Information Extraction and Unsupervised Methods for Streamlining Evidence Synthesis in International Development Grey Literature

Kristen Edwards, Jack Gammack, Luis Kumanduri, Dylan Lewis

Abstract

In fields like international development, decision-makers prioritize making evidence-based decisions for funding and implementing future projects. This aim is made difficult because of the plethora of information being published each year, and the nature of the research corpus as unstructured text or grey literature. To make informed decisions and understand the growing corpus of research available, researchers have turned to evidence synthesis - the process of compiling information and knowledge from many sources and disciplines to inform decisions. However, the manual evidence synthesis process takes extensive time (often 18 months to 3 years) and effort, and may soon be impossible at the world's increasing rate of research output. To address these problems, we employ natural language processing techniques on a international development literature corpus of 244 documents to extract information from the title and abstract of international development documents, and to automatically cluster documents based on their content. We classify documents by Country of Study using a pretrained transformer Named Entity Recognition model and achieve an accuracy of 91.0%. Using K-Means clustering, we uncover informative and distinctive groupings of the documents which share similar semantic content. These methods reduce the time it takes for manual evidence synthesis for international development grey literature by enabling country of study filtering and clustering documents by semantic similarity.

1 Introduction

Much of the research produced in the field of international development is unstructured text or "grey literature" - information produced outside of traditional publishing and distribution channels, such as reports, policy literature, newsletters, government documents, and so on (Libraries, 2021). Within this field, having an understanding of the entire current corpus of research in an area is particularly important to researchers and decision-makers (Donnelly et al., 2018). A full understanding of the current state of research allows for proper funding in areas with the most need, and also informs future research initiatives with past results. However, the manual assessment required to gain this full understanding takes extensive time and effort.

The term evidence synthesis refers to the process of compiling information and knowledge from many sources and disciplines to inform decisions (Donnelly et al., 2018). Evidence synthesis allows for the interpretation of an individual study within the context of the global knowledge about a topic (International). Thus, evidence synthesis is an incredibly valuable tool for decision-makers in the fields of policy and research funding. However, the nature of grey literature can hinder evidence synthesis. Manually understanding the plethora of research in the field of international development requires manually extracting information - like relevant country, study type, action employed, and population observed - for every document. Researchers must also manually cluster and classify documents in order to sort them.

We used several different Natural Language Processing (NLP) techniques to automatically extract information from international development documents, which can broadly be broken up into supervised and unsupervised methods. Since our dataset is fairly small, we used pre-trained models such as spaCy and the Universal Sentence Encoder as a preprocessing step for all of the above tasks. Our aim is to use NLP to expedite currently manual document understanding processes and to uncover relationships and broad classes among documents in this field.

2 Background and Related Work

In this section we discuss background information and previous related work in the areas of information extraction and uncovering semantic similarity of documents.

2.1 Information Extraction from Domain-Specific Free Text

Machine learning (ML) has produced state of the art results for many Information Extraction (IE) subtasks, e.g. Named Entity Recognition (NER), Relation Extraction, Temporal IE, etc. and produced models like BERT that are capable of performing multiple subtasks at once (Devlin et al., 2018; Liu et al., 2019; Raffel et al., 2020; Wimalasuriya and Dou, 2010). Researchers have successfully produced IE models that are specific to certain domains. For example, LexNLP was developed to perform NLP based IE on legal and regulatory texts (au2 et al., 2018). LexNLP's key functionality is to take unstructured legal text and a) segment documents, b) identify key text such as titles and section headings, c) extract over eighteen types of structured information like distances and dates, d) extract named entities such as companies and geopolitical entities, e) transform text into features for model training, and f) build unsupervised and supervised models such as word embedding or tagging models.

In addition to the domain-specific system mentioned above, there exist automated systematic literature review system prototypes, such as ExaCT, RobotReviewer, and NaCTeM which extract relevant data such as sample sizes, population, intervention, and outcomes from free-texts (Marshall and Wallace, 2019).

NLP systems such as LexNLP and RobotReviewer utilize tokenizers that are pre-trained on domain-specific text, such as legal documents or scientific research papers, so that the models are able to produce embeddings for concepts that are not common in general literature, such as "LLC" or "hydrophilic". Word embeddings and topic models are created using state of the art methods that rely on word co-occurrence, such as TF-IDF, or more sophisticated machine learning methods such as transformers and BERT. These word- and document-level embeddings are used to train standard ML models on supervised classification tasks such as document labelling.

2.2 Uncovering Semantic Similarity between Documents for Literature Review

In addition to using supervised machine learning algorithms such as those that perform information extraction, researchers have also investigated unsupervised methods such as topic modeling and clustering to uncover grouping of documents which share high semantic similarity as part of the literature review process.

2.2.1 Topic Modeling for Exploratory Literature Review

Latent Dirichlet Allocation (LDA) uncovers a fixed number of abstract topics, K, represented in a text corpus. Each topic has a distribution over the words in the corpus vocabulary, in which the words with the highest probability collectively convey a theme which can be used to describe the topic. LDA provides a distribution over topics for each document in the corpus, representing a document as a mixture of topics (lda, 2003). After determining the optimal number of topics in terms of perplexity, researchers in (Boye Asmussen and Møller, 2019) applied LDA on a corpus of 650 papers at K = 20 topics placing each document into the topic which had the highest probability for that document. The procedure was evaluated qualitatively by determining if the papers grouped together in a topic on the basis of paper titles and the 10 most probable words for that topic made semantic sense grouped together.

2.2.2 Clustering on Scientific Article Metadata

To assist in the search phase of systematic literature review (SLR), the authors in (Weißer et al., 2020) iteratively applied K-means clustering algorithm using article metadata including the title, keywords, and the abstract to form distinct topical clusters. A TF-IDF normalized term-document matrix is constructed using the words in the text corpus and documents represented by the words counts of their respective article metadata. Further, Latent Semantic Analysis (LSA) is applied to get a lower-dimensional representation of the words in the corpus from the term-document matrix. After applying K-means, the clusters are then semantically defined by their most relevant words determined by a cluster's centroid. This process is done iteratively so as to refine the search for specific and relevant corpora as part of SLR using the relevant keywords for a cluster of interest. Then the average TF-IDF score among the top 5 words (by TF-IDF score) from the cluster's centroid is used as a proxy for the cluster's informative value (Weißer et al., 2020).

