NIHR Policy Research Programme Reviews Facility

Supporting national policy development and implementation

Generative Large Language Model-Based Tools for Health and Social Care Applications: A Living Map and Critical Review (*Protocol*)

Ian Shemilt, Gareth Hollands, Claire Khouja, Gary Raine, Dylan Kneale, Alison O'Mara-Eves, Katy Sutcliffe and James Thomas

2024

NIHR Policy Research Programme Reviews Facility is a collaboration between:

EPPI Centre
Evidence for Policy & Practice

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

UNIVERSITY of York

Generative Large Language Model-Based Tools for Health and Social Care Applications:

A Living Map and Critical Review (*Protocol*)

Authors:
Ian Shemilt, Gareth Hollands, Claire Khouja, Gary Raine, Dylan Kneale,
Alison O'Mara-Eves, Katy Sutcliffe and James Thomas

June 2024

The authors of this protocol are:

Ian Shemilt[1], Gareth Hollands[1], Claire Khouja[2], Gary Raine[2], Dylan Kneale[1], Alison O'Mara-Eves[1], Katy Sutcliffe[1] and James Thomas[1]

[1] EPPI Centre, UCL Social Research Institute, University College London, United Kingdom.

[2] Centre for Reviews and Dissemination, University of York, United Kingdom.

Funding

This work has been commissioned by the National Institute for Health Research Policy Research Programme (NIHR PRP) for the UK Department of Health and Social Care (DHSC). It is being funded through the NIHR PRP contract with the EPPI Centre at University College London (UCL): Reviews facility to support national policy development and implementation, NIHR200701. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the DHSC.

Funder involvement

This work is part of an ongoing programme of research funded by the DHSC (see also 'Contributions', below). DHSC policy team members will be consulted at various stages to help us to understand the policy context of the issue under study and identify the focus and scope of the analysis.

Conflicts of interest

None.

Contributions

This work is being undertaken by a team of researchers based at the EPPI Centre and the Centre for Reviews and Dissemination (CRD) under the auspices of the UK NIHR Policy Research Programme Reviews Facility (NIHR PRP Reviews Facility). The NIHR PRP Reviews Facility is a collaboration between the EPPI Centre, University College London, the Centre for Reviews and Dissemination (CRD), University of York, and the Department of Public Health, Environments and Society (PHES), London School of Hygiene and Tropical Medicine.

Suggested citation: Shemilt I, Hollands GJ, Khouja C, Raine G, Kneale D, O'Mara-Eves A, Sutcliffe K, Thomas J (2024). *Generative Large Language Model-Based Tools for Health and Social Care Applications: A Living Map and Critical Review (Protocol)*. London: EPPI Centre, Social Science Research Unit, UCL Institute of Education, University College London.

ISBN: 978-1-911605-57-7

CONTENTS

Abbreviations 5

Terms used in this protocol 6

1 BACKGROUND 7

1.1 Generative large language model-based tools 7

1.2 Health and social care applications 8

1.3 Why it is important to map and review this evidence 9

2 AIMS, SCOPE AND METHODS 10

2.1 Aims and scope 10

2.1.1 Living evidence and gap map 10

2.1.2 Critical review 11

2.2 Methods 13

2.2.1 Living evidence and gap map 13

2.2.2 Critical review 27

2.3 Equity issues 34

2.4 Stakeholder engagement 36

2.5 Research ethics 37

2.5.1 Ethical issues related to the conduct of this project 37

2.5.2 Ethical issues related to the use of generative LLM-based tools in the field 37

3 REFERENCES 38

3.1 Appendix 1. MEDLINE (Ovid) and Embase (Ovid) search strategies 45

3.2 Appendix 2. List of specific generative large language models 48

3.3 Appendix 3. Equity issues checklist 50

Abbreviations

AI	Artificial intelligence
EGM	Evidence and gap map
LLM	Large language model
LAM	Large audio model
RAG	Retrieval augmented generation

Terms used in this protocol

Generative large language model-based tools	A class of artificial intelligence tools capable of generating text, images, audio and/or other media in response to user- or self-prompts.
Corpora	Corpus (pl. corpora). A large collection of (labelled or unlabelled) text, images or audio organised into a dataset, on which the LLM is trained, which is used by the model to infer and generate new content in response to inputs to LLM-based tools from user- or self-prompts
Parameters	Variables which comprise the corpora (see above) on which the LLM (or LAM) is trained.
Reinforcement learning	An approach to training and fine-tuning the LLM, which involves providing human feedback in the form of rewards or penalties, with the aim of enabling the model to adjust its outputs to better align with human instructions or intentions.
Deepfakes	Audio, images and videos that use digital manipulation to create or alter people’s facial appearance and/or speech, including attributes, identity, expression, or language, using image and/or audio generating LLM- and/or LAM-based tools.
Retrieval augmented generation	A method for enhancing the reliability or usefulness of outputs generated by LLM-based tools (LLMs) by anchoring the model to external knowledge sources, to complement the model's internal representation of information.

1 BACKGROUND

1.1 Generative large language model-based tools

Generative large language model- (LLM) based tools are a class of artificial intelligence (AI) tools capable of generating natural language text, images, audio[1], or other media[2] in response to user- or self-inputs or ‘prompts’. These tools can generate content - that is, perform 'generative decoder operations’ - because generative LLMs can process (‘encode’) - and thereby ‘learn’ to predict (or ‘decode’) - the syntax, semantics, and patterns of natural language (or the patterns of images or audio): that is, they are capable of ‘next word prediction’.

The prefix ‘generative’ distinguishes this class of LLM-based tools from another class that only encode, but do not decode or generate, text, images, audio, or other media content[3]. Encoder-only language model-based tools have been used in health care applications for several years; for example, for predicting oncology outcomes from structured radiology reports (see Fink 2022 for a recent example of this specific application). However, two key aspects which have changed with the advent of generative LLM-based tools are: 1) the models are (as their name suggests) much larger, that is they are trained on very large datasets (also known as ***corpora***), comprising millions, billions, or (in some cases) trillions of variables (also known as ***parameters***), using self-supervised, semi-supervised, or unsupervised machine learning techniques; and 2) ***reinforcement learning*** is being used to enable the models to appear to better ‘understand’ the context of a question (or prompt) and to natural language answers (or responses) (Karabacak 2023). Furthermore, ***retrieval augmented generation (RAG)*** can be used to enhance the reliability or usefulness of outputs generated by LLM-based tools by anchoring the model to one or more specific external knowledge sources. RAG aims to complement the underlying LLM’s internal representation of information, by ensuring the model has access to current, reliable facts and that users can access its sources.

Generative LLMs are typically designed to be general-purpose models, capable of performing a range of tasks across a wide variety of domains, without the need for task-specific training. As such, generative LLM-based tools have already found applications in diverse industries, including art, software development, product design, finance, gaming, marketing and fashion, as well as health care. Private sector multi-national companies, including OpenAI and Google, have developed generative LLM-based tools, like ChatGPT and Bard or Gemini, which have garnered widespread attention and adoption, bridging the gap between a limited computer science user base and the wider public.

However, while these models have gained a great deal of public attention and unprecedented growth in usership, due to their ability of the models to *appear* to

[1] Also known as large audio models (LAMs).

[2] For example, computer code.

[3] An example of a class of ‘non-generative’ LLMs is ‘encoder-only LLMs’, which includes Bidirectional Encoder Representation Transformer (BERT) models that are designed to learn embeddings for predictive modelling tasks, such as classification.

understand and respond to human language, generative LLMs are not yet (artificially) generally intelligent, in the sense that they are not yet capable of literal understanding nor intentionality. In fact, generative LLMs have no other capabilities beyond 'next word prediction'. Moreover, generative LLMs are not artificial either, in the sense that humans are often employed by model developers to provide manually labelled data for both training and reinforcement learning, aimed at improving the reliability of the responses generated by LLM-based tools.

Generative LLM-based tools also have other important limitations that temper the considerable excitement about their potential benefits. First, the models can produce incorrect or misleading information in an authoritative manner - a phenomenon known as 'hallucination' (Azamfirei 2023). Second, they can also generate biased content (Ferrara 2023), which may reflect and perpetuate spurious associations present in training datasets (see '2.3 Equity Issues' for further discussion of risk of bias in LLM inputs and outputs). Third, LLMs operate as black boxes, lacking transparency in explaining their responses. Serious concerns have also been raised about the potential for people to misuse generative LLMs and tools, including to propagate misinformation, and the creation of ***deepfakes*** which could deceive and manipulate people (Shahzad 2022, Khanjani 2023).

Rapid advancement of LLM-based tools therefore needs to be accompanied by careful evaluation to ensure their safe and responsible deployment. As the field progresses, there is also a need to strike a balance between innovation and the development of independent regulation and guidance to address ethical issues and concerns, alongside managing the risks of adverse effects, and/or unintended consequences, associated with using generative LLM-based tools.

1.2 Health and social care applications

Generative LLM-based tools - including those based on biomedical and/or health domain-specific LLMs (e.g. Luo 2022; Singhal 2022; Wornow 2023 - see also Section 2.2.1, '1. Model(s)'), have a wide range of potential clinical, public health or social care applications (Ali 2023, Briganti 2023, Hügle 2023, Li J 2023, Karabacak 2023, Thirunavukarasu 2023). Examples include: improving diagnostic accuracy (Wang 2023), predicting disease progression (Chen 2023, Jiang 2023), answering medical questions (Kung 2023), extracting information from patient records (Yang 2022), writing medical reports (Jeblick 2022, Patel 2023), and supporting shared decision-making between clinicians and patients (Callaghan 2023, Liu S 2023, Zhu 2023).

However, the rapid pace at which generative LLM-based tools and their applications are evolving, coupled with the lack of critical appraisal of the evidence base, make it challenging to evaluate the safe deployment of these tools in health and social care settings. Furthermore, few such tools have been thoroughly tested and validated on real-world patient data, indicating the need for further research and development tailored specifically to health care tasks (Wornow 2023).

Using generative LLM-based tools in health and social care comes with important risks of adverse and unintended impacts. These include potential harms stemming from the capacity of LLMs (and hence LLM-based tools) to produce 'hallucinations', which pose a special challenge to ensuring safe and reliable use in health and social care (Lee

2023, Au Yeung 2023), and risk of bias in the outputs of LLM-based tools, which may be experienced by, or pertain to, specific marginalised communities of people and/or inclusion health groups (NHS England 2023) (see also '2.3 Equity Issues').

To address these concerns, the Transformation Directorate at NHS England (NHSE) has released initial guidance emphasizing the need for accuracy, fairness, and transparency in AI systems (NHS England Transformation Directorate 2023). There are also growing calls for regulation, to ensure thorough evaluation and validation before active deployment (Harrer 2023) and, last year, The House of Lords Communications and Digital Committee published a call for evidence on regulation of LLMs and tools (House of Lords Communications and Digital Committee 2023).

In summary, while generative LLM-based tools offer exciting possibilities for health and social care applications, it is critical to address the risks associated with their use. Balancing the potential benefits of these models and tools with the need for rigorous evaluation and validation is therefore essential to ensure their safe and effective integration into decision-making processes in health and social care settings.

1.3 Why it is important to map and review this evidence

Decisions about whether, and when, to adopt generative LLM-based tools for health and social care applications need to balance evidence for their benefits against the associated risks of harmful or unintended consequences. A balanced understanding of the strengths and limitations of this technology is therefore essential to enable health and social care policy makers and practitioners to ask relevant questions and make informed adoption decisions about generative LLM-based tools.

Empirical research on the use and performance of generative LLM-based tools for health and social care applications has been accumulating at a fast pace, in a highly competitive and lucrative market, with high levels of investment into the development of generative LLMs and tools. This engenders the need for both continual surveillance and mapping, and a critical review, of these rapidly emerging bodies of evidence.

