Supplementary Information

Medical research responds better to disease burden and health shocks, yet global disparities persist

Hongyu Zhou[*]

[*]Renmin University of China, China
University of Cambridge, United Kingdom
University of Antwerp, Belgium

Prashant Garg[†]

[†]Imperial College London, United Kingdom

Thiemo Fetzer[‡]

[‡]University of Warwick, United Kingdom; University of Bonn, Germany

A Details on Bibliographic Dataset Construction

A.1 Journal Set Identification

To identify journals relevant to this study, we employed the 2023 edition of the *Journal Citation Reports* (JCR), provided by Clarivate JCR. JCR is a widely used journal evaluation tool built on the Web of Science Core Collection, reporting citation-based metrics such as the Journal Impact Factor and assigning each journal to one or more subject categories. These subject categories serve as proxies for disciplinary orientation and are commonly used in bibliometric and research evaluation studies. We linked the journals retrieved from JCR 2023 to OpenAlex using both ISSN and eISSN, to further obtain article-level bibliographic data.

- We identified **57 medicine-related categories** in JCR. Linking these journals to OpenAlex yielded **6,725 journals** with unique OpenAlex identifiers.
- We excluded journals with the word *"Review"* in the title to omit outlets primarily positioned to publish review articles rather than empirical studies. This reduced the set to **5,861 journals**.
- We focused on three broad categories that capture general medical research without explicit disease specialization:
  - *MEDICINE, GENERAL & INTERNAL*
  - *MEDICINE, RESEARCH & EXPERIMENTAL*
  - *PHARMACOLOGY & PHARMACY*

This refinement gave us **524 journals**, covering **1,065,683 papers** indexed in OpenAlex, restricted to documents of type "article" with at least one citation.

A.2 Article Corpus Filtering

Starting from 1,065,683 papers, we applied a set of filters to ensure coverage of relevant, attributable, and analyzable research outputs, while also producing a reasonable size of data for our LLM-based retrieval pipeline.

- **Domain filter:** Papers classified in OpenAlex as only belonging to the *Health Sciences* domain were retained (**741,264 papers; 69.6%**).
- **Authorship filter:** Papers with valid information on the country or territories of authors' affiliations were retained (**871,934 papers; 81.8%**).
- **Language filter:** Only papers written in English were retained (**1,050,170 papers; 98.5%**).
- **Combined filters:** Applying all three filters together produced a refined corpus of **603,447 papers (56.6%)**.

A.3 Geographic Context Filtering

To study the geographical dimension of medical research, we further filtered papers to identify those with potential geographical context in their titles or abstracts. Two complementary approaches were used:

- **GeoText method:** Applied the Python `geotext` library (https://github.com/elyase/geotext) to extract city and country mentions. This identified **202,790 papers (33.6%)**.
- **NER method:** Used a named entity recognition model (`dslim/distilbert-NER`) to extract *LOC* and *ORG* entities as potential location markers. This identified **309,950 papers (51.4%)**.
- **Combined filter:** Retaining papers identified by either method gave a total of **349,737 papers (58.0%)**.

A.4 LLM Retrieval and Linkage

We then processed the 349,737 papers with our LLM-based retrieval pipeline to enrich them with standardized biomedical and contextual information. This step successfully linked **308,135 papers (88.1%)** to at least one Medical Subject Heading (MeSH). Within this corpus, we report the following information on variables used in the study:

- **GBD categories:** Among these, **266,532 papers (86.5%)** were further linked to at least one level 2 Global Burden of Disease (GBD) category.
- **Funder information: 32,057 papers (10.4%)** had funder information disclosed and retrieved from OpenAlex.
- **Geographical context (LLM retrieval):** The LLM pipeline identified geographical context for **169,262 papers (54.9%)**.

Final dataset: The resulting curated dataset consists of **308,135 papers**, enriched with relevant disease categories, authorship information, funder information, and geographical context.

B Details on LLM Retrieval of Academic Papers

Our empirical design hinges on two ingredients for every article in our corpus:

1. *paper–data metadata*, which record whether the study uses primary or secondary data, the unit of analysis, temporal coverage, geographic scope, and ownership; and
2. *paper–disease links*, which anchor every disease mention to a controlled biomedical vocabulary so that research effort can be compared with national disease burdens.

We obtain both with a single large-language-model (LLM) call per paper. The model (gpt-4o-mini-2024-07-18) receives the article's title and abstract plus a structured *system prompt* that instructs it to respond with a JSON object. We use the default model hyperparameters.

Strict schema validation is enforced via OpenAI's structred output format, minimising hallucination and guaranteeing parseability.

B.1 LLM Prompt for Paper–Data Metadata Extraction

The prompt casts the model as an expert analyst of biomedical papers, tells it to ignore everything except data-related content, and spells out the required JSON keys (with explicit fall-backs such as NA or empty arrays) to guarantee parseable output.

System instructions We provide the following instructions to the model for each call:

```
You are an expert assistant analyzing research papers.
Task
- Given a paper's title and abstract, extract the data information mentioned in the
text.
- Return a JSON object with a single top-level field "data" containing the subfields
listed below.
Subfields
- uses_data (boolean): true if the paper uses any data, false otherwise.
- unit_of_analysis (array): e.g. individual, group, country.
- unit_of_analysis_details (string): extra detail or "NA".
- temporal_context (string): time period or "NA".
- start_year, end_year (arrays of integers): empty if absent.
- geographical_context_countries (array): ISO-3 codes or empty.
- ownership (array): choose from private company, public sector entity, academic
institution, non-profit organization, individual researcher, community-generated,
open source, other.
-ownership_details (array): extra owner details or empty.
Notes
- Extract only data-related information; ignore all other content.
- If a field is missing, use the specified default ("NA" or empty array).
-Produce only valid JSON matching the strict schema and add no extra keys.
```

User prompt For each paper, the following message is sent, where <TITLE> and <ABSTRACT> are the raw metadata:

```
Here is the title of the paper:
<TITLE>
Here is the abstract of the paper:
<ABSTRACT>
Extract the data information as specified.
```

B.2 Embedding-based Normalisation to MeSH

All disease strings appearing in the extracted metadata are converted into 1,024-dimensional embeddings using `text-embedding-3-large`. We pre-compute embeddings for the 30,836 preferred MeSH descriptor terms (including scope notes) and assign each disease the descriptor with the highest cosine similarity.