Our work bridges some of the NLP techniques discussed in the related work with other NLP techniques we discuss below by applying them in the domain of international development. Namely, we apply NER, pretrained embeddings, K-Means Clustering, and Multidimensional Scaling to assist in the process of gaining information from a corpus of international development grey literature. To analyze the results of these methods for evidence synthesis in the domain of international development literature, we use a text dataset of international development literature papers forming a corpus of 244 documents.

3 Data

In the following section we describe the data we are working with and the preparation pipeline utilized.

3.1 Dataset Description

Our dataset consists of a set of information describing 244 documents in the international development field, including titles, abstracts, descriptions of interventions and outcomes, manually extracted information, and labelled classes. The extracted information includes countries mentioned and evaluation methods used for studies. The labeled classes come from experts who identified the broad intervention type and outcome type of each document. Our dataset comes from a subset of an Evidence Gap Map (a visual form of evidence synthesis) created by the International Initiative for Impact Evaluation (3ie) (3ie, b). 3ie invented Evidence Gap Maps (EGMs) and is a leading producer of them (3ie, a).

3.2 Data Preparation Pipeline

To assist in preparation of the unstructured text data, we use NumPy, pandas, scikit-learn, and spaCy packages in Python. Our data preparation pipeline consists of extracting text fields and labels from Excel files using pandas and removing invalid values and characters.

4 Information Extraction

We begin our discussion of the supervised learning methodology we employ to assist in evidence synthesis on international development grey literature. More specifically, the country of study (CoS) associated with each paper is quite pertinent to the international development domain. Our dataset is labeled with the CoS of each paper in our corpus. However, since our dataset is rather small (244 documents), we sought to evaluate whether pretrained NER models which extract a variety of entity types from text (entitiy types noted in Appendix Section A), could accurately extract the CoS for the papers in our corpus, which have a variety of text fields.

4.1 Implementation and Methodology

To assess the accuracy of the CoS extraction process, we devise a baseline non-ML CoS extraction algorithm in addition to using pretrained NER models. We discuss below the implementations of the baseline algorithm as well as the pretrained NER models and the classification procedures which use extracted entities to predict the CoS for each paper in our corpus.

4.1.1 Country of Study (CoS) Extraction and Classification

We create a lower-cased, alphabetically-ordered, list of countries, which we construct using countryinfo[1], a Python package which contains a large dictionary of countries, their alternative names, and ISO information. We ensure the strings of the countries present in our corpus match their respective string in the alphabetically-sorted list of countries. We note that Myanmar and Kosovo are countries present in our corpus, but are not present in the countryinfo dictionary, so we add them to the final list of alphabetically-sorted countries. Since nationality is a type of named entity that NER models typically extract in addition to countries, using a comprehensive, open-source nationality-country mapping[2], we construct a lower-cased, alphabetically-ordered list of nationalities as well as a dictionary mapping nationality to country. We note that we use the words nationality and demonym interchangeably. These lists and dictionary are useful for performing the country of study (CoS) classification using extracted entities from input text or determining if a nationality or country is a substring contained in the input text string.

[1] Link to CountryInfo PyPI page

[2] Demonym-Country Mapping Link

4.1.2 Simple Substring Matcher (SSM) Algorithm Baseline & CoS Extraction & Classification

As a baseline to our CoS prediction task, we devise a simple, non-ML, deterministic algorithm, called the Simple Substring Matcher (SSM) Algorithm. This method begins by making the input text lower-cased. To predict a CoS, it then scans through the alphabetically-sorted list of countries and classifies the first country which is a substring in the input text as the CoS. If no country is found as a substring in the text, the method then scans the alphabetically-sorted list of nationalities. If a nationality is found as a substring of the input, the method maps the nationality to the corresponding country and classifies the paper as having that country as the CoS. If neither country nor nationality is found as a substring in the text, the method classifies the paper's CoS as a *None* value. We refer to this classification model as the Simple Substring Matcher (SSM) model.

4.1.3 spaCy CNNs with Residual Connections and the Bloom Embedding Strategy for NER

With the exception of the transformer model discussed in the next section, the pretrained NER models provided by spaCy are based on a deep CNN architecture with residual connections and the implementation of the transition-based parsers paradigm, which is borrowed from the shift-reduce parsers presented in (Lample et al., 2016).

The models employ a framework referred to as "Embed. Encode. Attend. Predict". With the exception of the pretrained NER English spaCy Small Model (ESMS) model, the models use static word embeddings and the bloom embedding strategy. Rather than use words as keys to an embedding dictionary, hashes computed for the words are used as keys in the embedding dictionary. This method of embedding is referred to as the Bloom Embedding strategy. An important implication of using the Bloom embedding dictionary is that different words with the same hash will yield the same embeddings, thus yielding a more compact embedding dictionary. Thus, a sentence is composed from a list of these word embeddings. A CNN is used for encoding this list into a sentence matrix. The sentence matrix enables context to be considered when forming the entity predictions. Using the sentence matrix and a provided query vector, the model attends to which parts of sentence are most informative and this process yields representations which are problem-specific. Finally, the prediction is made using a multilayer perceptron. This framework is discussed in great depth here.

We specifically use *en_core_web_sm* (ESMS), *en_core_web_md* (ESMM), and *en_core_web_lg* (ESML) implementations which are trained on various web text including blogs, news, and comments contained in the OntoNotes dataset[3]. The ESMS model uses one-hot encodings for words. Using static GloVe embeddings trained on Common Crawl[4], the ESMM model uses a word vector table that has 20k unique vectors (with dimension 300) for ∼500k words. Finally, the ESML model also uses static GloVe embeddings trained on Common Crawl but uses a word vector table which has 685k unique vectors (with dimension 300) for 685k words.

4.1.4 spaCy Transformer Model for NER

As an alternative to the CNN architecture discussed in section 4.1.3, spaCy offers a RoBERTa-base transformer model (Liu et al., 2019) pretrained on OntoNotes for NER. The implementation is referred to as *en_core_web_trf*, which we refer to as the pretrained NER English spaCy Transformer Model, or ESMT.

4.1.5 CoS Extraction & Classification by Pretrained NER Models

Although we utilize different pretrained NER models in our experiments as described above, the process for classifying CoS using predicted entities is the same. Each model takes the raw text as input, predicts various non-overlapping entities present in the text into one of the entity categories shown in Appendix Section A. For the CoS classification task, we only consider the predicted **NORP** (nationalities, or religious, or political groups) entities and the **GPE** (countries, cities, states) entities as we assume that these categories are the only ones which would contain the country or relevant demonym associated with the CoS. We now begin our discussion of the classification procedure for the pretrained NER models.

