

# PROTOCOL: Automatically mapping and assessing inequalities in public health research

*Finding Accessible Inequalities Research in Public Health (the FAIR Database)*

## Plain English Summary

Public health has been defined as “the art and science of preventing disease, prolonging life and promoting health through the organized efforts of society”. Most often, addressing public health problems requires complex strategies and programmes of work, implemented in diverse and sometimes rapidly changing situations around the world. As success is often dependent on people’s individual circumstances and capacity to act in a certain way, or utilise a service that is on offer, there is a risk that some public health interventions can actually increase health inequalities. For example, if people with more time and resource available are more likely to utilise a new service, then its introduction can increase inequalities in health in a way that is avoidable, remediable and considered unjust.

To effectively improve public health, and avoid increasing inequalities, we need a good understanding of the scientific evidence, which is mostly published as journal articles. Unfortunately, since there is a vast quantity of such articles, it is difficult to identify the key information needed. Researchers typically spend many months sifting through the literature in a systematic way to find all relevant studies. This has serious consequences because it restricts decision makers’ ability to make use of the research evidence. This is simply not an option in public health emergencies, such as the COVID pandemic, where time is of the essence and decisions must be made in days, or even hours.

The past ten years have seen significant advances in the use of Artificial Intelligence and Machine Learning to deal with the problem of accessing and interpreting the research literature to support decision-making. These approaches are effective in helping find evidence in clinical areas, but have not yet been applied to the problem of improving access to the evidence required for public health decision-making.

This project will address that gap by developing automated methods to find, organise and describe scientific literature relevant to public health and understanding its findings in relation to inequalities in health. Relevant papers will be identified from the Microsoft Academic database, one of the largest open access repositories of academic publications available. Machine learning algorithms will be used to automatically organise these documents into topics, identify the research method used and identify whether it contains information about factors related to health inequalities.

A freely available online search engine, that will continue to be available beyond the project, will be created to showcase these technologies. This will provide users with a coherent overview of the public health research literature in a way that allows them to explore it efficiently and understand the inequality foci associated with different areas of research. The project’s outcomes will be of use to a range of professionals involved in public health decision-making including public health researchers and funders, systematic reviewers, guideline developers and local public health policy-making teams as well as other stakeholders such as groups concerned with reducing health inequalities and patients and the public. We have plans to engage with these groups throughout the project, in order to ensure the work accurately reflects their perspectives and meets their needs.

## Background and Scientific Rationale

Timely access to the research literature is vital for informing a wide range of public health decisions. However, a number of important barriers exist. Public health interventions are often complex, and applied in diverse and rapidly evolving settings.(1,2) Often a very large number of studies of diverse

designs need to be considered (often thousands but sometimes hundreds of thousands).(3–5) Health inequalities have been defined as differences in health which are avoidable, remediable, and considered unjust.(6) Public health interventions risk exacerbating these inequalities,(7) and outcomes can be difficult to predict without a thorough understanding of the research literature. Understanding public health as a complex interconnected system increases the need to look holistically at the entirety of relevant evidence,(8,9) though also increases the volume of research to be examined. It is this problem, of how to quickly make sense of a large volume of heterogenous literature containing information about public health and inequalities, which this proposal targets.

The past ten years have seen increasing research attention on the use of computers to automate research synthesis through the application of methods from Artificial Intelligence and Machine Learning. This includes work by the applicants on using machine learning to speed up laborious parts of systematic review conduct, including search and study appraisal,(10–13) data extraction,(14) and synthesis.(15,16) To date, these methods have been largely focused on identifying and appraising randomized controlled trials (RCTs), and usually in the context of speeding up systematic reviews of medical interventions. **We aim to develop and evaluate automated research synthesis methods to identify, organise and classify the public health evidence base according to its impact on inequalities.** Public health research is a larger and more diverse collection of scientific literature that has been evaluated in prior work. We will develop techniques to identify health inequalities discussed within public health evidence using the PROGRESS-Plus framework and apply machine learning models that automatically identify common themes and topics discussed within a collection of documents. We will publish an online database that will allow decision makers get to grips with a large volume of research on public health topics rapidly including implications for health inequalities. All project outputs will be open access and open source, and we have a sustainability plan in place to enable the tool to be maintained beyond the lifetime of this project. (N.B. elements of this system have already been developed - see 'existing work' below.) As well as being a valuable resource and methodological advance in its own right, we see this work as being a much needed 'corrective' to the way that automation techniques are currently focused on medical interventions. The challenges in public health are arguably greater, but this area has received much less attention, increasing the relative costs of doing evidence synthesis of public health research in comparison to more clinical areas.

## Research Plan

This study seeks to develop methods to apply machine learning and Natural Language Processing approaches to support the review, assessment, evaluation and summarisation of large volumes of public health research to support decision making. It will develop and apply automatic methods for identifying information about inequalities, study types and common themes mentioned within large volumes of public health research. The output of these techniques will be made available through an online tool containing a continuously updated repository of public health research. Users will also be able to upload their own data for processing and download results.