Mapping the evidence will establish an overview of the emerging landscape of research on LLM-based tools for health and social care applications, by representing its key features and characteristics. A critical review of the evidence can assess the validity of key claims made about the performance of LLM-based tools for specific tasks in different classes of health and social care applications, alongside investigating any financial conflicts of interest among those making such claims. This will provide an assessment of the current readiness of these technologies to perform specific tasks in health and social care. We have decided to undertake a critical review of this emerging evidence base, instead of a systematic review of intervention effects, due to the need to focus attention on investigating emergent classes of health and social care applications, as opposed to specific tools.

2 AIMS, SCOPE AND METHODS

2.1 Aims and scope

The overarching aim of this work will be:

- To empower key policy and other stakeholders with the enabling knowledge and skills needed to ask relevant questions and make informed judgments about the utility, reliability, and potential risks of generative LLM-based tools, when considering these for potential adoption for specific health and social care applications.

We will achieve this aim by producing two main outputs, in consultation with key stakeholders and advisors (see also '2.4. Stakeholder Engagement'):

1) a living evidence and gap map (living EGM); and

2) a critical review of key articles.

Further details of the specific objectives, scope and methods of each of these outputs are provided below.

2.1.1 Living evidence and gap map

We will develop and maintain a living evidence and gap map (EGM) with the following objectives:

1. To maintain a continual surveillance of the landscape of accumulating research evidence for the use of generative LLM-based tools for health and social care applications;
2. To provide a regularly updated descriptive overview of the landscape of cumulative research evidence, and gaps in the evidence base, for the use of generative LLM-based tools for health and social care applications, classified in terms of its key features and characteristics;
3. To make cumulative research evidence for the use of generative LLM-based tools for health and social care applications more findable, accessible and reusable; and
4. To compile a glossary of key terms and concepts relating to generative LLMs and generative LLM-based tools for health and social care applications.

The following types of articles (reports) will be included in the living EGM:

- Empirical research studies[4] that evaluate the performance (i.e. beneficial, adverse and/or differential impacts) of generative LLM-based tools for any specific health care (including public health) and/or social care application; and
- Non-empirical commentary and/or non-systematic review-type articles (including editorial, perspective and viewpoint articles) that discuss the use of generative LLM-based tools for any specific health care (including public

[4] Primary research, systematic review, other research synthesis, or modelling studies.

health) and/or social care application(s) based on multiple empirical research studies[5].

Our rationale for including both types of articles (reports) in the map is that the incremental value of a well-argued commentary or review-type article, grounded in empirical research, to achieving our stated overall aim (see above in the opening paragraph of this section) may sometimes be higher than a single study evaluating the performance of a single LLM for a specific, relatively narrow application.

Similarly, while we will identify and code as many eligible articles as possible within the available resource, we will not primarily aim to produce a fully comprehensive EGM that includes every report that would meet its eligibility criteria (see '2.2.1. Living evidence and gap map'). Instead, we will iteratively: (i) develop and communicate a high-level, representative descriptive and conceptual overview of the 'landscape' of relevant research evidence and (ii) select articles (reports) that reflect key health and social care applications. We will also iteratively fine-tune the level of detail of coded information presented in the living EGM, through procedures of selection, simplification, and classification. This strategy is analogous to the process of 'cartographic generalisation' (Touya 2023) which is integral to the design of maps of physical space -

With regards to scope, this living EGM will exclude related research applications, encompassing primary research, evidence synthesis (including systematic reviews), modelling, technology appraisals, and clinical or public health guideline development processes and tasks. These include potential roles for generative LLM-based tools in performing (fully- or semi-) automated tasks that contribute to the surveillance, review and/or synthesis of bodies of health and social care evidence (Liu H 2023, Knafou 2023, Qureshi 2023, Sallam 2023, Tang 2023) as they emerge, known as living evidence synthesis (Elliott 2017). In addition, artificial intelligence tools that are not generative, including encoder-only LLMs (e.g. Bidirectional Encoder Representation Transformer (BERT) models (Devlin 2018)), will be out of scope.

2.1.2 Critical review

We will conduct a critical review to critically evaluate the literature and develop conceptual understanding of the use of generative LLM-based tools for specific tasks and emergent classes of health and social care applications. A key feature of critical reviews is their focus on examining a smaller collection of studies in detail, compared with a systematic review, at the expense of identifying and considering all studies on a given topic (systematicity) (Grant 2009). This approach is also consistent with evidence synthesis strategies that employ the metaphor of a mosaic or map, where

[5] Commentary and review-type articles (reports) have the capacity to describe complex ideas about the use of LLMs for health or social care applications in a format that can be understood across a range of different readers and audiences. These kinds of outputs may be influential because they outline and explain new ideas and trends emerging within a discipline and expound on the potential benefits; alternatively, some may adopt a different perspective and seek to encourage scepticism about a particular issue among readers.

each included study adds to the complete picture, and included studies complement each other by elucidating different aspects of the interventions and phenomena under investigation (Hammersley 2002).

This critical review will be conducted with the following specific objectives:

1. To describe how generative LLMs and LLM-based tools work and consider the range of different classes of health and social care applications.

2. To assess the validity of selected evidence claims made about the performance of generative LLM-based tools for specific tasks for different classes of health and social care applications, encompassing claims about both beneficial (or intended) and adverse (or unintended) effects/ impacts.
3. To identify and highlight any evidence for biases in the outputs of generative LLM-based tools, and/or any evidence for inequities in the benefits or harmful consequences of using such models that may be attributable to other factors.
4. To identify and highlight any financial or other conflicts of interest among people making evidence claims about the performance of generative LLM-based tools for selected classes of health and social care applications.
5. To summarise implications for policy and practice concerning the readiness of generative LLM-based tools to be adopted to perform or support specific tasks for different classes of health and social care applications.

At the outset, we expect the scope of the critical review to be the same as that of the living EGM with regards to our definitions of both health and social care applications and generative LLM-based tools (see section 2.1.1, above). However, at the outset, we also reserve the option of altering the scope of the critical review at an early stage of its development, contingent on emergent findings from mapping the evidence (living EGM). If we judge that any such alterations are minor, then we will report them in the critical review as minor changes from protocol. However, if we judge that any such changes are major, then we will report them in an updated protocol that would be re-published prior to commencing the review.

2.2 Methods

2.2.1 Living evidence and gap map

This section specifies the methods, procedures, and tools that we propose to use for maintaining the living EGM. As this will be a living map, to be continually updated and regularly republished, its eligibility criteria may change as the map evolves and we encounter, discuss, and resolve 'borderline eligible' articles (reports). We have previously used similar methods, procedures, and tools to maintain other living maps, including a large living map of research evidence on COVID-19 and its long COVID 'segment' (Lorenc 2020). The living EGM will be produced using EPPI Reviewer software for systematic reviews and other evidence synthesis (Thomas 2024); including its integrated automation and EGM visualisation features and tools .

Eligibility criteria

As stated in Section 2.1.1, this living EGM will include articles (reports) of:

- Empirical research studies that evaluate the performance (beneficial, adverse and/or differential impacts) of generative LLM-based tools for any specific clinical health, public health and/or social care application (empirical study reports); and
- Commentary and/or non-systematic review-type articles (including editorials, perspective and viewpoint articles) that discuss the performance of generative LLM-based tools for any specific clinical health, public health and/or social care application(s), based on evidence from multiple empirical research studies (non-empirical articles);

Articles (reports) published as full journal articles, including pre-print articles, will be considered for inclusion. Conference abstracts will be excluded. Studies reported in articles published before 1st January 2018 will be automatically excluded. Studies reported in languages other than English will also be excluded.

Searching for eligible reports

We initially assembled a small initial corpus of potentially eligible, or borderline eligible/ ineligible, articles (reports) - several of which are cited in this protocol - identified via preliminary scoping searches. We will also conduct initial targeted Boolean searches of MEDLINE (Ovid) and Embase (Ovid), which will intentionally be designed to prioritise precision over recall (see Appendix 1 for details of search strategies).

Our strategy will be to iteratively build upon our initial corpus and precise conventional electronic searches to identify further eligible articles (reports), by deploying three kinds of automated searches of a single, potentially comprehensive source -- the OpenAlex dataset[6] -- to be conducted using OpenAlex tools, hosted in EPPI Reviewer software (Thomas 2024):

- Custom search
- Network graph search
- Auto-update search

An OpenAlex *custom search* is similar to a conventional electronic database search, as it combines target sets of title-abstract keywords and OpenAlex 'concepts' using Boolean operators. OpenAlex 'concepts' are comparable to index terms in electronic databases (e.g. MeSH terms in MEDLINE) but they are automatically assigned to each record in the dataset by a machine learning algorithm, and they are not organised in a hierarchical 'tree' structure. For this living EGM, we will use an intentionally precise OpenAlex *custom search* strategy very similar to those used for our initial MEDLINE and Embase searches (see above and Appendix 1).

An OpenAlex *network graph search* retrieves all records that are connected, in the OpenAlex dataset, to a specified set of 'seed' records, via a specified set of network graph relationships, on the date of the search. For this living EGM, the specified 'seed' records will be the accumulating corpus of unique eligible records (reports) selected for inclusion in the living EGM (which have either been retrieved from OpenAlex or matched to a corresponding OpenAlex record from a source database record); starting with our initial corpus of eligible reports. The specified network graph relationships will be: (a) one-step forwards (i.e. records that cite 'seed' records) or backwards (records that are cited by 'seed' records) citation relationships (i.e. equivalent to citation searching); and (b) a one-step forwards ('recommended by' seed records) 'related publications' relationship[7]. We will run at least two rounds of OpenAlex *network graph searches* during initial phases of study identification for the living EGM.

Like *network graph searches*, OpenAlex *auto-update searches* will be 'seeded' by the accumulating set of unique eligible records (reports) selected for inclusion in this living EGM, starting from our initial corpus of eligible reports. For *auto-update searches*, 'seed' records (reports) will be subscribed to our novel machine learning 'recommender' model (the 'auto-update model') (Tenti 2021), which will automatically score all 'new' records (reports) added to the OpenAlex dataset every ~1 month and recommend those most likely to be eligible for inclusion in the living EGM. Initially, we

[6] OpenAlex is an open access dataset and knowledge graph comprising ~250 million bibliographic records of research articles (reports) from across science, connected in a very large network graph of conceptual, citation and other (e.g. author) relationships. The OpenAlex dataset is automatically, and continuously, updated with new records as new articles (reports) are published online. We have developed OpenAlex tools (in EPPI Reviewer) to enable regular automated searches of this dataset - primarily to support (continual) semi-automated updating of (living) systematic reviews, (living) evidence maps, and related use scenarios.

[7] OpenAlex 'related publications' (records) are recommended by 'seed' publications (records) based on their ranking (top ranked are those which score highest) on a composite metric that combines all of the various network graph relationships available within the OpenAlex dataset (knowledge graph).

will update each of these three kinds of automated OpenAlex searches every ~1-2 months, but we will keep this frequency under review.

Selecting included reports and studies

Records retrieved by our searches (i.e. bibliographic records, with abstracts if available) will be imported into EPPI Reviewer. Any records (reports) published before 1st January 2018 will be removed. Duplicates will be semi-automatically identified and discarded using 'manage duplicates' tools. We will prioritise records (reports) for manual screening using active learning (Miwa 2014, O'Mara-Eves 2015) ('priority screening mode') in EPPI Reviewer. Unique records from each round of updating searches will be added to the pool of records (reports) not yet screened, to be re-prioritised using active learning.

Screening will be undertaken in a single stage, with eligibility decisions based on the title and abstract of each record (report) wherever possible, or else on the corresponding full-text article (report) where more information is required to make a judgement (where available).

In an initial pilot screening phase, at least two researchers will independently screen each included article (record/ report), and we will compare and discuss any disagreements between their respective codes, before making final decisions. The objectives of this pilot phase will be to identify revisions to screening codes, as well as to foster consistency between researchers in applying the scheme. Beyond the pilot screening phase, ne trained researcher will screen each record ('single screening') to assess its eligibility based on the criteria specified in this section, and with the further option of referring records for a 'second opinion' if eligibility is judged uncertain.

