# Mapping Public Health Policy Options


Isaac Yen-Hao Chu
Rebecca Glover
Peter Weatherburn
Helen Burchett

LSHTM



SEPTEMBER 2024

## KEY STUDY CONTACTS


| | |
|---|---|
| Chief Investigator | Helen Burchett |
| Key Protocol Contributor | Isaac Yen-Hao Chu, Helen Burchett |
| Co-Investigator | Peter Weatherburn, Rebecca Glover |
| Funder | National Institute for Health Research |


## ROLE OF STUDY FUNDER


This report is independent research funded by the National Institute for Health Research (Reference number: PRU-PRU-1217-20901). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.


## ACKNOWLEDGEMENTS


We would like to express sincere gratitude to Claire Stansfield, Zak Ghouze, Sergio Graziosi and colleagues at the EPPI Centre, University College London for their technical support and cognitive contribution. We are very grateful to Mark Petticrew and Laurence Blanchard at LSHTM for their cognitive input and dedicated support.


## KEYWORDS


Health promotion
Public health intervention
Policy review
Bibliometric analysis
Equity
Artificial intelligence

## Table of Contents

## 1. Executive Summary

### Policy Implications

If DHSC wishes to enhance the effectiveness of public health interventions, it may consider encouraging a more balanced approach in implementation and evaluation. While capability-based interventions (e.g., education, information provision or knowledge acquisition) remain important, combining these with non-capability-based approaches could yield more comprehensive results. DHSC might promote research addressing multiple health determinants simultaneously, including structural and environmental factors.

If DHSC aims to ensure research priorities reflect current policy challenges, it might develop guidance on addressing health inequalities as well as complex system interventions. DHSC could also encourage public health researchers to adopt standardised frameworks to establish evidence on the consequences of under-explored health inequities.

### Background

This study aimed to explore what types of interventions, for which health topics, had been evaluated in trials (TRoPHI) and reviews (DoPHER).

### Methods

We conducted bibliometric analyses of two databases on trials and effectiveness reviews to explore what intervention functions, for which health topics, have been evaluated.

### Results

Most (59%) of trials involved both capability-based (e.g. knowledge provision) and non-capability-based intervention functions. Few (7%) trials aimed to change policies, laws or environments. Over half the trials aimed to change lifestyle behaviours. There were few policy-focused reviews of effectiveness (168; 2%). Of these, lifestyle behaviours were the most common health topic focus (46%), followed by substance misuse (29%). Equity-oriented reviews (i.e. that mention equity, inequalities or disparities in their title or abstract) were also scarce (188; 2%). Among equity-oriented reviews, socioeconomic status was the most commonly addressed factor (38%), followed by age (34%), sex/gender/sexuality (26%) and race/ethnicity (26%).

## Conclusions

This analysis highlights the need for a more diverse approach to public health interventions, moving beyond capability-based strategies to address wider determinants of health. The scarcity of equity-oriented reviews underscores the importance of adopting more inclusive research practices. While AI-assisted evidence synthesis shows potential, human-facilitated refinement is required to enhance its utility in public health policymaking.

## 2. Background

Three decades after the advent of the Dahlgren-Whitehead framework on health determinants,[1] there exists a wide range of policy and intervention options for improving well-being or reducing health disparities at national and local levels, varying from legislation, (dis-)incentivisation, service provision to education.[2] However, research evidence that supports feasible interventions for public health or health promotion (hereafter called public health)[3] or cross-topic learning in public health policy is not always available. The fast-growing and diverse nature of public health research has challenged both researchers and policymakers to grasp which evidence-informed interventions are available to use in specific contexts. Concerns have been raised that most interventions utilise capability-related approaches (e.g., information, knowledge and education) regardless of their research topics, despite evidence that knowledge provision alone is often insufficient to change behaviours and improve health.[4–6] Without a general understanding of what research has been conducted – and where research gaps exist – it can be challenging for policymakers to incorporate research evidence into their decision-making or address gaps through research commissioning. Although moving beyond topic-siloed boundaries has been recognised as a valuable principle in public health policymaking, in practice, it is often too challenging to achieve cross-topic learning with limited time, resources or personnel.

Open-access databases have great potential in facilitating cross-topic learning and in exploring feasible interventions for public health. Funded by the Department of Health and Social Care, the Trials Register of Promoting Health Interventions (TRoPHI)[7] and the Database of Promoting Health Effectiveness Reviews (DoPHER)[8] are two open-access online databases of more than 20,000 trials and 7,000 reviews related to public health (as of June 2024). They have been hosted and updated quarterly by the EPPI Centre at University College London (UCL) since 2010. Both TRoPHI and DoPHER can be searched for published public health trials and reviews respectively, for information on the year of publication, study type and relevant health topics (e.g., alcohol control, tobacco cessation and diabetes management). At present, TRoPHI is currently not searchable for key functions of interventions evaluated in trials (defined as *intervention function*, e.g. incentivisation, role modelling and restructuring environments). The type and extent of equity-focused or policy-oriented reviews in DoPHER also remain under-explored.

With technological advancements in artificial intelligence (AI) and machine learning models, the UCL EPPI Centre has incorporated AI into EPPI-Reviewer by using OpenAI's GPT-4o

Application Programming Interface (API)[9] to allow AI-assisted classification of keywords and health topics in TRoPHI and DoPHER.[10] It is believed that AI has great potential in assisting manual coding and in improving evidence synthesis and visualisation.[10,11] The integration of AI into these databases presents a unique opportunity to explore the nature of the studies included in the cross-topic evidence.