The key research question we aim to address is:

*Can text mining be used to maintain a 'living' database of public health research, including information about topics, methods and inequalities?*

We will address this research question by developing a 'living' database of public health research in collaboration with public health decision-makers, researchers, and patients and the public. The database will be populated by identifying public health records from the >240 million records in Microsoft Academic Graph and will therefore be a comprehensive repository. It will also be a 'living' database, as it will be updated every two weeks with newly published research. This comprehensive and high-quality dataset offers the right context for addressing our methodology research question, as while it can contain nearly all the world's published work in public health, the data will be unavoidably 'noisy', requiring the new approaches we propose to develop to become an efficient resource for our target users.

The work will be broken down into three main tasks, which are detailed below:









1. Developing and evaluating supervised machine learning methods for understanding the health inequalities perspectives of a diverse public health literature base and identifying the type of research being conducted.
2. Developing and evaluating unsupervised machine learning methods for identifying the topics covered by research.
3. Integrating these methods into a new freely available online tool, to allow users to interact with the literature. The user interface will allow automatic ranking of documents by relevance to public health, the specific health topic, study design, health inequalities, and geographical location.


Throughout each task we will engage with stakeholders and patients to ensure that the methods and tools meet their needs.

## Task 1: Identification of health inequalities and research type

In this task we propose to focus on classifying research in two ways: first, according to the “PROGRESS-Plus” criteria, and second according to the type of study that is being described.

**PROGRESS** refers to:

-  Place of residence
-  Race/ethnicity/culture/language
-  Occupation
-  Gender/sex
-  Religion
-  Education
-  Socioeconomic status
-  Social capital

 **Plus** refers to:

- 1) personal characteristics associated with discrimination (e.g. age, disability)
- 2) features of relationships (e.g. smoking parents, excluded from school)
- 3) time-dependent relationships (e.g. leaving the hospital, respite care, other instances where a person may be temporarily at a disadvantage)

### Box 1. The PROGRESS-Plus criteria<sup>1</sup>

Currently, researchers assessing health inequalities in the health literature conduct manual searches of biomedical databases (e.g. MEDLINE and Embase) using keyword terms, yielding a set (often numbering in the tens of thousands) of abstracts which require further manual sorting.(17) The system we aim to develop will go further, aiming to produce a higher precision set of articles (i.e. fewer irrelevant articles need manual assessment), with automatic assessment of the *type* of inequality being assessed, together with information on study design. All automatic classifications will be available as predicted probabilities, allowing instant ranking of documents to prioritise those with highest relevance.

PROGRESS-Plus is a conceptual framework that enables researchers to understand the social and personal factors which may be influencing health opportunities and outcomes (see Box 1).(18) Many

<sup>1</sup> <https://methods.cochrane.org/equity/projects/evidence-equity/progress-plus>

systematic reviews have applied an ‘equity lens’ to their analysis through applying this tool to the studies included, and it is recommended by the Campbell and Cochrane Equity Methods Group.(6) In addition, we will also extend existing tools (13) to classify each research record according to the type of study being described; for example, observational, qualitative study, randomized controlled trial.

## Methods

### Assembly of data

We will develop a system for finding and categorising research relevant to health inequalities for use in a supervised machine learning workflow. These machine learning models require ‘training’ data for development and evaluation. This dataset will be constructed through re-using data from published systematic reviews in public health that have used the PROGRESS-Plus criteria. In preparatory work we have identified over 80 relevant reviews (estimated to include more than 1000 studies). These reviews contain two types of relevant data: first, binary judgements as to whether the study addresses a particular inequality; second, a snippet of text (typically a sentence or two) taken directly from the original study text to justify the judgement (known as the rationale). In many cases, the data we need are available in the ‘characteristics of included studies’ tables included as appendices in published reviews. In some situations, we will contact authors to ask for the relevant tables (often saved in Excel or other tabular format). We note that article training will be done on annotated full text documents (i.e. PDFs of journal articles), while evaluation and real world application will be on research abstracts. This approach was chosen to allow us to leverage the large amount of existing data for training, rather than labelling new texts from scratch – the cost of which would be prohibitive; this also allows us to make use of the unparalleled dataset in Microsoft Academic Graph. We have used the same approach previously in assessing the risk of bias in research abstracts, and found that predicted probabilities were accurate and produced highly useful document ranking information (c statistic 0.8, Brier score 0.1).(15)

A combination of ‘data wrangling’ from published data and author contact is expected to yield a dataset of approximately 1,000 study abstracts some with associated PROGRESS-Plus text spans and relevance labels. We expect these data to allow training and evaluation of machine learning models; and will also be made available publicly for other groups to build upon.

As well as the PROGRESS-Plus criteria, we will classify research in our database according to the type of study being reported. We already have high-performing classifiers to identify randomized controlled trials,(13) systematic reviews and economic evaluations and will supplement these with further classifiers for qualitative research, observational studies, case series, clinical guidelines, editorials, and individual narrative accounts. The data to be used will be from PubMed, with labels the MeSH headings that have been used to catalogue the records.

The final set of data to be assembled will be the records that we use to ‘model’ the domain of public health in order to distinguish records in this domain from others in Microsoft Academic. These data will be identified from a range of sources: existing EPPI-Centre public health databases, Epistemonikos, PubMed, lists of journals that publish public health research, and using MeSH terms. New records entering the system will then be ranked according to their ‘distance’ from our public health records, where distance is computed according to the terms used, and their citation, journal and author networks.

