Marseille, 11–16 May 2020

HBCP Corpus: A New Resource for the Analysis of Behaviour Change Intervention Reports

Francesca Bonin*, Ailbhe N. Finnerty†, Candice Moore†, Charles Jochim*, Emma Norris†, Yufang Hou*, Martin Gleize*, Debasis Ganguly*, Alison J. Wright†, Emily Hayes†, Silje Zink†, Alessandra Pascale*, Pol Mac Aonghusa*, Susan Michie†

*IBM Research, † UCL

* Dublin, Ireland, † London, UK

{fbonin, charlesj, yhou, martin.gleize,debasis.ganguly1,apascale,aonghusa}@ibm.ie

{a.finnerty, candice.moore.11, emma.norris, alison.j.wright, emily.hayes.16, s.zink, s.michie}@ucl.ac.uk

Abstract

Due to the fast pace at which research reports in behaviour change are published, researchers, consultants and policymakers would benefit from more automatic ways to process these reports. Automatic extraction of the reports' intervention content, population, settings and their results etc. are essential in synthesising and summarising the literature. However, to the best of our knowledge, no unique resource exists at the moment to facilitate this synthesis. In this paper, we describe the construction of a corpus of published behaviour change intervention evaluation reports aimed at smoking cessation. We also describe and release the annotation of 57 entities, that can be used as an off-the-shelf data resource for tasks such as entity recognition, etc. Both the corpus and the annotation dataset are being made available to the community.

1. Introduction

There has been significant growth in the number of scientific publications in many disciplines in recent times (Larsen and Ins, 2010) and researchers can have difficulty keeping track of the state of the art in many fields. This is also true for behaviour change researchers, health professionals and consultants that explore the literature of behaviour change intervention reports, in order to understand the most effective methodology (or intervention) to help a certain population improve a specific target behaviour (for example, stopping smoking). The volume and rate at which research is produced about behaviour change is beyond the capability of human researchers to compare and understand which interventions are most effective and to be able to generalise the results to varying populations in different contexts (Michie and Johnston, 2017). More evidence is produced and published than it is possible for researchers to be able to use, synthesise and analyse effectively with current conventional research methods and the current waste in research is being increasingly recognised (Glasziou et al., 2014; Elliott et al., 2014; Macleod et al., 2014).

Systematic reviews seek to collate evidence that fits prespecified eligibility criteria in order to answer a specific research question (Higgins and Thomas, 2019). An ongoing challenge is the time taken to complete them, estimated at $\sim$ 1,000 hours of highly skilled manual work (Allen and Olkin, 1999) or 67 weeks (Borah et al., 2017), from preregistration stage to publication. It is clear from these studies that the current method of conducting systematic reviews is not sustainable. For this reason both domain experts in behaviour change and policymakers would benefit from automatic ways to analyse the literature.

Many technologies or tools have been developed to automate the process of conducting systematic reviews. However, it is not always clear to researchers how they should be used and at what stage of the reviewing process. A review of the current tools available, which are listed on SR Toolbox [1], a publicly available online catalogue of software tools to aid the production of systematic reviews, has produced a practical guide of how and when it is appropriate to use these tools to speed up the reviewing process (Marshall and Wallace, 2019).

Behaviour change literature analysis involves the synthesis of scientific publications that describe interventions which have been tested on select samples of the population. Analysing those reports means first extracting the relevant information with Information Extraction techniques, and then predicting or generalising from the information. Some effort in developing automatic ways to analyse behaviour change intervention evaluation reports has been conducted in the NLP and health communities, by applying rule-based (Hara and Matsumoto, 2007; Kiritchenko et al., 2010), or machine learning approaches (Summerscales, 2013; Kim et al., 2011; Hansen et al., 2008; Hassanzadeh et al., 2014) to the automatic extraction of information from evaluation reports. However, all these studies are difficult to compare due to the lack of a common benchmark dataset, and the diversity of the entities extracted.

To the best of our knowledge, no unique open access resource exists at the moment that is freely available and ready to use for advancing the research in the automatic analysis of behaviour change intervention evaluation reports.

The purpose of this paper is to present the community with a new corpus created with the help of behaviour change domain experts which aims to represent a standard for the future research on behaviour change interventions. With this corpus we intend to help the community in fostering research in the following areas: *1)* information extraction from behaviour change intervention reports, that represent

[1] http://systematicreviewtools.com/about.php.

a very complex and domain specific linguistic genre, *2)* intervention and outcome prediction, *3)* development tools for automatic systematic reviews.

We also provide an extensive annotation dataset of entities that domain experts have selected as the most relevant information in the study. Such annotations are used to guide the development and understanding of the Behaviour Change Intervention Ontology (BCIO)[2] (Michie et al., 2017; Norris et al., 2019).

