Systematic mapping of global research on health systems financing and health economics using machine learning

LIST OF AUTHORS WILL BE FINALIZED AT SUBMISSION

Pierre Marion[1,2]

Brian Hutchinson[1]

Sanghwa Lee[1]

Suvarna Pande[1]

Lucas Sempe[1]

Mark Engelbert[1]

Anilkrishna B. Thota[1]

[1]*International Initiative for Impact Evaluation (3ie)*

[2]*Economics Department, Business School, University of Sussex, Brighton, UK*

Protocol

June 2026

Abstract

WILL BE COMPLETED

Keywords: health systems financing; health economics; machine learning; systematic; mapping; global.

Plain language summary

WILL BE COMPLETED

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
COVID-19	Coronavirus Disease 2019
DEP	Development Evidence Portal
EGM	Evidence Gap Map
FCDO	Foreign, Commonwealth & Development Office
HICs	High-Income Countries
IE	Impact Evaluation
L&MICs	Low- and Middle-Income Countries
LLM	Large Language Model
ML	Machine Learning
OECD	Organisation for Economic Co-operation and Development
PFM	Public Financial Management
PICoST	Population/Problem, Interest, Context, Study Design, Time and Scope
PRISMA-P	Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols
RIS	Research Information Systems format
SR	Systematic Review
T/A	Title and Abstract
WHO	World Health Organization

Introduction

The design of national health systems has profound implications for population health. Countries that mobilize financial resources, pool funds centrally, prioritize cost-effective investments, and maintain strong public financial management are better positioned to maximize population health and longevity (Lastuka et al. 2025). Better health in turn can be an engine of development, fostering economic growth through its role as an investment in human capital (Shawa, Hollingsworth, and Zucchelli 2024).

The stakes are clear, but not without obstacles. The COVID-19 outbreak illustrated the substantial weaknesses and vulnerabilities of health systems globally (Tulenko and Vervoot 2020). In high income countries (HICs), aging populations, new health technologies, and other growing pressures on health systems mean that increases in health spending are projected to outpace economic growth (OECD 2025). Meanwhile in the face of 21 percent cuts in development assistance for health (Institute for Health Metrics and Evaluation 2025), many low- and middle-income countries (L&MICs) face urgent questions about how to raise new funds. All countries can benefit from managing existing ones more efficiently.

In this context, there is a need for transparent evidence to support the march toward stronger, more efficient, and resilient health systems. On the upside, the existing health financing knowledge base to draw from is massive. In one database alone (PubMed), a simple search using terms related to health finance and health economics yields nearly 450,000 records. Yet, the scale and diversity of the evidence mean it is increasingly difficult to make sense of using traditional methods.

Moreover, there have been few attempts to organize evidence in a way that facilitates its relation to—and assessment of—core health financing functions. Several studies have examined the rise in the number of studies and extracted bibliographic information (for example: Wagstaff and Culyer 2012; Gschwent et al. 2024). Munar et al. (2019) produced an Evidence Gap Map (EGM) of impact evaluations (IEs) and systematic reviews (SRs), examining the effects of performance measurement and management strategies in primary healthcare systems in L&MICs. Musiega et al. (2024) conducted a scoping review of studies focusing on the influence of public financial management (PFM) on health system efficiency.

Our study will adapt and extend recent advances in machine-learning–assisted evidence to deliver the first systematic mapping of the global health systems financing within the broad health economics literature. We will build on Berrang-Ford et al. (2021), who applied similar methods to map the literature exploring the relationship between climate change and health. We'll take a comprehensive approach to identify and map the published research on health systems financing and its linkages across the health economy. Specifically, we aim to document the universe of empirical evidence on how health resources are raised, pooled, managed, and allocated, and on the relationship of those functions to health and economic outcomes.

The resulting evidence map will outline the contours of the existing health evidence base, providing researchers, funders, and decision-makers with insight into which policy-

relevant areas are well studied and which remain underexplored. This protocol first presents the conceptual framework we rely on for this research. We then discuss the methods we will use for this global mapping, including the inclusion criteria, our search, screening, and data extraction strategy. We also explain how we intend to present the evidence base.

Conceptual framework

The World Health Organization (WHO) has promoted sustainable health systems financing (HSF) and best practices for reaching universal health coverage since the early 2000s, notably with the 2010 World Health Report (Alipouri Sakha et al. 2024). While many health financing frameworks exist (for example: Shakarishvili et al. 2010; Bertone and Meesen 2013), the Kutzin (2001) framework on health systems financing has been applied and revised in recent years (for example: Kutzin 2013; McIntyre and Kutzin 2016; Kutzin et al. 2017; Jowett et al. 2020). To set the foundation of our study, we use the latest iteration from WHO's 2022 Health System Performance Assessment Report (Papanicolas et al. 2022). The framework depicts core health financing functions: resource raising, pooling, purchasing, and governance and their relationship to other functions and goals of the health system.

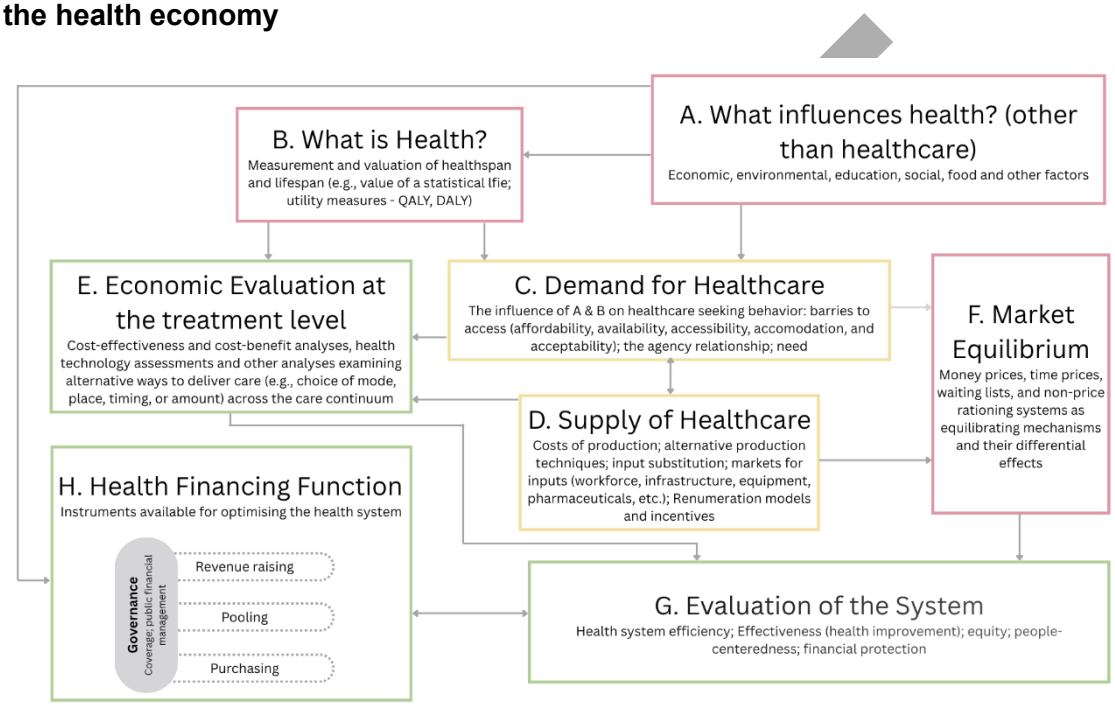
However, core health systems financing decisions are not made in a vacuum. They are informed by a broader health economy, the expanse of which is depicted in Williams (1987) widely cited schematic presentation of the main elements in health economics. The "plumbing diagram" depicts the relationship between: (A) Factors external to health (social determinants); (B) how individuals (and society) place a value on health; (C) demand and (D) supply of health care; (E) micro-economic evaluations of ways of delivering care, (F) market equilibrium conditions; (G) evaluation at the health system level; and (H) health financing planning, budgeting and monitoring mechanisms of the health economy.

As a broad overview, the plumbing diagram helps identify where to look for health-economic evidence that influences—but may not necessarily be taken up or evaluated within—the sphere of health system financing. For example, economic evaluations form a massive literature that can support decision-making. But, they do not sit neatly within any single HSF function and many health systems lack formal mechanisms for translating such evidence into policy. In turn, economic evaluations could not operationalize without the domain of research that supports the valuation of health: for example, quantifying utility weights and estimating the value of a statistical life.

