Knowledge Extraction and Prediction from Behavior Science Randomized Controlled Trials: A Case Study in Smoking Cessation

Francesca Bonin, PhD[1], Martin Gleize, PhD[1], Yufang Hou, PhD[1], Debasis Ganguly, PhD[1], Ailbhe N. Finnerty, PhD[2], Charles Jochim, PhD[1], Alessandra Pascale, PhD[1], Pierpaolo Tommasi[1], Pol Mac Aonghusa, PhD[1], Susan Michie, PhD[2]
[1]IBM Research Europe, Dublin, Ireland
[2]University College London, UK

Abstract

Due to the fast pace at which randomized controlled trials are published in the health domain, researchers, consultants and policymakers would benefit from more automatic ways to process them by both extracting relevant information and automating the meta-analysis processes. In this paper, we present a novel methodology based on natural language processing and reasoning models to 1) extract relevant information from RCTs and 2) predict potential outcome values on novel scenarios, given the extracted knowledge, in the domain of behavior change for smoking cessation.

1 Introduction

Medical evidence is disseminated in unstructured, natural language scientific evaluation reports that describe the content and results of randomized controlled trials (RCTs). The evidence-based research is so vast that, at present, more than 100 reports of RCTs are published on average every day[1].

Systematic reviews seek to collate evidence that fits pre-specified eligibility criteria in order to answer a specific research question[2]. However, the time taken to complete them is estimated at about 1,000 hours of highly skilled manual work[3] or 67 weeks[4], from pre-registration stage to publication. It is evident that systematic reviews across the vast amount of literature are very time consuming and cannot cope with the fast pace at which new research studies get published. More evidence is produced and published than it is possible for researchers to be able to use, synthesize and analyze effectively with these conventional methods[5–7]. This is also true for behavior change researchers, health professionals and consultants that explore the literature of behavior change intervention reports, in order to understand the most effective methodology (or intervention) to help a certain population improve a specific target behavior (for example, stopping smoking). The volume and rate at which research is produced about behavior change is beyond the capability of human researchers to compare and understand which interventions are most effective and to be able to generalize the results to varying populations in different contexts[8].

In this context, health consultants and policy makers, as well as researchers, would benefit from automatic ways to extract information from RCTs and synthesize it in order to predict reasonable estimates of outcomes for new trials without waiting for a real-world evidence-based study to be conducted. Some effort in developing automatic ways to analyze RCT reports has been conducted in the NLP and health communities, by applying rule-based[9,10], or machine learning approaches[11–14] to the automatic extraction of information from evaluation reports. However, none of these studies present a systematic way to both extract the information and make prediction based on that information. An additional challenge stands in the complex structures of RCT reports. Usually RCT reports compare two or more 'arms' representing groups of participants that receive different interventions, most commonly the intervention being evaluated versus a comparator condition. Therefore, it is necessary to associate features such as characteristics of the population being studied with each arm. The task is made more difficult by the propensity of authors of reports to refer to arms using many different terms at different points in the report. Few studies at the moment have investigated arm detection[15].

Our Contributions. In this work, we intend to take a step forward in the research on the automatization of meta-analysis, by combining information extraction and regression solutions studied to extract information from RCT reports and make inferences on the potential success of a new intervention. We introduce a novel system that first recognizes the relevant information in RCTs, and then uses this information to estimate outcome values in new situations with a given population and set of interventions.

2 Related Work

Conducting systematic reviews of published literature is a common method by which literature, including RCTs among other types of scientific studies, can be synthesized and analyzed effectively to answer specific research questions[16]. An ongoing challenge in conducting systematic reviews is the time taken to complete them and it has become clear that the current method of conducting systematic reviews is not sustainable given the amount of time and effort required by researchers to manually conduct each review. There are a number of ways in which researchers are coming together in an attempt to develop more efficient approaches to analyzing evidence at a rate which would reduce the waste which is increasingly being recognized[5,7], like rapid systematic reviews[17] or living systematic reviews[6]. As well as altering existing traditional methods to conducting systematic reviews in an attempt to speed up the systematic review process, technology is being used where possible to automate the process and reduce the manual work required by the researchers. A list of current tools available is the SR Toolbox[a], a publicly available online catalog of software tools to aid the production of systematic reviews. A review of such tools has produced a practical guide of how and when to use them[18].