In the rest of the paper, we describe the collection of the corpus of 407 PDFs, from literature in smoking cessation behaviour change. We call this the **HBCP-corpus**. We also describe the annotation of the HBCP-corpus for 57 relevant entities for behaviour change analysis and prediction. We describe the criteria for the data collection, the annotation scheme and annotation process. We are releasing two resources to the community:

1. A subset of the HBCP-corpus, comprising 97 open access papers, OA-HBCP.[3]

2. A dataset built on the OA-HBCP-corpus that contains annotation for 57 entities considered of importance for the analysis of behaviour change intervention reports. We call this the OA-HBCP-NE-dataset. Such entities, such as the population characteristics, the outcome of the study, the behaviour change technique (BCT) applied, and their annotation process, will be described in detail in the rest of the paper.

The corpus is released in JSON format.[4] The dataset is released in a CoNLL-like format with entities annotated using BIO labels. Both resources will be made available on the HBCP website[5] and will be added to LRE Map.

This work is conducted in the context of the Human Behaviour-Change Project (Michie et al., 2017), that aims to build a knowledge system of automatic meta-analysis of behaviour change evaluation reports.

2. Background

Many global threats to human health and well-being can only be solved by people, organisations and governments changing their behaviour. This includes behaviours directly relevant to health but also behaviours of policymakers and providers responsible for promoting health and delivering healthcare. Behaviour change researchers use Behaviour Change Interventions (BCIs) to improve how people, organisations and governments behave in a particular target behaviour. BCIs are policies, activities, services or products designed to cause people to act differently from how they would have done otherwise. Interventions involve attempting to change either members of the target population or their social or physical environment. They are constituted by a combination of Behaviour Change Techniques (BCTs). Intervention evaluation reports addressed in this work are scientific publications that provide invaluable knowledge to help with developing or selecting BCTs, once they are analysed and synthesised.

The typical behaviour change intervention is quite complex, constituted by the comparison of the application of different BCTs on a target population to solve a target behaviour (e.g., smoking cessation). Usually in a report, several study groups (interventions and control groups), on which the different BCTs are tested, are compared. Those groups are called *arms*. Each study then compares the outcome and effect sizes of the application of specific BCTs on the arms of the study.

3. Related Resources

The availability of open data in health, behavioural and social sciences is very sparse and part of an ongoing challenge in the field. This is true even though there is a movement towards open data (Munafò, 2016) and there is agreement that the availability of data would allow results to be reproduced, giving more confidence to patients, practitioners and policy advisors (Naudet et al., 2018). While one reason is the amount of time and effort needed to make data accessible, another reason is the fear of repercussions and retraction. According to (Packer, 2018), many researchers, who would be open to sharing data, fear being shamed if errors were found in the data or in the published results.

Resources. Nye et al. (2018) present a corpus of 5,000 richly annotated abstracts of medical articles describing clinical randomized controlled trials. Annotations include demarcations of text spans that describe the patient Population enrolled, the Interventions studied and to what they were Compared, and the Outcomes measured (the 'PICO' elements). Differently from our corpus, in (Nye et al., 2018) the annotation is conducted only on the abstracts, and, after the first pilot, all the labelling is crowdsourced. The HBCP-corpus presents annotation for the entire article and each article has been annotated by two domain experts (behaviour scientists). This ensures the quality of the annotation in a very complex task.

The Behaviour Change Technique Taxonomy v1 (BCTTv1) website[6] provides a resource which consists of the metadata of a corpus of 405 published reports coded with BCTs. Thirteen percent of the reports in the corpus are systematic reviews and 87% are interventions, covering a wide range of behaviours. However, only metadata is available to the public.

4. Data Collection

The intervention evaluation reports which were selected to create the corpus were taken from a variety of sources such as the Cochrane Library[7] and the IC-SMOKE project.[8] A

[2]Details on the ontology development summary can be found at `https://osf.io/86m75/`

[3]Due to copyright restriction we can only release the open access subset of the corpus

[4]Names of authors are removed for GDPR compliance.

[5]`https://www.humanbehaviourchange.org/`. For reviewing purposes the data can be found at: `https://github.com/HumanBehaviourChangeProject/Info-extract/blob/master/HBCP-Corpus.zip`

[6]https://www.bct-taxonomy.com/interventions

[7]`https://www.cochranelibrary.com/`

[8]`https://osf.io/23hfv/`

full list of where the papers came from can be found online.[9]

The main resource for the corpus were systematic reviews from the Cochrane Library. The library was searched for Cochrane systematic reviews on smoking cessation and all reviews were considered for inclusion in the corpus. Systematic reviews report the entire list of included studies in the review and also a list of relevant but not included studies. Studies from which outcome data can reliably be extracted are included in a meta-analysis.[10] From every systematic review we selected only those studies that were included in the meta analysis. This was done to allow us to eventually compare the results of any automated meta-analysis system to the ground truth results, which were produced from the review.