Together then, the Health Systems Financing framework and schematic of the health economy form a map of areas of interest. In Figure 1, we unify the two frameworks to illustrate our study's scope and apply filters to whittle it down to core topics of interest. Beyond examining HSF as a whole and relevant micro- and macro-level evaluations, we focus on evidence regarding how HSF directly influences financial decision-making among patients and healthcare providers. We exclude research on non-financial aspects of the health economy, such as the physical implementation of health systems and health-related decision-making driven by health system design, value-for-health considerations, or external factors.

Boxes in green indicate that the concepts are within this study's scope, red indicates that they are out of scope, and yellow indicates that only a subset of concepts is included. For example, we do not consider research within the "What is Health?" box. This research is outside the scope of this systematic mapping as it is only indirectly related to HSF. Within the "Demand for Healthcare" box, we consider studies focusing on own health financing, such as research on financial protection, but do not include studies related to the social determinants of demand for healthcare or dimensions of health care access, such as availability or acceptability. Appendix A provides the full decisions and justifications.

Figure 1. Framework for mapping the research on health systems financing within the health economy



Note: Adapted from Williams (1987) and Papanicolas et al. (2022). Boxes in green indicate that the concepts are included in this study's scope, red means that they are out of scope, and yellow highlights that only a subset of concepts are included. Box H, originally labeled "Planning, budgeting and monitoring mechanisms" in Williams's schematic, was renamed "Health Financing Functions" to represent the place of the Health Financing Framework within the original Williams schematic. Other minor adjustments were made to modernize language.

Health Financing Functions (box H)

Revenue-raising involves the financing of health systems. In most countries, national populations are major sources of health financing (Papanicolas et al. 2022) through contributions through domestic public financing sources (such as general taxation or social insurance), compulsory or voluntary health insurance schemes, and out-of-pocket payments. In low- or lower-middle income countries, development assistance for health—consisting of funding from foreign governments, international organizations, or private foundations—can also respectively form about five and 32 percent of total health expenditures (World Bank 2025). A central challenge of revenue-raising is the need for predictable and sustained funding over time (Zimmermann et al. 2025).

Revenue-pooling refers to how money that is raised is combined and managed "on behalf of some or all of the population" (McIntyre and Kutzin 2016). Risk pools can be wide in

coverage, for example formed as a single national pool, or fragmented. Smaller pools may be formed based on characteristics of belonging, for example employment status, affiliation with a social insurance fund or private insurer, or residence within a given geography. The size of risk pools and the diversity of the populations and services they cover shape how effectively health revenues can be redistributed to those in need. As a result, the design and integration of pooling arrangements are central topics in health financing policy (Mathauer et al. 2020).

Purchasing refers to paying health providers for the supply of health services. Foundationally, purchasing decisions include which services to buy; who to contract to deliver the services; the model, rates, and other terms of payment; and the extent to which payment is tied to performance (World Health Organization n.d.. These functions can have implications for how resources are distributed, and the extent to which they are distributed equitably, efficiently, and transparently (Cashin and Gatome-Munyua 2022). Important aspects here relate to the market structure of health care provision, which affects competition among providers, and the balance between private and public health providers across services.

Governance covers decisions about coverage and public financial management. Benefit coverage determines who is covered, which services are included, and how much of those costs are financed. These decisions may be made by payers, but are ideally set by higher-level authorities that weigh costs alongside other criteria in the interest of the public good (Papanicolas et al. 2022). Public financial management (PFM) refers to regulations, processes, and institutions that oversee how public funds are managed; their scope and effectiveness have significant implications for health system efficiency (Musiega et al. 2024).

Supporting evidence from the wider health economy

Models and mechanisms of HSF vary significantly across countries. The ability to efficiently raise, pool, and purchase, and govern, public resources for health systems determines the availability, quality, and sustainability of health services. In turn, HSF shapes health providers' and patients' decisions. The demand for health (box C) and the supply of healthcare (box D) operate in a complex environment and interact with each other. Grossman (1972; 1999) introduced the “human capital model of the demand for health”. Health is treated as a durable capital stock, equivalent to human capital, such as education or skills, in a household production function, which depreciates with age and increases with health investment. Many factors, such as the level of health technologies, incentive structures, and the market structure of health providers influence supply. Based on demand- and supply-side factors, market equilibrium conditions emerge (box F).

Within both the demand for and the supply of health care, aspects affected by HSF are present, which we aim to capture within our scope. In the absence of insurance or government funding, demand for health care is heavily influenced by patients’ ability to finance their own health (for example: Lagarde and Palmer 2011). Other non-financial aspects, such as care-seeking behaviors and barriers (time, psychological, and formal), are also central in demand for health care but not directly related to HSF. We will not consider research linked to individual or household decision-making arising resulting from health behaviors or external influences (boxes A and B). For supply, cost of delivery of

health care services, including the health workforce, equipment and infrastructure, are related to HSF purchasing function (for example: Prinja et al. 2016).

Evaluation at the health system level (box G) and micro-economic evaluations (box E) are relevant in this research. Evaluation systems are in place to ensure that the health economy is efficient, equitable and accessible at the macro-level (box G). Governments can improve equity in health systems by supporting access for patients from isolated or disadvantaged groups. Efficiency determines how many high-quality health services patients can access, given the available resources. Final outcomes include the coverage of high-quality services for patients by providers, financial sustainability, and financial protection for households and individuals. Micro-economic evaluations (box E) compare alternative interventions by examining their costs and consequences. These evaluations assess cost-effective strategies to deliver services across the care continuum, informing public financial management and other payer decisions.

Methods

This section outlines the methodology for this systematic mapping. In addition to the inclusion criteria and search strategy, we explain in detail how machine learning and LLMs will be used to screen studies, extract key characteristics, and analyze the data. We will build on the approach established by Berrang-Ford et al. (2021) to conduct this systematic mapping in health. Their study leveraged recent machine learning techniques, this study mapped and categorized more than 15,000 published studies on climate change and human health between 2013 and 2019. Drawing on similar methods, we will synthesize the extensive body of evidence within our scope. In addition, we aim to advance systematic mapping methods by integrating machine learning and LLM into new stages, such as full-text screening of included studies. Given the scale and complexity of the evidence base, this approach will be an iterative and adaptive process, with some components remaining exploratory and experimental.

We will follow principles of systematic mapping as developed by James et al. (2016) and the PRISMA-P checklist 2015 (Shamseer et al. 2015). We outline the methods employed in more detail below. The completed PRISMA-P checklist 2015 is in Appendix B. This systematic review was registered with PROSPERO (XXXID Number). We will develop an advisory group (AG) comprising evidence end-users and representatives from other global stakeholders. Criteria for including or excluding studies

Criteria for including or excluding studies

Table 1 summarises the inclusion criteria used to select studies for the systematic mapping following the PICoST approach: population/problem (P), interest (I), context (Co), Study Design (S), and time and scope (T). We included a study that meets all the inclusion criteria.

In this systematic mapping, we will include any study that empirically analyses interventions, phenomena, concepts, relationships or issues related to health systems financing and health economics. We will consider published evidence in English since 2010 to capture global research conducted following the publication of the WHO 2010

World Health Report. This report was the first major international effort to promote sustainable financing for health systems and to outline best practices for achieving universal health coverage. It put the focus on health systems financing and health economics in policy and research circles. Also, the relevance of evidence prior to 2010 may be low in current contexts.

Table 1. Summary of inclusion criteria

Criteria	Description
Population	We will focus on global populations
Interest	Evidence on health systems financing and its links through the wider health economy. Examples of specific areas of interest within box H in Figure 1 are presented in Appendix C.
Context	No limitations based on location, demographic, social, or health characteristics.
Study Design	We include studies that adopt qualitative, quantitative, or mixed-methods approaches, regardless of methodological rigor. This criterion excludes non-empirical works. The specific study designs included are presented in Appendix D.
Time and scope	Articles and reviews published since 2010. We will include studies published in English only (English-only search terms). We will restrict inclusion to published studies (we exclude grey literature).

We provide more detailed information on the inclusion and exclusion criteria, with specific examples, in Appendix E.

Search strategy

We will adopt a search strategy following guidelines for systematic literature searching (MacDonald et al., 2024). We will develop targeted search strings designed to capture the full range of financing and economic themes worldwide. We will conduct two distinct search processes, one related to health systems financing studies and the other related to health economics studies.

Search terms will be informed by a review of key primary studies, existing reviews, and reports produced by key institutions. Draft search strategies will be discussed and critiqued by team members and subject-matter experts and refined iteratively prior to deployment in major academic databases.