### Computational methods and evaluation

We will divide the labelled data randomly into ‘training’, ‘development’, and ‘evaluation’ subsets (representing 80%, 10%, and 10% of the full dataset respectively). The training and development sets will be used for developing model variants, and setting hyperparameters, with the evaluation set being withheld for that purpose. For classifications around study design and PROGRESS-Plus items, we will evaluate logistic regression, support vector machines, and neural classification models (BERT [Bidirectional Encoder Representations from Transformers], and variants of BERT pretrained on the scientific literature). We will evaluate binary classification accuracy (precision and recall), and

calibration accuracy of predicted probabilities (calibration curves, Brier score, and C statistics), in all cases with regard to the withheld evaluation dataset.

For extracting PROGRESS-Plus text spans, should sufficient training data be developed, we will evaluate the performance of various sequence labelling approaches for this task. Specifically, we have used previously a LSTM-CRF (Long Short-Term Memory neural network, with a Conditional Random Field layer), and BERT for a related sequence labelling task (extracting the Population, Interventions, and Outcomes; or 'PICO' elements from research abstracts), with good accuracy.(15) We will use an automatic concept extraction tool we have developed previously to extract structured vocabulary data from the text spans.(15) This system accurately translates free text strings into terms from vocabularies in the Unified Medical Language System (UMLS), and specifically into terms from SNOMED CT [Systematized Nomenclature of Medicine Clinical Terms], RxNorm, MeSH, MedDRA [Medical Dictionary for Regulatory Activities], and the World Health Organization ATC [Anatomic Therapeutic Chemical] classification system. These vocabularies are used by the Cochrane linked data project (<https://linkeddata.cochrane.org/>) and give good coverage for classifying research articles. This will allow structured search across the database for articles addressing particular areas of inequality.

## Task 2: Organising and understanding research literature

In addition to the type of study and PROGRESS-Plus classifications, users need a user-friendly way of identifying the topics of the research they are interested in finding. These might cover the type, or age, of population, the context or location of intervention (if any) and outcomes being assessed. Moreover, evidence needs to be interpreted and understood before it is useful for decision making.

Sensemaking, the process of organising information into an overall structure, is a key part of this process but is time consuming and slow for large collections of evidence.

Topic models use statistical analysis to identify the underlying themes being discussed and offer an automatic method for supporting sensemaking. They have been applied to a wide range of problems within text processing where they have been shown to be useful for a wide range of text processing tasks, extremely robust (19) and developing interfaces that support sensemaking, e.g. (20–23).

The project will apply topic models to identify the underlying structure in large collections of public health evidence so that they can be presented in a way that is straightforward to interpret. The models will identify the major themes discussed within a collection of research literature and create easily interpretable labels for each theme (e.g. “low income settings” and “pre-hospital care”). We build on previous experience of applying topic models to a diverse range of text types, e.g. (23–26).

### Methods

The process of building topic models is an ‘unsupervised’ machine learning approach so, unlike the assignment of PROGRESS-Plus and study type categories, this does not require a labelled dataset for model training; it will operate using the titles, abstracts and any ‘keywords’ present in the records. MeSH terms and the automatically generated ‘field of study’ terms from Microsoft Academic will be used to group records thematically prior to topic modelling.

**Topic inference.** The first stage in the analysis will be to identify the main themes mentioned in a collection of documents by applying a topic model, such as Latent Dirichlet Allocation.(27) The raw outputs from topic models can be made significantly easier to interpret by applying post-processing steps. Previous applications of topic models have shown that a small portion of the automatically created topics can be difficult to interpret (28) and, although uncommon, such topics can be confusing if presented to a user. We will apply our automatic methods based on distributional semantics to identify these topics automatically.(24) They will then be removed from the set generated by the topic model and not shown to the user. Topic models can also generate similar topics, particularly when a large portion of the documents in the collection are associated with closely related topics. We will identify these using our existing approaches, also based on distributional semantics, and combine these topics in the interface.(25)

**Topic representation.** The most common method for representing topics is to show the 10 keywords with the highest probability associated with each topic. For example, the keywords {*pain, disorder, symptom, depression, anxiety, patient, chronic, depressive, study, psychiatric*} may be used to represent documents discussing psychiatric disorders. It has been shown that associating meaningful labels with topics can reduce the cognitive load required to interpret them.(23) Researchers have developed a range of approaches to creating more interpretable topic labels including selecting the most appropriate keyword in the top 10, (29) identifying suitable images, (24) identifying the most discriminative keywords, (30) and generating short textual labels, e.g. (25,26,30). We shall experiment with two such techniques: reordering keywords (30) and generating short textual labels (26) to create meaningful labels for the automatically generated topics. The most successful of these will be used in the online interface.