We will apply one of the following screening codes (i.e. the first applicable code in this hierarchical list) when assessing title-abstract records and/or corresponding full-text articles (reports) for potential inclusion in the living EGM:

Excluded - Published before 2018
Apply this code to any records (reports) published before 1st January 2018 (i.e. with publication year of 2017 or earlier) that have not already been identified and removed prior to assigning records for screening.

Excluded - Duplicate record (report)
Apply this code to any records (reports), that are duplicates of one or more other (eligible or ineligible) records/ reports (i.e. further duplicates not already identified and removed prior to assigning records for eligibility screening).

Excluded - Reported in a language other than English
Apply this code when the corresponding full-text report is published in any language other than English. Also use this code when the abstract (or title and abstract) is reported in a language other than English and the corresponding full text is not available.

No major focus on generative LLM-based tool(s) OR no major focus on health or social care application(s)
Apply this code to records (reports), published/ reported in pre-print or full journal articles, that *EITHER*:

- Do not include a major focus on one or more generative large language model-based tool(s). Major focus is defined as a sole, predominant or substantive focus (judgement required). Generative LLM-based tools is a class of artificial intelligence tools capable of generating text, images, audio and/or other media (e.g. computer code) in response to user- (or self-) prompts. To be eligible, tool(s) must (explicitly or implicitly/ probably) be underpinned by LLMs *and* decoder (encoder and decoder = generative) architecture (i.e. LLM must perform one or more generative decoder operation(s)). If not explicitly generative, but explicitly an LLM-based tool *and* all issues/ topics covered are relevant to generative, then eligible on this criterion (but may still be ineligible on others, below). If focus is on BERT (or closely related) models, exclude.

OR

- Do not explicitly involve a major focus on the use of generative LLM-based tools for tasks directly related to one or more health or social care applications. Major focus is defined as a sole, predominant or substantive focus (judgement required). See also: 1) UK Clinical Research Collaboration (2024). *UKCRC Health Research Classification System - Health Categories* [Webpage]. Available from: https://hrcsonline.net/health-categories/ [Accessed: May 2024]; and 2) Care Quality Commission (2024). *Care Quality Commission (CQC) - Guidance for Providers > Service Types* [Webpage]. Available from: https://www.cqc.org.uk/guidance-providers/regulations-enforcement/service-types [Accessed: May 2024]. Social care is defined as the provision of social work, personal care, protection or social support services to children or adults in need or at risk, or adults with needs arising from illness, disability, old age or poverty. Health and social care applications encompasses medical education applications (including answering medical exam questions) and other applications related to workforce development or the training of health or social care professionals, including public health professionals. Health and social care applications also encompasses patient or service user education applications.

Do not apply this code to records (reports), published/ reported in pre-print or full journal articles, that explicitly include a major focus on applying generative LLM-based tools for tasks related to the prioritisation, planning, production, of empirical research studies, encompassing systematic reviews, other types of evidence synthesis, primary research and/or modelling studies – whether or not the application is to health or social care evidence synthesis/ research: apply the next code (below) instead.

<u>Excluded - Evidence synthesis or research application</u>
Apply this code to records (reports), published/ reported in pre-print or full journal articles, with a major focus on applying generative LLM-based tools for tasks related to the prioritisation, planning, production, of empirical research studies, encompassing systematic reviews, other types of evidence synthesis, primary research and/or modelling studies; and without a major focus on the use of generative LLM-based tools for tasks directly related to health and/or social care applications. Do not apply this code to reports that include a major focus on applications which explicitly have a health and/or social care professional educational/training/clinical

development role as well as a potential research/publication function (e.g. LLM-based tool is being used for helping to write case reports in healthcare).

Excluded - Letters/ correspondence, 'research highlights' (or similar) articles, or corrections / errata explicitly linked to a probably eligible single primary research study

Apply this code to records (reports) that explicitly refer to, summarise and/or 'highlight' a single eligible primary research report (article) only - published in the same issue, or elsewhere (see for example: https://doi.org/10.1038/s44222-023-00097-7). Do not apply this code to letters / correspondence that incorporate an original/ de novo formal evaluation (empirical studies) or substantive discussion (non-empirical) of performance of generative LLM-based tools for tasks related to one or more health or social care applications.

Excluded - No formal evaluation (empirical studies) or substantive discussion (non-empirical) of performance for one or more specific tasks

Apply this code to records (reports) that explicitly involve applying generative LLM-based tools for tasks related to one or more health or social care applications but which do not incorporate a formal evaluation (empirical studies) or major focus on discussing (commentary/ review-type articles) the performance of such tools. To qualify as a (record/ report of an) empirical study with formal evaluation, the report must include one or more (implicit or explicit) research aim (or research question), a description of the methods used, and at least some results. Primarily interested in articles that make claims about the performance of eligible tools for specific tasks.

Excluded - Not a full journal or pre-print article

Apply this code to records (reports) that are not either pre-print articles or full journal articles. For example, dissertation and theses, and grey literature publications, are both excluded. Conference proceedings articles are eligible if the report is a full-text article; but excluded if the report is an abstract only. Letters to editors and/or other correspondence articles are included if otherwise eligible. Apply this code to otherwise eligible retracted articles.

Included - Primary research

Apply this code to eligible (records/ reports of) primary research studies that formally evaluate the performance (i.e. beneficial, adverse and/or differential impacts) of generative LLM-based tools for any specific clinical health, public health and/or social care application. Include reports of studies that investigate the acceptability of, or people's experiences / views concerning, (use of) generative LLM-based tools for health and/or social care applications. To qualify as a (record/ report of an) empirical primary study with formal evaluation, the report must include one or more (implicit or explicit) research aim (or research question), a description of the methods used, and at least some results. Also apply this code to reports of protocols for eligible primary studies. If the article reports both an eligible primary research component and an eligible systematic review/ other evidence synthesis / modelling component, or both an eligible primary research component and an eligible commentary or review component, then apply the 'Included – Primary research' code.

Included - Systematic review / other research synthesis / modelling

Apply this code to eligible systematic reviews, other research synthesis, or modelling studies (records/ reports) that evaluate the performance (i.e. beneficial, adverse and/or differential impacts) of generative LLM-based tools for any specific clinical health, public health and/or social care application.

Define 'systematic review or other research synthesis' as any record (report) reporting secondary data that reports: some search terms; clearly defined inclusion criteria; and some information on the selection process (at least N of references located by searches and N of studies included). If the report is an otherwise eligible review article that does not meet the latter criteria for being considered a systematic review, then code as 'Included - Commentary or review-type article' instead (see next child code, below). Include any systematic review which aimed to include eligible studies, whether or not any were located. Include updates to eligible systematic reviews, living systematic reviews, and other research syntheses, if the record (report) presents new data and the original review/ research synthesis meets the criteria above. Include records (reports) of protocols for eligible systematic reviews/ other research synthesis.

Include modelling studies which are at least partly based on relevant empirical data (e.g. data used as inputs to the model, or data against which the model is being calibrated or tested). If the article reports both an eligible primary research component and an eligible systematic review/ other evidence synthesis / modelling component, then apply the 'Included – Primary research' code.

Included - Commentary or review-type article
Apply this code to eligible non-empirical commentary and/or non-systematic review-type articles (including editorial, perspective and viewpoint articles) that discuss the use of generative LLM-based tools for any specific health care (including public health) and/or social care application(s) based on multiple empirical research studies.

If the article reports both an eligible primary research component and an eligible commentary or review component, then apply the 'Included – Primary research' code.

Unsure - Refer for a 'second opinion' (eligibility)
Apply this code to any records (reports) you are unsure about with regards to eligibility. These records will be referred for a 'second opinion' – to be resolved by team discussion and consensus (involving two or more researchers, including the original screener). When referring a record/ report for a 'second opinion', please use the corresponding 'Info' box to record (+ 'save') initialled brief details of reason(s) for uncertainty; *and* also provisionally check (tick) the multiple candidate eligibility screening codes you are uncertain between (if applicable).

We will also (initially) apply a supplementary code to 'Included' records (reports) to distinguish between those reported in pre-print articles and those reported in full-text journal articles (see below).

Screening – Included Only – Article Type

Apply one code:

- Pre-print article
- Full journal article

'Second opinions' will be resolved by team discussion and consensus, involving two or more researchers (including the original screener). We will aim to retrieve the corresponding full-text article (usually PDFs) for each included record (report) - as well as for those with uncertain eligibility based on title and/or abstract alone - using either via manual searches or using the Zotero interface in EPPI Reviewer; and we will upload each retrieved article to EPPI Reviewer.

This will be a living EGM and therefore the eligibility screening guidance notes above (version last updated May 2024) will continue to evolve as we encounter, discuss and classify further eligible, borderline eligible, or ineligible articles (reports) in this rapidly evolving field of research.

We will investigate training, calibrating, evaluating and (contingent on evaluated performance) deploying a binary machine learning classifier, designed to distinguish between records (reports) of eligible (included or excluded but eligible) and ineligible (excluded and ineligible) studies. This binary machine learning classifier would be calibrated and deployed to automatically exclude (discard) further retrieved records (reports) prior to allocating records for potential (prioritised) manual screening. Finally, we will also investigate the potential to deploy GPT-4-enabled auto-coding tools in EPPI Reviewer in a human-in-the-loop workflow, designed to facilitate efficient, semi-automated eligibility screening at scale.

Classifying included studies

We will test, revise and apply the following pilot classification (coding) scheme to each article (report) included in the living EGM. The pilot scheme comprises eight dimensions:

1. Application class(es)
2. Study type
3. Model(s)
4. Task(s) Type(s)
5. Health / social care topic(s)
6. Service type(s)
7. Population(s)
8. Map version

Further information on each dimension is provided below.

1. Application class(es) [M]

Included articles (records) will first be partitioned into broad, top-level classes of health or social care applications, as follows.

- Clinical health care. Excludes medical education / medical exams, health professional development and training, patient or service user education, and public health. Includes reports focused on patient question answering tools that explicitly culminate in, or otherwise involve, one or more clinical decision (code the latter reports as ‘Clinical health care’).

- Public health. Excludes education for public health professionals / related exams, public health professional development and training, and public health-related patient or service user education).
- Social care. Excludes education for social care professionals / related exams, social care professional development or training, and social care-related service user education.
- Medical education, health or social care professional development and training.
- Medical question answering.
    - Answering medical (and related) exam questions.
    - Answering patient or service user questions / patient or service user education. Answering patient questions about medical conditions, treatments, or health services, unless use of the tool explicitly culminates in, or otherwise involves, one or more clinical decisions (code the latter reports as 'clinical health care'). Also use this code for patient or service user education applications.
    - Answering medical questions in general (no specific application).

Articles (reports) focused exclusively on (i) medical education applications and/or (ii) medical examination question answering and/or (iii) patient or service user education or question answering applications will be partitioned into their own 'segment(s)' and no further map coding will be applied, except for '10. Map version'. The 'segment(s)' will be published separately, in parallel to the main living EGM. This does not preclude the latter classes of applications being assessed in our critical review. If an article (report) focuses on (i) medical education applications and/or (ii) medical question answering applications *and* also on eligible health, public health and/or social care applications, then further map coding will be applied based solely on the clinical health, public health and/or social care applications (and no further coding will be applied that is only applicable to medical education and/or patient or service user education applications).

Our primary rationale for treating these classes of applications separately for the living EGM (including assigning them to their own separate 'segment(s)' is that our scoping searches and study identification to date clearly indicate there are very large numbers of published reports of these classes of applications. As such, including them in our main map of health and social care applications would be problematic from a map design perspective (i.e. the size of bodies of evidence in these cells would dwarf those in other cells, making it difficult or impossible for users to discern differences between those and other cells), as well as absorbing large amounts of time and resource for further coding.

2. Article type [S]

Included articles (reports) will be segmented into three broad types:

- Primary research
- Systematic reviews, other research synthesis, or modelling
- Commentary or review-type article

3. Model(s) [M]

Each article (report) will be classified according to the specific LLM(s) under investigation. Appendix 2 contains a list of >40 generative LLMs, reflecting a snapshot of the current state-of-the-art that is unlikely to be fully comprehensive, with new and updated models continually being launched (see also https://lifearchitect.ai/timeline/ and https://huggingface.co/spaces/mteb/leaderboard). Further specific models will be added to the coding scheme as we encounter them in the current work.