The full list of literature sources to be used and a preliminary example of the search strings for one database are presented in Appendix F. The precise strings and logic (e.g., index terms and truncation operators) will be adapted for each database and platform.

Screening protocol

The selection of studies for data extraction as part of the map will be managed using EPPI-Reviewer (Thomas et al., 2024) and Zotero, and will be completed by implementing the following steps:

- **Import study records:** All output files (e.g. RIS or .txt files) from the search strategy will be imported into Zotero/EPPI-Reviewer.

- **Removal of duplicate studies:** An automated process within EPPI-Reviewer will be used to remove duplicate files. A second layer of deduplication will be performed utilising R and Python libraries and *ad hoc* scripts.

- **Title and abstract human screening:** Titles and abstracts (T/A) of all imported and de-duplicated records will be screened in the following stages:

    - ***Machine-learning-assisted prioritisation.*** We will use EPPI-Reviewer's active learning functionality to prioritise the screening queue. An initial training set of [500–1,000] records will be screened manually to seed the classifier, after which records will be ranked by predicted probability of inclusion and screened in descending order. Screening will continue until a pre-specified stopping criterion is met — either (a) a target recall of 95% estimated against a held-out validation sample, or (b) N consecutive exclusions in the ranked queue (following Callaghan & Müller-Hansen, 2020). The probability threshold for inclusion in the screened set will be calibrated empirically against the validation sample rather than fixed a priori.

    - ***Human screening protocol.*** Screening will be performed single-blind with second-screener adjudication of uncertain items, following Shemilt et al. (2016). One screener will assign each record a status of "include," "exclude," or "undecided"; all "undecided" items, plus a 10% random audit sample of "include" and "exclude" decisions, will be independently screened by a second reviewer. Exclusion reasons will be coded hierarchically and applied in that order so the first failing criterion is recorded. The full code list is provided in Appendix G. Inter-rater agreement on the audit sample will be reported using Cohen's κ and prevalence-adjusted bias-adjusted kappa (PABAK), given the expected low inclusion prevalence. The output of this stage is a human-adjudicated ground-truth set used to benchmark the LLM pipeline.

    - ***LLM screening pipeline.*** We will develop Python scripts that call two LLMs from different model families (e.g., Claude and Mistral) using a shared structured prompt that mirrors the human protocol: the same hierarchical exclusion codes, an inclusion/exclusion decision, a rationale, and a verbatim supporting span extracted from the abstract. The verbatim span will be validated by exact string match against the source record; failures trigger a re-prompt. Each record is screened k=5 times per model with paraphrased prompt variants, yielding a per-model agreement rate (0/5 to 5/5) that serves as the operational confidence score for that decision. Per-model prompt variants and seed configurations are documented and version-controlled.

- ***Calibration against ground truth.*** We compare LLM outputs to the human-adjudicated set on three metrics: sensitivity, specificity, and Cohen's κ. We also assess how well the agreement-rate confidence scores track actual inclusion rates: predictions are grouped into deciles, and within each decile we compute the observed inclusion rate in the ground-truth set. This produces a calibration curve, summarised by two standard measures — expected calibration error (ECE) and the Brier score for the accuracy of probabilistic predictions. The pipeline is refined iteratively through prompt revision, clearer exclusion codes, and few-shot examples where needed, until it meets three pre-specified thresholds: sensitivity ≥ 0.95, κ ≥ 0.7, and ECE ≤ 0.10.

- ***Application to the remaining corpus.*** Once thresholds are met, the validated pipeline is applied to all remaining T/As. Each record receives a calibrated confidence score from each model; records on which the two engines disagree at the decision level, or on which either model returns a confidence score in a pre-specified uncertainty band (e.g., calibrated probability between 0.3 and 0.7), are routed to a third independently-prompted LLM from a third model family. Persistent disagreement after the third model, plus any record the calibration step flags as low-confidence, is escalated to the core team for human adjudication. The expected human workload at this stage is reported in the protocol as a percentage of total records based on calibration-set disagreement and uncertainty-band rates.

- **Full-text screening:** All records included at T/A stage proceed to full-text screening. PDFs will be retrieved through institutional access, and direct author contact; records for which full text cannot be obtained after [two] rounds of requests are recorded as "not retrievable" and reported in the PRISMA flow diagram

    - *Calibration set.* A random sample of [150–200] full texts is screened by a single reviewer, with a senior team member available to adjudicate cases the reviewer flags as uncertain. This set serves as the human-adjudicated ground truth against which the LLM pipeline is benchmarked. The hierarchical exclusion codes from T/A apply, with additional codes for criteria assessable only at full text (e.g., specific outcome measures, comparator definitions, analytic methods). The full list is in Appendix G. We acknowledge that single-reviewer ground truth introduces a known limitation: systematic errors by the reviewer will be propagated rather than caught. We mitigate this through the LLM audit described below and report the limitation transparently.

    - *LLM screening pipeline.* The two-model pipeline used at T/A is extended to full text with two adaptations. First, prompts include the full set of exclusion codes, including the full-text-only criteria. Second, because full texts exceed reliable single-prompt context, each document is segmented (abstract, methods, results, discussion) and the model is prompted to return a decision plus verbatim supporting spans from the methods and results sections, validated by exact string match. Spans drawn only from

the abstract are rejected and re-prompted, since the model has already seen the abstract at T/A and we need the full-text decision to reflect the full text. Each record is screened k=3 times per model with paraphrased prompt variants.

- *Calibration against ground truth.* We compare LLM outputs to the human-adjudicated set on sensitivity, specificity, and Cohen's κ. Agreement-rate confidence scores are calibrated by decile-binning against observed inclusion rates, yielding a calibration curve summarised by ECE and Brier score. The pipeline is refined through prompt revision, clearer exclusion codes, and few-shot examples until it meets three pre-specified thresholds: sensitivity ≥ 0.95, κ ≥ 0.7, and ECE ≤ 0.10. Because full-text exclusions are harder to recover from than T/A exclusions, sensitivity is the binding constraint and we accept lower specificity if needed to meet it.

- *Application to the remaining corpus.* Once thresholds are met, the validated pipeline is applied to all remaining full texts. Records on which the two engines disagree, or on which either model returns a calibrated confidence in the pre-specified uncertainty band, are routed to a third independently-prompted LLM. Persistent disagreement after the third model, plus any record flagged as low-confidence, is escalated to the core team for human adjudication. In addition, a 10% random audit of LLM-confident exclusions is conducted by the core team to detect systematic mis-exclusion that the calibration step would not catch.

- **Checks for linked publications:** Using LLMs, the project team will attempt to group publications that focus on the same intervention and study population (i.e., publications that report on the same study). This typically occurs when an author group publishes more than one paper related to a single study on a specific population, for example, a working paper before a journal article. The latest publication will be classified as the main record for that group of linked studies. Descriptive information will be extracted only once for each group of linked publications, drawing on all linked publications to ensure the extraction is as comprehensive as possible.

- **Retracted studies**: The team will use Zotero Retraction Watch data feature, which enables automatic flagging of retracted studies within the dataset we included after the screening. The process can support the identification and removal of a number of retracted studies from our dataset. The team will also explore additional automated approaches for retraction checking beyond Zotero, such as potential integration of an R-based retraction checking package into the workflow. The team recognizes that retracted studies represent only one subset of potentially unreliable evidence. Some fabricated, manipulated or otherwise critically flawed studies may not undergo retraction processes and may not be identifiable through the above measures. The team will continue exploring any additional feasible safeguards and consultation approaches throughout the project, and will report these efforts transparently. Limitations relating to the identification of such non-retracted but potentially unreliable studies will be acknowledged in the final outputs.

Each step in this process will be documented in detail and graphically presented in a flow chart in the final report for replicability and transparency.

Data extraction

Data will be extracted from all included studies using a structured template covering: bibliographic details, study context (country, setting, time period), population characteristics, study design and methods,[1] and financial and economic topics. Below, we discuss in more detail the use of machine-learning topic-mapping analysis to identify topics and thematic clusters across the included literature based on textual data. We will used the MeSH Tree Structures from the National Library of Medicine to guide the identification of topics. Based on this process and discussions with the AG, we will finalize the data extraction form.

Calibration set. A random sample of [10–30] included studies is extracted by a single reviewer, with a senior team member adjudicating uncertain fields. This set serves as the human-extracted ground truth against which the LLM pipeline is benchmarked. As with full-text screening, we acknowledge that single-reviewer ground truth introduces the possibility of propagated systematic error and mitigate this through the LLM cross-check described below.