On the other hand, to enable those technologies, work has been done in the area of Natural Language Processing and specifically information extraction to automate the task of extracting knowledge from existing RCT reports or systematic reviews. Jasch[19] provides an exhaustive survey of IE approaches on BCI literature. Previous work has mainly concentrated on identifying PICO attributes[10,20,21] (P = patient, population or problem, I = intervention, C = comparison and O = outcome) either at the word or sentence level. We do not use the PICO classification, but rather the more fine-grained Human Behaviour Intervention Ontology (BCIO)[22], described below. The reason is the necessity to extract fine-grained information to be able to provide more detailed information to the final user and to improve the feature space used later for prediction. Most research has extracted information from medical abstracts, although some studies have used the full text of the articles[23]. Some of the methods are rule-based[9,10], some are based only on abstracts manually chosen to be more suitable to the study[11]. A few works have been conducted on supervised approaches for medical information extraction, which usually concentrate only on abstracts[12–14]. Other studies have exploited the entire article, for the extraction of papers' metadata[24]: the authors propose a preliminary system based on CRF for extracting formulaic text (authors' names, email and institution) as well as some key study parameters in a free text form, from PubMed Central articles. They reach promising results for the formulaic text, but only moderate success for the free text attributes. Similarly to Lin et al.[24], we exploit the entire text of the article, but our task of extracting attributes and arms is much more challenging than the extraction of formulaic text, as it varies from Behavior Change Techniques (BCTs) to numbers, such as outcome values or ages, which are highly ambiguous.

Few studies to the best of our knowledge have tackled the problem of *arm extraction* (i.e., extracting arm names and associating entities to the correct arm). One of the seminal studies[25] in PICO extraction collapsed intervention arms and comparison arms. In another study[15], the authors experiment with the use of co-reference for arm identification. The authors try to identify if tokens in medical abstracts are part of an arm name. We also exploit a statistical approach to extract arm name mentions, and then further try to associate the extracted entities with the those arm names.

With respect to prediction, recent work[26] has looked into inferring whether a given treatment is effective with a specified outcome from clinical trial reports. Differently from our work, they do not take into account any information outside the treatment itself for inferring the outcome and manually annotate the treatments and the outcome, concentrating only on the inference part.

To summarize, given previous work in information extraction from RCTs and prediction of outcome of clinical trials, our contribution differs for the following aspects: *1)* to the best of our knowledge this is the first attempt to provide outcome predictions based on automatically extracted data; *2)* we extract a wide variety of information (around 60 entities) based on the BCIO; *3)* we propose a first method for extracting armified results.

3 Ontology and Dataset

Ontology. To evaluate the effectiveness of a behavior change intervention there are many characteristics of the study which are relevant to consider. To be able to compare across multiple published reports it is necessary to have a

[a] `http://systematicreviewtools.com/about.php`

common structure to extract the relevant information needed to understand which methodology is most effective for particular interventions (e.g., interventions for smoking cessation). To facilitate the extraction of relevant information from published reports, an ontology of behavior change interventions[22] was developed as part of the Human Behaviour Change Project.[b] The ontology was created following a methodology developed for the project[c] and provided a structured classification of terms relevant to behavior change interventions.

The upper level of the Behaviour Change Intervention Ontology (BCIO) was developed using a basic structure of key entities and causal relationships about behavior change. The lower levels were developed to a more granular level using both a top-down approach, i.e., searching for key relevant terms in other classification systems or ontologies, and a bottom-up approach, i.e., from expert manual annotation. The lower levels of the ontologies were used as the structure for information extraction of data from full text reports. The entities in the lower level ontologies can be grouped into a number of upper level entities, which represent the most relevant information to extract from published reports to compare and evaluate the effectiveness of interventions. The upper level entities are:

- **Population:** An aggregate of people who are exposed to a behavior change intervention.
- **Setting:** An aggregate of entities constituting the environment in which a behavior change intervention is provided.
- **Outcome behavior:** Human behavior that is an intervention outcome.
- **Effect estimate:** A behavior change evaluation finding that characterizes the difference between behavior change intervention outcome estimates of two behavior change intervention scenarios.
- **Source:** A role played by a person, population or organization that provides a behavior change intervention.
- **Delivery:** A part of a behavior change intervention that is the means by which behavior change intervention content is provided.
- **Reach:** The difference between the behavior change intervention study sample and the planned behavior change intervention population.
- **Content (BCTs):** A planned process that is part of a behavior change intervention and is intended to be causally active in influencing the outcome behavior. A behavior change technique (BCT)[27], specifically, is a planned process that is the smallest part of intervention content that is observable, replicable and on its own has the potential to bring about behavior change.

From these eight upper levels, 57 lower level entities were used to extract data from the published reports, these entities and examples of the data extracted can be found in Table 1. Of the hundreds of lower level entities across the BCIO these 57 were chosen as they were 1) found the most often in reports, 2) included in other relevant ontologies such as PICO[d] or 3) believed to be most relevant for predicting intervention effectiveness.