Mapping out health topics addressed in trials and reviews can support decision-making and cross-topic learning by highlighting what public health policy options have been evaluated. For example, it is plausible that innovative interventions targeting upstream, structural determinants of health to reduce health disparities have been evaluated for physical activity but not for mental health. We hope our findings can help ‘bridge the silos’[12] by improving interdisciplinary learning across diverse health topics, demonstrating the breadth of available options for public health interventions (and their evaluation), and may thus support policymakers to identify evidence-based options for public health policies.

## 3. Aim and Objectives

Our overarching aim is to identify what evidence exists to support the use of different public health interventions to improve health while reducing health disparities.

**Original Objectives**

1. Map which public health policies in DoPHER and interventions in TRoPHI have been evaluated for different health topics
2. Analyse if gaps or trends exist in DoPHER for policy reviews and in TRoPHI for intervention evaluations of particular health topics
3. Explore the distribution of published reviews in DoPHER on reducing disparities across health topics
4. Obtain DHSC’s feedback on the above analyses
5. Reflect on challenges and opportunities for using artificial-intelligence-assisted evidence mapping

**Changes to original Objectives**

Due to the sign-off process, our project timeframe was reduced from five months to approximately three months. Since the nature of Objective 4 involves presenting our project findings to DHSC, we decided to reposition Objective 4 to subsequent dissemination events organised by LSHTM's NIHR Public Health Policy Research Unit (PHPRU, expected by the end of 2024), so the potential impacts of our findings can be pragmatically maximised. Objectives 1, 2, 3 and 5 remain unchanged; our original Objective 5 was hereafter renumbered as Objective 4R.

## 4. Methods

**Project context**

We conducted rapid bibliometric analyses of TRoPHI and DoPHER after EPPI-Reviewer (version 6.15.3.0) officially launched GPT-4o-powered automatic coding in June 2024 (the version of GPT-4o model: 1 Feb 2024).

**Description of databases**

*Trials Register of Promoting Health Interventions (TRoPHI)*[7]

Established in 2004, this database includes both randomised and non-randomised trials evaluating public health interventions published worldwide. As of June 2024, it encompasses 19,380 public health trials. The register regularly identifies trials using three approaches: sensitive keyword search (manual search since 2004 and additional automated search since 2022), studies obtained from EPPI Centre's systematic reviews, and records from the Cochrane Health Promotion and Public Health field.

*Database of Promoting Health Effectiveness Reviews (DoPHER)*[8]

Created in 2004, this database aims to collect both systematic and non-systematic reviews of effectiveness in health promotion and public health topics around the world. As of June 2024, it includes 7,850 reviews with full titles and abstracts in public health fields. It has been quarterly updated since 2006 by the EPPI Centre.

**Eligibility criteria**

1. Any study registered in TRoPHI or DoPHER on or before 18 June 2024 (the date when EPPI-Reviewer officially launched AI-powered automatic coding) AND
2. Any study published in or after 1995 (to capture potential temporal changes across three decades, from 1995 to 2024)

**Inclusion criteria**

We included records that meet all conditions below for bibliometric analysis:
For TRoPHI (Objective 1 & 2):

- Trial with full information on its title and abstract (according to EPPI-Reviewer) AND
- Trial with information on health topic(s) AND
- Trial with information on intervention function(s)

For DoPHER (Objective 1, 2 and 3)

- Review with full information on its title and abstract AND
- Review was policy-oriented with information on health topic(s) OR
- Review with a focus on inequity, disparity or equivalent terms

We analysed all records that met both eligibility criteria and inclusion criteria in TRoPHI and DoPHER. As a result, our analyses included 17,279 trials, together with 168 policy-oriented and 188 disparity-focused reviews.

**Operational definitions for data analysis**

Intervention function of trials (Objective 1 & 2)
To ensure the theoretical rigour of our analyses, we first extracted relevant definitions from Michie et al.'s Behaviour Change Wheel (BCW) model,[2] which has been widely used for policy development and evaluation in the UK. Their nine-item intervention functions were defined as key strategies that one can utilise to change behaviour by altering individuals' capabilities, opportunities and/or motivation.[13] Next, we referred to McLeroy et al.'s Socio-Ecological Model[14] to ensure our classifiers cover areas beyond individual levels highlighted in the BCW, such as factors at interpersonal levels, (social networks and community mobilisation), institutional levels and policy levels. Our intervention function included eight categories as below (see Appendix 1 for details):

- *Capability-based intervention function*
  1. Education, skill, knowledge, information, training and equivalents

- *Non-capability-based intervention function (examples)*
  2. Therapeutic/motivational (counselling, consultation)

3. Non-medical product provision (supplement, screening)
4. Social/community network (peer support, community engagement)
5. Reinforcement (incentive, gamification)
6. Improve medical/social services (increase access, increase availability)
7. Change policy/law/environment (legislation, pricing)
8. Unspecified intervention (interventions without detailed information)

To identify intervention functions in trials, one researcher (IYC) applied text-search coding in EPPI-Reviewer. For example, a trial containing the text ‘education’ in its title or abstract was coded as a capability-based intervention function. Coding in this manner was justified because the performance of automatic coding using GPT-4o for coding intervention function(s) was unsatisfactory (Exact Match Ratio of 30%). Manual, human coding of 17,279 trial records was not feasible. To minimise potential concerns about inaccuracy in text-search coding, IYC randomly checked the abstracts and assigned codes (one per 100 records) to reduce misclassification. Appendix 2 lists all search terms with Boolean logic used for coding intervention functions.

Health topic in TRoPHI and DoPHER (Objective 1 & 2)

Ten categories were established to identify public health topics in TRoPHI and DoPHER (see Table 1). We applied GPT-4o-powered automatic coding to identify health topics across all 17,279 trials in TRoPHI and 168 policy reviews in DoPHER.