LLM extraction pipeline. Python scripts call two LLMs from different model families using a shared structured prompt that returns extraction in the template's schema (JSON), with a verbatim supporting span from the source document for each non-trivial field. Spans are validated by exact string match against the source; failures trigger a re-prompt. Documents are segmented (abstract, methods, results, tables, supplementary materials where available) and the model is directed to specific sections for specific fields — methods text for design fields, results text and tables for quantitative outcomes — to reduce the chance of the model inferring values from context rather than reading them. Each field is extracted k=3 times per model; agreement across runs serves as the per-field confidence score.

Field-level accuracy benchmarking. LLM outputs are compared to the human-extracted set field by field, with metrics chosen by field type:

- *Categorical fields* (study design, country, intervention type): exact-match accuracy and Cohen's κ against the human reference. This coding will focus on capturing the general characteristics of the study, including authors, publication date and status, study location, intervention type, level of intervention, outcomes reported, definition of outcome measures, population of interest, study and programme funders and first year of intervention delivery.
- *Numerical fields* (sample sizes, effect sizes, standard errors, p-values): exact-match accuracy for integers (e.g., sample size), and absolute and relative error for continuous values, with a pre-specified tolerance band per field (e.g., effect sizes within ±0.01, sample sizes exact).

[1] The initial data extraction form for bibliographic details, study context (country, setting, time period), population characteristics, and study design and methods is in Appendix H and was piloted on 10 studies.

- *Free-text fields* (intervention description, outcome definition): manual review of a random subsample for fidelity and completeness, since automated metrics are unreliable here.

The pipeline is refined through prompt revision, template clarification, and few-shot examples until it meets pre-specified thresholds per field type: categorical accuracy ≥ 0.95, numerical exact-match ≥ 0.95 within tolerance, and free-text fidelity judged acceptable by the senior reviewer on the subsample. Fields that fail to meet thresholds after refinement are flagged for full human extraction rather than LLM extraction.

Application to the remaining corpus. Once thresholds are met, the validated pipeline is applied to all remaining included studies. Every extracted field carries its confidence score and its verbatim supporting span. Three categories of output are routed to human review:

- Fields where the two models disagree, or where either model's k=3 runs disagree internally.
- All numerical fields used in quantitative synthesis (effect sizes, standard errors, sample sizes), regardless of confidence — these are 100% human-verified against the source document before pooling, given that a single extraction error here can materially shift a meta-analytic estimate.
- A 10% random audit of high-confidence categorical and free-text fields, conducted by the single reviewer with senior adjudication on flagged cases.

Records where the LLM cannot locate a required field return "not reported," with the verbatim absence noted; these are spot-checked by the reviewer against the source.

- Main-stage extraction for SRs: In the case of descriptive and equity-based information, studies will be coded by two coders independently. In the case of SR critical appraisal assessments, studies will first be single coded and then reviewed by a systematic review methods expert. Meetings will be held periodically with coders on the project to provide support and resolve queries.

Analysis and reporting

Analysis of the evidence

We will conduct a descriptive analysis to provide an overview of the included studies and key trends across the extracted dimensions outlined in the Data Extraction section above. The analysis will also identify evidence gaps across the mapped evidence base, including under-researched topics, geographic areas, outcomes, and combinations of these dimensions where empirical evidence clusters appear limited. Based on the findings, we will highlight implications for future health financing and economic evaluation research.

In addition, where feasible, we will explore machine learning-assisted topic mapping[2] to identify thematic clusters across the included literature based on textual data. This approach uses text mining methods to identify patterns of terms across documents, which can be visualised to explore conceptual relationships within the evidence base.

Given the complexity of high-dimensional topic model outputs, dimensionality reduction and clustering techniques may be employed to produce interpretable visual representations of thematic structure. Dimensionality reduction algorithms, such as UMAP (Uniform Manifold Approximation and Projection) or t-SNE (t-distributed Stochastic Neighbour Embedding), transform high-dimensional document-topic distributions into a lower-dimensional space suitable for visualisation. Clustering methods, such as HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)[3], can be used to identify and delineate clusters of semantically similar documents or topics in this reduced space. The resulting visual output allows to understand the structural relationships between emergent themes, facilitating a more intuitive interpretation of the overall thematic landscape of the data.

Reporting

We will produce the following analytical and reporting outputs:

- **Dynamic, filterable mapping:** A dynamic, filterable systematic evidence map that visually presents the current evidence base on the pre-determined topic-outcome framework. Filters will be incorporated into the map to enable a more targeted use – for example, by filtering the studies by specific geographic focus.
- **Interactive AI-powered chatbot:** A chatbot will be developed to allow users to query and explore the mapped evidence through questions and responses, which aim to improve user accessibility and engagement with the evidence base.
- **Data visualizations:** A range of complementary data visualizations, including heatmaps, geospatial maps, time-trend visualizations, and topic maps, will be developed to support rapid exploration and communication of key patterns and trends in the evidence base. Heatmaps will be used to visualise patterns of co-occurrence among topics and other key variables across the included studies. Geospatial and temporal visualizations will illustrate the geographic distribution and any emerging trends of the evidence base over time.

[2] Topic modelling is an unsupervised machine learning technique that automatically scans a collection of documents and identifies recurring patterns of words to uncover hidden thematic structures by grouping content into abstract "topics" without needing any predefined labels or human guidance. By analyzing patterns of word co-occurrence across documents, topic models assign probabilistic distributions of topics to each document without requiring predefined categories or manual labelling. A foundational method in this field is Latent Dirichlet Allocation (LDA), introduced by Blei, Ng and Jordan (2003), which treats each document as a mixture of topics and each topic as a probability distribution over words. More recently, transformer-based approaches such as BERTopic, which leverages Bidirectional Encoder Representations from Transformers (BERT), have advanced the field by incorporating contextual word embeddings to produce more semantically coherent and nuanced topic representations (Grootendorst, 2022). Together, these methods provide tools for systematically exploring, organizing, and interpreting thematic patterns across large textual data.

[3] Handles clusters of varying density and does not require the number of clusters to be specified in advance. (McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN: Hierarchical density based clustering. Journal of Open Source Software, 2(11), 205).

- **Presentation of findings:** The project team will deliver a presentation summarizing emerging findings and implications. It will provide an opportunity for the Wellcome Trust and other stakeholders to comment on the findings and collaboratively discuss additional analyses, interpretation of results and implications.
- **Publication-ready manuscript:** A manuscript will be prepared for submission to a peer-reviewed academic journal. It will present the methodological approach, study findings including key evidence trends and gaps, and policy-relevant implications for future research priorities.
- **Open and reusable research outputs:** Open, reusable datasets and code will be made available to the public, where feasible, to support data transparency and reproducibility.

Roles and responsibilities

WILL BE COMPLETED

Declaration of interest

All authors declare no conflicts of interest. While 3ie and Wellcome Trust have collaborated in developing the mapping scope and framework, Wellcome Trust will not influence its implementation analysis. 3ie and Wellcome Trust will collaborate again during the reporting and dissemination stages to help ensure the outputs are useful and relevant to Wellcome Trust and the wider sector.

Competing interests: No competing interests were disclosed

Sources of support

This project is funded by the Wellcome Trust.

Data availability

Underlying data:

No underlying data are associated with this protocol

Acknowledgement

WILL BE COMPLETED

The views and opinions expressed in this protocol are those of the authors and do not necessarily reflect those of Wellcome Trust or the organizations with which the advisory group members are affiliated.

References

Alipouri Sakha, Minoo, Mohammad Bazyar, and Arash Rashidian. 2024. 'Classification and Focus Comparison of Health Financing Frameworks: A Scoping Review'. *The International Journal of Health Planning and Management* 39 (4): 1146–71. https://doi.org/10.1002/hpm.3755.

Arrow, Kenneth J. 1963. *Uncertainty and the Welfare Economics of Medical Care*. 53 (5): 941–73. https://assets.aeaweb.org/asset-server/files/9442.pdf.

Baicker, Katherine, Sarah L. Taubman, Heidi L. Allen, et al. 2013. 'The Oregon Experiment — Effects of Medicaid on Clinical Outcomes'. *New England Journal of Medicine* 368 (18): 1713–22. https://doi.org/10.1056/NEJMsa1212321.

Berrang-Ford, Lea, Anne J. Sietsma, Max Callaghan, et al. 2021. 'Mapping Global Research on Climate and Health Using Machine Learning (a Systematic Evidence Map)'. *Wellcome Open Research* 6 (January): 7. https://doi.org/10.12688/wellcomeopenres.16415.1.

