Text Classification and Information Extraction from Abstracts of Randomized Clinical Trials:

One step closer to personalized semantic medical evidence search

Rong Xu
Yael Garten

Stanford Biomedical Informatics Training Program

Final Project: CS224N
Spring, 2006

Table of Contents

ABSTRACT

Patients desire medical information, such as efficacy of various treatments or side effects of drugs administered in clinical trials, that is specifically relevant to their demographic group and medical condition. Current search methods are usually keyword-based, and as such, do not incorporate semantics into the search algorithm. The results are either a wealth of information which contains a high degree of irrelevant hits, or at the other extreme, no results at all, because the specific query words were not found although their semantic equivalents do appear quite often in the database. Thus, the solution is to perform semantic search. In order to do this, information extraction must be performed *a priori*. The current study extracts the desired information from randomized clinical trial (RCT) papers. RCTs are one of the most important sources of scientific evidence on the safety and effectiveness of health interventions. Patients and the general public need such information, to help make treatment decisions.

Ideally, a person will be able to enter a query about the disease and population groups of interest to him, and receive synthesized information from all RCT papers that describe a study performed on a similar population or disease. The current work takes a step in that direction. We developed a system for information extraction from abstracts of RCT papers, of specific fields of interest: patient demographic descriptors ("males", "elderly"), medical conditions and descriptors ("multiple sclerosis", "severe diabetes"), and total number of patients in the study (which is indicative of the quality of the study).

In order to perform information extraction, we divided our task into three parts. First, we separated unstructured abstracts into five sections (Background, Objective, Methods, Results, Conclusions) using text classification and a Hidden Markov Model. We achieved a high accuracy rate of about 94% in average. Secondly, we classified the sentences in the Methods sections into two classes: those that are trial participant-related, and those that are not, and achieved an overall performance of 91%. Thirdly, we extracted specific types of information from participant-related sentences: total number of participants, demographic information related to the participants, and medical information (disease, symptoms). Accuracy for extraction of the number of participants is 92.5% and 82.5% for the demographic information.

INTRODUCTION

Today, when patients use search engines to obtain medical information relevant to a particular medical condition and demographic group, they often use search engines like Google an Pubmed, two keyword based search engines. For example, when an elderly Caucasian diabetic male uses the search query "diabetes Caucasian 70 year old male", he receives 166,000 search hits in Google, and zero in Pubmed. The results of Google may include pages upon pages of hits that are completely irrelevant. For example, one of the top-scoring hits is a document that discusses a 30-year old African American female with diabetes whose 70-year old father has Parkinson's disease. At the other extreme, Pubmed is so specific a search engine that it does not map "70 year old" to elderly and thus retrieves zero hits (whereas a search on "diabetes Caucasian elderly male" does retrieve 25 hits).

There is a very real and urgent need for the development of an authoritative personalized semantic medical-evidence search engine. This study takes a step in that direction. Pubmed, run by the National Library of Medicine, is the most reliable source of medical information today, and within it lies a subset of 204,000 papers called Randomized Clinical Trial papers, or RCTs, which provide reliable medical evidence. RCT papers usually report on the results of a treatment or intervention that was carried out on a specific small group of participants, usually as treatment to a particular disease.

Authors of RCT papers usually include a few sentences on each of five general topics in the abstract of the paper: Background, Objective, Methods, Results, and Conclusions. Two of these topics are most important to the patient seeking medical evidence: (1) the Methods, which describe the intervention itself and the demographics of the participants, and thus allows the user to decide whether the study is relevant to him, and (2) the Conclusions, which summarize the efficacy of the intervention on that particular group of participants. As we were interested in the extraction of information regarding the population that each study was performed upon, we focused on the Methods section, and sought to extract from this section all information that could assist in allowing for personalized semantic search in the future.

Within the sentences that discuss the Methods, there are generally five types of information conveyed about the methods used in the study. These are: settings, design, participants, interventions, and outcome measures. Again, as we are interested in extracting information about the trial population itself, we focused on analysis of only the "participants" sentences. Within the sentences that remain, lie the desired information, and most useful to patients using search engines are three types of information, which we sought to extract: (1) the total number of participants, which points to the quality of the study, because the larger the number of participants in a clinical study, the more reliable it is, (2) the demographic information regarding the participants (age, gender, etc.), and (3) the medical information such as diseases or symptoms which the participants had. These three pieces of information can allow personalized, medical evidence searching.

A subset of RCT papers has been written using structured abstracts, in which authors provide explicit subheading information for all the sentences in an abstract. That is, they separate the sentences with headings like "Objective:", "Methods:", etc. And within these structured abstract papers, some authors even separate their Methods section into subsections such as "Settings", "Participants", etc. These are extremely useful for our purposes. However, as only 20% of the RCT papers do use structured abstracts, and of those, only 8% have the "Participants" tag, our main efforts must still be devoted to automatic extraction of population information from unstructured abstracts. (See Appendix A for examples of structured and unstructured abstracts.)

Thus, in this study, we carried out three main steps:

1) Classify unstructured abstract into 5 sections (Background, Objective, Methods, Results, Conclusions)
2) Extract the 'Methods' section identified in step 1 and classify each of its sentences into two classes: those that discuss the population (or patients) and those that do not {PATIENTS, OTHER}
3) Using only those sentences identified as population-related sentences in step 2 (classified as PATIENTS), extract three specific types of information:
    1. Total number of participants (or all subgroups, when total number is not available)
    2. Patient descriptors (such as "males", "diabetics", "healthy", "elderly")
    3. Medical information (i.e. disease or symptoms)

METHODS

Our goal was to extract specific information from the abstracts of RCTs, which describe the population group in each study. The three main steps of our workflow were enumerated above, in the Introduction. We developed methods to perform each of the above three steps. Briefly, step 1 was performed by the combination of text classification and Hidden Markov Modeling (HMM) techniques. Step 2 was performed using a Maximum Entropy classifier, and step 3 was performed using a combination of rules, closed sets of words, Stanford parser [3,4], MetaMap (a tool used for textual medical information) [5], and a grammar we developed that is specific to this domain. A detailed description of each step follows.

Step 1: Classification of abstract into 5 sections

As described above, RCT paper abstracts can be generally separated in terms of style and content into 5 sections (Background, Objective, Method, Results and Conclusions). The sentences in each section will generally be more similar to one another than to sentences in other sections, and we can artificially tag sentences as belonging to one of these 5 sections. For example, patient population information is more likely contained in sentences in the Methods section, and effectiveness of intervention is usually contained in the Conclusions section. The sentence type information (e.g., which section the sentence "comes" from) as well as sentence content, is useful to automatically extract information.

