A DATA DRIVEN SEMANTIC FRAMEWORK FOR

CLINICAL TRIAL ELIGIBILITY CRITERIA

A THESIS IN
Computer Science

Presented to the Faculty of the University
of Missouri Kansas City in partial fulfillment of
the requirements for the degree

MASTER OF SCIENCE

by

SARANYA KRISHNAMOORTHY

B.Tech, Shanmugha Arts Science Technology and Research Academy (SASTRA) India, 2007

Kansas City, Missouri

2011

A DATA DRIVEN SEMANTIC FRAMEWORK FOR

CLINICAL TRIAL ELIGIBILITY CRITERIA

Saranya Krishnamoorthy, Candidate for the Master of Science Degree

University of Missouri – Kansas City, 2011

ABSTRACT

An important step in the discovery of new treatments for medical conditions is the matching of potential subjects with appropriate clinical trials. Eligibility criteria for clinical trials are typically specified in free text as inclusion and exclusion criteria for each study. While this is sufficient for a human to guide a recruitment interview, it cannot be reliably parsed to identify potential subjects computationally. Standardizing the representation of eligibility criteria can help in increasing the efficiency and accuracy of this process.

This thesis proposes a semantic framework for intelligent match matching to determine a minimal set of eligibility criteria with maximal coverage of clinical trials. In contrast to top down existing manual standardization efforts, a bottom-up data driven approach is presented that finds the canonical non-redundant representation of an arbitrary collection of clinical trial criteria set to facilitate intelligent match-making. The approach is based on semantic clustering.

The methodology been validated on a corpus of 708 clinical trials related to Generalized Anxiety Disorder containing 2760 inclusion and 4871 exclusion eligibility criteria. This corpus is represented by a relatively small number of 126 inclusion clusters and 175 exclusion clusters, each of which represents a semantically distinct criterion. Internal and external validation measures provide an objective evaluation of the method.

Based on the clustering, an eligibility criteria ontology has been constructed. The resulting model has been incorporated into the development of the MindTrial clinical trial

recruiting system. The prototype for clinical trial recruitment illustrates the real world effectiveness of the methodology in characterizing clinical trials and subjects, and accurate matching between them.

APPROVAL

The faculty listed below, appointed by the Dean of School of Computing and Engineering, have examined a thesis titled "A Data Driven Semantic Framework for Clinical Trial Eligibility Criteria" presented by Saranya Krishnamoorthy, candidate for the Master of Science degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Deendayal Dinakarpandian, MD, PhD, MS, Co-Chair
School of Computing and Engineering

Yugyung Lee, PhD, Co-Chair
School of Computing and Engineering

Praveen Rao, PhD
School of Computing and Engineering

Dennis P. Owens, M.D, PhD
Department of Psychiatry
University of Kansas School of Medicine

TABLE OF CONTENTS

ILLUSTRATIONS

TABLES

LIST OF ABBREVIATIONS

GAD	Generalized Anxiety Disorder
TF-IDF	Term Frequency – Inverse Document Frequency
MESH	Medical Subject Headings
ERGO	Eligibility Rule Grammar and Ontology
URL	Uniform Resource Locator
MCL	Markov Chain Clustering
SPARQL	SPARQL Protocol and RDF Query Language
MSSQL	Microsoft Structured Query Language
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
API	Application Programming Interface
TP	True Positive
FP	False Positive
FN	False Negative

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank the following people who have directly or indirectly helped me in academic achievements. Firstly, I would like to thank Dr. Dinakarpandian Deendayal and Dr. Yugyung Lee, my mentors and advisors, for their continuous support and guidance throughout my master's program in computer science. I sincerely thank Dr. Praveen Rao and Dr. Dennis Owens for accepting to be a part of my thesis committee and making time for me in their busy schedule. Finally, I would like to thank my family members and friends for all their encouragement and support.

The views and conclusions contained herein are those of the author's and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the University of Missouri - Kansas City.

CHAPTER 1

INTRODUCTION

1.1 Research Motivation

Clinical trials play a key role in the drug discovery process to validate a newly discovered drug. Human volunteers are involved in clinical trials to answer specific health questions like drug safety and efficacy. The recruitment of human subjects for clinical trials research is a critically important step in the discovery of new cures for diseases. Volunteers are recruited by a set of inclusion/exclusion criteria designed by the recruiters/researchers of a study in order to minimize the variation in the selection of subjects.

The eligibility criteria are basically a set of conditions which the volunteers or patients must satisfy in order to be recruited for the study. Eligibility criteria are commonly handwritten free text phrases or sentences. Criteria set vary from general criteria like “18 - 70 years of age” and “English speaking” to highly specific criteria like “A Hamilton Depression Rating Scale 17 Item (HDRS-17) score of > 17” or “Meets DSM-IV criteria for generalized social anxiety disorder (GSAD).” While this is sufficient for a human user to guide a recruitment interview, it cannot be reliably parsed or processed computationally.

Standardizing the representation of eligibility criteria can help in reducing ambiguity in their interpretation and also minimize the inter-recruiter variation in subject selection whereby the drug response can be measured more accurately. Standardizing criteria also helps in faster and identifying more matches between subjects and clinical studies as the process can be automated. This can eventually lead to saving cost and time in the planning and execution of clinical trials. Redundancy in recruitment efforts can be reduced if a new study uses criteria that can be precisely mapped to criteria from previous studies.

1.2 Problem Statement

In this thesis, an automated data driven approach to standardize clinical eligibility criteria set and an intelligent matchmaking technique to automatically identify potential participants is proposed. Given a set of textual inclusion/exclusion clinical trial eligibility criteria for a disorder of interest, our data driven model is able to (1) Identify non-redundant minimal set of eligibility criteria from the given collection of criteria, (2) Able to map the newly built criteria to the non-redundant set developed, (3) Identify associative criteria based on selected criteria and (4) Provide an intelligent matchmaking feature for discovering potential subjects for clinical trials

A visual prototype system has been developed as part of the thesis to aid recruiters in developing eligibility criteria set for a study, use an enhanced search engine to identify patients/subjects matching the criteria set selected.

1.3 Thesis Outline

In Chapter 2 we present related work on the various initiatives taken for standardizing the representation of clinical trial eligibility criteria. Chapter 3 describes the Data Driven model for clinical trials. Chapter 4 introduces the patient database system architecture in detail. Chapter 5 demonstrates the development of an ontology for eligibility criteria. Chapter 6 shows a case study using GAD clinical trials. Chapter 7 describes the online interface developed for the proposed system. Chapter 8 shows the evaluation and experimental results of measuring the accuracy of the model proposed. Chapter 9 concludes this thesis and provides information for future work on this system.

CHAPTER 2

RELATED WORK

In this chapter, we discuss about the various efforts carried out in standardizing the eligibility criteria. Standardizing core eligibility criteria is a key factor to achieve computability at semantic, syntactic and knowledge levels [1]. Advances in generic representations of eligibility criteria will provide the necessary semantic foundation for maximizing the ability of computers to help manage and apply complex clinical phenotypes as defined by eligibility criteria in clinical research [2]. Several solutions have been proposed for standardizing the representation of selection criteria ([3], [4], [5], [6], [7], and [8]). Weng et al [3] provides an overview of recent initiatives for establishing structured or semantic representation of eligibility criteria of clinical trials. These include Agreement on Standardized Protocol Inclusion Requirements for Eligibility (ASPIRE), that aims to differentiates "pan-disease criteria" (e.g., age, demographics, functional status, pregnancy, functional status, etc.) from disease-specific criteria using Common Data Elements (CDEs) for encoding or annotating medical concepts in eligibility criteria [4]. CDEs are standardized medical terms primarily developed by NCI (National Cancer Institute) for terms for the collection and exchange of data [5], CaMatch project is primarily focused on eligibility criteria representations for breast cancer using controlled vocabularies [6], ERGO (Eligibility Rule Grammar and Ontology) uses a template Based expression language to encode eligibility criteria [7] and Clinical Observations Interoperability (COI) Task Force by the W3C Interest Group (W3C 2008; [8]). The frontiers of biomedical text mining in general continue to present interesting challenges and opportunities for improvements.

The table below shows a comparison of related work in terms of categorization of criteria and approaches used.

Table 1: Comparison of Related Work

Citations	**Work**	**Criteria Representation**	**Standard Terminologies**
Samson et al [7]	ERGO	Structured template approach using noun phrases, expressions and temporal constraints to represent criteria in a recursive manner. Categorizes criteria by interventions, subject behavior etc.	SNOMEDCT, MESH & UMLS
Cohen et al [6]	CaMatch	Follows a vocabulary driven approach by a structured form compliant with HL7 structured protocol representation effort	No
Niland J et al [4]	ASPIRE	Classifies criteria by diseases and takes concepts from a list of accepted values to represent the criteria	CDISC
Sim et al [12]	TrialBank	Rule based representation of criteria by using logical constructors	No
John et al [11]	CRFQ	Categorizes criteria by demographic, disease and protocol information. Uses an HL7 based representation	No
Metz et al [9]	Onco link	Web based Questionnaire	No

Citations	Work	Criteria Representation	Standard Terminologies
Musen et al [10]	T-Helper	Structured representation of criteria using simple comparison, arithmetic combinations mapped to a template to convert to patient database queries	No

Distinct from the existing human intensive approaches, we propose an automated data driven approach to standardize the representation of eligibility criteria while retaining the flexibility of free text.

CHAPTER 3

DATA DRIVEN MODEL FOR CLINICAL TRIAL ELIGIBILITY CRITERIA

3.1 Introduction

In this chapter we have proposed an intelligent and dynamic model titled as “Data Driven Model for Clinical Trial Eligibility Criteria”. Given a set of clinical trial criteria from a list of studies of a disease, the model identifies the canonical set of eligibility criteria for the target disease. This can be used to guide the recruiter to develop eligibility criteria for a study and map them to patient information.

In this section we describe the sequence of steps underlying the methodology for extracting a representative set of eligibility criteria from a given study collection. The model involves retrieval of studies related to the target disease and extraction of clinical trial eligibility criteria from the study set followed by lexical pre-processing of criteria data to remove noise and finally grouping of criteria into semantic clusters. Figure 1 below shows the overall framework of the model.

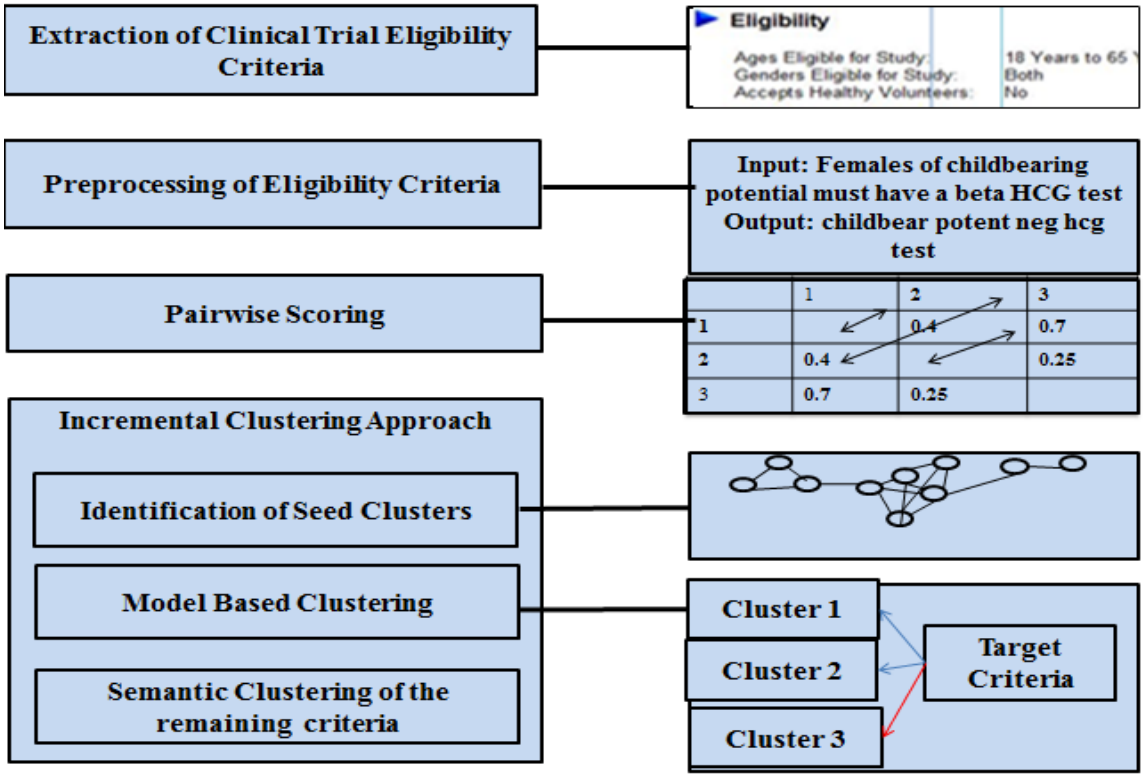


Figure 1: Model Flow

Several concepts are introduced while describing the whole process. The descriptions of some of the concepts widely used in the model are as follows:

Clinical Trial: A clinical trial is basically a research study in which human volunteers are involved to answer specific health questions. The recruitment of human subjects for clinical trials research is a critically important step in the discovery of new cures for diseases. Volunteers are recruited by a set of inclusion/exclusion criteria designed by the recruiters/researchers of a study.

Inclusion/Exclusion Eligibility Criteria: These are basically a set of conditions which the volunteer or patient must satisfy in order to be recruited for the study.

ClinicalTrials.gov: A registry of clinical trials created and maintained by the United State National Library of Medicine at the National Institutes of Health. It is one of the largest clinical trial databases where we can find for a wide range of diseases and conditions.

MCL: MCL **(**Markov chain clustering) is a fast, scalable and unsupervised learning cluster algorithm based on simulation flow in graphs and unlike approaches like *k*-means, MCL does not require an explicit number of clusters to be predefined. MCL identifies the cluster structure of graphs using a mathematical bootstrapping procedure. The MCL algorithm includes a parameter called the inflation factor that determines the granularity of clustering; higher the value of the inflation factor, larger the number of clusters.

Ontology: A formal representation of concepts and the relationship between the concepts in a domain of interest.

SNOMED CT : Systematized Nomenclature of Medicine Clinical terms is a concept oriented clinical terminology covering major areas of clinical information like diseases, substances, findings, procedures etc.

MESH: Medical Subject Heading is a comprehensive controlled medical vocabulary ideally developed by the United states NLM for the purpose of indexing journals and books.

Apriori Algorithm: Apriori algorithm is primarily used for learning association rules based on the following fact "A subset of frequent item set must also be a frequent item set".

Association: Association has a degree that defines how the clusters are related. This can be mathematically expressed as $A_{ij} = E(C_i, C_j)$ where E(x, y) represents an edge between Criteria x and Criteria y. The criteria association is determined based on 3 important factors.

Support: Support is calculated as the ratio of the number of study items where Criteria C_x and Criteria C_y occurs together divided by the total number of study list.

$$S = \frac{Number\ of\ studies\ containing\ C_x\ and\ C_y}{Total\ number\ of\ studies}$$

Confidence: Confidence is calculated as the ratio of the number of study items where Criteria C_x and Criteria Cy occurs together divided by the total number of study list containing Criteria C_x.

$$C = \frac{Number\ of\ study\ containing\ C_x\ and\ C_y}{Total\ number\ of\ studies\ containing\ C_x}$$

Lift: Lift is calculated as the ratio of the probability of Criteria C_x and Criteria C_y occurring together divided by their independent probabilities.

$$L = \frac{P(C_x \cap C_y)}{P(C_x)P(C_y)}$$

The criteria associative rules are identified by the top ranked LIFT scores. We picked the rules with LIFT threshold 1.1 and Support Threshold 0.1.

Porter Stemming Algorithm: The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the common morphological and inflectional endings from words in English. This algorithm is mainly used as a term normalization process that is usually done as a part of pre-processing system before undergoing Information Retrieval systems.

TF-IDF: Term frequency – Inverse Document frequency is a statistical text mining technique used to determine the importance of a term in a cluster. The importance of a term increases proportionally to the number of times a term appears in the cluster but is offset by the frequency of the term in the entire eligibility criteria set.

In this section we give a brief description about the functionality and importance of each phase which are elaborated in later sections.

Criteria Extraction : In this phase we retrieve the criteria from the clinical study of the target disease. The study descriptions are downloaded from ClinicalTrials.Gov (Additional details of the criteria extraction component are discussed in Section 3.2). This functionality was developed as an API which takes in the target disease of interest as input and retrieves the eligibility criteria set related to the target disease, grouping them into inclusion and exclusion criteria separately.

Pre-processing of Eligibility Criteria: The next step is to extract meaningful medical information from eligibility criteria set. This process is carried out to eliminate the noise in the criteria when inter-criteria similarity is identified. The pre-processing is lexical based and involves the following steps (1) Splitting of Criteria (2) Conversion of multi-word tokens into a common representation using lookup table (3) Removal of stop words (4) Mapping of criteria concepts SNOMED-CT and MESH ontological terms (5) Stemming of concepts to extract the stem concepts (Additional details of the pre-processing component are discussed in Section 3.3).

Symmetric Pairwise Scoring: The inter-criteria similarity is calculated at this step. This phase serves as an input to group semantically similar criterion into clusters. Similarity is determined by symmetric pairwise score. Based on the scores we can determine the degree of closeness of the criteria set. (Additional details of the symmetric pairwise component are discussed in Section 3.4).

Incremental Clustering of Eligibility criteria: Clustering is performed in 3 phases to group eligibility criteria. The first phase is to determine the seed clusters by a semantic clustering process. We have used Markov chain clustering to first determine the seed cluster, and then merged the remaining known criteria to the seed clusters using the maximum Ontological TF-IDF approach. The remaining non-clustered unknown criteria set are again clustered using MCL. (Additional details of the Incremental Clustering are discussed in Section 3.5).

Criteria Association: In this phase we predict the relation between any two criteria set based on how closely the two criteria goes together in the entire study set. We determine the association at the cluster level of criteria set. The obtained association rules depict the degree of association between criteria clusters on the overall study set. The association is identified by logic similar to Apriori algorithm. The degree of association is primarily determined by the LIFT and SUPPORT

score of the rules. The closer the clusters, the higher the LIFT score. (Additional details of the criteria association are discussed in Section 3.6).

3.2 Extraction of Clinical Trial Eligibility Criteria

Criteria extraction is the first phase of the model. In this step we download all clinical trial studies for a disease of interest from clinical trials.gov website. The eligibility criteria data are then extracted from the downloaded study set description. The complete process is undergone by a JAVA based API which takes the target disease name and downloads all the studies posted on clinical trials.gov. The clinicaltrials.gov provides an option to download the entire study set in HTML format. Our API uses the "GNU GetUrl" application to download the studies of interest.

The downloaded study documents are then run through another HTML parsing script API to extract eligibility criteria of each study from the study descriptions downloaded. Finally the parsed criteria are grouped separately into Inclusion and Exclusion criteria set based on the tags.

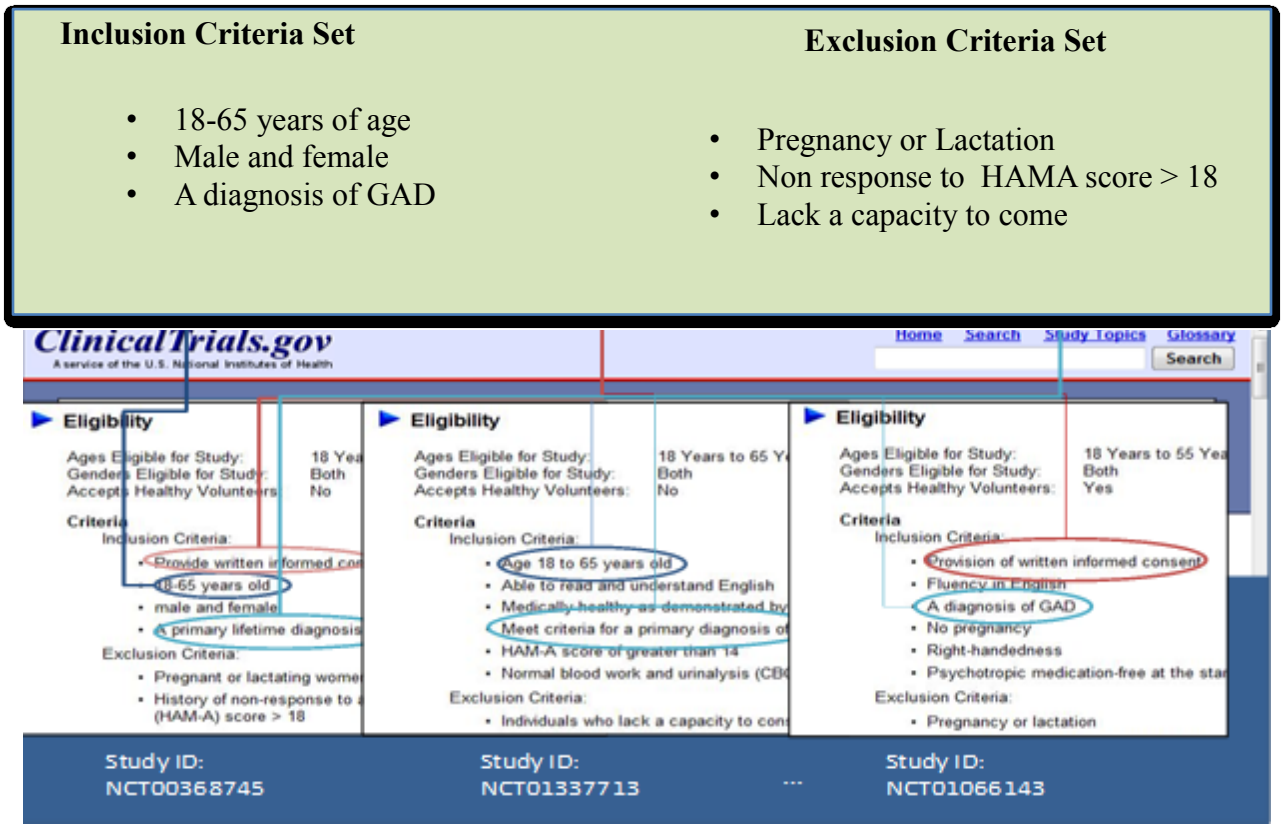


Figure 2: Extraction of Eligibility Criteria from the Study Set

Below is an example of the eligibility criteria of a study

Inclusion Criteria:

- *Provide written informed consent*
- *18-65 years old*
- *male and female*
- *A primary lifetime diagnosis of DSM-IV-TR (2000) GAD (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition)*

Exclusion Criteria:

- *Pregnant or lactating women*

- *Hamilton Anxiety Scale (HAM-A) score > 18*

The eligibility criteria for each study vary from general criteria to study specific criteria.