Bertone, M. P., and B. Meessen. 2013. 'Studying the Link between Institutions and Health System Performance: A Framework and an Illustration with the Analysis of Two Performance-Based Financing Schemes in Burundi'. *Health Policy and Planning* 28 (8): 847–57. https://doi.org/10.1093/heapol/czs124.

Cashin, Cheryl, and Agnes Gatome-Munyua. 2022. 'The Strategic Health Purchasing Progress Tracking Framework: A Practical Approach to Describing, Assessing, and Improving Strategic Purchasing for Universal Health Coverage'. *Health Systems & Reform* 8 (2): e2051794. https://doi.org/10.1080/23288604.2022.2051794.

Grossman, Michael. 1972. *The Demand for Health: A Theoretical and Empirical Investigation*. Columbia University Press. https://doi.org/10.7312/gros17900.

Grossman, Michael. 1999. *The Human Capital Model of the Demand for Health*. NBER Working Paper No. W7078. https://ssrn.com/abstract=206128.

Gschwent, Lorenz, Björn Hammarfelt, Martin Karlsson, and Mathias Kifmann. 2026. 'The Rise of Health Economics: Transforming the Landscape of Economic Research'. *Health Economics* 35 (1): 52–68. https://doi.org/10.1002/hec.70044.

Institute for Health Metrics and Evaluation. 2025. *Financing Global Health 2025: Cuts in Aid and Future Outlook*. Institute for Health Metrics and Evaluation. https://www.healthdata.org/sites/default/files/2025-07/FGH_2025_FINAL_incl_Translations_2025.07.31.pdf.

James, Katy L., Nicola P. Randall, and Neal R. Haddaway. 2016. 'A Methodology for Systematic Mapping in Environmental Sciences'. *Environmental Evidence* 5 (1): 7. https://doi.org/10.1186/s13750-016-0059-6.

Jowett, Matthew, Joseph Kutzin, Soonman Kwon, Justine Hsu, Julia Sallaku, and Juan Gregorio Solano. 2020. *Guidance Paper - Assessing Country Health Financing Systems: The Health Financing Progress Matrix*. Health Financing Guidance No. 8. World Health Organization. https://www.who.int/publications/i/item/9789240017405.

Kutzin, Joseph. 2001. 'A Descriptive Framework for Country-Level Analysis of Health Care Financing Arrangements'. *Health Policy* 56 (3): 171–204. https://doi.org/10.1016/S0168-8510(00)00149-4.

Kutzin, Joseph. 2013. 'Health Financing for Universal Coverage and Health System Performance: Concepts and Implications for Policy'. *Bulletin of the World Health Organization* 91 (8): 602–11. https://doi.org/10.2471/BLT.12.113985.

Kutzin, Joseph, Sophie Witter, Matthew Jowett, and Dorjsuren Bayarsaikhan. 2017. *Developing a National Health Financing Strategy: A Reference Guide*. Health Financing Guidance Series No. 3. World Health Organization. https://iris.who.int/server/api/core/bitstreams/e3d41923-bda2-455f-9fb4-b848495e7ee6/content.

Lagarde, Mylene, and Natasha Palmer. 2011. 'The Impact of User Fees on Access to Health Services in Low- and Middle-Income Countries'. *Cochrane Database of Systematic Reviews* 2011 (4). https://doi.org/10.1002/14651858.CD009094.

Lastuka, Amy, Michael R. Breshock, Simon I. Hay, et al. 2025. 'Global, Regional, and National Health-Care Inefficiency and Associated Factors in 201 Countries, 1995–2022: A Stochastic Frontier Meta-Analysis for the Global Burden of Disease Study 2023'. *The Lancet Global Health* 13 (8): e1349–57. https://doi.org/10.1016/S2214-109X(25)00178-0.

MacDonald, Heather, Cozette Comer, Margaret Foster, et al. 2024. 'Searching for Studies: A Guide to Information Retrieval for Campbell Systematic Reviews'. *Campbell Systematic Reviews* 20 (3): e1433. https://doi.org/10.1002/cl2.1433.

Mathauer, Inke, Lluis Vinyals Torres, Joseph Kutzin, Melitta Jakab, and Kara Hanson. 2020. 'Pooling Financial Resources for Universal Health Coverage: Options for Reform'. *Bulletin of the World Health Organization* 98 (2): 132–39. https://doi.org/10.2471/BLT.19.234153.

McInnes, Leland, John Healy, and James Melville. 2018. 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction'. Preprint, arXiv. https://doi.org/10.48550/ARXIV.1802.03426.

McIntyre, Diane, and Joseph Kutzin. 2016. *Health Financing Country Diagnostic: A Foundation for National Strategy Development*. Health Financing Guidance 16.1. World Health Organization. https://www.who.int/publications/i/item/9789241510110.

Munar, Wolfgang, Birte Snilstveit, Ligia Esther Aranda, Nilakshi Biswas, Theresa Baffour, and Jenniffer Stevenson. 2019. 'Evidence Gap Map of Performance Measurement and Management in Primary Healthcare Systems in Low-Income and Middle-Income Countries'. *BMJ Global Health* 4 (Suppl 8): e001451. https://doi.org/10.1136/bmjgh-2019-001451.

Musiega, Anita, Benjamin Tsofa, and Edwine Barasa. 2024. 'How Does Public Financial Management (PFM) Influence Health System Efficiency: A Scoping Review'. *Wellcome Open Research* 9 (October): 566. https://doi.org/10.12688/wellcomeopenres.22533.1.

OECD. 2025. *Health at a Glance 2025: OECD Indicators*. Health at a Glance. OECD Publishing. https://doi.org/10.1787/8f9e3f98-en.

Papanicolas, Irene, Dheepa Rajan, Marina Karanikolos, Agnes Soucat, and Josep Figueras. 2022. *Health System Performance Assessment - a Framework for Policy Analysis*. Health Policy Series No. 57. World Health Organization. https://www.who.int/publications/i/item/9789240042476.

Prinja, Shankar, Aditi Gupta, Ramesh Verma, et al. 2016. 'Cost of Delivering Health Care Services in Public Sector Primary and Community Health Centres in North India'. *PLOS ONE* 11 (8): e0160986. https://doi.org/10.1371/journal.pone.0160986.

Shakarishvili, George A., Atun Rifat, Peter Berman, William Hsiao, Chris Burgess, and Mary Ann Lansang. 2010. *Converging Health Systems Frameworks: Towards A Concepts-to-Actions Roadmap for Health Systems Strengthening in Low and Middle Income Countries*. 3. http://ghgj.org/Shakarishvili_Converging%20Health%20Systems%20Frameworks.pdf.

Shamseer, L., D. Moher, M. Clarke, et al. 2015. 'Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015: Elaboration and Explanation'. *BMJ* 349 (jan02 1): g7647–g7647. https://doi.org/10.1136/bmj.g7647.

Shawa, Ken Chamuva, Bruce Hollingsworth, and Eugenio Zucchelli. 2024. 'A Systematic Review and Meta-Analysis on the Effects of Ill Health and Health Shocks on Labour Supply.' *Systematic Reviews* 13 (1): 52. https://doi.org/10.1186/s13643-024-02454-y.

Shemilt, Ian, Nada Khan, Sophie Park, and James Thomas. 2016. 'Use of Cost-Effectiveness Analysis to Compare the Efficiency of Study Identification Methods in Systematic Reviews'. *Systematic Reviews* 5 (1): 140. https://doi.org/10.1186/s13643-016-0315-4.

Tulenko, Kate, and Dominique Vervoort. 2020. 'Cracks in the System: The Effects of the Coronavirus Pandemic on Public Health Systems'. *The American Review of Public Administration* 50 (6–7): 455–66. https://doi.org/10.1177/0275074020941667.

Van der Maaten, Laurens, and Geoffrey Hinton. 2008. *Visualizing Data Using T-SNE*. 9 (86): 2579–605. http://jmlr.org/papers/v9/vandermaaten08a.html.

Wagstaff, Adam, and Anthony J. Culyer. 2012. 'Four Decades of Health Economics through a Bibliometric Lens'. *Journal of Health Economics* 31 (2): 406–39. https://doi.org/10.1016/j.jhealeco.2012.03.002.

Williams, Alan. 1987. 'Health Economics: The Cheerful Face of the Dismal Science?' In *Health and Economics*, edited by Alan Williams. Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-18800-0_1.

World Bank. 2025. ‘External Health Expenditure (% of Current Health Expenditure) - Low Income, Lower Middle Income, Upper Middle Income’. World Development Indicators. https://data.worldbank.org/indicator/SH.XPD.EHEX.CH.ZS?locations=XM-XN-XT.

