OECD Science, Technology and Industry Working Papers

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to Directorate for Science, Technology and Innovation, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/STP/NESTI/MARIAD(2022)4/FINAL

Measuring governments' R&D funding response to COVID-19

An application of the OECD Fundstat infrastructure to the analysis of R&D directionality

This paper presents new evidence on the size and direction of governments' R&D funding response to the COVID-19 pandemic through the exploration of a novel data infrastructure, the OECD Fundstat initiative for the analysis of government-funded R&D projects. The document reports on the exploratory development and application of automatic classification tools to detect relevant COVID-19 R&D funding, map salient topics and classify and allocate project funding according to priorities in the WHO COVID-19 R&D Blueprint, as well as comparing results with similar analysis of scientific publication output data. The results provide new insights on which areas of enquiry were prioritised by governmental R&D funding bodies.

Authors: Leonidas Aristodemou, Fernando Galindo-Rueda, Kuniko Matsumoto, and Akiyoshi Murakami

Keywords: COVID-19, Government funding, Research and Development (R&D), directionality, topic modelling, classification, large language models

JEL codes: C38, C45, O32, O38

Acknowledgements

This report has been prepared by Leonidas Aristodemou, Fernando Galindo-Rueda, Akiyoshi Murakami, and Kuniko Matsumoto at the Science and Technology Policy Division in the OECD Directorate for Science, Technology, and Innovation (DSTI). Brigitte van Beuzekom facilitated the use of bibliometric data to complement the analysis.

This study has been conducted as part of the Programme of Work and Budget 2021-2022 of the Committee for Scientific and Technological Policy (CSTP) under aegis of the Working Party of National Experts on Science and Technology Indicators (NESTI) and entrusted to the Expert Group on the Management and Analysis of R&D and Innovation Administrative Data (MARIAD). Earlier versions of the document were presented for discussion in the September 2022 NESTI meeting, November 2022 MARIAD meeting, March 2023 OECD-MABIS project workshop and March 2023 CSTP meeting.

The authors would like to express their gratitude towards MARIAD and NESTI delegates and to their respective Bureaus for their feedback. In addition, some of the analysis in this study would not have been possible without the assistance of MARIAD delegates from several countries in making funding data available for analysis.

Voluntary contribution support by Japan's Ministry of Education, Culture, Sports, Science and Technology, the United States National Science Foundation (NSF) National Centre for Science and Engineering Statistics, and the European Commission's Horizon 2020 Programme through the Mapping Business Innovation Support (MABIS) project, is also gratefully acknowledged.

Additional information:

A spreadsheet workbook containing the figures and tables presented in the report is available for download at https://gitlab.algobank.oecd.org/OECD_FUNDSTAT/oecd_fundstat/covid19_rnd/public. For further questions on the data and methods used in this report, please contact: mariad@oecd.org.

Table of contents

Tables

Figures

Executive summary

The COVID-19 pandemic presented the world with a unique and major global public health emergency not seen in generations. Rising to the challenge, science, technology, and innovation (STI) systems have played a key role in containing the virus's spread, developing, and deploying vaccines and treatments in record time, and providing tools and knowledge to help combat the pandemic, mitigating against its negative impacts. Government support for research and development (R&D) in both public and private sectors has been instrumental.

Monitoring the governmental R&D funding response is a major priority to help inform collective action both while a crisis is ongoing and afterwards, to build an evidence base to foster increased resilience against future pandemics or shocks. Understanding the size and direction of government R&D funding response in crises like the COVID-19 pandemic and having the appropriate data infrastructures to do so are necessary conditions for realising that vision. The OECD Fundstat initiative emerged to fill this gap, prompted by the 2015 OECD Daejeon ministerial declaration and the 2016 OECD Blue Sky Forum, to pursue the creation of a flexible international analytical infrastructure to study government R&D funding directionality. By using data on publicly funded R&D projects, combining both quantitative and qualitative information, it enables a detailed, granular, and timely analysis of specific policy priorities.

The work presented in this document deployed a range of tools for the integrated analysis of project funding data to identify government financial support for COVID-19 R&D as an experimental study of R&D funding directionality. This project set out to: (i) help provide evidence on the composition of COVID-19 R&D funding provided by government agencies; (ii) demonstrate the use of natural language processing (NLP) methods to measure directionality for policy analysis; and (iii) provide a basis for scaling up the OECD Fundstat infrastructure and encourage country engagement, collaboration, and mutual learning.

This study provides an in-depth analysis of R&D support portfolios using data from 27 funding sources from 13 OECD countries and the European Commission (EC) and retrieving funding for COVID-19 R&D projects approved in 2019-21. The 11,886 projects identified add up to total government funding of USD 12.59 billion and average funding per project around USD 1.20 million. This represents 4% of R&D project funding registered in the Fundstat database over that period and 2% of projects. The application of topic modelling analysis of the corpus of COVID-19 R&D projects identified 34 distinct topics, grouped into 8 higher level topic clusters which have been labelled as follows: 'Coronavirus understanding, therapeutics, and vaccine development', 'Platforms and capabilities', 'Epidemiology and social intervention', 'Digital access and online education', 'Cancer (screening and treatment)', 'Public healthcare and other groups at risk', 'Mental health and addictions', and 'Environmental detection, transmission, and protection'.

While biomedical (including R&D on virus understanding, development of diagnostics, vaccines, and treatments) and social science-oriented topics are generally balanced in terms of numbers of projects, government funding for COVID-19 R&D is primarily focused on the former. There is evidence that on average biomedical projects are larger in terms of funding awards relative to social science-oriented topics.

Funding for R&D platforms and capabilities is significant, and co-occurrence patterns indicate that this topic plays a pivotal role across different areas of COVID-19 R&D. This is particularly relevant as health and R&D systems seek to build resilience capacity towards future pandemics or attempt to find ways to apply COVID-19 based discoveries and technologies to other pressing health challenges.

Based on the language used in project's text descriptions, market-oriented R&D projects, a concept that is proxied using a business and market vocabulary, accounted for 34% of COVID-19 R&D funding, representing 22% of COVID-19 R&D projects, with innovation-funding agencies predictably showing higher shares of market-oriented R&D projects and funding.

Analysis of language patterns in the COVID-19 Research Project Tracker by UKCDR & GloPID-R, which has manually labelled project data mapped to the WHO classification of research priorities, has COVID-19 R&D funding model. The application of this model to the OECD Fundstat database suggests that R&D on vaccines and therapeutics R&D have each received very similar funding allocations. On that basis, it is not possible to conclude that the priority topic of R&D therapeutics was relatively underfunded compared with vaccines as it has sometimes been claimed in the past. The results also help contextualise evidence obtained from analysing the thematic profile of scientific publications, since counts of papers do not provide an indication of resource intensity.

The study also provides several methodological insights for the conduct of analysis on the directionality of R&D funding with NLP methods:

- **Measuring R&D directionality towards specific challenges requires precise and implementable concepts.** Defining and implementing relevance is a key task for mapping R&D funding, and funding agencies could converge towards shared data standards. This can enhance transparency, stewardship, and facilitate international comparative analysis and coordination.
- **Relying on multiple indicators, such as the combination of funding with project descriptions, can provide insights into the directionality of public support for R&D.** Policymakers should consider complementary indicators, as solely relying on document-counts indicators can result in significant biases.
- Combining the insights of **classifications using machine-driven methods and formal taxonomies with expert labelling** can facilitate a more **comprehensive understanding of the government R&D funding landscape.**
- In light of reported coverage limitations and missing information, there would be major analytical benefits from encouraging funding agencies to **converge towards data openness and, whenever possible, use of common core metadata, for accountability and analysis purposes**.

The experience of using large language models for processing text data in this study showcases pivotal advancements in statistical and policy analysis as well as several implementation and interpretation challenges. This methodology **opens the possibility of application to other R&D policy challenges.** The careful integration of generative artificial intelligence (AI) tools can help **monitor trends swiftly and anticipate R&D needs in other urgent areas**. The work on Fundstat by the NESTI MARIAD group will continue to foster the responsible development of the underlying data infrastructure and application of these methodologies.

1 Introduction and background

1.1. The COVID-19 pandemic and the R&D funding response

The global pandemic that ensued the coronavirus outbreak, officially declared in January 2020 by the World Health Organisation (WHO) as public health emergency of international concern (PHEIC), presented the world with a major challenge that impacted on everyone's lives. While in early May 2023 the WHO's International Health Regulations Emergency Committee concurred that the PHEIC declaration should end, its effects have been profound and can still be felt at the global scale (WHO, 2023[1]). In May 2023, reports indicated that there had been over 767 million confirmed cases of COVID-19 and 6.9 million deaths (WHO, 2022[2]; WHO, 2023[3]).

The response of science, technology, and innovation (STI) systems to the COVID-19 crisis has been particularly vigorous, playing an essential role in generating the knowledge and technologies needed to respond to the COVID-19 crisis (OECD, 2023[4]). STI has been central in informing governments' efforts to limit the virus spread and underpinning the rapid development and deployment of effective vaccines and treatments (OECD, 2021[5]). As of May 2023, the WHO Coronavirus COVID-19 portal indicates that over 13 billion vaccine doses have been administered (WHO, 2023[3]). The pandemic has underscored the importance of science and innovation to societal capacity to both proactive prepare and reactively responds to future crises (WHO, 2020[6]; OECD, 2023[4]).

Government financial and non-financial support for R&D in both public and private sectors has been pivotal to the rapid development of tools (e.g., vaccines, therapeutics, and diagnostics) and knowledge (e.g., virus understanding, epidemiological monitoring, and social behavioural insights). This response would not have been possible without a solid foundation of scientific and technical knowledge built over decades and also supported by government funding programmes. All this combined has been crucial in helping the world to overcome the several challenges posed by the COVID-19 pandemic (Tietze et al., 2022[7]; OECD, 2021[8]; Agarwal and Gaule, 2022[9]; OECD, 2023[10]). Several tracker initiatives have reported how newly funded R&D initiatives worth billions of dollars have been set up in record time, and research and innovation have led to the rapid development of vaccines (EU/OECD, 2022[11]; Policy Cures Research, 2020[12]; INGSA, 2020[13]; Bucher et al., 2023[14]; UKCDR & GloPID-R, 2023[15]).

Looking ahead, while there are many concrete lessons to be drawn from the COVID-19 experience, the focus is turning towards ensuring preparedness and resilience of health systems, with specific emphasis on the health science and innovation subsystem. For example, the Japan-hosted G7 Science and Technology Ministers' Communique of May 2023 alluded to the possibility "[through international collaboration in research and innovation]…to collectively address urgent global health issues such as the need to develop safe and effective medical countermeasures (MCMs) in the event of a future pandemic as promoted through the 100 Days Mission, including vaccines, diagnostics, and therapeutics, to combat infectious disease threats, as well as tools to address other shared health burdens like cancer" (G7, 2023[16]). Given the plethora of policy questions about how best to deploy and direct R&D funding in response to and anticipation of crises as well as other longstanding challenges, it is imperative for societies to be able to count on a robust evidence base and set of tools that enable tracking of government R&D funding and assessment of its impacts.

1.2. Understanding the R&D COVID-19 response through project analysis: a motivation for conducting a proof of concept for the OECD Fundstat initiative

When responding to STI systems and policy monitoring, evaluation and appraisal needs, it is important to note that different types of data are best suited to different purposes. Official statistics on government R&D budgets are designed from a top-down perspective to capture from high-level administrative finance documents, on a regular and longitudinally consistent basis, the full range of high-level government policy priorities to which R&D funds are allocated. Government R&D budget statistics collected by national authorities and compiled by the OECD apply a mutually exclusive allocation of R&D funding to socioeconomic objectives that reflects top-level priority setting (OECD, 2015[17]). One downside is that these statistics are not designed to capture funding allocations on a granular basis and are therefore not suited to track funding directed to tackle the COVID-19 pandemic or other specific, context-contingent subjects.

Echoing discussions at the OECD meeting of science ministers held in Daejeon in 2015, the OECD Blue Sky Forum held in 2016 on the future of science and innovation data and indicators posited the possibility of developing complementary, micro-based pathways focused on the analysis of project-level data to complement more established means of statistical analysis of R&D funding (OECD, 2015[18]; OECD, 2018[19]). The notion of an 'R&D project' as unit of analysis had already been explicitly introduced in the 2015 edition of the Frascati Manual with the aim of facilitating the reporting of R&D data for R&D statistics (OECD, 2015[20]). Data about R&D projects can be extremely rich sources of information for policy analysis. The proposals at the OECD Blue Sky Forum compelled the OECD to promote the active and coordinated use of data about R&D projects (funding, metadata, and textual descriptions in abstracts), under what came to be described as the "*Fundstat"* initiative for a brand new analytical data infrastructure (OECD, 2018[21]).

In contrast with the approach of building *ad hoc* trackers to address one measurement priority *at a time*, a key part of the concept behind Fundstat is the aim to develop a standing but flexible infrastructure suitable for use as soon as such priorities arise. The semantic context of text-based project descriptions is a critical element for the implementation of this concept, in combination with the use of machine-supported methods of classification for projects and the allocated funding amounts to each one of those. This bottom-up, machine-guided, approach presents unique challenges of its own. These relate to the difficulties in building up an underlying international database to ensure that it provides an adequately representative basis for measurement, as well as ensuring that the machine-based or human-assisted methods of classification are fit for purpose, particularly in context where there is no pre-established consensus on what should be measured.

The OECD Fundstat infrastructure was first piloted through an analysis of government funding for R&D projects related to Artificial Intelligence (Yamashita et al., 2021[22]). Direct responsibility for this initiative of the OECD Working Party of National Experts on Science and Technology Indicators (NESTI) has been assigned to the recently established OECD Expert Group on the Measurement and Analysis of R&D and innovation administrative data (MARIAD), which was been created to assist in the pursuit of and quality assessment of statistical analysis based on administrative data. MARIAD included this initiative within its action plan going up to 2024, contributing to the overarching work of the Committee of Scientific and Technological Policy (CSTP) on science and innovation for resilience and transitions.

1.3. Aim and outline of this study

In this study, quantitative and qualitative tools have been used to identify government funding of COVID-19 R&D in the Fundstat database, an analytical data infrastructure under continuous development. The main objectives are to illustrate the level and composition of COVID-19 funding by government agencies,

characterise and address the methodological challenges of identifying COVID-19 relevant R&D projects from heterogeneous text-based data, and motivate the process of R&D project data sharing at an international scale. This provides additional motivation for the extension and consolidation of the Fundstat data infrastructure, helping assess the extent to which R&D project databases represent government R&D funding and providing the basis for future analysis of other topics. Furthermore, this study aims to demonstrate the potential of AI methods for the analysis of funding administrative data, as complementary detection, and classification tools for statistical measurement, enabling greater analysis uniformity and replicability, providing a tangible milestone in the scaling up of the OECD Fundstat infrastructure and the development of NLP methods to measure directionality.

The remainder of this paper is thus structured as follows.

- Section 2 provides a brief description of the R&D funding data used for the analysis and the methodology applied. It describes the pre-identification of R&D projects using key-terms, the automated machine-based procedure to eliminate projects whose abstracts only make contextual references to COVID-19 terms, and the topic modelling analysis using natural language processing (NLP) methods to infer relevant funded COVID-19 R&D topics.
- Section 3 presents the analysis of directionality of COVID-19 funding by topic, cluster, country/area, funding agency and market/commercial orientation.
- Section 4 compares the results on funding data and the machine-based classification to those of the COVID-19 Research Project Tracker by UKCDR & GloPID-R, which has been mapped by experts in global health research against the priorities identified in the WHO Coordinated Global Research Roadmap for COVID-19 (Bucher et al., 2023[14]; UKCDR & GloPID-R, 2023[15]). The expert classification data are used to train a model that enables the classification of Fundstat data according to WHO priorities as well as an in-depth analysis of machine and expert classification patterns. Furthermore, the results are also compared to scientific publications, accounting for the different factors that underpin differences between R&D input and publication output data.
- Section 5 concludes by outlining the key messages, limitations, and future work.

2 Data and methodology

2.1. Project funding data from the Fundstat infrastructure

Administrative data on R&D project funding by governments present outstanding opportunities as well as major challenges when intended for statistical analysis in an international comparative context. Governmental R&D funding procedures provide a key foundation for R&D project records to exist when information would otherwise be non-existent, instilling a basic degree of accountability and quality control with core features that are common across agencies and countries. However, these data are not systematically available for use across countries and their funding bodies and agencies. When available for analysis, a growing welcome trend obeying to transparency principles, they often come in very different formats and styles which makes them complex objects of analysis (OECD/Eurostat, 2018[23]).

It is important to note that, depending on the country, project level information may not be necessarily representative of all government R&D funding. Public R&D funding systems around the world adopt different approaches to allocate resources to the teams and individuals who conduct R&D work. In some countries, particularly those relying more on institutional block funding, a significant share of public funds may be passed on to institutions to decide on how to allocate the resources internally without an obligation on their part to report on how funds are allocated to specific projects. In many cases, the bulk of the salary and fixed capital component of R&D activity costs may fall outside the scope of project-level funding calls, thus leaving supplies, services, incidentals, and potentially additional hires – often at more junior levels – as the only elements of a project's total cost within scope for analysis. However, despite these multiple limitations, project-level data can be informative of the government's discretionary use of R&D funds as well as other non-discretionary instruments where there may also be project-level data available for analysis. This can shed some light on the intended and actual direction of government support for R&D (Veugelers, Wang and Stephan, 2022[24]).

Specific purposes of both funders and applicants underpin the data generation process and constrain the range of admissible interpretation of indicators based on such data, particularly those that are based on analysis of textual descriptions of projects. In the case of government funded R&D projects, the depth and breadth of the information requested by funders and provided by applicants depends on the incentives and constraints that each face. Tagging of projects by applicants or administrators is potentially subject to error and inconsistent applications of definitions, a problem that can impact indicators based on administrative data, particularly those that rely on machine-learning based analysis procedures. In an ideal scenario, project descriptions would be available in full and not just limited to abstracts, but confidentiality restrictions apply. In the same scenario, project abstracts should provide sufficient and standardised information allowing managers and analysts to discern at least the foundations, methods and key expected findings of a given project.

The Fundstat infrastructure of government R&D funding projects is an entity under continuous development within the OECD Directorate for Science, Technology, and Innovation under the oversight of the OECD Expert Group on the Management and Analysis of R&D and Innovation Administrative Data (MARIAD). The version of Fundstat (December 2022) used as the basis for the study of COVID-19 R&D comes from 27 organisational databases, originating from 13 OECD member countries and the European

Commission (EC). A full list of the contributing organisational databases and funding agencies is available in Annex A. It is worth noting that the database comprises of project descriptions in languages other than English, namely Dutch, German, French, Japanese, Latvian, Norwegian, and Swedish. Altogether, Fundstat contains 607,098 project awards granted from 2019 to 2021, accounting for a total funding amount of USD 287.58 billion (Table 1). Data collected for 2021 are partial due to different data reporting structures across different countries. About 5% of project observations in Fundstat do not include funding information or have assigned a value of nil funding. Such entries often obey to reporting disclosure controls, and sometimes, rather than standalone projects, they can reflect project award entries intended to register modifications on previous awards that do not have any funding implications (e.g., zero additional funding).

Table 1. Description of R&D awards and funding in the Fundstat database, 2019-21

	FY2019	FY2020	FY2021 (partial)[1]	Total
R&D project awards ("projects")	211,968	221,076	174,054	607,098
(Of which) R&D projects with missing funding information	9,665	11,412	9,256	30,333
R&D funding [USD million]	90,044	101,599	95,936	287,579
Mean project award [USD million per project][2]	0.45	0.48	0.58	0.50

Notes:
[1] FY2021 has partial data because of data reporting structures and data availability across different countries.
[2] The calculation of mean project award (in USD million per project) excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
Source: OECD analysis of Fundstat database, March 2023.

Table 2. Geographical distribution of R&D awards and funding in the Fundstat database, 2019-21

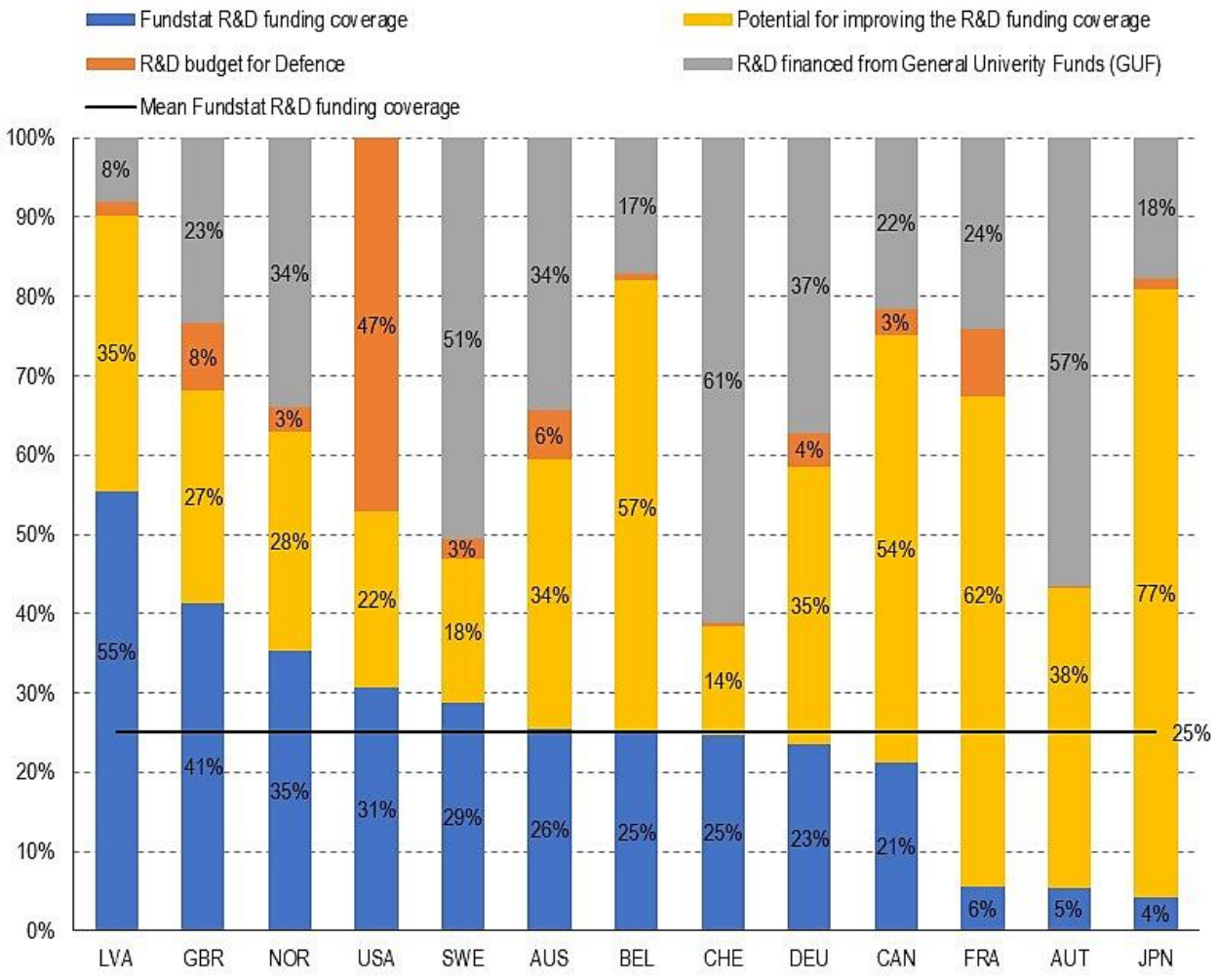
Country/area	R&D projects	(Of which) R&D projects with missing funding information	Share of R&D projects [%]	R&D Funding [USD million]	Share of R&D funding [%]	Mean project award [USD million per project][1]
AUS	6,487	0	1.07	4,718	1.64	0.73
AUT	2,080	0	0.34	759	0.26	0.36
BEL[2]	18,876	1,894	3.11	3,368	1.17	0.20
CAN	78,006	0	12.85	5,130	1.78	0.07
CHE	14,529	464	2.39	4,491	1.56	0.32
DEU	55,798	367	9.19	38,637	13.44	0.70
FRA	5,891	1	0.97	5,166	1.80	0.88
GBR	37,359	16,522	6.15	16,823	5.85	0.81
JPN	100,536	248	16.56	9,368	3.26	0.09
LVA	3,898	879	0.64	401	0.14	0.13
NOR	4,768	917	0.79	2,224	0.77	0.58
SWE	11,403	0	1.88	4,006	1.39	0.35
USA	253,105	8,533	41.69	150,003	52.16	0.61
EC-EU	14,362	508	2.37	42,483	14.77	3.07
Total	607,098	30,333	100.00	287,579	100.00	0.50

Notes:
[1] The calculation of mean project award (in USD million per project) excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
[2] For Belgium (BEL), the total R&D projects and funding is provided by the Belgian Science Policy Office (BELSPO), and by the Department of Economy, Science, and Innovation (EWI) of the Flemish government, and is an estimate based on extrapolating the project information with known funding (90% of R&D projects).
Source: OECD analysis of Fundstat database, March 2023.

