Supplementary Information

A static research enterprise decouples from changes in the burden of disease

Leo Schmallenbach, Maximilian Bley, Till Bärnighausen, Cassidy Sugimoto, Carolin Lerchenmüller, Marc J. Lerchenmueller

Table of contents

Section 1. Measures of imbalance

The Kullback-Leibler divergence (KLD), which is closely related to relative entropy, is a non-symmetric measure of the difference between two probability distributions *p(x)* and *q(x)*. The KLD is given by:

$$KLD(p(x)||q(x)) = \sum_{x \in X} p(x) \ln\left(\frac{p(x)}{q(x)}\right)$$

where *p(x)* represents the reference distribution, in our case the discrete distribution of DALYs per disease *x* (in percent), and *q(x)* represents the discrete distribution of research articles per disease *x* (in percent). The sum of the individual divergences of research articles to DALYs per disease *x* across all 16 diseases in the set *X* of level 2 diseases from the Global Burden of Disease (GBD) database yields the KLD as our key annual divergence measure. Again, we focus on the KLD mostly for its asymmetric property, considering the distribution of disease burden *p(x)* as the reference distribution providing normative orientation for the research distribution *q(x).*

In addition to the KLD, we calculate the Population Stability Index (PSI), another measure of distributional divergence that does not require a reference distribution. Unlike the KLD, which assumes that the distribution of research publications should follow the distribution of DALYs, the PSI places equal weight on deviations in both directions—comparing research to DALYs and DALYs to research. This approach provides a non-directional perspective on the alignment between research and disease burden over time (*Supplementary Figure 1*).

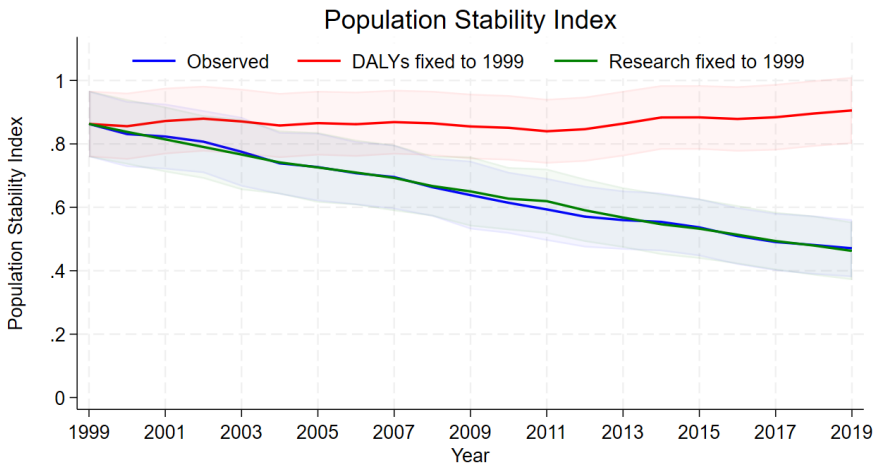
We also calculate the Hellinger distance, which ranges from 0 to 1, where 0 indicates perfect alignment and 1 represents maximal divergence. Unlike the KLD, which is asymmetric and relies on a reference distribution, or the PSI, which evaluates deviation in both directions equally, the Hellinger distance is inherently symmetric and emphasizes proportional differences between distributions. It is computed as the square root of the sum of squared differences between the square roots of corresponding probabilities in each distribution. This method provides an intuitive way to assess alignment by focusing on how close the distributions are in terms of their overall structure, yet it softens the normative evaluation of the degree to which research follows disease burden over time.

As a third measure, we calculate the Jensen-Shannon divergence as a mixture assessment that uses a combination of the disease and research distribution as a comparator distribution. It is derived from the Kullback-Leibler Divergence (KLD) but unlike KLD, it is symmetric and thus also does not consider DALYs to provide a normative orientation for the distribution of research.