In order to automatically label sentences in an abstract as belonging to one of the five classes, we combined text classification with a Hidden Markov Modeling (HMM) technique [6] to categorize sentences in RCT abstracts. We selected the 3896 structured RCT abstracts published between 2004 and 2005 and parsed them into 46,370 sentences. Each sentence was a labeled input to a multi-class (Introduction, Objective, Methods, Results and Conclusions) text classifier. Of the 46,370 sentences, 50% were used for model training (23,185 sentences) and 50% for testing. We used MALLET, a text classification toolkit, in our study [7].

To pick the best method to represent a sentence, we compared the results of text classifiers when the sentence was presented as a (1) N-gram, (2) bag-of-words with stemming (3) bag-of-words with no stop words, and (4) unprocessed bag-of-words. Performance was measured by classification precision, recall and F1 measure (a composite measure of classification precision and recall). We found that sentences represented as a bag of words (unigram model) without preprocessing gave the best performance. Therefore, the sentences in our subsequent analyses were represented by the unprocessed bag-of-words model.

The performance of classification algorithms is application specific and depends on the underlying theoretical foundations of each classifier. We comparatively evaluated the performances of a range of text classification algorithms, including Naïve Bayes, Maximum Entropy and Decision trees.

For each of the algorithms, boosting and bagging techniques were applied. The main idea of boosting and bagging is to generate many, relatively weak classification rules and to combine these into a single highly accurate classification rule. To compare the performance of each classification method, the same training and testing samples were used for all classifiers.

The text categorization methods of Naïve Bayes, Maximum Entropy and Decision trees assume that every sentence in an abstract is independent of other sentences. This is not the case, however: If the sentence is categorized to belong to the Background section of an abstract, then the probability of the next sentence to belong to the Objective or Background section is higher than the probability of that sentence belonging to a Results or Conclusions section.

To exploit the sequential ordering of sentences in an abstract, we used an HMM to label sentence types. HMMs are commonly used for speech recognition and biological sequence alignment. In our case, we have transformed the sentence categorization problem into a HMM sequence alignment problem.

The HMM states correspond to the sentence types. Labeling sentences in an abstract is equivalent to aligning the sentences to the HMM states (Figure 1).

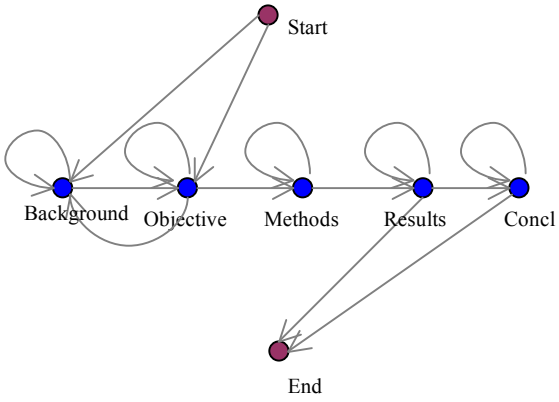


Figure 1. **HMM Model**. States represent the five sentence categories. Directed edges represent the direction of the transition probability.

There are five states in our HMM model: Background, Objective, Method, Result and Conclusion. The transition probabilities between these states were estimated from the training data by dividing the number of times each transition occurs in the training set by the sum of all the transitions. For example, the transition probability between the "Background" state and the "Objective" state is 0.2731, since in our training set, of the 4152 sentences in the background section, 1134 have a succeeding sentence in the objective section.

The state emission probabilities were calculated from the score output that were reported by the multi-class classifiers. For example, a Naïve Bayes classifier may report a probability of 0.48 for the given sentence to belong to the Background section, 0.42 for Objective section, 0.01 for Results section, 0.04 for the Methods section, and 0.05 that it belongs to the Conclusions section. To label this sentence using the HMM we assign these probability values to the respective states. Given the HMM model, state

emission probabilities, and the state transition probabilities, the Viterbi algorithm [6] was used to compute the most likely sequence of states that emits all the sentences in the abstract. Subsequently, the state associated with the sentence was extracted from the most likely sequence of states.

Step 2: Classification of sentences into population-related vs. other

The sentences describing the population of a clinical trial study usually appear in the Methods section of an abstract. However, as previously stated, along with this information, the author also describes other things in the Methods section, such as settings, design, intervention, and outcome measures. The work described in Step 1 allows us to extract a "Methods" section from an unstructured abstract that does not have an actual tagged Methods section. This powerful tool allows us to proceed, and to focus our information extraction efforts (recall the ultimate goal of extracting patient-specific fields) on a subset of sentences of the abstract, rather than the entire abstract. Step 1 retrieves for each abstract, a set of sentences probabilistically most relevant to being tagged "Methods". We now show how we can focus even further, by identifying within this section those sentences that are most relevant to our goal.

Given a Methods section comprised of several sentences, we used the same classification model described in Step 1 to classify sentences into two classes {PATIENT, OTHER} corresponding to whether or not the sentence contains information about the population group (as opposed to information about other things like settings, design, etc.). We tested all features and classifier models described in Step 1 above, and subsequently used the one with the best overall performance, which was a Maximum Entropy classifier, where sentences were represented by a combined model of unigram, bigram and unprocessed bag-of-words. (See Results section for F1 scores of the various methods.)

There is a very small percent of 3300 papers (out of the 204,000 total RCT papers) that have a section in the abstract titled "Participants", which specifically describes the participants of the clinical study. These sections may appear as a subsection of the Methods section, or as a separate section altogether, and are usually comprised of 1 or 2 sentences. (See Appendix E for an example abstract.) These sentences therefore contain information about the patient population, whereas sentences that appear in sections such as "Designs", "Analysis", "Settings" typically do not discuss the population. We initially used sentences from these two types of sections (extracted from the 3300 papers) as positive and negative training data for our classifier, rather than manually tagging sentences to be used as training data. Note that we selected subsections for our negative training data that would normally appear under the Methods section of a paper, thus "Conclusions" was not used even though it also would not discuss population information.

After running the classifier on 60% of the training data, and testing on the remaining 40%, we achieved an F1 score of 92%. However, when evaluating the classifier on a test set of sentences which came from 100 abstracts without a "Participants" subsection, the F1 was only 77%. As the majority of RCT papers do not have this "Participants" subsection, it was essential to train the text classifiers on the sentences from this type of abstracts. Thus, we manually tagged 1000 abstracts and used them as training and testing data for the classifier[1]. These 1000 abstracts do not contain a "Participants" subsection.