A second source of papers, IC-SMOKE, is a systematic review project of behavioural smoking cessation trials, which is funded by Cancer Research UK.

We used the following criteria for the selection of the papers:

- They are randomised control trials (RCT);
- They are included studies in a systematic review on smoking cessation;
- They are included in a meta-analysis in a systematic review on smoking cessation;
- They have a behavioural outcome value at a predefined follow up time point (in the case of smoking cessation, the percentage of participants who stopped smoking).

Once appropriate reviews and included studies were identified, the papers were uploaded to EPPI Reviewer (Thomas et al., 2010) software[11] for text extraction.

Some of these papers were eventually removed from the corpus as they were found to have issues that caused noise in the data. These included PDFs which could not be annotated correctly, (e.g., the highlighting functionality would not capture the section of text accurately) or could not be correctly processed by PDF extractors. The removal of these papers resulted in a cleaner annotation dataset.

As a result of this process, we created a full HBCP corpus of 407 papers from all relevant sources, including 120 reports from the IC-Smoke project and 287 reports from 15 systematic reviews from the Cochrane Library. At the moment we are releasing a subset of this corpus that comprises only Open Access papers which we have called OA-HBCP corpus. The OA-HBCP corpus is made up of 97 papers.

5. Annotation

5.1. Annotation Scheme

Our aim during annotation was to capture the most relevant features of smoking behaviour change intervention RCTs for predicting intervention effectiveness from published behaviour change intervention reports.

In order to achieve this, we created an annotation scheme based on the Behaviour Change Intervention Ontology (BCIO), developed as part of the Human Behaviour-Change Project, and the previously developed Behaviour Change Techniques (BCT) Taxonomy (Michie et al., 2013) to provide a structured classification of terms relevant to interventions.

The current full version of the ontology consists of hundreds of entities and it is divided in two levels: an upper-level and a lower-level.

The lower-level entities were selected to a more granular level using a top-down process, searching for key relevant terms in other classification systems or ontologies, and a bottom-up, annotation, process. The annotations were used as a bottom-up process by the researchers to further develop the lower-levels of the BCIO, by categorising data as it was described in reports. The lower-levels of the ontologies were used as the annotation scheme. This annotated data was double coded for quality and makes up the content of the dataset.

All the entities in the lower-level ontology can be grouped according to the following upper-level entities:

- **Population:** Aggregate of people whose behaviour an intervention is intended to change, in our case to stop smoking (for example, *women between 18 and 35 years old*).
- **Setting:** Aggregate of entities in the social and physical environment in which the intervention takes place (for example, *schools in the UK*).
- **Outcome Behaviour:** The behaviour that the intervention is targeting, in our case smoking, and its measurement, including timing of measurement, and type of measurement taken, (for example, *self-report* vs *biochemically verified*).
- **Estimated Effect:** The estimated difference in outcome behaviour between an intervention scenario and a comparator scenario (for example, *a control group*). This includes the type of statistic used to represent the difference (e.g., *odds ratio*), its value (e.g., effect size estimate *1.35*), and a measure of significance (e.g., p-value $p < 0.05$).
- **Source:** Who delivers one or more BCTs or who provides the target population with the behaviour change intervention materials that contain one or more BCTs (for example, *personnel delivering the intervention*, *a nurse*).
- **Delivery:** Method or methods by which the content is brought to the Population; includes mode of delivery, timing and manifested characteristics of the Source (for example, *face to face group counseling*, *stop smoking booklet*).
- **Reach:** That proportion and nature of the population that encounters an intervention in a behaviour change

[9] https://osf.io/myje6/

[10] A meta-analysis is a statistical approach used by researchers to generalise the results of a small number of studies

[11] http://eppi.ioe.ac.uk/eppireviewer4/

intervention scenario. (for example, *436 patients were randomly allocated to the intervention group*)

- **Content - BCTs:** Those parts of a behaviour change intervention that can be classified into specified behaviour change techniques and as such are intended to be active in bringing about behaviour change (for example, *giving information about the health consequences of the behaviour*).

We selected a focused prioritised list of $n = 57$ entities from the BCIO. They were selected because they 1) occurred most commonly in papers, 2) were included in other relevant ontologies such as PICO[12], or 3) were believed to be most relevant for predicting intervention effectiveness. Table 1 provides a complete list of the entities, grouped per upper-level class, with some annotation examples. The released dataset includes annotations for these 57 key entities.