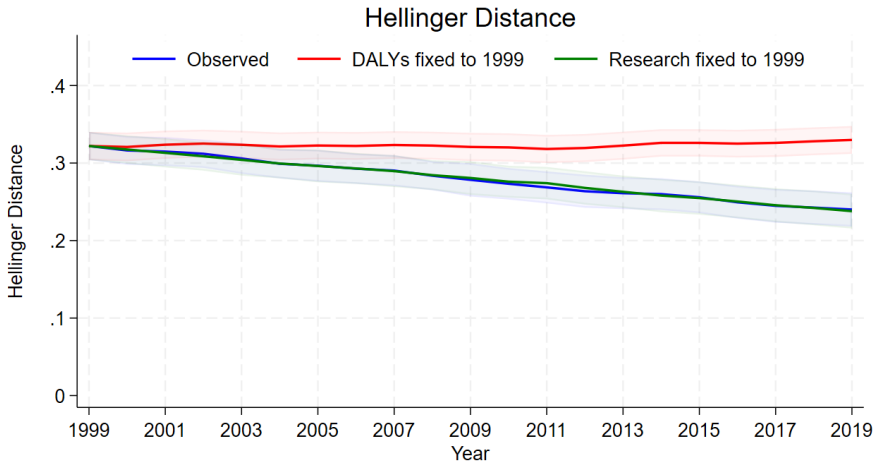
Across all four metrics – Kullback-Leibler divergence, Population Stability Index, Hellinger Distance, and Jensen-Shannon divergence – we observe a decline in the imbalance between research and disease over the past two decades. Also consistent across these measures is the fact, that the change in the distribution of DALYs accounts for the reduction in imbalance (indicated by the space between blue and red lines in Supplementray Figure 1), not the change in the distribution of research (indicated by the overlapping blue and green lines).

As an additional robustness test to the simulations presented in Supplementary Figure 1, we also provide descriptive visuals of the distribution of research across diseases and DALYs across diseases for the past two decades. Supplementary Figure 2 shows, that the distribution of research remained virtually unchanged across diseases for the past 20 years while the distribution of DALYs have changed over time.

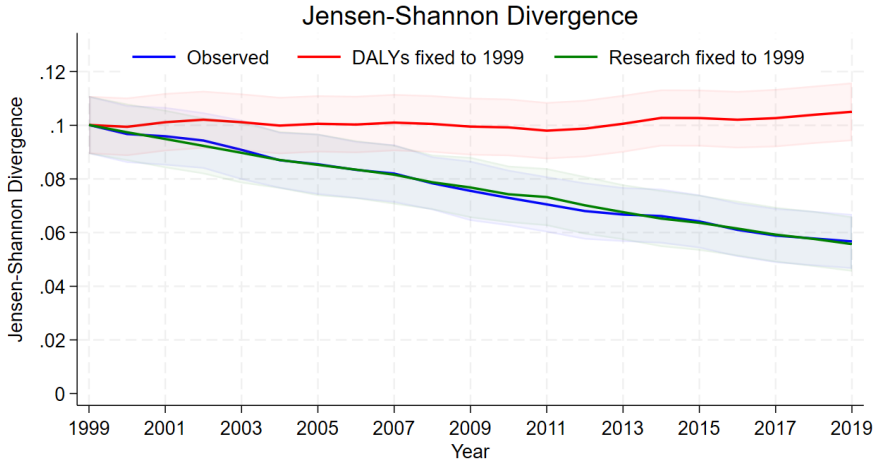
Supplementary Figure 1. Imbalance of research and disease burden across different measures of imbalance



Note: Bootstrapped 95%-confidence Intervals based on DALY estimates for 16 level 2 disease causes