World Health Organization. n.d. ‘Promoting Strategic Purchasing’. https://www.who.int/activities/promoting-strategic-purchasing.

Zimmermann, Julia, Charlotte McKee, Marina Karanikolos, Jonathan Cylus, and Members of the OECD Health Division. 2025. *Strengthening Health Systems: A Practical Handbook for Resilience Testing [Internet]*. European Observatory on Health Systems and Policies. https://www.ncbi.nlm.nih.gov/books/NBK618256/.

Appendices

Appendix A: Decisions on scope

Williams Plumbing Diagram	Link to WHO Health Financing Framework	Description	Inclusion decision and justification	Illustrative search-strategy keywords
(A) Other than health, what influences healthcare?	**Exterior to recurrent health financing functions**	Social determinants (e.g., income, education, housing) that shape population healthcare seeking	**EXCLUDE** as not clearly related to HSF	
(B) What is Health? What is its Value?	**Exterior to recurrent health financing functions**	Empirical measurement and valuation of health—for example of the value of a statistical life (VSL) or composite utility measures (DALY, QALY)—often quantified using evidence obtained from direct and indirect elicitation studies	**EXCLUDE** as tangential to recurrent financing of health systems. Health valuations are an essential input to the denominator of economic evaluations of health interventions, which may in turn inform quality public financial management or other payer decisions.	
(C) Demand for health care	**The affordability dimension of demands is linked to coverage decisions (i.e. through policies on financial protection)**	Demand for healthcare is shaped by social determinants of health, the perceived value of health, and the "5A" dimensions of access: availability, accessibility, accommodation, affordability, and acceptability.	**INCLUDE** empirical measurements of own health financing such as out-of-pocket (OoP) expenditure given that they are financial in nature and directly influence coverage decisions. **EXCLUDE:** Studies of other 5A components' influence on demand, given that they are more influenced by service delivery and are primarily non-financial in nature.	**Own health financing**: *("financial protection" OR "catastrophic health expenditure" OR "out-of-pocket" OR OOP OR "impoverishing health expenditure" OR "financial ruin")*

Williams Plumbing Diagram	Link to WHO Health Financing Framework	Description	Inclusion decision and justification	Illustrative search-strategy keywords
(D) Supply of health care	**Production costs link to purchasing of goods and services** **Non-monetized inputs are embedded under a separate health system function: resource generation**	The supply of healthcare covers the resources required for care delivery — including the workforce, infrastructure, equipment, and medicines. Strategies and incentives determine whether and how efficiently resources are deployed.	**INCLUDE:** Empirical studies on production costs. These are financial in nature with links to purchasing decisions. **EXCLUDE:** Other studies of expanding resource inputs, or more strategically deploying them, fall more clearly under other health system functions such as resource generation or service delivery.	**Financing/cost of health workforce:** *(workforce OR "human resources" OR HRH OR labo* OR staff OR personnel).* ***Financing/costs of health equipment/infrastructure:*** *(supplies OR pharmaceutical* OR technology OR equipment OR infrastructure OR facilities OR devices OR hospitals OR clinics OR "primary care facilities").* **Delivery:** *("service delivery" OR "health service delivery" OR "primary health care")* **AND Linking:** *(financing OR "provider payment" OR reimbursement)* **Exclude studies: s**uppliers' incentives, suppliers' market structure, information asymmetries, use of manpower, behaviour of suppliers, clinical trial, treatment efficacy.
(E) Micro-economic evaluation at the treatment level	**Direct link to governance and purchasing. Micro-economic evaluations ideally inform quality public financial management or other payer decisions.** **Direct link to intermediate and final outcomes**	Full economic evaluations that compare alternative interventions by examining their costs and consequences. Used to assess cost-effective strategies to deliver services across the care continuum, informing public financial management and other payer decisions.	**INCLUDE:** A key input to health financing decision-making	**Methods:** *("cost-effectiveness analysis" OR CEA OR "cost-utility analysis" OR "cost-benefit analysis" OR "economic evaluation" OR "budget impact analysis" OR "Return-on-investment analysis" OR "Efficiency analysis" OR "Data envelopment analysis" OR DEA "stochastic frontier" OR "Fiscal incidence analysis" OR "health technology assessment" OR HTA OR "allocative-efficiency analysis")* **Tools:** *("opportunity cost")* **Application:** *("priority setting" OR allocation OR distribution OR "needs assessment" OR "population needs" OR optimi* OR "funding formula")*
(F) Market Equilibrium	**??**	The point at which the supply of and demand for healthcare are balanced through a combination of financial costs, waiting times, and rationing practices. How the balance is struck has consequences for who gets care, when, and on what terms.	**TBD**	

Williams Plumbing Diagram	Link to WHO Health Financing Framework	Description	Inclusion decision and justification	Illustrative search-strategy keywords
(G) Evaluation at the system level	**Direct link to intermediate and final health system outcomes.**	Assessments of health system performance against stated goals (e.g., efficiency, effectiveness (health improvement), equity, safety, user experience, access, quality).	**INCLUDE: Assessment of instruments of recurrent health system financing instruments against health and societal system goals is a core element of the evidence mapping.**	**No separate searches conducted.** All studies identified through other searches are assessed to identify measured outcomes, including both health system goals (e.g., efficiency, effectiveness (health improvement, equity, safety, user experience, access, quality) and societal goals related to economic development.

(H) Planning, budgeting, & Monitoring Mechanisms	**Recurrent Financing of Health Services:**
	• **Revenue raising**
	  ○ **Domestic**
	  ○ **External**
	• **Pooling**
	• **Purchasing**
	• **Governance**
	  ○ **Coverage decisions (i.e. benefits design)**
	  ○ **Public financial management**
		WHO's core health financing sub-functions (see descriptions in protocol narrative)	**INCLUDE**	**Revenue:** *("revenue raising" OR taxation OR "tax simulation" OR "earmarked tax" OR premium* OR "contributions" OR "premiums" OR "payments" OR "foreign aid" OR "external funding" OR "development assistance" OR "collection" OR "financial gap analyses" OR level AND revenue OR trend AND revenue OR sources AND revenue)*
				Pooling: *("risk pooling" OR "risk arrangements" OR insurance OR SHI OR pool OR fund OR "risk sharing" OR "solidarity mechanisms" OR "redistributive mechanisms" OR "mutual health organisations" OR prepayment)*
				Purchasing: *("purchasing" OR "provider payment" OR payment OR reimbursement OR contract OR commission OR allocation OR budget OR "performance-based financing" OR "pay for performance" OR "strategic purchasing" OR "procurement process" OR "procurement systems", OR "costing" OR "pricing" OR "Multi-Criteria Decision Analysis" OR MCDA OR "Program Budgeting and Marginal Analysis" OR PBMA)*
				Governance: *("governance" OR stewardship OR planning)*
				PFM angle: *AND ("public financial management" OR PFM OR "budget execution" OR "fiscal space" or tracking OR "budget setting" OR "budget formula" OR "budget prioriti*ation" OR "budget reporting" OR "budget monitoring" OR "budget transparency" OR "budget accountability" OR "anti-corruption" OR "state-building" OR "state-legitimacy" OR "bottlenecks in fund flows" OR "fund reach" OR "fund disbursement" OR "budget execution challenges" OR "provider autonomy" OR "provider accountability")*
				Exclude no focus on: cost of accessing health.

Williams Plumbing Diagram	Link to WHO Health Financing Framework	Description	Inclusion decision and justification	Illustrative search-strategy keywords
				Exclude no focus on supply-side economics: *suppliers' incentives, market structure, information asymmetries* **Exclude no focus on demand-side economics:** *price of care, time to access case, barriers, need, agency,* **Exclude broad:** *clinical governance" OR corporate governance* Exclude health systems service implementation: *NOT* Manpower Allocations; Norms; Regulation, etc. and the Incentive Structures they generate.