The mean project award is about 0.50 USD million, with differences across countries (Table 2). The ensemble of awards by the United States, European Commission (EC)-EU and Germany account for more than 80% of the entire R&D funding covered by Fundstat at this point, significantly more than implied by pure counts of projects. Differences in estimates of funding per project are underpinned by compositional differences within countries regarding the coverage and nature of projects, as well as data reporting structures. For example, the range of eligible project costs covered by awards provided by the French ANR, which often top up the salaries of researchers in higher education and other research institutions, is relatively small in comparison with intramural projects of the US NIH, which in turn cover the entirety of project costs including established researchers. Documenting these differences is a priority for the Fundstat initiative, a task that would be much facilitated by separate reporting of total project costs relative to awards, as this would greatly facilitate international comparisons. Probably because they typically entail multi-country R&D consortia, EC-EU funded projects appear to display the largest awards on average, in the order of USD 3.07 million per project. Co-financing for these EU-funded projects can also exhibit considerable variation by programme.

For the reasons already noted, it is very important to contextualise the analysis of the Fundstat R&D projects in terms of the coverage of data sources within the broader national R&D funding landscape. Figure 1 shows how the Fundstat R&D funding for 2020 compares to estimates reported in the OECD statistics on Government Budget Allocations for R&D (GBARD). Fundstat R&D projects appear to cover the equivalent of approximately 25% of the official R&D funding estimates, a figure that varies across countries. The chart helps explain to what extent this may be due to the relative importance of institutional block funding within the country. It does so by identifying what proportion, shown at the top, is accounted for R&D funded through General University Funds. This helps provide an upper bound of how far the current Fundstat is from the ideal coverage target, although one should also note that general R&D funds may also be given to other R&D performing institutions which may be counted under the General Advancement of Knowledge objective. In the case of Austria, Fundstat coverage is only 5% but 57% of GBARD is accounted for by GUF, so relative coverage is higher than implied. In the case of Switzerland, after accounting for GUF, the 25% coverage represents a high coverage, with a similar picture in Sweden. In the case of Japan, coverage is still very limited although it is worth noting that a significant part of the 2020 GBARD funds were not disbursed to projects.[1]

Figure 1. R&D funding coverage in the Fundstat database as a percentage of Government Budget Allocations for R&D (GBARD), 2020



Notes:
[1] Sorted in order of highest ratio.
[2] The benchmark year is selected to be 2020, which is the first year of the COVID-19 pandemic and the data are complete.
[3] Data on GBARD, R&D budget for Defence, and R&D financed from General University Funds (GUF) are sourced through http://oe.cd/msti, in current USD PPP.
[4] For Canada (CAN), federal expenditures on science and technology are used (link), with the latest available data for R&D financed from General University Funds (GUF) in MSTI from 2016.
[5] USA COVID-19 R&D procurement data are not included in the above ratio calculation because they have been added separately to the analysis as additional COVID-19 specific projects.
[6] For Belgium (BEL), data on total R&D projects and funding is provided by the Belgian Science Policy Office (BELSPO), and by the Department of Economy, Science, and Innovation (EWI) of the Flemish government, and is an estimate based on extrapolating the project information with known funding (90% of R&D projects).
[7] For Japan (JPN), GBARD in 2020 includes: (i) a major University Endowment Fund, which has been reported in reference years 2020 and 2021 as General University Funds (GUF); and (ii) a 10-year green innovation fund, which has been reported in reference year 2020 (see the OECD GBARD Sources and Methods Database, available at https://rdmetadata.oecd.org/).
[8] The EC-EU data cannot be compared to a GBARD value and hence are not on the data coverage dashboard.
Source: OECD analysis of the Fundstat database and OECD R&D Statistics, March 2023.

2.2. COVID-19 R&D analysis methodology

Natural language processing and understanding (NLP) has been around for more than 50 years (Jurafsky and Martin, 2023[25]), accelerating in recent years due to the increase in computational capacity to analyse large corpora of text, and availability of algorithms (e.g. large language models), which have led to the high degree of popularity in text-generative AI tools. In this study, NLP text-based analysis has been applied to the Fundstat project-level R&D funding data. In the area of science, technology and innovation (STI), international analysis efforts have mainly concentrated on field-specific topic extraction mainly for scientific publication and intellectual property rights (IPRs) data (Aristodemou and Tietze, 2018[26]). There are limited precedents on R&D funding data, which mainly reflect challenges with data availability (Annapureddy et al., 2020[27]; Abadi, He and Pecht, 2020[28]; Yamashita et al., 2021[22]). Data analysis methods must be designed and adapted to take into consideration both objectives and data availability, with the aim to map the directionality of government support for COVID-19 R&D across topical areas.

2.2.1. Retrieval of COVID-19 R&D projects

COVID-19 R&D projects are initially retrieved through 'key term' matching. A 'key term' tagging approach is adopted to identify COVID-19 R&D projects in the text corpus of projects for each data source (Annex A). Key term matching presents several challenges, since there is at present no consensus on a standard set of key terms that comprehensively and unambiguously represent COVID-19 R&D. Such a set is bound to be specific to different corpora and vary over time. Failing to capture all relevant key terms risks overlooking many relevant COVID-19 projects, thus underestimating the total. There is both a risk of overlooking projects that do not contain the chosen key terms if the list is too narrow; and a risk that the title and abstract of project applications do not contain sufficient information from which to retrieve terms that might otherwise be presented in the full but not accessible project proposal. A key term approach can yield the opposite result when using an excessive and broad range key terms thus opening the way to including projects not necessarily connected in their substance to the pandemic. Key term matching from a pre-defined menu of COVID-19 terms does not ensure that the research is ultimately COVID-19 related. Potential key COVID-19 terms can feature in abstracts purely as elements of context descriptions. To address these problems, this study combines key term selection for 'candidate' COVID-19 project retrieval with (a) the application of a 'contextual error' detection/classifier to identify and remove contextual COVID-19 project, which includes manual inspections of contextual COVID-19 projects, and (b) the filtering of projects with insufficient content and that are not possible to classify in relation to COVID-19 relevance.

2.2.1.1. Selection of COVID-19 key terms and retrieval of 'candidate' projects

With no formal training database, the key term selection process requires an initial (base) list of terms that are COVID-19 relevant. The Medical Subject Headings (MeSH) taxonomy of the U.S. National Library of Medicine, which contains headings for COVID-19 and SARS-CoV-2, is used to obtain a baseline of 10 key terms. This structure does not provide a comprehensive source of all potentially relevant COVID-19 terms. Instead, it provides a basic structure for the categorisation of research activity and outputs in the health domain.[2] To reduce the risk of omitting relevant COVID-19 R&D projects, additional key terms that provide relevant signals of COVID-19 related research activity were retrieved using a Word2Vec methodology from 3 relevant data corpora: (i) Elsevier's Scopus Custom database, (ii) USA National Library of Medicine PubMed publications, and (iii) the World Health Organisation's COVID-19 database (OECD, 2021[5]; Yamashita et al., 2021[22]; Mikolov et al., 2013[29]). The Word2Vec methodology identifies associations of 'similar' words to the COVID-19 base set of key terms, resulting in an extended list of 24 key terms.[3] Finally, all the key terms were translated in all languages across databases.[4] The consolidated set of 34 key terms (Table 3) was used to search within the title and abstract of the Fundstat database, extracting 16,346 'candidate' COVID-19 R&D projects.[5]

Table 3. Identification and expansion of COVID-19 key-terms

Origin	Set	# key terms[1]	Key terms[2,3]
Medical Subject Headings (MeSH)	Base	10	Coronavirus disease, covid, covid19, ncov, novel coronavirus, sars coronavirus 2, sars cov 2, severe acute respiratory syndrome coronavirus 2, wuhan coronavirus, wuhan seafood market pneumonia virus
PubMed, CDC, and Scopus	Extension	24	Coronavirus2019, coronaviruscovid, ncov2019, ncovid, new coronavirus, novel coronaviruses, co vid, covd, recently discovered coronavirus, sar cov 2, sars co v2, 2019ncov, corona virus disease, cov 2, newly identified coronavirus, novel betacoronavirus, novel corona virus, sars cov2, sarscov 2, sarscov2, coronavirusdisease, newly emerge coronavirus, newly discovered coronavirus, previously unknown coronavirus
	Total	34	

Notes:
[1] A base set of key terms is identified from the Medical Subject Headings (MeSH), including the terms 'COVID-19' and 'SARS-COV-2'. The base terms are expanded using the Word2Vec similarity method, originating from other corpora (PubMed publications, WHO COVID-19 database, and the COVID-19 Scopus publications) (OECD, 2021[5]; Yamashita et al., 2021[22]; Mikolov et al., 2013[29]).
[2] The key terms and the text used for retrieval have been cleaned and processed by lowercasing and lemmatising. Lemmatisation involves grouping together the different forms of a word and analysing them as a single element (known as 'lemma').
[3] The list of key terms is not exhaustive, and other COVID-19 relevant key terms and translation variations could exist.
Source: OECD Fundstat infrastructure, March 2023.

2.2.1.2. Database processing for selecting relevant COVID-19 projects

To reduce the potential of false positives, e.g., projects being inappropriately tagged as COVID-19-related by the retrieval method, a supervised machine learning model was developed to identify projects where instances of key terms on COVID-19 are of a purely contextual nature.[6] These instances are defined as projects where COVID-19 terms are used to allude, for instance to, COVID-19 disruption influencing R&D work in other areas, or contributions of R&D projects to recovery from the economic consequences of the COVID-19 crisis or developing competitiveness through R&D in a post-COVID-19 world. In other words, COVID-19 key terms are *mentioned* in the project description but *do not reflect the object of the funded R&D*. The contextual error classifier is built on a random sample of manually tagged projects. The OECD team manually reviewed and labelled a random sample of 1,300 candidate projects into likely COVID-19 R&D projects and purely contextual projects. In this sample, 7% of projects were identified as being in the latter group. Using this dataset, a machine leaning classifier is built and used on the full set of 'candidate' COVID-19 R&D projects, to identify further 2,287 projects as contextual.[7]

Table 4 shows a list of 5 examples involving contextual projects that the algorithm places under such category. All cases, but one, illustrate contextual references to COVID-19 and would be correctly eliminated. The third example indicates a potential 'prediction' error, as the R&D on aerosol systems is of direct relevance to COVID-19 but it may have been predicted as 'contextual' because of the relatively small residual weight of the COVID-19 language around the COVID-19 allusion. As a final step, the analysis adopts a criterion to remove projects with very limited textual content for these may jeopardize the implementation of the prediction of contextual error and the topic modelling described in the next section. Projects with less than 5 meaningful words (excluding stopwords) after data cleaning and post-processing (Annex B.1) are removed from the set of selected COVID-19 projects (Annex B.2).[8]

The resulting COVID-19 R&D project database comprises 11,886 projects that account for close to USD 12.5 billion (Table 5). Within the initial 16,346 candidate projects, there are 2,082 projects with limited content information. A slightly higher number of projects is excluded based on being classified as having purely contextual references to COVID-19 key-terms. Awards for such projects appear to be on average larger, namely USD 1.34 million for excluded projects versus USD 1.20 million for retained ones.

Table 4. Examples of candidate R&D projects machine-tagged as having contextual COVID-19 descriptions

Country	Database	Title	Abstract (relevant excerpt)	R&D Funding [USD million]	COVID-19 status
USA	NIH	Statistical and Data Management Center (SDMC): Microbicide Trials Network	...To address these delays, the MTN Statistical and Data Management Center (MTN SDMC) is requesting a one-year extension of funding… which have been **interrupted due to the COVID-19 epidemic**...	3.47	Fits contextual definition – correctly excluded
GBR	GtR_Innovate UK	SPRITE 2 - Sustainable Plastics Recycling Innovation by Tagging Electronically	Plastic packaging waste is a $80Bn global opportunity according to the World Economic Forum**… Progress on all these has seen a significant setback during the COVID-19**...	1.75	Fits contextual definition – correctly excluded
GBR	GtR_EPSRC	University of Bristol Core Equipment Award 2020	...A scanning Aerodynamic Aerosol Classifier (s-AAC), enabling the elucidation of mechanisms behind accelerated reactions in aerosol systems…and **providing pioneering insights into the role of aerosols in disease transmission, including Covid-19**...	0.99	Exclusion may not be correct. Somewhat tentative relevance
SWE	SWECRIS_Forte	National project on the effectiveness of tobacco interventions	...**Due to Covid-19 it has been impossible to follow the original plan**, as other tasks have been prioritised in the clinics and therefore, we have extended the previous…	0.52	Fits contextual definition – correctly excluded
FRA	ANR-dos	Epigenetic immune subversion in Leishmania macrophage infection	...overcome important scientific, translational, and technical barriers **relevant to other intracellular infections, such as tuberculosis, candidiasis, AIDS or COVID**…	0.47	Appears to fit contextual definition. Very tentative relevance

Notes:
The table provides a list of examples where the classifier algorithm trained to detect spurious contextual references to COVID-19 results in a recommendation to exclude from the analysis. The GBR_GtR_EPSRC example appears to indicate a possible prediction error, namely a potentially relevant COVID-19 project (relevance of aerosols work) classified as contextual and therefore removed from the analysis.
Source: OECD analysis of Fundstat database, March 2023.

Table 5. Retention of COVID-19 R&D candidate projects following data cleaning, 2019-21

	R&D projects	(Of which) projects with missing funding data	R&D funding [USD million]	Mean project award [USD million per project][1]
Total retrieved 'candidate' COVID-19 R&D projects	16,346	2,823	16,604	1.23
(-) 'Candidate' projects with contextual descriptions[2]	2,378	429	2,605	1.34
(-) 'Candidate' projects with limited content[3]	2,082	980	1,411	1.28
Total retained COVID-19 R&D projects	11,886	1,414	12,588	1.20

Notes:
[1] The calculation of the mean project award (in USD million per project) excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
[2] 'Candidate' projects with contextual descriptions are identified by a prediction classifier as having purely contextual references to COVID-19 and a manual inspection.
[3] 'Candidate' R&D projects with limited content are projects that after data cleaning and post-processing (e.g., removal of stopwords, removal of language-specific stopwords, and removal of common phrases and highest frequency words), there is limited content (less than 5 meaningful words) in the text to infer any type of activity (Annex B.1 and B.2).
Source: OECD analysis of Fundstat database, March 2023.

2.2.2. Topic modelling analysis of COVID-19 R&D projects

The corpus of COVID-19 R&D projects has been analysed to determine the most salient COVID-19 R&D topics by examining their titles and abstracts. Machine learning-based topic modelling methods, specifically the BERTopic library based on the Bidirectional Encoder Representation from Transformers (BERT) model (Grootendorst, 2022[30]) have been used to this end. This model utilises transformer language models, also known as large language models (LLM), to handle large and diverse datasets without extensive text pre-processing, in combination with c-TF-IDF[9] criteria to create dense clusters that allow for easily interpretable topics while preserving important words in the topic descriptions.

The process starts by combining the title and abstract into a single data column for all the 11,886 retained COVID-19 R&D projects (Table 5). The text is pre-processed, cleaned, and standardised for all languages to address the diverse nature of R&D project descriptions and to eliminate all possible ambiguities due to the projects' diverse set of sources[10]. To maintain the semantic relationships of the original language in which the project text is written (Annex B.3), the analysis relies on the transformation of the text using a pre-trained multilingual sentence embedding 'distiluse-base-multilingual-cased-v2' (Reimers and Gurevych, 2019[31]). This is followed by the dimensionality reduction of the vector embeddings, using a tool, UMAP, that preserves local relatedness of similar data points. The analysis then employs a density-based hierarchical clustering method, HDBSCAN, to create topic clusters based on the dimensionality-reduced embeddings. HDBSCAN is a soft clustering technique that assigns each project a vector probability rather than a cluster label. This approach allows for the possibility of projects to belong to multiple clusters, enabling the identification of clusters of different shapes, without making any assumptions about the expected structure of the topic clusters (McInnes, Healy and Astels, 2017[32]). Once the clusters have been generated, a bag-of-words representation is created for each cluster, and the c-TF-IDF criterion is used to identify the most important and representative words within a cluster.

To help validate and interpret the topic model and topic representations, the study has incorporated comparative analysis with alternative sources (Section 4) such as the COVID-19 Scopus publications and the COVID-19 Research Project Tracker by UKCDR & GloPID-R (OECD, 2021[5]; Bucher et al., 2023[14]; UKCDR & GloPID-R, 2023[15]). The COVID-19 tracker data, which maps and analyses global COVID-19 funding to the priorities identified in the WHO Coordinated Global Research Roadmap: 2019 Novel Coronavirus, allows the comparison of machine- and expert manual-based classifications of COVID-19 R&D, while also providing a basis for assessing the performance of Fundstat as a data source. The expert-based prediction of WHO topics provides the training basis for building a new classifier that this work applies to the COVID-19 R&D projects from the Fundstat database, to predict the WHO labels on this dataset[11].

The same process of COVID-19 R&D retrieval has been applied to retrieve COVID-19 scientific publications from the Elsevier Scopus Custom database at the OECD. For the contextual removal of Scopus publications, a contextual classifier is constructed, which is firstly trained on the R&D contextual projects, followed by a set of publications that have been manually tagged by the OECD STI team. The trained topic model on R&D projects and the COVID-19 tracker WHO topic predictive models have then been used to infer the COVID-19 related topics in the scientific publication corpus.

3 Features of COVID-19 R&D funding

3.1. Aggregate estimates of COVID-19 R&D project funding

3.1.1. COVID-19 R&D project counts and funding

Using the methodology described in sub-section 2.2.1, the database of retained COVID-19 R&D projects contains 11,886 project awards corresponding to total funding worth about USD 12.59 million, with mean project funding award of USD 1.20 million (Table 6).

Table 6. Fundstat estimates of COVID-19 R&D, 2019-21

	FY2019	FY2020	FY2021 (partial)[2]	Total
Number of R&D projects[1]	150	6,648	5,088	11,886
(Of which) R&D projects with missing funding information	23	809	582	1,414
Total funding [USD million]	69	6,667	5,852	12,588
Mean R&D project funding award [USD million per project][3]	0.55	1.14	1.30	1.20

Notes:
[1] Analysis limited to the retained COVID-19 R&D projects.
[2] FY2021 is partial because of data reporting structures and data availability across different countries.
[3] The calculation of mean project award (in USD million per project) excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
Source: OECD analysis of Fundstat database, March 2023.

The annual breakdown of financial award commitments appears to indicate a slight slowdown in COVID-19 R&D funding in 2021 compared to 2020. However, this may be due in part to the fact that published funding databases for 2021 are not as complete as those for previous years, given grant data publication lags. With some degree of caution, it is possible to note that there appears to be a marked increase in the average size of COVID-19 funded projects (in terms of the size of the funding award provided) over the three-year observation period. The distribution of retained COVID-19 funded projects across countries and the EC-EU (Table 7) indicates a leading role in funding for the United States (68%), followed by Germany (9%), the EC-EU (8%), and the United Kingdom (6%), which reflects in large part the underlying scope and coverage features of the Fundstat database described in Section 2.1.

3.1.2. COVID-19 R&D funding in the broader landscape

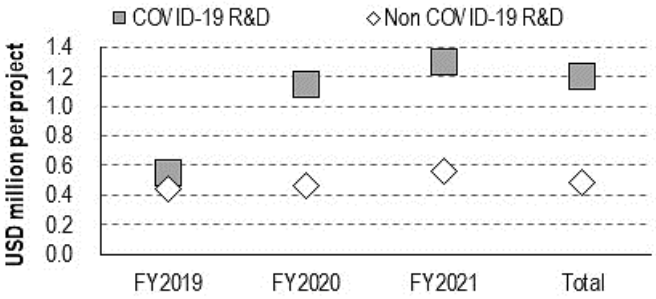
COVID-19 R&D projects constitute approximately 2% of the total number of projects in the Fundstat database, while they account for more than 4% of the R&D funding. This implies systematic disparities in the average funding per project for COVID-19 R&D in comparison to non-COVID-19 R&D from 2020 onwards. Average funding for COVID-19 R&D projects has more than doubled, while average funding for non-COVID-19 R&D projects remained relatively steady (Figure 2), suggesting that growth in R&D funding for COVID-19 has been relatively more concentrated in larger awards.

Table 7. Geographical distribution of Fundstat COVID-19 R&D funding, 2019-21

Country/area	R&D projects[1]	(Of which) R&D projects with missing funding information	Share of R&D projects [%]	R&D Funding [USD million]	Share of R&D funding [%]	Mean project award [USD million per project][2]
AUS	94	0	0.79	103	0.82	1.10
AUT	25	0	0.21	10	0.08	0.38
BEL[3]	363	86	3.05	49	0.39	0.18
CAN	1,148	0	9.66	282	2.24	0.25
CHE	220	6	1.85	90	0.71	0.42
DEU	647	258	5.44	1143	9.08	2.94
FRA	343	0	2.89	71	0.56	0.21
GBR	1,581	175	13.30	796	6.32	0.57
JPN	978	80	8.23	387	3.07	0.43
LVA	64	18	0.54	13	0.10	0.28
NOR	15	4	0.13	4	0.03	0.37
SWE	378	0	3.18	92	0.73	0.24
USA[4]	5,881	782	49.48 (*13.55)	8534	67.79 (*20.29)	1.67
EC-EU	149	5	1.25	1017	8.08	7.06
Total	**11,886**	**1,414**	**100.00**	**12,588**	**100.00**	**1.20**

Notes:
[1] Analysis limited to the retained COVID-19 R&D projects. Shares of countries reflect the coverage of R&D projects in the Fundstat database (Section 2.1).
[2] The calculation of mean project award (in USD million per project) excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
[3] For Belgium (BEL), the total R&D projects and funding is provided by the Belgian Science Policy Office (BELSPO), and by the Department of Economy, Science, and Innovation (EWI) of the Flemish government, and is an estimate based on extrapolating the project information with known funding (90% of R&D projects).
[4] *=of which Federal Procurement project data, awarded as contracts. Federal procurement R&D contract awards identified by US authorities as connected to COVID-19, and covering multiple government agencies. This is based on the USA COVID-19 Contract Obligation Tracking Dashboard, which provides all government-wide emergency acquisition spending for the COVID-19 pandemic.
Source: OECD analysis of Fundstat database, March 2023.