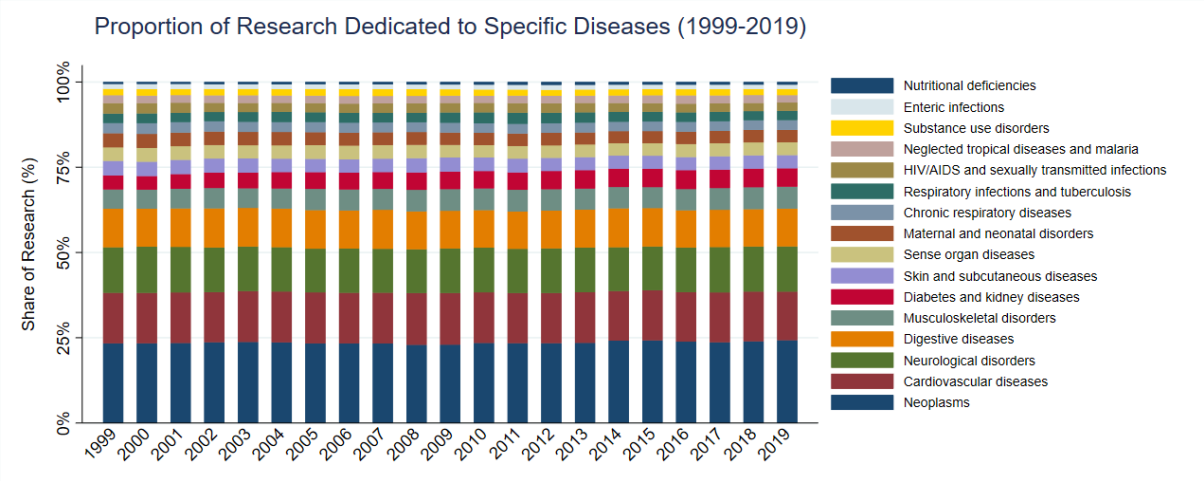


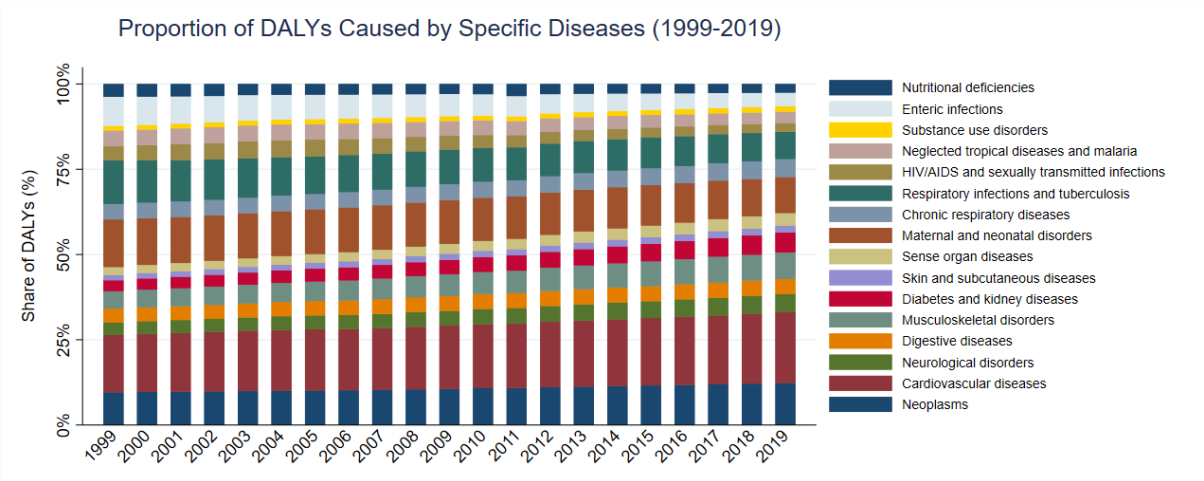
Note: Bootstrapped 95%-confidence Intervals based on DALY estimates for 16 level 2 disease causes



Note: Bootstrapped 95%-confidence Intervals based on DALY estimates for 16 level 2 disease causes

Supplementary Figure 2. Distribution of disease-specific research and disease burden