B.3 Limitations

Three caveats merit mention. First, the model works only with titles and abstracts: full-text mining could reveal additional entities but is possible only for open-access papers. Second, assigning each disease mention to a single MeSH descriptor ignores genuine polysemy. Third, information on temporal coverage, comprehensive disease attribution and data ownership is often missing; we therefore carry a dedicated missingness flag into all regressions. None of these constraints undermines the pipeline's ability to deliver a scalable, reproducible bridge from academic papers to policy-relevant, disease-coded metadata.

C Details on LLM Retrieval and Structuring of WHO Disease-Outbreak News

We download all Disease Outbreak News (DON) items published on `www.who.int/emergencies/disease-outbreak-news`. The snapshot used here covers alerts issued between 1996-2025 and yields 3,134 HTML pages. After stripping boiler-plate, we retain four fields per page: the alert title, publication date, canonical URL slug, and full prose.

LLM Prompt For Alert Extraction Each page is passed, one call per alert, to `gpt-4o-mini-2024-07-18` together with a system prompt that instructs the model to return an array of "alerts" containing, for every disease mention, seven slots: disease, geography, date, month, year, cases, deaths, and computed case-fatality ratio if provided. All slots default to `NA` (not available) when absent. The schema is declared as `"strict": true` and `"additionalProperties": false` so that any hallucinated field triggers an automatic retry.

System Instructions We provide the following instructions to the model for each call:

```
You are an expert assistant analyzing WHO Disease Outbreak News (DONs).  Each text
describes an outbreak or suspected outbreak.  Your goal is to extract every mention of
a disease along with the geography, date (day, month, year), and numerical details such
as cases, deaths, and case-fatality ratio (CFR).
Task
1. Identify every disease name explicitly mentioned.
2. Identify every geography (country, region, or area) explicitly mentioned.
3. For each mention, locate the date (if given) and any figures on cases, deaths, and
```

```
CFR. If multiple dates or data points are present, create a separate entry for each
(disease, geography, date, cases, deaths, CFR) tuple.
Output
Return an array under the key "alerts". Each item must contain: disease, geography,
date, month, year, cases, deaths, cfr. Use "NA" for any missing element. Produce only
valid JSON matching the strict schema and add no extra keys.
```

User Prompt For each alert, the following prompt is supplied, where <TITLE>, <DATE>, and <CONTENT> are the page's metadata and prose:

```
Title of the alert: <TITLE>
Published date: <DATE>
Full content:
<CONTENT>
Please extract all disease-geography-date mentions as per the instructions.
```

Normalising Diseases to MeSH Free-text disease names are embedded with `text-embedding-3-large` (1,024 dimensions) and matched to the nearest MeSH descriptor by cosine similarity, exactly as described in Section B.2. The resulting MeSH codes are subsequently linked to ICD-10 and GBD level-2 causes for empirical work.

Refining Geography Strings Alert prose often mixes coarse labels ("*Central Africa*") with granular places ("*Kasese District*"). To standardise these, we run a second LLM call on the unique set of geography strings (roughly 1,500 after de-duplication) guided by a schema that asks for: city, country (ISO-3), continent (seven-code system), region, and other. Each geography is assigned a deterministic geography identifier, enabling many-to-one joins back to the alert table.

System Instructions We provide the following instructions to the model for each call:

```
You are an expert assistant specialised in geographic data extraction and mapping. For
every input string, return a JSON object with the fields below.
- City: the city named in the string.
- Country: three-letter ISO code for the country mentioned or, if only a city is
given, inferred.
- Continent: use the codes AF, AN, AS, EU, NA, OC, SA.
- Region: sub-national unit such as a state or province.
- Other: any remaining context that does not fit the above.

If the input is coarse (e.g. \Africa''), fill only Continent and leave the other fields
empty. If the input is granular (e.g. a city), fill City and, where possible, infer
Country and Continent. Return valid JSON that matches the strict schema and contains
no extra keys.
```

User Prompt for Geography For each distinct geography string we send:

```
Geography:  <GEOGRAPHY_STRING>
Please extract City, Country (ISO3), Continent, Region and Other from the string above
and return them in the required JSON format.
```

Post-processing Steps

1. **Year back-fill.** Where the alert text lacks an explicit year (≈ 6%), we impute it from the publication date.

2. **MeSH tree levels.** A unique MeSH tree cross-walk maps every descriptor to up to three hierarchical levels (mesh tree level 1–3), facilitating aggregation in event-study regressions.

Limitations Three issues deserve note. First, case and death figures follow no common template, so treating every number that precedes "cases" or "deaths" as valid can blur distinctions between suspected and confirmed counts. Second, several DON items cover more than one disease (for instance dengue and chikungunya), and our extraction rightly records each mention but inherits any ambiguity in the source text. Third, where the prose names only a continent, finer geographic slots remain empty, potentially understating sub-national variation. Even so, the pipeline yields the high-frequency, geography-resolved shocks that underpin the event-study design.

D Details on LLM classification of research funders

We classify every one of the 32,437 funders listed in the 2024 OpenAlex snapshot. Each funder's display name is sent five times to `gpt-4o-mini-2024-07-18` (temperature 0.7). The model is instructed to return a JSON object with six Boolean keys: direct government, independent public, private corporate, private philanthropic, academic institution, and hybrid. Strict schema validation (`‘‘strict’’: true`, `‘‘additionalProperties’’: false`) guarantees parseable output.