The entities fit broadly into four data types which guide the style of the annotation:

1. **Binary:** These annotations are intended to indicate the presence of a particular entity in relation to the intervention. For example, the annotation of a mention indicating *goal setting*, indicates the presence of this BCT (1.1 Goal Setting) in the intervention.
2. **Numerical:** Numerical annotations assign a particular numerical value to an entity e.g. Mean Age = 45.
3. **Text:** Text annotations give additional information in relation to a particular entity. For example, a general text description of training given to interventionists (GP's received 100 hours of training, plus ongoing supervision) is annotated for the entity "Expertise of Source".
4. **Attribute-Value Pair:** These annotations are in the form multiple *attribute:value* and are used to assign values to lower-level categories of an entity which require the annotation of the label of the value, as there can be more than one label for these entities. For example, for the entity "Aggregate Relationship Status", an annotation would be *single:26%, divorced:15%*

5.2. Annotation Guidelines

We report here the general annotation guidelines that apply to all 57 entities or sub-groups of them. Due to the complexity of the task, we developed general rules that apply to all the entities and specific annotation rules for each entity, that are reported in an annotation guidance manual. Before annotating a report, annotators were required to read the abstract, the last paragraph of the introduction (where the overall aims of the study are usually reported), the methods, and results sections of the paper. This is because many extracted features are complex and their interpretation depends on features reported elsewhere in the paper. For example, BCTs maybe be reported in a paper but if they are not directed towards the target behaviour of the intervention they should not be annotated.

[12] https://linkeddata.cochrane.org/pico-ontology

Objectivity of the mention. Annotations should never be taken from the discussion section of a report as this typically describes the authors interpretation of the study or results.

Mention and context. Each mention, of a word, phrase, or numerical value relevant to the entity to which it is assigned, should be highlighted (the nature of the mention depends on the type of entity as in Table 1). At the same time, annotations should also be assigned a 'context'. For annotations found within the main body of text of papers, the context should be the full sentence from which the annotation was extracted. For annotations taken from tables, the context should be three rows (the row above the annotation, the row including the annotation, and the row below the annotation).

Annotating multiple mentions in EPPI. Where more than one annotation is assigned to an entity, annotators should take the context for each annotation, and separate the contexts with ';;;'.

Arms. For each paper, the intervention 'arms' or conditions (e.g., intervention, control) should be annotated. The words used to describe the arms within the paper (e.g., "Intervention Group", "CBT Counselling Group") should be highlighted, and the context included. All annotations are assigned to a study arm, unless the annotation applies to all arms in the study, in which case, they are assigned to "Whole Study". For example, if the CBT Counselling Group had 101 participants, *101* should be annotated and assigned to the *CBT Counselling Group* arm for the entity *Reach - Analysed*.

Annotating Numerical Entities. Annotations of numerical entities should include only the number and not associated symbols (e.g., %, £, +) or units (e.g., weeks, months, mg). These symbols or units should be included in the context. If multiple numbers are included in a numerical value (e.g., a range of values) only the first number should be annotated (e.g., if the age of participants is reported as 10-15, annotate *"10"*).

Attribute:value annotations. For some more complex entities, annotations are in the attribute:value form to allow values to be associated with lower-level categories for an entity. For example, annotations for "Proportion identifying as belonging to a specific ethnic group" should consist of text indicating the ethnic group and the value (e.g., *50%* of participants identified as *White*, whilst *20%* of participants identified as *Black*) will be annotated as *50:White, 20:Black*.

5.3. Annotation Process

Annotators selection. Annotation requires a level of expertise in interpreting behaviour change intervention reports so only annotators with (at minimum) an appropriate masters in psychology or a related discipline were recruited.

Annotation Process. The lower-level entities of the BCIO were used as an annotation scheme and were imported as a hierarchical classification *Codeset* into EPPI Reviewer.

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 [...]
Estimated Effect	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 behaviour** [...]
	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. In bold the mention annotation in its context. Context has been truncated at times, due to space restrictions.

During the annotation process each mention of an entity in the text is assigned to the entity code, specified in the codeset, along with the context (sentence or rows of a table) surrounding the specific mention as in Figure 1.
Seven researchers in total, with expertise in health, cognitive psychology and behaviour change, were involved in annotating the reports for this dataset and each report was independently manually annotated by two annotators. The pairs of annotators were varied for each group of 10-15 papers to ensure consistency in annotations across all reports annotated and to minimise any inconsistencies in the annotations between the different pairs of annotators.
The researchers annotated the full text reports in stages rather than annotating all entities at once, as the lower-levels were developed at a different pace depending on the granularity required for annotation. Entities belonging to the same upper-level entity were annotated at the same time e.g., all Population level entities were annotated together, while other ontologies were in development and only used to annotate once they had been sufficiently completed.
A log of annotation issues was maintained throughout the process where annotators could log any questions they had while annotating or reconciling in their pair. Regular meetings were held to discuss the issues which led to more rules being made around how to annotate specific entities and updates to the annotation guidance manual.