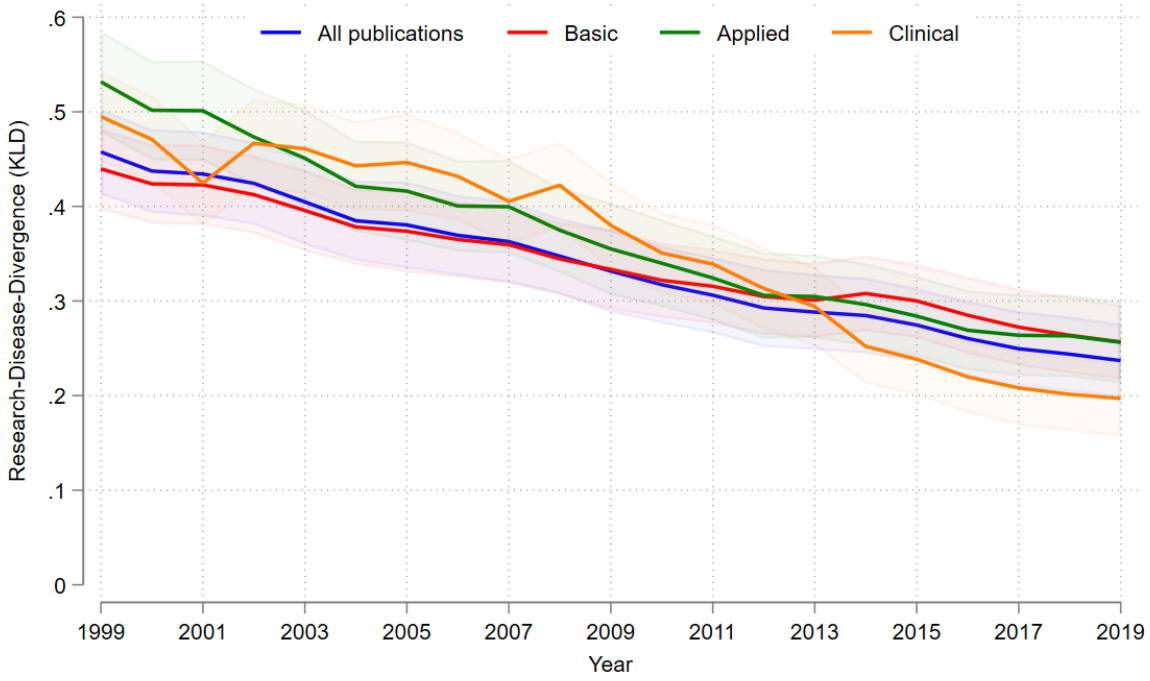




Section 2. Imbalance for different types of research

To test whether the research distribution remained static for different types of research over the past 20 years, we stratified our sample of research articles into (a) basic research (red line in Supplementary Figure 3), (b) applied research (green line in Supplementary Figure 3), (c) clinical research (orange line in Supplementary Figure 3). We also reproduce the estimate for research-disease divergence (KLD) from the main manuscript (blue line in Supplementary Figure 3). Across these different samples of research, we do not observe a statistically distinguishable reduction in imbalance.