General Example:

- *English – Speaking*

Study Specific Example:

- *Meets DSM-IV criteria for generalized social anxiety disorder (GSAD)*

3.3 Pre-processing of Eligibility Criteria

3.3.1 Introduction

Pre-processing is a crucial step in the model . Since the criteria are usually written on free english text by various recruiters/scientists it had been difficult to parse the information computationally. Also most of the criterion are ill formed and contain a lot of redundant information. Our goal is to capture non-redundant important medical concepts of each criterion and map them to potential patient information. The raw criterion are mostly complex sentences and contains a lot of noise like special characters, stop words and repetition of words. Another important hurdle was that the semantically same criterion is expressed in different ways in the study set.

Example:

- *Willingness to accept randomization.*
- *Consent to be randomized into treatment*

The pre-processing phase consists of several modules which are explained below. Figure below shows an example of the complete pre-processing flow.

Preprocessing with Eligibility Criteria

- **Original Criteria :** *Females of childbearing potential must have a negative serum or urinary beta HCG test and 25 years of age*
  - **Splitting of criteria :**
    - *Females of childbearing potential must have a negative serum or urinary beta HCG test*
    - *25 years of age*

- **Sample Input:** *Females of childbearing potential must have a negative serum or urinary beta HCG test*
- **Output**: *women childbear potent neg beta-hcg test*
  - **Conversion using Lookup Table** (e.g., *childbearing* ➔ *child-bearing*)
  - **Removal of Stop Words** (e.g., *of, have, a, or*)
  - **Mapping with SNOMED-CT and MESH**
    (e.g., *women, child-bearing , negative, serum, beta-hcg test*)
  - **Stemming of Concepts:** (e.g., *child bearing* ➔ *childbear*)

Figure 3 : Example Describing the Pre-processing Approach

3.3.2 Splitting of Criteria

The first step of the pre-processing phase is the splitting of complex criterion into simple criteria set. On analysis we found that most of the complex and compound criteria are formed with the 4 major connectors which are as follows

- *And /or*
- *with*
- *Who*

Therefore we split the criteria with multiple criteria into simple criteria set based on the connectors. However there were certain exceptions

Example:

- *Patient with age >= 18 years*

Such exceptional cases were treated as exceptions and treated manually.

3.3.3 Conversion using Lookup Tables

The major goal of this is to take care of multi-word token cases and certain domain speacific lexical variations which cannot be handled by Ontological mapping (Section 3.3.4). In this step all the criterion are processed through a set of rules where all the lexical variation and multi-tokens are converted to a common format so that inter-criteria similarity can be calculated more accurately. In this phase we also elminated inconsistent punctuated terms like quotes, semicolons, periods etc as this might impact in identifying conceptual similarity between criteria. Since the existing ontological terminologies hierarchy and their corresponding synonymns list were not able to handle these variations these rules are developed on manual analysis of the domain.

Below is the list of the original string considered and the corresponding replacements made.

- *Original Term : females Replaced Term : women*
- *Original Term : female Replaced Term : women*
- *Original Term : males Replaced Term : men*
- *Original Term : male Replaced Term : men*
- *Original Term : boy Replaced Term : male*
- *Original Term : girl Replaced Term : female*
- *Original Term : subject Replaced Term : patients*
- *Original Term : patient Replaced Term : patients*
- *Original Term : disorders Replaced Term : disorder*
- *Original Term : biopsy proven Replaced Term : biopsy-proven*
- *Original Term : beta hcg Replaced Term : beta-hcg*
- *Original Term : pre dose Replaced Term : pre-dose*

- *Original Term : co morbid Replaced Term : co-morbid*
- *Original Term : out-patient Replaced Term : out-patients*
- *Original Term : out patients Replaced Term : out-patients*
- *Original Term : out patient Replaced Term : out-patients*
- *Original Term : double barrier Replaced Term : double-barrier*
- *Original Term : semi structured Replaced Term : semi-structured*
- *Original Term : diagnostic and statistical manual of mental disorder 4th edition text revision dsm-iv-tr Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : diagnostic and statistical manual of mental disorder fourth edition text revision dsm-iv-tr Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : diagnostic and statistical manual of mental disorder fourth edition text revision dsm iv-tr Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : diagnostic and statistical manual of mental disorder fourth edition text revision dsm iv tr Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : diagnostic and statistical manual of mental disorder 4th edition text revision (dsm-iv-tr) Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*

- *Original Term : diagnostic and statistical manual of mental disorder 4th edition text revision dsm-iv-trâ® Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : diagnostic and statistical manual of mental disorder 4th edition text revision (dsm-iv tr) Replaced Term :*
- *Original Term : diagnostic and statistical manual-iv dsm-iv 30002 Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : dsm iv Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : dsmiv Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision DSM-IV-TRÂ® Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : dsm fourth edition Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : diagnostic manual 4th edition revision Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : diagnostic manual fourth edition revision Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : diagnostic manual fourth edition revision Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*

- *Original Term : dsm-iv Replaced Term : diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)*
- *Original Term : schizophrenia disorder sd Replaced Term : schizophrenia-disorder*
- *Original Term : sd Replaced Term : schizophrenia-disorder*
- *Original Term : schizophrenia disorder Replaced Term : schizophrenia-disorder*
- *Original Term : schizoaffective disorder Replaced Term : schizoaffective-disorder*
- *Original Term : obsessive compulsive disorder ocd Replaced Term : obsessive-compulsive-disorder*
- *Original Term : obsessive-compulsive disorde Replaced Term : obsessive-compulsive-disorder*
- *Original Term : obsessive compulsive disorder Replaced Term : obsessive-compulsive-disorder*
- *Original Term : ocd Replaced Term : obsessive-compulsive-disorder*
- *Original Term : post menopausal Replaced Term : post-menopausal*
- *Original Term : post-menopause Replaced Term : post menopause*
- *Original Term : non smoking Replaced Term : non-smoking*
- *Original Term : generalized anxiety disorder gad Replaced Term : generalized-anxiety-disorder*
- *Original Term : generalized anxiety disorder (gad) Replaced Term : generalized-anxiety-disorder*
- *Original Term : generalized-anxiety disorder Replaced Term : generalized-anxiety-disorder*

- *Original Term : generalized anxiety disorder Replaced Term : generalized-anxiety-disorder*
- *Original Term : gad Replaced Term : generalized-anxiety-disorder*
- *Original Term : panic disorder (pd) Replaced Term : panic-disorder*
- *Original Term : panic disorder pd Replaced Term : panic-disorder*
- *Original Term : panic disorder Replaced Term : panic-disorder*
- *Original Term : major depression disorder (mdd) Replaced Term : major-depression-disorder*
- *Original Term : major depression disorder mdd Replaced Term : major-depression-disorder*
- *Original Term : major depression disorder Replaced Term : major-depression-disorder*
- *Original Term : mdd Replaced Term : major-depression-disorder*
- *Original Term : nonpregnant Replaced Term : non-pregnant*
- *Original Term : non pregnant Replaced Term : non-pregnant*
- *Original Term : non nursing Replaced Term : non-nursing*
- *Original Term : post traumatic stress disorder (ptsd) Replaced Term : post-traumatic-stress-disorder*
- *Original Term : post traumatic stress disorder ptsd Replaced Term : post-traumatic-stress-disorder*
- *Original Term : post-traumatic-stress disorder Replaced Term : post-traumatic-stress-disorder*
- *Original Term : post-traumatic stress disorder Replaced Term : post-traumatic-stress-disorder*

- *Original Term : post-traumatic stress disorder ptsd Replaced Term : post-traumatic-stress-disorder*
- *Original Term : ptsd Replaced Term : post-traumatic-stress-disorder*
- *Original Term : hipaa Replaced Term : health-insurance-portability-and-accountability-act-(hipaa)*
- *Original Term : health insurance portability and accountability act Replaced Term : health-insurance-portability-and-accountability-act-(hipaa)*
- *Original Term : semi structured Replaced Term : semi-structured*
- *Original Term : social-anxiety disorder Replaced Term : social-anxiety-disorder*
- *Original Term : social anxiety disorder Replaced Term : social-anxiety-disorder*
- *Original Term : separation anxiety disorder sad Replaced Term : separation-anxiety-disorder-(sad)*
- *Original Term : separation anxiety disorder (sad) Replaced Term : separation-anxiety-disorder-(sad)*
- *Original Term : separation-anxiety disorder Replaced Term : separation-anxiety-disorder-(sad)*
- *Original Term : sad Replaced Term : separation-anxiety-disorder-(sad)*
- *Original Term : social phobia sp Replaced Term : social-phobia*
- *Original Term : social phobia (sp) Replaced Term : social-phobia*
- *Original Term : sp Replaced Term : social-phobia*
- *Original Term : social phobia Replaced Term : social-phobia*
- *Original Term : personality disorder (pd) Replaced Term : personality-disorder*
- *Original Term : personality disorder pd Replaced Term : personality-disorder*

- *Original Term : personality disorder Replaced Term : personality-disorder*
- *Original Term : personality disorder Replaced Term : personality-disorder*
- *Original Term : pd Replaced Term : personality-disorder*
- *Original Term : hamilton-anxiety-rating-scale-(ham-a) Replaced Term : hamilton anxiety rating scale*
- *Original Term : hamilton anxiety rating scale (ham-a) Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hamilton anxiety rating scale ham-a Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hamilton anxiety rating scale hama Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hamilton anxiety rating scale (hama) Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hamilton anxiety rating scale (ham a) Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hamilton anxiety rating scale ham a Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hamilton-anxiety-rating scale ham-a Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hamilton-anxiety-rating scale (ham-a) Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hama Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : ham-a Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*

- *Original Term : ham a Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hamilton-anxiety-rating scale Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hamilton depression rating scale (ham-d) Replaced Term : hamilton-depression-rating-scale-(ham-d)*
- *Original Term : hamilton depression rating scale ham-d Replaced Term : hamilton-depression-rating-scale-(ham-d)*
- *Original Term : hamilton depression rating scale ham d Replaced Term : hamilton-depression-rating-scale-(ham-d)*
- *Original Term : hamilton depression rating scale (ham d) Replaced Term : hamilton-depression-rating-scale-(ham-d)*
- *Original Term : hamilton-depression-rating-scale (ham-d) Replaced Term : hamilton-depression-rating-scale-(ham-d)*
- *Original Term : hamilton-depression-rating scale (ham-d) Replaced Term : hamilton-depression-rating-scale-(ham-d)*
- *Original Term : hamilton-depression-rating scale Replaced Term : hamilton-depression-rating-scale-(ham-d)*
- *Original Term : ham-d Replaced Term : hamilton-depression-rating-scale-(ham-d)*
- *Original Term : hamd Replaced Term : hamilton-depression-rating-scale-(ham-d)*
- *Original Term : hamilton anxiety scale Replaced Term : hamilton-anxiety-rating-scale-(ham-a)*
- *Original Term : hamilton depression rating scale Replaced Term : hamilton-depression-rating-scale-(ham-d)*

- *Original Term : montgomery asberg depression rating scale (madrs) Replaced Term : montgomery-asberg-depression-rating-scale-(madrs)*
- *Original Term : montgomery asberg depression rating scale madrs Replaced Term : montgomery-asberg-depression-rating-scale-(madrs)*
- *Original Term : montgomery asberg depression rating scale Replaced Term : montgomery-asberg-depression-rating-scale-(madrs)*
- *Original Term : montgomery-asberg-depression-rating scale madrs Replaced Term : montgomery-asberg-depression-rating-scale-(madrs)*
- *Original Term : montgomery-asberg-depression-rating scale Replaced Term : montgomery-asberg-depression-rating-scale-(madrs)*
- *Original Term : madrs Replaced Term : montgomery-asberg-depression-rating-scale-(madrs)*
- *Original Term : hospital anxiety and depression scale hads Replaced Term : hospital-anxiety-and-depression-scale-(hads)*
- *Original Term : hospital anxiety and depression scale (hads) Replaced Term : hospital-anxiety-and-depression-scale-(hads)*
- *Original Term : hospital anxiety and depression scale Replaced Term : hospital-anxiety-and-depression-scale-(hads)*
- *Original Term : hads Replaced Term : hospital-anxiety-and-depression-scale-(hads)*
- *Original Term : clinician administered post traumatic stress disorder (caps) Replaced Term : clinician-administered-post-traumatic-stress-disorder-(caps)*
- *Original Term : clinician administered post traumatic stress disorder caps Replaced Term : clinician-administered-post-traumatic-stress-disorder-(caps)*

- *Original Term : clinician administered post-traumatic-stress-disorder caps Replaced Term : clinician-administered-post-traumatic-stress-disorder-(caps)*
- *Original Term : clinician administered post-traumatic stress disorder caps Replaced Term : clinician-administered-post-traumatic-stress-disorder-(caps)*
- *Original Term : clinician administered post-traumatic-stress-disorder Replaced Term : clinician-administered-post-traumatic-stress-disorder-(caps)*
- *Original Term : clinician administered post-traumatic stress disorder Replaced Term : clinician-administered-post-traumatic-stress-disorder-(caps)*
- *Original Term : caps Replaced Term : clinician-administered-post-traumatic-stress-disorder-(caps)*
- *Original Term : child depression rating scale (cdrs) Replaced Term : child-depression-rating-scale-(cdrs)*
- *Original Term : child depression rating scale cdrs Replaced Term : child-depression-rating-scale-(cdrs)*
- *Original Term : child-depression-rating scale (cdrs) Replaced Term : child-depression-rating-scale-(cdrs)*
- *Original Term : child-depression-rating scale cdrs Replaced Term : child-depression-rating-scale-(cdrs)*
- *Original Term : child-depression-rating scale Replaced Term : child-depression-rating-scale-(cdrs)*
- *Original Term : child depression rating scale Replaced Term : child-depression-rating-scale-(cdrs)*
- *Original Term : cdrs Replaced Term : child-depression-rating-scale-(cdrs)*

- *Original Term : childrens depression inventory parent version (cdi-p) Replaced Term : childrens-depression-inventory-parent-version-(cdi-p)*
- *Original Term : childrens depression inventory parent version cdi-p Replaced Term : childrens-depression-inventory-parent-version-(cdi-p)*
- *Original Term : childrens depression inventory parent version cdi p Replaced Term : childrens-depression-inventory-parent-version-(cdi-p)*
- *Original Term : childrens depression inventory parent version cdip Replaced Term : childrens-depression-inventory-parent-version-(cdi-p)*
- *Original Term : cdi-p Replaced Term : childrens-depression-inventory-parent-version-(cdi-p)*
- *Original Term : cdi p Replaced Term : childrens-depression-inventory-parent-version-(cdi-p)*
- *Original Term : cdip Replaced Term : childrens-depression-inventory-parent-version-(cdi-p)*
- *Original Term : childrens depression inventory parent version Replaced Term : childrens-depression-inventory-parent-version-(cdi-p)*
- *Original Term : childrens depression inventory (cdi) Replaced Term : childrens-depression-inventory-(cdi)*
- *Original Term : childrens depression inventory cdi Replaced Term : childrens-depression-inventory-(cdi)*
- *Original Term : cdi Replaced Term : childrens-depression-inventory-(cdi)*
- *Original Term : childrens depression inventory Replaced Term : childrens-depression-inventory-(cdi)*

- *Original Term : children depression inventory Replaced Term : childrens-depression-inventory-(cdi)*
- *Original Term : childrens global assessment scale (cgas) Replaced Term : childrens-global-assessment-scale-(cgas)*
- *Original Term : childrens global assessment scale cgas Replaced Term : childrens-global-assessment-scale-(cgas)*
- *Original Term : childrens-global-assessment scale (cgas) Replaced Term : childrens-global-assessment-scale-(cgas)*
- *Original Term : childrens-global-assessment scale Replaced Term : childrens-global-assessment-scale-(cgas)*
- *Original Term : children-global-assessment scale Replaced Term : childrens-global-assessment-scale-(cgas)*
- *Original Term : childrens global assessment scale Replaced Term : childrens-global-assessment-scale-(cgas)*
- *Original Term : children global assessment scale Replaced Term : childrens-global-assessment-scale-(cgas)*
- *Original Term : cgas Replaced Term : childrens-global-assessment-scale-(cgas)*
- *Original Term : covi anxiety scale (cas) Replaced Term : covi-anxiety-scale-(cas)*
- *Original Term : covi anxiety scale cas Replaced Term : covi-anxiety-scale-(cas)*
- *Original Term : covi anxiety scale Replaced Term : covi-anxiety-scale-(cas)*
- *Original Term : cas Replaced Term : covi-anxiety-scale-(cas)*
- *Original Term : raskin depression scale (rds) Replaced Term : raskin-depression-scale-(rds)*

- *Original Term : raskin depression scale rds Replaced Term : raskin-depression-scale-(rds)*
- *Original Term : raskin depression scale Replaced Term : raskin-depression-scale-(rds)*
- *Original Term : rds Replaced Term : raskin-depression-scale-(rds)*
- *Original Term : simpson angus scale (sas) Replaced Term : simpson-angus-scale-(sas)*
- *Original Term : simpson angus scale sas Replaced Term : simpson-angus-scale-(sas)*
- *Original Term : simpson angus scale Replaced Term : simpson-angus-scale-(sas)*
- *Original Term : sas Replaced Term : simpson-angus-scale-(sas)*
- *Original Term : child bearing Replaced Term : child-bearing*
- *Original Term : bipolar disorder Replaced Term : bipolar-disorder*
- *Original Term : bd Replaced Term : bipolar-disorder*
- *Original Term : birth control Replaced Term : birth-control*
- *Original Term : comorbid anxiety disorder Replaced Term : comorbid-anxiety-disorder*
- *Original Term : cad Replaced Term : comorbid-anxiety-disorder*
- *Original Term : anxiety disorder Replaced Term : anxiety-disorder*
- *Original Term : depressive disorder Replaced Term : depressive-disorder*
- *Original Term : eating disorder Replaced Term : eating-disorder*
- *Original Term : dysthymic disorder Replaced Term : dysthymic-disorder*
- *Original Term : calgary depression scale (cds) Replaced Term : calgary-depression-scale-(cds)*
- *Original Term : calgary depression scale cds Replaced Term : calgary-depression-scale-(cds)*

- *Original Term : calgary depression scale Replaced Term : calgary-depression-scale-(cds)*
- *Original Term : cds Replaced Term : calgary-depression-scale-(cds)*
- *Original Term : coloured analogue scale Replaced Term : coloured-analogue-scale*
- *Original Term : cas Replaced Term : coloured-analogue-scale*
- *Original Term : diabetes mellitus Replaced Term : diabetes-mellitus*
- *Original Term : abnormal pap Replaced Term : abnormal-pap*
- *Original Term : affective disorder Replaced Term : affective-disorder*
- *Original Term : depression rating scale Replaced Term : depression-rating-scale*
- *Original Term : drs Replaced Term : depression-rating-scale*
- *Original Term : gds Replaced Term : geriatric-depression-score*
- *Original Term : geriatric depression score Replaced Term : geriatric-depression-score*
- *Original Term : epworth sleepiness scale Replaced Term : epworth-sleepiness-scale*
- *Original Term : ess Replaced Term : epworth-sleepiness-scale*
- *Original Term : mattis dementia rating Replaced Term : mattis-dementia-rating*
- *Original Term : mdr Replaced Term : mattis-dementia-rating*
- *Original Term : pre menopausal Replaced Term : pre-menopausal*
- *Original Term : left handed Replaced Term : left-handed*
- *Original Term : anti obesity Replaced Term : anti-obesity*
- *Original Term : follow up Replaced Term : follow-up*
- *Original Term : greater than or equal to Replaced Term : > =*
- *Original Term : greater than equal to Replaced Term : > =*
- *Original Term : more than or equal to Replaced Term : > =*

- *Original Term : more than equal to Replaced Term : > =*
- *Original Term : older than or equal to Replaced Term : > =*
- *Original Term : >/= Replaced Term : > =*
- *Original Term : </= Replaced Term : < =*
- *Original Term : >= Replaced Term : > =*
- *Original Term : >= Replaced Term : > =*
- *Original Term : <= Replaced Term : < =*
- *Original Term : greater than Replaced Term : >*
- *Original Term : more than Replaced Term : >*
- *Original Term : older than Replaced Term : >*
- *Original Term : less than or equal to Replaced Term : < =*
- *Original Term : less than equal to Replaced Term : < =*
- *Original Term : <= Replaced Term : < =*
- *Original Term : less than Replaced Term : <*

3.3.4 Removal of Stop Words

The next step in the criteria pre-processing is the removal of stop words from the criteria. We used Lingua-Stopword-0.09 to eliminate these stop words from the criteria. The complete list of 174 stopwords is given below.