Figure 2. Mean project funding award for COVID-19 R&D vs. non COVID-19 R&D

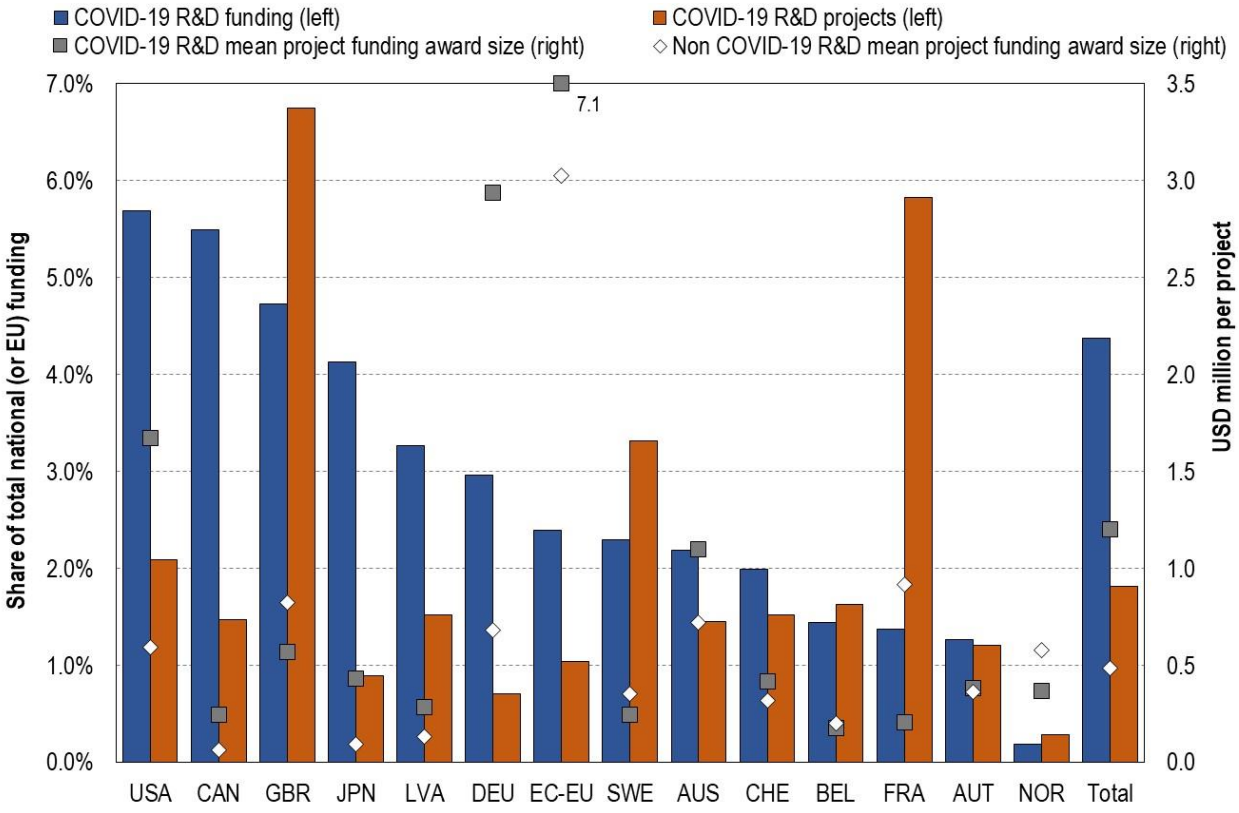


Notes:
[1] Analysis limited to the retained COVID-19 R&D projects. Calculations are based on Table 1 and Table 6.
[2] The calculations of mean project funding award for COVID-19 R&D and non-COVID-19 R&D exclude projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications.
Source: OECD analysis of Fundstat database, March 2023.

Marked geographical differences can be appreciated when looking at the percentage share of COVID-19 R&D funding relative to the funding covered within Fundstat (Figure 3). Funding allocated to COVID-19 R&D in the United States, Canada, United Kingdom, and Japan represent more than 4% of total funding

within those countries (Florio, Gamba and Pancotti, 2023[33]). The domestic share of numbers of COVID-19 R&D projects in United Kingdom, Sweden, Belgium, and France exceed the corresponding share of COVID-19 R&D funding. The case of France is noteworthy because nearly 6% of all Fundstat projects are identified as COVID-19 related while these correspond to less than 1.5% of funding. It is important to note that identified funding per project is not necessarily proportional to actual project size, particularly if major alternative institutional funding streams are available. EC-EU funded projects have the highest average funding levels for COVID-19 R&D, more than twice the funding for non-COVID-19 projects. Germany and the United States follow. In the case of Germany the mean project size for COVID-19 R&D is about six times the non-COVID-19 R&D (Figure 3).

Figure 3. COVID-19 R&D within national/EU R&D covered in Fundstat, 2019-21

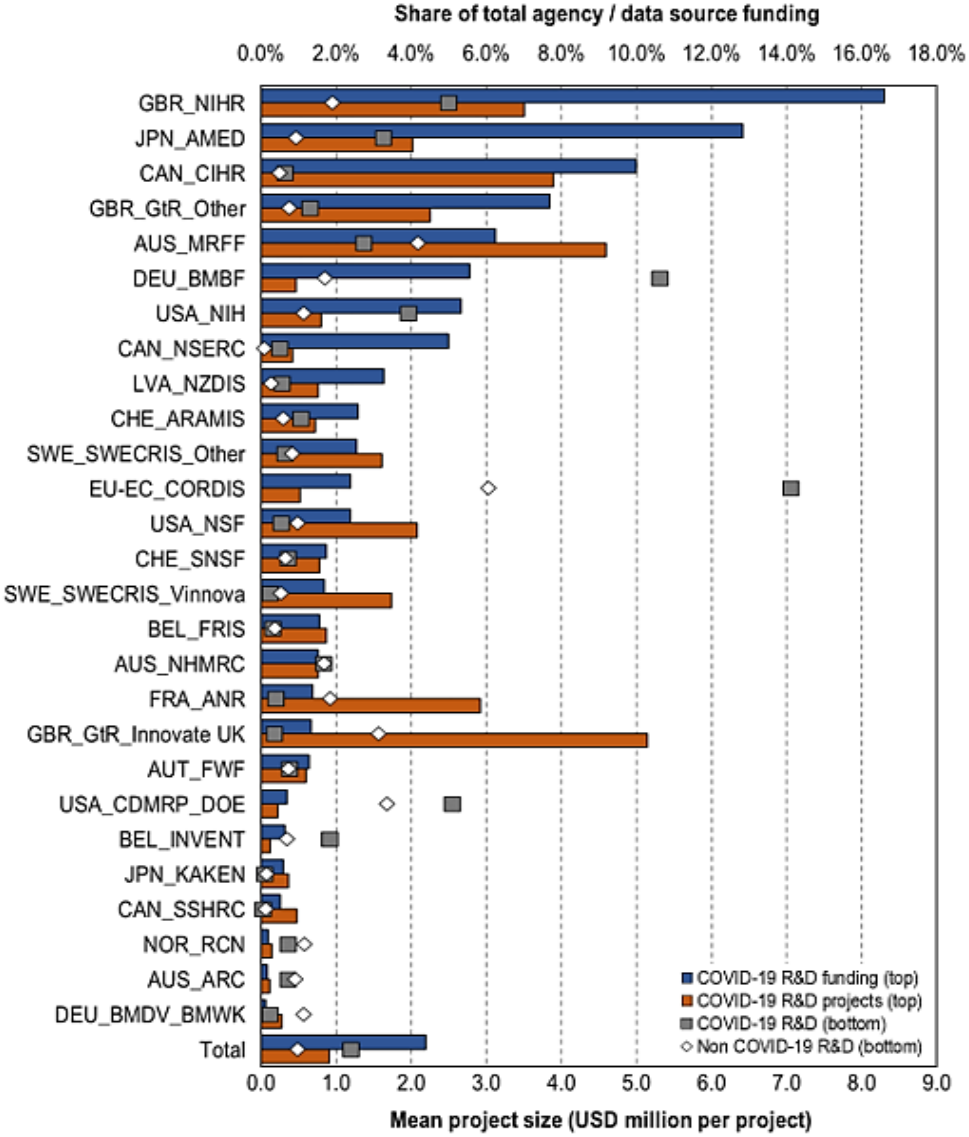


Notes:
[1] Analysis limited to the retained COVID-19 R&D projects. Differences across countries are indicative of differences in the coverage of Fundstat for each one of them (Section 2.1). Calculations are based on Table 2, and Table A.1.
[2] The calculation of the share of COVID-19 R&D to the total national (or EU) R&D excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
[3] The estimates for the United States include COVID-19 R&D procurement in the numerator but do not include all R&D procurement in the denominator since only the procurement of R&D services tagged as COVID-19 by the US authorities has been added to Fundstat.
[4] The calculations of mean project award for COVID-19 R&D and non-COVID-19 R&D exclude projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding). Project award size refers in this figure to funding award per project, as actual information on the size of projects is not available.
Source: OECD analysis of Fundstat database, March 2023.

Figure 4 illustrates the contribution of COVID-19 funding within the different funding agencies or the data sources in Fundstat. Data at this level are reported for individual funders but on occasions these data are only available pooled at source or it is convenient for presentational purposes to combine them to keep

the number of entities on display manageable. Specialised medical/health R&D funding agencies display the highest COVID-19 funding intensities. The UK National Institute of Health Research (labelled as GBR_NIHR) has the highest share of COVID-19 R&D funding, while Canada's Institute of Health Research (CAN_CIHR) has the highest count-based share of COVID-19 R&D projects.

Figure 4. Estimates of COVID-19 R&D within funding agency / source covered in Fundstat, 2019-21



Notes:

[1] Analysis limited to the retained COVID-19 R&D projects. The DEU_GEPRIS and the USA_COVID-FPDS data, which have been identified by the authorities connected to COVID-19 are not included. Differences across funding agencies and data sources are indicative of differences in the coverage of Fundstat for each one of them (Section 2.1). Calculations are based on Table 2, Table 7 and Table A.1.

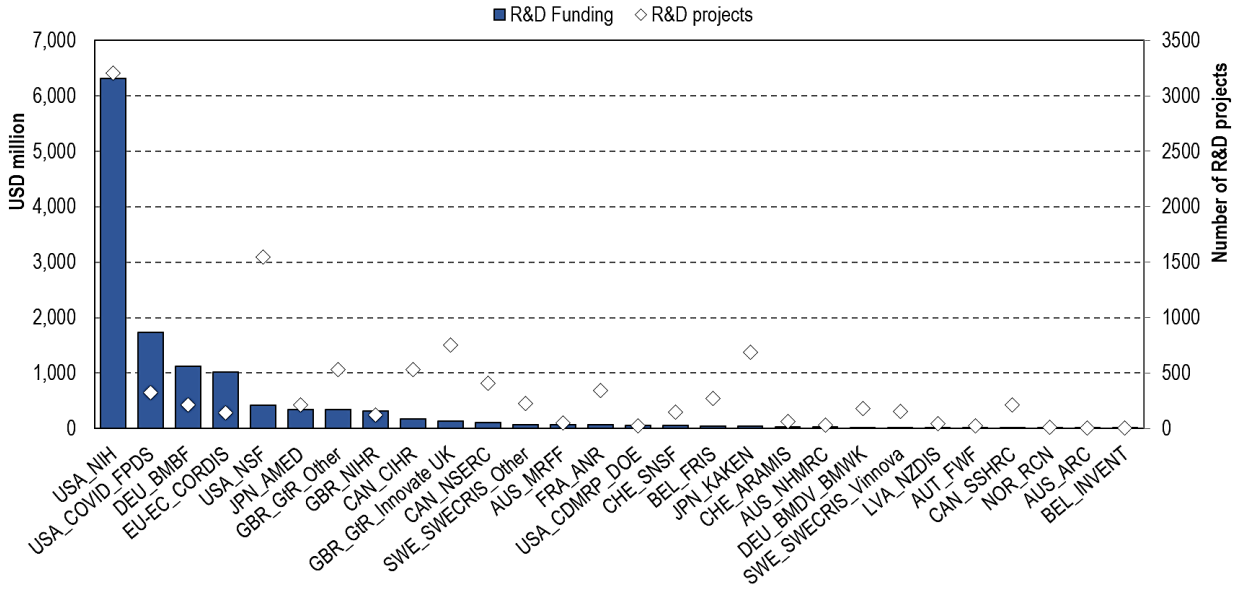
[2] FY2021 is partial because of data reporting structures and data availability across different countries. Specifically, FY2021 data are unavailable for BEL_INVENT, CAN_CIHR, CAN_NSERC, CAN_SSHRC, DEU_COVID-GEPRIS, JPN_AMED, and NOR_RCN.

[3] The calculation of the share of COVID-19 R&D to the total national (or EU) R&D excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding). Project size refers in this figure to funding award per project, as actual information on the size of projects is not available.

[4] The calculations of mean project award for COVID-19 R&D and non-COVID-19 R&D exclude projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).

Source: OECD analysis of Fundstat database, March 2023.

Figure 5. COVID-19 R&D by funding agency/data source



Notes:
[1] Analysis limited to 10,472 COVID-19 R&D projects in the Fundstat database with known funding information.
[2] FY2021 is partial because of data reporting structures and data availability across different countries. Specifically, FY2021 data are unavailable for BEL_INVENT, CAN_CIHR, CAN_NSERC, CAN_SSHRC, DEU_COVID-GEPRIS, JPN_AMED, and NOR_RCN.
[3] The labels include the country and the funding agency/data source (Annex A).
[4] The analysis does not include the DEU_COVID-GEPRIS database as there is no funding information available for those projects.
[5] GBR_Other includes ESRC, NERC, EPSRC, BBSRC, MRC, AHRC, STFC, NC3Rs, and UKRI.
[6] SWE_SWECRIS_Other includes FBEES, IFAU, VR, Forte, Formas, SHLF, RJ, and SWEA.
Source: OECD analysis of Fundstat database, March 2023.

Information on the total number projects and funding awards for COVID-19 R&D by agency or source, contributing to the country totals presented in Figure 3, are available in Figure 5. US sources feature as major funders in terms of counts of projects and total COVID-19 funding. US R&D procurement funding is separately itemised relative to grant awards by specific funders like NIH or NSF. DEU_BMBF and EU-EC_CORDIS are the largest non-US COVID-19 funding sources. The relative positions of agencies and sources differs if one looks at counts of projects or funding, in what appears to reflect systematic differences in the types of projects and initiatives funded.

3.2. Directionality of COVID-19 R&D funding

3.2.1. Funding analysis by machine-generated topic and topic cluster

The implementation of COVID-19 topic model analysis described in Section 2 returns a total of 34 COVID-19 funding topics (Table 8). For each of the 11,886 COVID-19 R&D projects analysed, the model delivers a matrix with the distribution of probabilities for each project "belonging" to each one of the 34 topics. In a perfect-world topic model, the sum of probabilities of the machine-generated topics should equal unity. However, as it is common in topic modelling analysis, this is not the case in the COVID-19 topic model as there are projects that have relatively low probability of assignment into the 34 cohesive topics generated by the model. "Niche" topics are removed with a restriction of a minimum topic cluster size of 40 projects. This constraint is intended to remove ad hoc topics that lack a minimum critical mass and that may distort the thematic mapping of COVID-19 projects. The probability that projects relate to 'non-specific' topics,

represented by the residual probability, provides an indication of the model's effectiveness in comprehensively capturing projects into cohesive and distinctive topics. Across all projects, the average residual probability of assignment to non-specific topics is just over 13%, a relatively small share but not trivial amount. This percentage is rather skewedly distributed across projects. The median probability of a project being allocated to non-specific topic is 0% (i.e., most projects have zero residual topic content) while 75% of projects have a residual probability below 20%.

The probability matrix generated by the topic model also serves as the basis for fractional project/award counting and funding allocation. As projects may belong with some probability to multiple topics, fractional counts and total funding amounts estimated for each topic are obtained by apportioning project funding amounts across topics using topic probabilities as the attributing shares for both counts and funded amounts. As shown in Table 8, the fractional equivalent of 10,298 out of 11,886 projects (87%) can be assigned to the 34 topics identified by the topic model, corresponding to a total funding amount of approximately USD 11 billion out of 12.59 billion.

Examination of projects with a high residual content that cannot be categorized into the 34 topics generated by the model indicates a high incidence of support for business R&D for economic and technological resilience (see section 3.2.4 and Figure 15) at the boundaries of the COVID-19 relevance definition operationalised in this study. Because of differences in the average project award across documents with different residual content, on a fractional calculation basis, COVID-19 R&D project content assigned to the core body of 34 well-defined COVID-19 topics displays a slightly higher average funding amount per project than content allocated to other non-specific topics.

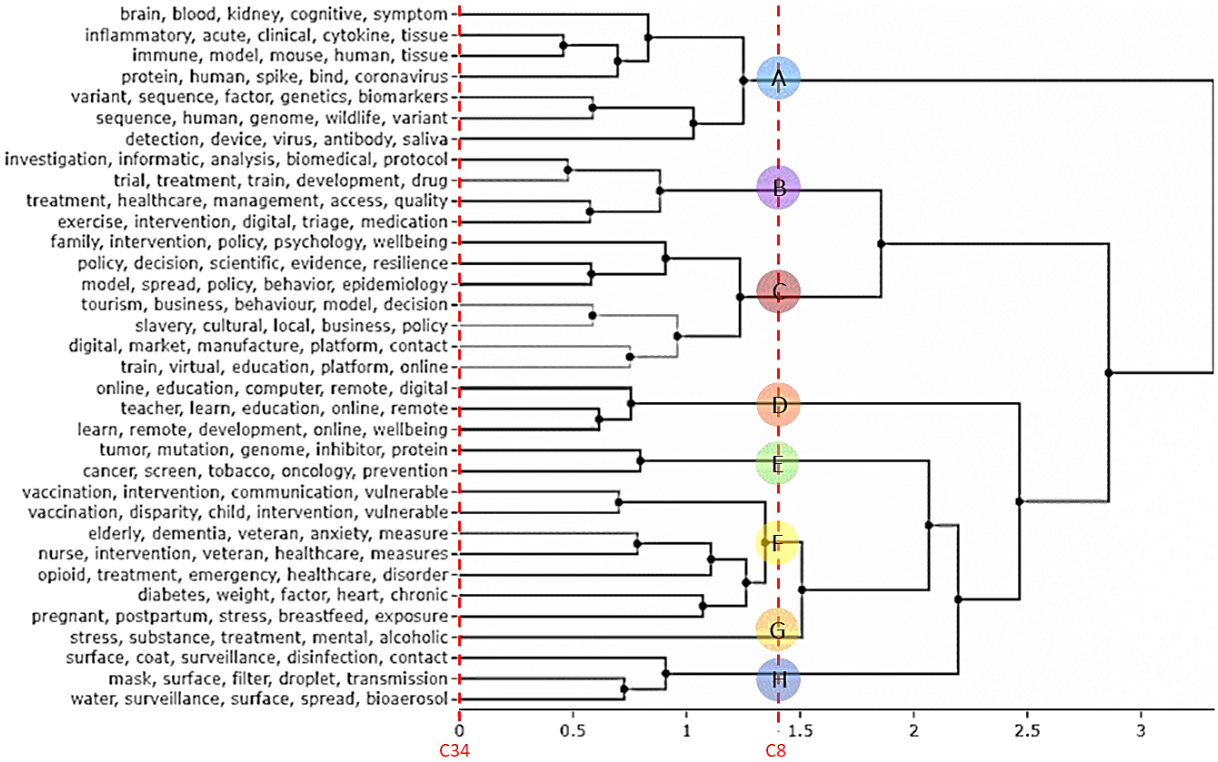
Table 8. COVID-19 R&D into defined (C34) and non-specific topics

Allocation of COVID-19 R&D projects (2019-2021) to [1]	Fractional count of R&D projects	(Of which) R&D projects with missing funding information	Fractionally allocated funding [USD million]	Implied mean project award [USD million per project][2]
34 topics arising from the topic model with minimum cluster size restriction [CORE C34]	10,298	1,211	10,999	1.21
Other non-specific topics (residual allocation)	1,588	203	1,589	1.15
[Residual percentage]	[13.3%]	[14.3%]	[12.6%]	
Total	11,886	1,414	12,588	1.20

Notes:
[1] Analysis limited to the retained COVID-19 R&D projects from Table 5. Breakdown based on the topics generated by the topic model and project-based probabilities estimated for each category.
[2] The calculation of mean project award (in USD million per project) excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
Source: OECD analysis of Fundstat database, March 2023.

A hierarchical representation of topic clusters enables an exploration of the structure of topics generated by the model. Figure 6 displays a hierarchical dendrogram based on topic similarities, where all 34 topics identified from the topic model are shown on the left with their 5 most salient words serving as detailed topic headings (see Annex C). Clusters of varying granularities can be extracted by selecting similarity thresholds. The selection of a higher similarity threshold results in more narrowly defined topic clusters. By drawing a vertical line through the dendrogram at the line marked as C8, it is possible to identify 8 distinct topic clusters. By moving to the left, there is an increase in topic granularity converging eventually with the C34 categories.

Figure 6. Hierarchical dendrogram of machine generated COVID-19 topics and topic clusters



Notes:

[1] Analysis limited to the retained COVID-19 R&D projects from Table 5.

[2] The horizontal axis represents the result of a Ward linkage function on the cosine similarity distance of the c-TF-IDF matrix of topic embeddings, with increasing similarity moving from right to left (less distance).

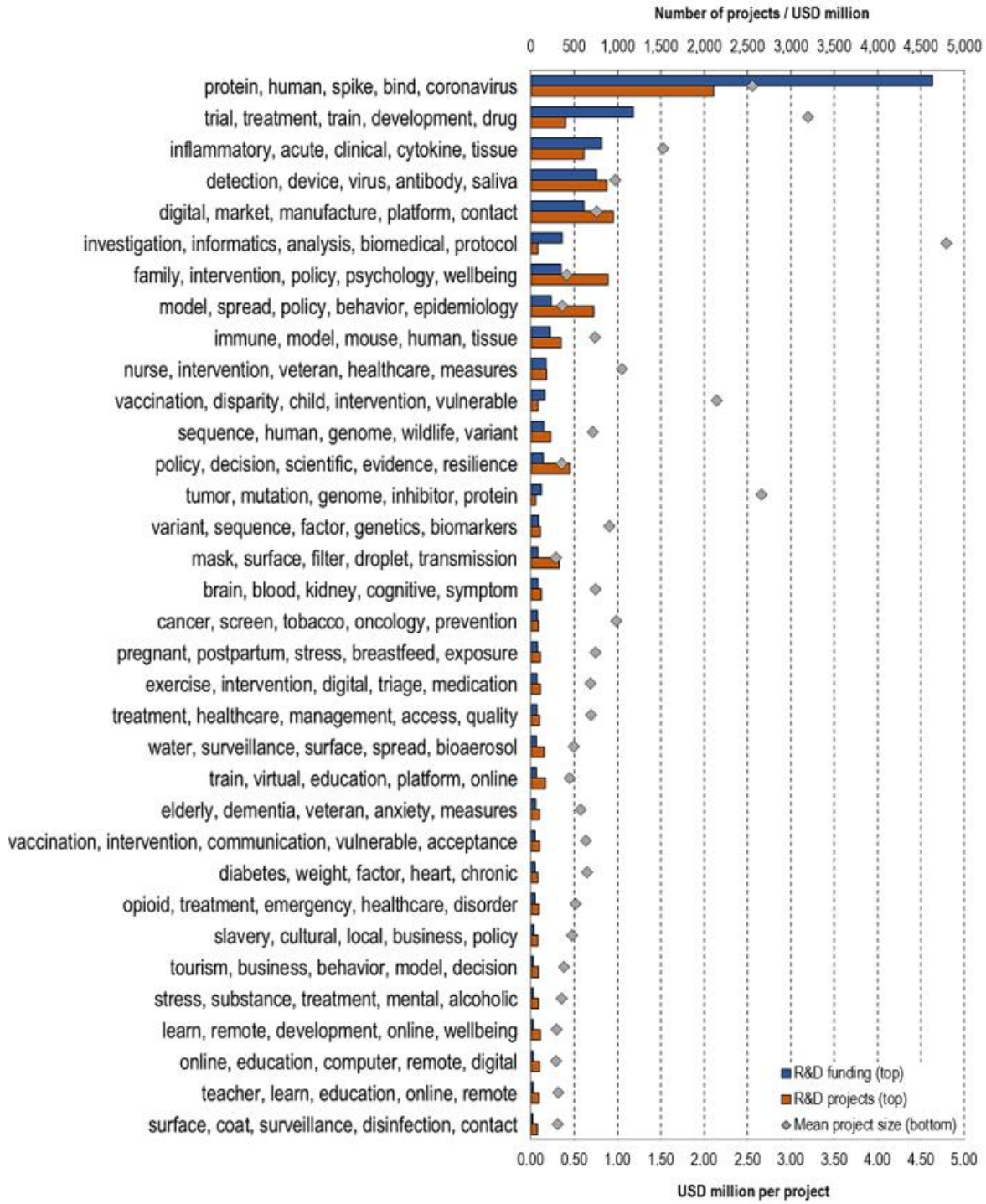
[3] C34 reflects the 34 topics produced by the topic model, represented by the top 5 salient words with the highest within topic importance scores (see Annex C for a complete characterisation of the C34 topics).