System Instructions We provide the following instructions to the model for each call:

```
You are an expert assistant in the analysis and classification of funding
organizations.  Given the display name of a funder from the OpenAlex dataset, your
task is to determine its type across several categories.  Specifically, based on the
provided display name, decide whether the organization is:
1.  Direct Government:  A government agency fully funded by public resources.
2.  Independent Public:  Receiving public funds but operating with independent
decision-making.
3.  Private Corporate:  A for-profit corporate entity, often funding research as part
of a strategic or CSR initiative.
4.  Private Philanthropic:  A philanthropic or non-profit organization that funds
research through endowed or donated resources.
5.  Academic Institution:  An academic or research institution funding research
internally or via affiliated research bodies.
```

```
6.  Hybrid:  An entity that combines multiple funding sources (public, private, etc.)
or does not clearly fit into the above categories.
Based solely on the funder's display name, classify it by returning boolean values for
each of the above categories.  If there is any ambiguity or insufficient information,
default to "false" for categories that are not clearly applicable.  Return only a
valid JSON object adhering strictly to the provided schema, with no additional keys or
commentary.
```

User Prompt For each funder, the following prompt is supplied, where ”<DISPLAY_NAME>” is the name of the funder as provided by OpenAlex:

```
Funder Display Name:  <DISPLAY_NAME>
Please classify this funder into the following categories based solely on its name:
- Direct Government
- Independent Public
- Private Corporate
- Private Philanthropic
- Academic Institution
- Hybrid
Return only a JSON object following the provided schema with boolean values for each
field.
```

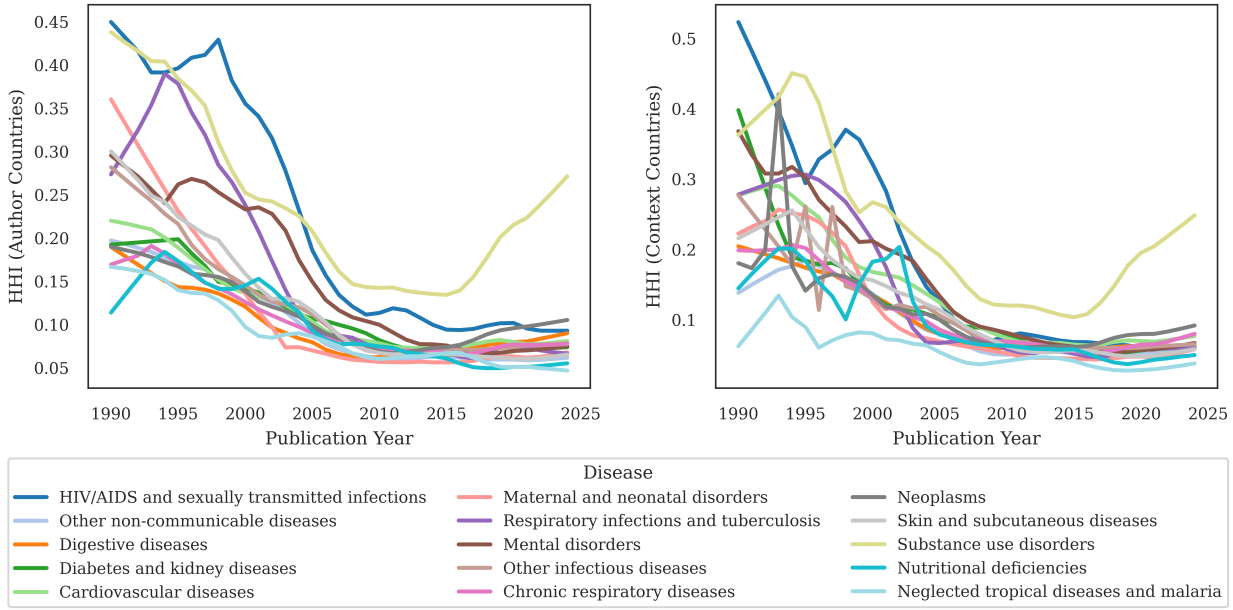
Aggregation and Consistency For each (*id*, category) pair we take the modal value across the five responses; ties are broken at random. Across the six categories, 96–99% of funders receive identical labels in all five calls, indicating high internal reproducibility despite the stochastic sampling (model temperature was set at 0.7).

Category Prevalence. In the universe of funders, 36% are labelled academic institution, 34% private philanthropic, 17% direct government, 9% private corporate, 4% independent public, and 2% hybrid. Because the categories are not mutually exclusive, a single funder can carry multiple `true` flags.

Potential Limitations Classification relies on the name alone, which can be ambiguous or language-specific; mixed-funding bodies may fit more than one bucket; and the six categories are an imperfect proxy for more granular legal forms. For our purposes (broadly contrasting public, corporate, philanthropic and hybrid sources) these coarse buckets are adequate, and the high agreement rate suggests that any residual misclassification is unlikely to change the qualitative results.