Reconciliation. Reconciliation was done after groups of approximately 10-15 reports were independently annotated, as this was a manageable number of issues for discussion between the pairs. Within the EPPI Reviewer software it is possible to view and compare the annotations of two annotators, as shown in Figure. 2. Using this functionality the pairs of annotators compared their annotations in two ways; first, they compared whether they had both selected the same entity and second, if they agreed on the text they chose for this entity. During the reconciliation process the annotators **1)** decide on which version of coding is to be accepted as complete based on which coding is the more "correct" version, **2)** any changes that need to be made to the completed coding to ensure no data is missing from the final record and **3)** any questions or issues they had in applying the entity as codes to the reports.
Double coding. Double coding is done for every report to ensure that the data extracted is always checked and verified by two expert annotators. This is necessary due to the complex nature of the data which can be subjective and require a judgement to be made by the annotator, that lead to differences in the text assigned to an entity. The reconciliation process allows the two annotators to compare their annotations and make a judgement of which is the most correct version, this version is then manually corrected for any missing or incorrect data. Only the correct coding record is produced to be included in the dataset to ensure a high quality of the annotations.
The reconciliation process also allows the annotators to raise any issues they encountered with the specific set of papers and to bring these issues to the team. The log of annotation issues which arose is maintained and the issues are resolved through team discussions. These annotation issues and the team discussions primarily led to specific rules being made with regards to how annotations are done as a general rule and for specific entities. The updates to these rules led to changes in the annotation guidance manual to refine the guidance and improve the quality of annotations. Once all the papers in the database have been annotated and reconciled the data is exported from the software as a JSON file.

5.4. Annotation Agreement

To ensure the dataset is of high quality, we needed to ensure that the data which was extracted was done so reliably. To assess this we calculated the Inter-Rater Reliability (IRR) for pairs of coders for specific groups of papers annotated.
The process of assessing reliability was done for two purposes **1)** to assess the reliability of the ontologies which were used as coding scheme and **2)** to asses the reliability of the annotations.
Krippendorf's alpha (Krippendorff, 2004) was chosen as the measure to calculate reliability and was used to quantify the extent of "agreement" (observed disagreement/expected disagreement) between the raters.
The agreement was calculated on binary data extracted from the JSON reports where "1" represented the presence of an entity and "0" represented the absence of an entity. The coders "agreed" when they both detected the presence or noted the absence of an entity in the report. They "disagreed" if only one coder detected the presence of an entity in the report and the second did not. We calculated agreement on the binary detection data and not on the selected text and its boundaries.
For the purpose of **1** (assessing the reliability of the ontology as a coding scheme), we recruited external annotators, that have not participated in the ontology development but who were familiar with the annotation software and behaviour change interventions. They annotated 50 papers which were randomly selected from a database of $\sim$ 200 papers which were from the BCT portal[13] and coded for BCTs using the BCT taxonomy (Michie et al., 2013). The papers were based on interventions of a wide variety of behaviours not just smoking cessation, such as, medication adherence, alcohol use, healthy eating etc. Only Population and Setting have been completed at the moment with an external IRR of α=0.85 and α=0.61 respectively.
For the purpose of **2** (assessing the reliability of the annotators), we calculated IRR for annotations done by pairs of annotators involved in the development of the ontology. Specifically, we calculated the IRR for 83 *new* papers which were not annotated as part of the ontology development. The IRR was calculated for these new papers to assess the reliability of annotators in applying the codeset, comprising the 57 entities included in this dataset, to smoking cessation reports. The IRR results for these entities averaged across their respective ontologies can be found in Table 2.
While the agreement for some upper-level entities is low, the overall agreement for these entities was α=0.74, which is above the acceptable threshold of agreement (α=0.67) recommended by Krippendorf (Krippendorff, 2009). The IRR is used here to identify problematic entities which have