{"i, me, my, myself, we, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, would, should, could, ought, i'm, you're, he's, she's, it's, we're, they're, i've,

you've, we've, they've, i'd, you'd, he'd, she'd, we'd, they'd, i'll, you'll, he'll, she'll, we'll, they'll, isn't, aren't, wasn't, weren't, hasn't, haven't, hadn't, doesn't, don't, didn't, won't, wouldn't, shan't, shouldn't, can't, cannot, couldn't, mustn't, let's, that's, who's, what's, here's, there's, when's, where's, why's, how's, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, though, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very"}

3.3.5 Mapping with SNOMED-CT and MESH

In this step the remaining criterion terms are mapped to medical ontologies (SNOMED CT and MESH). By this way we extract only the semantic information of each criterion. Based on our analysis we found some terms missed out in the SNOMEDCT and MESH ontologies but relevant to the domain so we also included them. Below is the 182 LOOKUP terms identified for the mental disorder clinical trial domain.

{"cohort-Part, co-morbid, pre-dose, biopsy-proven, K-SADS-E, CDI-P, SIGH-A, CY-BOCS, Y-BOCS, LSAS, BAI, beta-HCG, BSPS, out-patients, double-barrier, semi-structured, Diagnostic-and-Statistical-Manual-of-Mental-disorder-4th-Edition-Text-Revision-(DSM-IV-TR), schizophrenia-disorder, schizoaffective-disorder, Obsessive-Compulsive-Disorder, Post-menopausal, post-menopause, non-smoking, generalized-anxiety-disorder, Panic-Disorder, major-depression-disorder, non-pregnant, non-nursing, Post-Traumatic-Stress-Disorder, Health-Insurance-Portability-and-Accountability-Act-(HIPAA), semi-structured, social-anxiety-disorder, Separation-Anxiety-Disorder-(SAD), social-phobia, Personality-Disorder, Hamilton-Anxiety-Rating-Scale-(HAM-A), Hamilton-Depression-Rating-Scale-(HAM-D), Montgomery-

Asberg-Depression-Rating-Scale-(MADRS), Hospital-Anxiety-and-Depression-Scale-(HADS), clinician-administered-Post-Traumatic-Stress-Disorder-(CAPS), Child-Depression-Rating-Scale-(CDRS), Clinical-Global-Impression-Severity, Childrens-Depression-Inventory-Parent-Version-(CDI-P), Childrens-Depression-Inventory-(CDI), Childrens-Global-Assessment-Scale-(CGAS), Covi-Anxiety-Scale-(CAS), Raskin-Depression-Scale-(RDS), Simpson-Angus-Scale-(SAS), Personality-Disorder, child-bearing, CGI-ADHD-4, CGI-S, follow-up, bipolar-disorder, VCUG, understand, comply, requirements, study, written, consent, , item-1, item-2, item-3, item-5, 14-item, 17-item, 16-item, informed, voluntarily, signed , 21-Item, birth-control, comorbid-anxiety-disorder, anxiety-disorder, depressive-disorder, eating-disorder, Dysthymic Disorder, ASA, BMI, Regenerative nodules, walk, key, healthy, medically, Physically, DTS, reliable, HRSD, VAS, mMS, FIQ, HTQ, Calgary-Depression-Scale-(CDS), protocol, rIL-2, HARS, IBS, HDRS, HARS, MST, Coloured-Analogue-Scale, Diabetes-Mellitus, Skin-Picking, NeP, abnormal-pap, psychotropic, medications, cognitive, affective-disorder, Refractoriness, child-bearing, Depression-Rating-Scale, Geriatric-Depression-Score, Epworth-Sleepiness-Scale, EDSS, MMSE, MATTIS-Dementia-Rating, CDRS-R, Hachinski, YMRS, BDI, Renal, pre-menopausal, left-handed, anti-obesity, Claustrophobia, Pretreatment"}

3.3.6 Stemming Of Concepts

This is final processing step for the criteria set. In this phase we extract the root form of the ontological and lookup terms of each criterion abtained by a stemming technique. In this way we take of the morphological or inflexional suffixes of the terms before calculating the inter-criteria similarity. We have used Porter Stemming algorithm to carry out the stemming process.

3.4 Symmetric Pairwise Scoring

Symmetric pairwise scoring is the next phase of the model. This component is used to calculate the pairwise scores for each criteria to all the other members in the list, therby building a n*n dimensional pairwise score matrix.

Below is the formulae used to calculate the pairwise score matrix. Let us consider Criteria A and B for which the pairwise score is calculated. PS_A is calculated as the ratio of the number of terms present in both Criteria A and by the total number of terms in A. Similarly PS_B is calculated as the ratio of the number of terms present in both Criteria A and by the total number of terms in B. Finally PS_{AB} is the average of the scores A and B.

Symmetric Pairwise scoring technique is used to answer PS_A how close is Criteria A to B and PS_B is to answer how close is B to A. By considering both and taking their average score we prevent the similarity score falling too low for unequal criteria length.

The below figure describes the overall flow of the component with an example.

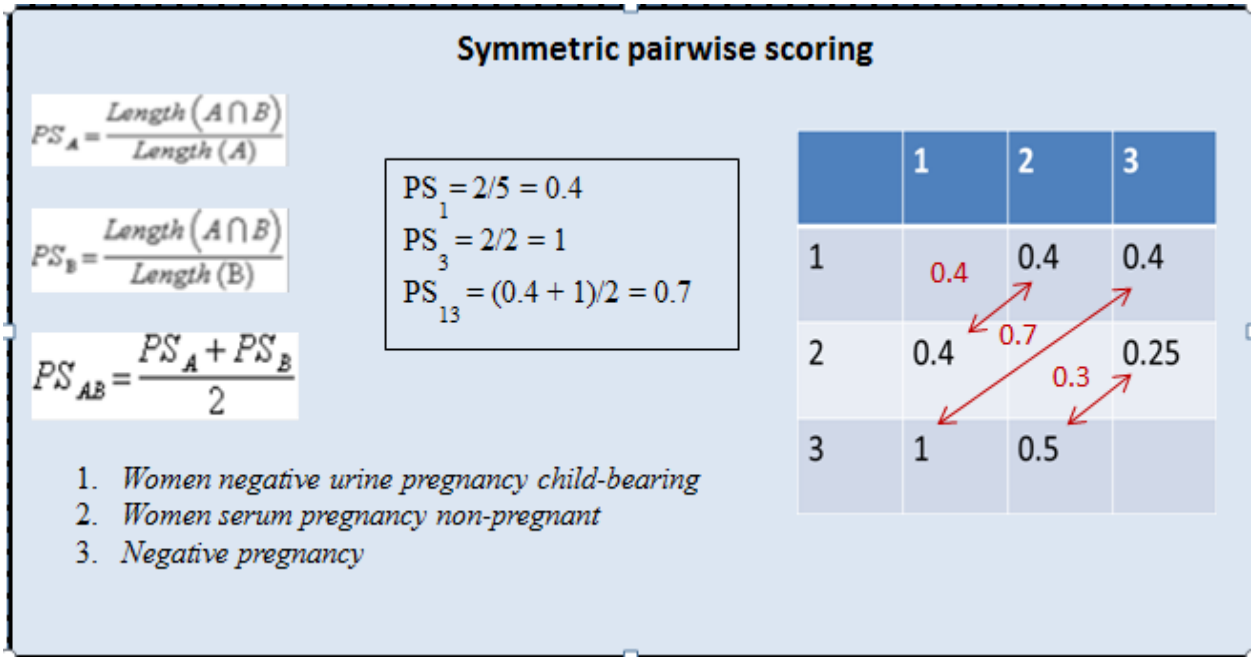


Figure 4 Pairwise Scoring Matrix of Eligibility Criteria Set

3.5 Incremental Clustering of Eligibility Criteria

3.5.1 Introduction

Clustering is an important phase in this model. Similar criteria concepts are grouped together into a single cluster. In this phase clustering is done in an incremental phase by first constructing seed clusters for a large sample data set. Markov Clustering was used to construct seed clusters from the given dataset. Then tried to merge other known criteria to the seed clusters by a model based clustering approach. Then the remanining unknown criteria are clustered separately by semanntic clustering approach as in phase 1.Finally all the clusters are put together to construct the canonical non-redundant set. The reason for following an incremental approach rather than clustering all at one go is because MCL is primarily based on simulation flow of graphs. More the number of criteria set MCL tend to provide coarse and unnatural cluster formation.

3.5.2 Semantic Clustering to Identify Seed Clusters

MCL is widely used algorithm based on network flow paradigms and doesn't need an explicit number of clusters to be predefined. The MCL clusters includes a parameter called the inflation factor that is used to determine the granualrity of clustering, higher the value of inflation factor , larger the number of clusters. The F-measure was used to determine the best value for the parameter discussed in Section 8.3. The above clustering algorithm takes in the pairwise score matrix obtained from the previous step as input and outputs the criteria clusters. The seed clusters are obtained from a large sample of criteria. As per our GAD case study (Section 6.2) we developed our seed clusters using inclusion criteria set.

3.5.3 Identication of the TF-IDF Terms for Seed Clusters

This phase is used to identify unique concepts specific to each seed cluster. We have Term Frequency- Inverse Document Frequency (TF-IDF) statistical text mining technique to determine the important concepts of a cluster. The TF-IDF term determines how important a term to the cluster normalized by the term frequency in the entire criteria set.

First we determined the TF (term frequency) of the ontological and lookup terms of each criterion in a cluster

$$\text{TF(term A)} = \frac{Number\ of\ times\ term\ A\ occurs\ in\ the\ critreria\ set}{Total\ number\ of\ \text{ontological} + \text{Lookup terms in the criteria set}}$$

$$\text{IDF (term A)} = \frac{Total\ Number\ of\ clusters\ formed\ by\ the\ criteria\ set}{(number\ of\ clusters\ where\ the\ term\ A\ is\ present)}$$

In the above formulae 1 is added to the denominator if the term A is not present in any other documents , therby preventing the denominator fallling to 0. Then the weighted TF-IDF score is calculated by the below formulae.

$$\text{TF} - \text{IDF (term A)} = \text{TF(term A)} * \text{IDF(Term A)}$$

Finally from the list of tf-idf terms of each cluster we set a threshold of 0.08 to pick the top ranked TF-IDF terms for each cluster.

3.5.4 Model based Clustering to Merge Known Criteria to Seed Clusters

Based on the TF-IDF terms obtained for the seed cluster set, we tried to merge the the other known criteria set to the seed clusters. We took advantage of the redundant criteria information and as our overall goal was to find the minimum non-redunndant canonical criteria set we followed a model based clustering approach. One example of exclusion and inclusion criteria overlap is demonstrated below.

Example:

- *Women not pregnant - Inclusion Criteria*
- *Pregnancy - Exclusion Criteria*

The only difference in the above criterion set is that the inclusion criterion is negated but both speaking about the same concept. The merging of known criteria to seed clusters is carried out in the below manner. The target criterion ontological and lookup terms are taken and compared with the TF-IDF of each cluster. Then it is merged to the seed cluster where the target criteria terms form the maximum subset of the TF-IDF of the merging seed cluster. Following figure below shows an example of the merging process

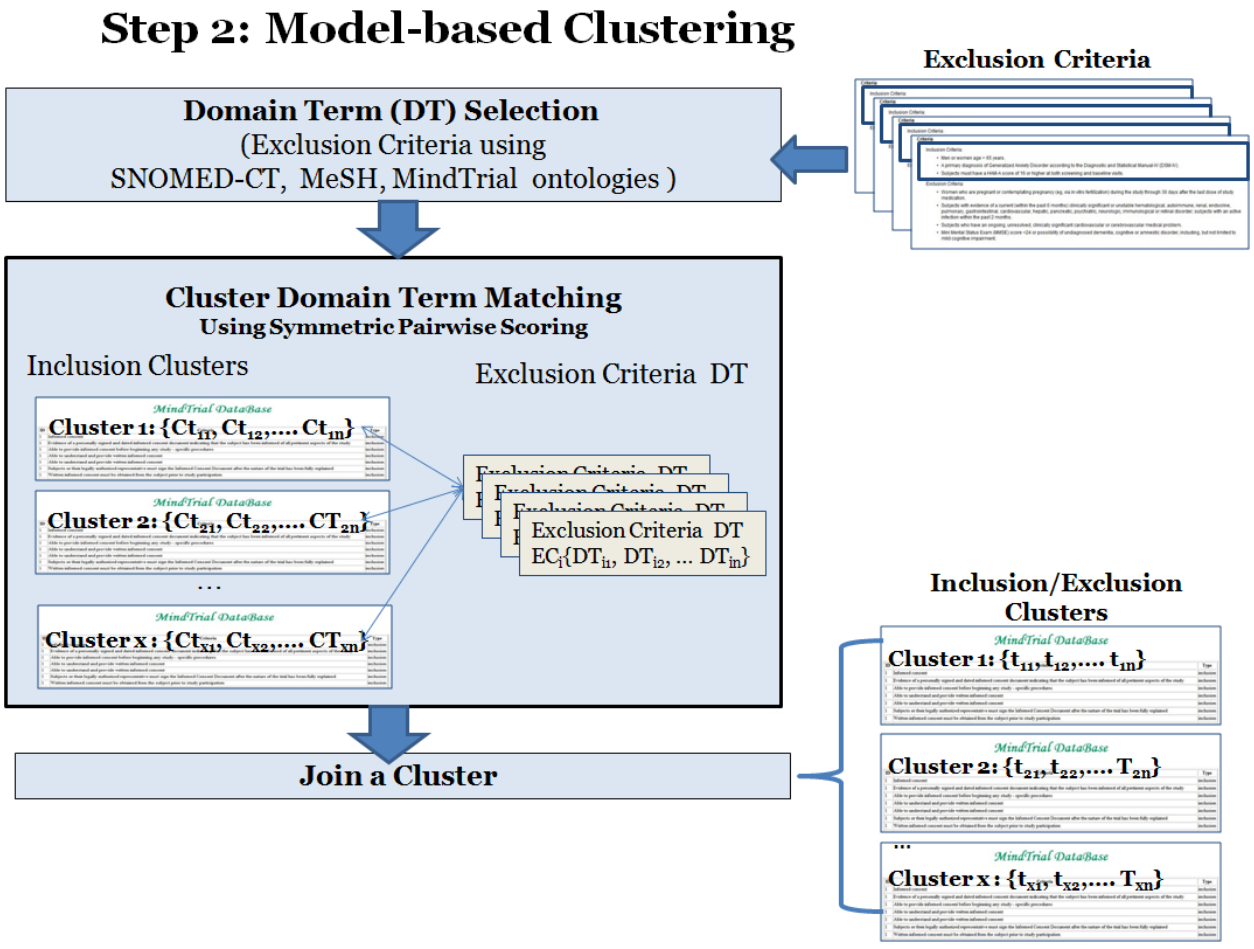


Figure 5: Demonstration of Model Based Clustering

The remaining unknown criteria which cannot be merged to the seed clustersn are clustered separately using the Markov Clustering approach.

3.5.5 Identification of the Representative Criteria for each Cluster

The final step of this clustering phase is to identify the representative criteria for each cluster. This is ideally a member of the cluster which has the maximum average symmentric pairwise scores with the other members of the cluster. In this process also we consider only the ontological and LOOKUP of each criteria obtained from the pre-processing phase.This representative criteria is ideally used in the frontend to depict each individual cluster.

3.6 Criteria Association

In this phase we identify the criteria association at cluster level. This process uses a apriori based algorithm to find the association between the clusters. The goal is to identify how closely does criteria members of cluster A and Cluster B occurs in the study set of the target disease. So we basically did a mapping of the members of the cluster with the study fromwhich the criteria set is extracted. Then a associative rule mining is carried out using WEKA tool. The best closest cluster of the target cluster is identified by the high LIFT score of the rules obtained. The LIFT threshold was set to 1.1 as below that the two clusters are considered independent. Following figure below is the sample associative rules obtained for a cluster set.

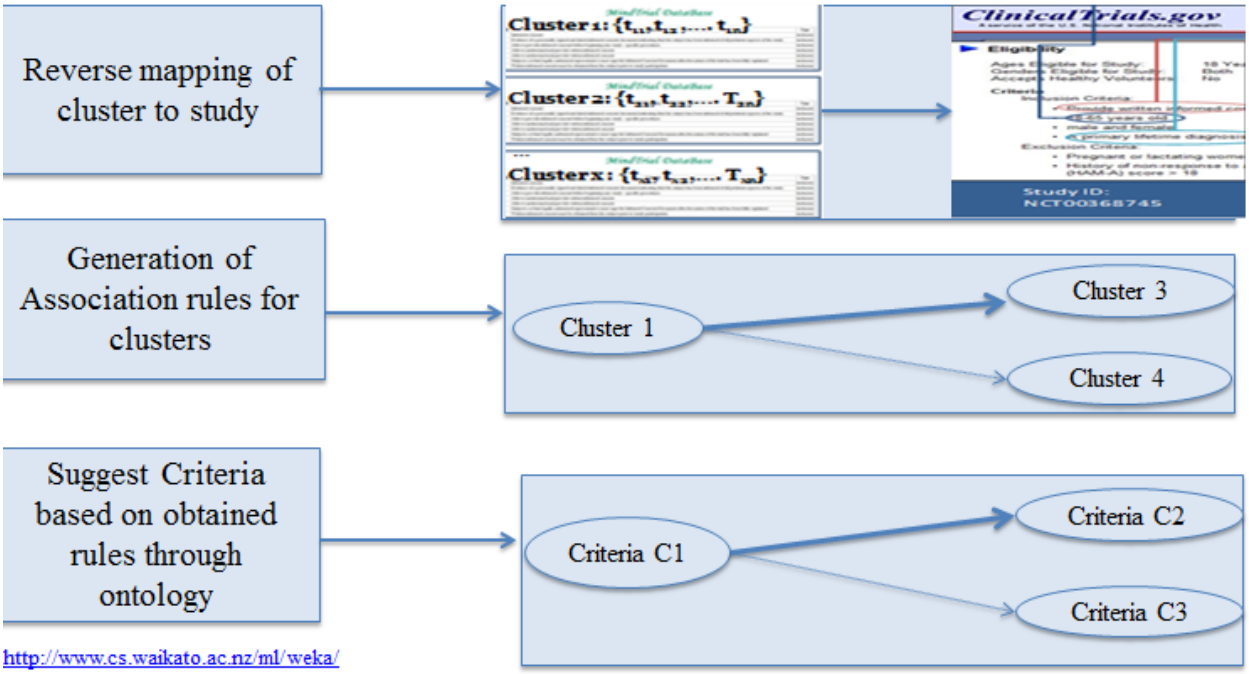


Figure 6: Sample Associative Rules Generated

CHAPTER 4

CREATION OF MOCK CLINICAL TRIAL SUBJECT DATABASE

4.1 Introduction

This component is to develop a mock database containing subject medical information. The database structure was designed to capture all information reuired to answer the major eligibility criteria set of a study. The Subject database was developed on a MS SQL server and mock subject information was generated based on the criteria of the study.

4.2 Database Model

The subject database is an important component of the model which is ideally used to map criteria set to subject medical information therby identifying subjects for a study.Below is the overall normalized structure of the subject database. We can observe from the figure that each subject information table is connected to the mt_user_profile table via user_id column which acts as a primary key for each subjects.