- Not about specific model(s)
- BioGPT
- Claude / Claude 2
- Falcon 40-B
- GPT-2 / GPT-3 / GPT-4 / ChatGPT (OpenAI family)
- LLaMA / LLaMA-2
- Gemini / Bard
- Med-PaLM
- Mistral / Mixtral
- PaLM / PaLM 2
- PanGu-α / PanGu-Σ
- PubMed GPT
- Other(s) - specify

4. Task(s) type(s)

Select all child codes that apply, up to 3 specific types (or select ‘various tasks’ if more than 3 specific types).

We will prospectively and iteratively develop and apply a coding scheme to classify included articles (records) based on the type(s) of task(s) on which the performance of generative LLM-based tool(s) is/are being evaluated (articles reporting empirical studies) or discussed (commentary/ review-type articles).

This code set will comprise a series of concise, output-focussed[8] task descriptors (child codes). We will prospectively formulate these child codes based on the article authors’ verbatim description(s) of the task(s) and output(s). To foster consistency between descriptors, we will adjust our task descriptors to using the present participle form of the verb (i.e. ending in -ing) when necessary. Since, by definition, all included articles are focused on generative LLM-based tools, we will also avoid using the verb

[8] ‘Output-focussed’ means ‘focussed on the output(s) that is/ are generated by the LLM-based tool’ (via one or more generative decoder operations).

'generating' (present participle) in our task descriptors, as it will not discriminate between different types of tasks.

Draft output-focussed task descriptors will first be discussed by during our regular coding team 'second opinions' meetings, before being added to the code set, or discarded, based on team consensus.

- Various types of tasks. Use this code for reports focused on **more than three** different types of tasks.
- Diagnosing / recognising / detecting (e.g. symptoms)
- Extracting information
- Predicting (e.g. disease risk(s))
- Recommending / suggesting (e.g. treatments)
- Summarising / simplifying (e.g. discharge summaries, simplified radiology reports)
- Supporting / assisting decisions (e.g. clinical decision aids)
- Transcribing
- Translating
- Triaging
- Writing (e.g. medical notes, letters or e-mails)
- Other(s) – specify
- Unclear / not reported

5. Health / social care topic(s) [M]

See: https://hrcsonline.net/health-categories/

Included articles (records) will also be partitioned into broad areas of health and disease, and/or social care, based on the following topic categories (adapted from UK Clinical Research Collaboration 2023b) or 'Not applicable':

- Generic health relevance. Apply this child code when the report is judged relevant to more than 5 health categories. Also apply this code when the report is judged of relevance to all diseases and conditions or to health and well-being in general.
- Blood
- Cancer and neoplasms
- Cardiovascular
- Congenital disorders
- Ear
- Eye

- Infection
- Inflammatory and immune system
- Injuries and accidents
- Mental health
- Metabolic and endocrine
- Musculoskeletal
- Neurological
- Oral and gastrointestinal
- Renal and urogenital
- Reproductive health and childbirth
- Respiratory
- Skin
- Stroke
- Public health
- Disputed aetiology
- Adult social care
- Children's social care
- Other(s) - specify
- Unclear / not reported
- Not applicable

6. Service type(s) [M]

Included articles (records) will also be classified according to the following (level 1 and 2) service type(s) (Care Quality Commission 2023), or 'Not applicable':

- Generic relevance / various service type(s). Apply this child code when the report is judged relevant to more than 5 service types (level 2 child codes). Also apply this code when the report is judged of relevance to all service types, or to health and/or social care services in general.
- Healthcare services
    - Acute services (ACS) 1: Accident and emergency (hospital in the UK)
    - Acute services (ACS) 2: Medicine (inpatient / outpatient hospital in the UK)
    - Acute services (ACS) 3: Surgery (inpatient / outpatient hospital in the UK)
    - Hyperbaric chamber services (HBC)

  - Hospice services (HPS)
  - Long-term conditions services (LTC)
  - Hospital services for people with mental health needs, and/or learning disabilities, and/or problems with substance misuse (MLS)
  - Prison healthcare services (PHS)
  - Rehabilitation services (RHS)
  - Residential substance misuse treatment/ rehabilitation services (RSM)
- Community or integrated healthcare (e.g. primary care in the UK)
  - Community healthcare services (CHC)
  - Doctors consultation services (DCS)
  - Doctors treatment services (DTS)
  - Dental services (DEN)
  - Diagnostic and/or screening services (DSS). Always use this code for all diagnostic and/or screening services regardless of the setting in which these services are being delivered.
  - Community-based services for people with a learning disability (LDC)
  - Mobile doctors services (MBS)
  - Community-based services for people with mental health needs (MHC)
  - Community-based services for people who misuse substances (SMC)
  - Urgent care services (UCS)
- Residential social care
  - Care home services with nursing (CHN)
  - Care home services without nursing (CHS)
  - Specialist college services (SPC)
- Community social care
  - Domiciliary care services including those provided for children (DCC)
  - Extra care housing services (EXC)
  - Shared Lives (formerly known as Adult Placement) (SHL)
  - Supported living services (SLS)
- Miscellaneous healthcare
  - Ambulance services (AMB)
  - Blood and transplant services (BTS)
  - Remote clinical advice services (RCA)

- Public health
    - Health protection. Action for clean air, water and food, infectious disease control, protection against environmental health hazards, chemical incidents and emergency response.
    - Health improvement. Wide ranging action to improve health and wellbeing and/or to reduce health inequalities.
    - Public health services. Action in service planning, commissioning and development, clinical effectiveness, clinical governance and efficiency working with Partners across the system in the NHS, Local Authorities and Voluntary Sector.
    - Public health intelligence. Surveillance, monitoring and assessment of health and the determinants of health, plus the development of the public health evidence base and knowledge.
- Other(s) – please specify.
- Unclear / not reported
- Not applicable

7. Population(s) [M]

In addition to health and social care categories that may reflect the specific health conditions experienced by study participants, we will also (if applicable: these child codes will only be applied to reports of primary research studies with human participants and SRs of studies with human participants) code the following participant characteristics for empirical studies included in the living EGM (code all that apply):

Age

- Infants
- Children
- Adults
- Older adults
- Unclear / not reported
- Not applicable

Infants = 0-24 months; Children = 2-17 years; Adults = 18 to 64 years; Older adults = 65+ years.

If the article does not report any information about the age of the participants = Select: 'Unclear / not reported'

If the article does not report in detail the age of the participants, but it describes the participants as "adults" = Select: 'Adults' AND 'Older adults'

If the article states that “children” were in the study population = Select: ‘Children’ BUT NOT ‘Infants’

If the age is reported as a mean, select the most appropriate check box. Example: “mean age 43 years” = Select: ‘Adults’

If the age is reported as a mean and its standard deviation, standard error, or other measure of spread or variability, select the most appropriate check box or boxes. Example: “mean age 18.0 ±10.0 years” = Select: ‘Children’ AND ‘Adults’

If the age is reported as "older than 18 years" and no upper age limit (nor any other clear indication that the upper age limit of participants is likely to be <65 years) is reported = Select: 'Adults' AND 'Older adults'

Sex

- Male
- Female
- Unclear / not reported
- Not applicable

Country/ Countries

Select all child codes that apply, or ‘Not applicable’. For systematic reviews and other research syntheses, select ‘Not applicable’ unless eligibility criteria for considering studies include specific geographical restrictions.

- United Kingdom
- Other Europe
- Canada
- United States of America
- Other North / Central America
- South America
- Africa
- Asia
- Oceania
- Unclear / not reported
- Not applicable

Further participant characteristics related to dimensions of health equity will be extracted for those studies (also included in the living EGM) selected for inclusion in the critical review (see ‘2.2.2. Critical review’ and ‘2.3. Equity issues’, below).

8. Map version

This code will be automatically assigned to each included record (report) based on the version number of the living EGM in which the record is first published (see also 'Publishing the living map', below in this section).

Coding procedures
In a pilot phase, at least two researchers will independently classify (code) each included article (record/ report), and we will compare and discuss their respective codes, before making final decisions. The objectives of this pilot phase will be to identify early revisions to the classification (coding) scheme, as well as to foster consistency between researchers in applying the scheme. After the initial pilot coding phase, we will switch to a procedure of one trained researcher classifying (coding) each included article (record/ report), with an option to refer it for a 'second opinion' if unsure. 'Second opinions' will be resolved by team discussion and consensus, involving at least two researchers (including the original coder).

From the inception of this living EGM, each included article (record/ report) will be manually classified by researcher(s), based on both title-abstract and the corresponding full-text article when available. However, once we have accumulated sufficient corpora of human labels (codes) for each dimension, we will investigate the potential to automate the process of assigning codes to new included records (articles/ reports), so far as possible, using GPT-4 based auto-coding tools hosted in EPPI Reviewer.

Publishing the living map

We will aim to publish the inaugural version ('Version 1') of this living EGM 6-9 months after the inception of our production workflows. Subsequently, 'new' classified (fully-coded) articles will be added to the living EGM, and an updated version will be published, once every 2-3 months (to be confirmed). The living EGM will be curated and published as an open access web database, using EPPI Reviewer and EPPI-Visualiser tools, and hosted on the EPPI Centre website. This web database will enable users of the living EGM to interact with, and search among, coded records of included articles, including links to corresponding full-text reports (where available).

2.2.2 Critical review

This section of our protocol specifies the methods, procedures, and tools that we will use to produce the critical review.

Eligibility criteria

Eligibility criteria for the critical review will provisionally match those of the living EGM (see section 2.2.1). As such, the critical review will assess both reports of empirical research studies (encompassing both primary research and systematic reviews) and non-empirical review-type articles. Likewise, full journal articles, including pre-prints will be considered for inclusion, while conference abstracts will be excluded; articles published before 1st January 2018 will be excluded; and articles reported in languages other than English will also be excluded.

Searching for and selecting included articles

We will exclusively select articles (reports) for inclusion the critical review from among reports of empirical research studies and non-empirical review-type articles already included in the living EGM (see section 2.2.1). Articles will be selected using researchers' judgement, based on two core principles: we will select articles that: (1) reflect key 'classes' of LLM-based tools identified in the living EGM, and (2) are judged to be of most instrumental value (once critically reviewed) to achieving the overall aim of this work (see also section 2.1 'Aims and Scope').

As such, the unit of analysis in the critical review will be 'class of LLM-based tool'. We have selected this unit of analysis based on two key premises. First, when mapping the evidence, we expect to encounter a range of different types of LLM-based tools, such as: those trained on different datasets; those which use a LLM to perform a specific task - e.g. transcription; and those that use a LLM to interrogate a database. Second, these different types of use scenarios for LLM-based tools will be more or less 'safe'. As such, we will first group the emergent 'classes' of LLM-based tools delineated by the evidence mapping process (see section 2.2.1); and then identify and select exemplar articles to illustrate how LLMs work and what they do. For each specific type of task encountered among health and social care applications, we will specify how LLMs would need to work in order for the engineering of the system to ensure that the outputs are reliable 'by design'.

For example, in an 'answering question' task (or a 'summarising' task), users of the LLM (e.g. via ChatGPT) for a health or social care application are likely to need to have confidence that its outputs both (i) draw on the sum of current, cumulative knowledge (and do not present an incomplete, or biased, representation of this) and (ii) discriminate between reliable and unreliable research evidence (and do not treat all evidence as equally reliable). In this context, our critical review would investigate the extent to which eligible articles contain valid evidence claims that LLMs meet these criteria (i and ii); and we would select, analyse and highlight archetypal example (or exemplar) articles to illustrate the extent to which these necessary and sufficient conditions are, or aren't, met.