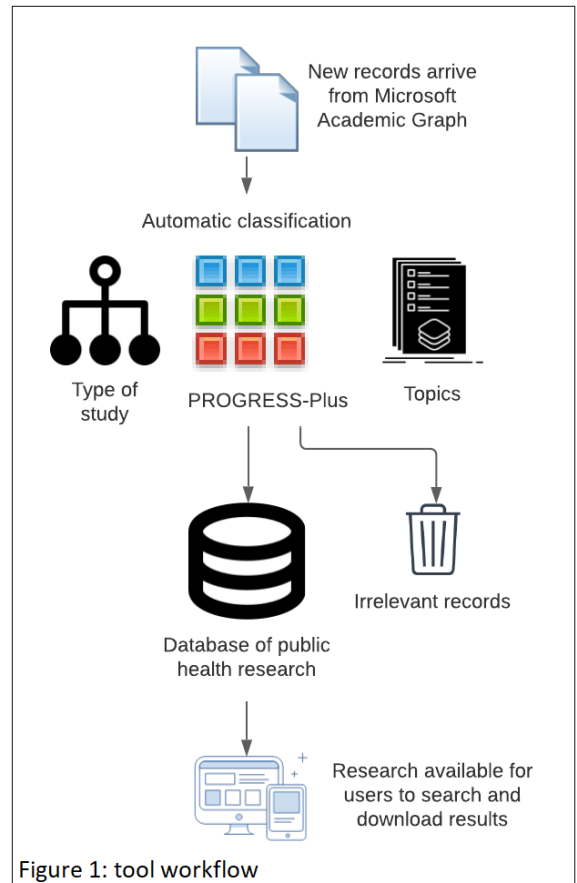
### Task 3: Development of Online Tool

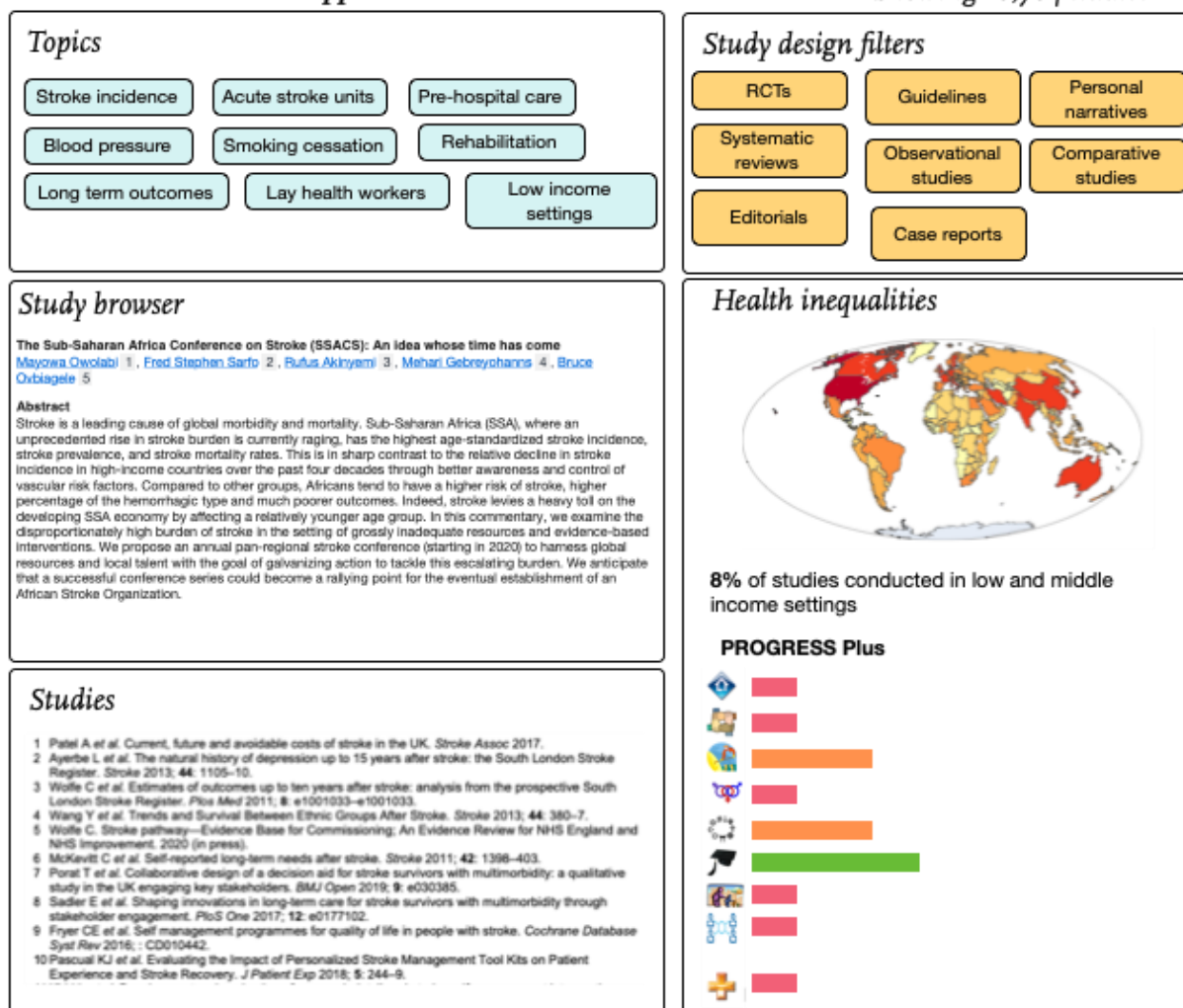
The best performing models from the previous tasks will be made available in the online tool, with the input of the Stakeholder group to guide the user interface and functionality. The tool will provide information on individual studies, and also across the full set of studies via data tables and visualisations. We have recently used choropleth maps to visualise automated data extractions from clinical trials (see the 'health inequalities' section in Figure 2). Applying this approach to the public health literature would allow an overview of which regions have attracted the most, and least, research attention.