After training on the manually tagged training data, the classifier obtained an F1 score of 91%, which is quite good. There is obviously a significant difference in the writing style authors tend to use when they have a structured abstract which specifies a "Participants" subsection, versus a general Methods section which includes the participants information embedded within. From observation of the data it appears that the former are much more focused and limited to several common templates, such as:

[1] Lesson learned: No free lunch.

- Sixty-two patients hospitalized for acute asthma
- A total of 83 women and 12 men with lung cancer
- Participants were 97 adults at risk for depression

The latter, that is, those abstracts without a specified "Participants" section, tend to include much longer, more complex sentences. (e.g. "A randomized, double-blind, placebo-controlled, multicenter, pilot study was performed to evaluate the effects of 32 mg/day ruboxistaurin for 1 year in persons (n = 123) with type 2 diabetes and persistent albuminuria (albumin-to-creatinine ratio [ACR] 200-2,000 mg/g), despite therapy with renin-angiotensin system inhibitors.")

Step 3: Information Extraction from sentences tagged as population-related

Our final goal was to extract specific types of information from the sentences which describe the population of each study. We use only those sentences identified as population-related (classified as PATIENTS) sentences in step 2, and extract three specific types of information from them:

1. Total number of participants (or all subgroups, when total number is not available)
2. Participant descriptors (such as "male", "diabetics", "healthy", "elderly")
3. Disease or Symptom

This step proved to be the most challenging, as there is great variability in the sentence structures. Step 3 was performed using a combination of rules, closed sets of words, and parsing using a CKY Parser with a grammar we developed that is specific to this domain. In parsing the sentences, we used the Stanford Parser trained on the Penn Treebank corpus. We also considered using the GENIA corpus, which is a biological corpus, however we found that the language usage in the clinical trial paper is actually much closer to the Penn Treebank corpus than to the GENIA corpus.

We randomly selected 240 sentences extracted from abstracts of papers published in 2005, which were tagged by Step 2 as population-related. 200 sentences were used as training data, and 40 sentences were set aside as test data.

A typical sentence is: "251 adult outpatients with acute major depression were identified".
We were interested in extracting three basic features characterizing a RCT: (1) the total number of patients or subgroups used ("251" in this case), (2) the patient descriptors ("adult outpatients"), and (3) the disease along with its modifiers ("acute major depression"). The total number of patients used in the study is of interest (to a greater extent than the actual numbers in each subgroup) because it points to the quality of the study, as larger studies have greater predictive value to clinicians and patients.

We now describe the steps used to extract each of these types of information.

Type 1: Extracting total number of participants

For each sentence, we tagged all words identified by the Stanford parser as "CD" with the tag [NUMBER]. However, we observed that this will not properly tag all numbers. For example, in the phrase "351 in intervention group and 411 in control group", the number 351 was tagged as "CD", but 411 was tagged as "NP". Thus, we augmented the number tagging by using a closed set of words, and several rules. The closed set of words included words like {one, two, three…, twenty, thirty, …, hundred, thousand…}. If any of these words appeared in the sentence, the word was tagged as [NUMBER] even if the Stanford parser did not tag it as "CD". We then added several rules, such as if a word is a combination of two words from the closed set connected by a hyphen, such as "fifty-two", it should also be tagged as [NUMBER]. Also, if any word consists of only digits, such as "411", tag it as a [NUMBER]. This combination of closed sets and rules allowed us to tag the majority of numbers that appear in the

sentence. See Appendix C for a full list of the closed sets.

As we were interested in extracting the total number of participants in the trial, our next step was to filter out all tagged numbers that cannot pertain to study size. For example, the following phrases contain words tagged as [NUMBER] that are obviously not the population size: “16 years old”, “May 15”, “in 2004”, etc. After filtering, the remaining set of words tagged as [NUMBER] should contain the information pertinent to total population (or subpopulation) size.

However, as each sentence at this stage could still have several words tagged as [NUMBER], how can we know which of these numbers is the one of interest? For example: in the sentence “Patients included 100 men and 100 women”, there are two numbers and we are interested in both (total number is 100 + 100). However, in the sentence “300 patients included 50 elderly men and 100 adolescent girls” we are interested in the total number “300”.

In addition to the [NUMBER] tags we also created [PATIENT] tags that would aid us in understanding the semantics of the sentence. The exact mapping of words to this tag will be described in greater detail below; suffice it to say that words were tagged [PATIENT] if they belonged to a closed set of words such as {subjects, participants, men, women, volunteers, …}. Thus sentences were transformed to include these two tags in place of the relevant words.

In order to extract the correct number of interest from each sentence, we began by experimenting with a rule-based system, in which we wrote rules based on template sentence structures that seemed to occur often. For example: “30 men with lymphoma were studied” is a common way of stating the population information (this sentence maps to the reduced sentence [NUMBER] [PATIENT]). However, after testing it appeared that there are many exceptions to these rules, as there is variability in sentence structure, number of subgroups, etc, and thus we did not obtain sufficient performance by the rule-based system. Often we needed specific recursive rules, which we had not implemented. Thus, we turned to the creation of a grammar which would encapsulate the way in which authors discuss patient populations in a more general way.

We created a grammar which can be seen in Appendix B. We converted our grammar to CNF form and adapted the CKY backtracking algorithm to commence not by looking at the whole span of the sentence, but rather at the largest left-aligned span that created a whole “sentence” based on our grammar. For an explanation of why we used the left-aligned span, we observe the following example: If the original sentence is “The study observed 300 men who underwent over 6 operations each.”, we first transform the sentence by tagging the words by the tags {[NUMBER], [PATIENT]}, and then extract only those words that are in our grammar, and omitting all others. As the grammar includes the words: {[NUMBER], [PATIENT], and, or, included, …}, the sentence would be transformed to “[NUMBER] [PATIENT] [NUMBER]” (corresponding to “300 men 6”). Our grammar does not have a rule S→ [NUMBER] [PATIENT] [NUMBER], but does have a rule S→ [NUMBER] [PATIENT], and thus only “300 men” will be maintained, and subsequently “300” will be returned as the correct total number of patients. If our sentence was “Patients included 100 men and 100 women, …” we would return the result [100, 100] corresponding to the subgroups [men, women].