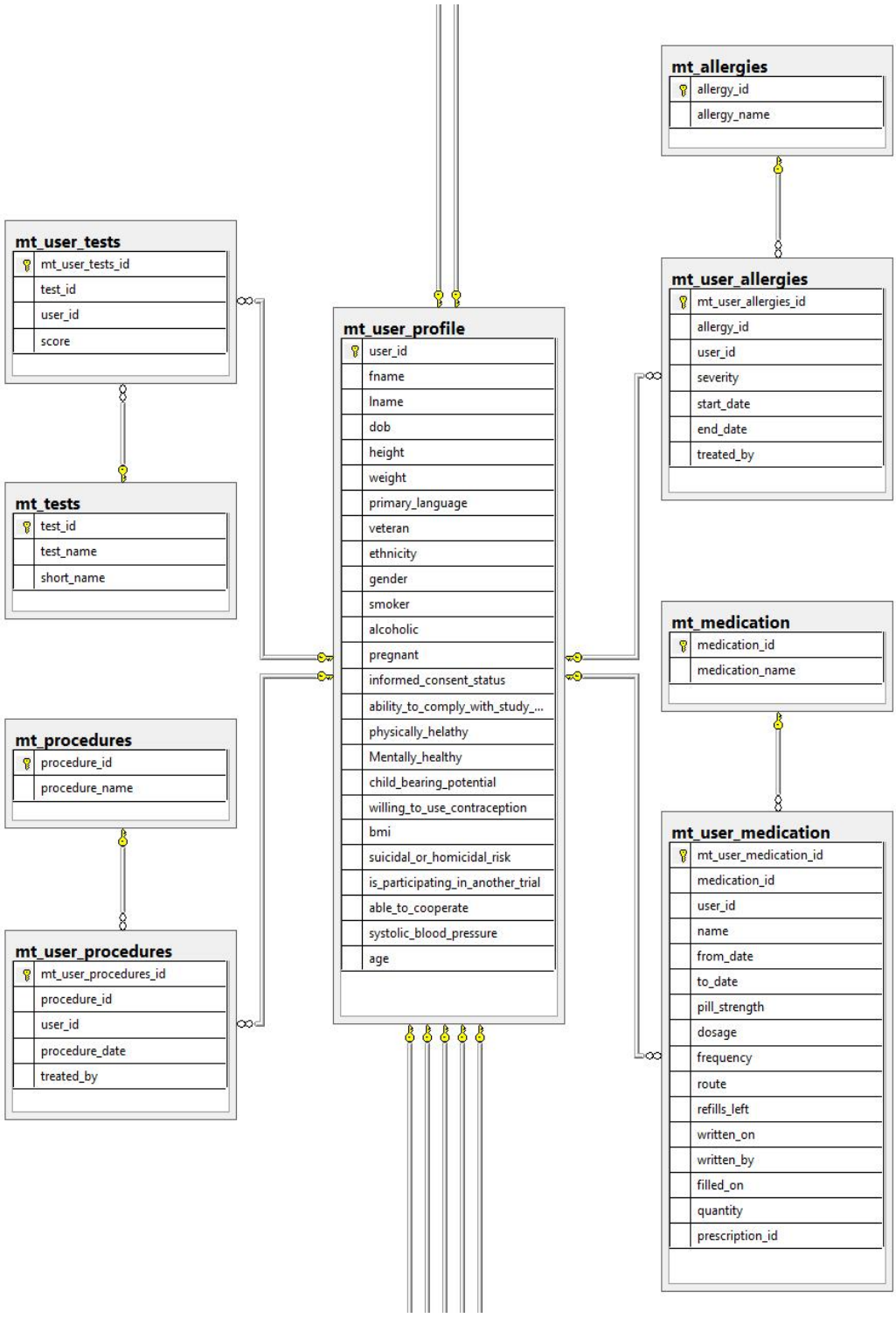


Figure 7: Database Design of Subject Medical Information

4.3 Mock DB Generation

In this phase we describe how mock subjects are generated dynamically based on the criteria set. The subject's general information is obtained by using the FAKEDB generator. This is an online service to extract mock person information like name, gender, age, address, height and weight. The medical information for the subject table is extracted by taking the studies inclusion and exclusion criteria set.First the inclusion set is taken and converted them to insertion queries assuming one subject for each study completing satisfying the criteria set. Then randomly converted 90% of inclusion criteria set of each study into subjects, followed by 80% and so on. However subjects created by this model might satisfy more than one study as the criteria set of the stuies overalap.The same procedure is undergone for exclusion criterion, therby framing subjects satisying some inclusion and exclusion eligibility criteria of studies. The entire process is carried out by an API which takes in the study criteria set as input and generates subject insertion queries as output. Finally the insertion queries are run on the SQL server to get the MOCK subject information onto the database.

CHAPTER 5

ONTOLOGY CREATION FOR CLINICAL TRIALS

5.1 Introduction

This component is used to construct ontology concepts for the clusters obtained from the previous phase.The main goal of the ontological development is to capture the semantic meaning of criteria and map it to the subject medical information The ontology involves a direct mapping to ontological TF-IDF concepts of clusters to the subject database tabels, therby enabling the development of semantic SQL queries to extract subject information for the target studies.

Below figure shows the higher level design of criteria to subject mapping.

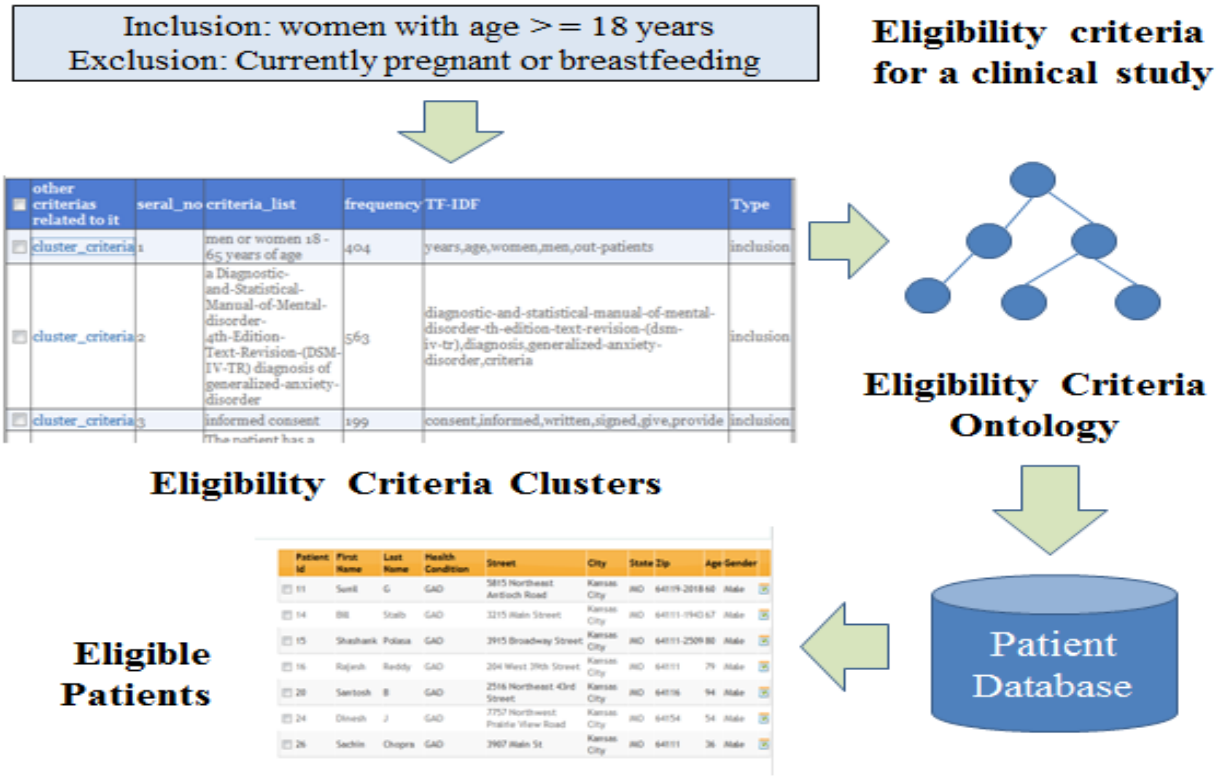


Figure 8: Subject MatchMaking Based on Criteria

All these concept creation was developed by an API which in turn calls JENA a java based ontology API toaccess the ontology. The Ontology visualization is done using Protégé editor to load classes and visualize them. We have used the inbuild modules to create concepts, and properties. We have used SPARQL which is similar to SQL query to query the ontological concepts via JENA API.

5.2 Ontology Creation

In this phase we describe the development of criteria ontology using the concepts from the criteria clusters obtained from the previous phase The developed ontology hierarchy is basicallly a terminological schema formed by mapping the top ranked TF-IDF terms of the clusters to the terminology hierarchy of SNOMEDCT and MESH. As the TF-IDF terms of clusters are mostly ontological terms, they were easily mapped to the ontologies. However some of the lookup terms developed by us which were a part of the top ranked TF-IDF terms are manually verified and plugged into the hierarchy developed by us. These concepts are created as child concepts under the corresponding database table and column names. The reason for this mapping is for the development of dynamic subject SQL queries for the targeted criteria. These parent table name concepts are indeed mapped to clusters via ontological properties. Below is the short description of the ontological model.

The cluster owl file constructed has 3 main parent concepts *inclusion_cluster, exclusion_cluster* and *database_tables.* The inclusion_cluster and exclusion_cluster will inturn have the clusters as subconcepts. For convinience purpose we have denotes the subconcepts for clusters by their cluster ids.Similar design process is carried out for exclusion clusters also.

The database_tables parent class consists of the names of all patient tables as subclasses. These subject tables in turn has the corresponding column names as subclasses. Below is the sample figure depicting the concepts developed in protege.

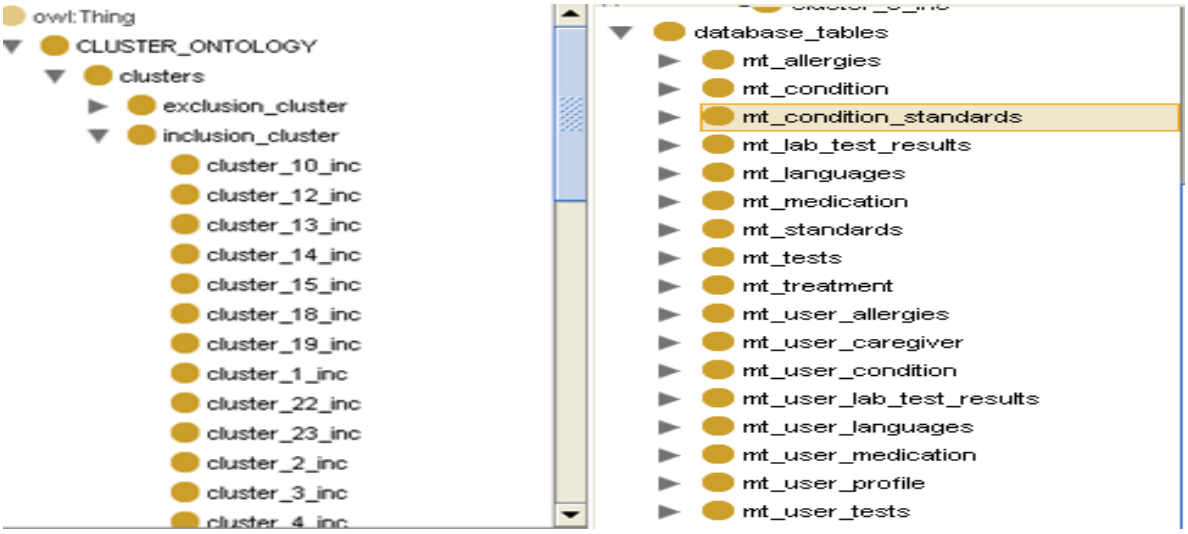


Figure 9: Criteria Cluster Ontology Concept Creation

We then identified the table cluster , apping and connected them via ontological properties. This is ideally build to integrate criteria cluster concepts to subject medical information. Below is a screenshot of a sample connection by an ontological property.

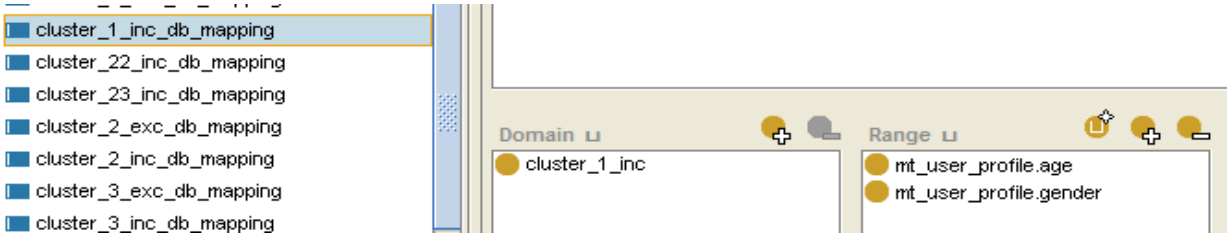


Figure 10: Criteria Cluster Ontology Concept Mapping

Then the cluster criteria top ranked TF-IDF concpts are mapped to the ontological hierarchy and developed as subconcepts.Following is an example where an cluster TF-IDF is mapped to a SNOMED CT hierarchy and plugged into the ontology as subconcepts of the

affecting tables.Below is the screenshot of how the TF-IDf concepts are converted to ontological classes.

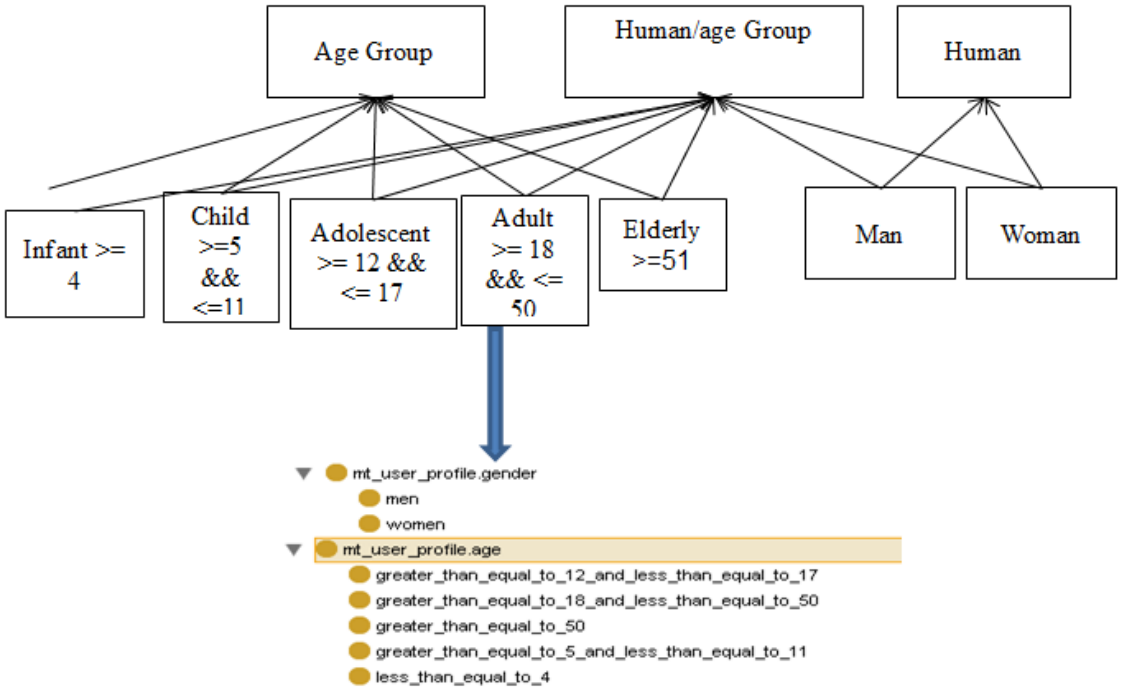


Figure 11: Demonstration of Cluster Concepts Mapped to SNOMEDCT Concept Hierarchy

5.3 Query Generation using Cluster Ontology

In this phase we basically tried to convert the targeted criteria into SQL queries by using Cluster Ontology. The first step of query generation is to identify, the cluster of the target criteria. This can either be picked up from the cluster information obtained from the previous phase or follow a model based clustering approach to determine the cluster ids (This approach is discussed in Section 3.5.3). Then the query is formed based on the cluster ids. That concept for the cluster id in our cluster ontology is identified then the corresponding database table and columns are identified by the property associated with the targeted cluster concept. Then a parsing process is undergone to construct the queries with their sub concepts and dynamic value

from the criteria. This criterion to query conversion is taken care by a java web service which takes in the selected criteria along with the mapped cluster id as input and frames the corresponding SQL query using knowledge base concepts.

The example below shows how a criterion is converted to a query

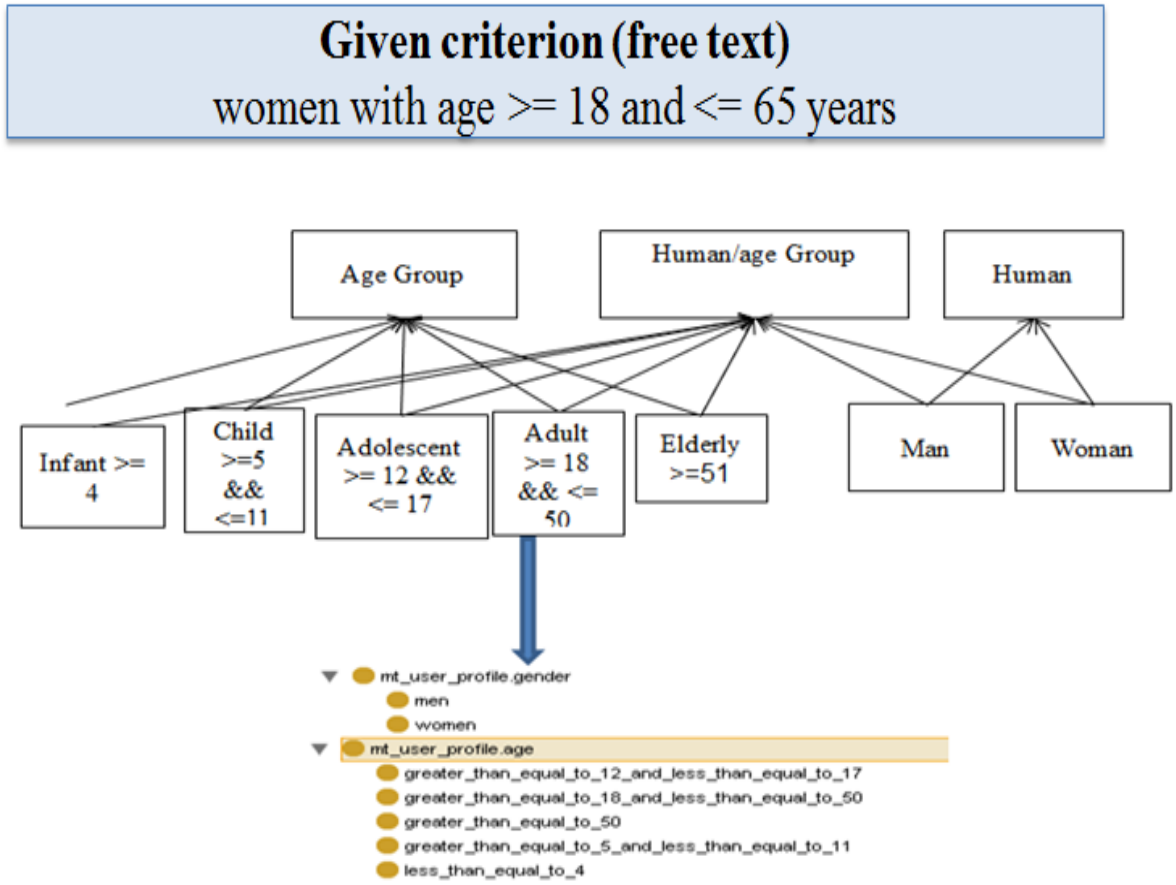


Figure 12: Exact SQL Query Formation using Criteria Ontological Concepts

5.4 Relaxed Query Formation

In this phase we describe the formation of dynamic Relaxed SQL queries using the knowledgebase developed for the clusters. The aim of this approach is to identify subjects with the eligibility criteria relaxed technique so that we could find more number of subjects for a

study. However a manual judgment by the researchers is required to decide whether the subjects identified by the relaxing technique are actually eligible for the study.

Relaxed queries are formed by relaxing the concepts to the subsequent higher levels thereby considering the siblings of the target criteria concept. As semantically similar siblings are grouped under a parent concept this technique is logical.

Below is an example taken to show how the relaxed query is formed for criterion involving numbers is explained below. If the criterion is “age >= 18 years” and the recruiter needs 1000 subjects while if we have only 900 subjects with age >= 18 years, on relaxing the query like “age >=15 years” we could find more number of subjects. However we provide this relaxed query results in a separate column as the recruiter can decide whether to include or not include those list of subjects.

Another way of relaxation for non-numerical criterion is done in a dynamic manner by using the ontological structure mentioned above in the exact query approach field. One example of query relaxation using ontological concepts is explained below. The relaxed query goes in a looped manner until it finds at least 50 subjects or when all the sub concepts are gone through relaxation.