**Table 1 Health topics in TRoPHI and DoPHER**

| Category | Original codes in TRoPHI and DoPHER |
|---|---|
| 1. Non-communicable disease (NCD) | Asthma<br>Cancer (all types)<br>Cardiovascular<br>Diabetes |
| 2. Substance misuse | Alcohol<br>Drugs<br>Solvents<br>Tobacco |
| 3. Mental health | Eating disorders<br>Mental health<br>Suicide |
| 4. Lifestyle behaviour | Healthy eating<br>Hygiene<br>Leisure<br>Obesity<br>Oral health<br>Parenting<br>Physical activity |
| 5. Inequality | Inequality |
| 6. Sexual and reproductive health (SRH) | Pregnancy prevention<br>Sexual health<br>Sexually transmitted infection |
| 7. (Non-)Accidental incidents | Abuse (all types)<br>Accidents<br>Delinquency (crime)<br>Injury |
| 8. Complex systems | Education systems (including environments)<br>Medical care<br>Workplace |
| 9. Disability | Disability |
| 10. Unspecified | Health promotion (general) |

Defining disparity/inequity

We defined disparity/inequity per Cochrane Collaboration's PROGRESS-Plus framework[15] to identify equity-related reviews: place of residence, race/ethnicity/culture/language, occupation, gender/sex/sexuality, religion, education, socioeconomic status, social capital (including social network) and two PLUS factors (i.e., age and disability). We also documented the number of reviews that employed the PROGRESS-PLUS framework. This task was completed by IYC's manual coding.

Evaluation of accuracy of automatic coding compared to human coding

To evaluate GPT-4o's performance in classifying health topics (Objective 4R), we contrasted its coding to human, manual coding (see Appendix 3 for more details).

## Ethical consideration

After consulting research ethics bodies at both LSHTM and UCL (where TRoPHI and DoPHER were maintained), we confirm that this study does not require ethical approval based on its nature as a bibliometric analysis of published records (titles/abstracts) in public databases.

## 5. Results

Our analyses revealed that, throughout three decades, most trials utilised a combination of capability-based (e.g., education, information provision, knowledge acquisition or training) and non-capability-based intervention functions, although the majority used individualistic, cognitive functions (e.g. capability, therapeutic or motivational). Few reviews focused on either policies or health disparities.

Characteristics of eligible trials in TRoPHI

Our analysis encompassed 17,279 eligible records from TRoPHI, spanning from 1995 to 2024. The temporal distribution of these records revealed a marked increase in public health research over the past three decades (Table 2).

**Table 2 Publication year and intervention functions of trials, 1995-2024 (n=17,279)**

| Characteristic | Count | Proportion* |
|---|---|---|
| *Year of Publication* | | |
| 1995-1999 | 286 | 2% |
| 2000-2004 | 665 | 4% |
| 2005-2009 | 2,796 | 16% |
| 2010-2014 | 3,355 | 19% |
| 2015-2019 | 5,281 | 31% |
| 2020-2024 | 4,896 | 28% |
| *Overview of intervention function* | | |
| Capability-based intervention only | 4,993 | 29% |
| Both capability and non-capability-based | 10,100 | 59% |
| Non-capability-based intervention only | 2,186 | 13% |
| *Category of Intervention function** | | |
| Capability-based | 15,093 | 87% |
| Therapeutic or motivational | 6,783 | 39% |
| Non-medical product provision | 3,713 | 22% |
| Social network or community | 3,577 | 21% |
| Reinforcement | 1,536 | 9% |
| Improve medical/social services | 1,360 | 8% |
| Change policy/law/environment | 1,202 | 7% |
| Unspecified intervention | 62 | <1% |

* The sum of proportions may exceed 100% as each record could contain more than one intervention function.

Intervention functions of trials

The majority of trials in TRoPHI (59%) employed a pluralistic approach by involving both capability-based and non-capability-based intervention functions (see Table 2). Interventions focusing solely on capabilities accounted for more than a quarter (29%) of analysed trials, while 13% evaluated exclusively non-capability-based interventions.

Capability-based interventions were the most prevalent (87%), followed by non-capability-based intervention functions: therapeutic or motivational interventions (39%); non-medical product provision (e.g. screening or provision of masks or supplements) (22%). Evaluations of social network or community-based interventions (e.g. peer or group-based) (21%) were less common, as were reinforcement (e.g. incentives) (9%), improving medical/social services (8%) and changing policy, law or the environment (7%).

There were no notable changes in the distribution of intervention functions over time.

Health topics of trials

Over half the trials (52%) evaluated interventions targeting lifestyle behaviours (e.g. healthy eating, parenting). Over one-quarter of interventions (28%) targeted complex systems (e.g. schools, workplace), whilst other health topics were less frequently targeted (Table 3).

**Table 3 Health topics of trials**

| **Health topic** | **Count (%)** |
|---|---:|
| Lifestyle | 8,921 (52%) |
| Complex systems | 4,842 (28%) |
| Mental health | 3,100 (18%) |
| Substance misuse | 2,749 (16%) |
| NCD | 2,562 (15%) |
| Inequality | 2,029 (12%) |
| SRH | 1,794 (10%) |
| (Non-)Accidental incidents | 1,540 (9%) |
| Disability | 261 (2%) |
| Unspecified (general health promotion) | 1,033 (6%) |
| Uncoded* | 444 (3%) |