Appendix B: PRISMA-P checklist 2015 completed

PRISMA-P (Preferred Reporting Items for Systematic review and Meta-Analysis Protocols) 2015 checklist:

Section and topic	Item No	Checklist item	Authors response
ADMINISTRATIVE INFORMATION			
Title:			
Identification	1a	Identify the report as a protocol of a systematic review	Yes
Update	1b	If the protocol is for an update of a previous systematic review, identify as such	NA
Registration	2	If registered, provide the name of the registry (such as PROSPERO) and registration number	
Authors:			
Contact	3a	Provide name, institutional affiliation, e-mail address of all protocol authors; provide physical mailing address of corresponding author	Yes
Contributions	3b	Describe contributions of protocol authors and identify the guarantor of the review	Yes
Amendments	4	If the protocol represents an amendment of a previously completed or published protocol, identify as such and list changes; otherwise, state plan for documenting important protocol amendments	NA
Support:			
Sources	5a	Indicate sources of financial or other support for the review	Yes
Sponsor	5b	Provide name for the review funder and/or sponsor	
Role of sponsor or funder	5c	Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol	
INTRODUCTION			
Rationale	6	Describe the rationale for the review in the context of what is already known	Yes
Objectives	7	Provide an explicit statement of the question(s) the review will address with reference to participants, interventions, comparators, and outcomes (PICO)	Yes
METHODS			
Eligibility criteria	8	Specify the study characteristics (such as PICO, study design, setting, time frame) and report characteristics (such as years considered, language, publication status) to be used as criteria for eligibility for the review	Yes

Information sources	9	Describe all intended information sources (such as electronic databases, contact with study authors, trial registers or other grey literature sources) with planned dates of coverage	Yes
Search strategy	10	Present draft of search strategy to be used for at least one electronic database, including planned limits, such that it could be repeated	Yes
Study records:			
Data management	11a	Describe the mechanism(s) that will be used to manage records and data throughout the review	Yes
Selection process	11b	State the process that will be used for selecting studies (such as two independent reviewers) through each phase of the review (that is, screening, eligibility and inclusion in meta-analysis)	Yes
Data collection process	11c	Describe planned method of extracting data from reports (such as piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators	Yes
Data items	12	List and define all variables for which data will be sought (such as PICO items, funding sources), any pre-planned data assumptions and simplifications	Yes
Outcomes and prioritization	13	List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale	Yes
Risk of bias in individual studies	14	Describe anticipated methods for assessing risk of bias of individual studies, including whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis	NA
Data synthesis	15a	Describe criteria under which study data will be quantitatively synthesised	NA
	15b	If data are appropriate for quantitative synthesis, describe planned summary measures, methods of handling data and methods of combining data from studies, including any planned exploration of consistency (such as I^2, Kendall's τ)	NA
	15c	Describe any proposed additional analyses (such as sensitivity or subgroup analyses, meta-regression)	Yes
	15d	If quantitative synthesis is not appropriate, describe the type of summary planned	Yes
Meta-bias(es)	16	Specify any planned assessment of meta-bias(es) (such as publication bias across studies, selective reporting within studies)	NA
Confidence in cumulative evidence	17	Describe how the strength of the body of evidence will be assessed (such as GRADE)	NA

Appendix C: Recurrent Financing of Health Services

Table C1: Key topics of interest

Categories	Topics
Revenue raising (domestic and external)	Studies on levels, trends, or sources of revenue (all sources) + financial gap analyses
	Domestic resource mobilisation (including fiscal space analysis, taxation, political economy of taxation, tax simulations, etc.)
	Aid / External funding (including foreign aid/ODA impact analysis, donor dependency, transitions, fungibility)
	Public – private partnership for healthcare funding, market shaping interventions
Pooling and financial risk protection	Studies on pooling mechanisms (including all types of health insurance, automatic, compulsory or voluntary participation in pools)
	Free healthcare / exemptions mechanisms for specific services/target groups
	Targeted exemptions mechanisms for poor (including identification/targeting approaches)
	Health insurance interventions on financial protection
	Fragmentation of pooling / multiple pooling mechanisms, cross-programmatic efficiency analysis
Purchasing & Provider Payment	Provider payment mechanisms, including incentives, performance based financing, direct facility funding
	Strategic purchasing
	Financing of procurement processes and systems (medical devices, pharmaceuticals, supply chain)
	Costing and pricing (services or packages, pharmaceuticals, claims analyses)
Governance - Public Financial Management	Budget setting / formulation (including prioritization of health, budget formatting, PBB etc.)
	Budget execution (public expenditure tracking, bottlenecks in fund flows, measurements of reach or disbursement (e.g., % of funds reaching facilities); qualitative assessments of coordination or budget execution challenges, etc.)
	Digitalization and technical PFM reforms (IFMIS, etc.)
	Provider autonomy and accountability
	Budget reporting and monitoring, transparency, accountability, anti-corruption, state-building/legitimacy through PFM systems
Governance – Coverage Decisions (i.e. benefits decisions)	Studies of waiting lists and other rationing practices
	Benefit-package design, including studies of co-payment/fee policies, cash transfers, and vouchers that influence the price that health system users pay.

	Prioritization processes and methods (e.g., Multi-Criteria Decision Analysis (MCDA)), Program Budgeting and Marginal Analysis (PBMA), resource allocation formulae
Overall	Health financing assessments (health financing matrices, etc.)
	Policy analysis and political economy of health financing choices and reforms

Appendix D: Study Designs

Included studies analysing an intervention:

- **Experimental and quasi-experimental impact evaluations** (randomized controlled trials, natural experiments, difference-in-differences, interrupted time series, event-study analyses, before-after comparisons, etc.),
- **Economic and efficiency analyses** (cost-effectiveness analysis, cost-benefit analysis, cost-utility analysis, efficiency analysis, fiscal incidence analysis, etc.),
- **Policy and implementation analyses** (process evaluations, realist evaluations, policy analyses, etc.),
- **Equity and distributional analyses** (equity impact analysis, distributional effects analysis, access analyses, etc.),
- **Qualitative and mixed-methods evaluations** (key informant interviews, focus groups, etc.),
- **Modelling studies linked to interventions** (microsimulation models, forecasting/scenario analyses, etc.),
- **Reviews of the studies above** (systematic reviews, scoping reviews, mapping, evidence gap maps, etc.).

Included studies which do not analyse an intervention:

- **Descriptive system analyses** (descriptive health financing analyses, health expenditure tracking, resource allocation analyses, etc.),
- **Cross-sectional and correlational analyses** (cross-sectional regressions, correlational analyses, panel-data observational analyses, time-trend analyses, etc.),
- **Comparative health systems analyses,**
- **Equity and population analyses** (distributional analyses, access disparities analyses, financial burden analyses, etc.),
- **Provider/payment system analyses,**
- **Political economy and governance studies** (governance/institutional analyses, etc.),
- **Qualitative exploratory studies** (interview studies, focus groups, etc.),
- **Forecasting studies,**
- **Reviews of the studies above.**

Excluded studies:

Any study which does not analyse empirical data and instead presents opinions, concepts, arguments, or summaries rather than original evidence.

- News article,
- Blogs,
- Briefs,
- Commentaries,
- Opinion pieces,
- Editorials,
- Perspectives,
- Essays,
- Purely theoretical/conceptual/methodological/modelling/measurement studies,
- Protocols without results,
- Advocacy pieces,

- Think pieces/discussion papers,
- Reviews of the studies above.

Appendix E: Detailed screening

Table E1: Detailed screening decisions

Inclusion	Exclusion	Clarification
Interest (I) and Context (Co)		
WILL BE COMPLETED		
Study design (S)		
The study is based on a randomized controlled trial.	The study refers to an impact study but it did not conduct it.	Protocols, editorials, commentaries, opinion pieces will be excluded.
Modelling which uses empirical data	Modelling but no data	
Time and scope (T)		
The study was published on or after 2010	The study was published before 2010	N/A
The study is in English	The study is not in English	N/A
The study is published in a journal.	The study is a conference paper or a thesis.	N/A

Appendix F: Search strategy

Table F1: Academic/bibliographic databases

Academic databases
Agricola
Applied Social Sciences Index and Abstracts (ASSIA)
CAB Abstracts
EBSCO Discovery Service (including GreenFILE, Academic Search Complete, Science Direct, AGRIS, RePEc, World Bank e-Library).
EconLit
Epistemonikos
Global Health
Medline
Embase
PsycINFO
SciELO (Scientific Electronic Library Online)
Scopus
Web of Science (including Web of Science Core Collection (Social Sciences Citation Index (SSCI), Science Citation Index Expanded (SCI-EXPANDED), Conference Proceedings Citation Index – Science (CPCI-S), Conference Proceedings Citation Index – Social Science & Humanities (CPCI-SSH), Emerging Sources Citation Index (ESCI)).

Example of search strings:

Database/Platform: **Ovid MEDLINE**.