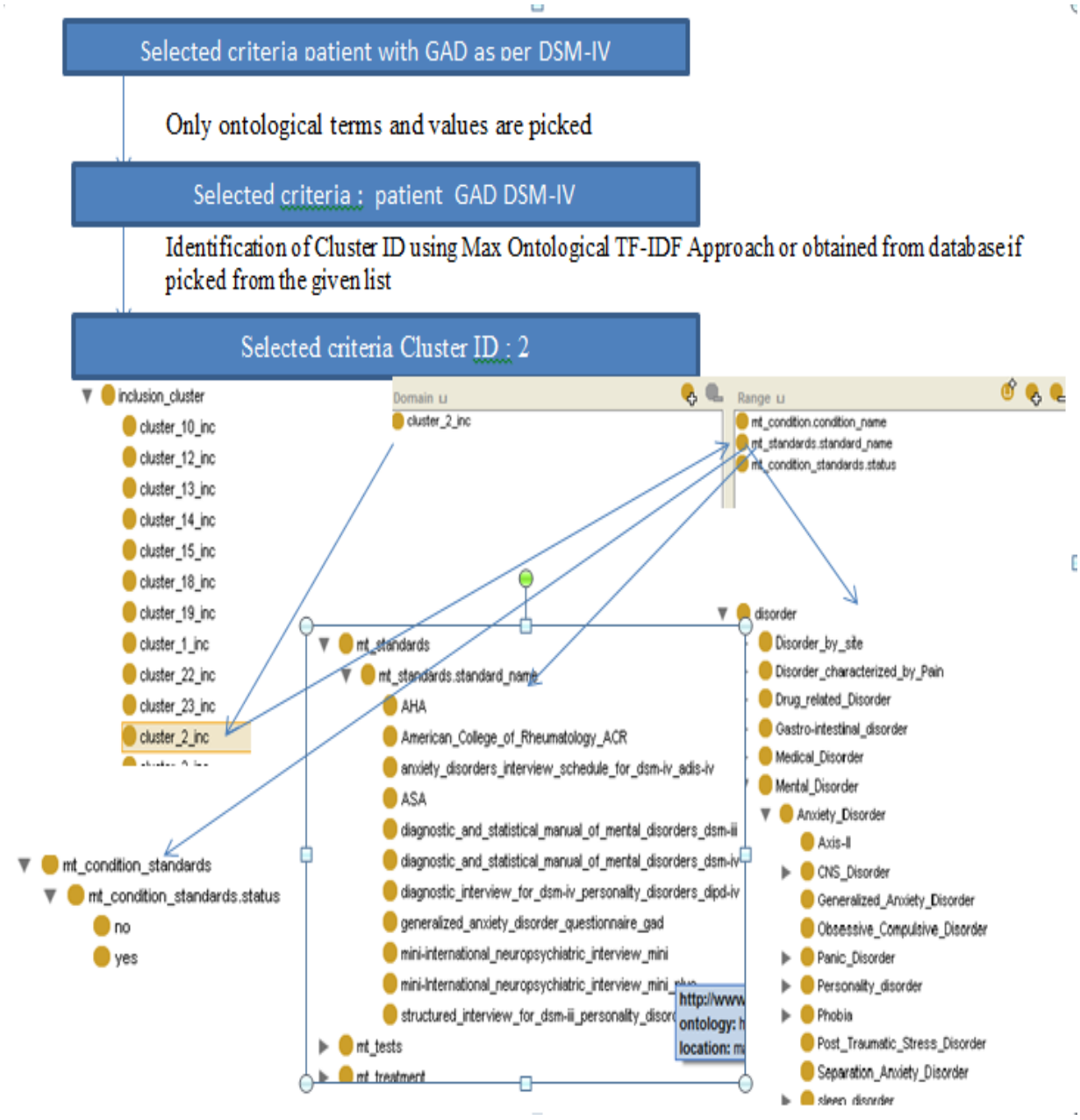


Figure 13: Relaxed Query Concept Mapping using Criteria Ontological Concepts

The relaxed query is formed by considering framing the query to the next higher level of the exact concept based on the ontological hierarchical structure. In the above example the relaxed query is first formed by considering all the guidelines apart from DSM-IV

2) Ontology based Query Relaxation

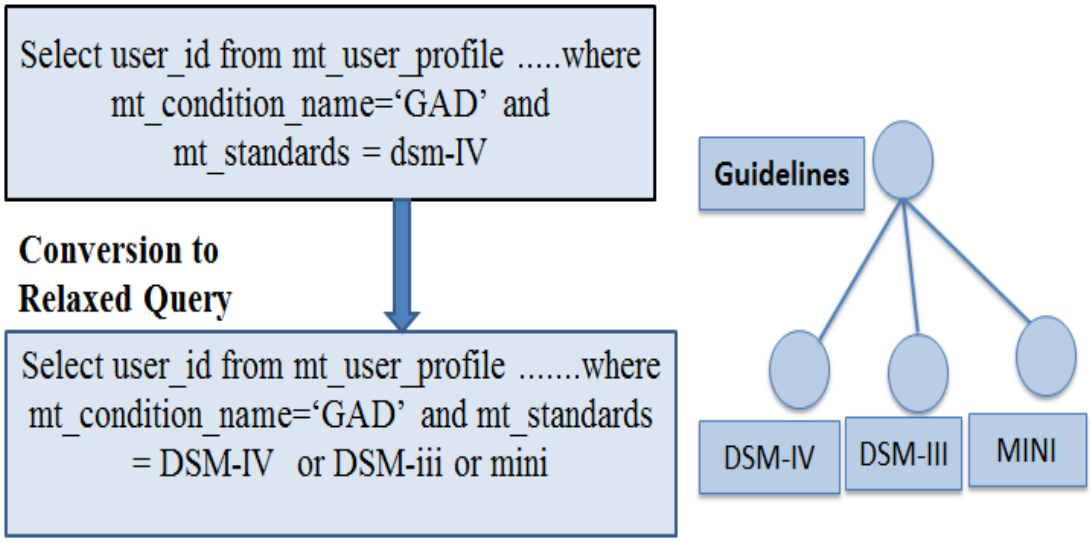


Figure 14: Conversion of Exact Query to Relaxed Query

Thus on relaxation we can pick more number of subjects than the exact query. This is suggested in a separate column in frontend so that the recruiters can select if required.

CHAPTER 6

CASE STUDY ON GAD CLINICAL TRIALS

6.1 Introduction

Recruitment of subjects for clinical trial is an important step in the drug development process. Recruitment of subjects using the traditional approach like face to face interviews or phone interviews is inefficient. In order to take of this issue we have proposed a semi-automated online intelligent model to aid in the recruiting process. We have presented a case study on clinical trials of Generalized Anxiety Disorder to illustrate our data driven model.

6.2 Data Driven Model on GAD Clinical Trials

We downloaded all clinical trials, matching the query term "Generalized Anxiety Disorder". They were on total 708 clinical trial studies as of December 2010. From the clinical trial studies we extracted the eligibility criteria both inclusion and exclusion criteria set using the "Criteria Extraction" component of our model. This API basically does all the parsing and extracts the inclusion and exclusion criteria set of the GAD study list. There were on total of 2395 inclusion criteria and 4871 exclusion criteria which were used for further analysis.

This criteria set are then sent to the pre-processing component of the model where we first split the complex criterion into simple criteria set by using three basic rules (And /or, with, Who). On application of these splitting rules the total number of inclusion criteria obtained was 2477 and the total number of exclusion criteria set was 5000.These criteria set are then sent to the web service which takes care of further pre-processing. The API basically starts with replacing the certain terms with the lookup terms developed by us on analysis of the domain, followed by stop words removal on each criterion, then extracting the only the SNOMEDCT and MESH ontological terms and also the 182 LOOKUP terms that were missed out in the ontology.

By filtering the criterion in this manner we filter out all the noise in the criteria. Finally, stemming of the terms is done to take care of all morphological variations.

Followed by the pre-processing of the criteria we developed the symmetric pairwise scoring matrix of all the criteria set. This pairwise scoring was done for the inclusion criteria set followed by the exclusion criteria set. So on total (n (n+1)/2) pairwise scores were obtained where n = the total number of criteria in inclusion/exclusion set. As mentioned in the model the pairwise score matrix was developed using a 'Symmetric pairwise scoring technique'.

The clustering is from the pairwise scores obtained from the previous phase. As discussed earlier MCL clustering was done with the pairwise scores of the inclusion criteria set. We obtained 126 inclusion clusters were obtained. On analysis of the clusters we found that nearly 56 clusters are singletons and the top most frequent 15 clusters could cover 90% of the inclusion criteria set. Then we tried to merge the 5000 exclusion criteria set to the inclusion clusters by considering only their ontological + Lookup terms and the TF-IDF terms of the cluster with their weight score above 0.08. We picked up this threshold as on analysis for the top 15 most frequent clusters the TF-IDF terms above 0.08 thresholds did a good job in representing the semantic concepts of the cluster. Using the model based clustering we were able to merge about (3091/5000) 63.7% of the exclusion criteria set to the inclusion clusters. The remaining 1909 of the exclusion criteria set are clustered separated by MCL clustering approach and we got 175 exclusion clusters. On analysis of it we found that top most frequent 18 of the exclusion clusters could cover 90% of the exclusion criteria set. The clustering of exclusion criteria set was followed by the identification of top ranked TF-IDF concepts for the newly formed exclusion clusters.

The below table depicts the summary of the clustering result.

Table 2: Overall Summary of Clustering of Clinical Eligibility Criteria

Methods	Criteria Type	Criteria#	Cluster#	Singleton Cluster#	Average Criteria# (Cluster)	Max Criteria# (Cluster)
Phase 1 MCL Semantic Clustering	Inclusion Criteria	2852	126	61	22.63	563
Phase 2 Model-based Clustering (Max Ontological TF-IDF approach)	Exclusion Criteria	3095	46 (80)	3	24.56	1299
Phase 3 MCL Semantic Clustering		1755	175	76	10.02	316

The below histogram depicts the frequency of the inclusion/exclusion clusters in the given study set.

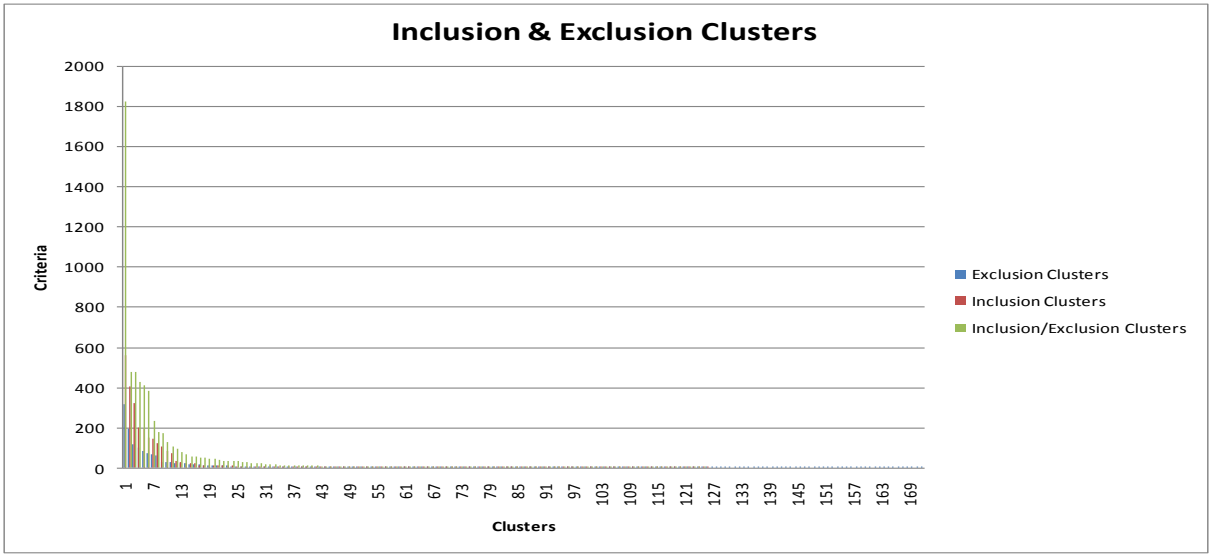


Figure 15: Cluster Frequency

Associative rule mining of GAD clusters are carried out by basically mapping the criteria members to the cluster ids and in turn to the study from which the criteria set id extracted. Associated rule mining is done to identify how many GAD studies do a subset of criteria occur together. Identifying this association helps in suggesting the recruiters while building the criteria.

The ontology for the most frequent 15 inclusion and 18 exclusion clusters were developed using the TF-IDF terms of the clusters which were mostly SNOMEDCT or MESH terms .The lookup terms which are also a part of TF-IDF terms of clusters are manually plugged into the ontology hierarchy. The major categories include demographic information, written consent, substance use history, birth control verification, and previous trial participation. Though the ontology was developed using GAD eligibility criteria the ontology design is done in a more elaborated manner that can suit to most of the mental disorders. The ontology framework is modular as some subsets (such as demographic data and previous trial participation) are common to the majority of clinical trials while others (such as alcohol-specific questions of substance abuse history) may be specific to a smaller number of studies.

Finally the subject MOCK database is developed by using a subset of the GAD criteria. As a proof of concept of this approach we randomly picked 100 studies out of the original 708 studies list. Did reverse engineering by extracting the criteria of those 100 studies and mapping it to the clusters identified by us. With the guidance of knowledge base designed we designed insertion queries for the criteria set. Initially we designed 100 subjects satisfying the inclusion criteria set of 100 studies. That is developing one subject satisfying the study 100%. However one subject can map too many studies as the criteria overlap. Then we developed subjects satisfying 75% of the criteria set, if a study has 4 inclusion criteria our script randomly pick 3 criteria and develops subject information thereby making the subject satisfy 75% of the criteria

set and so on. Similarly we developed subjects satisfying both inclusion and exclusion criteria set some part of inclusion and exclusion set and so on. The names, address, DOB, age are retrieved using FAKEDB generator. Thus we designed 1000 subjects using the eligibility criteria of 100 GAD studies. The application of the developed model is demonstrated using the Mindtrial web interface where a search engine eligibility criteria is developed and subject mapping is done to the selected or developed criteria set (Additional information on the interface is explained in Chapter 7).

CHAPTER 7

WEB INTERFACE FOR GAD ELIGIBILITY CRITERIA

7.1 Introduction

The web interface is mainly designed for the clinical trial recruiters and researchers to develop a list of eligibility criteria for a study and identify subjects mapping their criteria. The web application is built using ASP.NET 3.5 technologies. The frontend internally calls JAVA web services. The web services are used for reusability and interoperability. In order to store information to and from the web interface we have used MySQL 2009 and Microsoft SQL server express 2010 databases as our backend. We used C# as our code behind programming language which acts as a controller for our web front end.

The web interface is primarily designed taking GAD as our case study. As mentioned in the table of contents the web interface involves 5 different phases. A detailed overview of each of the interface is discussed in the following sessions.

7.2 Web Interface for Criteria Search

7.2.1 Introduction

The initial step is to select the target disease for which the eligibility criteria set is to be developed. We have implemented the interface model for GAD as a case study. Figure 16 shows the screenshot of the interface.

Figure 16: Web Interface for Selecting the Disease

The next step for the recruiter is to select a set of inclusion and exclusion clinical trial for the target study. The subjects recruited for the study should satisfy the criteria selected in order to participate in the study Following figure below is the search page which primarily displays the list of search options for eligibility criteria set.

Figure 17: Web Interface for Search Options

7.2.2 Keyword Based Criteria Search

Figure 18 and Figure 19 shows the interface for keyword based search. In this page we select the criteria based on the keyword entered by the recruiter. The recruiter can either type in the keyword or pick the keyword concepts from the dropdown. This keyword is ideally used search in the representative criteria and in the top ranked TF-IDF list of each cluster. The searched criteria are restricted by the type (i.e. inclusion or exclusion) selected. Appropriate message is prompted when the required fields are not selected.

Figure 18: Web Interface for Keyword Search by Typing in the Key

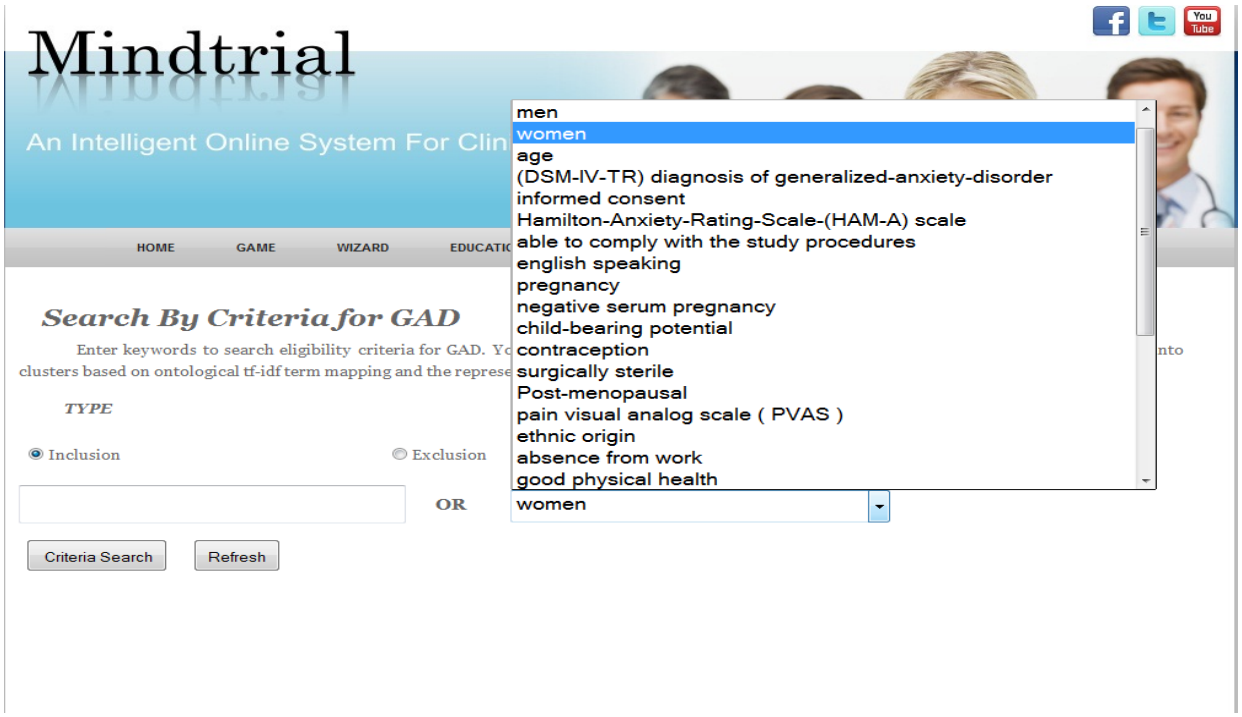


Figure 19: Web Interface for Keyword Search by Picking from Dropdown

Figure 20 is the search result for the keyword selected. All the representative criteria containing the keyword is displayed. In order to view the members of the cluster we have click on the “Member of the cluster” hyperlinked column. Appropriate error message is thrown if there is no search result matching the query.

Figure 20: Web Interface for Keyword Search Results

7.2.3 Cluster ID Based Criteria Search

Figure 21 shows the interface for cluster id based search. In this page we select the criteria based on the cluster id entered by the recruiter. The cluster ids are given dynamically to each cluster based on the frequency of the members when initially formed using MCL. The type is also a mandatory field to restrict the criteria search. Appropriate message is prompted when the required fields are not selected.

Figure 21: Web Interface for Cluster ID Based Search

Figure 22 is the search result for the cluster id selected. In order to view the members of the cluster we have click on the “Member of the cluster” hyperlinked column. Appropriate error message is thrown if there is no search result matching the query.

Figure 22: Web Interface for Cluster ID Search Result

7.2.4 Frequency Based Criteria Search

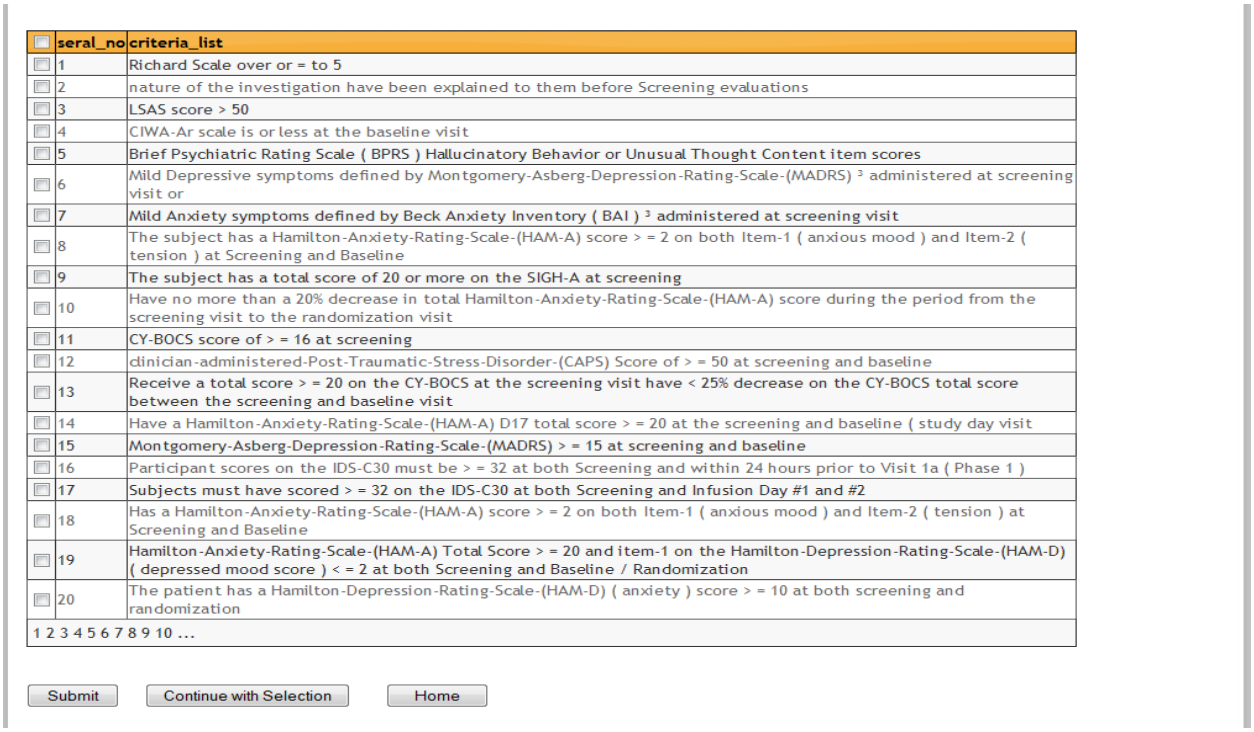
Figure 23 shows the interface for frequency based criteria search. The top most frequent criteria are displayed as search result based on the number and type selected. Appropriate message is prompted when the required fields are not selected.