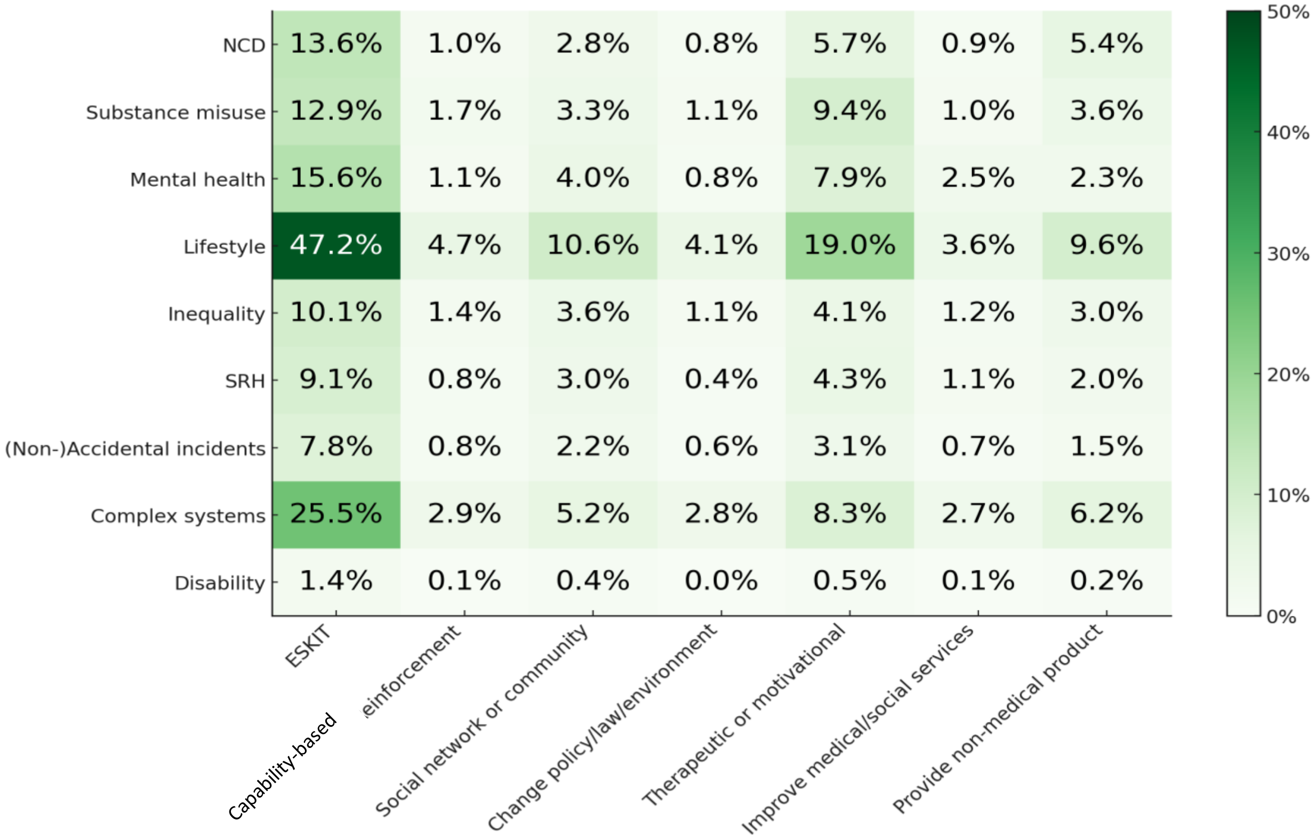
*GPT-4o was unable to code a small minority of records. NCD: non-communicable diseases; SRH: Sexual and reproductive health.

Distribution of health topics across intervention functions

Almost half of trials (47%) employed capability-based interventions for lifestyle behaviour, followed by a quarter (26%) for complex systems (e.g., educational or work environments) and 16% for mental health issues, respectively. Therapeutic or motivational functions were the second most commonly used, particularly for lifestyle behaviour (19%) and substance misuse (9%). Notably, interventions targeting policy, law or environmental changes were infrequent, with the highest proportion (4%) observed in trials about lifestyle behaviour. Disability-focused interventions were the least represented across all categories, consistently negligible (below 2%).

Figure 1 illustrates the distribution of intervention functions by various health topics among 16,775 trials with specific information (97% of the included trials). To more easily visualise and explore these trials, we created an interactive map (Appendix 2) with detailed records of the trials by intervention function and health topic.