[4] C8 reflects the 8 high level topic clusters (see Table 9 for a complete characterisation of the C8 clusters).

Source: OECD analysis of Fundstat database, March 2023.

The fractional allocation of R&D projects and funding to the 34 most granular topics is summarised Figure 7, revealing a preeminent role for biomedical R&D topics The topic on 'protein, human, spike, bind, coronavirus' accounts for the largest allocated volume of R&D funding and number projects. Its salient keywords, such as 'protein' and 'spike', are reflective of R&D on the structure and function of the coronavirus, which are also relevant for the development of vaccines and treatments (Gaviria and Kilic, 2021[34]). The topic with the second-highest level of R&D funding has 'trial, treatment, train, development, drug' as salient words, focusing on the capabilities needed for clinical trials for COVID-19 treatments, secure evidence, and provide training. The topic with the highest level of implied average funding award has 'investigation, informatic, analysis, biomedical, protocol' as top words, apparently connected with the design, implementation, and analysis of protocols for biomedical research on COVID-19 (Baden et al., 2021[35]; Xu et al., 2020[36]).

Figure 7. Estimates of COVID-19 R&D by machine generated "C34" topics



Notes:
[1] Analysis limited to the retained COVID-19 R&D projects from Table 5. Breakdown based on the topics generated by the topic model.
[2] C34 reflects the 34 topics produced by the topic model, represented by the top 5 salient words with the highest within topic importance scores (see Annex C for a complete characterisation of the C34 topics).
[3] The calculation of mean project award (in USD million per project) excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
Source: OECD analysis of Fundstat database, March 2023.

Table 9 shows the machine-generated 8 cluster categories that aggregate the 34 more granular COVID-19 R&D topics according to a fixed similarity threshold. For interpretation purposes, these have been manually assigned a 'descriptive label'[12]. The top 10 salient words per cluster with the highest degree of within-cluster similarity are displayed. Annex C provides examples of COVID-19 R&D projects allocated to each of the C8 topic clusters.

Table 9. COVID-19 R&D "C8" topic clusters and their most salient terms

Fundstat COVID-19 R&D clusters and labels[1,2]	Top 10 salient words per cluster
A. Coronavirus understanding, therapeutics and vaccine development	protein, vaccine, human, antibody, coronavirus, model, respiratory, therapeutic, sequence, detection
B. Platforms and capabilities	medical, train, management, rehabilitation, access, intervention, platform, digital, evidence, model
C. Epidemiology and social interventions	model, policy, intervention, family, economic, access, worker, digital, distance, analysis
D. Digital access and online education	education, online, teach, remote, digital, skill, engage, virtual, access, platform
E. Cancer (Screening and treatment)	cancer, screen, trial, disparity, intervention, delivery, chemotherapy, tobacco, diagnosis, prevention
F. Public healthcare and other groups at risk	vaccine, intervention, maternal, diabetes, factor, healthcare, treatment, stress, exposure, measure
G. Mental health and addictions	stress, consumption, treatment, substance, mental, harm, factor, behavioral, exposure, alcoholic
H. Environmental detection, transmission, and protection	surface, mask, material, water, transmission, droplet, particle, protection, respiratory, environment

Notes:
[1] Analysis limited to the retained COVID-19 R&D projects from Table 5. Top 10 salient words are ordered by highest degree of within-cluster importance
[2] The OECD-STI team have labelled the C8 cluster topics by reviewing: (i) the top 10 salient words; (ii) the complementarity of these clusters with the WHO priority topics; and (iii) the top funded R&D projects per cluster.
Source: OECD analysis of Fundstat database, March 2023.

Cluster A, '**Coronavirus understanding, therapeutics, and vaccine development**', captures key elements of **biomedical R&D on COVID-19**, with 'protein' and 'vaccine' as the 2 most salient words, followed by 'human, antibody, coronavirus, model, respiratory, therapeutic, sequence, detection'. This cluster is broadly concerned with aspects of COVID-19 human virology and immunology, development of vaccines and therapeutics. It is worth noting that terms related to the development of these key solutions are not present among the top 5 important terms that help characterize the 7 detailed topics within this cluster (Annex C). Such applications are transversal to the extracted topics since they are only captured by conducting the analysis at a more aggregate level.

Cluster B, '**Platforms and capabilities**', captures funding for platforms and networks in the biomedical space with a rather transversal nature. These include elements such as trials, large population studies, bioinformatics, development of protocols, training, and other capabilities for R&D on COVID-19. These projects refer to investments made to bring together actors and provide them with the capabilities to engage with the new R&D challenges at scale (Keelara and Haywood, 2023[37]).

Cluster C, '**Epidemiology and social interventions**', encompasses a broad class of topics in the space of diffusion modelling and predictive research, scientific evidence, and advice for policymaking, as well as a wide range of topics covering social distancing and remote working, mobility, markets, business, and family impacts. The salient words for this cluster include 'access, worker, digital, distance'. The social sciences feature prominently in this topic alongside psychology and statistics. Cluster D, '**Digital access**

and online education', displays terms linked to R&D on remote interactions and learning forced by the COVID-19 physical distancing measures (Vincent-Lancrin, 2022[38]).

A series of interconnected clusters under a common 'branch' bring together themes dealing with COVID-19 R&D and population groups at risk (Berchet, Barrenho and de Bienassis, 2023[39]). Cluster E, '**Cancer (screening and treatment)**', is distinctive enough to be separately identified by the automatic classification process, probably owing to the existence of large and consolidated funding streams in this area. The presence of R&D funding and performer organisation in project descriptions may have played a role in this outcome. These funders appear to have reacted to COVID-19 through specific projects under their area of discretion given the increased exposure of cancer patients to the disease and preventive measures (e.g., screening deferrals and surgery cancellations). Cluster F, '**Public healthcare and other groups at risk**', represents a broader category that includes R&D connected to healthcare workers, pregnant women, diabetes, the elderly, and other vulnerable groups, as well as aspects concerning exclusion from healthcare provision. Completing the set of topic clusters on groups at risk, Cluster G, '**Mental health and addictions**', emerges as a distinct top-level cluster category, dealing in part with the psychological and psychiatric implications of COVID-19 (Astorga-Pinto, Hewlett and Haywood, 2023[40]; OECD/European Union, 2022[41]).

A final Cluster H, '**Environmental detection, transmission, and protection**', is concerned with environmental surveillance and developing methods of protection from transmission (e.g., PPE).

Table 10. Estimates of COVID-19 R&D by machine-generated "C8" topic clusters

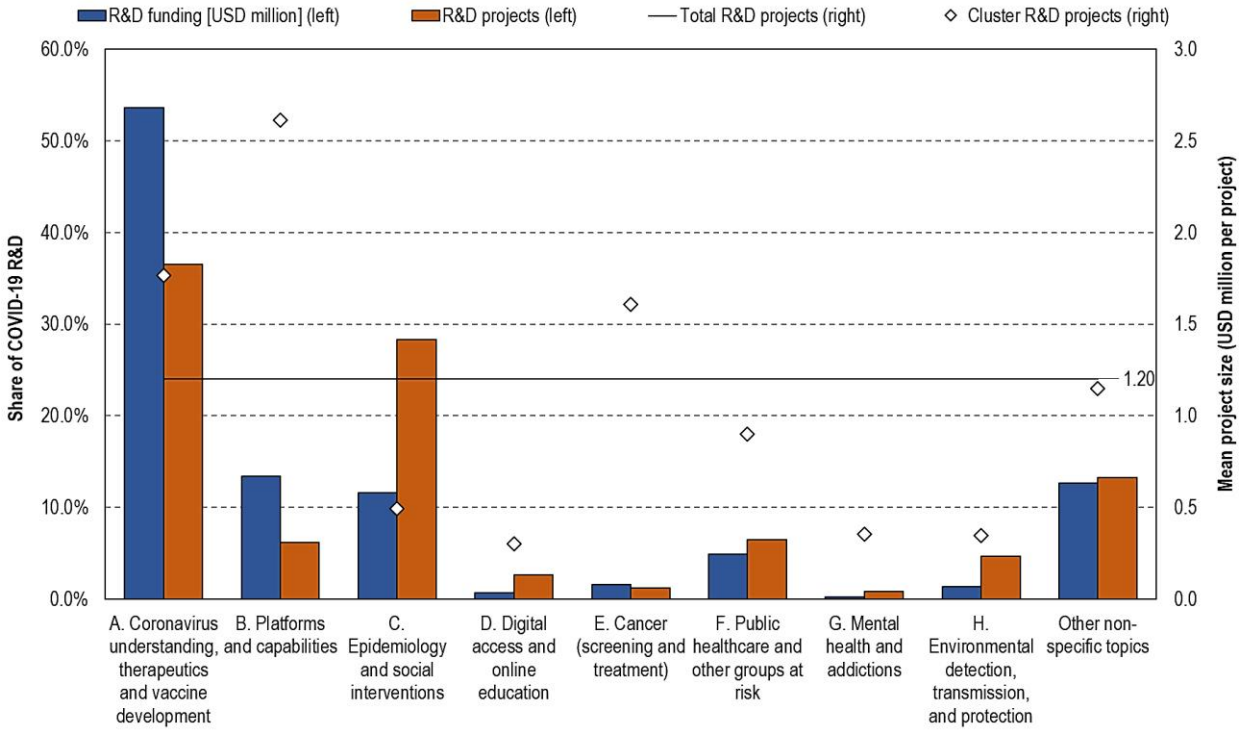
C8 topic cluster labels[1]	# R&D projects[2]	(Of which) R&D projects with missing funding information	R&D funding [USD million]	Implied mean project award [USD million per project][3]
A. Coronavirus understanding, therapeutics and vaccine development	4,395	573	6,752	1.77
B. Platforms and capabilities	694	50	1,685	2.61
C. Epidemiology and social interventions	3,357	392	1,464	0.49
F. Public healthcare and other groups at risk	758	79	613	0.90
E. Cancer (screening and treatment)	149	22	203	1.61
H. Environmental detection, transmission, and protection	558	68	169	0.35
D. Digital access and online education	298	22	83	0.30
G. Mental health and addictions	89	5	30	0.35
Other non-specific topics	1,588	204	1,589	1.15
Total	11,886	1,414	12,588	1.20

Notes:
[1] C8 reflects the 8 high level topic clusters (see Table 9 for a complete characterisation of the C8 clusters).
[2] Analysis limited to the retained COVID-19 R&D projects from Table 5.
[3] The calculation of mean project award (in USD million per project) excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
Source: OECD analysis of Fundstat database, March 2023.

Table 10 provides information on the R&D projects and funding allocated to the C8 clusters. Cluster A, '**Coronavirus understanding, therapeutics and vaccine development**', exhibits the highest R&D funding with more than USD 6.75 billion (more than 50% of the total), followed by cluster B, '**Platforms and capabilities**', with more than USD 1.69 billion (13%), and cluster C, '**Epidemiology and social interventions**' with USD 1.46 billion (12%). These results confirm the dominance **of biomedical R&D** in the overall COVID-19 R&D funding. However, an examination of project counts would provide a different qualitative picture, with many projects under cluster A (36%) and C (28%).

Differences between the distribution of R&D funding and projects can be explained by differences in funding awards across projects with different topic orientation. Implied average funding by topic clusters is calculated on a fractional basis and excludes projects with missing or nil funding information. Confirming the earlier results on R&D funding at the C34 topic level, biomedical R&D projects are significantly larger funds recipients on average (Figure 8) (OECD, 2023[42]). Awards for projects in Cluster B are the largest with average awards of USD 2.61 million per project, potentially reflecting the scale and scope pursued by this cluster's projects (Guan et al., 2020[43]; He et al., 2020[44]). Clusters A and E follow in terms of average award at USD 1.77 million and USD 1.61 million respectively. The smallest projects by funding are in clusters D, G, and H with mean awards in the order of USD 0.30 million. Cluster C, the one with higher social science content, has also relatively low average award levels at about USD 0.50 million.

Figure 8. Distribution of COVID-19 R&D project and funding by "C8" topic cluster

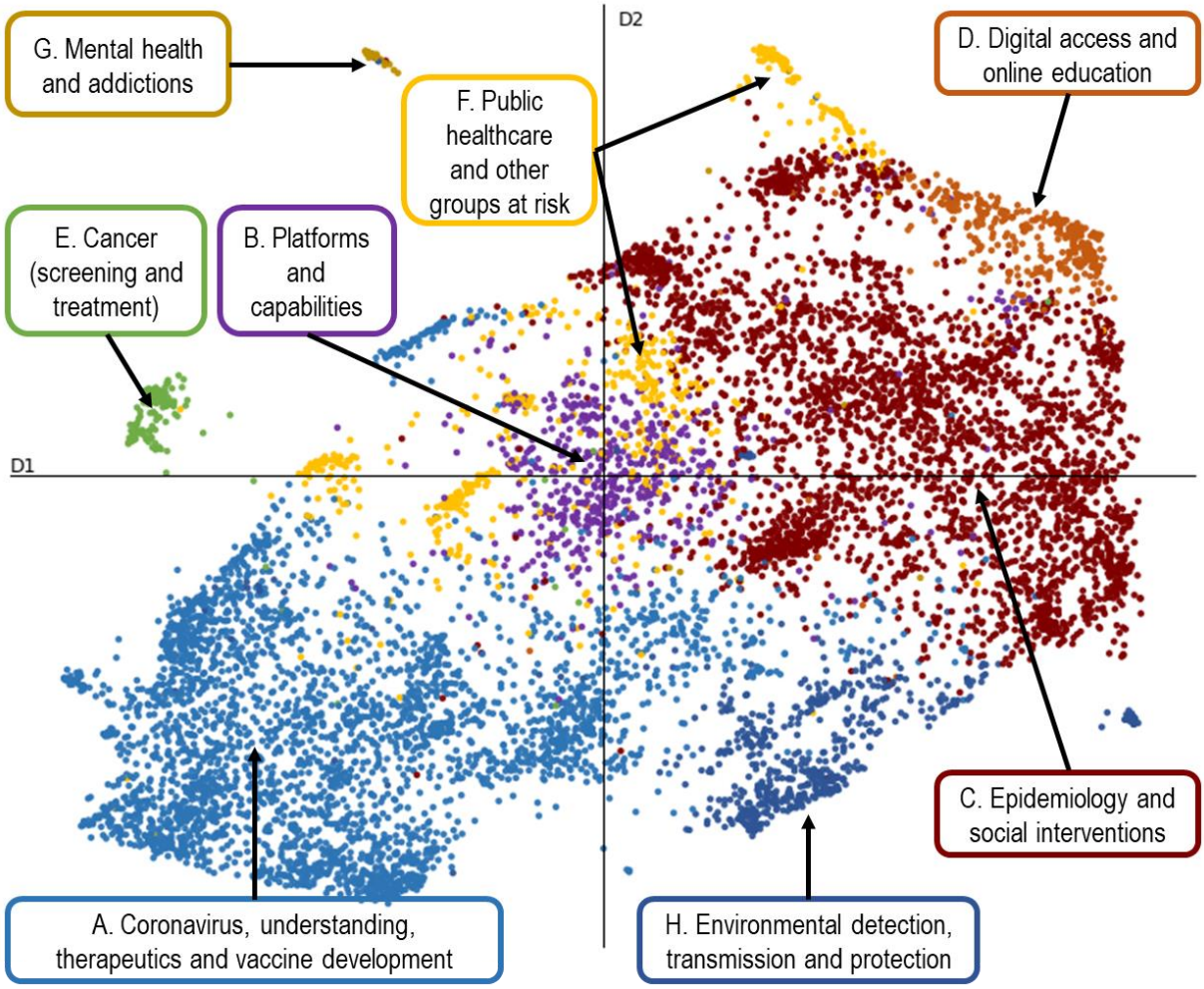


Notes:
[1] Analysis limited to the retained COVID-19 R&D projects from Table 5 and based on Table 10.
[2] The calculation of mean project award (in USD million per project) excludes projects with missing or nil funding information, which often obey to disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
Source: OECD analysis of Fundstat database, March 2023.

The probability distribution of top-level topics allows visualising the interconnectedness of projects and topics according to co-occurrence patterns. Figure 9 plots projects in a reduced two-dimensional space representation of multidimensional similarities between projects in terms of their topic profile distributions, in which the dominant topic has been used to assign a colour coding to each project. The visual representation places topic B on Platforms and capabilities at the centre of the project funding landscape, apparently bridging Topic A- "heavy" projects in the biomedical space towards the left of the chart with Topic C- social science intensive projects on the right side of the figure, and digital education at the very extreme of topic C topics at the opposite extreme. Topic E "Cancer" and Topic G "Mental health" appear in peripheral spaces with the former closer to biomedical A type projects and the latter closer to Topic C.

Topic H projects on "Environmental detection, transmission and protection" are more closely integrated in the funding landscape, in the space between Topic A and Topic C projects but more peripheral than Topic B projects. Projects on topic F on "Public healthcare and other groups at risk" inhabit both central and peripheral positions in the chart.

Figure 9. Visualisation of COVID-19 R&D projects and their dominant high-level topic clusters



Notes:
[1] Analysis limited to the retained COVID-19 R&D projects and their dominant high-level topic cluster.
[2] Project text (the combined processed title and abstract) is represented as a text embedding vector, which is reduced to two dimensions and visualised. Each dot represents a project in the vector space in two dimensions.
[3] The labels reflect the identified high-level topic clusters from Table 9, and topic cluster colours follow Figure 6.
[4] It is important to note that when there is dimensionality reduction (in this case to two dimensions), there is inherently loss of information.
Source: OECD analysis of Fundstat database, March 2023.

3.2.2. Funding analysis by agency/data source

Funding agency and source-based analysis can help inform a better understanding of funding portfolios and contribute towards explaining aggregate results, whilst also allowing to identify opportunities for improving data coverage and representativeness within and across countries. Figure 10 and Figure 11 show the top-level topic distribution by funding agency/source for COVID-19 R&D projects and funding amounts, respectively. Analysis is limited to the 10,472 COVID-19 R&D projects in the Fundstat database with known funding information.

As it may be expected from the diversity of funding sources covered, there is considerable topic heterogeneity in COVID-19 funding portfolios across different agencies/data sources, reflecting differences in funding mandates, strategy, and choices in response to the crisis. For most agencies, the distribution of R&D project content (in project count equivalents) across topics appears to be more evenly distributed, whereas the distribution of R&D funding is more concentrated.

In most cases, topic cluster A with its all-encompassing biomedical heading on Coronavirus understanding, vaccines and therapeutics, accounts for most of the funding, but several exceptions can be found among funding bodies and sources with a focus in social sciences or even in non-biomedical areas. This is particularly the case for Health and medical R&D funding agencies such as AUS_MRFF, AUS_NHMRC, CAN_CIHR, JPN_AMED, GBR_NIHR, and USA_NIH.

Topic cluster A is clearly dominant for the agency with the largest amount of COVID-19 funding in the Fundstat database, the US National Institutes of Health (USA_NIH) (Figure 5). As shown in Figure 10 and Figure 11, topic A accounts in this case for over 50% of funding and slightly less in terms of fractional project counts (Lalani et al., 2023[45]). This is followed by topic B, "Platforms and capabilities", with nearly 20% of funding. Topic A is even more dominant for US Federal COVID-19 R&D procurement. (USA_COVID-FPDS), to the detriment of Topic B.

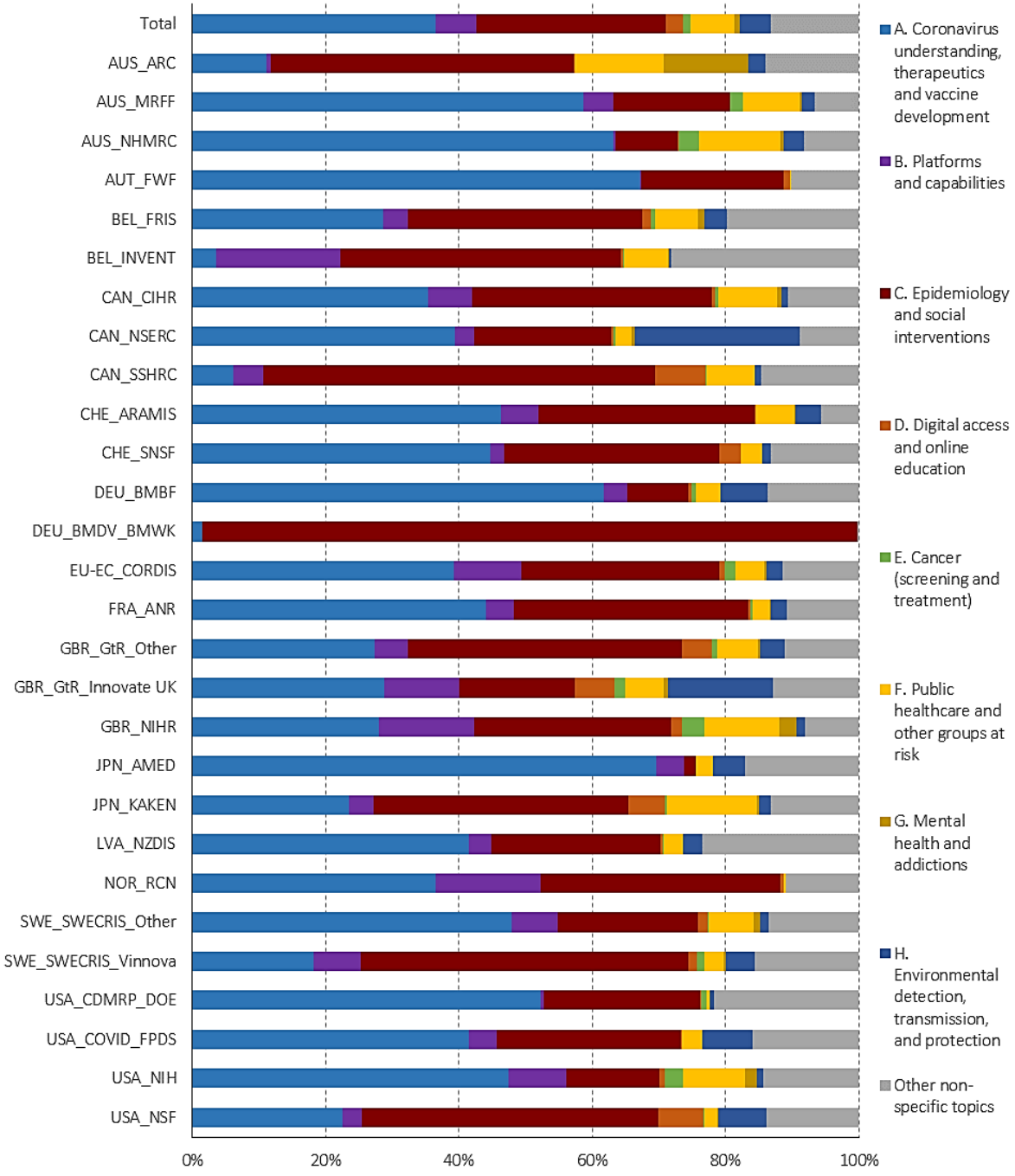
Not all Health and medical R&D funding agencies devote similar allocations to topic B. GBR_NIHR and JPN_AMED, together with USA_NIH are among those assigning higher funding levels to this topic on "Platforms and capabilities". AUS_NHMRC and GBR_NIHR support for topic F, "Public healthcare and other groups at risk" can be noted, while biomedical topics such as E, "Cancer (screening and treatment)" represent relatively small but significant components of the COVID-19 R&D allocation by GBR_NIHR and USA_NIH.