Dataset. A total of 407 behavior change intervention reports were manually annotated, according to the structure of the BCIO, to create a database of relevant information related to smoking cessation intervention RCTs. The reports to annotate were identified through the Cochrane Database of Systematic Reviews[e] and through a collaboration with the IC-Smoke project[f]. A full account of the source of reports can be found online on OSF[g]. To be included for annotation the reports had to be 1) a randomized controlled trial (RCT), 2) included in a systematic review on smoking cessation, 3) included in a meta-analysis in a systematic review, and 4) have a behavioral outcome value related to smoking cessation.

A total of seven researchers with expertise in psychology and behavior change independently annotated the reports using EPPI-Reviewer.[h] Each report was annotated by two researchers and the pairs were varied across each 10-15 papers to minimize inconsistencies in the data. An annotation is a chunk of text assigned to its corresponding entity in the ontology (e.g., *19.5 years of age* represents the entity "Mean age"). As well as capturing the value of the entity, annotators label the context in which it appears. A full account of the annotation procedure was provided by Bonin

[b] https://www.humanbehaviourchange.org/
[c] https://osf.io/86m75/
[d] https://linkeddata.cochrane.org/pico-ontology
[e] https://www.cochranelibrary.com/cdsr/about-cdsr
[f] https://osf.io/23hfv/
[g] https://osf.io/myje6/
[h] http://eppi.ioe.ac.uk/eppireviewer4/

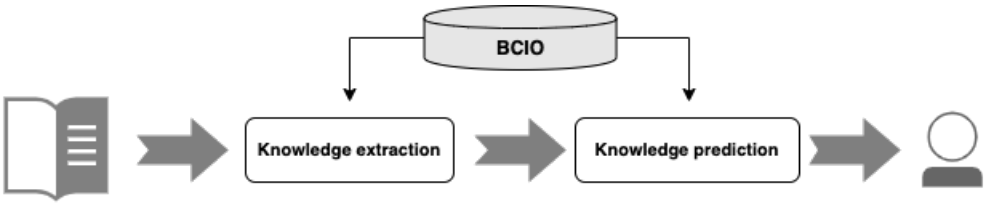

Figure 1: Overall pipeline of the knowledge system

et al.[28] Given the complexity of the task, double coding was required. Each annotated report was reconciled between two researchers and discrepancies were resolved. Finally, to ensure a high quality of data the inter-rater reliability (IRR) for pairs of annotators was assessed at various time points during the process. Krippendorf's alpha[29] was used to quantify the agreement (observed disagreement/expected disagreement) between the researchers. The agreement for individual entities varied, but overall a score above the acceptable threshold of agreement (α=0.67) recommended by Krippendorf[30] was achieved. The *HBCP-corpus*, constituting 407 published reports, was created as a part of our previous work[28]. To alleviate legal issues regarding open access, a subset of this collection, named *OA-HBCP-corpus*, was also created, comprising a sub-collection of fully open-access papers with annotation for 57 entities.

4 Knowledge System

In order to extract information from unstructured RCT reports and to make inferences over the extracted information, the methodology developed relies on two core components:

- a **knowledge extraction** module, which takes as input the text extracted from RCT PDFs, and returns relevant entities associated with the respective arms; and
- a **knowledge prediction** module, which uses the extracted information to estimate outcome values in unseen settings.

Both modules leverage the BCIO ontology. Figure 1 shows the entire workflow and, in the subsequent sections, we describe the methodologies behind the two components.

4.1 Knowledge Extraction Methodology

The aim of the knowledge extraction module is to extract structured knowledge from RCT reports in the form of entities (described in Table 1) with an associated arm. We first extract each entity and then associate it to the correct arm. The association of an entity with the corresponding arm is conducted by the *arm associator* described below. Specifically, the knowledge extraction module starts by taking as input a parsed PDF document. It has been tested using the ABBYY PDF parser[i] and a parser based on GROBID[31] with improved table parsing[32]. However, the code is easily adaptable to other PDF parsers. Then extraction is done by framing the problem as a *named entity recognition* task, for which we can use supervised machine learning. Figure 2 shows the pipeline of this module. The advantage of supervised approaches is their scalability and the fact that they do not require handcrafted rules. However, they also require large amount of data, which can be difficult to acquire. One of the challenges of our task is the variety in the entities to extract. Some of them, like BCTs are strings of text, others like 'outcome value' are numbers (usually percentages). In addition, usually those relevant entities are mentioned only once in a paper, so the variety of context from which a system can learn is limited.