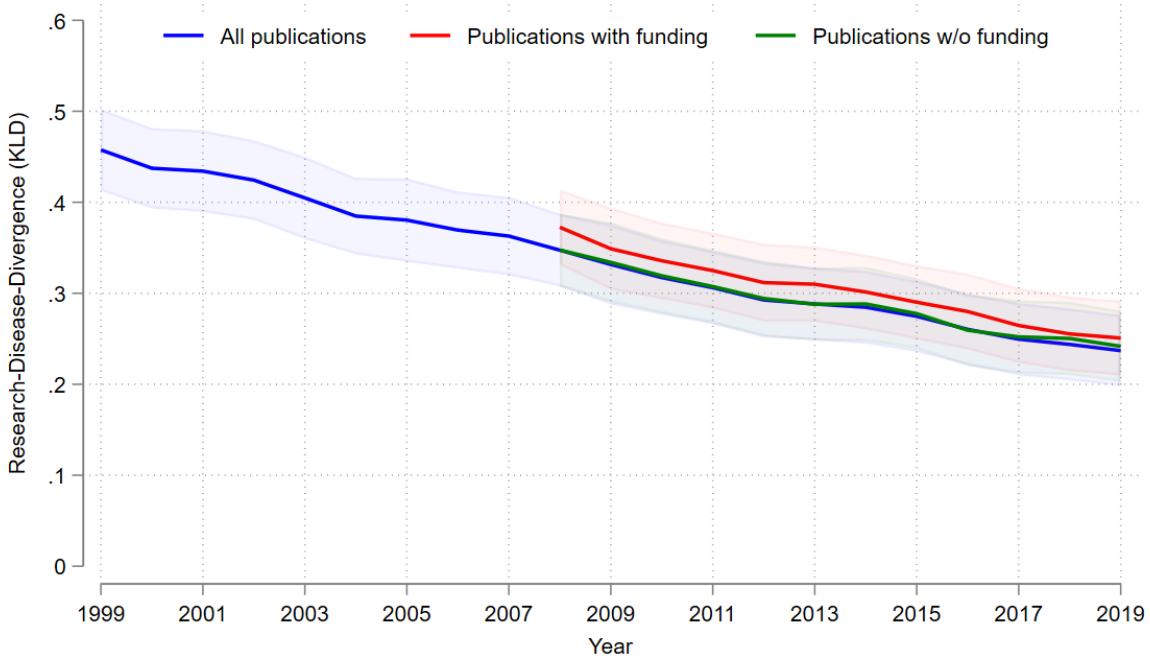
Supplementary Figure 3. Imbalance of research and disease burden across different types of research



Note: Bootstrapped 95%-confidence Intervals based on DALY estimates for 16 level 2 disease causes.

We also test whether the reduction in imbalance is correlated with research funding, stratifying research articles according to whether the article acknowledges funding versus not. Consistent with our main finding – changes in research did essentially not contribute to a reduction in imbalance – we also obtain no statistically significant difference between research that acknowledges funding (red line in Supplementary Figure 4) versus research that does not acknowledge funding (green line in Supplementary Figure 4).

Supplementary Figure 4. Imbalance of research and disease burden differentiating research that acknowledges funding versus not