The contrast between agencies within a country can often match similar comparisons within other countries in terms of their specialisation patterns. This for instance the case of Japanese agencies JPN_AMED and JPN_KAKEN (data source principally covering JSPS funding) vis a vis USA_NIH and USA_NSF.

Business R&D and innovation-funding agencies such as covered under GBR_GtR_Innovate_UK and SWE_SWECRIS_Vinnova exhibit a broad spectrum of project topic allocations as well as somewhat significant shares of non-specific content. The EU-EC CORDIS source, which combines research and experimental development funding, presents a more balanced distribution across topics. It also displays a rather significant share of Topic B funding.

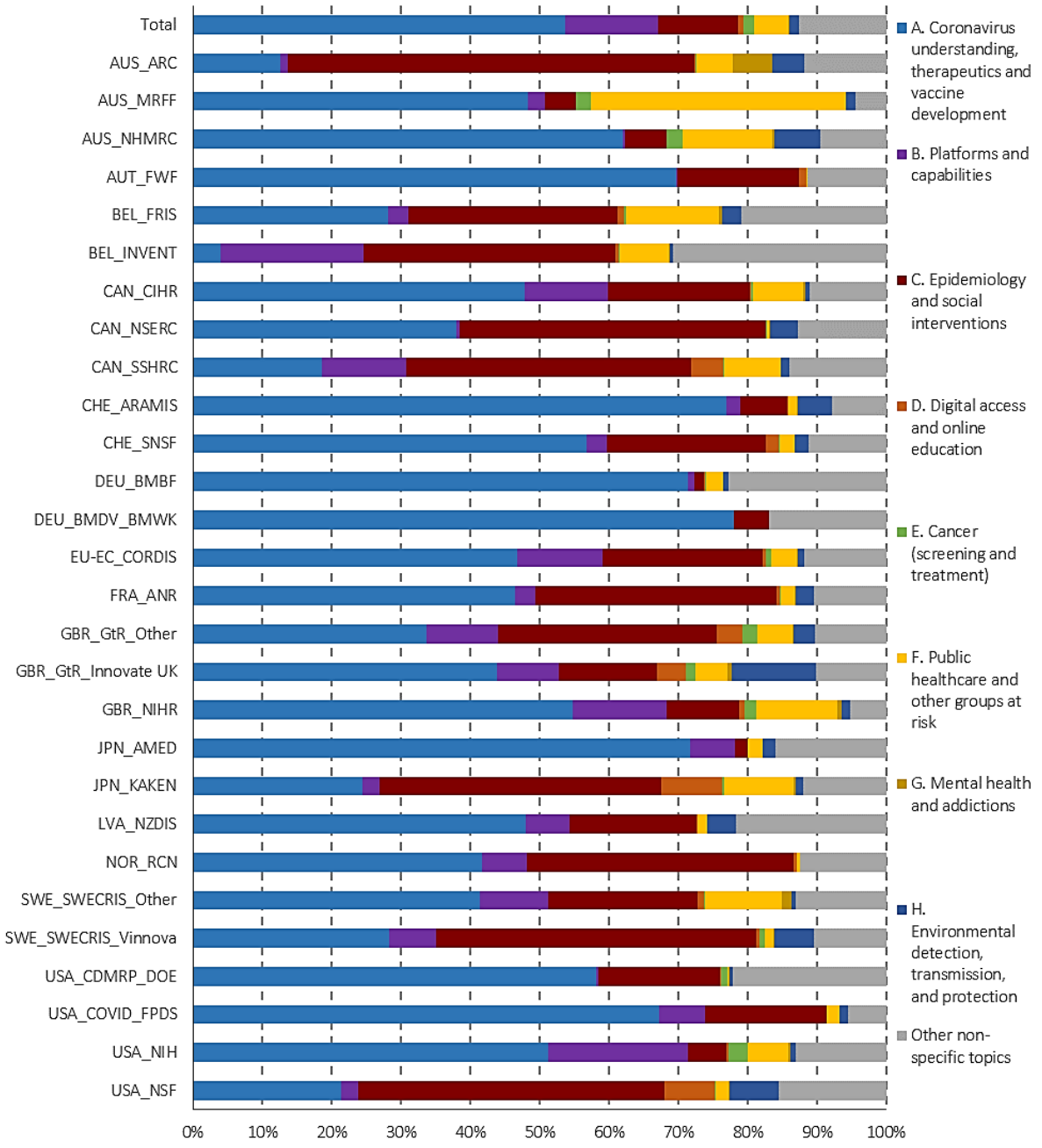
Social science-oriented agencies such as CAN_SSHRC, allocate higher project and funding towards topic C "Epidemiology and social interventions".

Figure 10. COVID-19 R&D projects by C8 clusters and funding agency/data source



Notes:
[1] Analysis limited to 10,472 COVID-19 R&D projects in the Fundstat database with known funding information. Biases due to the data coverage and the topic model development may distort the estimate of support breakdown by funding agency/data source.
[2] FY2021 is partial because of data reporting structures and data availability across different countries. Specifically, FY2021 data are unavailable for BEL_INVENT, CAN_CIHR, CAN_NSERC, CAN_SSHRC, DEU_COVID-GEPRIS, JPN_AMED, and NOR_RCN.
[3] Sorted by country followed by agency/data source alphabetical order. Colours follow the cluster colours introduced in Figure 6.
[4] The analysis does not include the DEU_COVID-GEPRIS database as there is no funding information available for those projects.
[5] GBR_Other includes ESRC, NERC, EPSRC, BBSRC, MRC, AHRC, STFC, NC3Rs, and UKRI.
[6] SWE_SWECRIS_Other includes FBEES, IFAU, VR, Forte, Formas, SHLF, RJ, and SWEA.
Source: OECD analysis of Fundstat database, March 2023.

Figure 11. COVID-19 R&D funding by C8 clusters and funding agency/data source



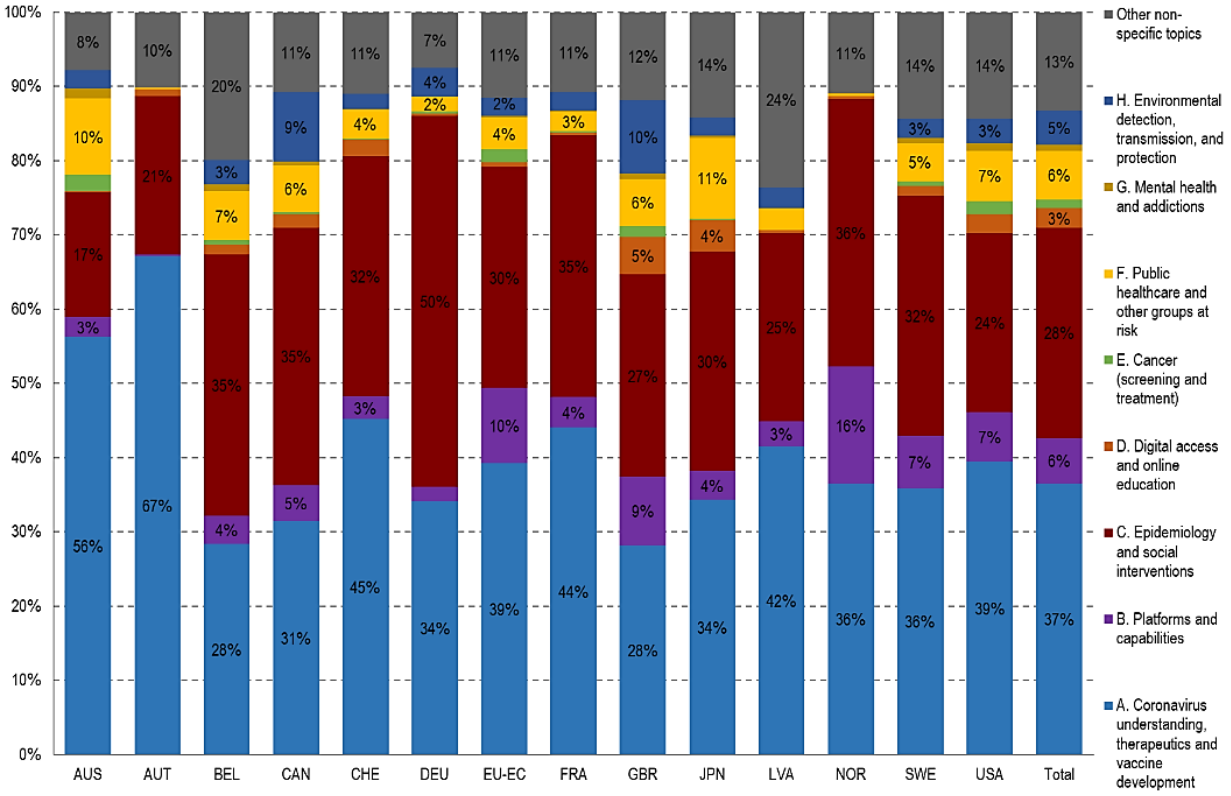
Notes:
[1] Analysis limited to 10,472 COVID-19 R&D projects in the Fundstat database with known funding information. Biases due to the data coverage and the topic model development may distort the estimate of support breakdown by funding agency/data source.
[2] FY2021 is partial because of data reporting structures and data availability across different countries. Specifically, FY2021 data are unavailable for BEL_INVENT, CAN_CIHR, CAN_NSERC, CAN_SSHRC, DEU_COVID-GEPRIS, JPN_AMED, and NOR_RCN.
[3] Sorted by country followed by agency/data source alphabetical order. Colours follow the cluster colours introduced in Figure 6.
[4] The analysis does not include the DEU_COVID-GEPRIS database as there is no funding information available for those projects.
[5] GBR_Other includes ESRC, NERC, EPSRC, BBSRC, MRC, AHRC, STFC, NC3Rs, and UKRI.
[6] SWE_SWECRIS_Other includes FBEES, IFAU, VR, Forte, Formas, SHLF, RJ, and SWEA.
Source: OECD analysis of Fundstat database, March 2023.

3.2.3. Funding analysis by geographical area

Based on the agency and source-based results in the previous sub-section, Figure 12 and Figure 13 provide an aggregated breakdown of R&D projects and funding by C8 topics for the EU and each of the countries included in the study as a step towards building country-level funding indicators. Belgium and Latvia exhibit larger shares of non-specific project content, just over 20% compared to other countries. Clusters A and C combined account for well over 60% of projects (in fractional equivalent terms). With exceptions like Australia and Austria where there is a much larger number of projects for Cluster A (Coronavirus…) than for C (Epidemiology and social), numbers for these two topics tend to be balanced and the picture across most countries rather similar.

Figure 12. Topic distribution of COVID-19 R&D projects, by country

Shares based on fractional project counts based on project-topic probabilities



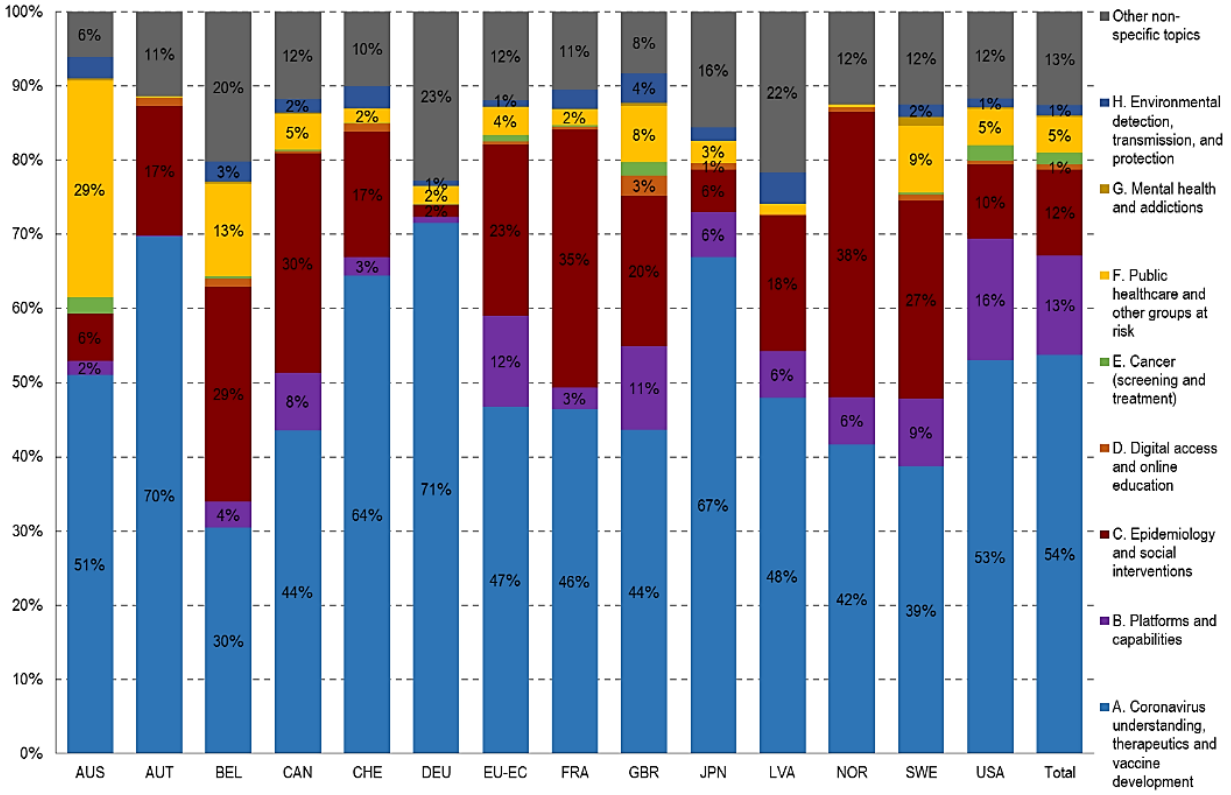
Notes:
[1] Analysis limited to the retained COVID-19 R&D projects from Table 5. Biases due to the data coverage (Figure 1) and the topic model development may distort the estimate of support breakdown by topic for any given country.
[2] Colours follow the cluster colours introduced in Figure 6.
[3] The analysis excludes projects with missing or nil funding information, which obey disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
Source: OECD analysis of Fundstat database, March 2023.

Examination of the funding distribution (Figure 13) shows that Germany joins Belgium and Latvia as countries with a relatively high share of funding in projects with content that cannot be allocated to any of the specific topics. Consistent with the aggregate picture, cluster A accounts for the largest share of funding in all countries by a significant margin, except for Belgium where topic A and C are very close. Cluster C is typically ranked second in funding for all countries but United States, where topic B (Platforms) prevails

by a significant margin, contributing through its large weight to the aggregate position of topic B as the second largest overall across countries.

Figure 13. Topic distribution of COVID-19 R&D funding, by country

Shares based on fractional funding allocation based on project-topic probabilities



Notes:
[1] Analysis limited to the retained COVID-19 R&D projects from Table 5. Biases due to the data coverage (Figure 1) and the topic model development may distort the estimate of support breakdown by topic for any given country.
[2] Colours follow the cluster colours introduced in Figure 6.
[3] The analysis excludes projects with missing or nil funding information, which obey disclosure controls or database entries intended to register project modifications without funding implications (zero additional funding).
Source: OECD analysis of Fundstat database, March 2023.

Several other idiosyncratic aspects emerge. Australia displays the largest share of funding under topic F "Public healthcare and other groups at risk" with nearly 30%. Most of the funding in Germany has been allocated to the cluster A, with relatively no funding allocated to any of the other topics. This appears to be due to major systematic differences in award size, combined with very large projects on RNA vaccines funded by BMBF[13]. Cancer-related funding (topic E) is most visible in the United States and Australia, but still relatively minor. The United Kingdom dedicates the largest share of funding to R&D on topic D "Digital education". Funding for topic C "Epidemiology and social sciences" is proportionally largest in Belgium, France, and Norway.

3.2.4. Analysis of market orientation in COVID-19 R&D funding

One key aspect of R&D funding directionality concerns its potential orientation towards developing products (goods or services) and processes for use in the marketplace. One potential identification would be to identify R&D support beneficiaries in the business sector. However, this information is not widely currently available in the sources deployed in Fundstat, and even if available, a subject-based approach would represent an underestimate of market orientation since R&D funded in other sectors might be directly contributing to market-based activities. As an alternative, the study explored the possibility of using the text-based information in abstract to draw inferences about potential market orientation, as potential guide to drawing a demarcation between funding for research and experimental development.

To implement the identification of market-orientated COVID-19 R&D funding, this study adopts a simple key-term matching approach for illustrative purposes. A market-oriented COVID-19 R&D project is defined as one that has two or more key-terms from a base vocabulary in the project's post-processed text for all data sources, except for the case of the USA_COVID-FPDS procurement data, where an extended vocabulary is used to consider the specific language of procurement award descriptions. A base and an extended business and market-oriented vocabularies consisting of key-terms are defined, which are then lowercased, lemmatised, and translated in all languages in the Fundstat database (Section 2.2.1). The key-terms are then used to search within the projects' processed text that combines the title and abstract and are shown in Table 11.

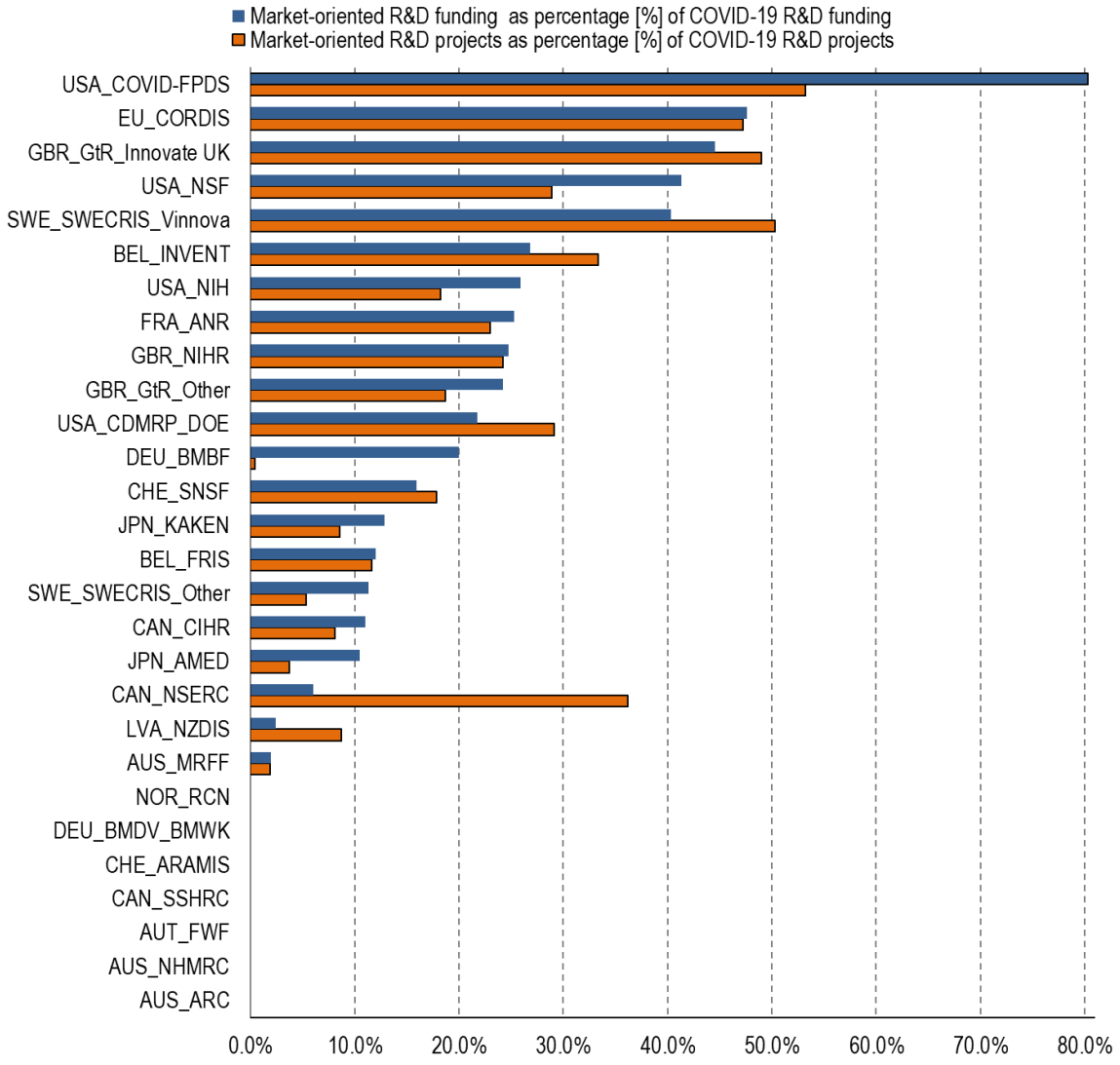
Table 11. Business and market-oriented experimental vocabulary of key-terms

Set[1]	Number of key terms	Key terms[2,3]
Base	33	adoption, business, commercialisation, competitive, consumer, corporate, corporation, diffusion, enterprise, entrepreneurship, expenditure, feasibility study, firm, industry, innovation, intellectual property, investment, IPR, joint venture, license, management, manufacture, market, model, process, product, production, service, spinoff, start-up, strategy, technology, technology transfer
Extension	13	adjustment, clause, condition, deliverable, delivery, extension, external, issue, platform, procure, purchase, requirement, transaction
Total	46	

Notes:
[1] A market-oriented COVID-19 R&D project is defined as one that has two or more key-terms from the base vocabulary in the project's post-processed text for all data sources, except for the case of the USA_COVID-FPDS procurement data, where an extended vocabulary is used.
[2] The key-terms are cleaned and processed by lowercasing and lemmatising. Lemmatisation involves grouping together the different forms of a word and analysing them as a single element (known as 'lemma'). The key-terms are translated in all languages in the Fundstat database (Section 2).
[3] The experimental list of key terms is not exhaustive, and other relevant key-terms and translation variations could exist.
Source: OECD Fundstat infrastructure, March 2023.

Following this procedure, this study identifies 2,705 market-oriented COVID-19 R&D projects (437 of them have zero funding or no funding information), which represent 22% of all COVID-19 R&D projects. These projects account for nearly 4.3 USD billion worth of funding, namely just over one third (34%) of estimated COVID-19 R&D funding. Figure 14 shows how the share of market-oriented R&D projects and funding in total COVID-19 R&D awards varies across funding agencies and sources.