**Figure 1 Intervention functions and health topics among trials**



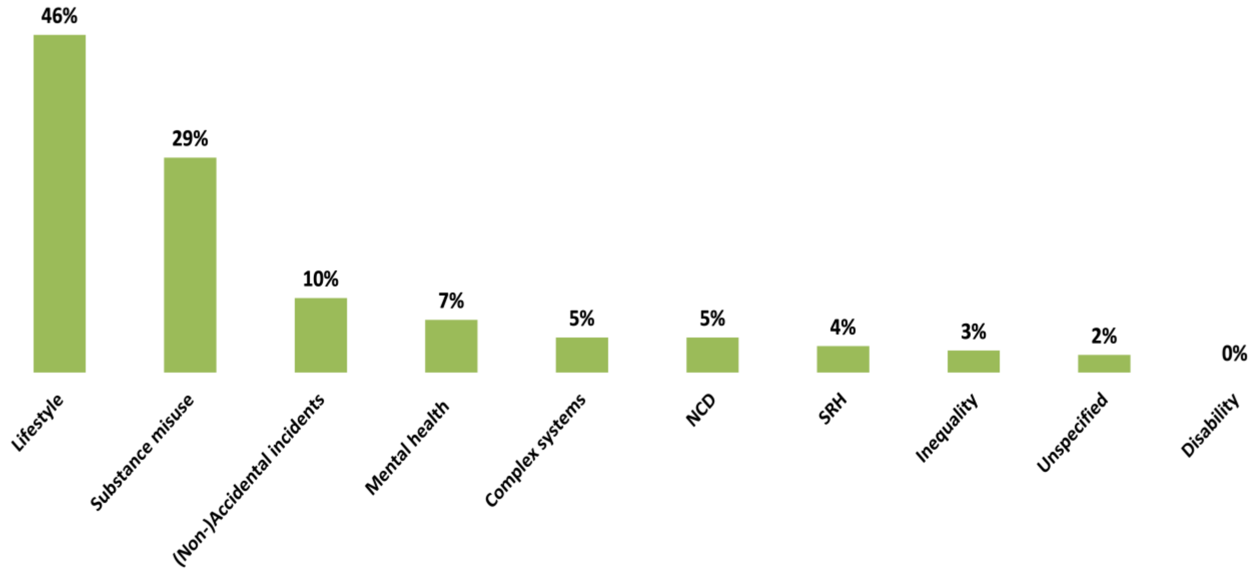

The sum of proportion in each column or row can exceed 100% as a trial may involve more than one intervention function and/or health topic.

NCD: non-communicable diseases; SRH: Sexual and reproductive health.

Characteristics of eligible reviews

Only a minority of reviews in DoPHER were policy-oriented and therefore included in our analysis (2%, 168 out of 7,560). As with the trials, lifestyle behaviours (e.g. healthy eating or hygiene) were the most common focus, of almost half the policy reviews (77 records, 46%). Substance misuse was the second most frequent (49 records, 29%), followed by (non)-accidental incidents (17 records, 10%). Figure 2 presents the distribution of health topics involved in policy reviews.

**Figure 2 Health topics in 168 policy reviews**




NCD: non-communicable diseases. SRH: sexual and reproductive health.

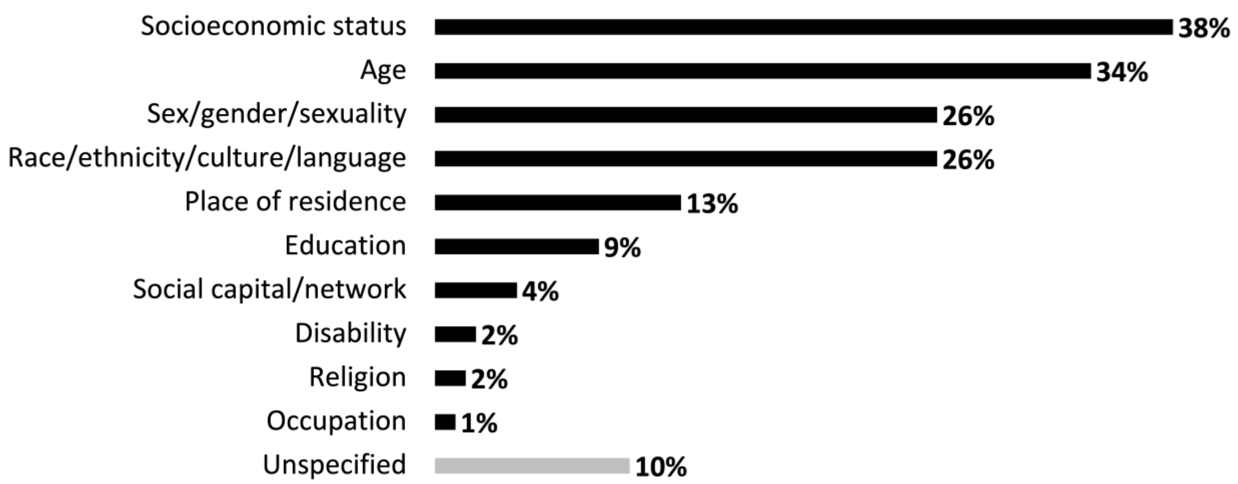
Addressing health disparities in reviews

Few reviews focused on health disparities (2%, 188 of 7,560). Figure 3 illustrates the distribution of PROGRESS-PLUS characteristics among the analysed reviews. Six per cent (12 records) explicitly mentioned the PROGRESS-PLUS framework in titles or abstracts, whereas 10% (19 records) only mentioned the term disparities (or inequalities or inequities) in titles or abstracts without specifying which type of disparities. The most commonly addressed characteristics were:

1. Socioeconomic status (38%): socially disadvantaged or deprived populations
2. Age (34%): children, adolescents or seniors

3. Sex/gender/sexuality (26%): women, girls or men who have sex with men
4. Race/ethnicity/culture/language (26%): minorities or people with culturally or linguistically diverse backgrounds

**Figure 3 PROGRESS-PLUS characteristics in 188 reviews addressing health disparities**




<u>Examining the performance of GPT-4o-assisted coding of health topics</u>

It was not possible to use GPT-4o to code the intervention functions used within the trials, since the agreement rate with human coding was poor (30%). However coding health topics was a simpler task, for which we used GPT-4o and used this as the basis for our assessment of its performance. The health topics of the 168 policy-oriented reviews were coded both manually and by GPT-4o (see the full details in Appendix 3). Despite being a less complex task than coding intervention functions, its performance was less than ideal, with GPT-4o correctly predicting health topic codes in less than half of all reviews.

## 6. Discussion

Our bibliometric analysis of public health research published from 1995 to 2024 revealed that most interventions combined capability and non-capability-based intervention functions (59%). However, over a quarter (29%) of trials evaluated interventions that only addressed capability (e.g. education, information provision). While capability-based intervention functions remain crucial, their dominance across three decades raises questions about the breadth of interventions being evaluated. The fact that a majority of studies target capability, or therapeutic/motivational intervention functions, suggests that evaluations continue to focus on individual, cognitive factors affecting health and health behaviour. This is despite the fact that knowledge provision alone is rarely sufficient to change behaviours and that broader determinants of health are important to consider. It is noteworthy that few (7%) trials in TRoPHI aimed to change policies, laws or aspects of the environment. As the UK's public health strategy has increasingly recognised the importance of addressing wider determinants of health, and redoubling efforts to invest in the prevention of ill-health,[16,17] evaluations of interventions targeting determinants beyond individual-level, cognitive factors are warranted.