In practice, we trained a named entity recognition model to extract RCTs listed in Table 1. We use BIO tagging for our task, where B, I and O represent the beginning, inside and outside of an entity, respectively. We employ a BiLSTM-CRF model, using the recent *Flair* framework[33] based on the concatenation of the following embeddings: GloVe (pre-trained on Wikipedia and Gigaword)[34], flair news-forward and news-backward contextual string embeddings (pre-trained on 1-billion word corpus). It is worth noting that annotations were created per document and, for each annotated entity, a snippet of text that contains the annotated entity is also highlighted as the corresponding context

[i] https://www.abbyy.com/

UL BCIO Class	Entities	Example annotation
Population	Mean age	The mean age of participants in the smoke-less-app group was **45**
	Proportion identifying as female gender	Sixty-one participants (**65.6**% female; mean age of 47.3 years)...
	Proportion identifying as male gender	Seventy (62%) participants were female and 43 (**38**%) were male
	Proportion identifying as belonging to a specific ethnic group	**Latinos** accounted for **83.4**% (n = 371) of the participants
	Proportion belonging to specified individual income category	**15**% of participants have annual incomes of <**£10000**
	Proportion belonging to specified family or household income category	**15**% of participants had household annual incomes of <**£10000**
	Mean number of years in education completed	Participants had completed **10** years of education on average.
	Proportion achieved university or college	**60**% of participants had obtained university degrees.
	Proportion employed	In the intervention group, **75**% of participants were in paid employment.
	Aggregate relationship status	**60**% of participants reported being **single or never married**
	Proportion in a legal marriage or union	Most participants (**95**%) were married.
	Aggregate patient role	[...] a smoking cessation intervention for hospital **patients** with COPD.
	Aggregate health status type	[...] a smoking cessation intervention for hospital patients with **COPD**.
	Mean number of times tobacco used	Participants smoked on average **20** cigarettes per day.
Setting	Country of intervention	The intervention took place in 18 GP clinics in Greater Manchester, **UK**.
	Lower-level geographical region	[...] took place in 18 GP clinics in **Greater Manchester**, UK.
	Healthcare facility	[...] **health centre** within easy access of participant's homes.
	Hospital facility	Hospital inpatients were given brief advice at their **hospital** bedside
	Doctor-led primary care facility	The intervention took place in 18 **GP clinics** in Greater Manchester, UK.
Outcome behaviour	Smoking	We measured **smoking** cessation through a self-report questionnaire.
	Longest follow up	[...] smoking status at 1 month,[...], **12** month follow-up points.
	Self report	Smoking status was assessed via a **self-report** questionnaire
	Biochemical verification	Abstinence was defined as **expired CO** below 10ppm
	Outcome value	**54**% of participants were biochemically verified abstinent at 6 months [...]
Effect estimate	Odds Ratio	**Odds ratios** were calculated to test the effectiveness [...]
	Effect size estimate	The intervention was effective (OR **1.07**, (0.47, 0.9)
	Effect size p value	The intervention was effective (OR 1.07, (0.47, 0.9), p<**0.05**)
Delivery	Face to face	the three interventions consisted of ten 90-min **sessions**
	Distance	counselling included an initial intake and counselling **phone call**
	Printed material	All five **booklets** compared in this study were identical
	Digital content type	Patients also received [...] and a relaxation **audio tape**.
	Website / Computer Program / App	[...] plus access to a smoking cessation **website** [...]
	Somatic	Those who smoked were offered **nicotine replacement therapy**
	Patch	[...] in the form of the nicotine **patch**
	Pill	Participants began taking one **pill** (150-mg of bupropion SR or placebo)[...]
	Individual	Participants [...] received up to four **one-on-one** sessions [...]
	Group-based	All participants received 10 weeks of **group-based** CBT [...]
Source	Health Professional	All patients attended a 30-min individual counselling by the study **nurse**.
	Psychologist	Therapists were a male clinical **psychologist** [...]
	Researcher not otherwise specified	All instructions were provided by trained **research assistants** [...]
	Interventionist not otherwise specified	Two **patient navigators** received 10 hours [...]
	Expertise of Source	Counsellors were three **Master's-level** professionals
Reach	Individual-level allocated	Smokers (n = **94**) from 26 states [...]
	Individual-level analysed	Psychodrama group (n= **61**) Control group (n= 52).
Content - BCT	1.1.Goal setting (behavior)	During the counselling sessions, [...] solutions, **set a goal to quit** [...]
	1.2 Problem solving	[...] encouraged to **reflect on barriers to change and identify solutions**,
	1.4 Action planning	[...] **come up with a detailed action plan to help them quit**
	2.2 Feedback on behaviour	[...] GPs gave participants **feedback on their current smoking levels**
	2.3 Self-monitoring of behavior	[...] to **closely monitor their smoking** [...]
	3.1 Social support (unspecified)	During the **counselling** sessions, [...]
	4.1 Instruction on how to perform the behavior	In addition to being offered NRT and a **quit smoking self-help guide**, [...]
	4.5. Advise to change behavior	[...] participants were **advised to quit**,[...]
	5.1 Information about health consequences	[...] informed of the **negative health effects of smoking** ,[...]
	5.3 Information about social and environmental consequences	[...] **social impact of smoking**, and were informed [...]
	11.1 Pharmacological support	In addition to being **offered NRT** [...]
	11.2 Reduce negative emotions	[...] informed about meditation as a useful **stress-reduction** tool.