Figure 14. Market-oriented COVID-19 R&D by funding agency/data source



Notes:
[1] Analysis limited to 10,472 COVID-19 R&D projects in the Fundstat database with known funding information.
[2] Sorted by highest share of business-oriented R&D funding. Biases due to the data coverage and the topic model development may distort the estimate of support breakdown by funding agency/data source, which follow from Figure 1 and Figure 2.
[3] A business-oriented COVID-19 R&D project is defined as one that has two or more key-terms from a base vocabulary in the project post-processed text for all data sources, except for the case of the USA_COVID-FPDS procurement data, where an extended vocabulary is used. Lack of identifiable market-oriented R&D projects is in no way indicative of a lack of market-relevance, but a reflection of a lack of market-oriented language in the short project descriptions found in abstracts for these agencies. Agencies with shorter project descriptions are also more likely to fall under this category.
[4] FY2021 is partial because of data reporting structures and data availability across different countries. Specifically, FY2021 data are unavailable for BEL_INVENT, CAN_CIHR, CAN_NSERC, CAN_SSHRC, DEU_COVID-GEPRIS, JPN_AMED, and NOR_RCN.
[5] The analysis does not include the DEU_COVID-GEPRIS database as there is no funding information available for those projects.
[6] GBR_GtR_Other includes ESRC, NERC, EPSRC, BBSRC, MRC, AHRC, STFC, NC3Rs, and UKRI.
[7] SWE_SWECRIS_Other includes FBEES, IFAU, VR, Forte, Formas, SHLF, RJ, and SWEA.
Source: OECD analysis of Fundstat database, March 2023.

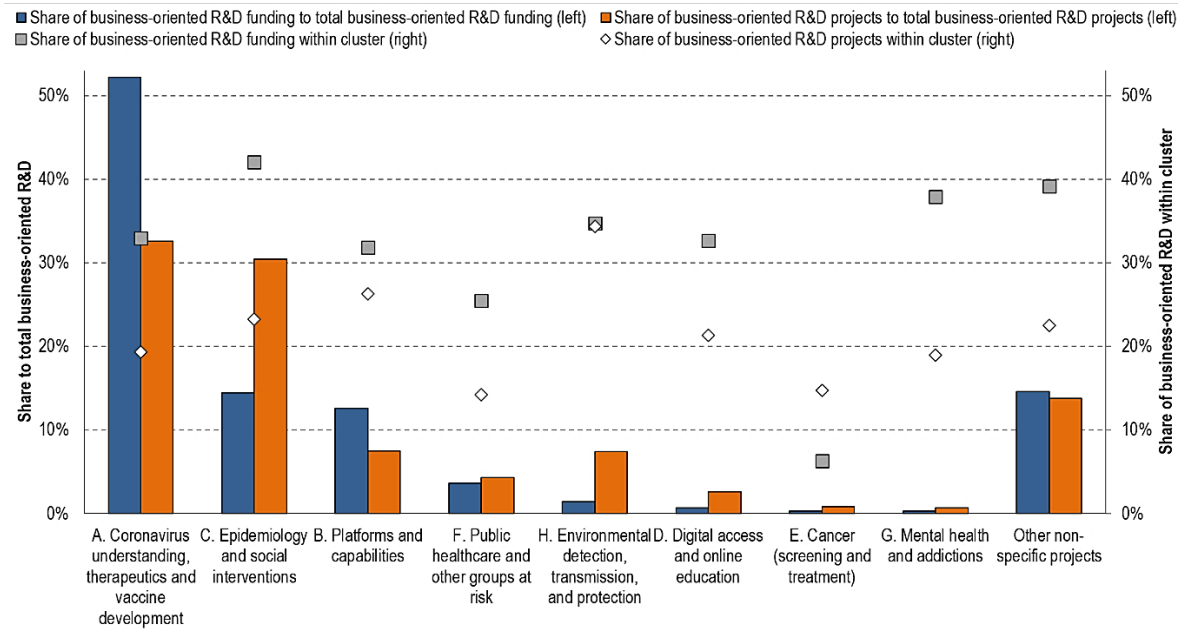
Agencies and sources with a high share of business-oriented R&D projects and funding tend to be defined by the explicit market orientation of procurement activity (as USA_COVID-FPDS) and funding for

innovation agencies such as captured under SWE_SWECRIS_Vinnova and GBR_GtR_Innovate UK. Funding covered by EU-EC_CORDIS also features under the sources with a higher degree of market-orientation.

The procedure also results in several agencies appearing as having no identifiable market-oriented R&D projects. This is in no way indicative of a lack of market-relevance, but a reflection of a lack of market-oriented language in the short project descriptions found in abstracts for these agencies. Agencies with shorter project descriptions are also more likely to fall under this category. This is another stark reminder that, before reaching any firm conclusions about text-based analysis, it is important to advance in ensuring greater homogeneity in the content that is captured in project descriptions used for comparisons across different funding programmes, agencies, and countries.

Figure 15 shows the share of business-oriented R&D projects and funding to the total business-oriented R&D projects and funding across different topic clusters, as well as the share of business-oriented R&D projects and funding within each cluster. The cluster with the highest share of business-oriented R&D projects and funding to the total business-oriented R&D is the 'Coronavirus understanding, therapeutics and vaccine development', with 33% of business-oriented R&D projects and 52% of business-oriented R&D funding. However, the share of business-oriented R&D projects within this cluster is lower than the average at 19%, indicating that other types of R&D (such as academic or government-funded research) could also be prevalent in this area. This cluster is followed by cluster C 'Epidemiology and social interventions', and cluster B 'Platforms and capabilities' for both the share of business-oriented R&D projects and funding. In both these clusters, the share of business-oriented R&D is higher than the average, which potentially signal a drive towards business R&D. It is also interesting to note that clusters D 'Digital access and online education', G 'Mental health and addictions', H 'Environmental detection transmission and protection' also have a relatively high share of business-oriented R&D projects and funding within cluster. This also applies to projects classified in other non-specific topics.

Figure 15. Business-oriented COVID-19 R&D by C8 cluster



Notes:
[1] Analysis limited to the retained COVID-19 R&D projects. Data are based on Table 10 and Figure 8.
[2] The calculations of share of business-oriented R&D projects both to the total business-oriented R&D projects and within cluster exclude projects with missing or nil funding information.
Source: OECD analysis of Fundstat database, March 2023.

4 Comparing Fundstat COVID-19 R&D with alternative data sources and expert classifications

4.1. Mapping Fundstat results to the WHO classification of COVID-19 research priorities

Analysis of Fundstat COVID-19 R&D funding and its allocation to different topics using an unsupervised machine-driven approach raises two main questions, namely:

- are the results consistent with other studies of COVID-19 R&D funding and what reasons lie behind any differences?
- how do the results of a machine-based "classification" of R&D projects map onto expert-based categories?

Addressing these questions is the purpose of this sub-section, in which the COVID-19 Research Project Tracker by UKCDR & GloPID-R (Bucher et al., 2023[14]; UKCDR & GloPID-R, 2023[15]) is used as a comparator for funding estimates as well as the basis for a comparative analysis between the machine-based labels of the C8 clusters and the WHO research priority areas. The latter is done by:

- First developing a machine-learning model trained on the WHO labels and text descriptions in the COVID-19 tracker database. Since R&D projects collected by the COVID-19 tracker are manually labelled by experts to the WHO classification of research priorities, this provides a valuable contrast to the entirely machine-based classification in the previous sections.
- Using the model to predict (assign probabilities to) WHO priority topics for each R&D project in the Fundstat database and map how different categories relate at the project and aggregate level.

4.1.1. Comparing COVID-19 R&D funding estimates from different sources

Coverage comparisons between the Fundstat database on COVID-19 R&D funding and the COVID-19 Project Tracker by UKCDR & GloPID-R (February 2023, version 9) are presented in Table 12.The latter is the largest database in terms of numbers of projects, with 17,955 projects compared to 11,886 for Fundstat. The COVID-19 Project Tracker covers a larger number of funders as its scope is not geographically or institutionally limited to government bodies as Fundstat. In contrast, it captures R&D investment in the order of USD 6.5 billion (Bucher et al., 2023[14]; UKCDR & GloPID-R, 2023[15])[14]. Observed funding amount differences with the amounts reported for Fundstat under Section 3.1 appear to be partly explained by the inclusion of the US COVID-19 R&D procurement data[15] in the Fundstat database, not present in the COVID-19 tracker. Furthermore, funding amount data is available for only 61% of all projects in the COVID-19 tracker relative to 88% in the Fundstat database.

Table 12. Description of COVID-19 R&D awards and funding in the Fundstat and COVID-19 Project Tracker databases

	OECD Fundstat database	COVID-19 tracker
R&D project awards ("projects")	11,886	17,955
(Of which) R&D projects with missing funding information	1,414	6,747
R&D funding [USD billion]	12.6	6.5
Mean project award [USD million per project]	1.2	0.6
Main distinctive coverage features		
	Only government bodies and EU	Also includes major non-profit funders and consortia
	Limited coverage, focus on funding data availability	Larger coverage of countries/agencies
	Includes US Federal R&D procurement	Entirely focused on grants

Note: The calculation of mean project awards (in USD million per project) excludes projects with missing or nil funding information.
Source: OECD analysis of Fundstat database and COVID-19 Research Project Tracker by UKCDR & GloPID-R database (February 2023, version 9), March 2023.

Distribution of projects and funding by WHO priorities

One of the main advantages of the COVID-19 Project Tracker stems from its availability of manually tagged projects to WHO priorities (WHO, 2020[46]; WHO, 2020[6]). This not only allows describing the distribution of funding using the Tracker data, but also provides a basis for analysing the composition of the Fundstat database in terms of an expert-defined classification system. Since the projects in the two databases cannot be easily and unequivocally matched, the strategy adopted estimates the probability that any given Fundstat project should be allocated to each WHO topic with the help of a machine-learning model. This model has been developed using the latest available COVID-19 tracker data (February 2023, version 9) and a fine-tuned BERT model ('bert-base-multilingual-cased') for multi-label classification. To use as inputs for the model, the COVID-19 tracker project data is transformed into embeddings using the same sentence transformer as the BERTopic topic model. The model then produces outputs in the form of probability scores indicating the likelihood that the project in question is aligned with WHO priorities (WHO, 2020[46]). These probabilities are based on the primary WHO priority label assigned to each project by experts compiling the COVID-19 tracker data[16]. Finally, the model is used to predict the WHO priority topics for the COVID-19 R&D projects in the Fundstat database and the estimated probabilities for the COVID-19 tracker are used to provide like-for-like comparisons with the predicted probabilities for the Fundstat database.

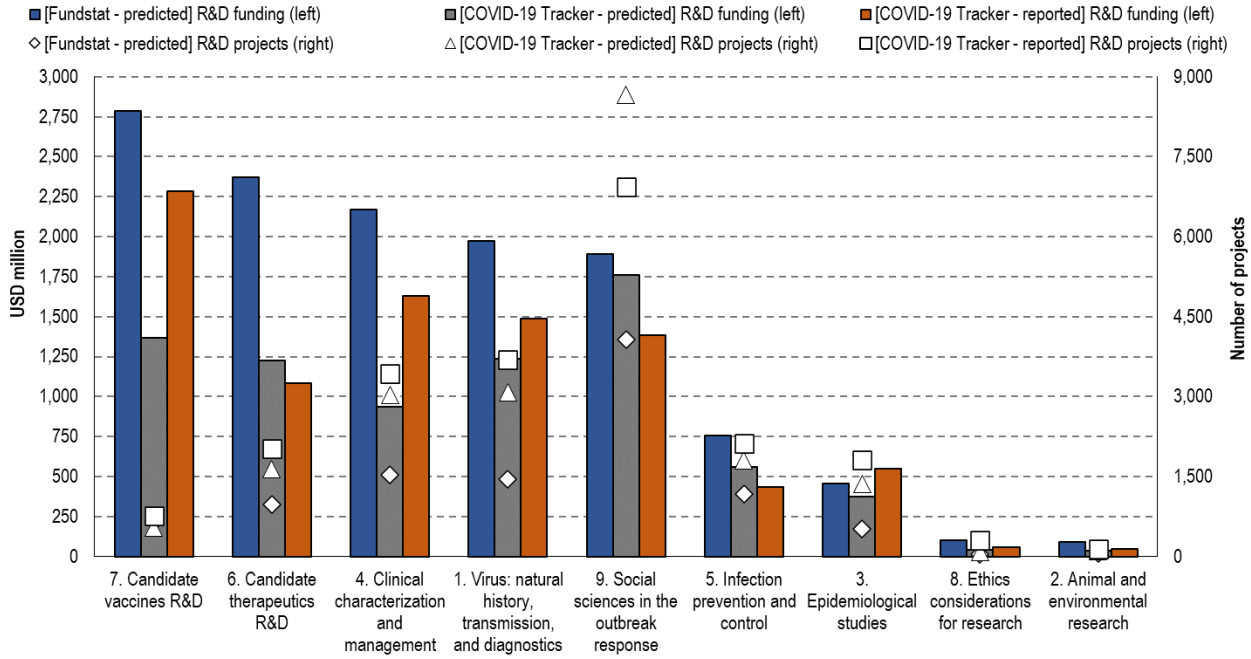
Figure 16 shows the predicted funding allocation for each WHO priority topic (in USD million) for the COVID-19 R&D projects in the Fundstat database and compares those to the reported funding allocation for each WHO priority topic as provided by the COVID-19 Tracker team. Since Tracker projects can be reported in the database as assigned to more than one topic, resulting in non-additive totals, a more like-for-like comparison is also made with estimates based on apportioning project counts and funding using the model-predicted topic probabilities for each project.

Comparing reported and modelled estimates within the COVID-19 tracker, the model-based fractional allocation appears to reduce the relative importance of candidate vaccines relative to other topics. The result of applying the model results in estimates that show vaccines and therapeutics receiving very similar funding allocations, potentially because projects involving vaccines also include text elements and patterns that can be potentially tagged as well to other topics, thus reducing the fractional allocation of any given vaccine-related project to that topic. On that basis, it is not possible to conclude that the priority of R&D therapeutics was relatively underfunded compared with vaccines as it might otherwise be claimed. Some degree of caution is required though. Observed distributional differences may also be influenced by the

modelling process due to potential prediction errors, as models are reliant on the available text inputs that may be incomplete and manual tagging decisions which can obey to multiple criteria.

Figure 16. COVID-19 R&D Funding allocations to WHO priorities in Fundstat and COVID-19 tracker

Funding and project distribution by WHO reported (Tracker) and model-predicted topics (Fundstat and Tracker)



Notes:

[1] For the COVID-19 Research Project Tracker by UKCDR & GloPID-R, version 9 (February 2023) contains 17,955 projects with a funding investment of about USD 6.5 billion. Reported project and funding are based on tagged priorities areas. Since a given project can be assigned to more than one WHO priority topic, the total sum across categories exceeds the total number of projects and funding. Model based probability estimates of a project belonging to a topic have been used to generate additive estimates per topic for comparison with Fundstat data.

[2] The Fundstat calculations are based on 10,472 projects in the Fundstat database with known funding information. The topic model trained on COVID-19 tracker data has been used to predict topic probabilities for Fundstat projects.

[3] Known differences in coverage relate to: (i) the sourcing of procurement data identified by US authorities as connected to COVID-19 (in Fundstat); (ii) the coverage of multiple government agencies with smaller funding covered (in COVID-19 tracker); (iii) the coverage of non-for-profit or philanthropic entities that have provided funding grants during the pandemic (in COVID-19 tracker); (iv) the data availability for funding information, with 61% in the COVID-19 tracker, and 88% in the Fundstat database;

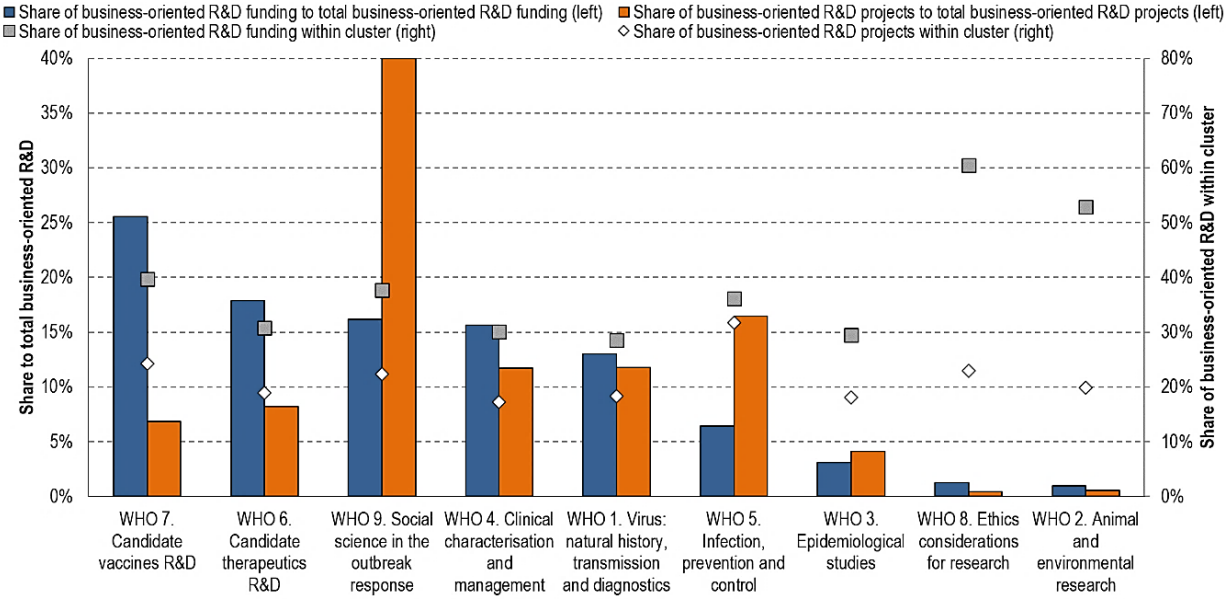
Source: OECD analysis of Fundstat database and COVID-19 Research Project Tracker by UKCDR & GloPID-R (February 2023, version 9), March 2023.

Comparing like-for-like model-based estimates across the two databases, the main differences in terms of funding between Fundstat and UKCDR ultimately attributable to funding coverage are found under Vaccines, Candidate therapeutics and Clinical characterisation and management, with Fundstat estimates significantly higher. The inclusion of Federal procurement R&D data that has been identified and published as COVID-19 related by US authorities in the Fundstat database makes a significant difference. Many of these projects involve re-purposed drug trials for COVID-19 purposes that were implemented as procurement actions and originated from the US Office of Assistant Secretary for Preparedness and Response. However, these projects are not covered under the COVID-19 tracker. Under "Social sciences", total results are very similar across databases in terms of absolute funding amounts, although UKCDR exhibits a much higher number of projects (both reported and modelled for fractional counting).

Market orientation of COVID-19 funding by WHO priority topics

Extending the analysis presented under sub-section 3.2.4, with the approximative method to identify market orientation, Figure 17 shows estimates of business-oriented COVID-19 R&D according to the WHO priority topics. The cluster on WHO 7 'Candidate vaccines R&D' accounts for the largest share of market-oriented R&D funding out of total COVID-19 R&D funding, followed by WHO 6 'Candidate therapeutics R&D' and WHO 9 'Social science in the outbreak response'. These results are again very different compared to results based on project counts because of the unequal distribution of funding across projects across clusters. Cluster WHO 9 'Social science in the outbreak response' has the highest share of market-oriented R&D projects to total R&D projects, which could also be driven by its broad definition (WHO, 2020[6]). In terms of market orientation intensity within topics, WHO 5 'Infection, prevention and control' has the highest share of market-oriented R&D projects, whereas WHO 8 'Ethics' has the highest share of market-oriented R&D funding, albeit based on rather small body of project content. Vaccines exhibits the highest market orientation intensity among the priorities accounting for a significant share of the funding. It is important to note that this analysis does not use a predictive model to identify and distinguish contextual appearances of the market-oriented keyterms.

Figure 17. Business-oriented COVID-19 R&D by WHO priority topic

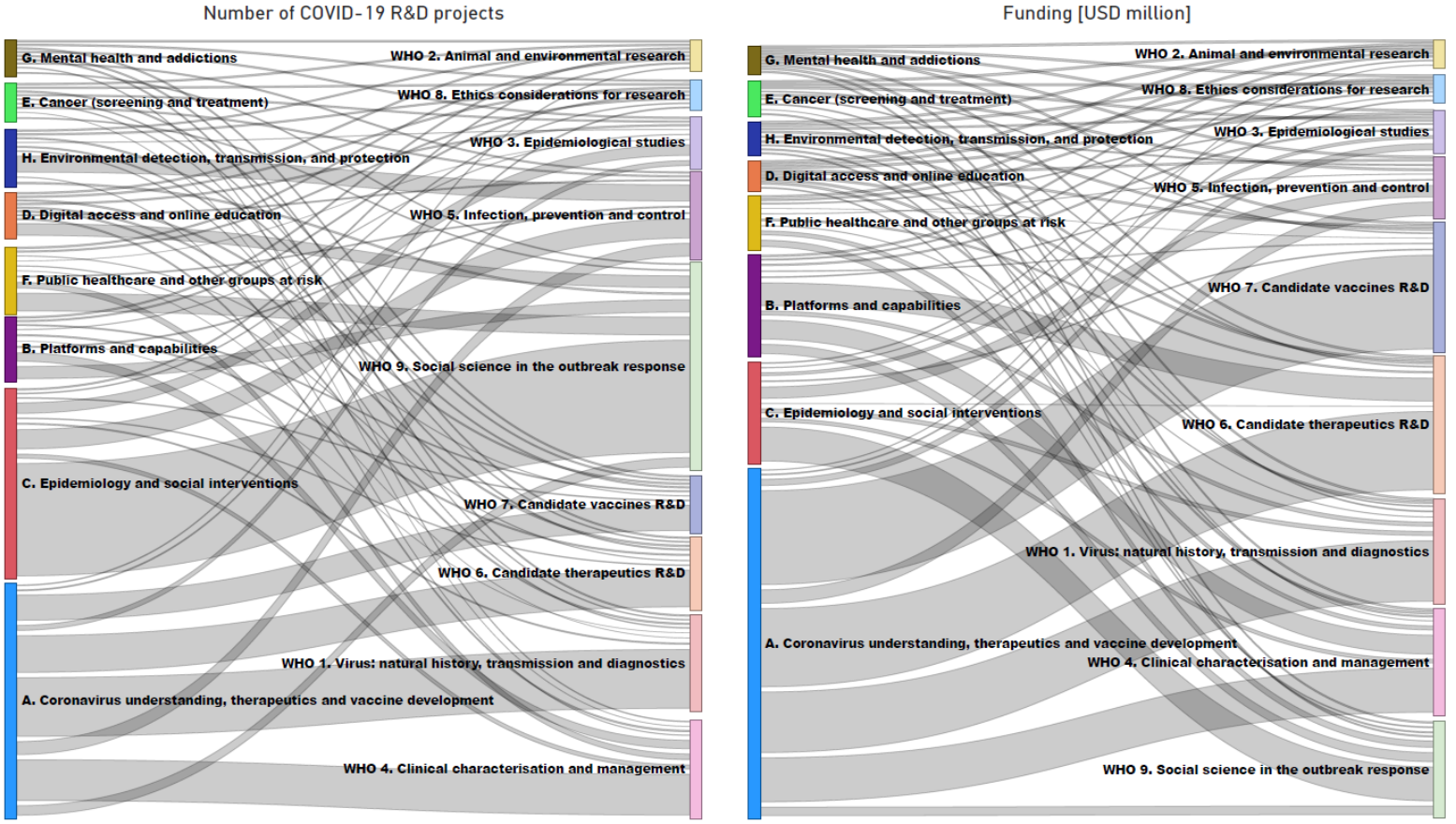


Notes:
[1] Analysis limited to the retained Fundstat COVID-19 R&D projects, based on modelling estimates reported under Figure 16.
[2] The calculations of share of business-oriented R&D projects both to the total business-oriented R&D projects and within cluster exclude projects with missing or nil funding information.
Source: OECD analysis of Fundstat database, March 2023.