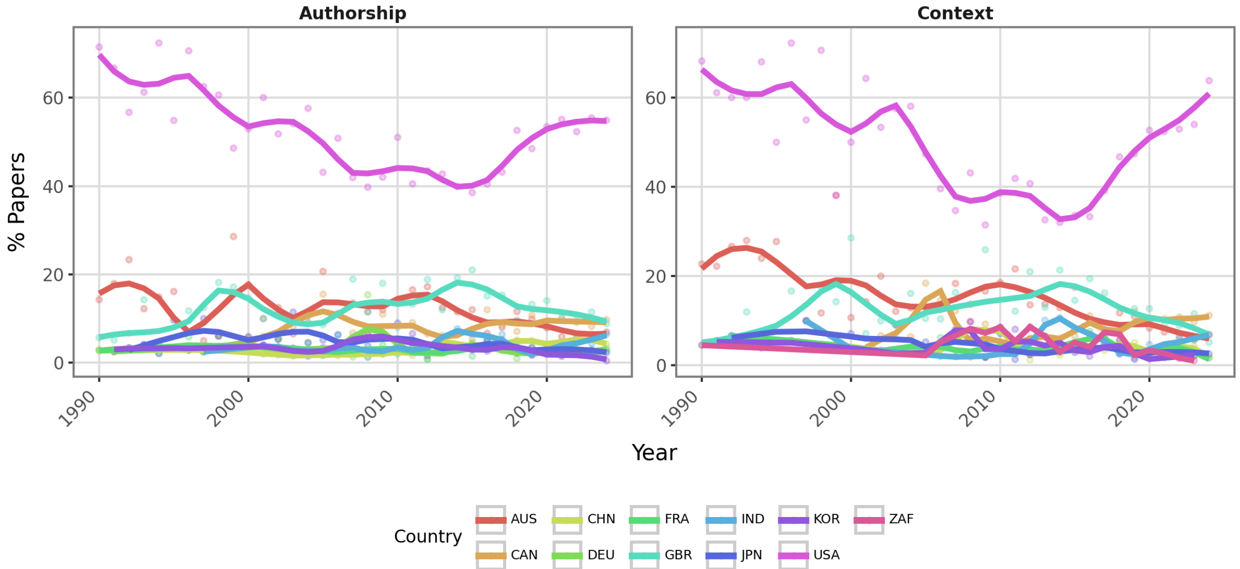
E Appendix Figures

Figure E1: The concentration of disease research output and contextual mentions of countries over time.



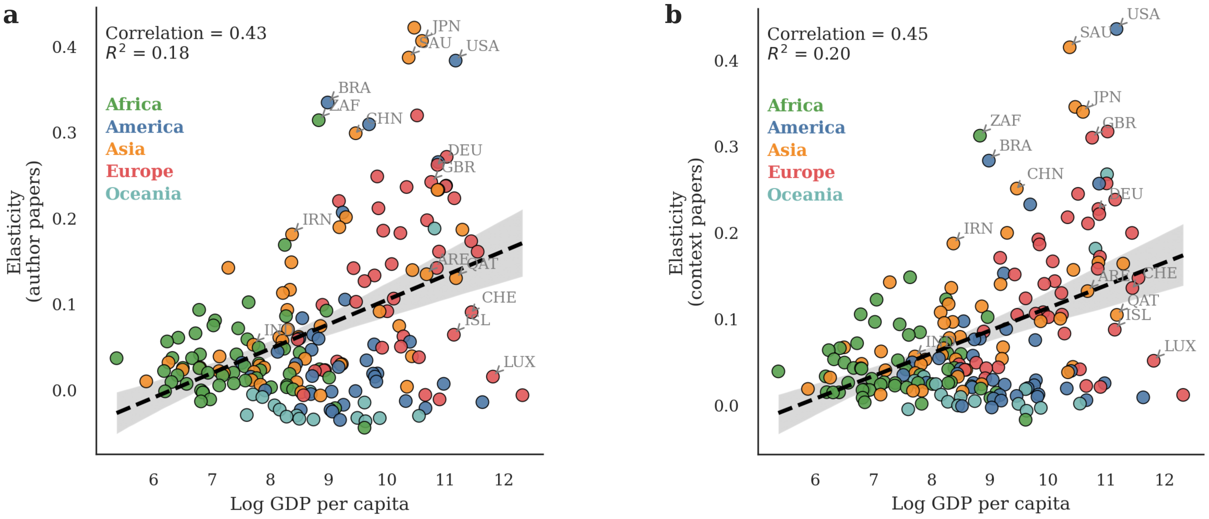
The Herfindahl-Hirschman Index (HHI) is displayed across three decades to illustrate how the concentration of country involvement evolves for various disease categories. The left panel shows HHI for countries as authors ("Author Countries"), while the right panel shows HHI for countries mentioned in titles or abstracts ("Context Countries"). Higher HHI values indicate a greater concentration among fewer countries, whereas lower values suggest broader participation or references.

Figure E2: Evolution of the top 11 most prolific countries publishing on substance use disorders.