Table 1: Extracted entities grouped according to the higher level ontology classes with example annotation. Bold-face text represents the annotated text within its context (which is truncated to fit the table width).

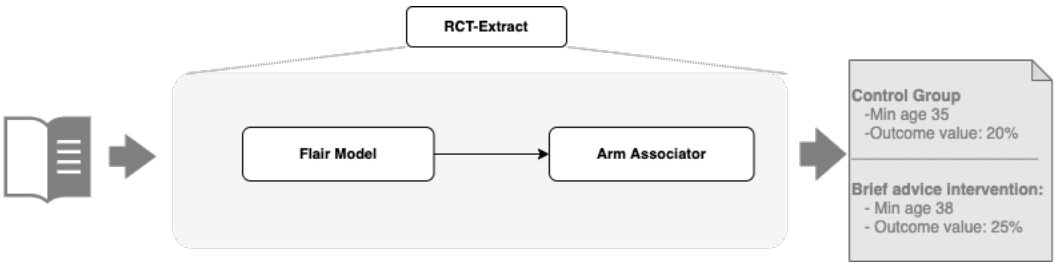

Figure 2: Knowledge extraction module workflow

(Table 1). One of the challenges is that 40% of the information to be extracted occurs in tables. Since tables usually exhibit diverse formatting structures across a collection, it is particularly difficult to correctly extract the values from them. To overcome this issue, we transform the content, structured in table format, into text. Specifically, to generate the training instances for our model, we first generate pseudo-sentences for each table element transforming every cell in a table in the following format: $< rowheader >$ *of* $< columnheader >$ *has a value of* $< cellvalue >$. We then map all annotated entities back to their contexts in the BIO tagging format.

During the inference stage, given a test document in PDF, we first extract all sentences from the document. We then augment this with the pseudo-sentences generated from each table. Next, we apply the trained model to these sentences (both original and pseudo) to extract the RCT entities.

We also extract arm names as one of the entities. Once the list of potential arm name mentions and the list of entities are extracted the *arm associator* module creates tuples of <*armName*,*entity*>, by detecting, for each entity, the closest instance of an arm name in a window of size t (t is set empirically). If no instance of arm name is found, the entity is associated to the *default arm*. The default arm indicates that the entity value is common to all arms in the study.

Since each arm can have many mentions in the RCT report, we need to cluster the arm name mentions and as a consequence the associated extracted entities. We use a complete-link clustering algorithm where the similarity of two clusters is the similarity of their most dissimilar members[35]. We cluster the different arm names into n classes with n corresponding to the number of arms. n is detected by exploiting the common pattern that authors often use to indicate the number of groups, e.g., $'into/in' + n + groups$ and extracting the value n corresponding to the number of arms. In the end, for each cluster, the more frequent arm name mention is elected as cluster label.

It is worth noting that associating predicted entities to the corresponding arm names is a very challenging problem. The task requires performing cross-sentence information extraction and inference. It is left to future work to explore novel arm association methodologies.

4.2 Knowledge Extraction Evaluation

In this section, we describe how we test the knowledge extraction module. We use the corpus described in Section 3 and focus on extracting the 57 entities mentioned there (plus the arm names). We split the corpus into training and test sets so that we can learn an extraction model on the training data and then test that model with unseen test data. We have 300 PDFs in training and 97 in test, while 10 reports were excluded that did not have relevant annotation[j]. After training the BiLSTM-CRF model on the 300 RCTs of the training set, we extract entities from the test set and evaluate whether the extracted entities are correct or not. We compare our model with an unsupervised rule-based baseline[36], which involves first, retrieving passages that are more likely to contain a relevant entity, and then extract the value within this passage using a series of ad-hoc extractors based on pattern matching and regular expressions. Our choice of baseline is motivated by previous work[37] showing that an unsupervised approach (i.e., without machine learning but using expert queries) performs well against supervised and semi-supervised approaches. Our goal in this work is not to optimize the machine learning approach but show that the system we propose, automatic information extraction feeding into a prediction system, is a viable alternative to handcrafted manual approaches.

