoke' will return entries in the document for 'smoking', 'smoked' etc.) and will match whole word phrases like "average units per week". This led to fewer but more accurate synonyms found as well as different suffixes for each word.

As with the WordNet approach, we performed the same four step matching process, although this time, with a far smaller set of terms per risk factor. We took the terms at each of the three levels of the ontology, stemmed

each term and found the appropriate synonyms, which lead to more accurate matching and better input for the structural matching phase. As part of the evaluation, the same audit of all matching operations was performed with one difference: with WordNet RF_{ij} was a single term, while with the Lucene method, RF_{ij} could be a set of terms.

4.3 Structure-Based Mapping

Due to the nature of the questions in clinical studies, there remain many unmatched questions after the first three rounds of word matching. 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 $T_s = 0.3$.

5. Evaluation

The In-MINDD project uses the MAAS dataset for hypothesis testing. MAAS is an epidemiological study into biological, medical and psychosocial aspects of normal and pathological cognitive aging [12], with 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. The In-MINDD Ontology was generated using 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) [21] and WordNet [31] technologies were incorporated into a Python 2.7 application. We used the MySQL database with a data warehouse schema model to capture the result of every comparison operation and its result (matching status and threshold score).

5.1 Evaluation Methodology

Recall that each sub-category or 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. A word or phrase is identified by the triple $\{C_m,Q_n,W_o\}$. Every RF_{ij} is matched against each W_o (excluding stop words) in each question $\{C_m,Q_n\}$ contained in the clinical study. For all 4 steps (concept match, property match, vocabulary match and structure match) 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.

Definition 1. **Fact Table Structure**
FT_{wm} = {CID, QID, RFID, Step, RF_{ti}, $C_mQ_nW_o$, Tti, Result}

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 or phrase identifier; Tt_i, the threshold (used for synonym match only); 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 also have an XML-based approach using [8] for web-based source data.

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
lowCognitiveActivity	72	52	141	60	290	65	633	65
physicalInactivity	0	0	201	63	234	68	358	68
depression	21	4	102	30	160	35	309	35
midlifeObesity	0	0	92	16	272	70	625	70
cholesterol	3	2	3	2	39	16	43	16
alcohol	5	3	37	11	83	17	91	17
midlifeHypertension	2	2	28	8	41	14	45	14
coronaryHeartDisease	25	11	25	11	34	14	38	14
chronicKidneyDisease	0	0	0	0	23	8	27	8
diabetes	5	4	18	7	26	9	74	9
diet	0	0	96	22	105	22	150	22
smoking	3	2	16	11	23	13	30	13

Table 1: Synonym based match with thresholds $Tt_1 = 0.3$; $Tt_2 = 0.3$; $Tt_3 = 0.3$; $Tt_4 = 0.2$

Evaluation forms are automatically generated for all risk factors. The database of matched questions is queried using SQL with the `rand()` function to randomly select 16 questions per risk factor, together with details of which (and how many) of the four steps results in the match. This sample of questions matched to Risk Factors were presented to dementia experts and they were asked to indicate those matches which were relevant (true positives) and those which were not (false positives). 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.

5.2 Evaluation: Degree of Matching

For the WordNet approach using synonyms, we 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. The results of the initial run are shown in Table 1.

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
lowCognitiveActivity	0	0	99	17	99	17	216	17
physicalInactivity	0	0	25	9	38	13	86	13
depression	3	3	6	4	15	7	42	7
midlifeObesity	0	0	11	3	14	4	35	4
cholesterol	3	2	3	2	6	3	10	3
alcohol	7	3	9	3	25	5	29	5
midlifeHypertension	2	2	13	6	41	13	185	13
coronaryHeartDisease	0	0	0	0	37	10	41	10
chronicKidneyDisease	0	0	0	0	30	7	165	7
diabetes	5	4	5	4	17	7	21	7
diet	0	0	7	2	16	3	16	3
smoking	12	2	47	2	51	4	55	4

Table 2: Lucene Matching Results

For Lucene, there were a smaller number of matches as the matching criteria was higher to some degree. While stemming was used to increase matches, no synonyms were used and where phrases were found in the ontology, phase-based matching was used. While thresholds were not required at steps 1 to 3, the threshold of 0.1 was used for structural matching, meaning that if 10% of the questions in a cluster were matched, then the entire cluster was matched.

5.3 Evaluation: Quality of Matching

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.

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'*;

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 3: Synonym based match with thresholds $Tt_1 = 0.3$; $Tt_2 = 0.8$; $Tt_3 = 0.8$; $Tt_4 = 0.1$

5.3.1 Wordnet/Synonym approach

The system was used to run query expressions for all false positives found in the dementia expert's 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% for the matches shown in Table 1. 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 table 3. 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. As Table 3 shows, although a lower number of matches were found, the number of false positives (found through random sampling and evaluation of a dementia expert) decreased to 23%.

5.3.2 Lucene Stemming/Phrase approach

The Lucene approach had a superior quality of matching as a lower number of false positives were found and a second run was not required.. Less than 1% of the matches were found to be false positives in most of the factors, although within some this was not the case. Some risk factors had a significant (> 60 %) number of false positives, i.e. questions having no direct relation to the risk factor. When these were examined it was found that some of the words used for the match did relate to concepts that were close to the risk factor but not directly while a small number were not all related. An example of the latter were words like 'intelligence', 'thinking', 'remembering' that were used when searching for questions related to the risk factor 'low cognitive activity'. These words were matched against many questions, but these questions related to 'cognitive ability' and 'cognitive testing' but not necessarily the stated dementia risk factor, i.e. cognitive inactivity. Similar situations were found in other risk factors and this was found to be the cause of many false positives throughout the application.