Figure 23: Web Interface for Frequency Based Search

Figure 24 is the search result for the frequency based search. The results are sorted on descending order of frequency. In order to view the members of the cluster we have click on the "Member of the cluster" hyperlinked column. Appropriate error message is thrown if there is no search result matching the query.

Figure 24: Web Interface for Frequency Based Search Results

Figure 25 shows the sample screenshot of the page displayed when a particular cluster's "members of the cluster" is clicked. This page basically displays all the members corresponding to the selected clusters. The clusters are grouped initially using MCL and the remaining criteria are merged using maximum ontological tf-idf mappping aproach.



	seral_no	criteria_list
☐	1	Richard Scale over or = to 5
☐	2	nature of the investigation have been explained to them before Screening evaluations
☐	3	LSAS score > 50
☐	4	CIWA-Ar scale is or less at the baseline visit
☐	5	Brief Psychiatric Rating Scale (BPRS) Hallucinatory Behavior or Unusual Thought Content item scores
☐	6	Mild Depressive symptoms defined by Montgomery-Asberg-Depression-Rating-Scale-(MADRS) ³ administered at screening visit or
☐	7	Mild Anxiety symptoms defined by Beck Anxiety Inventory (BAI) ³ administered at screening visit
☐	8	The subject has a Hamilton-Anxiety-Rating-Scale-(HAM-A) score > = 2 on both Item-1 (anxious mood) and Item-2 (tension) at Screening and Baseline
☐	9	The subject has a total score of 20 or more on the SIGH-A at screening
☐	10	Have no more than a 20% decrease in total Hamilton-Anxiety-Rating-Scale-(HAM-A) score during the period from the screening visit to the randomization visit
☐	11	CY-BOCS score of > = 16 at screening
☐	12	clinician-administered-Post-Traumatic-Stress-Disorder-(CAPS) Score of > = 50 at screening and baseline
☐	13	Receive a total score > = 20 on the CY-BOCS at the screening visit have < 25% decrease on the CY-BOCS total score between the screening and baseline visit
☐	14	Have a Hamilton-Anxiety-Rating-Scale-(HAM-A) D17 total score > = 20 at the screening and baseline (study day visit
☐	15	Montgomery-Asberg-Depression-Rating-Scale-(MADRS) > = 15 at screening and baseline
☐	16	Participant scores on the IDS-C30 must be > = 32 at both Screening and within 24 hours prior to Visit 1a (Phase 1)
☐	17	Subjects must have scored > = 32 on the IDS-C30 at both Screening and Infusion Day #1 and #2
☐	18	Has a Hamilton-Anxiety-Rating-Scale-(HAM-A) score > = 2 on both Item-1 (anxious mood) and Item-2 (tension) at Screening and Baseline
☐	19	Hamilton-Anxiety-Rating-Scale-(HAM-A) Total Score > = 20 and item-1 on the Hamilton-Depression-Rating-Scale-(HAM-D) (depressed mood score) < = 2 at both Screening and Baseline / Randomization
☐	20	The patient has a Hamilton-Depression-Rating-Scale-(HAM-D) (anxiety) score > = 10 at both screening and randomization

1 2 3 4 5 6 7 8 9 10 ...

Submit | Continue with Selection | Home

Figure 25: Web Interface for Criteria Members Display

7.3 Web Interface for Criteria Selection

7.3.1 Introduction

In this phase we select or build the criteria required for the targeted study. Before identifying the subjects satisfying the criteria the selected set is once confirmed with the recruiter. We primarily provide two ways of criteria selection. Either selecting from the existing criteria set or building own criteria set using auto-suggestive toolbox.

Apart from this the recruiter is given option to see all the set of inclusion and exclusion criteria for the target disorder existing in the database and select from the list provided. Figure 26 is the interface which shows all the list of criteria both for inclusion and exclusion.

Inclusion set

Members of the Cluster	ID	Representative Criteria	frequency	TF-IDF	Type
cluster_criteria	1	men or women 18 - 65 years of age	404	years age women men out-patients	inclusion
cluster_criteria	2	a Diagnostic-and-Statistical-Manual-of-Mental-disorder-4th-Edition-Text-Revision-(DSM-IV-TR) diagnosis of generalized-anxiety-disorder	563	diagnostic-and-statistical-manual-of-mental-disorder-th-edition-text-revision-(dsm-iv-tr) diagnosis generalized-anxiety-disorder criteria	inclusion
cluster_criteria	3	informed consent	199	consent informed written signed give provide	inclusion
cluster_criteria	4	The patient has a total score of at least on the Hamilton-Anxiety-Rating-Scale-(HAM-A) scale	326	score total hamilton-anxiety-rating-scale-(ham-a) clinical-global-impression-severity	inclusion
cluster_criteria	5	Patients must be willing and able to comply with the study procedures	145	study comply	inclusion
cluster_criteria	6	negative serum pregnancy test for women of child-bearing potential	107	negative pregnancy women child-bearing urine serum	inclusion
cluster_criteria	7	English speaking	72	english speaking language native	inclusion
cluster_criteria	8	Women must be of non child-bearing potential	126	contraception women reliable child-bearing method	inclusion
cluster_criteria	9	have a score of on the pain visual analog scale (PVAS) score at screening	86	pain	inclusion
cluster_criteria	10	any ethnic origin	5	origin ethnicity races race gender without will	inclusion

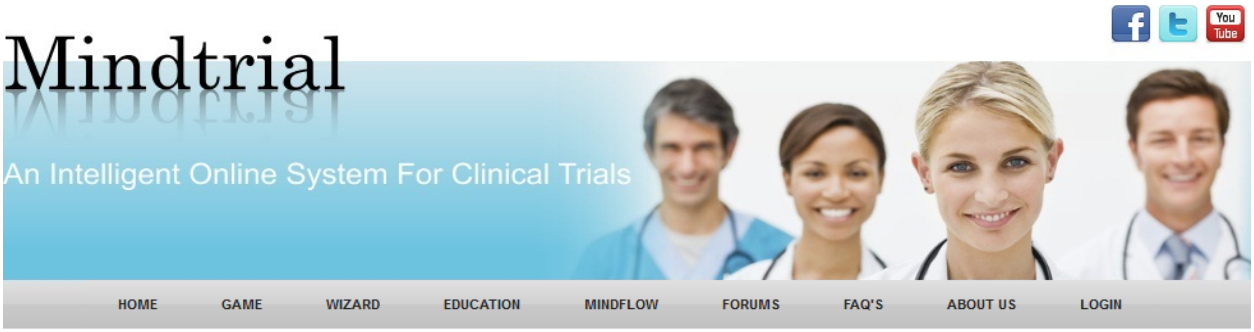
Exclusion set

Members of the Cluster	ID	Representative Criteria	frequency	TF-IDF	Type
cluster_criteria	1	Patients with severe depression	71	study severe illness	exclusion
cluster_criteria	2	Treatment with a monoamine oxidase inhibitor tricyclic SSRI antidepressant (with the exception of fluoxetine) or lithium within 2 weeks prior to beginning study medication	316	treatment study medication	exclusion
cluster_criteria	3	Participation in any clinical trial 30 days prior to entering the study	100	study clinical drug	exclusion
cluster_criteria	4	Current suicidal or homicidal risk	197	suicidal suicide risk ideation serious current significant intent active	exclusion
cluster_criteria	5	pulmonary disease	88	disease significant	exclusion
cluster_criteria	6	Has received electroconvulsive therapy within 6 months prior to Screening	76	therapy screening	exclusion
cluster_criteria	7	Known history of intolerance or hypersensitivity to pioglitazone	120	hypersensitivity allergy known reaction sensitivity paroxetine intolerance quetiapine	exclusion
cluster_criteria	8	Known urethral stricture	2	urethral known reconstruction stricture	exclusion
cluster_criteria	9	Known pelvic malignancy	2	malignancy known suspected pelvic	exclusion
cluster_criteria	10	Other protocol-defined inclusion/exclusion criteria may apply	28	criteria apply following specific common diabetes-mellitus includes every protocol	exclusion
cluster_criteria	11	psychotic symptoms	65	psychosis psychotic symptoms current illness mania psychiatric active	exclusion
cluster_criteria	12	Patients unable to cooperate	22	unable child regular intervention questionnaires physical scales unwilling parents complete benzodiazepines signs follow	exclusion
cluster_criteria	13	Patients with uncontrolled narrow-angle glaucoma	25	glaucoma angle narrow uncontrolled acute retention	exclusion

Figure 26: Display of Entire Criteria Set

7.3.2 Selection from the Existing Criteria Set

Once the list of criteria is displayed, the recruiter can start picking the criteria from the list displayed. The recruiter is given freedom to pick either from the representative criteria of a cluster or go forth the members of the cluster to select the criteria set wanted for the targeted study. Figure 27 is the sample screenshot showing the way criteria is selected.



The representative criteria of each cluster are displayed below. Please click on the "Member of the cluster column" to view all members of the cluster

	Members of the Cluster	ID	Representative Criteria	frequency	TF-IDF	Type
☑	cluster_criteria	1	men or women 18 - 65 years of age	404	years age women men out-patients	inclusion
☑	cluster_criteria	6	negative serum pregnancy test for women of child-bearing potential	107	negative pregnancy women child-bearing urine serum	inclusion
☑	cluster_criteria	8	Women must be of non child-bearing potential	126	contraception women reliable child-bearing method	inclusion
☐	cluster_criteria	23	Women with a history of Stage I II or III breast cancer for at least months	36	cancer breast	inclusion

Submit | Create your own | show all | Home

Figure 27: Sample Criteria Set Selection

7.3.3 Autosuggestion Based Criteria Development

The recruiter is also provided an option to build his own criteria. However we help in building criteria by autosuggestion. The user receives suggestions for every third letter he types in. The suggestions are based on the top ranked TF-IDF terms of each cluster. Figure 28 shows the autosuggestion interface

Figure 28: Build your own Criteria Interface

Once the criterion is typed and submit button is clicked the stop-words typed are automatically excluded. The stop-words like ‘and, is, or, the’ are ideally considered as noise and automatically excluded. Then the remaining are split into individual terms and mapped to SNOMED and MESH ontological terms. The reason to do this is to capture variations and synonyms. Then the max ontological TF-IDF merging approach is used to identify the cluster, corresponding to the entered the criteria. In other words the obtained ontological term of the entered criteria is compared with the identified TF-IDF term list of each cluster. Finally it is merged to the cluster where the target criteria’s ontological term has the maximum subset of the TF-IDF terms of the cluster. If we were not able to plugin to any of the existing cluster we add it as a new singleton cluster for future reference.

Figure 29: Mapping of Criteria to the Cluster

7.3.4 Web Interface for Criteria Filtering

This page is a comprehensive page showing the list of criteria selected by the recruiter using the above mentioned search options. Irrespective of what search options we use it is finally redirected to this confirmation page where we again provide an option of filtering the criteria. Figure 30 is the sample screenshot of the interface mentioned.

criteria_selected	ID	Type
men or women 18 - 65 years of age	1	inclusion
informed consent	3	inclusion
The patient has a total score of at least on the Hamilton-Anxiety-Rating-Scale-(HAM-A) scale	4	inclusion

Figure 30: Comprehensive List of Selected Criteria Information

7.4 Web Interface for Criteria Association

7.4.1 Introduction

Associative rule mining technique is used to suggest related criteria to the recruiter selected ones. It is primarily used to aid the recruiter in developing the eligibility criteria set for a study. The suggestions are given based on Lift and Confidence association scores of the clusters corresponding to the criteria.

7.4.2 Web Interface for Associated Criteria Suggestion

Associativity rules are calculated at the cluster level and implied to its members. As each criteria is associated with a cluster id the associated cluster is found using the mining rules and the representative criteria of the corresponding cluster is suggested. We set a cut off to the suggested list by providing only the top 3 closely associated criteria. The closeness is identified by high lift scores and reasonable confidence. We decided to use LIFT score as a threshold

because we wanted to focus on how much interest the association is rather than frequency. Figure 31 shows the screenshot of the associated rules for the selected criteria set.

Figure 31: Comprehensive Selected List with Associated Criteria Information

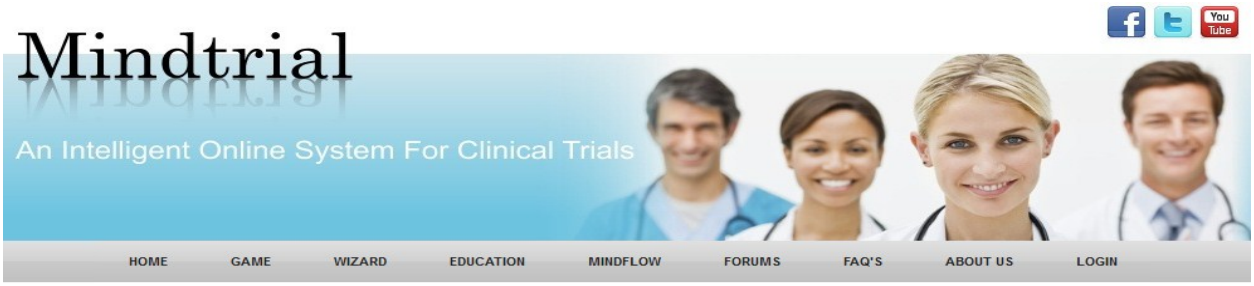
7.5 Web Interface for Subject Information

7.5.1 Introduction

This step is to identify subjects matching the criteria set selected. The subject information is retrieved from the MSSQL database which ideally is created dynamically using the MOCKDB approach explained below. Basically we have created a set of mock subjects who either completely or partially satisfies the criteria set of GAD. There are two ways of querying subjects for each criterion. Exact query match and Relaxed query match approach.

7.5.2 Web Interface for Criteria- Subject Mapping

In this interface Figure 32 we could see the number of subjects satisfying the criteria exactly and the number of subjects satisfying the criteria on query relaxation. The technique of how the exact and relaxed query is formed is explained in Section 5.4



Serial No	Selected Criteria	Number Of Patients	Relaxed query
1	men women > = 18 and <= 65 years age	697	1443
2	informed consent	75	75
3	all criteria satisfied	58	60

Figure 32: List of Criteria Selected and the Number of Subjects Satisfying the Criteria

7.5.3 Web Interface for Subject Distribution Information

The subject information can be viewed by clicking on the 'number of subjects hyperlink set either in the exact query column or the relaxed query column. This in turn has the 3 different interfaces to view the subject information which is shown below.

7.5.3.1 Subject Information – Detail View

The first interface is the "Detail view" where the subject basic information is provided in a tabular format. Further detailed information like address can be viewed on clicking the 'View

Detail' hyperlink. Figure 33 and Figure 34 show the sample screenshot of it. We also have a dynamic Location, Age and Gender filter which can be used to filter according to the study requirements.

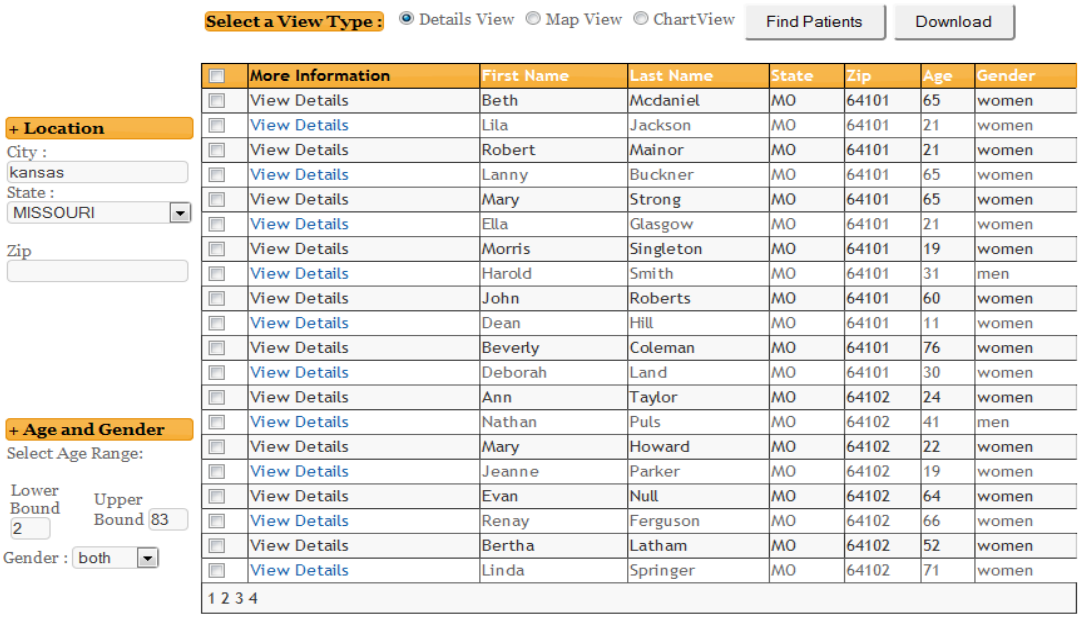


Figure 33: Subject Information in Details View Interface

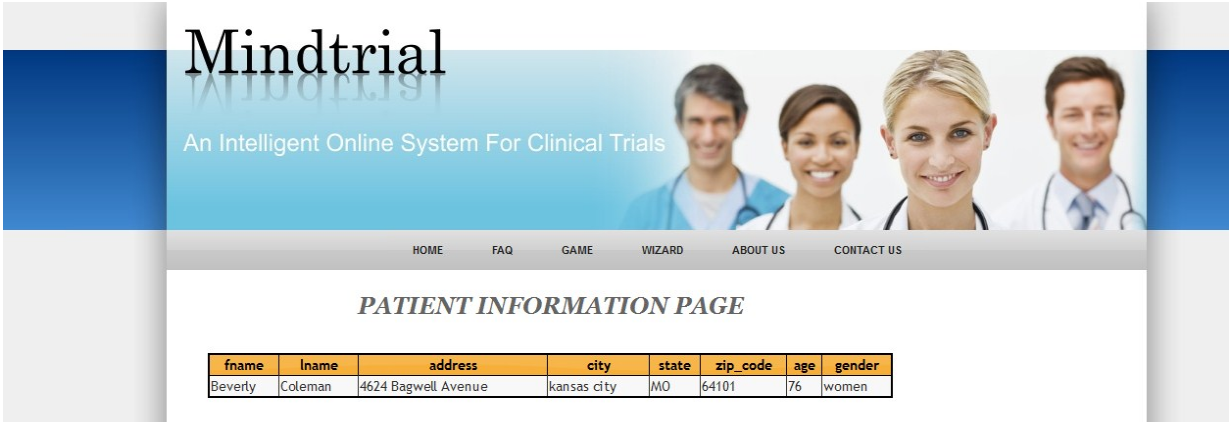
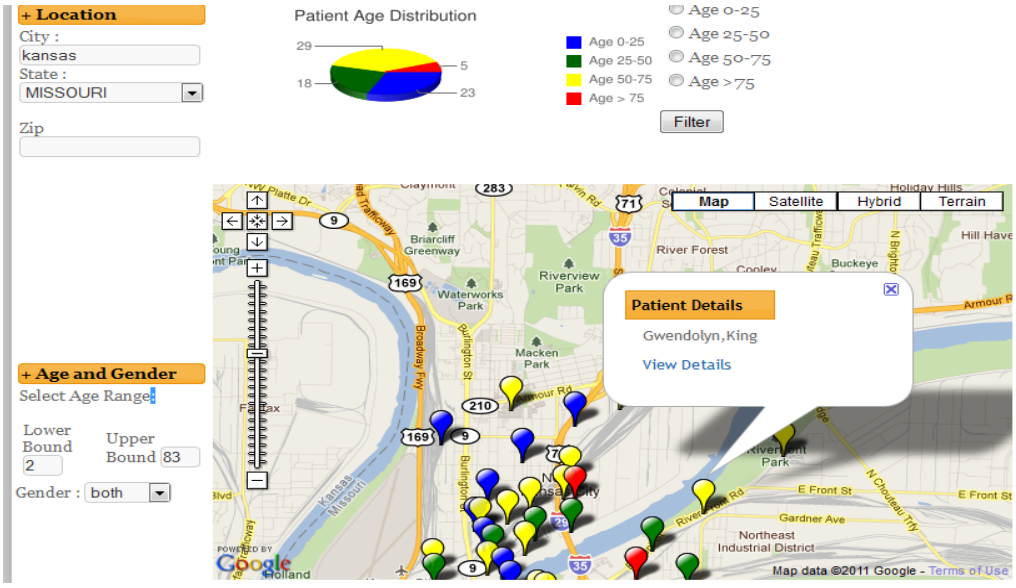


Figure 34: Subject Address Information

7.5.3.2 Subject Information – Location View

The second interface id the ‘Map view’ where the subject distribution information is depicted using Google maps API. On clicking the key of the Google maps the subject detailed information can be seen. We also provide age distribution chart which can be filtered. The filter in turn affects the Google maps also. Figure 35 is the sample screenshot of the map view



.