First, we make all NORP and GPE entities lower-cased. Next, we map any demonyms present among the NORP entities to their corresponding country. We then combine the resulting unique

[3] OntoNotes Link

[4] Link to Common Crawl

NORP and GPE entities into an alphabetically-ordered list. We scan this list of NORP and GPE entities checking if any of them exist in the countries list mentioned in Section 4.1.1, classifying the CoS as the first entity-country match found. If no match is found, we make a final attempt to determine the CoS by providing each of the entities as input to the GeoPy Geocoder[5] object, which provides an address-location object if a location is found for the provided entity or no value otherwise. We do this for each entity, and if a location is found for a particular entity, we classify the paper's CoS as the country associated with the found address-location. If no country is found for all the entities, we classify the paper's CoS as *None*.

4.2 Evaluation

For each paper in our text corpus, the country for which the study is took place is provided as a label. Though our corpus is relatively small (244 documents), for larger scale international development literature corpora, it would be useful to partition the countries on the basis of the country where the study took place. This partitioning would be useful for filtering corpora for studies which took place in particular countries, e.g. searching for studies which took place in Kenya. We thus formulate the classification task of predicting the Country of Study (CoS) associated with a paper using different text fields as input.

To perform this classification task, we leverage various pretrained NER models provided by the spaCy[6] python package, which encapsulate a pipeline that extracts various types of named entities from raw text strings. The full list of named entity types that are extracted from these models can be found in Appendix Section A. Specifically, we evaluate whether the list of unique extracted entities corresponding to geopolitical entities (GPEs) or nationality, or religious or political groups (NORPs) entities found by the pretrained NER models, if any, contain the ground-truth label for the country of study (CoS) for each paper in our corpus. We compare the performance, by accuracy, of the pretrained NER models and a non-ML baseline. In addition to testing different classification models, we experiment with different input strings to see how results change with various text fields and concatenations between them. These various inputs to the models include the title, abstract, intervention description, outcome description, and various concatenations of these text fields. We list all combinations of various input text and models and their respective accuracy on the corpus in Table 1.

We use the following abbreviations for the various inputs and models in Table 1:

Input Text Abbreviations:

T = Title of Paper,
A = Abstract of Paper,
ID = Intervention Description of Paper,
OD = Outcome Description of Paper,
X + Y = Text field **X** concatenated with text field **Y**

NER Model Abbreviations:

Baseline:
SSM = Simple Substring Matcher,
Pretrained spaCy English NER Models:
ESMS = CNN Model (One-hot Word Encodings),
ESMM = CNN Model (Word-Vector table of 20k unique vectors (Dimension of 300) for ∼500k words),
ESML = CNN Model (Word-Vector table of 685k unique vectors (Dimension of 300) for 685k words,
ESMT = RoBERTa-base Transformer Model

4.3 Results

We detail the accuracies achieved by the various input and NER model combinations below. Additionally, we describe use-cases which can assist in the evidence synthesis process using the predicted CoS from the most accurate input, model combination.

4.3.1 Accuracy of Text Input & NER model Combinations for Predicting CoS

We analyze various inputs and how different models of varying complexity perform relative to each other. The CoS classification accuracy of various input, model combinations can be found in Table 1. All of the pretrained spaCy NER models have 0.0% accuracy when using the just the intervention description, however the SSM model achieves 13.9% accuracy on the intervention description. All models attain an accuracy of 2.9% when using just the outcome description. The title and abstract individually appear to be good input fields for predicting the CoS, however the concatenation of title and abstract appears to be the most informative input, as this is the input that yields the highest performance across all of the models. Overall, we observe that the baseline simple substring checker

[5] Link to GeoPy API
[6] https://spacy.io/

Text Input	Baseline: SSM	ESMS	ESMM	ESML	ESMT
T	0.676	0.434	0.520	0.488	0.725
A	0.762	0.787	0.807	0.820	0.836
ID	0.139	0.0	0.0	0.0	0.0
OD	0.029	0.029	0.029	0.029	0.029
T + A	0.840	0.844	0.869	0.885	**0.910**
T + A + ID	0.832	0.828	0.861	0.873	0.893
T + A + OD	0.832	0.840	0.857	0.881	0.898
T + A + ID + OD	0.824	0.824	0.848	0.869	0.881

Table 1: NER Input & Model Combination Accuracy Results on predicting Country of Study (CoS) for the International Development Literature Corpus. Bolded accuracy shows most performant input & model combination.

is a fairly competitive model against the pretrained ML models, outperforming all the ML models on intervention description, performing the same as the ML models on outcome description, and only falling a few percentage points below even the best ML model on the other inputs. With the exception of the title, intervention description, and outcome description, the ML models in increasing order of complexity, do increasingly better on the CoS extraction task, in the following order from least performant to most performant: ESMS, ESMM, ESML, and ESMT. With the exception of the intervention description and output description inputs, we observe that the ESMT model performs best across all other inputs. Furthermore, we see that the concatenated title and abstract input and the ESMT model combination performs the best across all input-model combinations with 91.0% accuracy. We note two misclassified concatenated title and abstract inputs to the pretrained transformer NER model in Appendix Section B.

In the first misclassified example, observe that there are no countries mentioned in the concatenated title and abstract input, thus there are no countries for pretrained transformer NER model to extract and thus the classification method predicts *None*, although the true CoS is Iran.

In the second misclassified example, we observe that there are technically multiple countries as part of the study. The Democratic Republic of the Congo is the ground truth CoS. In the input, we see that the Democratic Republic of the Congo, Guatemala, India, and Pakistan are mentioned. However, there is an acronym adjacent to the Democractic Republic of the Congo, DRC, which is likely why the classification procedure predicted Guatemala rather than the DRC. Thus, a natural extension to the classification procedure would be to allow for multiple labels to be predicted for the CoS as well as the robust ability to predict CoS even when variations of country names are present in a text, not necessarily the one we have in the list mentioned in Section 4.1.1.

We describe the potential use cases for using the predicted CoS in assisting in the evidence synthesis process below.

4.3.2 Map of Predicted CoS

Using the CoS predictions from the pretrained NER transformer model on the concatenated title and abstract input, we construct a geographical map of the corpus as shown in Figure 1. For each paper, which had a non-null prediction by the ESMT model, we place a tooltip at location coordinate associated with the predicted CoS. These location coordinates were pulled using the GeoPy Geocoder object mentioned in section 4.1.5. We added slight, uniform random jitter to each of the coordinates, so papers with the same predicted CoS don't directly overlap. When a user hovers over the tooltip, they will see the title of the paper associated with that tooltip as seen in Figure 2. The webpage for this map can be downloaded here for view in a browser.