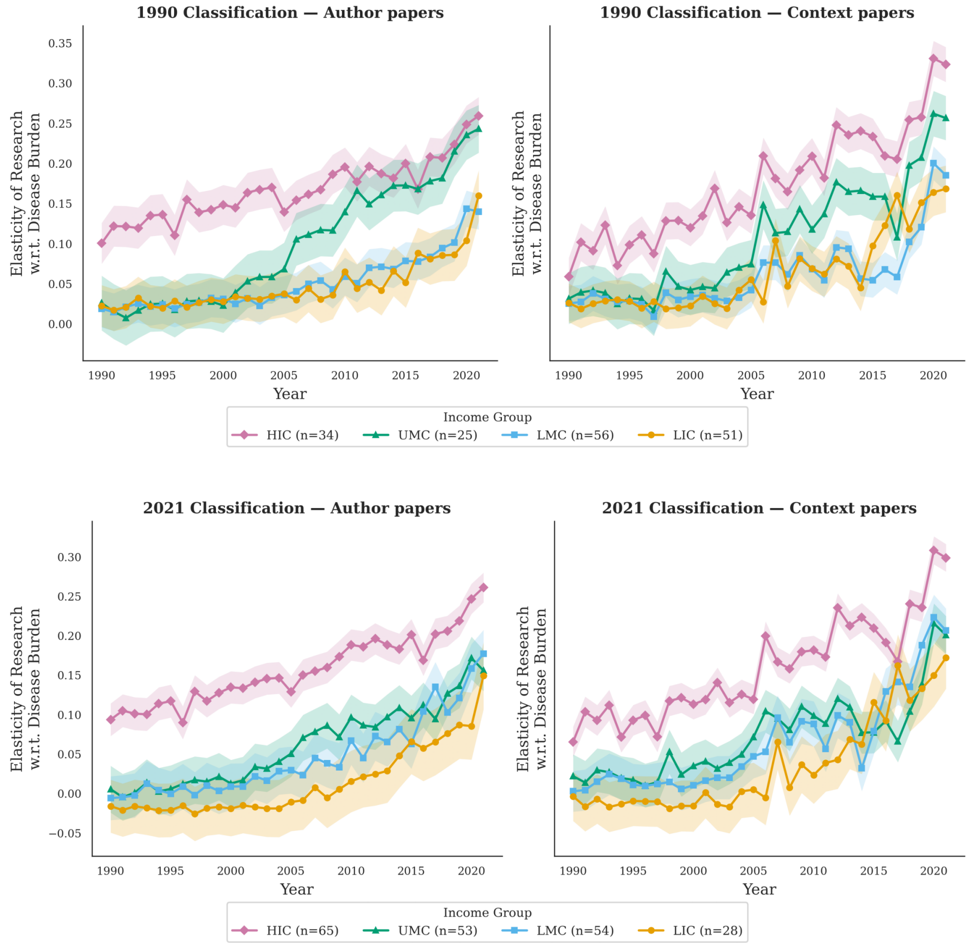
The figure depicts the percentage of articles on substance use disorders attributed to the eleven most prolific countries in this area. The United States consistently accounts for the largest share of publications, with its proportion declining prior to 2015 but rising substantially in subsequent years. Since around 2015, the U.S. share has increased from approximately 40% to 57% by authorship and from 32% to 61% by country context in our dataset.

Figure E3: Country-level research responsiveness with relations to GDP per capita by authored papers and contextual mentions.



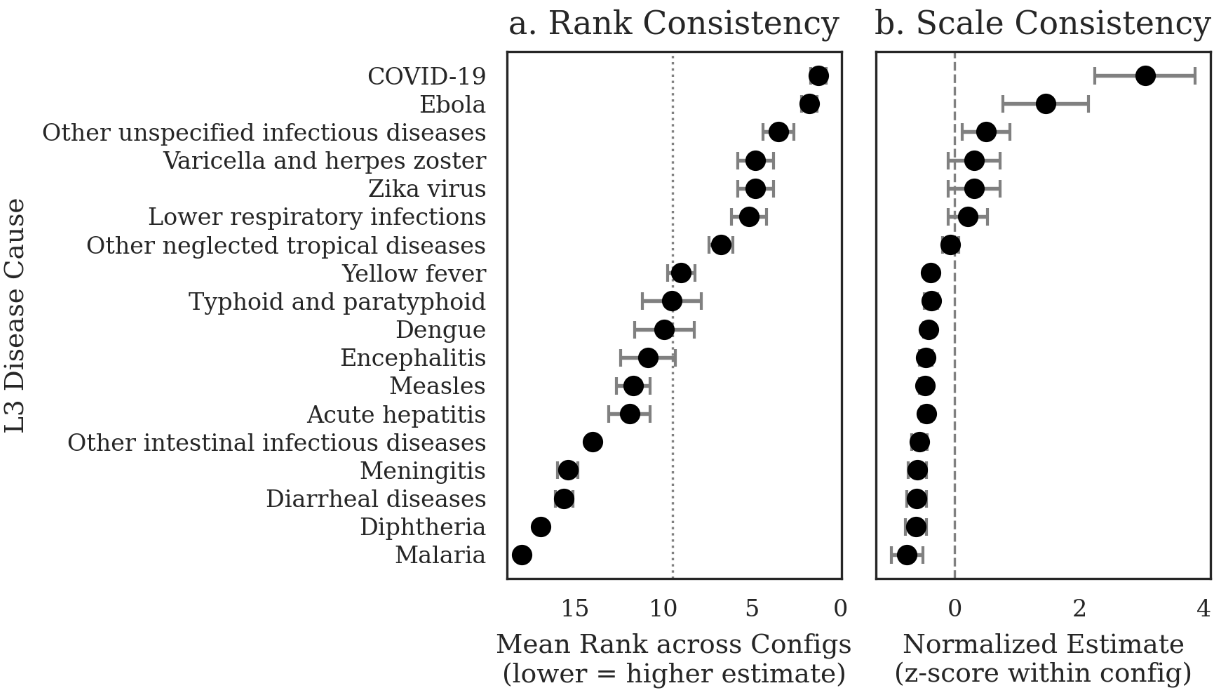
We estimate countries' responsiveness (elasticity) of research output to disease burden using a regression with an interaction between log-transformed disease burden and country, including fixed effects for year and disease category. The elasticity represents the percentage change in research output associated with a 1% change in disease burden, yielding a responsiveness coefficient for each country. We then examine its association with GDP per capita in scatter plots. Panel a reports estimates using the log of countries' authored paper counts as the dependent variable, while panel b uses the log of paper counts in which countries are mentioned in titles or abstracts. Both panels indicate a positive relationship between research responsiveness and GDP per capita, with correlations of 0.43 and 0.45, respectively.

Figure E4: Evolution of elasticity of research output to disease burden by country income group (1990 vs. 2021 classifications; author papers vs. contextual mentions).