Figure 35: Map View of Subjects

7.5.3.3 Subject Information – Chart View

The third one is the chart interface where the subject distribution using a pie and a bar chart. The charts can be modified by the filter options provided. By default the age is selected as the first distribution and gender is selected as the second distribution. Figure 36 and Figure 37 is the sample screenshot of the chart view

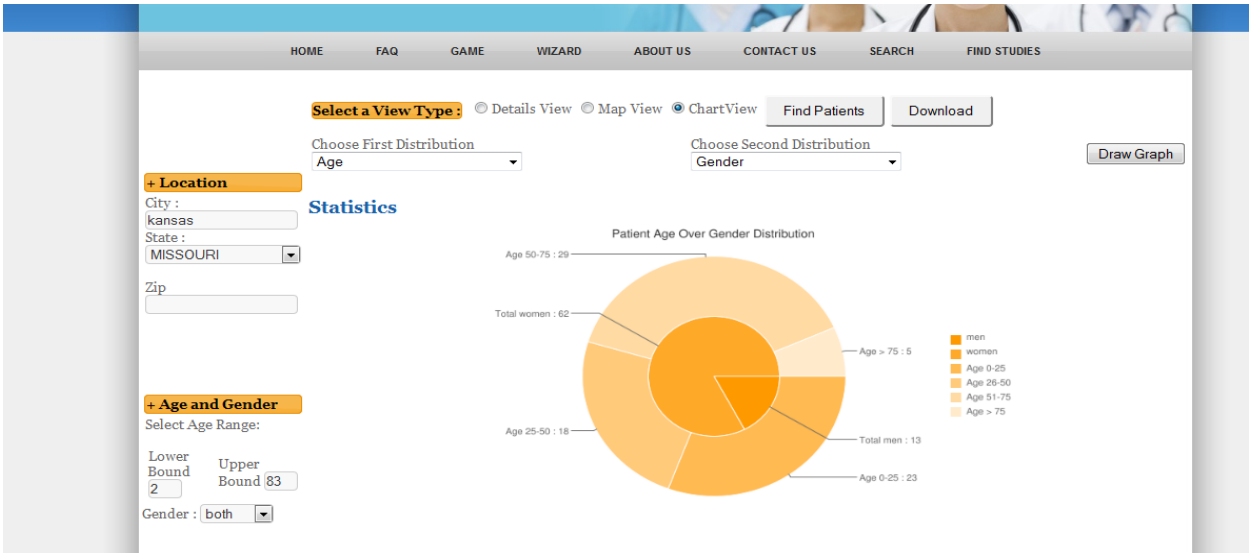


Figure 36: Subject Information Pie Chart

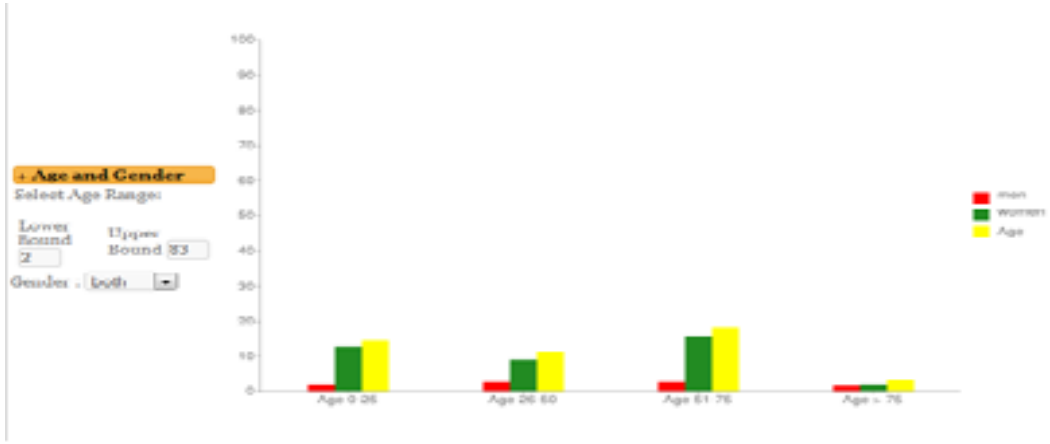


Figure 37: Subject Information Bar Graph

7.5.4 Web Interface for Subject Filtering

Finally once the subjects are selected by the checkbox selected, the detailed information can be downloaded in the csv format for further process. The downloaded files contain the detailed information of the selected patients including address, phone number, medical disorder etc. which are used for further processing.

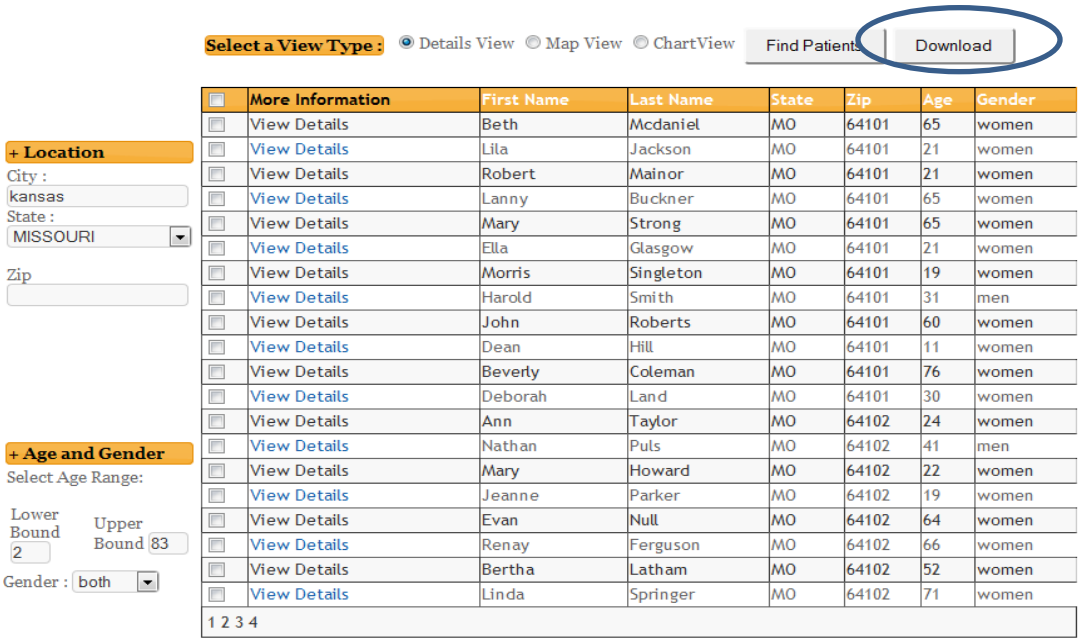


Figure 38: Subject Information Download Option

CHAPTER 8

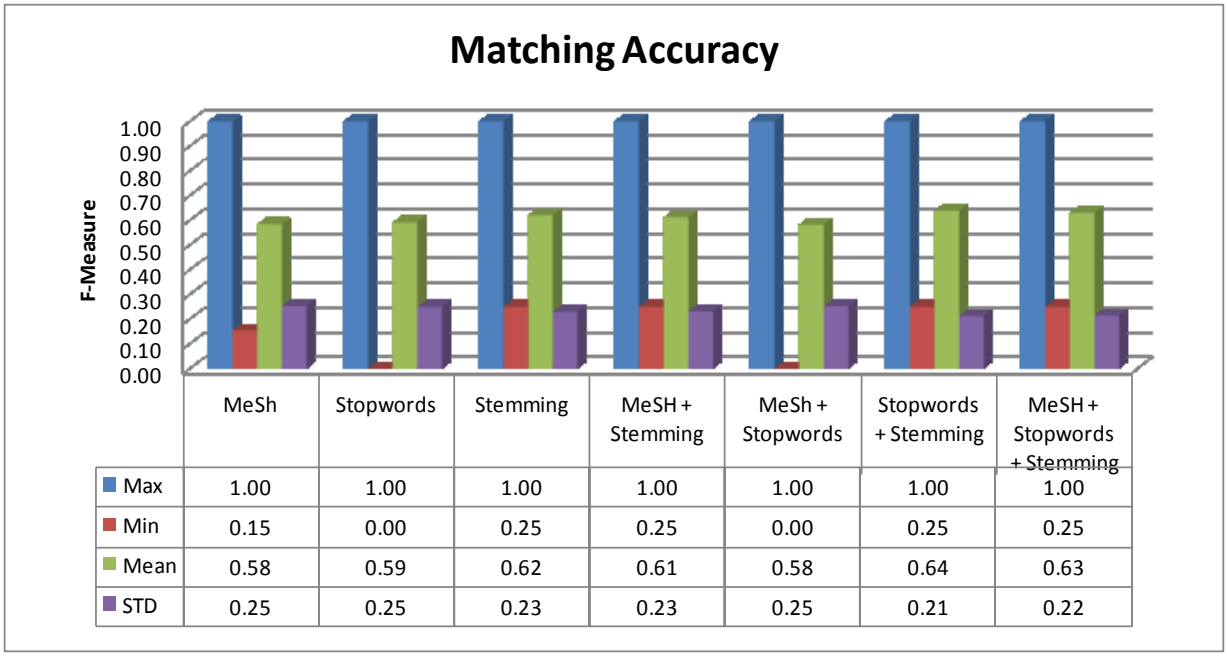
. VALIDATION

8.1 Introduction

This chapter plays an important role in this thesis. We have demonstrated below the validation of the model at each phase and its corresponding results. The validation results give us a clear picture to defend the reason to choose the model. Section 8.2 describes the comparison of our criteria pre-processing model with 8 different models tested. We took a sample test criteria set from GAD and performed the analysis. Section 8.3 describes the optimal pairwise threshold taken for clustering and the best inflation factor selected for MCL clustering using a small set of clusters with 1 to 4 criteria per cluster. Section 8.4 describes the overall cluster evaluation using F-measure an external cluster validation technique and Silhouette width an internal cluster validation technique. Section 8.5 describes the validation of the merging of criteria to cluster using the model based clustering technique.

8.2 Optimal Inter-Criteria Similarity Metric

In this section we compared 8 different models to identify the best pairwise scoring strategy. This is identified as the model that gives the maximum pairwise score. In order to validate the models we selected 60 semantically similar pairwise criteria with the lexical similarity score ranging from 0.1 to 1.0. All 8 models were applied and the best model was identified as the one which gave the maximum average pairwise scores and overall all pairs scores moving towards 1. On analysis of this approach we found that Stop words removal+Stemming+Ontological term model proved to be the best. The below histogram shows the mean scores and standard deviation of all the 8 models.



	MeSh	Stopwords	Stemming	MeSH + Stemming	MeSh + Stopwords	Stopwords + Stemming	MeSH + Stopwords + Stemming
Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Min	0.15	0.00	0.25	0.25	0.00	0.25	0.25
Mean	0.58	0.59	0.62	0.61	0.58	0.64	0.63
STD	0.25	0.25	0.23	0.23	0.25	0.21	0.22

Figure 39: Comparison of Scoring Metrics

The graph below shows the depiction of how the entire score set move towards 1 for the best model which is 8.

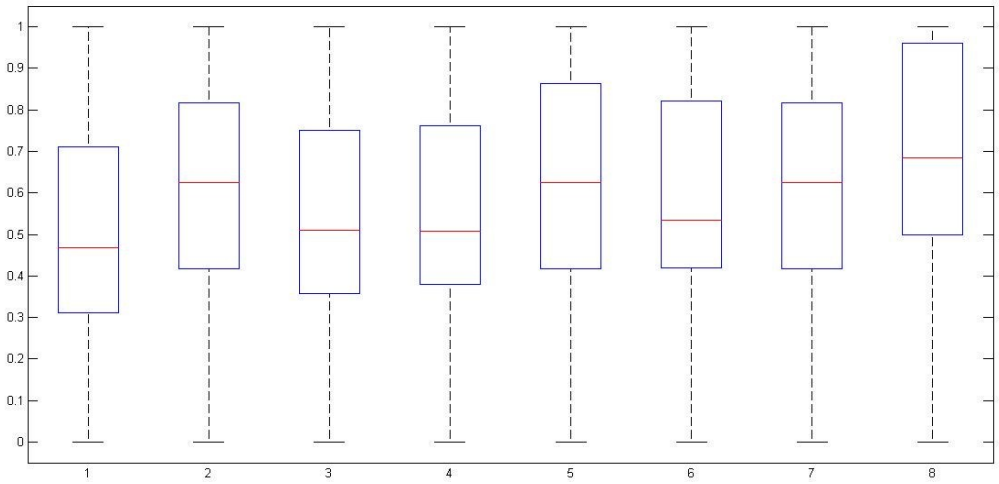


Figure 40: Box Plot of all 8 Model Scoring Metrics

Comparative histograms of the distribution of pair-wise similarity scores based only on lexical matches versus combining all approaches are shown in Figure 41. The latter is skewed towards a higher value and has lower variance.

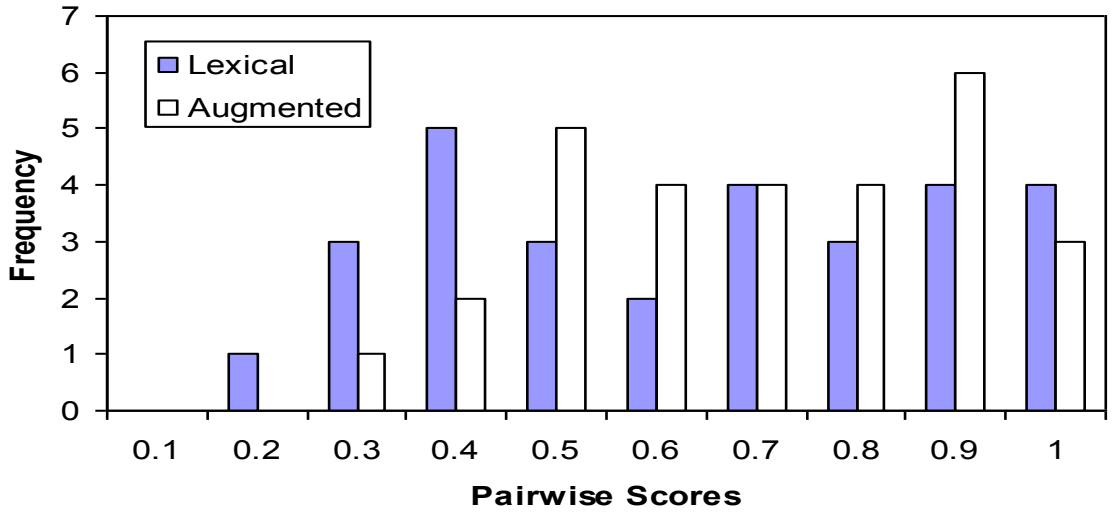


Figure 41: Comparison of Lexical and Semantic Model

8.3 Optimal Pairwise Score and Inflation Factor Computation

This section is used to determine the best cut off for the semantic scoring strategy. In this case a second dataset with 60 clusters each cluster containing 1 to 4 criteria and the pairwise score ranging from 0.1 to 1 was taken. All the criteria are put together and then clustered again to check whether spurious clusters are formed. This validation was done by comparing the F-measure (2*Precision*Recall/ [Precision+Recall]) while varying the pairwise score cut-off (threshold) to define membership in a cluster.

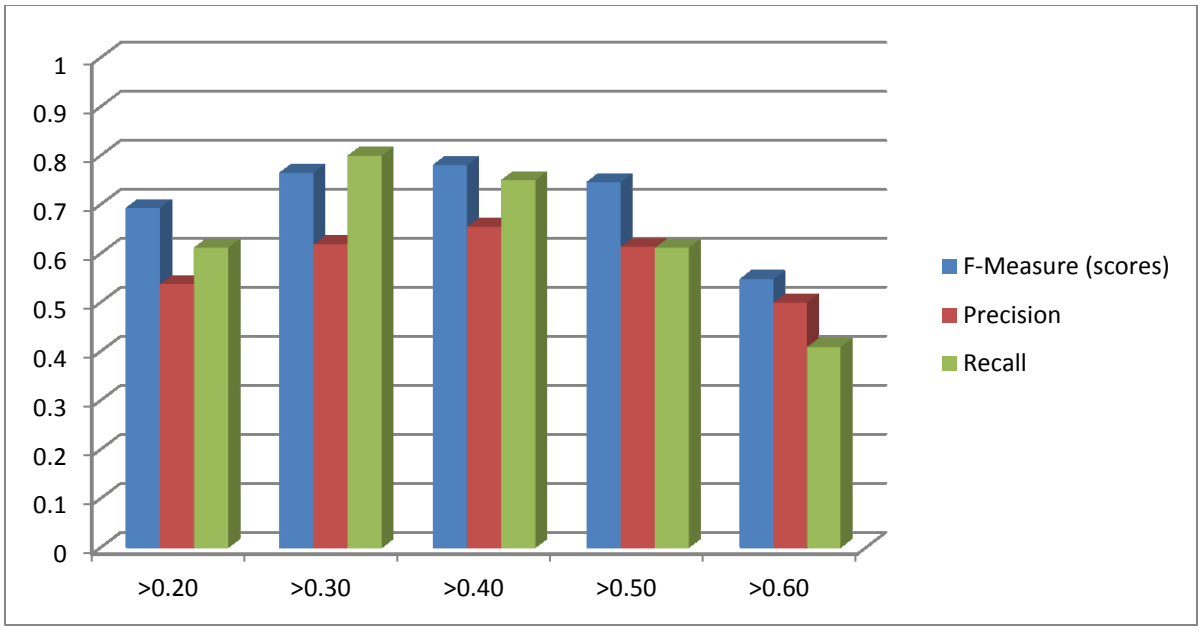


Figure 42: Comparison Chart to Select the Optimal Threshold

Similarly the MCL algorithm includes a parameter called the inflation factor that determines the granularity of clustering; higher the value of the inflation factor, larger the number of clusters. The F-measure was used to determine the best value for this parameter.

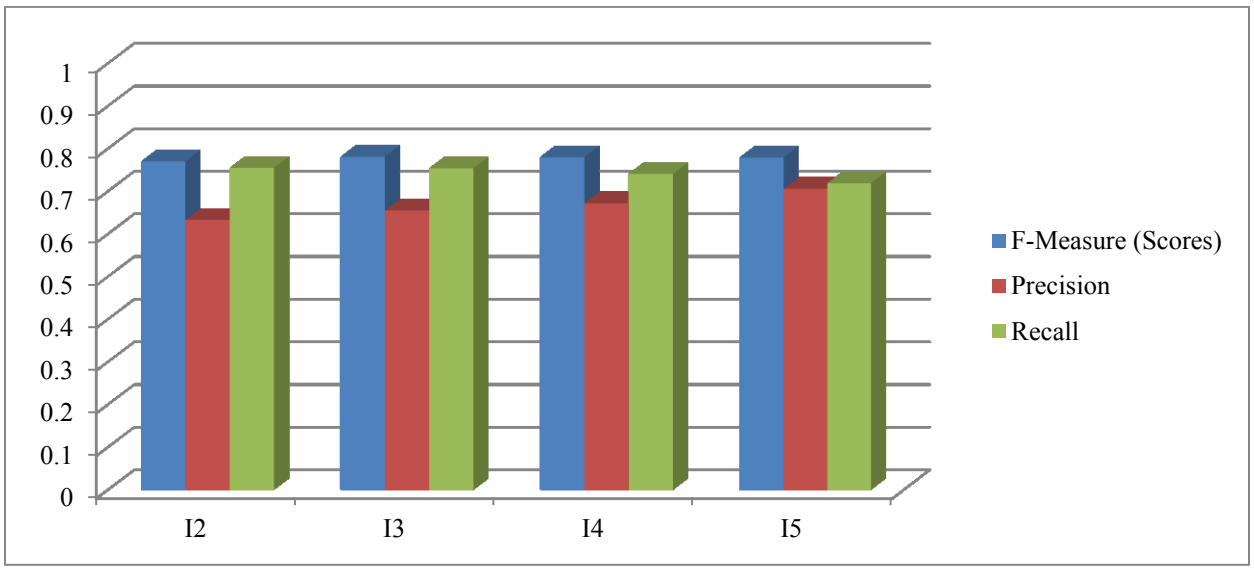


Figure 43: Comparison Chart to Identify the Best Inflation factor

8.4 Cluster Evaluation

The accuracy of clusters is determined by using External and Internal cluster validation technique.