Removing these types of keywords from the vocabulary eliminated most of the false positives in the risk factors where they were present. Similarly, use of words like 'activity' when searching for matches against the risk factor

'physical inactivity' produced matches against questions concerning the concept of 'social activity' or 'physical health'. It was harder to eliminate words from the vocabulary that produced such false positives in this case.

When searching for questions related to risk factors such as diet and obesity, we initially identified some questions found as false positives, but as the question did not have to be strictly related to the risk factor, many of these false positives became positives. For example, searches for questions related to *obesity* resulted in questions about *weight*, *physical activity*, lists of conditions that include those associated with obesity, *appetite* and *food consumption*. Whereas *weight* is strictly related to *obesity*, the rest are broadly related to *obesity* and once both types of questions were identified as positives, the number of false positives dropped sharply.

As the quality of matching was superior for the Lucene-based approach, it is worth examining each of the results for the risk factors in this case. After making modifications to the rules and vocabulary, the matching was as follows: Depression, Obesity, Cholesterol, Alcohol, Renal, Diabetes and Diet had no false hits; Smoking and Coronary had less than 10%; while Low Cognitive Activity and Physical Inactivity had almost 50% false hits due to the reasons provided above. To reduce the false hits for the poorest performing risk factors, we adopted a system which looks at how matching occurred at each of the four levels. A threshold was set for specific risk factors whereby unless matching was made at three steps or more, the match was removed. This has the effect of removing a significant number of the false hits for the two worst risk factors.

7. Conclusions

Many of the clinical studies into the various effects of aging commenced ten or twenty years ago. As a result, they are not well suited to modern approaches of defining risk factors and ontologies as a mechanism for better processing data and extracting knowledge. As these studies provide a wealth of information, a strategy for matching older clinical studies with new representations for knowledge is necessary. In this paper, we presented an approach which maps older clinical studies to modern ontologies by building a small set of algorithms on top of existing natural language utilities. Using our evaluation framework, we compared an approach used in earlier work [28] with a new approach and different technology, presented in this paper. Unlike [28] where a high volume of false positives were identified, our current approach (with results in Tables 2 and 3) creates a high level of matching between the ontology's risk factors and knowledge in clinical studies. Together with clinical experts, we demonstrated a low level of false matches through a system of automated evaluation forms and easy detection of where erroneous matches occurred.

While our ontology also provides an opportunity for interoperability across clinical studies and for an XML-based integration with online clinical data, we do not present a discussion here. Instead this forms part of current research building upon the optimization strategies presented in [14,15]. While only a brief description of the In-MINDD Ontology was discussed here, it has evolved from the initial ontology in [28] and is now stable to the point where it is used as a foundation for the Profiler and Support Forum Cloud-based apps. This forms part of a second stream of current research [23] where Cloud-based medical data is anonymised.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2012) under grant agreement no. 304979.

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. Brookmeyer, R., E. Johnson, K. Ziegler-Graham, and H. M. Arrighi (2007). Forecasting the global burden of Alzheimer's disease. Alzheimer Disease & Associated Disorders 3(3), 186-191.
3. 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.
4. Comas-Herrera A, Northey S, Wittenburg R, Knapp M, Bhattacharyya S, Burns A. Future costs of dementia-related long-term care: exploring future scenarios. International Psychogeriatrics, 2011; 23: 20-30.
5. 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.
6. 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.

7. Ding, Yihong, et al. "Generating ontologies via language components and ontology reuse." Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2007. 131-142.
8. 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.
9. 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.
10. Huz S. and Karras B., A Proposed Ontology for Online Healthcare Surveys, Proceedings of AMIA Annual Symposium Proc., pp.304–308, 2003.
11. In-MINDD - INnovative, Midlife INtervention for Dementia Deterrence, at: http://www.inmindd.eu/, 2012.
12. 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.
13. 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.
14. A heuristic approach to selecting views for materialization. Software: Practice and Experience. Article first published online: 13 MAR 2013, DOI: 10.1002/spe.2192.
15. 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.
16. Liu, Yao, et al. "Research on automatic construction of medical ontology." Biomedical Engineering and Computer Science (ICBECS), 2010 International Conference on. IEEE, 2010.
17. Lucene. url: lucene.apache.org/core/index.html, download, September 2013.
18. 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.
19. Meteer, Marie, et al. "MedLingMap: A growing resource mapping the Bio-Medical NLP field." Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, 2012.
20. Miller George et. al. WordNet: An on-line lexical database. Int. Journal of Lexicography, 3(4), pp. 235-244, 1990.
21. Natural Language Toolkit V2.0. At: nltk.org, July, 2012.
22. Novacek V., Laera L. and Handschuh S. Dynamic Integration of Medical Ontologies in Large Scale, In Proceedings of WWW2007/HCLSDI, ACM Press, 2007.
23. Neil Donnelly, Kate Irving, Mark Roantree. Cooperation across Multiple Healthcare Clinics on the Cloud. 14th IFIP Conference on Distributed Applications and Interoperable Systems, LNCS, Springer, pp.82-88, 2014.
24. 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.
25. Prince, M., Guerchet, M. and Prina, M. (2013) Policy Brief for Heads of Government: The Global Impact of Dementia 2013–2050, ADI, London. [at: ww.alz.co.uk/research/GlobalImpactDementia2013.pdf.
26. 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.
27. 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.
28. Mark Roantree, Jim O'Donoghue, Noel E. O'Kelly, Martin van Boxtel, Sebastian Köhler. Automating the Integration of Clinical Studies into Medical Ontologies. 47th Hawaii International Conference on System Sciences, IEEE Press, pp. 2938-2947, 2014.
29. 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.
30. 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.
31. wordnet.princeton.edu/wordnet/documentation/, 2012.
32. 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.