This figures shows how the responsiveness (elasticity) of research output to disease burden has evolved across countries with different income levels. Elasticity estimates are derived from a fixed-effects regression with interaction terms for income group and year, allowing the relationship between research output and disease burden to vary by both dimensions. Specifically, the model estimates the percentage change in research output associated with a 1% change in disease burden for each income group and year, controlling for country, disease category, and year fixed effects. The lines represent the estimated elasticity for each income group over time, with shaded bands denoting 95% confidence intervals. The left-hand panels use countries' authored paper counts as the dependent variable, while the right-hand panels use counts of papers in which countries are mentioned in titles or abstracts. The top row classifies countries by their 1990 income groups (the start year of our analyses), whereas the bottom row uses 2021 classifications. The upward trend in responsiveness is robust across all income groups and dependent variable specifications.

Figure E5: Robustness of pooled event-study estimates of outbreak impacts on research effort, averaged across configurations and GBD level 3 disease categories.



Panel a presents the mean rank of disease categories across specifications (lower ranks correspond to larger estimated effects), with horizontal bars showing variability across configurations. Panel b reports the mean normalized estimates (z-scores within each configuration) for each category, again with horizontal bars indicating variability. Consistency in both rank and scale across diseases demonstrates that the observed various outbreak effects on research effort by diseases are robust to alternative model specifications.

Figure E6: Mean dynamic estimates by time to event across specifications.

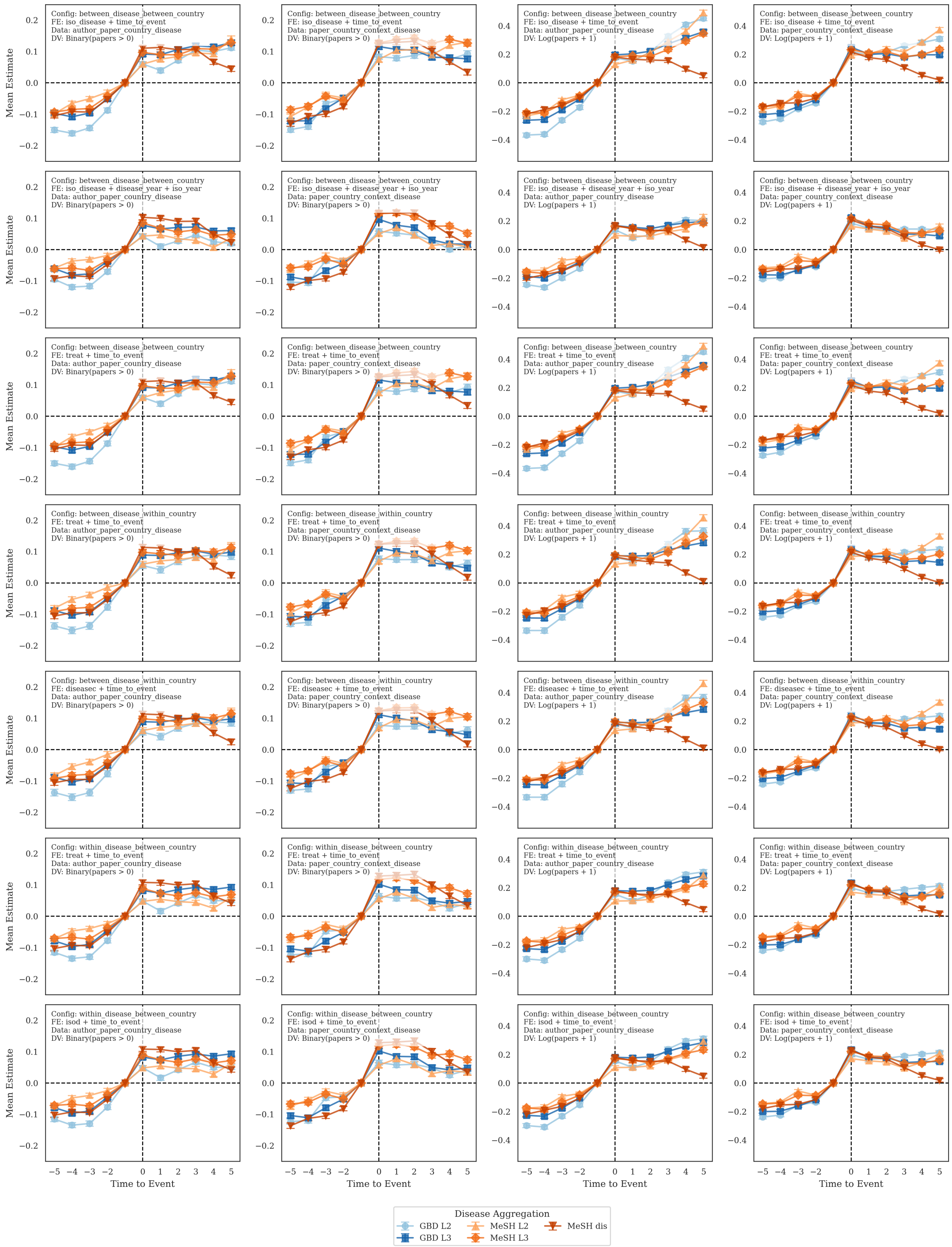


Figure E7: Mean dynamic estimates by decades across specifications.