If, however, the sentence were similar to the above, but did specify the total number of patients ("300 patients included 100 men and 100 women”), we would return the number 300, which represents the total number of patients. The sentence would first be transformed to “[NUMBER] [PATIENT] included [NUMBER] [PATIENT] and [NUMBER] [PATIENT].” Again, we see left alignment: Since we have the rule S→ [NUMBER] [PATIENT], this would be aligned to “300 patients”, the rest would be truncated, and 300 would be returned as the total number of patients.
(The rule S→ [PATIENT] included [NUMBER] [PATIENT] and [NUMBER] [PATIENT] only deals

with cases in which the subgroup numbers were specified, but the total number was not.)

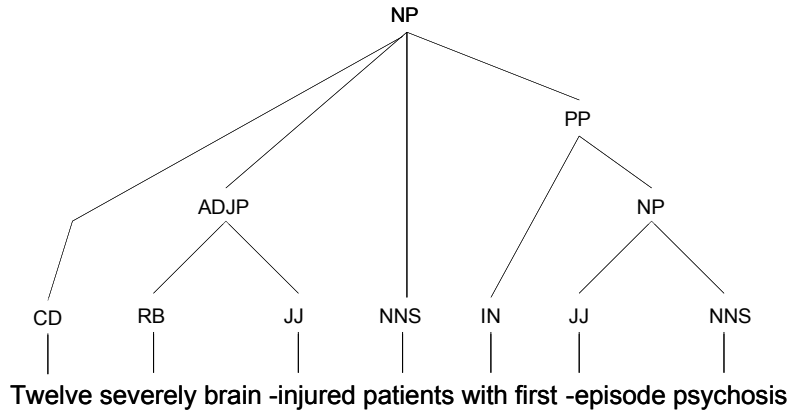
Type 2: Extracting participant descriptors

For each sentence, we were interested in extracting participant descriptors (such as "male", "diabetics", "healthy", "elderly"). These words usually appear before words such as {patients, participants, subjects} in a sentence (e.g. "we studied 300 elderly diabetic patients").

We created a closed set of words which identify the population anchor, which includes words such as the above mentioned {patients, participants, subjects…}. For a complete list, see Appendix C. These words were tagged as [PATIENT] in the original sentence.

In order to extract the descriptors that usually modify the word tagged [PATIENT], we used the parse tree of the sentence created by the Stanford parser. We find the word tagged as [PATIENT] in the parse tree, and then go up in the unary chain (usually [PATIENT] is an N which is a child of a NP) and then extract all other children of the NP, which are actually siblings of the [PATIENT] word. These are the descriptors of the study participants.

There could be descriptors on either side of the word tagged [PATIENT]. For example, in the tree shown below, the left sibling phrase of the word "patient" (which is tagged [PATIENT]) is "twelve severely brain-injured", and the right sibling phrase is "with first-episode psychosis". We also experimented with the option of taking only descriptors preceding the tagged word (to the left of [PATIENT]), and of taking descriptors preceding and following the tagged word (both sides of [PATIENT]).



Type 3: Extracting disease

In order to extract the disease, we first used the Stanford Parser to parse each sentence. Then, for each subphrase at each level of the tree, we used the subphrase as input to MetaMap [5]. MetaMap is a tool that can identify whether or not a phrase relates to a disease, sign, syndrome, or symptom, and it does so by mapping the input phrase onto concepts in the UMLS Metathesaurus, a metathesaurus of all clinically related terms (e.g. "rhinosinusitis" is mapped to the semantic type "Disease or Syndrome").

For examples of semantic types used by MetaMap, see Appendix D.

We run each phrase against the UMLS via MetaMap, to retrieve the smallest phrase that maps to "Disease

or Syndrome". We do this for each subphrase in the sentence rather than the whole sentence, because MetaMap takes a phrase as input, and checks if any word/phrase within it is contained in UMLS dictionary. However, if we enter the whole sentence, it may return the result that the sentence does match to the "Disease or Syndrome" concept, but it will not say which words in the sentence mapped to this concept.

In addition to the "Disease or Syndrome" concept, we also retrieved results which mapped to symptoms (concept "Sign or Symptom"). This process extracts the clinical information from the sentence, which is otherwise very difficult to do, as it is extremely domain specific.

As explained above, since MetaMap doesn't give an upper bound on phrase length and just checks if any words or subphrase in a phrase is a disease, we must limit it by starting with the smallest phrase in the parse tree and slowly expanding to include siblings and parents. When we find a phrase that is identified as a disease or symptom, we stop. However, such a technique will tag "diabetes" as the [DISEASE], when the phrase in the original sentence that encompasses the full disease is actually "acute juvenile diabetes". Because we want some descriptors around the disease, we can use the parse tree to help us collect a larger phrase than just "diabetes", in a way similar to what was done when extracting participant descriptors in the previous section. Thus, in this example we extract the whole phrase following the preposition "with", which is "acute juvenile diabetes". We extract modifiers surrounding the noun phrase [DISEASE] to the left and to the right.

RESULTS AND DISCUSSION

We describe below the results of each of the three steps described in the Methods section of the paper, along with error analysis and discussion.

Step 1: Classification of abstract into 5 sections

Results:
We have evaluated the performance of classifying sentences in RCT abstracts using three widely-used text classification algorithms, namely, Naïve Bayes (NB), Maximum Entropy (ME) and Decision Tree (DT): basic (without modifications), with boosting, and with bagging. Precision, recall, and F1 are compared across five type of abstract sections with and without HMM augmentation. The results of F1 measure are summarized in Figure 2. Each comparison is made on F1 before and after using HMM augmentation.

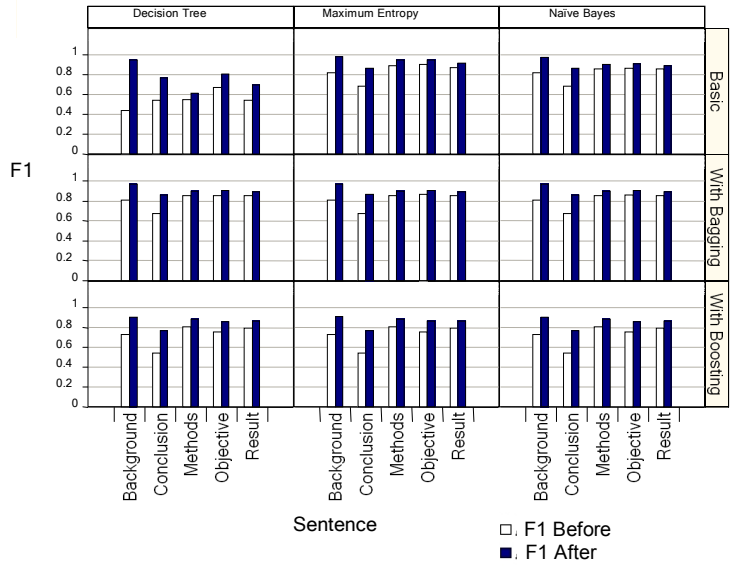


Figure 2. Comparative plot of F1 measure on text-classification algorithms before HMM and after using HMM augmentation.