A key challenge in analysing evidence claims about the performance of generative LLM-based tools is that the models - and therefore the outputs of models and their performance on a given task – will evolve and change over time (e.g. successive generations or versions of the same LLM). As such, evaluation results are likely to become quickly out of date, with corollary uncertainty about the stability of evaluation findings and related evidence claims. We will aim to address this challenge by aiming to select, for each class of LLM-based tool we identify, archetypal example articles that illustrate the capacity, or lack of capacity, of generative LLM-based tools to produce outputs for health and social care applications that are reliable and trustworthy 'by design'.

Collecting data from included studies

Included empirical research studies and commentary or review-type articles will be treated similarly in the data collection and analysis stages of the critical review.
For both types of article (report), we will pilot, revise and apply a data collection framework adapted from one that we originally developed for another (published) NIHR PRP Reviews Facility critical review focused on 'precision public health' (Kneale

2020). This phase of data collection will also build on the coding of each report already undertaken in the mapping phase (i.e. for the living EGM – see '2.2.1 Living EGM', above).

Provisionally, further data to be extracted (at this stage[9]) from full-text reports selected for inclusion in the critical review will include:

1. Research Question(s)

We will extract brief details of the research question(s) investigated in each included empirical study or commentary/review article using a standard formulation/ format:

Model(s) versus *Comparator(s)* or compared with *Gold standard(s) or* outputs assessed by *human expert(s)* for *Specific Task(s)* in *Context(s)/Setting(s)* [*Research activities +/- Health and/or social care categories +/- Service type(s)*] and/or [Population(s)]

Here, we are exclusively interested in research questions that concern the performance of generative LLM-based tools for one or more health and/or social care applications (encompassing beneficial, adverse and/or differential impacts or effects).

Selected articles will already have been coded according to the specific models, tools and tasks they investigated as part of the coding scheme applied to the larger set of articles included in the living EGM (see '2.2.1 Living evidence and gap map'). For the critical review, we will additionally code selected articles according to whether they report details of the dataset(s) (***corpora***) used to train the specific model(s) under investigation; and each selected article will be further classified (coded) according to whether (a) the model(s), and (b) their training data, are (i) open access, or (ii) closed access (if reported).

2. Outcome measure(s) (performance metrics)

Based on preliminary scoping of this research literature, we are likely to encounter a diverse range of performance metrics (outcome measures) among articles included in our living EGM. For example, there is a large class of metrics reported in evaluations of text generating LLMs that are essentially counts of words that match with the specified 'gold standard' comparator text (examples include 'bleu' and 'rouge'). For the critical review, we will aim to select articles that can add to the 'cartography' of the 'landscape' of performance metrics. Then we will extract and code the specific performance metrics being used, with the aim of producing plain language definitions of metrics in common use across the range of classes of health and social care applications.

For the EGM, we will have coded articles reporting empirical studies according to whether the evaluation study datasets were (i) open access or (ii) closed access. For the critical review, we will also further distinguish between 'prompts data' and 'performance data'. With regards to 'prompts data', we will extract details of 'prompt design' (which aims to specify the right prompts for the specific task at hand and the

[9] The data described in this section, to be extracted for the critical review, will be supplemented by corresponding study characteristics data that we will have already extracted (coded) for the living EGM (see section 2.2.1 for further details).

required type(s) of output(s)) and of 'prompt engineering' (which aims to improve model performance for the specific task at hand by adjusting or adding prompts).

3. Evidence claim(s) - Identification

Evidence claims made about the performance of LLMs may be positive, negative, or neutral. We will identify and extract details of the principal evidence claim(s)[10] being made by the investigators/ authors, with regards to each investigated research question, in each included empirical study (report(s)) or commentary/review article.

- What claim is being made? (Describe)
- What is the claim being made in author's words? (If applicable)
- Is the claim positive, negative, or neutral?
- Does the claim concern a beneficial, adverse, or differential impact/ effect?

Here, we are exclusively interested in evidence claims being made about the performance of generative LLM-based tools for one or more health and/or social care applications (encompassing beneficial, adverse and/or differential impacts/effects).

4. Evidence claim(s) - Components

We will extract relevant details of each specific component of each evidence claim, namely:

- Qualifier(s)[11]: Are there any qualifiers to the claim?
- Rebuttals(s)[12]: Have the authors pre-empted any rebuttals to the claim?
- Grounds[13] - What are the grounds for the claim?
- Warrant(s)[14] - What is the warrant(s) for the claim? Does/do the study report(s) explicitly state the warrant(s)?
- Backing for warrant(s)[15] - What is the backing for the warrant(s)? Does/do the study report(s) explicitly provide appropriate backing for the warrant(s)?

[10] Judgment required in cases in which there are multiple candidate claims and/or multiple formulations of the same claim (see example - Kung 2023).

[11] Qualifiers are indicators of the strength of the 'leap' from the data to the claim and may limit the universality of the claim.

[12] Rebuttals are made when an investigator/ author anticipates potential counterarguments to the claim and outlines why potential counterarguments may not be valid.

[13] The grounds is/are the basis for the claim; in this case, the empirical study at hand, or the studies cited as supporting examples (evidence). We will only extract details of grounds that are based in evidence.

[14] The link between the grounds and the claim is established through a warrant. The warrant should explain how the grounds support the claim (i.e. the reason(s) to accept the claim).

[15] Warrant(s) and their backing are expected to be implicit in most cases, rather than explicitly stated.

5. Bias, equity and trustworthiness

This critical review will include consideration of issues of bias, equity and trustworthiness in several ways - see section ‘2.3 Equity issues’ for full details. To inform these assessments, we will extract the following data and information from included studies (when available):

i. Evidence for any biases in model inputs and/or outputs pertaining to specific marginalised or at-risk subgroups;
ii. Brief details of any approaches used to address risk of such biases;
iii. Brief details of any approaches used to evaluate the (broader) trustworthiness of model outputs and/or their alignment with human intentions (Liu Y 2023);
iv. Evidence for any differential performance of generative LLM-based tools pertaining to specific marginalised or at-risk subgroups of people (within-study comparisons reported by authors only);
v. Study characteristics data on the percentages and numbers of each subgroup identified in i or iv within included study datasets and/or participant samples (if applicable)

6. Funding source(s)

We will extract brief details of the sources of any funding used to support the empirical study or commentary/ review.

7. Investigator(s) / author(s) affiliations

We will extract brief details of the investigators’ / authors’ institutional affiliations.

8. Conflict of interest statements

We will extract verbatim details of any financial or other conflicts of interest explicitly stated by the investigator(s) / author(s).

This provisional data collection plan will be piloted and refined during initial stages of the critical review, and also amended based on consultation with stakeholders, including via our proposed advisory group.

Analysing and synthesising data from included studies

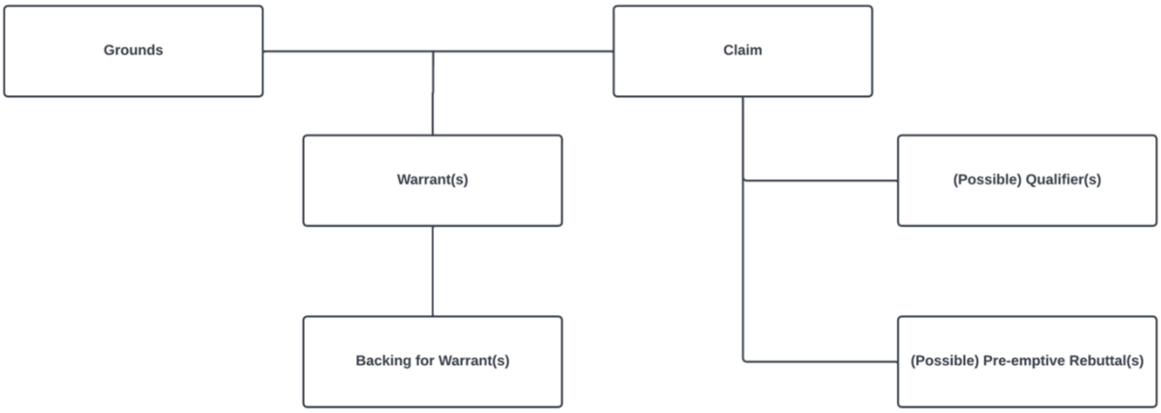
In our report of this critical review, analyses will be prefaced by a general description of how each class of generative LLM-based tools work, supplemented with an updated glossary of terms. We will also draw on our living EGM to briefly summarise the range of potential health or social care applications, highlighting any key gaps in cumulative bodies of evidence.

In line with our provisional plans for data collection, we will treat included empirical research studies and non-empirical commentary/ non-systematic review-type articles similarly in the analysis stage of the critical review. We will present extracted data in tables, using a format adapted from another critical review we have conducted on ‘precision public health’ (Kneale 2020); grouped by types of task(s) and/or research questions (if sufficiently similar - i.e. similarly formulated - research questions are addressed across multiple included studies and/or commentaries/ reviews).

We will then proceed to evaluate the validity of each evidence claim about the performance of generative LLM-based tools for health and/or social care applications, in empirical studies and commentaries/reviews - in relation to the specific research question(s) already identified - using an adapted version of Toulmin's model of argumentation tool (Kneale 2020, Toulmin 2003).

For each identified evidence claim (and corresponding research question), the overall objective of this analysis will be to establish the extent to which the claim is substantiated by the evidence presented; that is, the *validity* of the claim (i.e. whether the conclusion follows from the premises of the argument - see also Figure 1). As stated above (section 2.2.2 - 'Collecting data from included studies), we will restrict our focus exclusively to claims specifically concerning the performance of generative LLM-based tools for health and/or social care applications (and we will not analyse any other kinds of claims).

Figure 1. Line of argument and its components (adapted from Kneale 2020)



With the line of argument for each selected evidence claim broken down into its constituent parts (components), we will first consider any qualifiers and/or (pre-emptive) rebuttals to the claim (see Figure 1). This will help to give us a sense of the extent to which the investigator(s)/ author(s) endorse the claim, and under what conditions.

Next, we will consider the grounds for the claim. Here, we are most interested in grounds for which the investigator(s)/ author(s) provide some 'data' or 'fact'; in this case, either the empirical study at hand, or (for commentaries/reviews) the study (or studies) cited as supporting examples.

The link between the grounds (the study or studies) and the claim is established through a warrant (which may be explicit or, more often, implicit); and we will judge whether the (explicit or implicit) warrant explains how the grounds support the claim (that is, the extent to which the warrant is credible). We will also judge whether the backing for the warrant is (implicitly or explicitly) substantiated.

We will narratively summarise the results of our main analysis of evidence claims (as described above) supplemented by tabulated data. In addition, this critical review will include consideration of risk of bias in model outputs and other equity issues (see '2.3 Equity issues' for further details).

Finally, we will use the results from our analyses to draw provisional implications for policy and practice, focusing especially on inferences concerning the stage of readiness of generative LLM-based tools to be adopted to perform, or to support the performance of, different specific tasks (and/or types of tasks) for health and social care applications. These provisional implications for policy and practice will be shared with relevant groups of stakeholders and refined based on any feedback before final publication (see also ‘2.4 Stakeholder engagement’).

2.3 Equity issues

Health inequities are unfair, socially produced, systematic disparities in health outcomes between population groups, associated with their social, economic or personal characteristics (Whitehead 2006, Dahlgren 2006, Hollands 2024). Communities of people (population subgroups) experience health inequities when they face significant collective barriers to attaining good health, health-related quality of life, or participation in society.

Our critical review will consider issues of equity, bias and trustworthiness in several ways. First, in our general description of how generative LLMs and tools work, we will highlight risk of bias in the inputs and outputs of LLMs, which may be experienced by, or pertain to, specific marginalised, at-risk, or socially excluded groups (subgroups) of people, including (but not limited to) members of inclusion health groups (NHS England 2024). In the context of generative LLM-based tools, bias in model outputs has been defined as:

"*the presenc4 of systematic misrepresentations, attribution errors, or factual distortions that result in favouring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions based on learned patterns*" (Ferrara 2023).