4.1.2. Mapping machine-based topics to WHO research priority areas

The information on predicted WHO priorities covered in projects provides an additional mechanism for interpreting the outcomes of the machine learning classification analysis in Section 3.2. Figure 18 shows a Sankey diagram linking the different topic distributions of the Fundstat C8 clusters to the WHO priority topics. To map the Fundstat topic modelling labels to the predicted WHO priority topics, flows represent the connection between C8 clusters to the predicted WHO topics that results from apportioning the share of a project (in terms of project counts) in each C8 topic to different WHO topics. This allows for the calculation of the joint probability that a project belongs to both the C8 topic and the connected WHO topic.

Figure 18. Mapping between Fundstat C8 clusters and WHO priority topics by R&D projects and R&D funding



Notes: [1] The Fundstat C8 clusters are on the left, and the WHO priority labels on the right. The thickness of the relationship link represents the number of projects or funding, and the thickness of the node represents the total. [2] The mapping includes 10,472 projects in the Fundstat database with known funding information. [3] The total at the input (left) is equal to the total at the output (right) for both number of project counts and funding.
Source: OECD analysis of Fundstat database, March 2023.

This mapping provides a broad view of the connection between the Fundstat labels and the WHO priority topics for the COVID-19 R&D projects with known funding information. The machine-generated topic cluster A 'Coronavirus understanding, therapeutics, and vaccine development' maps almost entirely onto the WHO 1 ('Virus natural history, transmission, and diagnostics'), 4 ('Clinical characterization and management'), 6 ('Candidate therapeutics R&D'), and 7 ('Candidate vaccines R&D') for project counts and covers a significant amount of funding. Some aspects of WHO 4 and 6 also connect to cluster B ('Platforms and capabilities'), which lacks a clear WHO equivalent category. The largest flows for cluster B ('Platforms and capabilities') connect with WHO 6 ('Candidate therapeutics'), and WHO 4 ('Clinical characterisation and engagement').

WHO 9 ('Social sciences in the outbreak response') encompasses a wide range of projects, including those related to cluster C ('Epidemiology and social interventions'), cluster D ('Digital access and online education'), and cluster F, ('Public healthcare and other risk groups'). Most of the projects assigned to WHO 3 ('Epidemiological studies), are classified under the machine cluster C ('Epidemiology and social interventions'). This may obey to the potential use of common tools, such as statistical analysis and modelling, across these projects and those related to social sciences in the outbreak response.

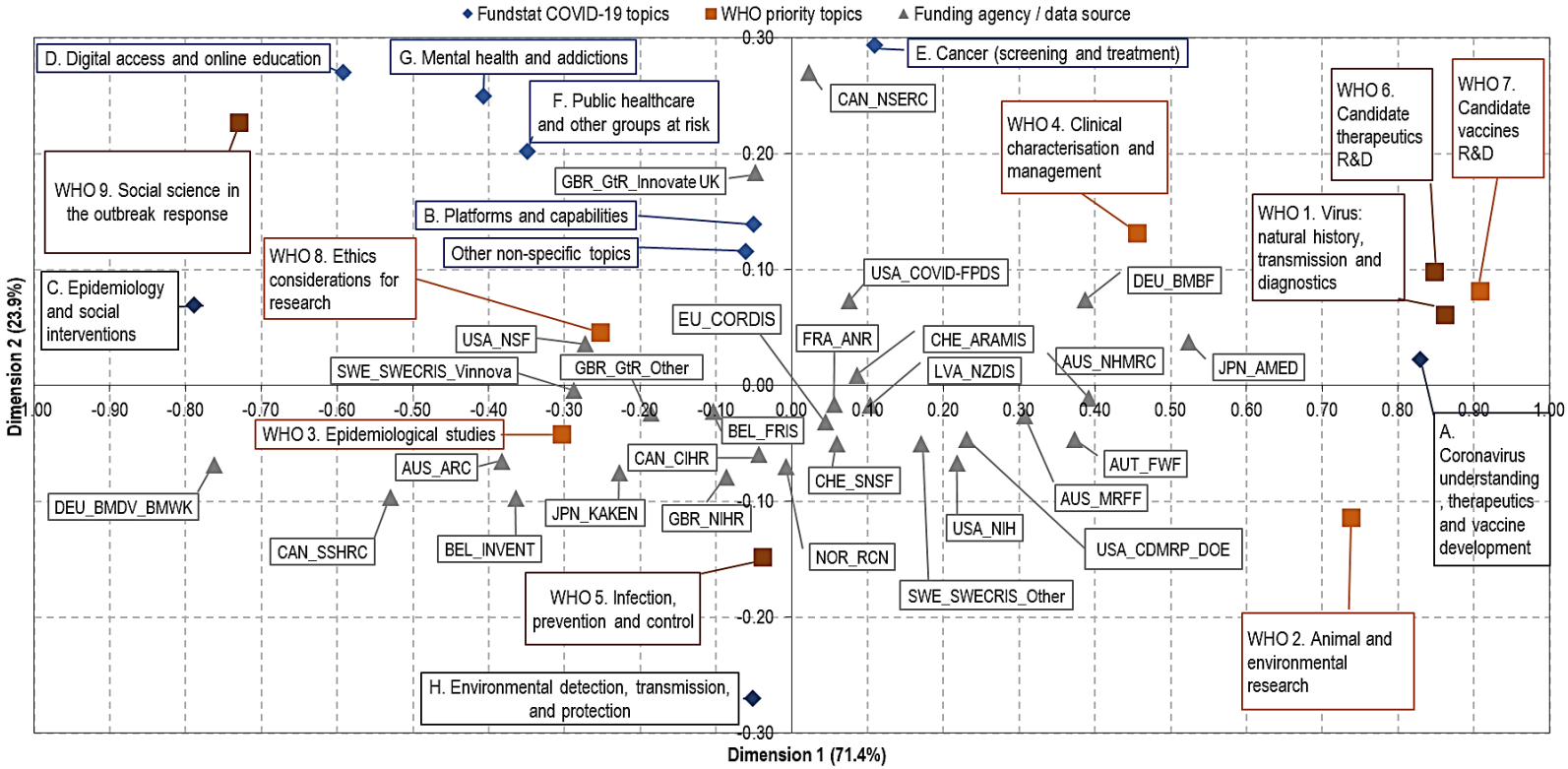
The multiavariate correspondence analysis depicted in Figure 19 attempts to simplify on a 2-dimensional space the respresentation of topic saliency and how the classification systems interrelate as two potentially correlated categorical variables, adding as a third variable the identity of funding agencies (or funding data source), thus helping illustrate the latter's funding orientation. 95.2% of the variance in the associated correspondence matrix can be explained with only two factors.

The position of the combined profiles of the Fundstat C8 classification and WHO topics reveals some relevant association patterns. For instance, cluster A ('Coronavirus understanding, therapeutics, and vaccine development') has a strong positive alignment (both in terms of correlation and intertia contribution) with Dimension axis 1, as well as WHO 1 ('Virus: natural history, transmission and diagnostics'), and WHO 6 ('Candidate therapeutics R&D'). At the opposite end of the same scale, Cluster C ('Epidemiology and social interventions') and WHO 9 ('Social science in the outbreak response') occupy a similar space with reference to Dimension axis 1. Both classifications appear to partition project content in similar ways along a continum that separates biomedical and social science R&D related to COVID-19 (OECD, 2023[42]), with WHO being more detailed in relation to biomedical topics that the machine classification does not identify in an equally distinctive fashion.

The Fundstat classification provides greater granularity of insight in some topics that are not appreciable in the headline list of WHO priorities. For example, Fundstat cluster E ('Cancer (screening and treatment'), similarly to cluster G ('Mental health and addictions') (Astorga-Pinto, Hewlett and Haywood, 2023[40]) are particularly salient but show no clear match other than a loose connection with social sciences (WHO 9) through the Dimension 2. Items with high values in dimension 2 have in common a focus on target and at risks groups, like those already mentioned and Cluster D ('Digital access and online education'). Low values of Dimension 2 are found for cluster H ('Environmental detection, transmission and protection) and WHO 5 ('Infection, prevention and control') which appear to exhibit a high mutual correlation and fall way in the middle of the Dimension 1.

Furthermore, the correspondence analysis allows for a better understanding of the relationship between the different topics and profile of the funding agencies, assisting with the interpretation of the topic classification results presented in section 3.2.2. The analysis shows that there is a strong association between Fundstat topics related to cluster A ('Coronavirus understanding, therapeutics and vaccine development') and WHO priority topics related to WHO 1 ('Virus natural history, transmission and diagnostics'), WHO 6 ('Candidate therapeutics R&D'), and WHO 7 ('Candidate vaccines R&D') with medical/health agencies. Similarly, there is a strong association between Fundstat topics related to cluster C ('Epidemiology and social interventions'), WHO 3 ('Epidemiological studies') and WHO 9 ('Social science in the outbreak response'), with innovation agencies.

Figure 19. Correspondence analysis of COVID-19 R&D projects across funding agencies, Fundstat C8 clusters and WHO priority topics



Notes: [1] Multi-variate correspondence analysis based on correspondence table for 10,472 projects in the Fundstat database with known funding information, with 3 categorical variables: WHO priority topics, Fundstat machine-generated topic clusters and agencies/sources. The two factors depicted account for 95.2% of the variance in the correspondence table. The following category items have been scaled to fit within the chart, with original coordinates presented in brackets: D. Digital access and online education (-0.99, 0.45); H. Environmental detection, transmission, and protection (-0.33, -1.76); WHO 5. Infection, prevention, and control (-0.27, -1.05); and CAN_NSERC (0.06, 0.39).

[2] The most important profiles have a correlation with the dimension axis greater than 0.13 and a contribution to the dimension axis inertia greater than 0.11. For Dimension 1 axis, the most important profiles are WHO 1, 6, and 9 (dark orange), and Clusters A and C (dark blue). For Dimension 2 axis, the most important profiles are WHO 5 and 9 (dark orange) and Cluster H (dark blue).

Source: OECD analysis of Fundstat database, March 2023.

There is significant evidence supporting the rationale for combining information from the machine-driven Fundstat classification and the WHO taxonomy. While the correspondence analysis shows that there are areas of overlap and alignment between the two systems, each reveals unique aspects of analytical and policy relevance. The unsupervised classification scheme provides unique and valuable insights into salient topics like some application domains and platforms (topic B), while the WHO taxonomy provides a more formal structure and greater granularity within the biomedical domain that attracts the bulk of the funding.

Figure 20 tabulates funding and projects using combined categories for C8 clusters and the WHO priorities, indicating that funding for R&D on COVID-19 was largely focused on the biomedical response to the coronavirus pandemic, with vaccines and therapeutics attracting most of the funding followed by efforts to build understanding of the virus in terms of biology, diagnostics, and clinical characterisation. Funding for social science funding was also significant but significantly less than implied by the sheer number of projects. R&D in support of Platforms and capabilities appears to have been disseminated across a wider range of WHO topics.

By combining unsupervised classification with models that emulate expert classficiation systems, researchers and funders can gain a more comprehensive understanding of the R&D response to COVID-19, identifying areas that may not be receiving as much attention in either classification system alone, and informing research priorities and funding decisions.

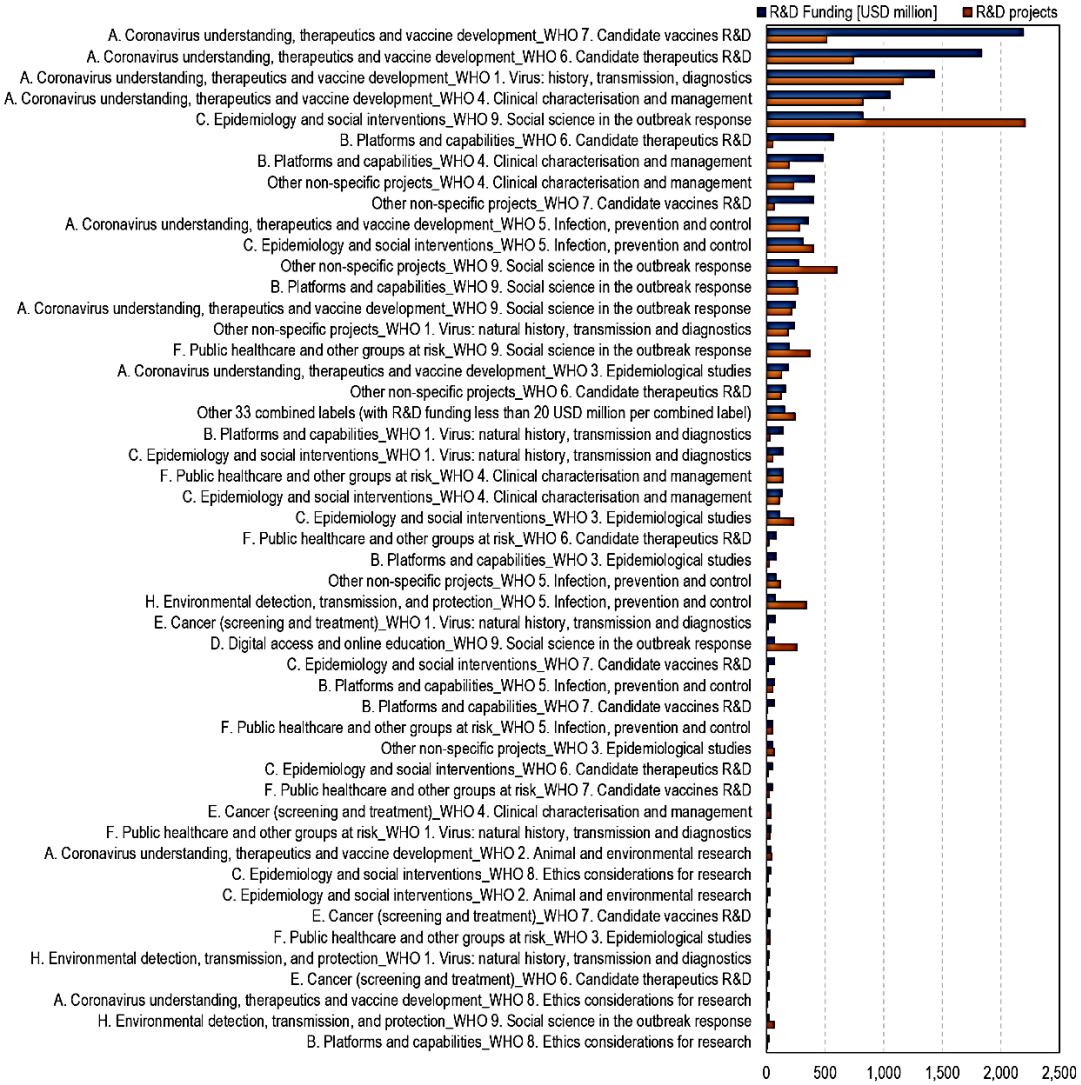
4.2. R&D funding and scientific publications on COVID-19

Data on scientific publications provide an additional benchmark for the funding analysis presented in this paper. While they lack the funding information that is key to the Fundstat initiative, topic classification comparisons can help illustrate both issues of coverage and intrinsic differences between R&D inputs and outputs data for topic modelling. The scientific publication data, with its own coverage biases, is more geographically comprehensive than the Fundstat database on funding and it is potentially inclusive of R&D that has been funding through institutional mechanisms leaving no R&D project funding trace. Their relevance as measure of R&D directionality is however more confined to academic research, as it does not necessarily capture all outputs from government funded R&D projects, particularly those with a more commercial or practical orientation.

4.2.1. Identification of COVID-19 publications in the Scopus database

The methodology used to extract COVID-19 publications from the Scopus database is analogous to that used for COVID-19 R&D projects, with both involving the use of key terms. In addition, data on the presence of funding acknowledgement in publications provides an additional indicator of the type of funding supporting the research leading to the publication. Separating the analysis for publications with and without such acknowledgements assists in the attempt to compare publications that may be funded by projects in the Fundstat database and those that result from institutionally funded research alone. A total of 205,053 'candidate' COVID-19 publications indexed between 2019-2021 were extracted and carefully screened to eliminate documents with only contextual references to COVID-19. The screening involved the construction of a contextual classifier, similar but not identical to the classifier used for COVID-19 R&D project awards.[17] A total number of 200,763 COVID-19 publications was finally retained (Table 13). The share of contextual publications to the total 'candidate' publications is 2%, significantly less than for R&D projects.

Figure 20. Combined label classification of Fundstat C8 clusters and WHO priority topics



Notes:
[1] Labels represent the Fundstat topic modelling C8 clusters combined with WHO priority topics.
[2] Analysis is based on 10,472 projects in the Fundstat database with known funding information.
Source: OECD analysis of Fundstat database, March 2023.

Table 13. Retention of COVID-19 Scopus publications, 2019-21

COVID-19 Publications (2019-2021)[1]	Number of publications
Total 'candidate' publications	205,053
(-) 'Candidate' contextual publications manually tagged by the OECD-STI team	1,324
(-) 'Candidate' contextual publications tagged by classifier[2]	2,966
Total 'retained' COVID-19 publications	200,763

Notes:
[1] Data retrieval with key terms from Table 3, from Elsevier's Scopus database (Scopus 2023), constrained to publication articles (ar), review articles (re), and conference papers (cp).
[2] The screening involved the construction of a contextual classifier, which is firstly trained on the R&D contextual projects (from Table 5), followed by training and fine-tuning on a set of publications manually reviewed and tagged as contextual by the OECD-STI team.
Source: OECD analysis of Elsevier's Scopus database (Scopus 2023), March 2023.

Table 14 provides a summary of the COVID-19 'retained' publications by year and the acknowledgement of funding. The number of publications increased dramatically from 2020 to 2021, particularly for those with funding acknowledgements. Well over a third of COVID-19 publications acknowledge some form of funding.

Table 14. COVID-19 Scopus publications, 2019-21

Funding acknowledgement	2019	2020	2021	Total
Total publications	343	63,050	137,370	200,763
(Of which) Publications with funding acknowledgement	203	20,324	54,234	74,761

Notes: Data retrieval with key terms from Table 3, from Elsevier's Scopus database (Scopus 2023), constrained to publication articles (ar), review articles (re), and conference papers (cp).
Source: OECD analysis of Elsevier's Scopus database (Scopus 2023), March 2023.

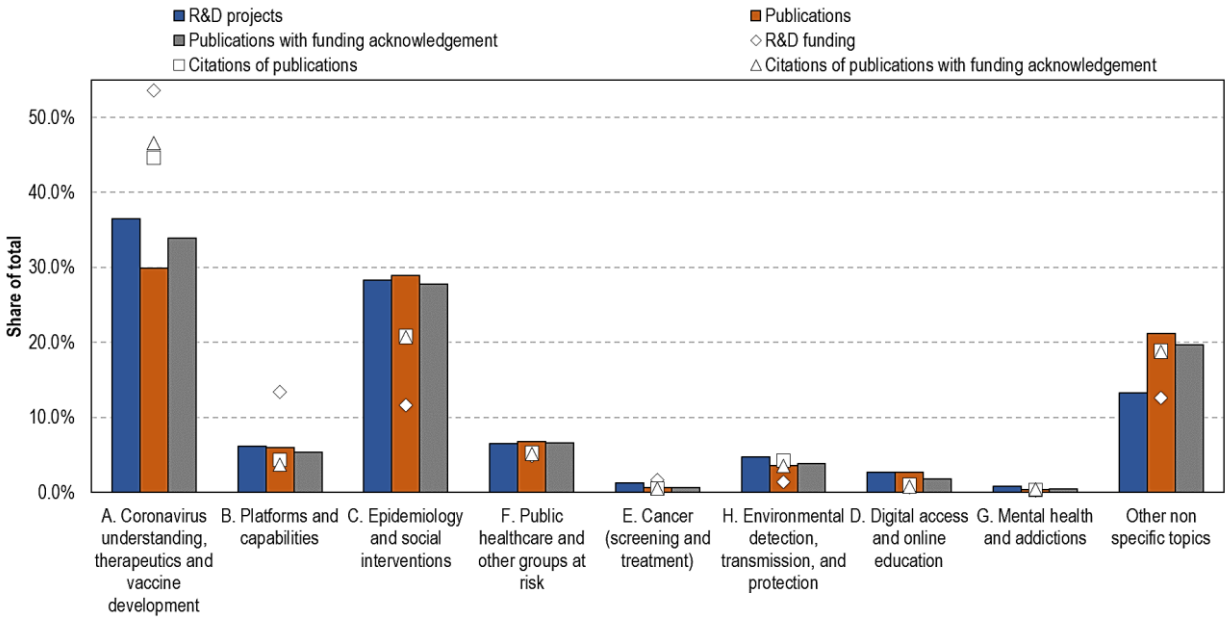
4.2.2. Comparison with COVID-19 R&D project data

The COVID-19 retained publications from the Scopus database were allocated to (i) the unsupervised Fundstat C8 clusters by an 'out-of-corpus' prediction using the model initially developed for COVID-19 R&D projects; and (ii) the WHO priority topics using the model developed on the COVID-19 tracker and applied to the Fundstat data in section 4.1. The results of the analysis presented in Figure 21 reveal that scientific publications appear to present a slightly higher degree of non-specificity than R&D projects, which is understandable since the unsupervised topic model was developed drawing on a different corpus. Asides from that, topic allocations in terms of counts of records are very similar for R&D projects and publications, particularly if one focuses on publications with funding acknowledgements. Publication counts therefore provide similar information to project counts and are as a result not a particularly accurate indication of funding allocations by topic. A publication-based indicator exhibits a somewhat closer relationship with R&D funding, namely the topic distribution of counts of forward citations to papers. The share of citations heading towards publications in key topic cluster A is over 45% for papers with funding acknowledgements, still below 55% of funding but higher than its 35% share of publications. Citations however do not appear to reflect the full funding significance of Topic B on Platforms, whereas Topic C still attracts 20% of citations compared with about 12% of funding.

A similar analysis of COVID-19 scientific publications is performed based on the WHO priority topics (Figure 22). Scientific publishing on WHO 9 ('Social sciences in the outbreak response') accounts for slightly over the equivalent share of R&D projects, and almost the same in the case of publications acknowledging funding. The distribution of citations to scientific publications is dominated instead by WHO 4 ('Clinical characterisation and management') with about 30% of the total, but this is still almost twice the share of this topic in project funding. The two largest fund recipient areas (WHO 6 'Candidate therapeutics R&D' and WHO 7 'Candidate vaccines R&D') are far from those attracting the most citations. Within the biomedical sphere there is considerable heterogeneity, as WHO1 ('Virus natural history, transmission, and diagnostics') receives a disproportionate share of citations relative to the share of publications. Citation results should however be interpreted with caution, as citation windows for publications differ by domain and papers on the nature of the virus and the disease may have had an early citation advantage over other scientific papers which may be cited in years to come.