The results show F1 improvements across five types of abstract sections with HMM augmentation. Best overall F1 measures were achieved with Maximum Entropy Model (basic) with HMM augmentation, 98.20% (95% CI 97.95-98.24%) for Background, 94.2% (93.62-94.47%) for Objective, 94.40% (94.02-94.53%) for Methods, 91.5% (89.1-92.8%) for Results and 85.7% (84.1-87.5%) for Conclusion sentence type.

Step2: Classification of sentences into population-related vs. other

Results:
In classifying sentences into the two classes {PATIENT, OTHER}, we experimented with a number of different features and classifiers, F1 scores appear below. On average, the best performance was obtained by the Maximum Entropy and Naïve Bayes classifiers, which performed much better than the Decision

Tree. The best results were obtained when sentences were represented by a combined model of unigram, bigram, and unprocessed bag-of-words.

In general we find that the Maximum Entropy classifier is better on average than the Naïve Bayes classifier, because one may use redundant features, and this won't affect performance because the method doesn't assume independence between features (such as bigram and unigram which are dependent, or the number of words in a sentence and the sentence length, which are also dependent).

	Naïve Bayes	Maximum Entropy	Decision Tree
Unigram	0.90465	0.9063	0.8034
Bigram	0.6671	0.8154	0.8034
Trigram	0.5882	0.7661	0.7545
Uni + Bi	0.9153	0.91159	0.8249
Uni + Bi +Tri	0.90598	0.90816	0.8272

Table 1: **F1 measure** of Naïve Bayes, Maximum Entropy, and Decision Tree as classifiers of sentences into the classes {PATIENT, OTHER}.

In our case Naïve Bayes might be minutely better (0.9153 vs 0.9115) because we have a small amount of data and few features. However on average their performance is quite similar.

Error analysis:
An F1 score of 91% may not be adequate for our needs, as we will use the results of this step as the substrate for information extraction in Step 3. Thus, future work will include improvement of the classifier. For example, we tried adding or removing stop words, transforming all words to lowercase, stemming/not stemming. Other features we might explore are perhaps more domain-specific, such as phrases or structures that appear often in the medical context describing patients, as opposed to other words.

Step 3: Information Extraction from sentences tagged as population-related

As briefly explained in the Methods section, Step 3 of our workflow was the most challenging. We experimented with a variety of information extraction methods, from simple rules to creation of grammars, and in this section we show some the results, and some errors that arise when extracting each of the three types of information.

Type 1: Extracting total number of participants

Results:
Our accuracy of identification of the correct number was 80% in the training data, and 92.5% on the test data. The higher score achieved in the test data seems to be because the random separation into two groups of 200 and 40 sentences happened to select very simple, regular sentences for the test data, whereas the training data contains more structurally complex and irregular sentences.

Error analysis:
Because our number extraction scheme begins by using the Stanford parser, and tags all words identified as "CD" with the [NUMBER] tag, if this tagging is wrong, we introduce errors. For example, the Stanford parser parses the sentence "Participants were GP patients…" and tags the word "GP" as part of speech "CD". Thus, when transforming the sentence we wrongly include the tag [NUMBER] for the word non-number word "GP".

Another interesting phenomenon we encountered in our error analysis of number identification was with positional displacement:

- This sentence: "We checked 20 patients for different diseases and symptoms and we found that 6 of the patients had …" would be transformed to:
- "[NUMBER] [PATIENT] and [NUMBER] [PATIENT]"
- Since our grammar has the rule S→ [NUMBER] [PATIENT] and [NUMBER] [PATIENT] because this rule often allows us to identify two subgroups (e.g."20 men and 30 women"), the grammar in this case interprets the sentence as including two subgroups of 20 patients and 6 patients, which of course is wrong, and simply results from the fact that by chance we have the word "and" in the sentence, followed by a number and the word "patients" with words interspersed between them, that are removed due to the sentence transformation algorithm. Thus, we added a heuristic in which if there is a distance of over 10 words between the 2 numbers, we truncate the transformed sentence before the second number. This solves the problem well for many cases. However, one confound is that we may then lose instances that are real. "We checked 20 patients with a variety of mild or acute types of heart diseases and also checked 6 patients with lung disease…". In this sentence the two true subgroups are in fact separated by many words.

Another problem we came across relates to the comma included in our grammar. We included the comma because often, subgroups are separated from one another by commas, and since we wanted to maintain the subgroup information, it was necessary to recognize the comma. A common sentence structure is "Participants included 100 men, women, and children" and so we made grammar rule:
S → [PATIENT] included [NUMBER] [PATIENT] comma [PATIENT] comma [PATIENT].
An additional type of sentence is "Participants included 100 men and 200 women", which is handled by the rule: S→ [PATIENT] included [NUMBER] [PATIENT] and [NUMBER] [PATIENT].
However, because we try to recognize both example sentence structures, we introduce noise that causes us to miss other types of sentence structure. For example, if we have the sentence "Participants included 100 men and 200 healthy, young women.", this sentence is transformed to [PATIENT] included [NUMBER] [PATIENT] and [NUMBER] comma [PATIENT]. Since the grammar does not handle such a sentence, in which the comma separates the second number from the patient word, our grammar will wrongly return only 100.

For complete results, see the file named "number.txt" posted at http://stanford.edu/~xurong/output/ (which for each sentence includes the original sentence, the sentence with some words tagged as [NUMBER], and the extracted total number of participants).

Type 2: Extracting participant descriptors

Results:
After tagging the relevant word or words in the original sentence with the [PATIENT] tag based on our closed set of words {patients, participants, subjects, etc.}, we then used the parse tree to extract siblings of this word which describe it. We experimented with extraction of descriptors to the left, right, and on both sides of the tagged word. Our accuracy of identification of the correct descriptors was 84% in the training data, and 82.5% on the test data.

Error analysis:
Errors may easily arise at this stage if the parse tree is wrong, as our method of extraction of descriptors depends solely on the parse. For example, the extraction may take a phrase that is too long. If the parent noun phrase of the word tagged [PATIENT] includes many words as leaves of its parse tree, the words extracted as descriptors of the participants will be noisy and include irrelevant words.