Specific types of potential bias in the inputs and outputs of LLMs and tools include demographic cultural, linguistic, temporal, confirmation, and ideological, political or organisational biases (Ferrara 2023, Navigli 2023, Shramowski 2022). Potential sources of bias in the outputs of generative LLM-based tools encompass the datasets used to train the models (including labels or annotations from reinforcement learning), model algorithms (including product design and policy decisions), and interactions with the user (Pagano 2023, Ferrara 2023). For example, if an LLM is trained on a dataset/ input data (corpus) which includes biased language, perspectives, or stereotypes about people from specific ethnicity groups, then its outputs are likely to reflect those biases. Second, we will extract and summarise any evidence for such biases, as well as brief details of approaches to addressing them (e.g. Nozza 2023), that may be highlighted, or identified, in either included review-type articles, or included studies, evaluating the performance of LLM-based tools for health and social care applications (critical review). Bias is one aspect of the trustworthiness of generative LLM-based tools and their outputs, and we will also extract (critical review) details of any approaches used to evaluate the trustworthiness of model outputs or their alignment with human intentions (Liu Y 2023).

Third, due to our uncertainty about the presence and impacts of bias in the outputs of generative LLM-based tools for specific health and social care applications, we are also unsure whether their performance on specific tasks is likely to differ among specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people in important or meaningful ways - see also 'Appendix 3. Equity issues checklist'. We will therefore extract (from relevant within-study comparisons among studies included in the critical review) and highlight any evidence for differential performance of generative LLM-based tools pertaining to specific subgroups.

Fourth, due to our corollary uncertainty about the range of health and social care applications we will encounter in this work, and which will be selected for the critical review, we are also unsure whether the health conditions, or public health issues, addressed by the tasks being performed by generative LLM-based tools under investigation among included studies are more (or less) likely to be experienced by one or more specific marginalised, socially excluded and/or inclusion health group(s) of people (see also Appendix 3). Likewise, we are unsure whether aspects of the ways the LLMs are trained and deployed, or the ways included studies have been designed, may make it less (or more) likely that specific marginalised, at-risk, socially excluded, or inclusion health group(s) of people are represented in study datasets or (if applicable) among study participant samples (see also Appendix 3). We will therefore extract any study characteristics data on the percentages and numbers of each subgroup (Appendix 3) represented within included study datasets and/or participant samples when applicable (critical review), and then use these data to: (a) describe participants included in the studies and the final critical review; (b) highlight any clear evidence for under- (or over-) representation of specific subgroups of people; (c) assess whether sampling is good enough for study and critical review findings to be applicable to the United Kingdom population; and (d) summarise any implications for the extent and limits of applicability of our critical review findings.

2.4 Stakeholder engagement

This living EGM (evidence surveillance) and critical review will be conducted by an independent team of academic researchers based at the EPPI Centre (University College London) and the Centre for Reviews and Dissemination (CRD, University of York), under the auspices of the NIHR PRP Reviews Facility. This facility undertakes evidence synthesis projects commissioned by the NIHR Policy Research Programme (NIHR PRP) to support decision-making by the UK Secretary of State for Health and Social Care, Ministers, and Senior Officials in the UK Department of Health and Social Care (DHSC) and its Arm's Length Bodies (ALBs).

This project was commissioned following consultation with policy stakeholders on a concept note, in which we briefly outlined its need and purpose, along with its proposed aims, scope and methods. The concept note was discussed during an NIHR PRP Research and Development (R&D) Committee meeting (May 2023), which included representatives of the DHSC, the Care Quality Commission (CQC), the National Institute for Health and Care Excellence (NICE), NHS England (NHSE), and the UK Health Security Agency (UKHSA). We have used the initial feedback received from these stakeholders to inform the development of the draft version of this protocol for the living EGM and critical review.

When we have prepared an inaugural version of the living EGM and the critical review, we will elicit further feedback from the same policy stakeholders, via the NIHR PRP and its R&D Committee; and we will use this to: (i) inform revisions to the critical review prior to its publication (as a report on the EPPI Centre website) - in particular, provisional implications for policy and practice will be shared and refined based on any feedback; and (ii) inform any modifications to this protocol, for use to guide subsequent updates of this living EGM.

We will establish an advisory group for this project, comprised of invited policy stakeholders, academic experts, and representatives of industry, patients and the public. The project advisory group will meet virtually: first, when we are in the process of developing an inaugural version of the living evidence map; and second, when we are finalising the design of the critical review (including our provisional typology of classes of health and social care applications of LLM-based tools). We will also send a draft version of the final report of the critical review to advisory group members, with an invitation to provide formative feedback, before its publication on the EPPI Centre website.

2.5 Research ethics

2.5.1 Ethical issues related to the conduct of this project

The EPPI Centre is a research centre based in the Social Research Institute (SRI) of the University College London (UCL) Institute of Education (IOE). All research projects undertaken by staff, students or visitors of the UCL IOE which collect or use data from human participants, including secondary data analysis, systematic reviews and pilot studies, are required to gain ethical approval before data collection begins. For the current project, we applied to the IOE Research Ethics Committee for a standard ethics review and we did not start pilot coding or extracting data from included study reports until the committee had approved our application. Research ethics approval was received on 6th December 2023 and the letter of approval is uploaded to the project webpage.

This project is classed as a systematic review (systematic reviews and related forms of evidence synthesis) and it will not have any human participants, except for key policy stakeholders with whom we are engaging to inform the design, conduct and reporting of this work (see Section 2.4 'Stakeholder engagement). We have therefore not identified any ethical issues related to the conduct of this project with regards to: methods; sampling; recruitment; gatekeepers; informed consent; potentially vulnerable participants; safeguarding and child protection; sensitive topics; international research; risks to participants or researchers; confidentiality and anonymity; disclosures and limits to confidentiality; data storage and security; reporting; dissemination and use of findings; nor impact or public engagement.

2.5.2 Ethical issues related to the use of generative LLM-based tools in the field

Ethical issues related to the use of generative LLM-based tools in the field include:

- Ethical concerns arising from the unintended consequences of biased model outputs (see, for example: https://arxiv.org/abs/2304.03738);

- Ethical concerns relating to the terms and conditions of employment of human annotators who provide training or reinforcement learning data to support the development of AI systems including LLMs (see, for example: https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots); and

- Ethical concerns related to the environmental impacts of power consumption when training and using LLMs (see, for example: https://www.statista.com/statistics/1384401/energy-use-when-training-llm-models).

We will describe these ethical concerns in our general description of how generative LLMs and tools work, as well as highlighting any evidence for the nature and extent of these phenomena, or any other ethical issues, encountered during this work.

3 REFERENCES

Ali 2023
Ali H, Qadir J, Shah Z. (2023). ChatGPT and Large Language Models (LLMs) in Healthcare: Opportunities and Risks [Pre-print]. *TechRxiv*: 22579852.v2 https://doi.org/10.36227/techrxiv.22579852.v2

Au Yeung 2023
Au Yeung J, Kraljevic Z, Luintel A, Balston A, Idowu E, Dobson RJ, Teo JT. (2023). AI chatbots not yet ready for clinical use. *Frontiers in Digital Health*; 5: 1161098. https://doi.org/10.3389/fdgth.2023.1161098

Azamfirei 2023
Azamfirei R, Kudchadkar SR, Fackler J. (2023). Large language models and the perils of their hallucinations. *Critical Care*; 27: 120. https://doi.org/10.1186/s13054-023-04393-x

Briganti 2023
Briganti G. (2023). A clinician's guide to large language models. *Future Medicine AI* 1; 1. https://doi.org/10.2217/fmai-2023-0003

Callaghan 2023
Callahan A, Gombar S, Cahan EM, Jung K, Steinberg E, Polony V, Morse K, Tibshirani R, Hastie T, Harrington R, Shah NH. (2023). Using Aggregate Patient Data at the Bedside via an On-Demand Consultation Service. *NEJM Catalyst*; 2(10). https://doi.org/10.1056/CAT.21.0224

Care Quality Commission 2024
Care Quality Commission (2024). *Care Quality Commission (CQC) - Guidance for Providers > Service Types* [Webpage]. Available from: https://www.cqc.org.uk/guidance-providers/regulations-enforcement/service-types [Accessed: May 2024]

Chen 2023
Chen Z, Micsinai Balan M, Brown K. (2023). Language Models are Few-shot Learners for Prognostic Prediction [Pre-print]. *arXiv*: 2302.12692 https://doi.org/10.48550/arXiv.2302.12692

Dahlgren 2006
Dahlgren G, Whitehead M. (2006). *Levelling up (part 2): a discussion paper on European strategies for tackling social inequities in health.* World Health Organization: Studies on social and economic determinants of population health. Copenhagen: WHO Regional Office for Europe. Available from: https://iris.who.int/handle/10665/107791 [Accessed: November 2023]

Devlin 2018
Devlin J, Chang M-W, Lee K, Toutanova K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Pre-print]. *arXiv*: 1810.04805 https://doi.org/10.48550/arXiv.1810.04805

Elliott 2017
Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, Salanti G, Meerpohl J, MacLehose H, Hilton J, Tovey D, Shemilt I, Thomas J. (2017). Living systematic

review: 1. Introduction-the why, what, when, and how. *Journal of Clinical Epidemiology*; 91: 29-30. https://doi.org/10.1016/j.jclinepi.2017.08.010

Ferrara 2023
Ferrara E. (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models [Pre-print]. *arXiv*: 2304.03738
https://doi.org/10.48550/arXiv.2304.03738

Fink 2022
Fink MA, Kades K, Bischoff A, Moll M, Schnell M, Küchler M, Köhler G, Sellner J, Heussel CP, Kauczor H-U, Schlemmer H-P, Maier-Hein K, Weber TF, Kleesiek J. (2022). Deep Learning–based Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports. *Radiology: Artificial Intelligence* 2022; 4: 5. https://doi.org/10.1148/ryai.220055

Grant 2009
Grant MJ, Booth A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information Libraries Journal* 26(2): 91-108.
https://doi.org/10.1111/j.1471-1842.2009.00848.x

Guyatt 2011
Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, et al. (2011). GRADE Guidelines: 8. Rating the Quality of Evidence—Indirectness. *Journal of Clinical Epidemiology* 64 (12): 1303–1310.
https://doi.org/10.1016/j.jclinepi.2011.04.014

Guyatt 2023
Guyatt G, Zhao Y, Mayer M, Briel M, Mustafa R, Izcovich A, Hultcrantz M, et al. (2023). GRADE Guidance 36: Updates to GRADE's Approach to Addressing Inconsistency. *Journal of Clinical Epidemiology* 158: 70–83.
https://doi.org/10.1016/j.jclinepi.2023.03.003

Hammersley 2002
Hammersley M. (2002). *Systematic or unsystematic, is that the question? Some reflections on the science, art and politics of reviewing research evidence*. London: Health Development Agency Public Health Steering Group.

Harrer 2023
Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*; 90: 104512
https://doi.org/10.1016/j.ebiom.2023.104512

Health Services Research Unit 2023
Health Services Research Unit. (2023). *Tools to help reviewers make equity, diversity and inclusion assessments (PRO-EDI)* (Web page). Aberdeen: University of Aberdeen. Available from: https://www.abdn.ac.uk/hsru/what-we-do/research/projects/tools-to-help-reviewers-make-equity-diversity-and-inclusion-assessments-339 (Accessed: November 2023).