4.4 Filtering by Predicted CoS

For large corpora of International Development Literature, the utility of the CoS prediction task is most evident by the robust filtering capability it enables. For instance, by concatenating only the title and abstract of papers in the corpus, and using them as input to generate CoS predictions by the pretrained transformer model used in this study, this enables the ability for unlabeled papers in the

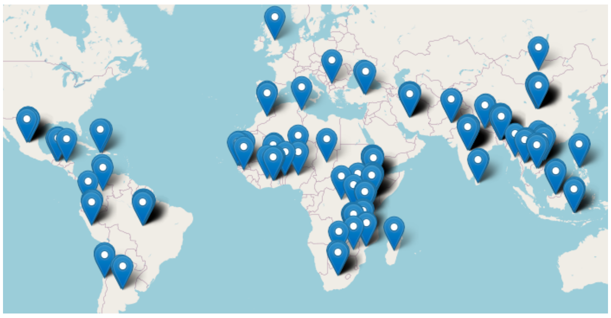
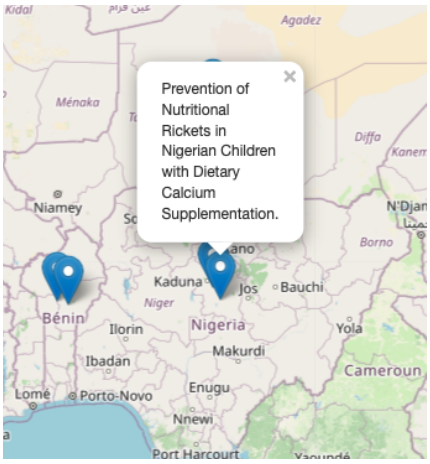


Figure 1: Map of NER Transformer CoS Predictions for papers in the corpus



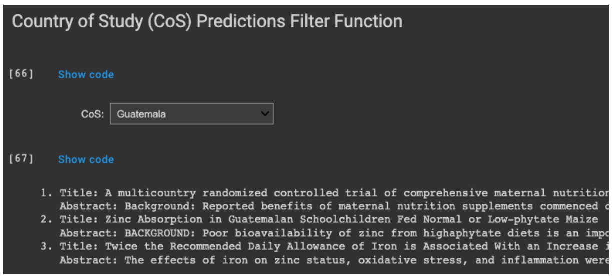
Prevention of Nutritional Rickets in Nigerian Children with Dietary Calcium Supplementation.

Figure 2: Tooltip showing title of the paper in predicted CoS on hover.

corpus to be accurately filtered to identify studies which had a specific CoS. This method would greatly reduce the time necessary for manual CoS annotation while also yielding higher accuracy than a simple substring matcher, simplifying a step in the international literature review process with high accuracy. An example of this filtering functionality is shown Figure 3 for papers in the corpus, which were predicted as having Guatemala as the CoS. The CoS was predicted using the pretrained transformer NER model with the concatenated title and abstract as input. We display the corresponding title and abstract for quick scanning of results for relevancy to research topic. Additionally, we provide the option to filter the corpus for papers which had a predicted CoS as *None*.

5 Semantic Similarity

In this section, we seek to uncover semantic similarity between documents to understand groupings of different topics as well as get representations that describe what these groupings of topics discuss.



Country of Study (CoS) Predictions Filter Function

CoS: Guatemala

```
1. Title: A multicountry randomized controlled trial of comprehensive maternal nutritior
   Abstract: Background: Reported benefits of maternal nutrition supplements commenced c
2. Title: Zinc Absorption in Guatemalan Schoolchildren Fed Normal or Low-phytate Maize
   Abstract: BACKGROUND: Poor bioavailability of zinc from highaphytate diets is an impc
3. Title: Twice the Recommended Daily Allowance of Iron is Associated With an Increase i
   Abstract: The effects of iron on zinc status, oxidative stress, and inflammation were
```

Figure 3: Filtering the International Literature Corpus for papers which were predicted as having Guatemala as the CoS

5.1 Implementation and Methodology

Here we discuss our methods for grouping papers of a similar topic together and for capturing words that describe the topic of each cluster in order to aid researchers in evidence synthesis.

5.1.1 Document Clustering and Topic Modeling

Each of the documents in our dataset have labels that roughly categorize the intervention domain and outcome domain, but it may be useful for researchers to be able to group the documents based on the semantic meaning of the text in the document. Additionally, clustering the documents with an algorithm may be easier using a model than having experts doing the labelling manually, especially as the dataset grows larger.

Using Google's Universal Sentence Encoder (Cer et al., 2018), we are able to easily capture the semantic meaning of each description into a 512-dimensional vector that may be used for clustering, as the vector space is built such that vectors closest to each other have the most similar semantic meaning. Ideally, using only the semantic meaning contained in the text description, the model should be able to group papers that are about similar topics together.

For each description of each paper (either abstract, intervention description, or outcome description), we encode the description into a vector and use K-Means clustering to get a user-chosen number of clusters, where clusters may not necessarily have the same size.

In order to visualize the clustering, we project the vector space into 2 dimensions using UMAP, a dimensionality reduction algorithm that has been shown to capture global and local structure well (McInnes et al., 2020).

After clustering papers together, we view words that occur more frequently within a certain cluster than they do in other clusters by using TF-IDF. We then label each cluster with the top N words that identify that cluster to get an idea of the topics within each cluster.

5.2 Results

Here we discuss the results of our clustering and topic modeling methods.

5.2.1 Clustering Results

Figure 4 shows the result of clustering (N=8) based on the encodings of the intervention description for each paper in the dataset. We are able to observe some separation of clusters in the figure despite it being a 2-dimensional projection of a 512-dimensional vector space.

Viewing the top 5 words for each cluster, we can see some clear topics such as the ”flour, rice, fortified, powder, sachets” topic in the bottom right corner of the graph. Viewing the intervention descriptions that fall into this cluster, we observe that the papers predominantly discuss directly fortifying rice or flour with nutrients and providing it to communities. This is a very different topic than the topics in the top left corner of the chart, which discuss breastfeeding, health, and nutritional and physical education.