Note: lines in bold are those that will be exported for de-duplication and screening

Date: May 26th 2026 (This example search was conducted before some latest updates to Appendix A. The final example search will be updated before submission)

No.	Concept	Search	Results
1	Health	(health or healthcare).ti,ab,kf	3,589,303
2	Financing	(financ* or revenue* or spend* or expend* or purchas* or pay* or resourc* or allocat* or cost*).ti,ab,kf	2,005,997
3	Health + financing	1 AND 2	649,447
4	Own health financing	(protection or catastrophic or out of pocket or oop or impoverish* or financial ruin).ti,ab,kf	475,464
5	**Own + health + financing**	**3 AND 4**	**24,008**
6	Health workforce	(workforce or human resources or hrh or labo?r or staff or personnel or doctor* or nurse*).ti,ab,kf	976,628
7	**Workforce + health + financing**	**3 AND 6**	**106,066**

8	Health equipment/ infrastructure	(supplies or pharmaceutical* or technolog* or equipment or infrastructur* or facility or facilities or device* or hospital* or clinic?).ti,ab,kf	4,249,031
9	**Equipment + health + financing**	**3 AND 8**	**256,198**
10	Service delivery	(service delivery or (primary adj3 care)).ti,ab,kf	238,679
11	**Service delivery + health + financing**	**3 AND 10**	**47,014**
12	**Cost analysis methods**	**("cost-effectiveness analysis" or CEA or "cost-utility analysis" or "cost-benefit analysis" or "economic evaluation" or "budget impact analysis" or "Return-on-investment analysis" or "Efficiency analysis" or "Data envelopment analysis" or DEA "stochastic frontier" or "Fiscal incidence analysis" or "health technology assessment" or HTA or "allocative-efficiency analysis" or opportunity cost).ti,ab,kf**	**82,678**
13	**Cost analysis application**	**("priority setting" or ((resource* or service*) adj2 (allocation or distribution)) or "needs assessment" OR "population needs" or "funding formula").ti,ab,kf**	**43,532**
14	System performance	((system adj3 (performance or effien* or responsive* or productiv*)) or performance measurement or benchmark* or matrix or analysis).ti,ab,kf	7,762,017
15	**Performance + health + financing**	**3 AND 14**	**204,249**
16	Evaluation	(impact evaluation or reform or policy evaluation).ti,ab,kf	47,714
17	**Evaluation + health + financing**	**3 AND 16**	**12,207**
18	**Comparison**	**(cross country comparison or international comparison).ti,ab,kf**	**2,451**
19	**Adequacy**	**(sufficiency of funds or fiscal space for health or domestic resource mobilization).ti,ab,kf**	**76**
20	**Efficiency**	**("administrative efficiency" or "technical efficiency" or "allocative efficiency").ti,ab,kf**	**1,438**
21	**Coverage/PFM**	**("benefit package" OR "comprehensive coverage" OR "public financial management" OR "universal health coverage" OR UHC OR "health coverage").ti,ab,kf**	**10,154**
22	**Equity**	**("equity in health financing" OR "progressivity" OR "Kakwani index").ti,ab,kf**	**356**
23	Revenue	(revenue raising OR taxation OR tax simulation OR earmarked tax OR premium* OR contributions OR premiums OR payments OR foreign aid OR external funding OR development assistance OR collection OR financial gap analys#s OR (revenue adj3 (trend? or revenue or source?))).ti,ab,kf	560,533

24	**Health + Revenue**	**1 AND 23**	**133,757**
25	Pooling	("risk pooling" OR risk arrangements OR insurance OR SHI OR pool OR fund OR risk sharing OR solidarity mechanisms OR redistributive mechanisms OR mutual health organisations OR prepayment).ti,ab,kf	285,868
26	**Pooling + health**	**1 AND 25**	**121,019**
27	Purchasing	("purchasing" OR "provider payment" OR payment OR reimbursement OR contract OR commission OR allocation OR budget OR performance-based financing OR pay for performance OR strategic purchasing OR procurement process OR procurement systems, OR costing OR pricing OR Multi-Criteria Decision Analysis OR MCDA OR "Program Budgeting and Marginal Analysis" OR PBMA).ti,ab,kf	269,579
28	**Purchasing + health**	**1 AND 27**	**105,566**
29	Governance	(governance or stewardship or planning).ti,ab,kf	498,131
30	**Governance + health**	**1 AND 29**	**131,854**
31	Public financial mgmt. (PFM)	("public financial management" OR PFM OR "budget execution" OR "fiscal space" or tracking OR budget setting OR budget formula OR budget prioriti*ation OR budget reporting OR budget monitoring OR budget transparency OR budget accountability OR anti-corruption OR state-building OR state-legitimacy OR bottlenecks in fund flows OR fund reach OR fund disbursement OR budget execution challenges OR provider autonomy OR provider accountability).ti,ab,kf	152,965
31	**PFM + Health**	**1 AND 31**	**20,978**
Total across all exported lines			**823,016**

Appendix G: Full screening code list

Each record receives exactly one exclusion code - the first criterion in the hierarchy that is not met.

Study characteristics	Criterion / Reason for Exclusion	Applies at Stage
Inclusion		
Record meets all inclusion criteria: (1) global population, (2) empirical evidence on health systems financing or its links through the wider health economy, (3) no disqualifying context limitation, (4) eligible study design (qualitative, quantitative, or mixed-methods or listed in Appendix D), and (5) published in English since 2010 in a peer-reviewed outlet.	Include	• Title and abstract • Full text
Unclear		
An LLM cannot determine the inclusion or exclusion status based on the available information. Item is escalated to a human reviewer at the text and abstract stage, either/or the full-text stage.	Unclear: Insufficient information	• Title and abstract • Full text
EXCLUSION CODES - apply in hierarchical order; record the first failing criterion only		
Population - does not meet population criteria		
Study does not focus on any human population (e.g., animal studies, purely theoretical modelling without human data).	Exclude: Population not eligible	• Title and abstract • Full text

Study characteristics	Criterion / Reason for Exclusion	Applies at Stage
Interest - does not address health systems financing or its links through the wider health economy		
Study does not provide empirical evidence on health systems financing or on its connections through the wider health economy.	Exclude: Topic outside scope	• Title and abstract • Full text
Abstract does not provide enough information to judge whether the topic falls within scope. Apply only when unclear cannot be applied.	Exclude: Topic insufficiently described at title and abstract	• Title and abstract
Study design - ineligible study design		
Study does not report original empirical data, empirical calibration or a synthesis of empirical data (e.g., editorials, opinion pieces, conceptual/theoretical papers, policy documents without original data).	Exclude: Does not analyse empirical data	• Title and abstract • Full text
Record is an opinion piece/ conference abstract, purely modelling/theoretical/ conceptual or registered trial without results.	Exclude: Does not analyse empirical data	• Title and abstract • Full text
Record is a duplicate publication or addendum already captured by a primary record.	Exclude: Duplicate report	• Title and abstract • Full text
Time and scope - outside temporal or scope boundaries		
Study was published prior to 1 January 2010.	Exclude: Published before 2010	• Title and abstract • Full text
Full text is not available in English.	Exclude: Not in English	• Title and abstract • Full text

Study characteristics	Criterion / Reason for Exclusion	Applies at Stage
Record is grey literature (e.g., government report, working paper, thesis, NGO publication) rather than a published peer-reviewed article or systematic review.	Exclude: Grey literature	• Title and abstract • Full text
Full-text specific codes - applied at full-text screening only		
Full text does not provide enough methodological information to confirm that the study design is eligible or to support data extraction.	Exclude: Insufficient methodological detail	• Full text
Not retrievable		
Full text could not be retrieved after two rounds of attempts (institutional access + direct author contact). Recorded in the PRISMA flow diagram.	Exclude: Full text not retrievable	• Full text

Appendix H: Initial data extraction

Table H1. Initial data extraction form for bibliographic details, study context, population characteristics, and study design and methods

Code	Subcode
Bibliographic details	Study ID
	Title
	Foreign Title
	Short title
	Language
	Author Name
	Author Affiliation Institution
	Author Affiliation Country
	Publication Type
	DOI
	Study status
	Abstract
	Keywords
	Journal name
	Other journal name
	Journal volume
	Journal issue
	Pages
	Year of Publication
	URL
	Publisher location
	Open access
Study context	Continent name
	Country name
	Additional country
	Country income level
	Region name
	State/province name
	District name
	City/town name
	Time period
Population characteristics	Age
	Sex
	Setting
	Sexual orientation
	Specific population group
Study design and methods	Evaluation Design
	Evaluation Method
	Mixed Method
	Additional quantitative Methods

Code	Subcode
	Unit of Observation
	Data source
	Data Type