Our findings revealed that equity-oriented reviews remain scarce. Almost half of the equity-oriented reviews examined socioeconomic status, which has been the cornerstone of the UK’s policies and guidance on tackling health disparity.

Using GPT-4o to code titles and abstracts was challenging; more so for complex coding like intervention functions. Although it was more accurate for simple coding tasks (health topics), it nevertheless achieved agreement with human coding for less than half the records assessed. Particular challenges were for those health topics for which only a few reviews were identified, since this provided only limited data inputs to optimise AI’s predictive power.

## 7. Policy implications

If DHSC wishes to enhance the effectiveness of public health interventions, it may consider encouraging a more balanced approach in implementation and evaluation. While capability-based interventions (e.g., education, information provision or knowledge acquisition) remain important, combining these with non-capability-based approaches could yield more comprehensive results. DHSC might promote research addressing multiple health determinants simultaneously, including structural and environmental factors.

If DHSC aims to ensure research priorities reflect current policy challenges, it might develop guidance on addressing health inequalities as well as complex system interventions. DHSC could also encourage public health researchers to adopt standardised frameworks to establish evidence on the consequences of under-explored health inequities.

## 8. Conclusions

Our bibliometric analysis of 30 years of public health trials and reviews reveals a predominance of capability-based interventions targeting lifestyle behaviours, highlighting potential gaps in implementing and evaluating interventions targeting broader determinants of health. Equity-oriented reviews were scarce, with limited application of comprehensive frameworks. The evaluation of GPT-4o in health topic identification demonstrates that, although there is potential of AI-assisted coding for evidence synthesis, challenges remain in optimising performance, particularly for topics with limited data.

## 9. References


1. Dahlgren G, Whitehead M. The Dahlgren-Whitehead model of health determinants: 30 years on and still chasing rainbows. Public Health. 2021 Oct 1;199:20–4.
2. Michie S, van Stralen M, West R. The Behaviour Change Wheel: a new method for characterising and designing behaviour change interventions. Implementation Science. 2011 Apr 23;6(1):1–12.
3. World Health Organization. Health promotion [Internet]. 2024 [cited 2024 Sep 16]. Available from: https://www.who.int/westernpacific/about/how-we-work/programmes/health-promotion
4. Golden SD, Earp JAL. Social Ecological approaches to individuals and their contexts: twenty years of health education & behavior health promotion interventions. Health Education and Behavior. 2012 Jun 20;39(3):364–72.
5. Walters R, Leslie SJ, Polson R, Cusack T, Gorely T. Establishing the efficacy of interventions to improve health literacy and health behaviours: a systematic review. BMC Public Health. 2020 Dec 30;20(1):1040.
6. van Achterberg T, Huisman-de Waal GGJ, Ketelaar NABM, Oostendorp RA, Jacobs JE, Wollersheim HCH. How to promote healthy behaviours in patients? An overview of

evidence for behaviour change techniques. Health Promot Int. 2011 Jun 1;26(2):148–62.

7. Trials Register of Promoting Health Interventions (TRoPHI) [Internet]. EPPI Centre, UCL Social Research Institute, University College London. London; [cited 2024 Jun 16]. Available from: https://eppi.ioe.ac.uk/webdatabases4/Intro.aspx?ID=12
8. Database of promoting health effectiveness reviews (DoPHER) [Internet]. EPPI Centre, UCL Social Research Institute, University College London. London; [cited 2024 Jun 16]. Available from: https://eppi.ioe.ac.uk/webdatabases4/Intro.aspx?ID=9
9. OpenAI. Models [Internet]. 2024 [cited 2024 Sep 30]. Available from: https://platform.openai.com/docs/models
10. Stansfield C, Thomas J. Applying automation to maintain research registers in health promotion. Journal of EAHIL. 2024 Jun 27;20(2):26–9.
11. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics. 2019 Feb 1;21(2):167–79.
12. Dooris M. Expert voices for change: bridging the silos—towards healthy and sustainable settings for the 21st century. Health Place. 2013 Mar 1;20:39–50.
13. Michie S, Atkins L, West R. The Behaviour Change Wheel: a guide to designing interventions. 2014.
14. McLeroy KR, Bibeau D, Steckler A, Karen Glanz D, Kenneth McLeroy MR. An ecological perspective on health promotion programs. Health Educ Q. 1988;15(4):351–77.
15. O’Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol. 2014 Jan;67(1):56–64.
16. Chapter 6: wider determinants of health [Internet]. [cited 2024 Sep 30]. Available from: https://www.gov.uk/government/publications/health-profile-for-england-2018/chapter-6-wider-determinants-of-health
17. Keir Starmer’s big idea: the preventative state [Internet]. [cited 2024 Sep 30]. Available from: https://www.newstatesman.com/politics/labour/2024/09/keir-starmers-big-idea-the-preventative-state
18. Heydarian M, Doyle TE, Samavi R. MLCM: Multi-Label Confusion Matrix. IEEE Access. 2022;10:19083–95.
19. Zhang ML, Zhou ZH. A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng. 2014;26(8):1819–37.