#### Methods

##### Evidence Identification and Processing

The records to populate our database will be sourced from Microsoft Academic Graph (MAG), one of the largest open access repositories of academic publications available. MAG contains records of >240 million research articles, derived from a number of sources including biomedical databases (PubMed and others), and is updated every two weeks. We have found in preliminary work that MAG has excellent coverage for studies included in public health systematic reviews; and the single MAG dataset is comparable to using multiple biomedical databases (being conventional practice in systematic reviews). Records will be enriched by applying the methods developed in the previous tasks and stored in a database.





**Figure 2: Example of user interface of new tool, showing automatic topic mapping (top left), study design filter (top right), inequalities browser (bottom right), and study browser (bottom left)**

### Development of Tool Interface

Building on an existing tool, we will build a user interface to present the information identified using the methods developed earlier in the project in an easily interpretable way and allow users to interact with the evidence. A representation of the interface as currently envisioned is shown in Figure 2 (the final version will be adapted and built based on stakeholder feedback). The interface will contain multiple types of information, including the following:

- **Inequalities data.** Users will be provided with information about the health inequalities (within the PROGRESS Plus framework) discussed within the collection of documents. They will also be able to view this information for an individual topic or single document, thereby allowing them to compare the variation in discussion of inequality factors across the evidence.
- **Automatic topic mapping.** Users will be able to view the set of topics derived from the collection (e.g. “Stroke incidence”, “acute stroke units” in Figure 4) to provide them with a summary overview of the general themes discussed within the collection. They will also be able to access a list of studies associated with each topic (see bottom left of Figure 4), providing access to this information.
- **Study design filter.** Users will have the option to filter studies by their type (e.g. “systematic review”, “guideline”).

- **Study browser.** Users will be able to view information about individual studies (e.g. publication details), the topics associated with the study and details about health inequality information mentioned within them.

## Existing work and team track record

We have outlined an ambitious schedule of tasks that would not normally be possible to deliver within a project lasting less than a year. Fortunately, the major architectural elements of the online tool have already been built and can be utilised in this project. As a result of previous work (including the Human Behaviour-Change Project funded by Wellcome Trust (31)) we already have the main elements described in Figure 1 in place — apart from the three components under ‘automatic classification’, which will be created in Tasks 1 and 2. The user interface also already exists, though we expect to develop it in accordance with the needs identified in this project and the specific requirements inherent in publishing a large database of this type.

More broadly, the team is well-placed to do this work. The EPPI-Centre information science team, led by James Thomas, already publishes a suite of online tools to support evidence identification and synthesis. These include: EPPI-Reviewer, software for conducting systematic reviews; EPPI-Mapper, software for visualising ‘maps’ of research; and two databases containing systematic reviews and trials in public health. EPPI-Reviewer is one of the tools used by Cochrane and NICE, and the EPPI-Mapper software is used by numerous organisations internationally including the Campbell Collaboration. The EPPI-Centre also develops and maintains a set of machine learning APIs that are integrated within Cochrane’s Central Register of Studies.(13)

Members of the EPPI-Centre team have been maintaining a ‘living’ database of COVID-19 research throughout the pandemic.(32) Updated weekly, and downloaded tens of thousands of times over the last few months, this tool is now using the new workflow of records coming from Microsoft Academic and applying machine learning to identify those most relevant to the map. The work proposed in this project requires more sophisticated machine learning techniques, but the COVID-19 ‘map’ can be considered as a proof of principle demonstration of the approach we propose to use in this project.

Claire Stansfield is an information specialist at the EPPI-Centre who specialises in conducting, developing and evaluating strategies for locating ‘hard to find’ public health research evidence. Patrick O’Driscoll is a senior research software developer with expertise in the range of technologies we propose to use, and Lionel Openshaw is UCL Social Research Institute’s graphic designer, who works on publications and digital design.

Iain Marshall is a clinical academic at King’s College London with a focus on automated methods for research synthesis. He co-leads the RobotReviewer project,(12) which focuses on the evaluation of artificial intelligence methods for speeding up evidence synthesis. Software from this project (RobotReviewer and RobotSearch) has been made available as open source, and is currently being used by NICE to speed up their evidence surveillance. With James Thomas he led the ‘State of the evidence’ study, which used machine learning methods to assess global health disparities in RCTs, which has recently been published in BMJ Global Health.(16)

Mark Stevenson and Suzy Paisley co-lead Sheffield’s TePHTA group which aims to improve healthcare decision making through the application of text processing technologies. Mark Stevenson works on applications of topic models and the development of tools to support the interpretation of large document collections, including scientific publications.(10,23) Suzy Paisley’s interests are in the design of novel search and review methods for decision-analytic models of cost-effectiveness evidence and systematic reviews of complex policy topics.(33)

## Dissemination, Outputs and anticipated Impact

Results of the project will be disseminated through two main routes: (1) computational methods/resources, including a publicly available web site; and (2) academic publications.



The most significant contribution, in terms of impact, is expected to be the online tool. This will be a complete version of the system developed within the project and be freely available through a dedicated website. The website will be hosted on the web pages of UCL's EPPI-Centre, which is already widely known within the evidence synthesis community. This provides a good route for long-term sustainability; the site has been in existence for more than 20 years and is a core resource of the Centre. The tool will be integrated within the EPPI-Centre's platform of resources which is used by numerous partners, facilitating its longer-term development and maintenance. Links from the websites of project partners (King's and Sheffield) and advertising through relevant public health fora (e.g. mailing lists) will increase the webpage's visibility. The project will also release the source code for the tools developed and the dataset of studies annotated with PROGRESS-Plus information to support open science and allow other researchers to build on the technologies developed during the project.