Internal Cluster Validation: Internal cluster validation score is calculated by the information intrinsic to the data alone. The validation process is carried out using the silhouette validation technique. Formula for the approach is

$$\text{Silhouette method } = \frac{b(i) - a(i)}{\max(b(i) - a(i))}$$

where a(i) denotes the average distance between i and all data items in the same cluster, b(i) denotes the average distance between i and all data items in the closest other cluster. The silhouette width is between (-1, 1) and should be maximized. The average distance is computed by considering only the ontological terms of each criterion. The distance between a(i) and b(i) is calculated as (1-pair wise score). . The silhouette value for an individual data item i reflects the confidence of the data item in this cluster assignment. The below table depicts the validation scores for the approach.

Table 3: Summary of Internal Cluster Validation Using Silhouette Width

Clusters	Number of Clusters	Silhouette scores
Inclusion Cluster Set	126	0.275
Exclusion Cluster Set	175	0.472

External Cluster Validation: External cluster validation technique is performed based on the external knowledge about the data. This is a semi-automatic procedure as the precision and recall calculation requires manual intervention. We used F-measure to validation the clusters. The formula for F-measure calculation for each cluster is as follows

$$F_i = \frac{((\beta + 1) * \text{Precision}_i * \text{Recall\_i})}{\beta * \text{Precision}_i + \text{Recall\_1}}$$

where F_i is the measure F-measure of cluster "I", Precision_i is the precision of cluster "i" and Recall_i is the recall for cluster i. and β is taken as 1 and is used for normalization. The precision and recall are found as below

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision tell the accuracy of the cluster. Recall depicts the completeness of the cluster. The precision and recall values are in turn used to calculate F-measure for each cluster. Once we obtain the F-measure for individual clusters, the overall F-measure is calculated using the formulae.

$$\text{Overall } F_{\text{measure}} = \Sigma(\frac{C_i}{C * F_i})$$

where Ci is the number of criteria set in cluster "i".

C is the total number of criteria in the entire set

Fi is the F-measure of cluster "I"

The below table depicts the validation scores for the approach.

Table 4: Summary of External Cluster Validation Using F-measure

Clusters	Number of Clusters	Precision	Recall	F-measure
Inclusion Cluster Set	126	0.915	0.908	0.939
Exclusion Cluster Set	175	0.929	0.910	0.956

The histogram below depicts the Inclusion Criteria clustering accuracy using F-measure.

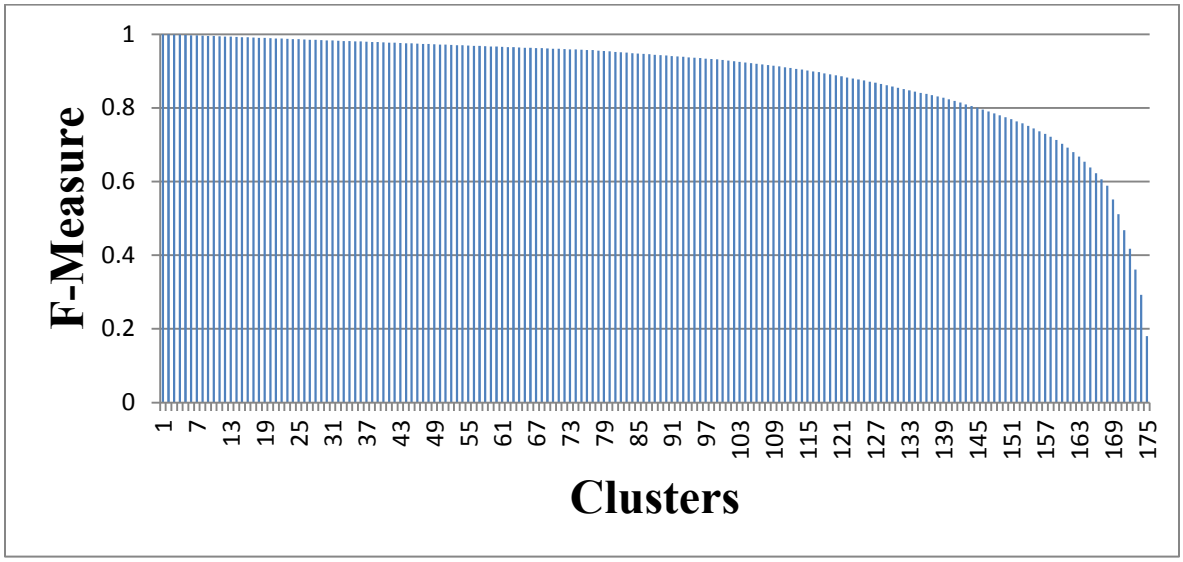


Figure 44: Inclusion Cluster Accuracy

Similar histogram is plotted to depict the exclusion criteria clustering accuracy.

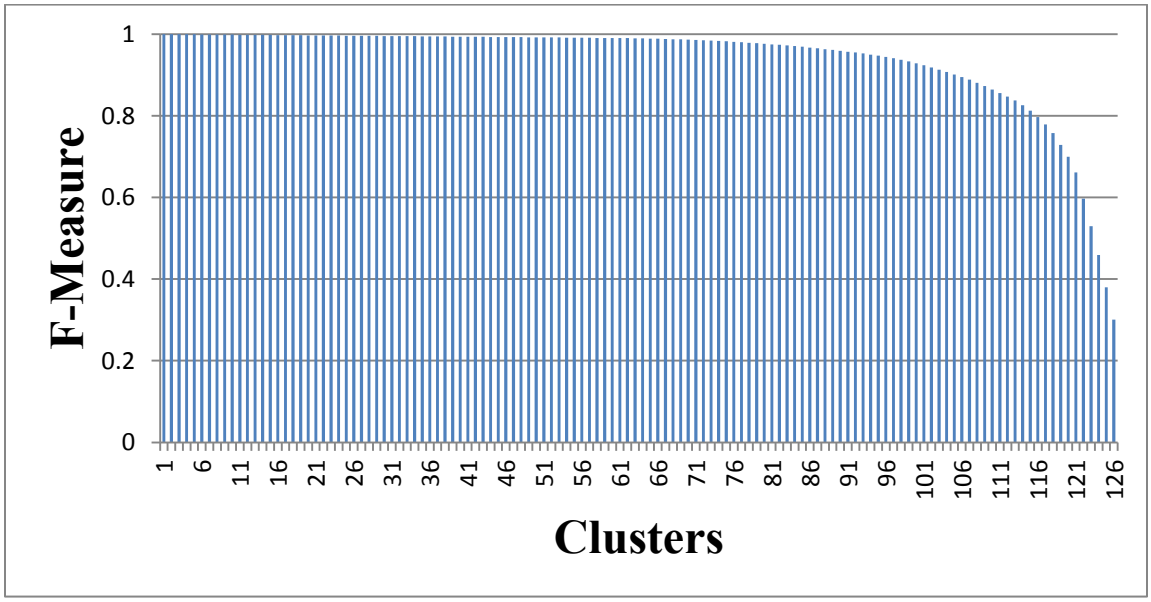


Figure 45: Exclusion Cluster Accuracy

8.5 Validation of Model Based Clustering Approach

Criteria merging to Cluster using Ontological TF-IDF Mapping Approach is another important phase in our model. We have demonstrated the application of this approach for 2 cases. The approach was first used to do model based clustering of merging a subset of exclusion criteria to inclusion cluster. In the second case it is used in the front end where the recruiter's build in criteria are mapped to the target. This approach was also used to do associative rule mining where the first step is used to map the target criteria to its corresponding cluster.

So validation of this approach is required. In order to validate the model we tried out 2 techniques.

Average Match Score Technique: In this process we calculated the symmetric pairwise score of the target criteria with the inclusion cluster and compared with the symmetric pairwise score inside the inclusion cluster. When the symmetric pairwise score of the target is greater than equal to symmetric pairwise of the inclusion cluster, then it is considered as a "correctly joined criteria". In that case 1460 criteria are correctly clustered. The overall accuracy was 1460/3027=48%. Used the same approach with the exclusion of Cluster 2 (disorder cluster) i.e. any criteria which has got merged to disorder cluster is considered as a valid. The reason is that the disorder cluster is extremely divergent and there were some new disorder that was not present in the inclusion cluster.

Maximum Match Score Technique: The other technique followed to validate the approach was the Maximum Match Score Technique. : In this approach instead of computing symmetric pairwise score taking the maximum pairwise score of the criteria with the inclusion cluster and if that is greater than equal to symmetric pairwise score of the inclusion cluster then it is considered as a "correctly merged criteria". The accuracy is 2674/3027 = 88%.

CHAPTER 9

CONCLUSION AND FUTURE WORK

9.1 Summary

In this thesis, we present a generalized bottom-up data driven approach to generate a minimal set of canonical non-redundant eligibility criteria for a given collection of clinical trials. The proposed model allows the flexibility of using free text while capturing the semantics of the criteria for computer readability. This approach complements with the top down specification of eligibility criteria based on formal ontologies.

This approach can aid in better characterization of both human volunteers and clinical study requirements, thus leading to accurate and efficient matching of subjects with clinical studies. The results obtained by the application of our approach can be used to generate an atomic set of eligibility criteria which can be used as to create successively refined or complex eligibility criteria by semantically combining the atomic set and can be readily incorporate into subject search engines of clinical trials.

9.2 Future Work

The data driven model leaves a lot of scope for the future. Some of the features that might be extended are

1. Extending the model to support inclusion/exclusion eligibility criteria from other mental disorders.
2. Develop a comprehensive rule set to split multiple criteria.
3. Classification of the criteria based on the application phase of clinical trials.

4. Propose a more random based approach for patient MockDb Generation.
5. Integrate the patient medical history from Microsoft HealthVault & Google Health into our patient database.

APPENDIX

The most frequent inclusion and exclusion clusters of GAD eligibility criteria are shown below.

Table 5: Top Most Frequent Inclusion clusters

ID	Cluster Title	Frequency	TF-IDF
2	a Diagnostic-and-Statistical-Manual-of-Mental-disorder-4th-Edition-Text-Revision-(DSM-IV-TR) diagnosis of generalized-anxiety-disorder	563	diagnostic-and-statistical-manual-of-mental-disorder-th-edition-text-revision-(dsm-iv-tr), diagnosis, generalized-anxiety-disorder, criteria
1	men or women 18 - 65 years of age	404	years, age, women, men, out-patients
4	The patient has a total score of at least on the Hamilton-Anxiety-Rating-Scale-(HAM-A) scale	326	score, total, Hamilton-anxiety-rating-scale-(ham-a), clinical-global-impression-severity
3	informed consent	199	consent, informed, written, signed, give, provide
13	stable dose of non-exclusionary medications and psychotherapeutic treatment for at least weeks prior to randomization	194	weeks, medication
12	Be in generally good physical health as determined by the Investigator on the basis of medical history physical examination and screening laboratory results	153	good, medical, health
5	Patients must be willing and able to comply with the study procedures	145	study, comply
8	Women must be of non-child-bearing potential [i.e. Post-menopausal be surgically sterile (hysterectomy or tubal ligation)] or Women must be given a pregnancy test (HCG)	126	contraception, women, reliable, child-bearing, method

ID	Cluster Title	Frequency	TF-IDF
6	negative serum pregnancy test for women of child-bearing potential	107	negative, pregnancy, women, child-bearing, urine, serum
9	have a score of on the pain visual analog scale (PVAS) score at screening	86	Pain
7	English speaking	72	English, speaking, language, native

Table 6: Top Most Frequent Exclusion clusters

ID	Cluster Title	Frequency	Total Frequency	TF-IDF
2	a Diagnostic-and-Statistical-Manual-of-Mental-disorder-4th-Edition-Text-Revision-(DSM-IV-TR) diagnosis of generalized-anxiety-disorder	1299	1823	obsessive-compulsive-disorder, post-traumatic-stress-disorder, generalized-anxiety-disorder, eating-disorder, co-morbid, biopsy-proven, comorbid-anxiety-disorder, separation-anxiety-disorder-(sad), semi-structured, mini, diagnostic-and-statistical-manual-of-mental-disorder-4th-edition-text-revision-(dsm-iv-tr)
13	stable dose of non-exclusionary medications and psychotherapeutic treatment for at least weeks prior to randomization	368	479	Medication
6	negative serum pregnancy test for women of child-bearing potential	281	385	negative, pregnancy, child-bearing, urine, serum

ID	Cluster Title	Frequency	Total Frequency	TF-IDF
12	Be in generally good physical health as determined by the Investigator on the basis of medical history physical examination and screening laboratory results	272	411	Health
4	The patient has a total score of at least on the Hamilton-Anxiety-Rating-Scale-(HAM-A) scale	101	427	clinician-administered-post-traumatic-stress-disorder-(caps), clinical-global-impression-severity, childrens-depression-inventory-parent-version-(cdi-p), skin-picking, simpson-angus-scale-(sas)
30	Moderate to severe brain injury operationalized as a GCS of or less in TBI participants and cognitive impairment of at least two standard deviations below expected levels in at least one cognitive domain in all participants	67	82	injury, brain, surgery
45	diagnosis of epilepsy with partial seizures as defined in the International League Against Epilepsy (ILAE) classification of seizures partial seizures may be simple or complex with or without secondary tonic-clonic generalization	57	67	seizures, partial, seizure, complex, epilepsy, simple, generalized, classification, generalization
10	any ethnic origin	56	50	origin, ethnicity, races, race

ID	Cluster Title	Frequency	Total Frequency	TF-IDF
33	At least an average of one - symptom panic attack per week over the last weeks prior to screening	52	55	panic, attack, weeks, symptom, one, week, baseline, per
8	Women must be of non child-bearing potential [i.e. Post-menopausal be surgically sterile (hysterectomy or tubal ligation)] or Women must be given a pregnancy test (HCG)	50	176	contraception, child-bearing, birth-control

REFERENCES

[1] Newell A. The Knowledge Level. AI Magazine. 1981; 2:1–33.

[2] Niland J, Dorr D, El Saadawi G, Embi P, Richesson RL, et al. Knowledge Representation Of Eligibility Criteria in Clinical Trials. in: American Medical Informatics Association Annual Symposium. 2007. Chicago.

[3] Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. Journal of Biomedical Informatics. 2010; 43(3): 451-67

[4] Niland J. ASPIRE: Agreement on Standardized Protocol Inclusion Requirements for Eligibility. 2007. Available from: http://hsspcohort.wikispaces.com/space/showimage/APIRE+CDISC+Intrachange+July+10+2007+Final.ppt.

[5] Fink E, Kokku PK, Nikiforou S, Hall LO, Goldgof DB, et al. Selection of patients for clinical trials: an interactive web-based system. Artificial Intelligence in Medicine 2004; 31: 241-254.

[6] Cohen EEA. caMATCH: A Patient Matching Tool for Clinical Trials. caBIG Annual Meeting: 2005.

[7] Olasav B, Sim I. RuleEd, a Web-based Semantic Network Interface for Constructing and Revising Computable Eligibility Rules. American Medical Informatics Annual Symposium. November 11-15, 2006: Washington, D.C.: 2006:1051.

[8] Kashyap V et al. Clinical Observations Interoperability: A Semantic Web Approach. American Medical Informatics Annual Symposium. Spring Congress 2009 Orlando.

[9] Butte A, Weinstein D, Kohane IS. Enrolling patients into clinical trials faster using RealTime Recuiting. American Medical Informatics Annual Symposium. 2000; 111–115

[10] Musen M, Carlson R, Fagan L, Deresinski S, Shortliffe E. T-HELPER: automated support for community-based clinical research. Annu Symp Comput Appl Med Care: 1992; 719-23.

[11] Koisch J, Mead C and Velezis M. Clinical Research Filtered Query. Available from: http://hssp-cohort.wikispaces.com/space/showimage/SFM_CRFQ_v1.0.doc.

[12] Sim I, Olasov B, Carini S. An ontology of randomized controlled trials for evidence based practice: content specification and evaluation using the competency decomposition method. Journal of Biomedical Informatics 2004; 37: 108-119.

[13] Costantino G, Ceriani E. Eligibility criteria of randomized controlled trials. Journal of the American Medical Association 2007; 297(11): 1233-40.

[14] Davis AM. Editorial. Study Eligibility Criteria: The Perils of Feasibility Based Decision Making The Journal of Rheumatology, 2005; 32(3):403.

[15] Noy NF, Crubezy M, Fergerson RW, et al. Protege-2000: an open-source ontology-development and knowledge-acquisition environment. American Medical Informatics Annual Symposium. Proc. 2003:953.

[16] Porter MF. An algorithm for suffix stripping. Program: Electronic Library and Information Systems, 2006; 40(3): 211-218.

[17] Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. August 21, 2007: 14(6): 687-96

[18] Tu SW, Kemper CA, Lane NM, Carlson RW, Musen MA. COI. Clinical Observations Interoperability: EMR+Clinical Trials, W3C (2008). Available from: http://esw.w3.org/topic/HCLS/ClinicalObservationsInteroperability.

[19] Handl J, Knowles J, Kell DB. Computational Cluster Validation in Post-genomic Data Analysis. Bioinformatics, 2005; 21(15):3201-3212

[20] Metz MJ, Coyle C, Hudson C, Hampshire M. An Internet-Based Cancer Clinical Trials Matching Resource. Journal of Medical Internet Research, 2005; 7:e24

[21] Patel C, Khan S, Gomadam K. TrialX: Using Semantic Technologies to Match Patients to Relevant Clinical Trials Based on Their Personal Health Records. Proc. of the International Semantic Web Conference (ISWC), 2009.

[22] Lee Y, Dinakarpandian D, Katakam N, Owens D. MindTrial: An Intelligent System for Clinical Trials, American Medical Informatics Association (AMIA). 2010 Annual Symposium.

[23] Tu SW, Peleg M, Carini S, Bobak M, et al. A practical method for transforming free-text eligibility criteria into computable criteria. Journal of Biomedical Informatics. 2010; 44(2):239-250

[24] Dongen SV. Graph clustering via a discrete uncoupling process. Siam Journal on Matrix Analysis and Applications, 2008; 30-1: 121-141.

[25] Lee Y, Katakam N, Dinakarpandian D, Owens D, Mathur S, Krishnamoorthy S, Wubbenhorst J. An Intelligent Online system for Enhanced Recruitment of Patients for Clinical Research, Missouri Regional Life Sciences Summit, 2010, Kansas City, MO.

[26] Krishnamoorthy S, Lee Y, Dinakarpandian D. Data Driven Derivation of Canonical Eligibility Criteria for Clinical Trials, CSHALS. 2011, Boston, MA.

[27] Lee Y, Dinakarpandian D, Katakam N, Owens D. MindTrial: An Intelligent System for Clinical Trials, American Medical Informatics Association (AMIA). 2010:442-6 (2010) PMID 21347017.

[28] Clinical Trials. Available from: http://www.clinicaltrials.gov/: accessed 30 Dec 2010.

[29] Mindtrial GAD Online Search Engine. Available from: http://c.mindtrial.com/ssearch/Default.aspx.

[30] Clinical Eligibility Criteria web services Available from: http://ec2-50-16-83-185.compute-1.amazonaws.com:8085/web_serv_for_mindtrial_ec2/

[31] Associative rule mining with WEKA Available from: http://maya.cs.depaul.edu/classes/ect584/weka/associate.html

[32] Zitko B, Stankov S, Rosic M, Grubisic A. Dynamic test generation over ontology-based knowledge representation in authoring shell. Expert Systems with Applications, 2009; 36(4):8185-8196.

VITA

Saranya Krishnamoorthy was born on May 13th, 1986, in Tamilnadu, India. She spent most of her childhood and teenage years in Salem, India. She then completed her Bachelors in Bioinformatics from Shanmugha Arts Science and Research Academy (SASTRA) University at Tanjore, India. After her under graduation she worked with INFOSYS for 2 years and then moved to United States for her higher studies. She is currently pursuing her Masters in Computer Science at the University of Missouri – Kansas City.

She worked as a Research intern for Midwest Psychiatric Research Group during FALL 2010 and then with Philips Research during summer 2011. She is currently working as an intern with Medical Knowledge Group since fall 2011. She continued working with Dr. Dinakarpandian Deendayal and Dr. Yugyung Lee on the research project titled Mindtrial funded by National Institute of Mental Health (NIMH). As a part of this research work Saranya has presented a poster at the Missouri Regional Life Sciences Summit -2010 and CSHALS--2011 (Conference on Semantics in Healthcare and Life Sciences). She is a recipient of GAF (Graduate Assistant Fund) from UMKC and is a member of Upsilon Pi Epsilon honor society

Posters:

1. Yugyung Lee, Nikhilesh Katakam, Deendayal Dinakarpandian, Dennis Owens, Sachin Mathur, Saranya Krishnamoorthy, & John Wubbenhorst, An Intelligent Online system for Enhanced Recruitment of Patients for Clinical Research, Missouri Regional Life Sciences Summit, 2010, Kansas City, MO.
2. Saranya Krishnamoorthy, Yugyung Lee & Deendayal Dinakarpandian, Data Driven Derivation of Canonical Eligibility Criteria for Clinical Trials CSHALS--2011 Boston, MA.