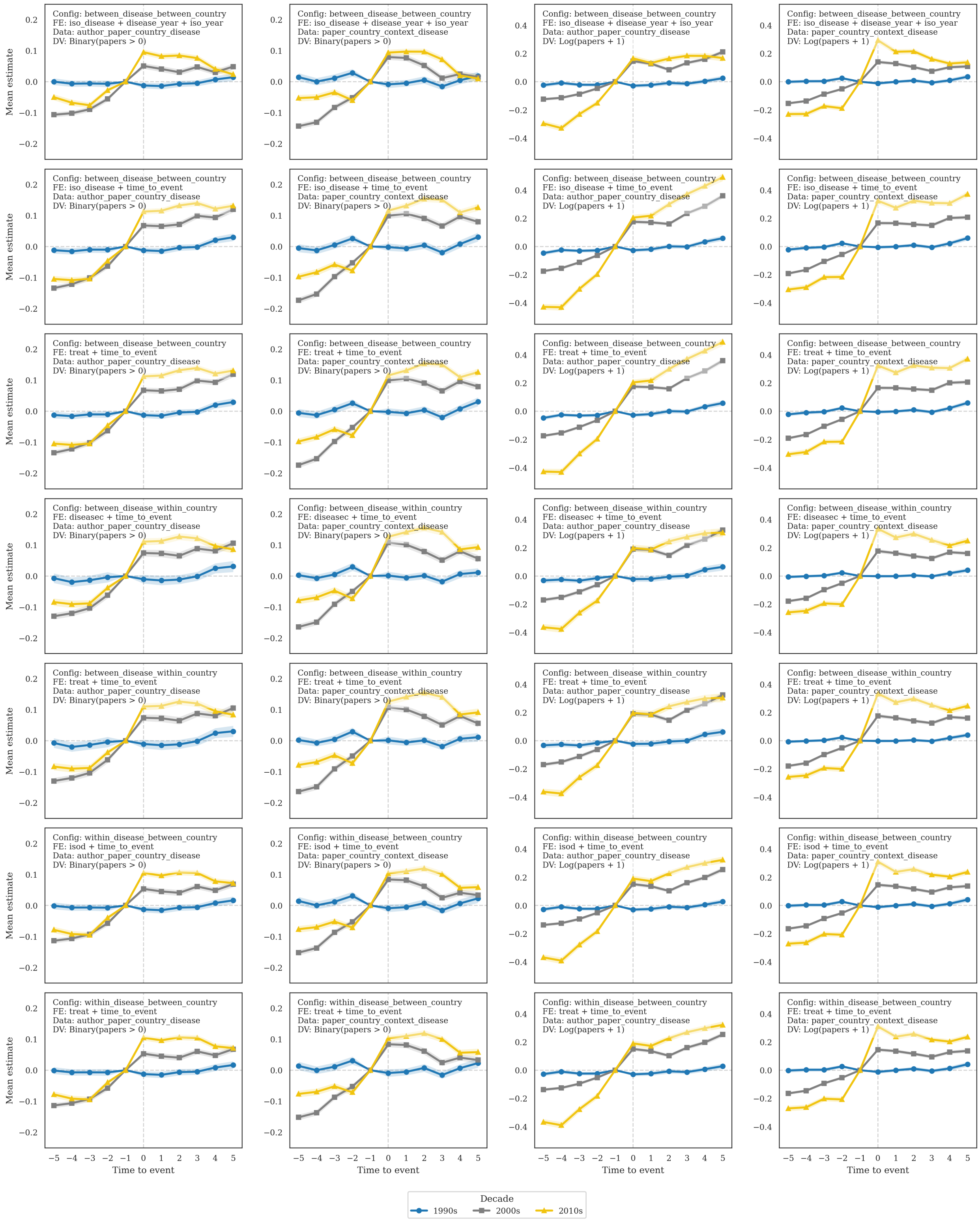
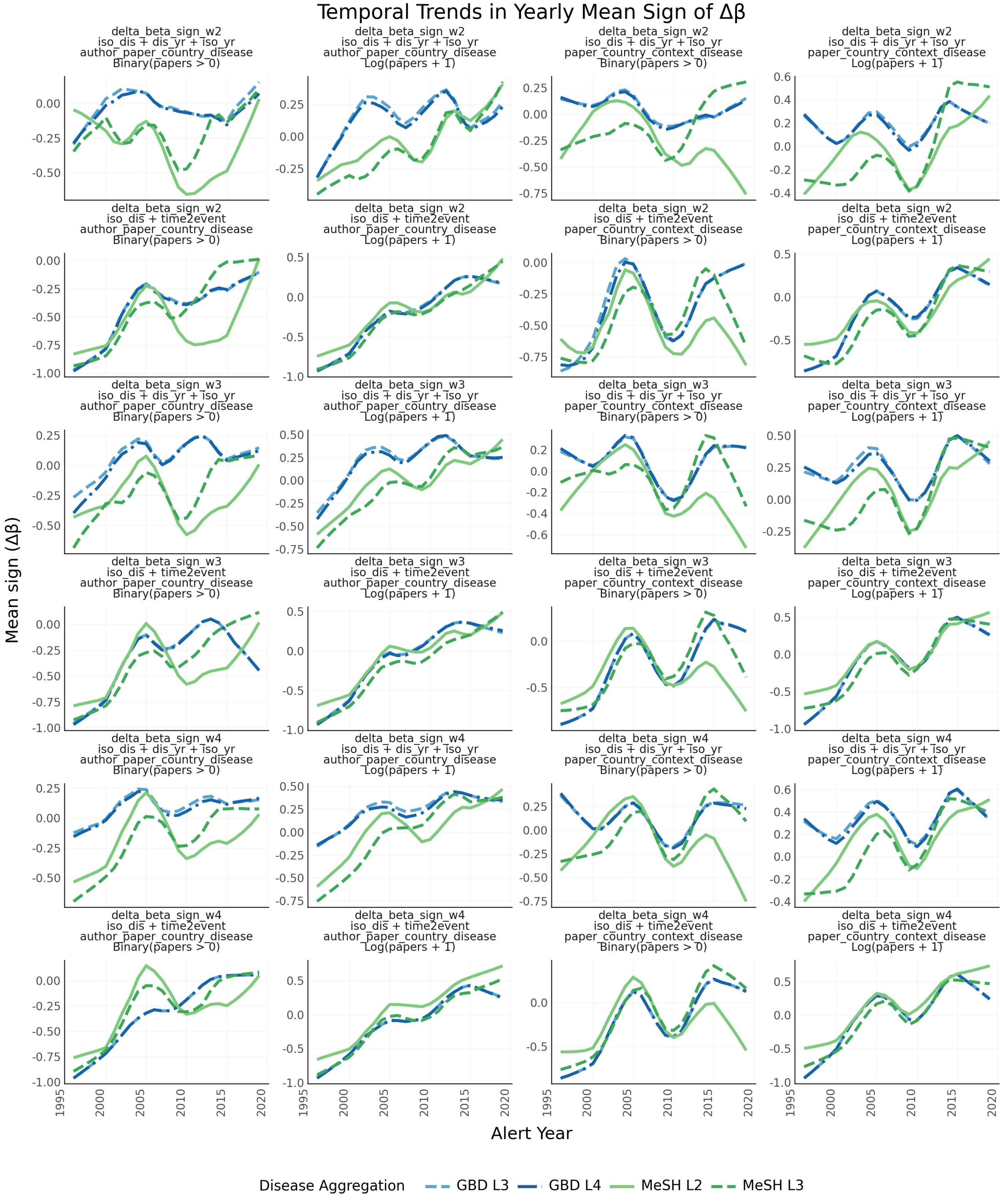