[j]The dataset is available at `https://github.com/HumanBehaviourChangeProject/Info-extract/blob/master/HBCP-Corpus.zip` and more details about the dataset are provided by Bonin et al.[28]

	Rule-based			*Flair* (w/o Table)			*Flair* (w/ Table)		
Upper level ontology	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
Population	14.37	29.40	17.33	17.24	13.88	11.21	27.42	31.99	**25.51**
Content	41.37	83.33	53.04	60.61	55.89	53.62	57.97	59.00	**56.52**
Source	14.25	12.72	12.11	19.13	11.37	13.41	21.52	17.28	**17.86**
Delivery	17.89	24.75	15.74	55.14	38.30	37.25	50.10	48.08	**46.37**
Outcome Behaviour	22.18	26.65	23.93	43.00	53.48	44.43	37.63	64.62	**44.92**
Effect estimate	28.87	41.42	**33.86**	31.75	40.92	32.51	28.55	47.07	31.70
Reach	29.92	48.74	**37.07**	11.17	21.71	14.74	9.92	14.94	11.92
Overall	24.12	38.14	27.58	34.00	33.65	29.60	33.30	40.42	**33.55**

Table 2: Mean precision, recall, and F_1 per upper level ontology for the *Flair* model trained with (w/ Table) and without the pseudo-sentences from tables (w/o Table) compared with the unsupervised rule-based baseline. Macro-averages are used for precision, recall, and F_1 of presence value-type attributes.

Table 2 reports arm-agnostic performance, grouped by upper level ontology categories, for the baseline and the *Flair* approaches. For the *Flair* approach we report both results of the model trained with and without the table pseudo-sentences. For each value attribute, recall is calculated as the number of attributes where the target values are correctly predicted divided by the number of these annotated attributes. Precision is calculated as the number of predicted attributes with correct values out of all the ones predicted for this attribute. F_1 score is the harmonic mean of the precision and recall. The attributes listed in Table 1 are associated with different value types. For example, content entities (e.g., BCTs) are Boolean variables ($\{0, 1\}$) denoting the presence or absence of the entity in the arm, while the population and outcome behavior attributes are real numbers. The matching criteria for *true positives* therefore depends on the type of attribute, e.g., Boolean values must match true or false for presence, but mean age can match the real value to a given level of precision.

From Table 2, we observe that the supervised machine learning approach (*Flair*) outperforms the rule-based one in almost all the upper level categories (with some exceptions, such as Effect estimate and Reach, for which there exists only a small number of annotations). In addition, we notice that using the table contents for training the model improves the results. Specifically, table content makes a difference for those entities, such as outcome value or population, that are often reported in table format. As expected, recall is higher for *Flair* with table sentences because it can recover more annotation, but this is at the expense of precision, which is generally better without tables. Recall is better still with the rule-based approach, which could be altered to have even better coverage, however the current baseline strikes a balance between precision and recall.

The F_1 results, even with our state-of-the-art statistical approach remain fairly low. There are a few contributing factors, some of which we are addressing in future work. The entity extraction is a feasible, but still quite difficult task as evidenced by the IRR in Section 3. The heterogeneity of the entity representation make it difficult to learn without sufficient data. We also note that current results come from a unique trained model, but improvements are possible when training dedicated models. It remains the case that the supervised approach presents two main advantages *1)* it is more flexible and scalable, as in it can learn new context, given new documents, and *2)* it shows better results for the majority of the entities.

4.3 Knowledge Prediction Methodology

After extracting relevant information from each study arm of an RCT on smoking cessation, we use it as input to predict outcome values, i.e., the percentage of people who stopped smoking. The aim is to provide a user with an estimated prediction of the success of a set of BCTs on a specific population with a given set of outcome qualifiers, particularly useful for situations where no RCT with the given situation exists. To obtain the outcome value predictions, we employ a regression based approach, which models the outputs, $y \in \mathbb{R}$, as a function ϕ of input vectors $\mathbf{x}$ of feature values. Given a training set of feature vectors associated with the outcome values, the function ϕ can be estimated with standard approaches such as minimizing the hinge-loss (for SVM). Specifically, in our experiments, we employed SVM as our regression approach (i.e., Support Vector Regression (SVR)[38]) because it can handle any non-linearities to be introduced for estimating the regression coefficients. One of the rationales for employing feature-based models

Trained with	RMSE	MAE
Ground truth	11.99	8.53
RB extraction	12.96	9.03
Flair extraction	12.78	8.94

Table 3: Comparison of outcome prediction as a regression problem, using different training sets. For both RMSE and MAE, a lower value indicates a more effective outcome prediction.

in our study (over data-driven ones) is that, firstly, the feature-based models are more conducive to debugging and explanations (e.g., by observing what set of input features are strongly correlated with the output values), and secondly, they can be trained on small quantities of data (which is the situation in our case). We also experimented with other regression models but they were exhibiting the same trends as SVM, so, in this paper, we chose to focus on comparing configurations of the training data rather than of the machine learning algorithms.