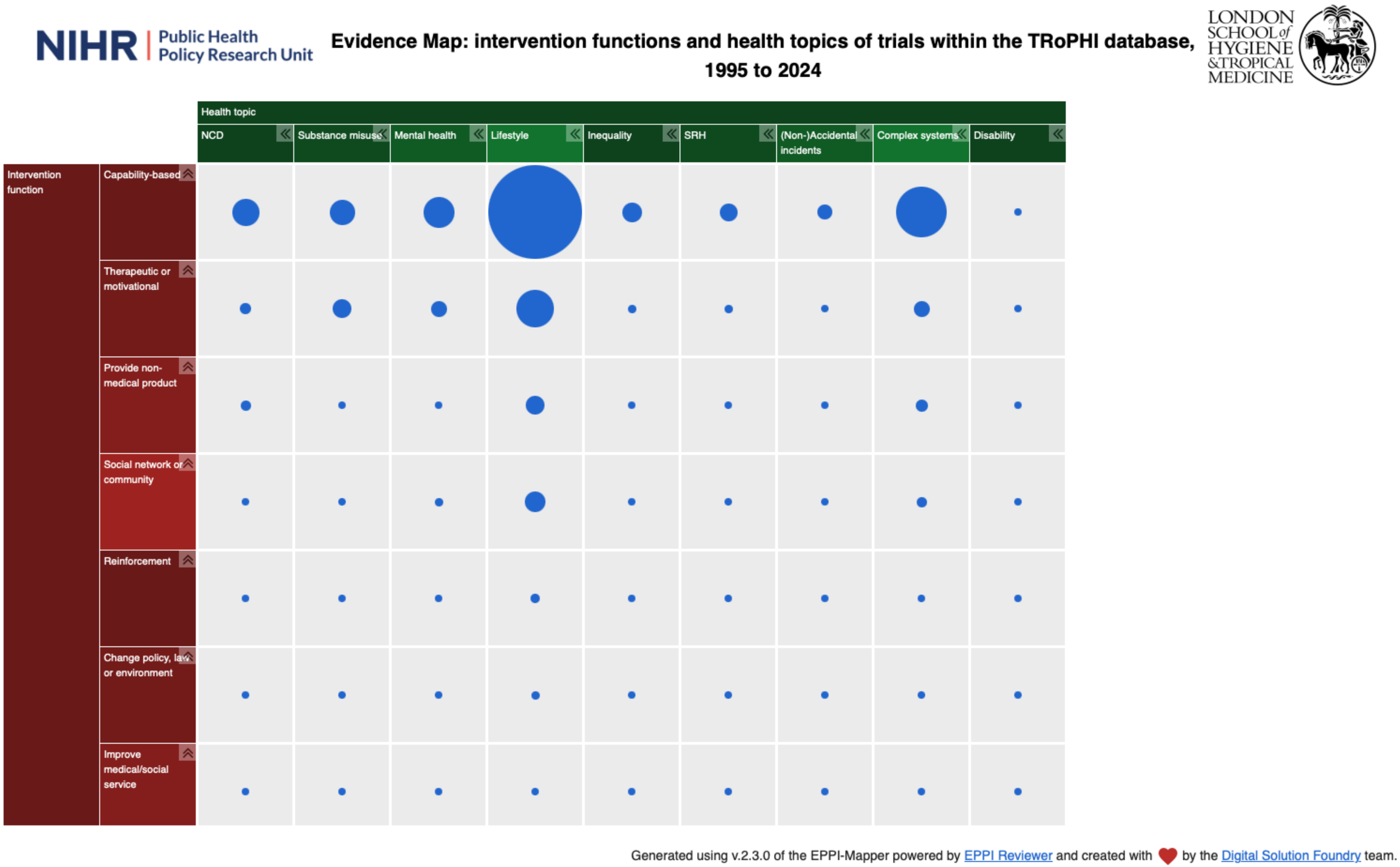
## 10. Appendices

### Appendix 1. Text search strategy for classifying intervention function in 17,279 trials

| Category | Search strategy |
|---|---|
| **Capability-based** | |
| Education | Education or Educational or educate or curricula or curriculum |
| Knowledge | Knowledge or awareness or learn |
| Information provision | Information or leaflet or manual or brochure or lecture or advice or advise or instruction or communication or campaign or advocacy or guideline or guidance or material |
| Skill | Skill or ability or competency or proficiency |
| Training | Training or workshop or course or session |
| Self-help | Self-help or self-controlled or self-paced or self-efficacy or self-effacing |
| Physical activity | Sport or exercise or fitness or "physical activity" or "physical program" |
| Memory | Reminder or text or message or notification or SMS or memory or memorise |
| **Reinforcement** | Incentive or incentivise or disincentive or disincentivise or reward or game or gamification or punish or punishment or contingent or contingencies or reinforce or reinforcement or voucher or coupon or cash or "cash transfer" or "money transfer" or nudge |
| **Social network or Community mobilisation** | Network or "social network" or "social support" or "social norm" or "group support" or "peer support" or "peer" or "peer-led" or "peer-oriented" or "peer educator" or "community support" or "community network" or "community engagement" or "group" or "group identity" or "group-based" or "role-model" or "community-based" or "community-oriented " or "home visit" |
| **Change policy, law or environments** | "Law amendment" or "law change" or legislation<br>"Policy change" or "Policy reform" or reform<br>"practice change" or "regulation change" or "regulation amendment"<br>pricing or levy or tax or taxed or label or labelling<br>"organisational culture" or "organisational change" or "leadership change" or "environmental change" or "environmental improvement" or "urban planning" or infrastructure |

| | |
|---|---|
| **Therapeutic or motivational** | Mindfulness or meditation or yoga or conversation or cognitive interview or motivational or coach or consultative or consultation or "brief intervention" or medication or medicine or treatment or therapy or therapeutic or counselling or counsel or feedback |
| **Improve medical or social services** | "Health service" or "social service" or "service improvement" or "service optimisation" or "increase access" or "increase availability" or "medical service" or "service delivery" or "service performance" or "service satisfaction" or satisfaction or "service upgrade" or "service change" or "service provision" or "better service" or "service quality" or "quality improvement" or "quality assessment" |
| **Provide non-medical product** | "Screen (non-SBIRT)" or provision or supply or mask or "the use of" or product or supplement |