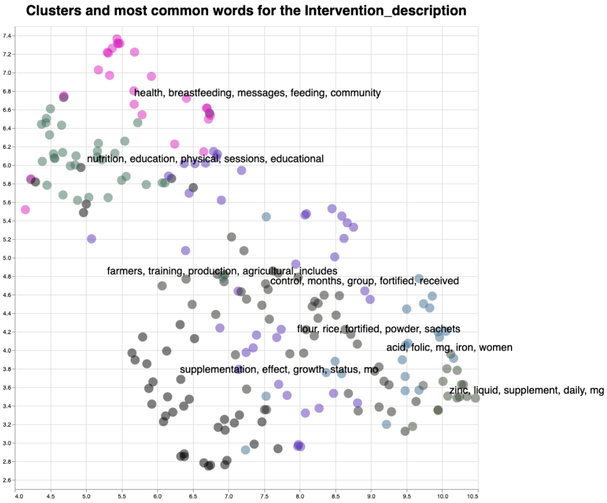


Figure 4: K-means clustering of the embeddings of intervention descriptions of papers.

5.2.2 Evaluation of Clustering

Figure 5 shows the performance of K-means clustering algorithm for different number of clusters when clustering embeddings of the intervention outcome for each document. In practice and by this metric, we found that 6-10 clusters suffices to capture broad relationships between different clusters.

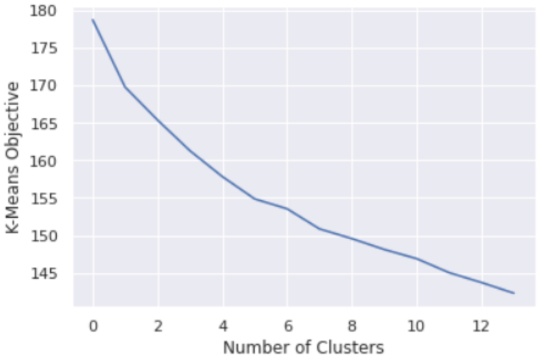


Figure 5: Performance of K-Means Clustering by intervention outcome with different number of clusters

Our data naturally breaks up into clusters of outcome groups, and the K-Means algorithm matches up quite well with these categories. When clustering the outcome description using 8 clusters, we capture distinct clusters for nutrient fortification, agriculture, diet, maternal health, iron deficiency, and child development as seen in Appendix Section C. The algorithm does not do perfectly, as these categories have some overlap in description (i.e. diet and child development). As the number of clusters increases, we primarily see splitting of these natural categories. Some of these are helpful, for instance a distinctive breastfeeding cluster forms, but others are much less useful. For example, the iron deficiency cluster is now split into three different clusters, one keying in on the word ’hemoglobin’, another on the word ’blood’, and the other capturing the rest. Similar clusters form when clustering by intervention description.

Even with relatively small number of clusters we also see clusters that correspond to relationships we weren’t trying to capture. For example, when clustering by intervention description there is a distinct ’randomized’ cluster that forms, for text which includes a description of a randomized experiment. When clustering by outcome description the first cluster that forms is a ’z-score’ cluster, which consists of text describing statistical methods used to analyze the outcomes. In order to obtain more effective clustering, it might be effective to delete certain common words which correspond to undesired clusters, and then see what our unsupervised methods do for our new text.

5.2.3 Multidimensional Scaling Plot

Figure 6 shows a multidimensional scaling (MDS) plot, which displays the textual similarity of intervention classes via their proximity in the plot. The closer two points are, the more similar the intervention classes that they represent are. Each point is labelled with a number which corresponds to its intervention class, as listed below:

1. Provision of improved water access and management systems
2. Provision of free or reduced-cost access to improved seed varieties
3. Provision of free or reduced-cost access to livestock
4. Provision of free or reduced-cost access to other/unspecified agricultural inputs
5. Education/information - Farmer field schools
6. Education/information - Agricultural extension programs
7. Education/information - information/ guidance
8. Education/information - other educational programs
9. Other efforts to improve the production system - insurance
10. Support for creating storage structures at farms
11. Fortification
12. Governmental price manipulations (excluding tariffs)
13. Direct provision of foods
14. Cash-for-food programs
15. Provision or use of supplements
16. Efforts to increase women's empowerment
17. Peer support/counsellors
18. Professional services (dietitians / nurses)
19. Community meetings
20. Classes
21. Healthy food social marketing campaigns

In Figure 6, the size of each point corresponds to the number of documents in the intervention class. For example, point 11 corresponds to class Fortification which includes 74 documents.

We observe a cluster in the MDS plot of intervention classes 11,13, and 15, which correspond to intervention classes Fortification, Direct provision of foods, and Provision or use of supplements.

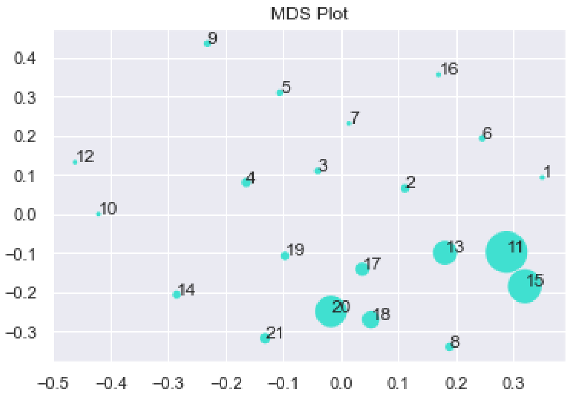


Figure 6: Multidimensional scaling (MDS) plot of similarities among intervention classes. Each point is labelled with a number which corresponds to its intervention class. All intervention classes are listed by number in this section.

6 Conclusion

The manual evidence synthesis for international development grey literature is a time-consuming process. We have demonstrated that certain components of the evidence synthesis process in international development grey literature such as filtering corpora for papers which have a specific country of study or grouping similar documents together can benefit greatly from the use of methods of information extraction and unsupervised learning. More specifically, we have utilized a pretrained transformer NER model to accurately predict the country of study for the papers present in the corpus used in this study, thus enabling accurate filtering of the corpus for papers with a specific predicted country of study. After tuning to find the optimal number of clusters in K-Means clustering, we uncovered informative and distinctive clusters of documents with similar content in the corpus. The automation of these components in the evidence synthesis process for international development grey literature mitigates the effort and time that is required for manual evidence synthesis.

6.1 Future Work

The dataset used in this study is only 244 documents, which is small. Thus a natural extension of this study would be to construct larger international development grey literature corpora and run similar analysis to that which has been done in this study. A larger corpus would enable the finetuning of models to the problems in this domain, such as finetuning a pretrained NER model for the CoS prediction task.