Figure E8: Robustness check on characterizing temporal shifts before and after the alert



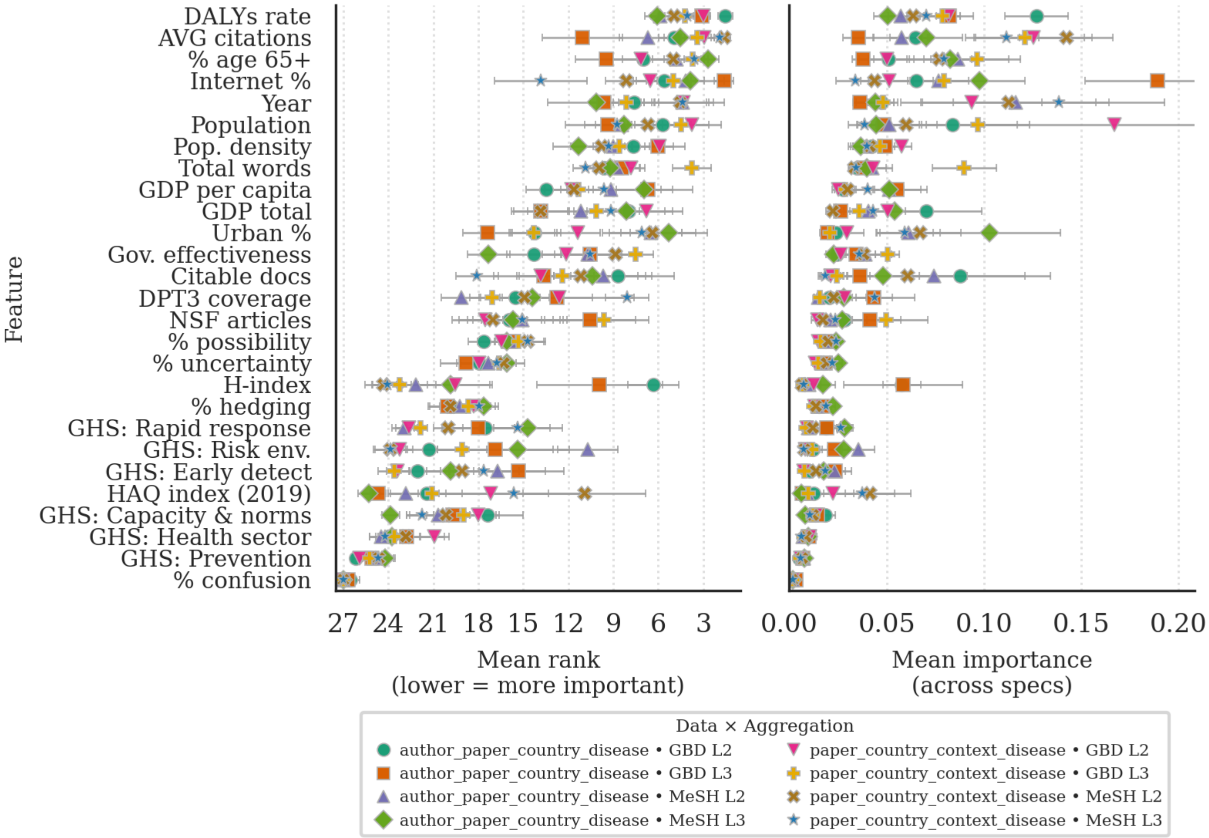
To assess how research activity shifts in response to a WHO alert, we calculate the change in the estimated event-study coefficients (β_τ) before and after the alert. For each alert event, we define the **delta beta** ($\Delta\beta$) as:

$$\Delta\beta = \frac{1}{w}\sum_{\tau=1}^{w}\beta_{\tau,\text{post}} - \frac{1}{w}\sum_{\tau=1}^{w}\beta_{-\tau,\text{pre}}$$

where w is the window size (e.g., 2, 3, or 4 years), $\beta_{\tau,\text{post}}$ are the coefficients for years after the alert, and $\beta_{-\tau,\text{pre}}$ are those for years before the alert. This statistic summarizes the average change in research output associated with the alert across different time windows.

By averaging $\Delta\beta$ across all events, we characterize the typical direction and magnitude of change in research response to alerts.

Figure E9: Robustness check on feature importance across specifications



Mean feature ranks (left, lower = more important) and mean absolute importances (right) estimated from Random Forest models. Each error bar corresponds to one dataset × disease aggregation combination and shows the mean and 95% confidence interval of pooled estimates, calculated across variations in dependent variable, fixed effects, and event-study configuration. This robustness check demonstrates which predictors remain consistently important across alternative modeling choices.