An opposite type of error may occur if the parse tree does not include enough words as siblings of the word tagged [PATIENT], and thus the descriptors phrase is not wide enough. For example, in a sentence which includes the phrase "20 healthy, ventilated young women", the parse tree returned "young women" as the noun phrase (NP), because it identified "ventilated" as a verb phrase (VP). The correct patient descriptor phrase would have included "healthy" and "ventilated".

In calculating accuracy, we only used extraction of information on the left of the word tagged as [PATIENT]. This is because in almost all sentences, the descriptor on the right side was a phrase such as "with diabetes", and was not correctly attached to the [PATIENT] word by the parse tree. This is probably due to the fact that the parser expects data that is in the form of full sentences, and tries to fit the input to match a full sentence structure, whereas our training and test data were mostly sentence fragments. This is because the 240 sentences we used were sentences from the abstracts that included a "Participants" subsection. We used these sentences in our information extraction step because we wanted to learn to do the process in a setting where sentences are simpler and more regular. However, although this is the case, our "sentences" are not in fact full sentences, and this causes problems such as the one described here. It may be that had we used training data that included complete sentences, as generally tend to appear in completely unstructured abstracts, we could have better performance and extraction of participant descriptors. This will be examined in the future. One possible rule that could be added is that if a prepositional phrase such as "with X" follows the word tagged as [PATIENT], attach it as a participant descriptor.
An example of this type of error: In the sentence "Twelve severely brain-injured patients with acute lung injury and intracranial pressure higher than applied PEEP.", the phrase "Twelve severely brain-injured patients" was extracted as the participant descriptor, when actually the "with acute lung injury and intracranial pressure" also describe the participants.

An additional type of error we found was in cases where a comma causes the parse tree to attach only part of the descriptors to the word tagged as [PATIENT]. For example, observe the following two similar sentences:

- Forty ASA physical status I and II adult patients scheduled for conventional pulmonary lobectomy.
- Forty ASA physical status I , II , and III patients presenting for primary total hip replacement .

In the first sentence, the phrase "Forty ASA physical status I and II adult patients" was extracted as the patient descriptor. However, in the second sentence, only the phrase "III patients" was extracted. This is because the comma before "and III patients" causes it to separate this phrase as a new noun phrase.

One additional error caused by incorrect parsing by the Stanford parser is demonstrated in the following example: The sentence "Participants were 649 people aged 12--39 years , all with mild to moderate inflammatory acne of the face" extracts "649 people" as the participant descriptor, instead of including the age of the participants. This is because the word "aged" is tagged as verb phrase (VP), and thus is not attached to the word "people".

For complete results, see the file named "patient_both_withTrees.txt" posted at http://stanford.edu/~xurong/output/ (which for each sentence includes the original sentence, the sentence with some words tagged as [PATIENT], the extracted PATIENT phrase, and the parse tree).

Type 3: Extracting disease

Results:

In extracting the disease, we used the results of MetaMap which tagged phrases in the sentence as mapping to a "Disease or Syndrome", or "Sign or Symptom". We then used the parse tree to expand the phrase, including descriptors of the disease. Overall, the recall of the disease-extraction is not good, because in most cases MetaMap cannot find any disease in the sentence, even if they do appear in the sentence. Specific examples are given below in the Error Analysis.

Error Analysis:
The most basic type of error arises when MetaMap does not recognize a disease as being a disease. To our surprise, the word "cancer" did not map to "Disease or Syndrome", but rather mapped to "Invertebrate". (e.g. "All patients attending the centres with oesophageal cancer" is not mapped to "Disease or Syndrome") As we currently have no other way to identify diseases, we do not overcome this problem. One possibility is to create a closed set of words that could be used in addition to MetaMap to identify diseases. However, if this closed set included the word "depression", which can also be used to describe an impression or dent, this may introduce another kind of error.

Additional errors arise if the parse is not correct, because then modifiers of the disease may not actually be relevant to the disease.

There are cases in which MetaMap recognizes a very general term as a "Disease or Syndrome", such as the words "critically" and "ill", as in the sentence: "Thirty-five randomly selected , mechanically ventilated , enterally fed critically ill patients ...".

In the sentence "Overweight men and women with mild-to-moderate dyslipidemia were recruited.", MetaMap tags "dyslipidemia" as a disease, and our algorithm also correctly extracts the descriptor "mild-to- moderate". However, the word "Overweight" is also tagged as a disease, and extending it to the right, based on the parse tree, gives incorrect results ("Overweight men and women" is tagged as a disease by our algorithm).

For complete results, see the file named "disease_both.txt" posted at http://stanford.edu/~xurong/output/ (which for each sentence includes the original sentence, the words tagged as [DISEASE], and the extracted DISEASE phrase).

Once MetaMap identified a word and tagged as [DISEASE], if we want to extend, and include descriptors, we do it based on the parse tree. However, since the parser was trained on the Wall Street Journal corpus, which does not contain many diseases, symptoms, and clinical findings, if it identifies a word never seen in the training corpus, it does not know the correct part of speech and thus gives equal probability to all POS tags. Thus we might incorrectly tag the disease as something like a verb, in which case the extension will not be the modifiers of a noun phrase and will lead to incorrect results, such as returning the entire sentence, as it is attached to the verb (subject on the left, object on the right).

Similar errors occur when looking at phrases tagged "Sign or Symptom" instead of "Disease or Syndrome".

CONCLUSIONS

We set out to perform information extraction of population information from the abstracts of randomized clinical trial papers. This task was divided into three parts. We first classified sentences in unstructured abstracts into five types (Background, Objective, Methods, Results, Conclusions), using a combined technique of text classification and a Hidden Markov Model. The results (precision of 0.94, recall of 0.93) of our approach are a significantly improvement over previously reported work on automated sentences categorization in RCT abstracts. We next classified the sentences in the Methods sections into two classes: those that are participant-related, and those that are not. This the first attempt in classifying sentences in clinical trial method sections. To do this we used a Maximum Entropy classifier, and achieved an overall performance of 91%. Finally, we extracted specific types of information from participant-related sentences: total number of participants, demographic information related to the participants, and medical information (disease, symptoms). Accuracy for extraction of the number of participants is 92.5% and 82.5% for the demographic information.

Our effort of automatic information extraction from RCT studies will significantly benefit the World Health Organization (WHO) International Clinical Trial Registry Platform effort [1]. International Clinical Trials Registry Platform is a major initiative aimed at standardizing the way information on medical studies is made available to the public through a process called registration. But the registration is voluntary and manual, and, currently, there are only few hundred registered clinical trials on the world. Our study can automatically populate the RCT bank with standard trial information extracted from the studies and provide a single point of access to the information.