Additional future work could involve training or fine-tuning a base language model on international development literature in order for the models to understand what the words mean in their specific contexts. Although the vast majority of words used in our dataset are common and likely within the vocabulary of the pre-trained NLP models, it would be more useful for the models to understand the contexts better.

As discussed in the clustering results, we observed issues with clustering on semantic meaning for papers that discussed statistics in their descriptions, because the statistical words caused these papers to be grouped together even when the underlying topics were very different. Some future work could involve creating a list of stop words to be removed from consideration when encoding the semantic meaning of a paper.

References

2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

3ie. a. Evidence gap maps. Available at `https://www.3ieimpact.org/evidence-hub/evidence-gap-maps` (2021/08/04).

3ie. b. Evidence mapping. Available at `https://www.3ieimpact.org/evidence-hub/evidence-gap-maps` (2021/08/04).

Michael J Bommarito II au2, Daniel Martin Katz, and Eric M Detterman. 2018. Lexnlp: Natural language processing and information extraction for legal and regulatory texts.

Claus Boye Asmussen and Charles Møller. 2019. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Christl A. Donnelly, Ian Boyd, Philip Campbell, Claire Craig, Patrick Vallance, Mark Walport, Christopher J. M. Whitty, Emma Woods, and Chris Wormald. 2018. Four principles to make evidence synthesis more useful for policy. *Nature*, 558:361–364.

Evidence Synthesis International. What is evidence synthesis? Available at `https://evidencesynthesis.org/what-is-evidence-synthesis/` (2021/08/04).

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360.

Simon Fraser University Libraries. 2021. Grey literature: What it is & how to find it.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ian J. Marshall and Byron C. Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Tim Weißer, Till Saßmannshausen, Dennis Ohrndorf, Peter Burggräf, and Johannes Wagner. 2020. A clustering approach for topic filtering within systematic literature reviews. *MethodsX*, 7:100831.

Daya Wimalasuriya and Dejing Dou. 2010. Ontology-based information extraction: An introduction and a survey of current approaches. *J. Information Science*, 36:306–323.

A Named Entity Types

PERSON = People including fictional, **NORP** = Nationalities, or religious, or political groups,
FAC = Buildings, airports, highways, bridges, etc., **ORG** = Companies, agencies, institutions, etc.,
GPE = Countries, cities, states, **LOC** = Non-GPE locations, mountain ranges, bodies of water,
PRODUCT = Objects, vehicles, foods, etc., **EVENT** = Named Hurricanes, battles, wars, sport events,
WORK OF ART = Title of books, songs, etc., **LAW** = Named documents made into laws,
LANGUAGE = Any named language, **DATE** = Absolute or relative dates or periods,
TIME = Times smaller than a day, **PERCENT** = Percentage, including '%',
MONEY = Monetary values, including unit, **QUANTITY** = Measurements, as of weight or distance,
ORDINAL = "first", "second", etc., **CARDINAL** = Numerals that do not fall under another type

B Misclassified CoS Examples

Misclassified CoS Example 1: Classified as another country

Title + Abstract: *The Effect of Educational Intervention Based on The Theory of Planned Behavior on Nutritional Behavior with Regard to Cardiovascular Diseases Among Health Volunteers. Background: We sought to evaluate the effect of educational intervention based on the theory of planned behavior (TPB) on nutritional behaviors in relation to the cardiovascular disease (CVD) among health volunteers. Methods: In this quasi-experimental study, the participants included 128 active health volunteers. To conduct the study, 65 and 63 participants were randomly assigned into the intervention and control groups, respectively. Data were collected before and six weeks after the intervention using a validated researcher-made questionnaire. The questionnaire consisted of demographic variables, knowledge questions, and TPB constructs. Data were analyzed by Chi-square, t-test, Mann-Whitney U, and Wilcoxon test. Results: No significant difference was observed between the intervention and control groups with regard to the demographic characteristics, knowledge mean scores, and TPB constructs at the beginning of the study. However, the mean scores of knowledge, attitude, subjective norms, perceived behavioral control (PBC), and nutritional behavior increased significantly (P¡0.001, P¡0.001, P=0.018, P=0.007, and P¡0.001, respectively) in the intervention group six weeks after the beginning of study. Significant differences were observed in nutritional performance of the intervention group, in other words the nutritional behavior of the intervention group members changed during the intervention. Conclusion: The PBC was the strongest construct in attitude. To optimize nutritional interventions in preventing the CVD, TPB should be implemented in educational interventions.*

True CoS: Iran; **Predicted CoS:** *None*

Misclassified CoS Example 2: Multicountry Study

Countries mentioned in the text are bolded to show presence in the text.

Title + Abstract: *A multicountry randomized controlled trial of comprehensive maternal nutrition supplementation initiated before conception: the Women First trial. Background: Reported benefits of maternal nutrition supplements commenced during pregnancy in low-resource populations have typically been quite limited. Objectives: This study tested the effects on newborn size, especially length, of commencing nutrition supplements for women in lowresource populations a3 mo before conception (Arm 1), compared with the same supplement commenced late in the first trimester of pregnancy (Arm 2) or not at all (control Arm 3). Methods: Women First was a 3-arm individualized randomized controlled trial (RCT). The intervention was a lipid-based micronutrient supplement; a protein-energy supplement was also provided if maternal body mass index (kg/m2) was ¡20 or gestational weight gain was less than recommendations. Study sites were in rural locations of the* ***Democratic Republic of the Congo*** *(DRC),* ***Guatemala***, ***India***, *and* ***Pakistan***. *The primary outcome was length-for-age z score (LAZ), with all anthropometry obtained ¡48 h post-delivery. Because gestational ages were unavailable in DRC, outcomes were determined for all 4 sites from WHO newborn standards (non-gestational-age-adjusted, NGAA) as well as INTERGROWTH-21st fetal standards (3 sites, gestational age-adjusted, GAA). Results: A total of 7387 nonpregnant women were randomly assigned, yielding 2451 births with NGAA primary outcomes and 1465 with GAA outcomes. Mean LAZ and other outcomes did not differ between Arm 1 and Arm 2 using either NGAA or GAA. Mean LAZ (NGAA) for Arm 1 was greater than for Arm 3 (effect size: +0.19;*

95% CI: 0.08, 0.30, P = 0.0008). For GAA outcomes, rates of stunting and small-for-gestational-age were lower in Arm 1 than in Arm 3 (RR: 0.69; 95% CI: 0.49, 0.98, P = 0.0361 and RR: 0.78; 95% CI: 0.70, 0.88, P ¡ 0.001, respectively). Rates of preterm birth did not differ among arms. Conclusions: In low-resource populations, benefits on fetal growtha related birth outcomes were derived from nutrition supplements commenced before conception or late in the first trimester. This trial was registered at clinicaltrials.gov as NCT01883193.