[13] http://www.bct-taxonomy.com/interventions

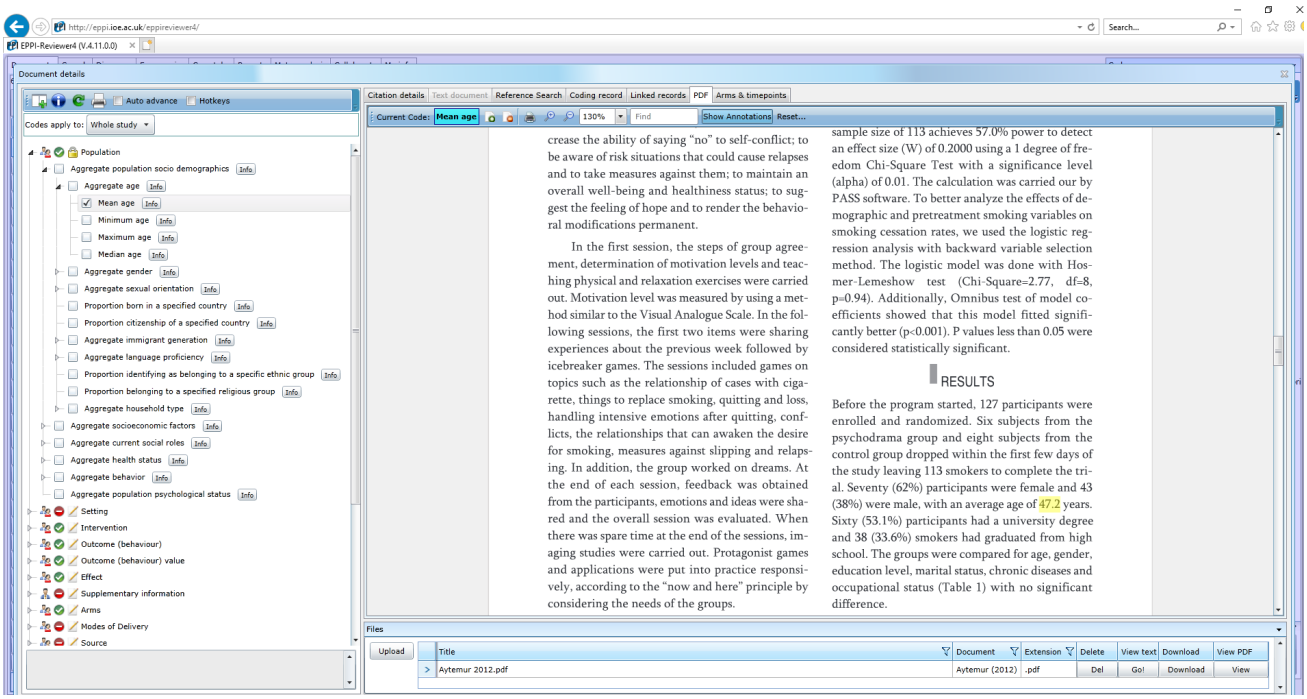

Figure 1: Example of an annotation in EPPI Reviewer

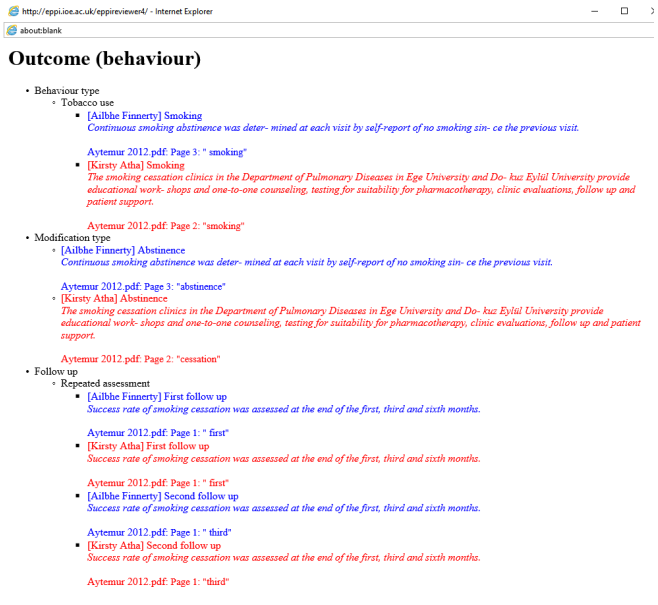

Figure 2: Comparison of two annotators coding records on EPPI

caused difficulty for the annotators as they are either not well reported in the papers or are not clearly defined in the guidance. Using these values we made significant improvements to the annotation guidance by clarifying terms and definitions and using more examples of text annotations. *Reach* for example showed to be a very challenging annotation group; this does not affect the annotation released as double coded has been done and disagreement resolved before release.

To measure the agreement between pairs of researchers and to determine if it would improve over time we compared the alpha scores for annotated data from the entities at two

Upper-Level Entity	***Krippendorf's alpha***
Population	0.67
Setting	0.66
Outcome	0.37
Estimated Effect	0.49
Delivery	0.62
Source	0.77
Reach	-0.14
Content - BCTs	0.44

Table 2: Inter-rater Reliability Scores for Groups of Entities

time points. At Time 1 (**T1**) 185 reports were annotated and at Time 2 (**T2**) 80 different reports were annotated, for the same entities. The results of the comparison found that the overall agreement increased from α=0.60 at T1 to α=0.74 at T2. 87% of the individual entities had improved alpha scores from T1 to T2, with an average improvement of α=0.36, showing that clear and concise guidance reduces discrepancies between the annotators.

The process of calculating IRR was automated as it is time consuming to extract the data and manually compute IRR. The scripts which are used to automate the IRR calculation process can be found on the GitHub repository.[14]

6. Resource Description

We release two resources: OA-HBCP corpus, and a CoNLL format of the annotations for the presented entities of the OA-HBCP-corpus, the OA-HBCP-NE-Dataset.