A feature vector $\mathbf{x}$, in our regression experiments, is constructed with all the entities corresponding to one arm of one study. The entities can be numerical or categorical, in which case they are turned into numerical values, e.g., a BCT is treated as a categorical variable with two values, namely $\{0, 1\}$. To give specific details about the experiment setting, the total number of features (dimensions of the input vectors) is 168, whereas the total number of instances used in our experiments is 819, which we split into a 4:1 ratio of train and test, respectively. The outcome values, which correspond to the percentage of people who stopped smoking, were normalized in the range $[0, 1]$, the mean being 0.16. We conduct the regression experiments under three settings, each using a different training sets:

- with ground-truth input features and outcome values;
- with input features and outcome values automatically extracted using the baseline rule-based system (RB);
- with input features and outcome values automatically extracted using *Flair*.

The objective is to investigate to which extent the automatic extraction process (Section 4.1) affects the effectiveness of outcome value prediction. Predicting from automatically extracted input indeed corresponds to the realistic scenario where the prediction system can be regularly re-trained using the most recent publications in the field without going through a long and costly manual annotation process first.

4.4 Knowledge Prediction Evaluation

Table 3 presents the results of our experiments. We use standard metrics for regression: root mean square error (RMSE)[k] and mean absolute error (MAE)[l], both evaluated using the ground-truth outcome values of the test instances. In all cases, the SVM parameters are identical: the regularization parameter (c - higher values prevent over-fitting on the training set) and the kernel width (γ - lower values associate more importance to local effects in the input feature space) are both set to 0.1, determined by a grid-search we conducted using the ground-truth input and outcome values.

As expected, the best prediction is obtained by the model trained on the ground-truth entities, that uses manually annotated data. However, the performance gap between the model trained on ground truth data and the models trained on automatically extracted data is small. The only small increase in error for the systems trained on automatically extracted entities indicates that we can output predictions of comparable quality without needing to manually annotate an ever increasing quantity of new studies. We also note that the prediction systems trained on automatically extracted entities have a similar performance drop compared to ground-truth training, but we still note that the model trained on *Flair* extractions, which was our better approach in the task of extraction (Section 4.1), performs better than the rule-based baseline on the downstream task of outcome value prediction.

5 Conclusions and Future Work

In this paper, we presented a novel methodology based on information extraction and prediction models to extract relevant information from RCTs and predict potential outcome values on unseen setting scenarios in the domain of

[k]RMSE measures how spread out these residuals are with respect to the actual values.

[l]MAE measures the absolute value of the difference between the predicted and the ground-truth outcome values.

behavior change for smoking cessation. We showed that statistical information extraction approaches can be used to extract a wide variety of behavior change entities, with the advantage of being more adaptable to new contexts and new data than rule-based approaches. We also showed that the data automatically extracted can be used to train a prediction model for the identification of outcome values with similar performance to the one trained on manually-extracted entities. The work described in this paper is part of an ongoing effort, where we are focusing on improving the effectiveness of both components. Future works along this direction could involve improving the arm association, extracting multiple outcome values at different follow-up points and incorporating text information from documents to further improve outcome value prediction.

6 Acknowledgments

This work was supported by a Wellcome Trust collaborative award as a part of the Human Behaviour-Change Project (HBCP): Building the science of behavior change for complex intervention development (grant no. 201,524/Z/16/Z).

References

[1] Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLOS Medicine. 2010 09;7(9):1–6.

[2] Higgins JPT, Thomas J, editors. Cochrane Handbook for Systematic Reviews of Interventions. 2nd ed. Wiley Cochrane Series. Wiley-Blackwell; 2019.

[3] Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. Jama. 1999;282(7):634–635.

[4] Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ open. 2017;7(2):e012545.

[5] Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. The Lancet. 2014;383(9913):267–276.

[6] Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JP, Mavergames C, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS medicine. 2014;11(2):e1001603.

[7] Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JPA, et al. Biomedical research: increasing value, reducing waste. The Lancet. 2014;383(9912):101–104.

[8] Michie S, Johnston M. Optimising the value of the evidence generated in implementation science: the use of ontologies to address the challenges. Implementation Science. 2017;12(1):131.

[9] Hara K, Matsumoto Y. Extracting Clinical Trial Design Information from MEDLINE Abstracts. New Generation Computing. 2007 May;25(3):263–275.

[10] Kiritchenko S, de Bruijn B, Carini S, Martin J, Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Medical Informatics and Decision Making. 2010;10(56).

[11] Summerscales RL. Automatic Summarization of clinical abstracts for evidence-based medicine. Illinois Institute of Technology. Chicago, Illinois; 2013.