True CoS: Democratic Republic of the Congo; **Predicted CoS:** Guatemala

C Sample Clusters

Sample clusters using 8 clusters by output description **Cluster 1** Behaviors included: mouthfuls taken, mouthfuls refused, self-feeding, motheraTMs verbal responses to a child signal, and hand-washing. More than 90

The primary outcomes were the prevalence of being wasted (weight-for-height z-score [WHZ] ¡ a2) and mean WHZ at 6 mo and at 1 y amongst children less than 5 y. Page 6; mothers BMI

Six months after child birth, duration of exclusive breastfeeding was assessed. Data were analyzed by means of descriptive and inferential statistics. The breastfeeding self-efficacy in the intervention group increased significantly compared to the control group one month after delivery (123.6 versus 101.7, i¡0.001). The duration of exclusive breastfeeding was significantly higher in the intervention group (5.03 mo versus 2.7 mo, i¡0.001). Also, there was a significant relationship between breastfeeding self-efficacy and duration of exclusive breastfeeding (i¡0.001). p.3

At 8 weeks postpartum, participants in the intervention group had significantly higher mean Breastfeeding Self-Efficacy ScaleaShort Form scores and rates of exclusive breastfeeding than those in the control group. No significant group differences were found with regard to breastfeeding duration. p.1

Infant feeding practices including rates of initiation of the breastfeeding within one hour of birth; exclusive breastfeeding and bottle-feeding during the first 6 months of life. P.114

The primary outcome of this study is the birthweight of the newborn infants". "The secondary outcome is the maternal dietary behaviour, which includes daily caloric intake and dietary diversity score". (p.7)

The expected main outcome of the study was a high rate of EBF practice at 6 months, especially in the intervention group. (p. 51) The research team developed the measures of demographic characteristics and breastfeeding practice. This included motheraTMs age, marital status, educational level, family yearly income, motheraTMs working status, and three questions on breastfeeding practice. (p. 56)

The main outcomes of the study were the effect of the interventions on: (i) dietary diversity score (DDS) and (ii) the proportion of mothers feeding the child an unhealthy snack in the last 24 h. (p. 355) The DDSs and the proportion of mothers who exclusively breastfed were measured using standardised tools recommended by the WHO. (p. 356)

The primary outcome in this study will be the exclusive breastfeeding rate at month 1, month 3, and month 6 postpartum in both the intervention group and the control group. (p. 7)

We obtained outcome information through face-to-face interviews at enrolment (in the post-partum ward 2a3 days after birth), age 1 week (at the post-partum clinic visit), and age 6, 10, 14, 18, and 24 weeks (at well-child visits). (a) We assessed two coprimary outcomes: the proportion of mothers who initiated breastfeeding within 1 h of birth (a) (p. e549)

the ICFI scores are (i) breastfeeding (whether or not the mother is currently breastfeeding her child); (ii) use of a baby bottle in the previous 24 h (yes or no) p. 24

The LATCH mnemonic scores were assessed by Latching on, Audible swallowing, Type of nipples, Comfort and Help necessary for the mother to hold the baby to breast. P. 42

Outcome variables, or parental investments, were grouped into five domains: family planning, breastfeeding, health, education, and paternal financial support. Each of these outcome variables is a parental investment insofar as each of them represents a choice some parents made to allocate scarce resources (energy, time, money, etc.) toward their childaTMs health and human capital that could have been allocated elsewhere.

In the breastfeeding domain, two parental investments were analyzed to capture adherence to these recommendations. The first was a dichotomous variable reflecting, by maternal recall at the time of the follow-up study, whether the child received his/her first complementary food or drink at 6 months of age versus earlier or later than 6 months. This outcome was analyzed for index children and for younger siblings who were at least six months of age at the time of follow-up data collection. Duration of breastfeeding, which was analyzed for index children only and was again reported by mothers retrospectively, was calculated as a count variable equal to the number of months the child was breastfed before completely weaning. p 6.

Fifty-three point eight percent of respondents were 20—35 years old, 80.8

Postnatal BSES (Breastfeeding Self-efficacy Scale)hort Form, a 14-question scale evaluates the extent to which mothers feel adequate in their breastfeeding. P.2

The practice of exclusive breastfeeding (dependent variable) was verified on a monthly basis in accordance with the food consumption data of the newborn contained in the monthly telephone interview forms. P.4

”Blood samples (4 ml) were collected from the mothers only at the enrollment and from both mothers and babies at the 4th month of the intervention. Page 302”

The primary outcomes studied are vitamin A intake among children aged 6a35 months and their mothers; notably, these children were not yet born at the time the REU intervention ended. P.1176

The agricultural intervention significantly improved maternal selfaefficacy (beta = 0.82; p ¡0.001) and this was partially mediated by increasing gains in spousal support (beta = 0.43; p ¡0.001) and food security (beta = 0.78; p ¡0.001). We found similar results in women who additionally received nutrition education, but there were no additional gains in selfaefficacy when compared to women who received the agricultural intervention without nutrition education. We discovered substantial variability in the delivery of nutrition education due to low CHV participation.

”Dietary intake at recruitment (4-13 weeks) and during the third trimester of pregnancy (26-40 weeks) was estimated using repeat 24-h recalls (2) on non-consecutive days, excluding weekends and holidays. Page 192”

”Main Outcomes and MeasuresaNeonatal mortality rate.” ”Gains in home delivery practices, essential newborn care, and feeding practices were also observed in the intervention clusters compared with the control clusters, which may explain part of the observed reduction in neonatal mortality.”. (p.7)

Trained nutritionists did 24 h dietary recalls at the 3-month visit. Infants were revisited at age 9 months to ascertain the duration of exclusive breastfeeding and to assess the effect of the complementary feeding intervention. (p. 1419)

All motherainfant pairs were analysed at birth, 2, 5, 7, 9, and 12 months postpartum. (a) Primary endpoints were: maternal and infant plasma vitamin A concentrations (a). (p. 2089) We also measured secondary endpoints: (...) breastmilk levels of vitamin A, (a). (p.2089) To assess the concentration of Na+, K+ , and vitamin A, breastmilk was obtained from both breasts by maternal manual expression. (p. 2090)