Hollands 2024
Hollands GJ, South E, Shemilt I, Oliver S, Thomas J, Sowden AJ (2024). Methods used to conceptualize dimensions of health equity impacts of public health interventions in

systematic reviews. *Journal of Clinical Epidemiology*; 111312 https://doi.org/10.1016/j.jclinepi.2024.111312

House of Lords Communications and Digital Committee 2023
House of Lords Communications and Digital Committee (2023). *How will AI large language models shape the future and what is the right regulatory approach?* (Webpage). Available from: https://www.parliament.uk/business/lords/media-centre/house-of-lords-media-notices/2023/july-2023/how-will-ai-large-language-models-shape-the-future-and-what-is-the-right-regulatory-approach/

Hügle 2023
Hügle T (2023). The wide range of opportunities for large language models such as ChatGPT in rheumatology. *Rheumatic & Musculoskeletal Diseases Open*; 9: e003105. https://doi.org/10.1136/rmdopen-2023-003105

Jeblick 2022
Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, Weber T, Wesp P, Sabel B, Ricke J, Ingrisch M. (2022). ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports Pre-print]. *arXiv*: 2212.14882 https://doi.org/10.48550/arXiv.2212.14882

Jiang 2023
Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, Eaton K, Riina HA, Laufer I, Punjabi P, Miceli M, Kim NC, Orillac C, Schnurman Z, Livia C, Weiss H, Kurland D, Neifert S, Dastagirzada Y, Kondziolka D, Cheung ATM, Yang G, Cao M, Flores M, Costa AB, Aphinyanaphongs Y, Cho K, Oermann EK. (2023). Health system-scale language models are all-purpose prediction engines. *Nature*; 619: 357–362. https://doi.org/10.1038/s41586-023-06160-y

Khanjani 2023
Khanjani Z, Watson G, Janeja VP. (2023). Audio deepfakes: A survey. *Frontiers in Big Data* 9; 5: 1001063. https://doi.org/10.3389/fdata.2022.1001063

Karabacak 2023
Karabacak M, Margetis K. (2023). Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus*; 15(5): e39305. https://doi.org/10.7759/cureus.3930

Knafou 2023
Knafou J, Haas Q, Borissov N, Counotte M, Low N, Imeri H, Ipekci AM, Buitrago-Garcia D, Heron L, Amini P, Teodoro D. (2023). Ensemble of deep learning language models to support the creation of living systematic reviews for the COVID-19 literature. *Systematic Reviews*; 12: 94. https://doi.org/10.1186/s13643-023-02247-9

Kneale 2020
Kneale D, Lorenc T, O'Mara-Eves A, Hong QN, Sutcliffe K, Sowden A, Thomas J. (2020). *Precision public health – A critical review of the opportunities and obstacles*. London: EPPI Centre, Social Science Research Unit, UCL Institute of Education, University College London. https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3788

Kung 2023
Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J,Tseng V. (2023). Performance of ChatGPT on

USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*; 2(2): e0000198. https://doi.org/10.1371/journal.pdig.0000198

Lee 2023
Lee P, Bubeck S, Petro J. (2023). Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*; 388: 1233-1239. https://doi.org/10.1056/NEJMsr2214184

Li J 2023
Li J, Dada A, Kleesiek J, Egger J. (2023). ChatGPT in Healthcare: A Taxonomy and Systematic Review [Pre-print]. *medRxiv*: 2023.03.30.23287899 https://doi.org/10.1101/2023.03.30.23287899

Liu H 2023
Liu H, Peng Y, Weng C. (2023). How Good Is ChatGPT for Medication Evidence Synthesis? *Studies in Health Technology and Informatics*; 302: 1062-1066. https://doi.org/10.3233/SHTI230347

Liu S 2023
Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, McCoy AB, Sittig DF, Wright A. (2023). Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association*; 30(7): 1237-1245. https://doi.org/10.1093/jamia/ocad072

Liu Y 2023
Liu Y, Yao Y, Ton J-F, Zhang X, Guo R, Cheng H, Klochkov Y, Taufiq MF, Li H. (2023). Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment [Pre-print]. *arXiv*: 2308.05374 [cs.AI] https://doi.org/10.48550/arXiv.2308.05374

Lorenc 2020
Lorenc T, Khouja C, Raine G, Shemilt I, Sutcliffe K, D'Souza P, Burchett H, Hinds K, Khatwa M, Macdowall W, Melton H, Richardson M, South E, Stansfield C, Thomas S, Kwan I, Wright K, Sowden A, Thomas J. (2020). *COVID-19: living map of the evidence.* London: EPPI-Centre, Social Science Research Unit, UCL Social Research Institute, University College London. Available from: https://eppi.ioe.ac.uk/cms/Projects/DepartmentofHealthandSocialCare/Publishedreviews/COVID-19Livingsystematicmapoftheevidence/tabid/3765/Default.aspx

Luo 2022
Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu T-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining, *Briefings in Bioinformatics*; 23(6): bbac409. https://doi.org/10.1093/bib/bbac409

Miwa 2014
Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*; 51: 242-253. https://doi.org/10.1016/j.jbi.2014.06.005

Navigli 2023
Navigli R, Conia S, Ross B. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality*; 15(2): 10. https://doi.org/10.1145/3597307

NHS England 2024
NHS England 2024. *Inclusion health groups* (Webpage). Available from: https://www.england.nhs.uk/about/equality/equality-hub/national-healthcare-inequalities-improvement-programme/what-are-healthcare-inequalities/inclusion-health-groups [Accessed: November 2023]

NHS England Transformation Directorate 2023
NHS England Transformation Directorate (2023). *Artificial Intelligence* (Webpage). Available from: https://transform.england.nhs.uk/information-governance/guidance/artificial-intelligence

Nozza 2023
Nozza D, Bianchi F, Hovy D. (2022). Pipelines for Social Bias Testing of Large Language Models. *Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large Language Models*: 68–74. Association for Computational Linguistics.

O'Mara-Eves 2015
O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*; 4:5. https://doi.org/10.1186/2046-4053-4-5

Pagano 2023
Pagano TP, Loureiro RB, Lisboa FVN, Peixoto RM, Guimarães GAS, Cruz GOR, Araujo MM, Santos LL, Cruz MAS, Oliveira ELS, Winkler I, Nascimento EGS. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*; 7(1): 15. https://doi.org/10.3390/bdcc7010015

Page 2021
Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, et al. (2021). PRISMA 2020 Explanation and Elaboration: Updated Guidance and Exemplars for Reporting Systematic Reviews. *British Medical Journal* 372: n160. https://doi.org/10.1136/bmj.n160

Patel 2023
Patel SB, Lam K. (2023). ChatGPT: the future of discharge summaries? *The Lancet Digital Health*; 5(3): e107-e108. https://doi.org/10.1016/S2589-7500(23)00021-3

Qureshi 2023
Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Reviews*; 12(1): 72. https://doi.org/10.1186/s13643-023-02243-z

Sallam 2023
Sallam M. (2023). The Utility of ChatGPT as an Example of Large Language Models in Healthcare Education, Research and Practice: Systematic Review on the Future Perspectives and Potential Limitations [Pre-print]. *medRxiv* 2023.02.19.23286155 https://doi.org/10.1101/2023.02.19.23286155

Shahzad 2022
Shahzad HF, Rustam F, Flores ES, Luís Vidal Mazón J, de la Torre Diez I, Ashraf I.

(2022). A Review of Image Processing Techniques for Deepfakes. *Sensors* 16; 22(12): 4556. https://doi.org/10.3390/s22124556

Shamseer 2015
Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. (2015). Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015: Elaboration and Explanation. *British Medical Journal* 349: g7647. https://doi.org/10.1136/bmj.g7647

Shramowski 2022
Shramowski P, Turan C, Andersen N, Rothkopf CA, Kersting K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*; 4: 258-268. https://doi.org/10.1038/s42256-022-00458-8

Singhal 2022
Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Scharli N, Chowdhery A, Mansfield P, Aguera y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. (2022). Large Language Models Encode Clinical Knowledge [Pre-print]. *arXiv*: 2212.13138. https://doi.org/10.48550/arXiv.2212.13138

Tang 2023
Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, Xu Z, Ding Y, Durrett G, Rousseau J, Weng C, Peng Y. (2023). Evaluating Large Language Models on Medical Evidence Summarization [Pre-print]. *medRxiv* 2023.04.22.23288967. https://doi.org/10.1101/2023.04.22.23288967

Tenti 2021
Tenti P, Thomas J, Peñaloza R, Pasi G. (2021). Using an ensemble of features for personalized recommendations of scientific publications. CEUR Workshop Proceedings. Available from: https://ceur-ws.org/Vol-2947/paper30.pdf

Thirunavukarasu 2023
Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. (2023). Large language models in medicine. *Nature Medicine*; 29(8): 1930-1940.

Thomas 2014
Thomas J, O'Mara-Eves A, Brunton G. (2014). Using Qualitative Comparative Analysis (QCA) in Systematic Reviews of Complex Interventions: A Worked Example. *Systematic Reviews* 3(1): 67. https://doi.org/10.1186/2046-4053-3-67

Thomas 2017a
Thomas J, O'Mara-Eves A, Kneale D, Shemilt I. (2017). Chapter 9: Synthesis Methods for Combining and Configuring Quantitative Data. In: Gough D, Oliver S, Thomas J (editors). *An Introduction to Systematic Reviews* (2nd Edition), pp. 211-250. London: Sage.

Thomas 2017b
Thomas J, O'Mara-Eves A, Harden A, Newman M. (2017). Chapter 8: Synthesis Methods for Combining and Configuring Textual or Mixed Methods Data. In: Gough D, Oliver S, Thomas J (editors). *An Introduction to Systematic Reviews* (2nd Edition), pp. 181-210. London: Sage.

Thomas 2024
Thomas J, Graziosi S, Brunton J, Ghouze Z, O'Driscoll P, Bond M (2024). *EPPI Reviewer: advanced software for systematic reviews, maps and other evidence synthesis* [Software]. Available from: https://eppi.ioe.ac.uk/cms/er

Toulmin 2003
Toulmin SE. (2003) *The uses of argument*. Cambridge: Cambridge University Press.

Touya 2023
Touya G, Potié Q, Mackaness WA. (2023). Incorporating ideas of structure and meaning in interactive multi scale mapping environments. *International Journal of Cartography*; 9(2): 342-372. https://doi.org/10.1080/23729333.2023.2215960

UK Clinical Research Collaboration 2024
UK Clinical Research Collaboration (2024). *UKCRC Health Research Classification System - Health Categories* [Webpage]. Available from: https://hrcsonline.net/health-categories/ [Accessed: May 2024]

Wang 2023
Wang S, Zhao Z, Ouyang X, Wang Q, Shen D. (2023). ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models [Pre-print]. *arXiv*: 2302.07257 https://doi.org/10.48550/arXiv.2302.07257

Welch 2016
Welch V, Petticrew M, Petkovic J, Moher D, Waters E, White H, Tugwell P, et al. (2016). Extending the PRISMA Statement to Equity-Focused Systematic Reviews (PRISMA-E 2012): Explanation and Elaboration. *Journal of Clinical Epidemiology* 70: 68–89. https://doi.org/10.1016/j.jclinepi.2015.09.001

Whitehead 2006
Whitehead M, Dahlgren G. (2006). *Levelling up (part 1): a discussion paper on concepts and principles for tackling social inequities in health, in World Health Organization: Studies on social and economic determinants of population health*. Copenhagen: WHO Regional Office for Europe. Available from: https://iris.who.int/handle/10665/107790 [Accessed: November 2023]

Wornow 2023
Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. (2023). The Shaky Foundations of Clinical Foundation Models: A Survey of Large Language Models and Foundation Models for EMRs [Pre-print]. *arXiv*: 2303.12961 https://doi.org/10.48550/arXiv.2303.12961

Yang 2022
Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG, Zhang Y, Magoc T, Harle CA, Lipori G, Mitchell DA, Hogan WR, Shenkman EA, Bian J, Wu Y. (2022). A large language model for electronic health records. *npj Digital Medicine*; 5: 194. https://doi.org/10.1038/s41746-022-00742-2

Zhu 2023
Zhu L, Mou W, Chen R. (2023). Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *Journal of Translational Medicine;* 21: 269. https://doi.org/10.1186/s12967-023-04123-5

3.1 Appendix 1. MEDLINE (Ovid) and Embase (Ovid) search strategies

MEDLINE I

1. large language model$.ti.
2. LLM$.ti.
3. generative artificial intelligence.ti.
4. generative ai.ti.
5. generative pre-trained transformer$.ti.
6. gpt$.ti.
7. chatgpt.ti.
8. or/1-7
9. limit 8 to yr="2018 -Current"
10. limit 9 to english language

Embase I

1. large language model$.ti.
2. LLM$.ti.
3. generative artificial intelligence.ti.
4. generative ai.ti.
5. generative pre-trained transformer$.ti.
6. gpt$.ti.
7. chatgpt.ti.
8. or/1-7
9. limit 8 to yr="2018 -Current"
10. limit 9 to english language
11. limit 10 to conference abstract status

12. 10 not 11

MEDLINE II

1. large language model$.ti.
2. LLM$.ti.
3. generative artificial intelligence.ti.
4. generative ai.ti.
5. generative pre-trained transformer$.ti.
6. gpt$.ti.
7. chatgpt.ti.
8. bard.ti.
9. llama.ti.
10. claude.ti.
11. med-palm.ti.
12. mistral.ti.
13. mixtral.ti.
14. or/1-13
15. limit 14 to yr="2018 -Current"
16. limit 15 to english language

Embase II

1. large language model$.ti.
2. LLM$.ti.
3. generative artificial intelligence.ti.
4. generative ai.ti.
5. generative pre-trained transformer$.ti.
6. gpt$.ti.