[12] Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support Evidence Based Medicine. BMC Bioinformatics. 2011 Mar;12(2):S5.

[13] Hansen MJ, Rasmussen NØ, Chung G. A method of extracting the number of trial participants from abstracts describing randomized controlled trials. Journal of Telemedicine and Telecare. 2008;14(7):354–358. PMID: 18852316.

[14] Hassanzadeh H, Groza T, Hunter J. Identifying scientific artefacts in biomedical literature: The Evidence Based Medicine use case. Journal of Biomedical Informatics. 2014;49:159–170.

[15] Ferracane E, Marshall I, Wallace BC, Erk K. Leveraging coreference to identify arms in medical abstracts: An experimental study. In: Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis. Austin, TX: Association for Computational Linguistics; 2016. p. 86–95.

[16] Gough D, Oliver S, Thomas J. An introduction to systematic reviews. Sage; 2017.

[17] Schünemann HJ, Moja L. Reviews: Rapid! Rapid! Rapid! ...and systematic. Systematic Reviews. 2015;4(4). Available from: https://doi.org/10.1186/2046-4053-4-4.

[18] Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic reviews. 2019;8(1):163.

[19] Jasch D. Information Extraction from Clinical Trials. Australian Institute of Health Innovation (AIHI). Sydney, Australia; 2016.

[20] Jin D, Szolovits P. PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks. In: Proceedings of the BioNLP 2018 workshop. Melbourne, Australia: Association for Computational Linguistics; 2018. p. 67–75.

[21] Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics. 2007;33(1):63–103.

[22] Michie S, West R, Finnerty A, Norris E, Wright A, Marques M, et al. Representation of behaviour change interventions and their evaluation: Development of the Upper Level of the Behaviour Change Intervention Ontology [version 1; peer review: awaiting peer review]. Wellcome Open Research. 2020;5(123).

[23] Wallace BC, Kuiper J, Sharma A, Zhu M, Marshall IJ. Extracting PICO Sentences from Clinical Trial Reports Using Supervised Distant Supervision. J Mach Learn Res. 2016 Jan;17(1):4572–4596. Available from: http://dl.acm.org/citation.cfm?id=2946645.3007085.

[24] Lin S, Ng JP, Pradhan S, Shah J, Pietrobon R, Kan MY. Extracting Formulaic and Free Text Clinical Research Articles Metadata Using Conditional Random Fields. In: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents. Louhi '10. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010. p. 90–95. Available from: http://dl.acm.org/citation.cfm?id=1867735.1867749.

[25] Lumley T. Network meta-analysis for indirect treatment comparisons. Statistics in medicine. 2002;21(16):2313–2324.

[26] Lehman E, DeYoung J, Barzilay R, Wallace BC. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 3705–3717.

[27] Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Annals of behavioral medicine. 2013;46(1):81–95.

[28] Bonin F, Gleize M, Finnerty A, Moore C, Jochim C, Norris E, et al. HBCP Corpus: A New Resource for the Analysis of Behavioural Change Intervention Reports. In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2020. p. 1967–1975.

[29] Krippendorff K. Content analysis: An introduction to its methodology Thousand Oaks. Calif: Sage. 2004;.

[30] Krippendorff K. Testing the reliability of content analysis data. The content analysis reader. 2009;p. 350–357.

[31] GROBID. GitHub; 2008–2020. https://github.com/kermitt2/grobid.

[32] Hou Y, Jochim C, Gleize M, Bonin F, Ganguly D. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 5203–5213. Available from: https://www.aclweb.org/anthology/P19-1513.

[33] Akbik A, Blythe D, Vollgraf R. Contextual String Embeddings for Sequence Labeling. In: COLING 2018, 27th International Conference on Computational Linguistics; 2018. p. 1638–1649.

[34] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1532–1543.

[35] Jain AK, Dubes RC, et al. Algorithms for clustering data. vol. 6. Prentice hall Englewood Cliffs; 1988.

[36] Ganguly D, Deleris LA, Mac Aonghusa P, Wright AJ, Finnerty AN, Norris E, et al. Unsupervised Information Extraction from Behaviour Change Literature. Studies in health technology and informatics. 2018;247:680–684.

[37] Ganguly D, Hou Y, Deleris LA, Bonin F. Information Extraction of Behavior Change Intervention Descriptions. In: Proceedings of AMIA Joint Summits on Translational Science. American Medical Informatics Association; 2019. p. 182–191.

[38] Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support Vector Regression Machines. In: Proceedings of the 9th International Conference on Neural Information Processing Systems. NIPS'96. Cambridge, MA, USA: MIT Press; 1996. p. 155–161.