FUTURE WORK

There are many improvements that can be made to our system. Some such examples:

- We can improve our grammar to make it both comprehensive and specific enough to achieve better performance on both training and testing data, for total number of participant extraction.
- We can improve the sentence parsing by training the Stanford parser on a corpus other than the Penn Treebank corpus. This should allow us to achieve better performance and more accurate descriptor extraction for participants and diseases.
- We must improve upon the MetaMap method, as it currently does not do well on extraction of the disease or symptoms. One alternative is to create a closed set of words such as {cancer, depression}. However, this is a crude alternative that we prefer to avoid.[2]
- As a result of this study, we learned that there is a very significant difference in sentence structure when authors write structured vs. unstructured abstracts, and even when structured, the presence of a subheading "Participants" has a significant effect on the sentence structures and thus information extraction performance. Therefore, we plan to repeat all analyses and grammar creation, using solely unstructured Methods sections.

[2] Suggestions on this point are particularly welcome, as we struggled quite a lot with it. ☺

REFERENCES

1. http://www.who.int/ictrp/en/
2. Dan Klein and Christopher D. Manning. 2002. Fast Exact Inference with a Factored Model for Natural Language Parsing. In Advances in Neural Information Processing Systems 15 (NIPS 2002), December 2002.
3. Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics.
4. Metamap: http://mmtx.nlm.nih.gov/
5. Rabiner,, L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition Proc IEEE 77 (2) 257-286
6. McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." 2002.
7. McKnight L and Srinivasan P. 2003. Categorization of sentence types in medical abstracts. In Proceedings of the 2003 AMIA conference.

APPENDIX A

Example abstracts from Ranomized Clincial Trial papers:

Structured abstract with "Methods" section:

BACKGROUND: Perforated necrotizing enterocolitis is a major cause of morbidity and mortality in premature infants, and the optimal treatment is uncertain. We designed this multicenter randomized trial to compare outcomes of primary peritoneal drainage with laparotomy and bowel resection in preterm infants with perforated necrotizing enterocolitis. METHODS: We randomly assigned 117 preterm infants (delivered before 34 weeks of gestation) with birth weights less than 1500 g and perforated necrotizing enterocolitis at 15 pediatric centers to undergo primary peritoneal drainage or laparotomy with bowel resection. Postoperative care was standardized. The primary outcome was survival at 90 days postoperatively. Secondary outcomes included dependence on parenteral nutrition 90 days postoperatively and length of hospital stay. RESULTS: At 90 days postoperatively, 19 of 55 infants assigned to primary peritoneal drainage had died (34.5 percent), as compared with 22 of 62 infants assigned to laparotomy (35.5 percent, P=0.92). The percentages of infants who depended on total parenteral nutrition were 17 of 36 (47.2 percent) in the peritoneal-drainage group and 16 of 40 (40.0 percent) in the laparotomy group (P=0.53). The mean (+/-SD) length of hospitalization for the 76 infants who were alive 90 days after operation was similar in the primary peritoneal-drainage and laparotomy groups (126+/-58 days and 116+/-56 days, respectively; P=0.43). Subgroup analyses stratified according to the presence or absence of radiographic evidence of extensive necrotizing enterocolitis (pneumatosis intestinalis), gestational age of less than 25 weeks, and serum pH less than 7.30 at presentation showed no significant advantage of either treatment in any group. CONCLUSIONS: The type of operation performed for perforated necrotizing enterocolitis does not influence survival or other clinically important early outcomes in preterm infants. (ClinicalTrials.gov number, NCT00252681.). Copyright 2006 Massachusetts Medical Society.

Moss RL, et al., N Engl J Med. 2006 May 25;354(21):2225-34. PMID: 16723614

Structured abstract with "Participants:" section:

BJECTIVE: To assess effects on condom use and other sexual behaviour of an HIV prevention programme at school that promotes the use of condoms with and without emergency contraception. DESIGN: Cluster randomised controlled trial. SETTING: 40 public high schools in the state of Morelos, Mexico. PARTICIPANTS: 10 954 first year high school students. INTERVENTION: Schools were randomised to one of three arms: an HIV prevention course that promoted condom use, the same course with emergency contraception as back-up, or the existing sex education course. Self administered anonymous questionnaires were completed at baseline, four months, and 16 months. Students at intervention schools received a 30 hour course (over 15 weeks) on HIV prevention and life skills, designed in accordance with guidelines of the joint United Nations programme on HIV/AIDS. Two extra hours of education on emergency contraception were given to students in the condom promotion with contraception arm. MAIN OUTCOME MEASURES: Primary outcome measure was reported condom use. Other outcomes were reported sexual activity; knowledge and attitudes about HIV and emergency contraception; and attitudes and confidence about condom use. RESULTS: Intervention did not affect reported condom use. Knowledge of HIV improved in both intervention arms and knowledge of emergency contraception improved in the condom promotion with contraception arm. Reported sexual behaviour was similar in the intervention arms and the control group. CONCLUSION: A rigorously

designed, implemented, and evaluated HIV education course based in public high schools did not reduce risk behaviour, so such courses need to be redesigned and evaluated. Addition of emergency contraception did not decrease reported condom use or increase risky sexual behaviour but did increase reported use of emergency contraception.

Walker, D, et al., BMJ. 2006 May 20;332(7551):1189-94. Epub 2006 May 8. PMID: 16682420

Unstructured abstract:

Authors performed a comparative study of replacement of captopril (2-3x daily) therapy with once daily enalapril. Blood pressure was measured by 24-hour ambulatory monitoring. The study enrolled 62 patients with mild to moderate hypertension. Captopril was previously administered 2-4 times per day, in mean dose of 74.4 mg, for 3.2 years as an average. After a 4 x 4 weeks study period the final enalapril dose was 15 mg, once in the morning. Enalapril was administered as monotherapy in 36 cases. During the trial mean blood pressure decreased from 140 +/- 14/85 +/- 9 mmHg to 125 +/- 12/76 +/- 7 mmHg ($p < 0.01$), diurnal index increased from 12/10% to 15/11%. Enalapril treatment lowered daytime percent time elevation index (PTI) from 56 +/- 30% to 27 +/- 23% ($p < 0.05$), nighttime PTI from 64 +/- 34% to 37 +/- 35% ($p < 0.05$), and hyperbaric impact values from 183 +/- 152 mmHg x hour to 97 +/- 128 mmHg x hour ($p < 0.01$). Adverse effects of both drugs were rare and mild, enalapril caused no changes in overall quality of life. Improvement of antihypertensive efficacy after a switch to enalapril treatment could be related to better compliance achieved by once-daily dose, and tight out-patient blood pressure control. The authors concluded that after a 16 weeks of therapy, once-daily enalapril administration was more effective and compliant in reducing blood pressure, than 2-4 times per day captopril treatment, when measured by 24-hour ambulatory blood pressure monitoring.