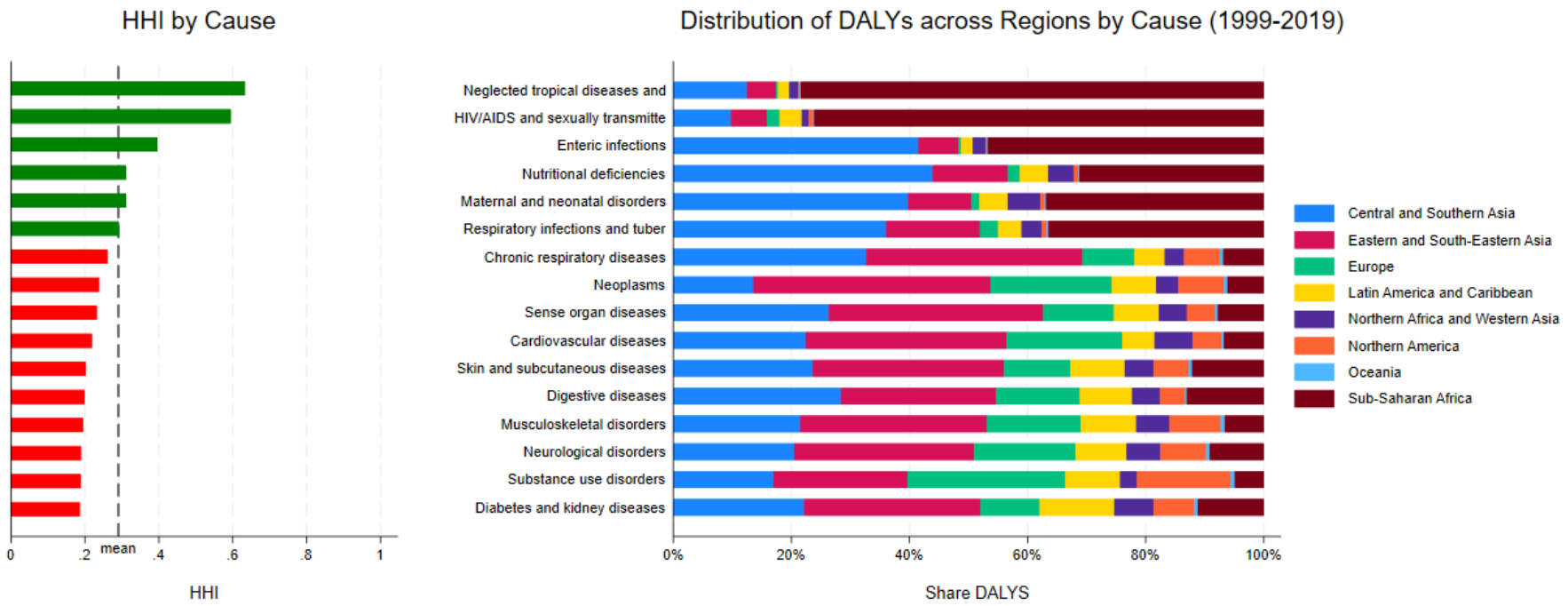


Note: Bootstrapped 95%-confidence Intervals based on DALY estimates for 16 level 2 disease causes.
Funding data available as of 2008.

Section 3. Geography of research and disease burden and future scenarios

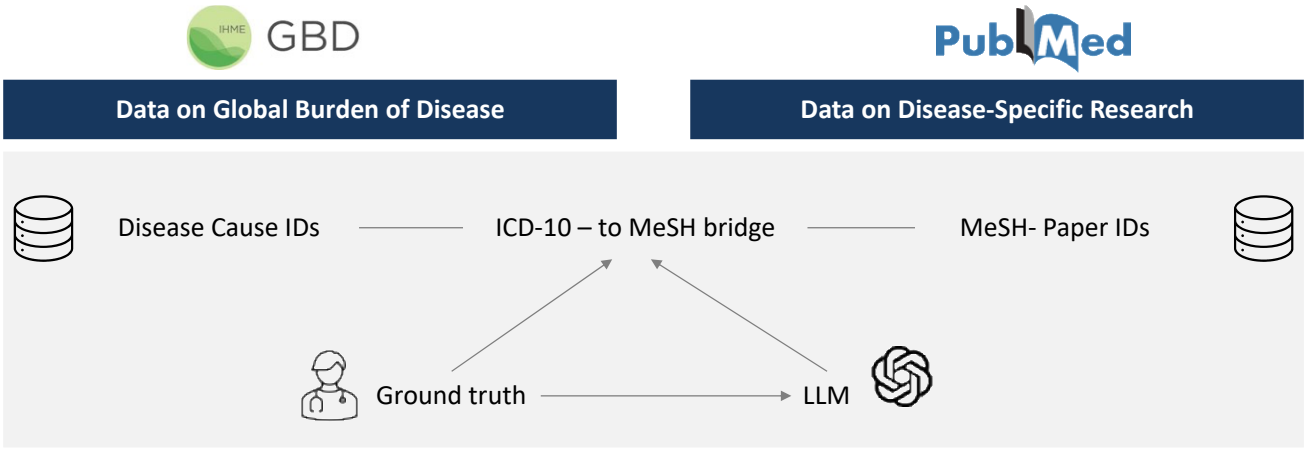
To assess the geographic distribution of diseases, we first calculated the share of DALYs per world region for each Level 2 disease cause across the eight world regions defined by the United Nations. The left panel of Supplementary Figure 5 ranks diseases in descending order based on their contribution to reducing the research-disease burden imbalance over the past 20 years, with the respective disease burden stratified by world region. The right panel of Supplementary Figure 5 presents the Herfindahl-Hirschman Index (HHI), a widely used measure of concentration. The HHI is calculated by squaring the DALY share of each region for a given Level 2 disease and summing the resulting values, where higher scores indicate a more regionally concentrated disease. The green shading of the bars in this panel corresponds to the green shading in Figure 4 of the main text, indicating diseases that contributed to reducing the imbalance. These diseases are all locally concentrated, as their HHI exceeds the average HHI in the sample, and, as noted in the main text, they are all communicable diseases. In contrast, the red-shaded diseases, which are mostly non-communicable, tend to be more globally distributed.