## Appendix 2. Interactive evidence map



## Appendix 3. Assessing the performance of AI coding

### Methods

We employed the Multi-Label Confusion Matrix (MLCM)[18] approach to assess the performance of GPT-4o coding because all records could be simultaneously coded to more than one health topic (e.g., a review could focus on both substance misuse and sexual/reproductive health). A human coding of 168 policy reviews in DoPHER (conducted by IYC) was considered as the reference group. We applied three key indicators to a comprehensive assessment of GPT-4o's performance:[19]

1. Exact Match Ratio: the proportion of instances where all categories were correctly predicted by AI. This ratio offered a stringent evaluation of AI models' performance as it was harder to predict all categories correctly in multi-label classifications).
2. Hamming Loss: the fraction of misclassified labels (categories), with values ranging from 0 to 1 (lower indicating fewer errors).
3. Weighted F1-score (the harmonic mean of positive predictive value and sensitivity): a balanced measure that reflected an overall predictive power of AI models, with values ranging from 0 to 1 (higher indicating stronger predictive power).

### Results

An Exact Match Ratio of 0.447 indicated that GPT-4o correctly predicted all health topics among 44.7% of all coding attempts, A Hamming Loss of 0.102 meant that GPT-4o erroneously classified 10% of codes across all health topics. Table A1 summarises the key MLCM indicators compared the performance of GPT-4o with human coding on health topics. See Table A2 and Table A3 for more details of our MLCM analyses.

## Table A1. GPT-4o performance in coding health topics

| **Indicator** (see Methods above for operational definitions) | **Value** (Between 0 and 1) | **Meaning** (Human coding as reference) |
|---|---|---|
| Exact Match Ratio* | 0.447 | GPT-4o correctly coded health topics in 44.7% of all coding attempts |
| Hamming Loss** | 0.102 | Considering GPT-4o's predictions across all health topics, 10.2% were wrongly predicted. |
| Weighted-average F1-Score*** | 0.450 | After considering the frequency of each health topic, the overall predictive power of GPT-4o (correctly identifying and classifying health topics) was 45%. |

*Exact Match Ratio = [Number of samples where all labels match (N.B.: grids in green colour in Table 1)] / Total samples = 152 / 340 = 0.4471

**Hamming Loss = (Total FP + Total FN) / (Total predictions * Number of labels) = (191 + 188) / (340 * 11) = 0.1019

***Weighted-average F1-Score = [(2 * (Precision * Recall) / (Precision + Recall)], weighted by considering imbalance) =0.4506

**Table A2. Performance of GPT-4o coding versus human coding using 168 reviews in DoPHER**

| GPT-4o vs. Human | NCD | Mental health | Substance misuse | Lifestyle | Inequality | SRH | Incident | Complex system | Disability | Unspecified | None of these |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NCD | 6 | 0 | 7 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mental health | 0 | 12 | 1 | 4 | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| Substance misuse | 2 | 0 | 31 | 7 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Lifestyle | 6 | 1 | 5 | 74 | 1 | 3 | 2 | 4 | 0 | 1 | 0 |
| Inequality | 1 | 1 | 8 | 23 | 4 | 2 | 3 | 1 | 0 | 0 | 0 |
| SRH | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| Incident | 0 | 5 | 7 | 3 | 1 | 0 | 14 | 0 | 0 | 0 | 0 |
| Complex system | 3 | 7 | 15 | 24 | 2 | 4 | 2 | 6 | 0 | 3 | 0 |
| Disability | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Unspecified | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 0 |
| None of these | 0 | 0 | 9 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

DoPHER: the Database of Promoting Health Effectiveness Reviews; NCD: non-communicable diseases; SRH: sexual and reproductive health

### Table A3. Detailed matrices from each sub-category of our 11-label classification matrix

| Category | True positive | False negative | False positive | True negative | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|---|---|---|---|
| NCD | 6 | 14 | 14 | 306 | 0.300 | 0.300 | 0.300 | 0.918 |
| Mental health | 12 | 13 | 14 | 301 | 0.462 | 0.480 | 0.471 | 0.921 |
| Substance misuse | 31 | 10 | 53 | 246 | 0.369 | 0.756 | 0.496 | 0.815 |
| Lifestyle | 74 | 23 | 69 | 174 | 0.518 | 0.763 | 0.617 | 0.729 |
| Inequality | 4 | 39 | 4 | 293 | 0.500 | 0.093 | 0.157 | 0.874 |
| SRH | 2 | 1 | 10 | 327 | 0.167 | 0.667 | 0.267 | 0.968 |
| (Non-)Accidental incidents | 14 | 16 | 11 | 299 | 0.560 | 0.467 | 0.509 | 0.921 |
| Complex systems | 6 | 60 | 9 | 265 | 0.400 | 0.091 | 0.148 | 0.797 |
| Disability | 0 | 1 | 1 | 338 | 0.000 | 0.000 | 0.000 | 0.994 |
| Unspecified | 3 | 1 | 6 | 330 | 0.333 | 0.750 | 0.462 | 0.979 |
| None of the above | 0 | 10 | 0 | 330 | 0.000 | 0.000 | 0.000 | 0.971 |

NCD: Non-communicable disease; SRH: Sexual and reproductive health

Precision: [(True Positive) / (True Positive + False Positive)]

Recall: [(True Positive) / (True Positive + False Negative)]

Accuracy: [(True Positive + True Negative) / (all coding attempts per sub-category)]