Project results will also be disseminated through the standard route of academic publications. Two journal papers are anticipated based on the work completed during the project. The first will describe the development of automated methods to annotate papers with information from PROGRESS-Plus, including the process of dataset creation as well as training and evaluation of computational models. The second paper will describe the overall system and report its evaluation.

The main impact anticipated from the project are the development of methods and tools that support the interpretation of large volumes of public health evidence, including explicit information about inequality factors, that will enable decision-makers to identify robust evidence of intervention effectiveness (and ineffectiveness). The project will also raise the profile of text mining methods on public health evidence synthesis within the computer science community and encourage further work on the topic. Such approaches offer the potential to significantly reduce the effort required to interpret public health evidence, thereby allowing more evidence to be considered in a shorter time and, ultimately, evidence-based decisions to be made more rapidly.

## What do you intend to produce from your research?

**Data:** Dataset of PROGRESS-Plus annotated study reports, for development and evaluation of machine learning models.

**Tools:** The online database, publicly available via a website; RESTFUL API access to the service to enable developers to integrate the data and machine learning models into their own tools; source code publicly available under open source license.

**Papers:** Journal article describing tool development and evaluation; journal article describing the PROGRESS-Plus annotated dataset and model performance; Powerpoint slides that summarise the methods developed and how to use them.

## What are the possible barriers for further research, development, adoption and implementation?

Although the project has been designed to be a complete program of research that will generate valuable outputs in itself, there are also ample opportunities for extension. An example would be conducting trials of the tools developed within relevant user communities (e.g. public health researchers) in order to evaluate its effectiveness and identify the enhancements that would be most useful. The tool could also be extended with the addition of further text analysis methods that provide users with additional information (e.g. the findings of the research). Much of what is required for this is readily available including the necessary data (large repositories of research publications) and Natural Language Processing algorithms that can be applied to it. However, there are still some barriers to further research and impact. There has been some reluctance to make use of tools based on Natural Language Processing within the evidence synthesis community, although this appears to be decreasing given the accumulation of evidence of the significant reductions in workload that can be achieved by applying these approaches, including those developed by the applicants. Another potential barrier is access to the relevant user groups for the technologies developed. This is mitigated

through the proposer's contacts with large groups of potential users for the project outputs and fact that it will be made available through a web service. Finally, obtaining funding also represents a practical barrier to further work. The interdisciplinary nature of the work (which includes contributions from the fields of public health and Computer Science) provides a number of routes through which this might be obtained.

## What do you think the impact of your research will be and for whom?

Evidence-based decisions should be made on the basis of an understanding of the range and totality of relevant research available because individual studies may be atypical and biased. Moreover, when taking a holistic view of evidence in order to assess the potential efficacy of intervention in a complex (and adaptive) system, a wide range of research may need to be consulted. Given that interventions can increase, as well as decrease, inequalities, it is vitally important that we take better account of this when deciding which interventions should be commissioned and implemented. In short, before embarking on new research or implementing a new policy, we should be able to take stock of what we already know. Decision-makers' ability to do this though is hampered by the sheer volume of public health research available, the length of time it takes to distinguish between relevant, and irrelevant, studies, and the fact that current resources simply do not provide the means to understand the existing evidence in terms of the impact possible options may have on inequalities.

This research will contribute to alleviating the above problems for public health decision-makers, practitioners, guideline developers, research commissioners and researchers. For those needing to find research quickly, it will provide a means of accessing relevant research precisely. It will enable research commissioners to better understand what has already been done, and it will enable systematic reviewers to reduce the quantity of irrelevant research they currently need to sift through.

We also aim for the research to raise the profile of the PROGRESS-Plus framework for understanding the impact of interventions on inequalities with two down-stream impacts. First, we hope to support the critical analysis of public health interventions in terms of the dimensions this framework provides. This, we hope, will raise awareness of the importance of considering inequalities when designing and evaluating interventions. Second, we hope that the use of this framework will feed through into the future reporting of research studies, and that authors will analyse and report explicitly on the impact their interventions might have had on inequalities. Finally, we hope that the use of a systematic framework for understanding inequalities will feed through, in time, into major classification systems such as MeSH, so that research is classified – and therefore more easily findable and analysable – according to dimensions of possible inequality.

This work also aims to raise the profile of public health with computer and data scientists who are developing automation tools to identify and synthesise research evidence. Public health research has been neglected in this area and this project aims to make some progress in remedying this situation. By providing the datasets and source code that are developed in this project, and by engaging with this community when disseminating results, we will foster a greater awareness of the challenges and opportunities that are present in public health.

## Project management, governance and patient and public involvement

The project will be led by UCL. A management committee will be formed consisting of a representative from each partner (Thomas for UCL, Marshall for King's and Stevenson for Sheffield). The management committee will ensure that the project runs according to schedule, scientific risks are mitigated and impact maximised.

Regular video conference calls between project partners will be scheduled each fortnight to discuss and monitor progress. In addition, three longer project meetings will be held to facilitate deeper discussion and project planning. These meetings will be held face-to-face in either London or Sheffield if possible (given the current pandemic) or substituted with extended video conference calls if not.