Ruszty L, et al, Orv Hetil. 1996 Dec 22;137(51):2851-4., PMID: 9679620

APPENDIX B

Lexicon and grammar created to identify number of participants in study. The grammar has been converted to CNF form.

Lexicon:

```
PP      -> [PATIENT]
NN      -> [NUMBER]
COMMA   -> ,
AND     -> and
OR      -> or
INCLUDE -> include
ARE     -> are
EQUAL   -> =
N'      -> n
N'      -> N
```

Grammar:

```
ROOT    -> S
S       -> PP
S       -> PP N'_EQUAL_NNT_
S       -> NNT PP
S       -> NNS_PP
S       -> PP ARE_NNS_PP
S       -> PP INCLUDE_NNS_PP

N'_EQUAL_NNT_   -> N' EQUAL_NNT_
EQUAL_NNT_      -> EQUAL NNT_

NNS_PP  -> NNS PP
NNS_PP  -> NNS_PP AND_NNS_PP
NNS_PP  -> NNS_PP COMMAR_AND_NNS_PP
COMMAR_AND_NNS_PP -> COMMAR AND_NNS_PP_
NNS_PP            -> NNS_PP COMMAR_NNS_PP
AND_NNS_PP_       -> AND NNS_PP_

COMMAR_NNS_PP  -> COMMA NNS_PP
ARE_NNS_PP     -> ARE NNS_PP
INCLUDE_NNS_PP -> INCLUDE NNS_PP
NNS_PP_        -> NNS PP

PP      -> PP AND_PP
PP      -> PP OR_PP
AND_PP  -> AND PP
OR_PP   -> OR PP

NNT     -> NN
NNT_    -> [NUMBER]
NNS     -> NN
NN      -> NN AND_NN
AND_NN  -> AND NN
NN      -> NN NN
```


```
NN      -> [NUMBER]
COMMA   -> ,
AND     -> and
OR      -> or
INCLUDE -> include
ARE     -> are
EQUAL   -> =
N'      -> n
N'      -> N
PP      -> [PATIENT]
```

APPENDIX C

These are the closed sets of words used to augment the labeling of participant words and numbers.

PATIENT =

{"patients", "men", "women", "subjects", "volunteers", "persons", "people", "participants", "children", "infants", "newborns", "teens", "students", "adults", "residents", "smokers", "neonates", "veterans", "individuals", "donors", "males", "boys", "girls", "seniors", "adolescents", "workers", "atheletes", "users", "babies", "recipients", "addicts", "diabetics", "outpatients", "inpatients", "overweight", "clients", "physicians"};

NUMBER_DIGITS=

{"0", "1", "2", "3", "4", "5", "6", "7", "8", "9"};

NUMBER_WORDS =

{"zero", "one", "two", "three", "four", "five", "six", "seven", "eight", "nine", "ten", "eleven", "twelve", "thirteen", "fourteen", "fifteen", "sixteen", "seventeen", "eighteen", "nineteen", "twenty", "thirty", "forty", "fifty", "sixty", "seventy", "eighty", "ninety", "hundred", "thousand", "million", "billion", "trillion"};
};

APPENDIX D

Below is a table of some of the semantic types used in MetaMap.

Abbreviation	Unique Identifier (TUI)	Full Name
acab	T020	Acquired Abnormality
acty	T052	Activity
aggp	T100	Age Group
anab	T190	Anatomical Abnormality
bpoc	T023	Body Part, Organ, or Organ Component
.	.	.
.	.	.
.	.	.
chem	T103	Chemical
clnd	T200	Clinical Drug
diap	T060	Diagnostic Procedure
dsyn	T047	Disease or Syndrome
edac	T065	Educational Activity
horm	T125	Hormone
humn	T016	Human
invt	T009	Invertebrate
medd	T074	Medical Device
moft	T044	Molecular Function
.	.	.
.	.	.
ortf	T042	Organ or Tissue Function
patf	T046	Pathologic Function
sosy	T184	Sign or Symptom
spco	T082	Spatial Concept
strd	T110	Steroid
tisu	T024	Tissue
tmco	T079	Temporal Concept
topp	T061	Therapeutic or Preventive Procedure
vtbt	T010	Vertebrate

APPENDIX E

Below is an example of an abstract that contains a "Participants" section, as well as other sections that relate to the Methods that do NOT contain participant information, such as Design, Setting, Intervention, and Main Outcome Measures.

OBJECTIVE: To assess the effect of family style mealtimes on quality of life, physical performance, and body weight of nursing home residents without dementia. DESIGN: Cluster randomised trial. SETTING: Five Dutch nursing homes. PARTICIPANTS: 178 residents (mean age 77 years). Two wards in each home were randomised to intervention (95 participants) or control groups (83). INTERVENTION: During six months the intervention group took their meals family style and the control group received the usual individual pre-plated service. MAIN OUTCOME MEASURES: Quality of life (perceived safety; autonomy; and sensory, physical, and psychosocial functioning), gross and fine motor function, and body weight. RESULTS: The difference in change between the groups was significant for overall quality of life (6.1 units, 95% confidence interval 2.1 to 10.3), fine motor function (1.8 units, 0.6 to 3.0), and body weight (1.5 kg, 0.6 to 2.4). CONCLUSION: Family style mealtimes maintain quality of life, physical performance, and body weight of nursing home residents without dementia. TRIAL REGISTRATION: Clinical trials NCT00114582.

Nijs KA, et al., BMJ. 2006 May 20;332(7551):1180-4. Epub 2006 May 5. PMID: 16679331

DESIGN:

Cluster randomised trial. SETTING: Five Dutch nursing homes.

PARTICIPANTS:

178 residents (mean age 77 years).

Two wards in each home were randomised to intervention (95 participants) or control groups (83).

INTERVENTION:

During six months the intervention group took their meals family style and the control group received the usual individual pre-plated service.

MAIN OUTCOME MEASURES:

Quality of life (perceived safety; autonomy; and sensory, physical, and psychosocial functioning), gross and fine motor function, and body weight.