OA-HBCP corpus: The corpus is released in JSON format, one file per paper. Each file is the output of a

[14] https://github.com/HumanBehaviourChangeProject/Automation-InterRater-Reliability

```
Redding_2015.pdf        BA      NN      B-Expertise_of_Source
Redding_2015.pdf        or      CC      I-Expertise_of_Source
Redding_2015.pdf        MA      NN      I-Expertise_of_Source
Redding_2015.pdf        level   NN      I-Expertise_of_Source
Redding_2015.pdf        counselors      NNS     B-Health_Professional
```

Figure 3: Extract from the OA-HBCP-NE-Dataset

PDF parser,[15] formatted to be both human and machine-readable. A few useful metadata are available, such as the title of the paper and its introduction. Then the general representation for the content is a succession of *TEXT* and *TABLE* elements. The former is simply a string of text in the paper, the latter is a table as a list of its cells with their row and column headers. More details are available in the corresponding README for this resource.

OA-HBCP-NE-dataset: The dataset comprises of 3173 contexts (of one or more sentences each), for a total of 91,097 tokens. The file has four TAB separated columns, with the following header: <PDF-file-name,token,PoS tag,BIO tag>, as in shown in Fig. 3.

```
Redding_2015.pdf BA NN
B-Expertise_of_Source
Redding_2015.pdf or CC
I-Expertise_of_Source
Redding_2015.pdf MA NN
I-Expertise_of_Source
Redding_2015.pdf level NN
I-Expertise_of_Source
Redding_2015.pdf counselors NNS
B-Health_Professional
```

7. Experiments and Preliminary Results

Some preliminary experiments have been conducted with the released dataset. We frame the experiment as an information extraction task where we aim to automatically extract a subset of the entities. We use a baseline system based on a hybrid information retrieval based unsupervised approach and rule-based approach. First, we identify the passages that are more likely to contain an entity or an arm name with the unsupervised approach described in (Ganguly et al., 2019). Once the passages are selected, the system uses a combination of modules to detect entities and arms and associate entities to arms. The results in Table 3 show precision, recall, and F_1 for the extraction of a subset of the more frequently annotated entities.

As mentioned in Section 5.1., entities are associated with different value types: binary, numerical, text, and attribute:value pairs. The matching criteria for *true positives* is therefore dependent on the type of entity, e.g., binary values must match true or false for presence, but numerical can match the real value to a given level of precision. Binary *presence* attributes (i.e., BCTs) and real value attributes also differ in the information retrieval passage size used by our unsupervised approach. BCTs are found with shorter passage windows, while mentions of, for example, *mean age* require longer passage windows. Passage window sizes for the results in Table 3 are 10 for binary types and 50 for the others.

As the results show, some of the entities are detected with high to fair F1, as is the case for a lot of the relevant information extracted from the reports, such as, *odds ratio*, *aggregate patient role*. Interestingly, some of the entities are particularly challenging, specifically *outcome value*. The reason being that in a paper, numerous mentions of outcome value are reported, together with several other measures of outcome, but only one (the most significant for the domain expert) is annotated. This task, as shown in Table 2 is challenging also for human annotators, evidenced by low alpha scores, who reconcile their annotations manually. More annotations and new algorithms are under development in an attempt to improve the scores.

Entity	Prec	Rec	F_1
Mean Age	25.7	26.8	26.2
Prop. female	84.2	35.6	50.0
Prop. male	57.9	25.0	34.9
Prop. ethnic group	40.0	43.5	41.7
Prop. achieved uni.	45.5	21.7	29.4
Prop. employed	100.0	38.9	56.0
Agg. patient role	51.0	80.6	62.5
Mean tobacco used	19.2	8.9	12.2
Country and lower-level geographical region	25.6	28.9	27.2
Smoking	79.7	61.8	69.6
Longest follow up	24.1	16.7	19.7
Self report	19.2	32.3	24.1
Biochemical verification	35.6	42.9	38.9
Outcome value	30.5	20.9	24.8
Odds Ratio	54.3	92.6	68.5
Effect size p value	21.7	19.6	20.6
1.1.Goal setting (behaviour)	71.4	83.3	76.9
1.2 Problem solving	54.8	88.9	67.8
1.4 Action planning	15.4	88.9	26.2
2.2 Feedback on behaviour	26.3	83.3	40.0
2.3 Self-monitoring	23.1	37.5	28.6
3.1 Social support	72.6	93.8	81.9
4.1 Instruction	50.8	80.5	62.3
4.5. Advise to change behavior	44.4	57.1	50.0
5.1 Information health consequences	36.4	91.4	52.0
5.3 Information social and environmental consequences	42.9	57.7	49.2
11.1 Pharmacological support	66.7	91.7	77.2
11.2 Reduce negative emotions	37.5	60.0	46.2

Table 3: Results with unsupervised extraction.