Day-to-day internal communications will be undertaken by email, video conferencing and telephone, as needed. Web-based storage will be used to share key information relevant to the project including internal documents, research papers, data sets, prototype tools and meeting minutes.

We will convene a stakeholder group containing expertise in public health decision-making at local and national level, research commissioning, academics and members of civil society and campaign groups who are working on health equity. This group will meet virtually three times during the project: at its outset to inform the work as a whole; when initial results from the research are available, along with a prototype system; and towards the end, when all elements of the research system are in place.

In preparation for the first meeting, we will undertake the work already outlined to identify and classify systematic reviews which have previously used the PROGRESS-Plus tool for describing their included studies. The group will be invited to consider how well the tool – and how it has been used – aligns with their experiences, and whether additional dimensions, or configurations are needed. We will also present our initial user interface to the database at this meeting both to seek input into its development, and also to offer it as an ongoing resource to participants and their organisations. This engagement may result in a revised schema for describing factors related to health inequalities, which will also be useful for work beyond that of this project. These meetings will allow us to understand the needs of potential users, and to discuss the technical possibilities and limitations, to ensure the final tool is both useful and achievable. Once the near-final version of the database is live, we will again invite these groups to meet to input into the final ‘tweaks’ that we make to the system before the end of the year. As the tool is live online, we will include a feedback facility that users can use to send the team comments at any point.

In a similar way as we outlined above, we propose to hold a patient engagement meeting with the Patient Participation Group (PPG) at the Greyswood Practice, south London to obtain the views of patients and the public. This is an active and diverse group of patient representatives, who meet regularly (virtually during current pandemic) to improve the care of the local population. We will host this meeting in month four or five, to allow a presentation of an early version of the tool and the data we have collected, which would allow public contribution to the development, design, and outcomes, with the opportunity for us to refine the tool afterwards.

## Project / research timetable

The project will run between April and December 2021 (9 months). The project plan contains three tasks, as described in the Research Plan (see above).

Project timetable	Month								
	1	2	3	4	5	6	7	8	9
Formation of stakeholder group	x								
Stakeholder and civil society group meeting		x				x			x
Patient and public engagement meeting				x	x				
Data preparation	x	x	x	x					
Task 1: identification of health inequalities and research type	x	x	x	x	x	x			
Task 2: organising and understanding research topics	x	x	x	x	x				
Task 3: development of online tool			x	x	x	x	x	x	
Writing up results							x	x	x

## Ethics / Regulatory Approvals

The research will involve patients as a research advisory group to contribute to the design of the methods, according to the INVOLVE framework, which does not require ethical approval. There will be no involvement of patients as research participants.

The project has already passed faculty ethics approval at the time of writing.

## Success criteria and barriers to proposed work

Success criteria begin with the internal evaluation the effectiveness and appropriateness of the automation methods being developed. Once deployed, the success of the project will be determined by the use of the online system and its inclusion as a search source in evidence synthesis activities such as rapid and systematic reviews. Our objective is not to provide 'yet another' database to search, but to offer a resource that provides more information about research than can be found elsewhere.

The short timescale for the research naturally presents some risks. For example, time is limited for the creation of new datasets, but we have mitigated this by using pre-existing review data (including data already held by the EPPI-Centre), rather than spending considerable time generating new data. Likewise, we propose to use existing contacts to form the core of our advisory group, and agreement has already been given for us to consult with an existing patient and public consultation group.

We are aware that our plans require us to run millions of records through our machine learning processes, and will manage the computational demand by utilising parallelisation (e.g. Spark and/or the use of GPU machines). We already maintain the Trialstreamer database, which uses computationally intensive machine learning models to classify and process the full contents of PubMed (>30 million articles), demonstrating the feasibility of the current proposal. While many of the machine learning tasks we have planned are relatively low risk (e.g. classifying study type and topic modelling), we are aware that classifying research according to PROGRESS-Plus will be challenging. We consider the potential benefits in terms of field advancement to warrant undertaking the work and have mitigated risk here by already beginning the process of assembling the available data. This work has shown us both that we will be able to assemble sufficient data for machine learning and also that the quality of the data is high. Moreover, as outlined in 'Assembly of data', members of the team have successfully undertaken a similar task before, when modelling risk of bias characteristics in randomized trials.

## NIHR funding acknowledgement/disclaimer

This study/project is funded by the National Institute for Health Research (NIHR) Public Health Research Programme (NIHR133603). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care