Supplementary Figure 5. Geographic concentration of diseases



Section 4. Supplementary Methods

Supplementary Table 1. Gold standard, ICD bridge, and LLM bridge for linking disease causes to MeSH terms



Compared to experts[1]	LLM	ICD-10
Accuracy	86.1	50.5

- 9.7M cause-papers
- ~50M location-specific authorships

Compared to experts[2]	LLM	ICD-10
Accuracy	94.9	67.0

1. 467 CVD MeSH term evaluations

Supplementary Table 2. LLM custom prompt

```
Model: ChatGPT, gpt-4o

role = "You are a world-class algorithm specialized in medical
terminology analysis."

prompt = Your task is to determine if a MeSH term unambiguously
corresponds to a global disease burden cause.

Does the specific MeSH term "{mesh_term}" unambiguously belong to
the global disease burden cause "{disease_cause}"?

For example, "Brain Ischemia" is a good MeSH term for the disease
"Stroke," but "Lung Cancer" is not for "Diabetes."

Answer with Yes or No and provide a confidence score between 0 (low)
and 100 (high) for your judgment. Provide your response and the
confidence score separated by a comma. Do not explain your answer.
```

Supplementary Table 3. Triangulated performance metrics

Evaluation ChatGPT vs. ICD-10 - Cardiovascular Diseases - at the Paper Level, Publications 1999-2021

ChatGPT

		Annotated	Not Annotated	
Expert	Annotated	1,324,710	34,589	**1,359,299**
	Not Annotated	46,216	182,625	**228,841**
		1,370,926	**217,214**	**1,588,140**

ICD-10

		Annotated	Not Annotated	
Expert	Annotated	867,638	491,661	**1,359,299**
	Not Annotated	32,735	196,106	**228,841**
		900,373	**687,767**	**1,588,140**

Compared to Experts	**ChatGPT**	**ICD-10**
Accuracy	94.9%	67.0%
Precision	96.6%	96.4%
Recall	97.5%	63.8%

Supplementary Table 4. LLM versus ICD recall analysis

		ChatGPT		ICD	
Journal Cause	**# articles**	**# articles with same cause**	**% articles with same cause**	**# articles with same cause**	**% articles with same cause**
Cardiovascular diseases	5053	4843	95.84%	3482	68.91%
Chronic respiratory diseases	1323	1306	98.72%	1278	96.60%
Diabetes and kidney diseases	4645	4233	91.13%	3538	76.17%
Digestive diseases	1194	1124	94.14%	670	56.11%
HIV/AIDS and sexually transmitted infections	8793	8319	94.61%	7680	87.34%
Musculoskeletal disorders	7485	6343	84.74%	4147	55.40%
Neglected tropical diseases and malaria	5876	5738	97.65%	5742	97.72%
Neoplasms	10294	10092	98.04%	4429	43.03%
Neurological disorders	19754	19043	96.40%	15873	80.35%
Total	**64417**	**61041**	**94.76%**	**46839**	**72.71%**

Supplementary Table 5. Sample creation process