7. chatgpt.ti.
8. bard.ti.
9. llama.ti.
10. claude.ti.
11. med-palm.ti.
12. mistral.ti.
13. mixtral.ti.
14. or/1-13
15. limit 14 to yr="2018 -Current"
16. limit 15 to english language
17. limit 16 to conference abstract status
18. 16 not 17

3.2 Appendix 2. List of specific generative large language models

The following alphabetical list of >40 open access or proprietary generative LLMs - including health domain-specific LLMs (denoted by an asterisk [*]) - comprises 'text-to-text' and 'text-to-image' models only. See also:

- https://lifearchitect.ai/timeline
- https://huggingface.co/spaces/mteb/leaderboard; and
- https://github.com/liusongxiang/Large-Audio-Models

Further specific models will be added to the coding scheme for the 'model(s)' dimension when launched and as we encounter them in the current work:

- AlexaTM - Alexa Teacher Model(s) (Amazon)
- Alpaca 7B (Stanford University)
- BioGPT * (John Snow Labs)
- BioMedLM * (Stanford Center for Research on Foundation Models)
- BioMegatron * (Nvidia)
- BLOOM - BigScience Large Open-science Open-access Multilingual Language Model (Hugging Face and others)
- Cerebras-GPT (Cerebras)
- Chinchilla (DeepMind)
- Claude (Anthropic)
- DALL·E 2 (Open AI)
- Ernie 3.0 Titan (Baidu)
- Falcon (Abu Dhabi Technology Innovation Institute)
- Galactica (Meta)
- GatorTron-OG * (Nvidia)
- GatorTron-S * (Nvidia)
- GLM-130B (Tsinghua University)
- GLaM - Generalist Language Model (Google)
- Gopher (DeepMind)
- GPT-2 (Open AI)
- GPT-3 (Open AI)
- GPT-4 (Open AI)
- GPT-J (EleutherAI)
- GPT-Neo (EleutherAI)
- GPT-NeoX (EleutherAI)

- Imagen (Google)
- LaMDA - Language Models for Dialog Applications (Google)
- LLaMA - Large Language Model Meta AI (Meta)
- Med-PaLM * (Google)
- Megatron-Turing NLG (Microsoft and Nvidia)
- Minerva (Google)
- Mistral / Mixtral (Mistral AI)
- OpenAssistant (LAION)
- OPT - Open Pre-trained Transformers (Meta)
- PaLM - Pathways Language Model (Google)
- PaLM 2 - Pathways Language Model 2 (Google)
- PanGu-α (Huawei)
- PanGu-Σ (Huawei)
- Parti - Pathways Autoregressive Text-to-Image Model (Google)
- PubMed GPT * (Stanford Center for Research on Foundation Models & MosaicML)
- Stable Diffusion (StabilityAI)
- Switch T5 (Google)
- Switch Transformer (Google) [Incl. Switch-Base, Switch-Large, Switch-XXL and Switch-C]
- Wu Dao 2.0 (Beijing Academy of Artificial Intelligence)
- YaLM 100B (Yandex)

3.3 Appendix 3. Equity issues checklist

N.B. This is the first, provisional, pilot version of a checklist that we are currently in the early stages of developing as part of a new suite of tools (also including flow charts) for use to help guide consideration issues of health equity in systematic reviews and related forms of evidence synthesis, which draws on several existing sources of guidance and tools (Guyatt 2011, Guyatt 2023, Health Services Research Unit 2023, Page 2021, Shamseer 2015, Thomas 2014, Thomas 2017a, Thomas 2017b, Welch 2016).

The pilot checklist comprises a set of draft signalling questions (1a, 1b, 2, 3a, 3b, 3c) that we have applied to the current work, as summarised in section '2.3. Equity Issues'. There are three tables below in this appendix (A, B and C), each of which contains the same set of signalling questions, but which differ in terms of the specific equity-related dimensions/ participant characteristics covered (see bullet points and table columns). This checklist is likely to undergo further development, with corollary changes in its structure, format, and content. Please contact the authors for further details.

Table A

- Age
- Sex
- Gender
- Race, ethnicity and ancestry
- Religion
- Socio-economic status
- Level of education
- Location

Signalling question	***Response options***	***Equity-related dimension(s) / participant characteristic(s)***							
		Age	*Sex*	*Gender*	*Race, ethnicity and ancestry*	*Religion*	*Socio-economic status*	*Level of education*	*Location*
	Yes								

1a. Is the intervention under investigation targeted at specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people? (e.g. school feeding for children from low-income families)									
	No	x	x	x	x	x	x	x	x
	N/A								
1b. Is the intervention under investigation aimed at reducing social gradients across populations or among subgroups of the population? (e.g. interventions to reduce the social gradient in smoking, obesity prevention in children, interventions delivered by lay health workers)	*Yes*								
	No	x	x	x	x	x	x	x	x
	N/A								
2. Are the impacts of, or responses to, the intervention(s), or the experiences of the phenomenon, under investigation, expected to differ among specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people *in important or meaningful ways*?	*Confident that effects differ*								
	Confident that effects do not differ								
	Unsure whether effects differ	x	x	x	x	x	x	x	x

	N/A								
3a. Is the health condition, public health issue, or phenomenon, being addressed by the review (and/or map) more likely to be experienced by one or more specific marginalised, socially excluded and/or inclusion health group(s) of people?	*Yes*								
	No								
	Unsure								
3b. Are aspects of the intervention(s) and/or comparator(s), including how they are provided, expected to make it harder for some specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people to take part in eligible studies?	*Yes*								
	No								
	Unsure								
	N/A								
3c. Are elements of study design, such as eligibility criteria or recruitment and consent processes, expected to make it harder for some specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people to take part in eligible studies?	*Yes*								
	No								
	Unsure	x	x	x	x	x	x	x	x

*****

Table B

- Sexual orientation
- Disability
- People experiencing homelessness
- Drug or alcohol dependence
- Vulnerable migrants

Signalling question	***Response options***	***Equity-related dimension(s) / participant characteristic(s)***				
		Sexual orientation	*Disability*	*People experiencing homelessness*	*Drug or alcohol dependence*	*Vulnerable migrants*
1a. Is the intervention under investigation targeted at specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people? (e.g. school feeding for children from low-income families)	*Yes*					
	No	x	x	x	x	x
	N/A					
1b. Is the intervention under investigation aimed at reducing social gradients across populations or among subgroups of the population? (e.g. interventions to reduce the social gradient in smoking, obesity prevention in children, interventions delivered by lay health workers)	*Yes*					
	No	x	x	x	x	x

	N/A					
2. Are the impacts of, or responses to, the intervention(s), or the experiences of the phenomenon, under investigation, expected to differ among specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people *in important or meaningful ways*?	*Confident that effects differ*					
	Confident that effects do not differ					
	Unsure whether effects differ	x	x	x	x	x
	N/A					
3a. Is the health condition, or public health issue, being addressed by the review (and/or map) more likely to be experienced by one or more specific marginalised, socially excluded and/or inclusion health group(s) of people?	*Yes*					
	No					
	Unsure	x	x	x	x	x
	N/A					
3b. Are aspects of the intervention(s) and/or comparator(s), including how they	*Yes*					

are provided, expected to make it harder for some specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people to take part in eligible studies?	*No*					
	Unsure	x	x	x	x	x
	N/A					
3c. Are elements of the design of eligible studies, such as their eligibility criteria, or their recruitment and/or consent processes, expected to make it harder for some specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people to take part in eligible studies?	*Yes*					
	No					
	Unsure	x	x	x	x	x

*****

Table C

- Gypsy, Roma and Traveller communities
- Sex workers
- People in contact with the justice system
- Victims of modern slavery
- Other marginalised, at-risk, socially excluded and/or inclusion health group(s)

Signalling question	***Response options***	***Equity-related dimension(s) / participant characteristic(s)***				
		Gypsy, Roma and Traveller communities	*Sex workers*	*People in contact with the justice system*	*Victims of modern slavery*	*Other marginalised, at-risk, socially excluded and/or inclusion health group(s)*
1a. Is the intervention under investigation targeted at specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people? (e.g. school feeding for children from low-income families)	*Yes*					
	No	x	x	x	x	x
	N/A					
1b. Is the intervention under investigation aimed at reducing social gradients across populations or among subgroups of the population? (e.g. interventions to reduce the social gradient in smoking, obesity prevention in	*Yes*					
	No	x	x	x	x	x

children, interventions delivered by lay health workers)						
	N/A					
2. Are the impacts of, or responses to, the intervention(s), or the experiences of the phenomenon, under investigation, expected to differ among specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people *in important or meaningful ways*?	*Confident that effects differ*					
	Confident that effects do not differ					
	Unsure whether effects differ	x	x	x	x	x
	N/A					
3a. Is the health condition, or public health issue, being addressed by the review (and/or map) more likely to be experienced by one or more specific marginalised, socially excluded and/or inclusion health group(s) of people?	*Yes*					
	No					
	Unsure	x	x	x	x	x
	N/A					

3b. Are aspects of the intervention(s) and/or comparator(s), including how they are provided, expected to make it harder for some specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people to take part in eligible studies?	*Yes*					
	No					
	Unsure	x	x	x	x	x
	N/A					
3c. Are elements of the design of eligible studies, such as their eligibility criteria, or their recruitment and/or consent processes, expected to make it harder for some specific marginalised, at-risk, socially excluded and/or inclusion health group(s) of people to take part in eligible studies?	*Yes*					
	No					
	Unsure	x	x	x	x	x

*****

The NIHR Policy Research Programme Reviews Facility puts the evidence into development and implementation of health policy through:

- Undertaking policy-relevant systematic reviews of health and social care research
- Developing capacity for undertaking and using reviews
- Producing new and improved methods for undertaking reviews
- Promoting global awareness and use of systematic reviews in decision-making

The Reviews Facility is a collaboration between the following centres:
EPPI Centre (Evidence for Policy and Practice Information Centre),
UCL Social Research Institute, University College London;
CRD (Centre for Reviews and Dissemination), University of York;
and the London School of Hygiene and Tropical Medicine.

The NIHR Policy Research Programme Reviews Facility collaboration has grown out of a previous 'reviews facility' in Health Promotion and Public Health based at the EPPI Centre, and has been funded by the Department of Health and Social Care since 1995.

The views expressed in this work are those of the authors and do not necessarily reflect the views of the collaborating centres or the funder. All errors and omissions remain those of the authors.

First produced in 2024 by:

Evidence for Policy and Practice Information Centre (EPPI Centre)
Social Science Research Unit, UCL Social Research Institute
UCL Institute of Education, University College London
18 Woburn Square
London WC1H 0NR

https://eppi.ioe.ac.uk
https://www.ucl.ac.uk/ioe/departments-and-centres/departments/ucl-social-research-institute

ISBN: 978-1-911605-57-7

This document is available in a range of accessible formats including large print. Please contact the Social Science Research Unit for assistance.

Email: ioe.ssru@ucl.ac.uk
Telephone: +44 (0)20 7331 5263