## References

1. Petticrew M, Knai C, Thomas J, Rehfuss EA, Noyes J, Gerhardus A, et al. Implications of a complexity perspective for systematic reviews and guideline development in health decision making. *BMJ Glob Heal*. 2019;4(Suppl 1):e000899.
2. Higgins JPT, López-López JA, Becker BJ, Davies SR, Dawson S, Grimshaw JM, et al. Synthesising quantitative evidence in systematic reviews of complex health interventions. *BMJ Glob Heal*. 2019;4(Suppl 1):e000858.
3. Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Syst Rev* [Internet]. 2016;5(1):140. Available from: <http://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-016-0315-4>
4. Donnelly C, Boyd I, Campbell P, Craig C, Vallance P, Walport M, et al. Four principles for synthesizing evidence. 364. 2018;
5. Petticrew M, Rehfuss E, Noyes J, Higgins JPT, Mayhew A, Pantoja T, et al. Synthesizing evidence on complex interventions: How meta-analytical, qualitative, and mixed-method approaches can contribute. *J Clin Epidemiol* [Internet]. 2013;66(11):1230–43. Available from: <http://dx.doi.org/10.1016/j.jclinepi.2013.06.005>
6. Welch VA, Petkovic J, Jull J, Hartling L, Klassen T, Kristjansson E, et al. Equity and specific populations. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ WV, editor. *Cochrane Handbook for Systematic Reviews of Interventions*. 2019.
7. Lorenc T, Petticrew M, Welch V, Tugwell P. What types of interventions generate inequalities? Evidence from systematic reviews. *J Epidemiol Community Health*. 2013;67(2):190–3.
8. Rutter H, Savona N, Glonti K, Bibby J, Cummins S, Finegood DT, et al. Viewpoint The need for a complex systems model of evidence for public health. *Lancet* [Internet]. 2017;6736(17):9–11. Available from: [http://dx.doi.org/10.1016/S0140-6736\(17\)31267-9](http://dx.doi.org/10.1016/S0140-6736(17)31267-9)
9. McGill E, Er V, Penney T, Egan M, White M, Meier P, et al. Evaluation of public health interventions from a complex systems perspective: a research methods review. *Soc Sci Med* [Internet]. 2021;113697. Available from: <http://www.sciencedirect.com/science/article/pii/S0277953621000290>
10. Alharbi A, Stevenson M. Refining Boolean queries to identify relevant studies for systematic review updates. *Journal of the American Medical Informatics Association*. 2020.
11. Alharbi A, Stevenson M. A Dataset of Systematic Reviews Updates. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Paris, France; 2019.
12. Marshall I, Kuiper J, Wallace BC. RobotReviewer : evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Informatics Assoc*. 2015;1–10.
13. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *J Clin Epidemiol* [Internet]. 2020; Available from: <https://doi.org/10.1016/j.jclinepi.2020.11.003>
14. Marshall IJ, Kuiper J, Banner E, Wallace BC. Automating biomedical evidence synthesis: Robotreviewer. In: *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*. 2017.
15. Marshall IJ, Nye B, Kuiper J, Noel-storr A, Marshall R, Maclean R, et al. Trialstreamer: A living, automatically updated database of clinical trial reports. *J Am Med Informatics Assoc*. 2020;00(0):1–10.
16. Marshall IJ, L'Esperance V, Marshall R, Thomas J, Noel-Storr A, Soboczenski F, et al. State of the evidence: a survey of global disparities in clinical trials. *BMJ Glob Heal*. 2021;6(1):e004145.
17. Prady SL, Uphoff EP, Power M, Golder S. Development and validation of a search filter to identify equity-focused studies: reducing the number needed to screen. *BMC Med Res Methodol*. 2018;
18. O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. Applying an equity lens

- to interventions: Using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. *J Clin Epidemiol*. 2014;
19. Boyd-Graber J, Hu Y, Mimno D. Applications of topic models. *Found Trends Inf Retr*. 2017;
  20. Chaney AJB, Blei DM. Visualizing topic models. In: *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. 2012.
  21. Hinneburg A, Preiss R, Schroder R. TopicExplorer: Exploring document collections with topic models. In: *In Machine Learning and Knowledge Discovery in Databases*. 2012. p. 838–841.
  22. Ganguly D, Ganguly M, Leveling J, Jones GJF. TopicVis: A GUI for topic-based feedback and navigation. In: *SIGIR 2013 - Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2013.
  23. Aletras N, Baldwin T, Lau JH, Stevenson M. Evaluating topic representations for exploring document collections. *J Assoc Inf Sci Technol*. 2017;
  24. Aletras N, Stevenson M. Evaluating topic coherence using distributional semantics. In: *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013 - Long Papers*. 2013.
  25. Aletras N, Stevenson M. Labelling topics using unsupervised graph-based methods. In: *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*. 2014.
  26. Alokaili A, Aletras N, Stevenson M. Automatic Generation of Topic Labels. In: *In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020.
  27. Blei DM. Probabilistic topic models. In: *Communications of the ACM*. 2012.
  28. Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei DM. Reading tea leaves: How humans interpret topic models. In: *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*. 2009.
  29. Lau JH, Newman D, Karimi S, Baldwin T. Best topic word selection for topic labelling. In: *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*. 2010.
  30. Alokaili A, Aletras N, Stevenson M. Re-ranking words to improve interpretability of automatically generated topics. *arXiv*. 2019.
  31. Michie S, Johnston M, Thomas J, Aonghusa P Mac, West R, Kelly MP, et al. The Human Behaviour-Change Project : An artificial intelligence system to answer questions about changing behaviour [ version 1 ; peer review : not peer reviewed ]. *Wellcome Open Res*. 2020;1–5.
  32. Lorenc T, Khouja C, Raine G, Shemilt I, Sutcliffe K, D'Souza P, et al. COVID-19: living map of the evidence. London: EPPI-Centre, Social Research Institute, University College London; 2020.
  33. Paisley S. Identification of Evidence for Key Parameters in Decision-Analytic Models of Cost Effectiveness: A Description of Sources and a Recommended Minimum Search Requirement. *PharmacoEconomics*. 2016.