Our secondary outcomes were exclusive breastfeeding prevalence (a). The prevalence of breastfeeding was assessed using 24-hour recall (a). (p. 6)

Cluster 2

Evaluate the impact of oportunidades (CCT) on the birth weight of children from poor rural familiesand examine the pathways by which the improvements occurred. P. 51-52

The study evaluated the effect on micronutrient status of a micronutrient-fortified biscuit, given to primary school children, over a period of 2.5 years. P.1204

Data on the attainment of milestones were collected using a test previously developed for Bangladeshi children that included motor milestones taken from the WHO Multicentre Growth Reference Study(21). Mothers were asked at each monthly visit whether and at what age their child had achieved any of the following developmental milestones: (1) motor skills: hands come together, picks up small objects, transfers spoon from hand to hand, sits unsupported, sits supported, crawls, standing with support, standing unsupported, walks supported, walks unsupported; (2) language skills: babbles, says three or four clear words; and (3) social development: smiles when smiled at, shows apprehension at strangers. Children were asked to demonstrate any new milestone reported by the mothers. The milestone was recorded as achieved only after project staff witnessed it (see page 558). BSID II were used to assess child development. The test has two subscales: the mental development index (MDI) and psychomotor development index (PDI) and was standardised in the USA. Each index is age-normalised with respect to a population mean of 100 and SD of 15. The test was administered by two trained psychologists. A few of the items were adapted to the different setting (for example, appearance of the dolls or pictures of the houses) but the underlying constructs were unchanged. (see page 558).

Motor development of the infants was assessed by the mothersaTM reporting of gross motor milestones, a method known to be accurate and sensitive for identifying developmental delays. (1034)

Gross motor development (assessed by using the Peabody Developmental Motor Scale, Second Edition, instrument) was the primary outcome. Secondary outcomes were neurologic integrity, evaluated by using the Infant Neurologic International Battery (INFANIB), 20 and motor quality, assessed by using the Behavior Rating Scale (BRS) of the Bayley Scales of Infant Development, Second Edition. pp3

WAZ

Weight measurements of the children on the intervention aged 6 to 59 months at the entry, exit and graduation stages were retrieved from Kenya Medical Research Institute Family AIDS Care and Education Services programme activities reports. Anthropometry (height measurements) for the children on the intervention and comparison children was taken.

Cognitive, receptive and expressive language, and fine and gross motor scores on the Bayley scales of infant development-III; height, weight, and hemoglobin levels measured at the baseline and end of intervention. P.1

Age- and sex-standardized z scores were computed for ... weight-for-height through the use of the WHO growth reference tables (p. 1072)

”Primary outcomes were height-forage z-scores (HAZ), according to WHO growth standards and cognitive composite scores at 36 months of age” (p.1).

The Bayley Scales of Infant Development 3rd edition (BayleyaTMs III) pp8 The Family Care Indicators is a self-report questionnaire pp8

height m; HFA z score

Cognitive ability was evaluated dynamically (a) at baseline and after 9 months of intervention. (p. 2)

Birth weight was the primary outcome of this trial. (p. 2)

To assess the overall effect on the interventions on anthropometric out-comes of preschool age children, we test whether the estimated impact on and height-for-age p. 23

The primary outcome of this analysis was the change in LAZ (length for age z-score) of intervention group compared to children in comparison group p. 395

The primary outcome of the trial is to assess change in the percentage of stunted infants (lengthaforaage ¡ a2 zascore) at 6, 12, and 18 months p. 7

The primary study outcome was linear growth as measured by the monthly change in length-for-age z score (LAZ) over the 12-mo intervention follow-up. Secondary outcome measures included monthly changes in weight-for-length z score (WLZ), midupper arm circumference (MUAC), and head circumference (HC); the longitudinal prevalence of diarrhea, vomiting, fever, acute respiratory illnesses, and any illness that required medical attention during the intervention period; the longitudinal development of stunting (LAZ ¡a2), wasting (WLZ ¡a2), anemia, and systemic inflammation, and blood LC-PUFAs concentrations at midterm (after 6 mo of intervention) and at endline (after 12 mo of intervention). p 456. 47583282

The primary outcome, newborn length-for-age z score (LAZ), was based on length measurements obtained by the assessment teams before 48 h of age.

At 24 mo of age, the offspring in the IFA group had significantly higher length-for-age z scores (LAZs) (0.14; 95

”The primary outcome of this study was child size attained at 24 mo of age. Weight, length, and head circumference measurements were obtained in the participantaTMs home. Page 1127”

”Anthropometric measurements (weight, length, and mid-upper arm circumference) of the young children were taken by the principle investigator and trained health workers at baseline and after 3 and 6 months of intervention."

Wt (kg) Ht (cm) MUAC (cm) WHZ HAZ WAZ MUACZ

"The educational intervention was able to maintain BF at two and four months, showing to be effective in maintaining BF in both the short and long term." (p.6)

"The primary outcomes were the cognitive, expressive language, receptive language and fine motor scores on the Bayley Scales of Infant and Toddler Development, which are most commonly used internationally to assess skills acquired in the first 3 years of life. Page 802"

"This will be calculated from the childaTMs age in months and length on associated ages. Page 7"

(a) length of children was determined with the child wearing minimal clothing. (p. 8) (a) analyses do not focus on the primary outcome that was formulated for the baseline survey (prevalence of stunting), but on the primary outcome: HAZ changes from baseline to endline. (p. 9)

The primary outcome was cognitive development as measured by a test of school readiness, a verbal reasoning test and a nonverbal reasoning test. (p. 4)

(a) child development outcomes (assessed by the Bayley Scale for Infant and Toddler Development, Third EditionaBSID III) (a) (p. 154) (a) cognitive, language, and motor development at 12 and 24 months evaluated by the BSID III. (p. 156)

The main outcomes were monthly change in LAZ, weight-forage (WAZ) and WLZ scores from 6 to 24 months of age, and absolute risk differences in prevalences of stunting (LAZ¡-2), underweight (WAZ¡-2) and wasting (WLZ¡-2) at a 12-month follow-up. (p. 3) "Change in length cm/month" (table 3, p. 10)

Assessment of growth was by comparison of final attained weight, height, and associated Z scores at 18 months of age. (p. 1866)

Anthropometric measurements (height and weight) were taken for children 6a41 months of age and their caregivers; only anthropometric data for children 12a41 months were used in the impact evaluation. (p. 591)

stunting, underweight, and wasting