8. Conclusion

In this paper we have described the construction of a corpus of behaviour change intervention evaluation reports and the annotation of 57 entities relevant to the behavioural scientist domain experts. Annotations guidelines and scheme have been described. All the papers are annotated by two expert annotators and reconciliation is performed. Both the corpus and the annotation dataset are being made available to the community.

9. Acknowledgements

This work was supported by a Wellcome Trust collaborative award as a part of the Human Behaviour-Change Project (HBCP): Build-

[15] A modified version of GROBID (GRO, 2008 2019), available at https://github.com/IBM/science-result-extractor

ing the science of behaviour change for complex intervention development (grant no. 201,524/Z/16/Z).

10. Bibliographical References

Allen, I. E. and Olkin, I. (1999). Estimating time to conduct a meta-analysis from number of citations retrieved. *Jama*, 282(7):634–635.

Borah, R., Brown, A. W., Capers, P. L., and Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7(2):e012545.

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P., Mavergames, C., and Gruen, R. L. (2014). Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, 11(2):e1001603.

Ganguly, D., Hou, Y., Deleris, L. A., and Bonin, F. (2019). Information extraction of behavior change intervention descriptions. In *Proceedings of AMIA Joint Summits on Translational Science*, pages 182–191. American Medical Informatics Association.

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., Michie, S., Moher, D., and Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*, 383(9913):267–276.

(2008 — 2019). Grobid. `https://github.com/kermitt2/grobid`.

Hansen, M. J., Rasmussen, N. Ã., and Chung, G. (2008). A method of extracting the number of trial participants from abstracts describing randomized controlled trials. *Journal of Telemedicine and Telecare*, 14(7):354–358. PMID: 18852316.

Hara, K. and Matsumoto, Y. (2007). Extracting clinical trial design information from medline abstracts. *New Generation Computing*, 25(3):263–275, May.

Hassanzadeh, H., Groza, T., and Hunter, J. (2014). Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of Biomedical Informatics*, 49:159–170.

Julian P. T. Higgins et al., editors. (2019). *Cochrane Handbook for Systematic Reviews of Interventions*. Wiley Cochrane Series. Wiley-Blackwell, 2nd edition.

Kim, S. N., Martinez, D., Cavedon, L., and Yencken, L. (2011). Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*, 12(2):S5, Mar.

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

Krippendorff, K. (2004). Content analysis: An introduction to its methodology thousand oaks. *Calif.: Sage*.

Krippendorff, K. (2009). Testing the reliability of content analysis data. *The content analysis reader*, pages 350–357.

Larsen, P. and Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84:575–603, 09.

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A., Salman, R. A.-S., Chan, A.-W., and Glasziou, P. (2014). Biomedical research: increasing value, reducing waste. *The Lancet*, 383(9912):101–104.

Marshall, I. J. and Wallace, B. C. (2019). Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8(1):163.

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

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M. P., Cane, J., and Wood, C. E. (2013). 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*, 46(1):81–95.

Michie, S., Thomas, J., Johnston, M., Mac Aonghusa, P., Shawe-Taylor, J., Kelly, M. P., Deleris, L. A., Finnerty, A. N., Marques, M. M., Norris, E., et al. (2017). The human behaviour-change project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. *Implementation Science*, 12(1):121.

Munafò, M. (2016). Open science and research reproducibility. *ecancermedicalscience*, 10.

Naudet, F., Sakarovitch, C., Janiaud, P., Cristea, I., Fanelli, D., Moher, D., and Ioannidis, J. P. (2018). Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in the bmj and plos medicine. *bmj*, 360:k400.

Norris, E., Finnerty, A. N., Hastings, J., Stokes, G., and Michie, S. (2019). A scoping review of ontologies related to human behaviour change. *Nature human behaviour*, 3(2):164.

Nye, B., Li, J. J., Patel, R., Yang, Y., Marshall, I., Nenkova, A., and Wallace, B. (2018). A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia, July. Association for Computational Linguistics.

Packer, M. (2018). Data sharing in medical research.

Summerscales, R. L. (2013). *Automatic Summarization of clinical abstracts for evidence-based medicine*. Ph.D. thesis, Illinois Institute of Technology, Chicago, Illinois.

Thomas, J., Brunton, J., and Graziosi, S. (2010). Eppi-reviewer 4.0: software for research synthesis. eppi-centre software. london: Social science research unit. *Institute of education, University of london*.

11. Language Resource References

Nye, B., Li, J. J., Patel, R., Yang, Y., Marshall, I., Nenkova, A., and Wallace, B. (2018). EBM-NLP: Nlp models in support of evidence based medicine. https://github.com/bepnye/EBM-NLP.