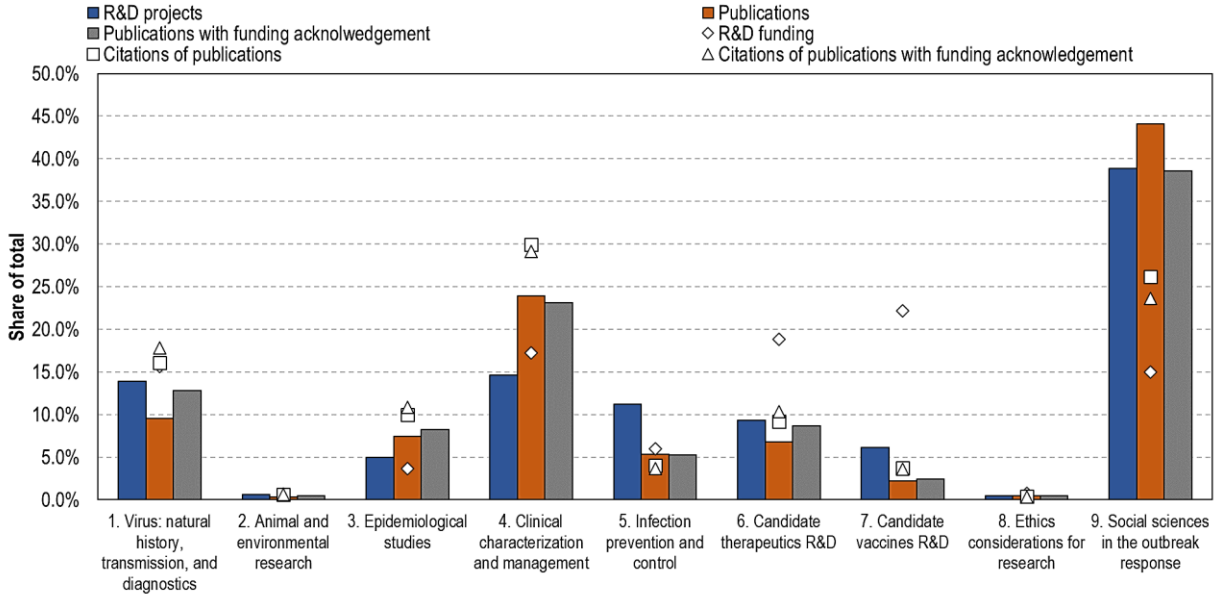
The analysis demonstrates the importance of considering funding data in addition to publication data to gain a more complete understanding of the landscape of R&D related on COVID-19. Although the number of publications with funding acknowledgements is similar to the number of R&D projects in terms of topic distribution from both Figure 21 and Figure 22, the funding allocations reveal a significant difference from count indicators. Specifically, there appears to be a greater investment in biomedical topics regardless of the classification system than what is indicated by project or paper counts alone.

Figure 21. Distribution of COVID-19 R&D and scientific publications by Fundstat C8 cluster



Notes: The analysis is based on (i) 10,472 R&D projects in the Fundstat database with known funding information, and (ii) an 'out-of-sample/corpus' prediction of topics on the retained COVID-19 publications using the COVID-19 R&D project topic model.
Source: OECD analysis of Fundstat and Elsevier's Scopus (Scopus 2023) databases, March 2023.

Figure 22. Distribution of COVID-19 R&D and scientific publications by WHO priority topic



Notes: Analysis is based on (i) 10,472 R&D projects in the Fundstat database with known funding information, and (ii) an 'out-of-sample/corpus' prediction of topics on the retained COVID-19 publications using the COVID-19 tracker WHO topic predictive model.
Source: OECD analysis of Fundstat and Elsevier's Scopus (Scopus 2023) databases, March 2023.

5 Concluding remarks

The work presented in this document has deployed a range of tools for the integrated analysis of quantitative and qualitative data to identify government financial support for COVID-19 R&D. This has been part of an exploratory study of R&D funding directionality, testing the potential of the OECD Fundstat initiative as an analytical data infrastructure for comparative analysis. This project set out to: (i) help provide evidence on the level and composition of COVID-19 R&D funding provided by government agencies in OECD countries and the European Union (EU), contributing to shed light on key aspects of governments' response to the pandemic; (ii) help demonstrate the use of natural language processing (NLP) methods to measure directionality for policy analysis; and (iii) provide a basis for scaling up the OECD Fundstat infrastructure and encourage country engagement, collaboration, and mutual learning.

This study has provided an in-depth analysis of funding for COVID-19 R&D projects approved in 2019-21 representing total funding in the order of about USD 12.59 billion and average funding per project of ca USD 1.20 million. While several of the results presented confirm existing literature and public debate, particularly in relation to the broad understanding of the allocation of COVID-19 R&D resources between biomedical and other topics, there are many findings in this report that call for a re-assessment of some widely held views. The results for example dispel claims that funding for R&D on therapeutics fell considerably short of funding for vaccines. Concrete evidence has also been provided on the relative lack of scale for social-science intensive R&D projects, at least as proxied by size of the funding awards. Another important insight concerns the saliency and economic significance of R&D projects that seek to build platforms and infrastructures for work on different pandemic R&D priorities. This is particularly relevant as health and R&D systems seek to build resilience towards future pandemics or attempt to find ways to apply and repurpose COVID-19 based discoveries and technologies to address other pressing health challenges.

This study has also provided several methodological insights, from addressing the challenges of defining and implementing studies that seek to capture R&D related to specific policy interest, reconciling evidence from multiple data sources (e.g., funding databases and publications, different agencies, and countries, and implementing and reconciling machine-driven unsupervised classification procedures with expert classifications, finding synergies across them. The analysis demonstrates the importance of combining funding information with detailed qualitative project descriptions. In the absence of funding data, the analysis of counts of projects or other items such as publications tends to overstate the relative importance of some topics.

A large body of project abstracts lacks the sufficient semantic detail and structure to provide reliable topic predictions, and the analysis also highlights the challenges of defining the boundaries for the body of projects that are effectively concerned with addressing a particular societal challenge such as COVID-19. One implication is that there would be major analytical benefits from having funding agencies converging towards better harmonised project summary descriptions while progressing towards greater data openness for accountability and analysis. The use of large language models to process text data has provided several valuable lessons on the conduct of such type of work for statistical and policy analysis. With the rising awareness of the challenges and opportunities of generative AI tools, it is imperative to develop and adopt a set of good practices in relation to the production and use of intelligence that draws upon these techniques. The work on the underlying data and methodological infrastructure that has underpinned this work continues in the framework of the NESTI MARIAD group.

Endnotes

[1] For Japan (JPN), GBARD in 2020 includes: (i) a major University Endowment Fund, which has been reported in reference years 2020 and 2021 as General University Funds (GUF); and (ii) a 10-year green innovation fund, which has been reported in reference year 2020 (Source: OECD GBARD Sources and Methods Database, available at https://rdmetadata.oecd.org/).

[2] The base set of key terms in Table 3 are lowercased and lemmatised. Lemmatisation involves grouping together the different forms of a word and analysing them as a single element (known as 'lemma').

[3] The Word2Vec methodology converts words into vectors, known as embeddings, which represent their context/meaning. This approach learns relationships between words and utilizes those relationships to identify similar words. Specifically, a Word2Vec model is trained for each of the 3 medical-related corpora (Elsevier's Scopus Custom database, PubMed publications, and the World Health Organisation's COVID-19 database). The base set of key terms (Table 3) are then projected in the embedding space using each of the trained Word2Vec model, and the words that appear near the base set of key terms represent their context. Using this relationship, additional similar key terms are identified from each of the three corpora, which comprise the extension set of key terms in Table 3.

[4] The English terms are translated into Dutch, German, French, Japanese, Latvian, Norwegian, and Swedish.

[5] During the experimentation stage, the base set of key terms (from Table 3) retrieved 95% of the 'candidate' COVID-19 R&D projects and the extension set of key terms retrieved the remaining 5%, from an earlier version of the Fundstat infrastructure (May 2022) that included data from 10 OECD member countries and the European Commission (originating from 16 funding databases/organisations).

[6] Based on its subject content, a contextual R&D project appears to be unrelated to the pandemic (e.g., not directly addressing the crisis), but contains direct and indirect mentions of COVID-19 either as examples or as references to the post-pandemic recovery and impact.

[7] The R&D contextual project classifier is based on a 2-layer artificial neural network (ANN) (Murphy, 2012[49]). The input to the model is the textual information (combined title and abstract) of the manually tagged R&D projects, which is transformed into an embedding representation using the pre-trained sentence transformer 'distiluse-base-multilingual-cased-v2' (Reimers and Gurevych, 2019[31]). A sentence transformer model maps sentences and paragraphs of a text into a dense vector space. The output to the model is the manual tag of the R&D project being contextual or not, which results in the probability of a project being contextual. 80% of the manually tagged dataset is used for training with k-fold cross validation, while the remaining 20% is used for testing (Aristodemou, 2020[48]). The classifier has a theoretical accuracy of 93% (with std of 1%) in the testing set for identifying contextual projects. The precision of the classifier on a random sample of 100 R&D projects tagged as contextual is 96%.

[8] After data cleaning and post-processing (e.g., removal of stopwords, removal of language-specific stopwords, removal of common phrases, removal of highest frequency non-informative words, lowercasing and lemmatising), there is limited content in the text of the combined title and abstract to infer any type of activity. For example, a project from NIH in USA has only the following in its title 'SARS-CoV-2 Pathogenesis'. Another example, a project from GtR_EPSRC in the UK has only the title 'COVID 19 Grant Extension Allocation University of Cambridge' and an abstract is not available. No attempt has been made to link awards that may correspond to an identical project as unique identifiers were not generally available.

[9] c-TF-IDF (class-Term Frequency - Inverse Document Frequency) is a statistical measure that helps to determine the importance of a term within a specific class of documents, such as a category or a topic. c-TF-IDF considers what makes the documents in one cluster different from documents in another cluster.

[10] Text processing and cleaning includes lower casing the text, and removal of punctuation. It also includes the removal of generic stopwords, language-specific stopwords, common phrases and highest frequency non-informative words. The text is then lemmatised.

[11] To predict the WHO priority topics of the Fundstat COVID-19 R&D projects, a fine-tuned BERT model is utilized with transfer learning, with an additional linear layer for multi-class and multi-label text classification. The COVID-19 tracker data (February 2023, version 9) is used to train the model, using as input the text (combined title and abstract), and as output the WHO primary topic label. The dataset is divided into 80% for training, and 20% for validation. The resulting classifier achieves a F1 score of 73% and a ROC AUC score of 84%.

[12] The OECD-STI team subjected the descriptive labels for the C8 clusters to an interpretation test using ChatGPT (OpenAI, 2023[47]). This is an exploratory test to ensure the coherence of the labels and is performed as a sense check in addition to the complementarity analysis in Section 4. The results are presented in Annex B5.

[13] Two vaccine projects [BNT-Covid-19-Vaccine - Beschleunigte Entwicklung und Bereitstellung eines mRNA-basierten COVID-19-Impfstoffs (BNT162) & (ACE-mR-CoV - Entwicklung, Testung und Produktion eines SARS-CoV-2 Impfstoffes auf Basis der mRNA Technologie)], account for USD 652 million.

[14] The COVID-19 tracker includes projects from regional and national funding agencies, and from non-profit or philanthropic organisations. It also includes projects that have been repurposed to address priorities related to COVID-19. These projects are obtained either through direct communication with research funders by completing a template spreadsheet, or via searching online databases belonging to research funders using the terms 'COVID', 'nCOV', and 'sars-cov-2' (Bucher et al., 2023[14]).

[15] USA Federal procurement for COVID-19 Contract Obligation Tracking Dashboard.

[16] The trained multi-label classifier for predicting the WHO priority topic labels is evaluated on a 10% out-of-sample test dataset from the COVID-19 tracker data, with an F1-score of 70% and a ROC AUC 81%.

[17] The publication contextual classifier is firstly trained on the R&D contextual projects (from Table 5), followed by training and fine-tuning on a set of publications manually tagged as contextual by the OECD-STI team. The resulting classifier achieves a precision of 82%, recall of 82% and accuracy of 80%, on a random sample of 200 publications, of which 50% are predicted as contextual and 50% are predicted as non-contextual.

References

Abadi, H., Z. He and M. Pecht (2020), "Artificial Intelligence-Related Research Funding by the U.S. National Science Foundation and the National Natural Science Foundation of China", *IEEE Access*, Vol. 8, pp. 183448-183459, https://doi.org/10.1109/access.2020.3029231. [28]

Agarwal, R. and P. Gaule (2022), "What drives innovation? Lessons from COVID-19 R&D", *Journal of Health Economics*, Vol. 82, p. 102591, https://doi.org/10.1016/j.jhealeco.2022.102591. [9]

Annapureddy, A. et al. (2020), "The National Institutes of Health funding for clinical research applying machine learning techniques in 2017", *npj Digital Medicine*, Vol. 3/1, https://doi.org/10.1038/s41746-020-0223-9. [27]

Aristodemou, L. (2020), *Identifying Valuable Patents: A Deep Learning Approach*, University of Cambridge, Cambridge, https://doi.org/10.17863/CAM.69403. [48]

Aristodemou, L. and F. Tietze (2018), "The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data", *World Patent Information*, Vol. 55, pp. 37-51, https://doi.org/10.1016/j.wpi.2018.07.002. [26]

Astorga-Pinto, S., E. Hewlett and P. Haywood (2023), "Protecting mental health", in *Ready for the Next Crisis? Investing in Health System Resilience*, OECD Publishing, Paris, https://doi.org/10.1787/0f76c6be-en. [40]

Baden, L. et al. (2021), "Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine", *New England Journal of Medicine*, Vol. 384/5, pp. 403-416, https://doi.org/10.1056/nejmoa2035389. [35]

Berchet, C., E. Barrenho and K. de Bienassis (2023), "Preserving continuity of care", in *Ready for the Next Crisis? Investing in Health System Resilience*, OECD Publishing, Paris, https://doi.org/10.1787/5d015a71-en. [39]

Bucher, A. et al. (2023), "A living mapping review for COVID-19 funded research projects: two year update", *Wellcome Open Research*, Vol. 5, p. 209, https://doi.org/10.12688/wellcomeopenres.16259.9. [14]

EU/OECD (2022), *STIP Compass COVID-19 Watch*, https://stip.oecd.org/covid/ (accessed on 1 May 2023). [11]

Florio, M., S. Gamba and C. Pancotti (2023), "Mapping of long-term public and private investments in the development of Covid-19 vaccines, publication for the special committee on COVID-19 pandemic: lessons learned and recommendations for the future (COVI)", [33]

Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, Luxembourg.

G7 (2023), *G7 Science and Technology Ministers' Communique*, Sendai, Hiroshima Summit, https://www8.cao.go.jp/cstp/kokusaiteki/g7_2023/230513_g7_communique.pdf (accessed on 17 May 2023). [16]

Gaviria, M. and B. Kilic (2021), "A network analysis of COVID-19 mRNA vaccine patents", *Nature Biotechnology*, Vol. 39/5, pp. 546-548, https://doi.org/10.1038/s41587-021-00912-9. [34]

Grootendorst, M. (2022), "BERTopic: Neural topic modeling with a class-based TF-IDF procedure", https://doi.org/10.48550/arxiv.2203.05794. [30]

Guan, W. et al. (2020), "Clinical Characteristics of Coronavirus Disease 2019 in China", *New England Journal of Medicine*, Vol. 382/18, pp. 1708-1720, https://doi.org/10.1056/nejmoa2002032. [43]

He, X. et al. (2020), "Temporal dynamics in viral shedding and transmissibility of COVID-19", *Nature Medicine*, Vol. 26/5, pp. 672-675, https://doi.org/10.1038/s41591-020-0869-5. [44]

INGSA (2020), *INGSA COVID-19 Evidence-to-Policy Tracker*, https://covid.ingsa.org/covid/policymaking-tracker-landing/ (accessed on 1 May 2023). [13]

Jurafsky, D. and J. Martin (2023), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Pearson Education, https://web.stanford.edu/~jurafsky/slp3/ (accessed on 1 May 2023). [25]

Keelara, R. and P. Haywood (2023), "Building the data and digital foundations of health systems", in *Ready for the Next Crisis? Investing in Health System Resilience*, OECD Publishing, Paris, https://doi.org/10.1787/9b8a7ce8-en. [37]

Lalani, H. et al. (2023), "US public investment in development of mRNA covid-19 vaccines: retrospective cohort study", *BMJ*, p. e073747, https://doi.org/10.1136/bmj-2022-073747. [45]

Lopert, R. et al. (2023), "Incentivising the Development of Global Public Goods for Health", in *Ready for the Next Crisis? Investing in Health System Resilience*, OECD Publishing, Paris, https://doi.org/10.1787/e117f1d2-en. [51]

McInnes, L., J. Healy and S. Astels (2017), "hdbscan: Hierarchical density based clustering", *The Journal of Open Source Software*, Vol. 2/11, p. 205, https://doi.org/10.21105/joss.00205. [32]

Mikolov, T. et al. (2013), "Efficient estimation of word representations in vector space", *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, https://doi.org/10.48550/arXiv.1301.3781. [29]

Murphy, K. (2012), *Machine Learning: A Probabilistic Perspective*, The MIT Press. [49]

OECD (2023), "Mobilising science in times of crisis: Lessons learned from COVID-19", in *OECD Science, Technology and Innovation Outlook 2023: Enabling Transitions in Times of Disruption*, OECD Publishing, Paris, https://doi.org/10.1787/855c7889-en. [52]

OECD (2023), *OECD Science, Technology and Innovation Outlook 2023: Enabling Transitions in Times of Disruption*, OECD Publishing, Paris, https://doi.org/10.1787/0b55736e-en. [4]

OECD (2023), *Ready for the Next Crisis? Investing in Health System Resilience*, OECD Health Policy Studies, OECD Publishing, Paris, https://doi.org/10.1787/1e53cf80-en. [10]

OECD (2023), "Science, technology and innovation policy in times of global crises", in *OECD Science, Technology and Innovation Outlook 2023: Enabling Transitions in Times of Disruption*, OECD Publishing, Paris, https://doi.org/10.1787/d54e7884-en. [42]

OECD (2021), *OECD Science, Technology and Innovation Outlook 2021: Times of Crisis and Opportunity*, OECD Publishing, Paris, https://doi.org/10.1787/75f79015-en. [5]

OECD (2021), *Strengthening Economic Resilience Following the COVID-19 Crisis: A Firm and Industry Perspective*, OECD Publishing, Paris, https://doi.org/10.1787/2a7081d8-en. [8]

OECD (2018), "Blue Sky perspectives towards the next generation of data and indicators on science and innovation", in *OECD Science, Technology and Innovation Outlook 2018: Adapting to Technological and Societal Disruption*, OECD Publishing, Paris, https://doi.org/10.1787/sti_in_outlook-2018-19-en. [19]

OECD (2018), *OECD Science, Technology and Innovation Outlook 2018: Adapting to Technological and Societal Disruption*, OECD Publishing, Paris, https://doi.org/10.1787/sti_in_outlook-2018-en. [21]

OECD (2015), *Daejeon Declaration on Science, Technology, and Innovation Policies for the Global and Digital Age*, https://www.oecd.org/sti/daejeon-declaration-2015.htm (accessed on 1 May 2023). [18]

OECD (2015), *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris, https://doi.org/10.1787/9789264239012-en. [20]

OECD (2015), "Government budget allocations for R&D", in *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, OECD Publishing, Paris, https://doi.org/10.1787/9789264239012-14-en. [17]

OECD/European Union (2022), "Coping with COVID-19: Young people's health in an age of disruption", in *Health at a Glance: Europe 2022: State of Health in the EU Cycle*, OECD Publishing, Paris, https://doi.org/10.1787/d64085ce-en. [41]

OECD/Eurostat (2018), *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition*, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris/Eurostat, Luxembourg, https://doi.org/10.1787/9789264304604-en. [23]

OpenAI (2023), *ChatGPT*, Version 3.5, Personal Communication, https://chat.openai.com/chat (accessed on 1 May 2023). [47]

Policy Cures Research (2020), *COVID-19 R&D tracker*, https://www.policycuresresearch.org/covid-19-r-d-tracker (accessed on 1 May 2023). [12]

Reimers, N. and I. Gurevych (2019), "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 3982-3992, https://doi.org/10.48550/arxiv.1908.10084. [31]

Tietze, F. et al. (2022), "Crisis-Critical Intellectual Property: Findings From the COVID-19 Pandemic", *IEEE Transactions on Engineering Management*, Vol. 69/5, pp. 2039-2056, https://doi.org/10.1109/tem.2020.2996982. [7]

UKCDR & GloPID-R (2023), *COVID-19 Research Project Tracker*, http://www.doi.org/10.7910/DVN/ARPEED. [15]

Veugelers, R., J. Wang and P. Stephan (2022), *Do Funding Agencies Select and Enable Risky Research: Evidence from ERC Using Novelty as a Proxy of Risk Taking*, National Bureau of Economic Research, Cambridge, MA, https://doi.org/10.3386/w30320. [24]

Vincent-Lancrin, S. (2022), "Educational innovation and digitalisation during the COVID-19 crisis: lessons for the future", in *How Learning Continued during the COVID-19 Pandemic: Global Lessons from Initiatives to Support Learners and Teachers*, OECD Publishing, Paris, https://doi.org/10.1787/93c3dc5e-en. [38]

WHO (2023), *Statement on the fifteenth meeting of the IHR (2005) Emergency Committee on the COVID-19 pandemic*, https://www.who.int/news/item/05-05-2023-statement-on-the-fifteenth-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-coronavirus-disease-(covid-19)-pandemic (accessed on 16 May 2023). [1]

WHO (2023), *WHO Coronavirus (COVID-19) Dashboard*, https://covid19.who.int/ (accessed on 1 May 2023). [3]

WHO (2022), *COVID-19 Research and Innovation: Powering the world's pandemic response – now and in the future*, https://www.who.int/publications/m/item/covid-19-research-and-innovation---powering-the-world-s-pandemic-response-now-and-in-the-future (accessed on 1 May 2023). [50]

WHO (2022), *WHO's response to COVID-19: 2022 Mid-Year Report*, https://www.who.int/publications/m/item/who-s-response-to-covid-19-2022-mid-year-report (accessed on 1 May 2023). [2]

WHO (2020), *A Coordinated Global Research Roadmap*, 2019 Novel Coronavirus, https://www.who.int/publications/m/item/a-coordinated-global-research-roadmap (accessed on 1 May 2023). [6]

WHO (2020), *World experts and funders set priorities for COVID-19 research*, https://www.who.int/news/item/12-02-2020-world-experts-and-funders-set-priorities-for-covid-19-research (accessed on 1 May 2023). [46]

Xu, Z. et al. (2020), "Pathological findings of COVID-19 associated with acute respiratory distress syndrome", *The Lancet Respiratory Medicine*, Vol. 8/4, pp. 420-422, https://doi.org/10.1016/s2213-2600(20)30076-x. [36]

Yamashita, I. et al. (2021), "Measuring the AI content of government-funded R&D projects: A proof of concept for the OECD Fundstat initiative", *OECD Science, Technology and Industry Working Papers*, No. 2021/09, OECD Publishing, Paris, https://doi.org/10.1787/7b43b038-en. [22]

Annex A. Features of the OECD Fundstat database

Table A.1. Coverage of the OECD Fundstat database, 2019-2021

Geo code	Country/ area name	Government funding organisation / data source	Number of projects	Funding [USD million]
AUS	Australia	Australian Research Council (ARC)	3,683	1,692
		Medical Research Future Fund (MRFF)	567	1,150
		National Health and Medical Research Council (NHMRC)	2,237	1,876
AUT	Austria	Austrian Science Fund (FWF)	2,080	759
BEL	Belgium	Flanders Research Information Space (FRIS)	16,399	2,933
		Belgian Science Policy Office (BELSPO) INVENT database[1]	2,477	435
CAN	Canada	Canadian Institute of Health Research (CIHR)	6,814	1,730
		Natural Sciences and Engineering Research Council (NSERC)	48,895	2,039
		Social Sciences and Humanities Research Council (SSHRC)	22,297	1,361
CHE	Switzerland	Administration Research Actions Management Information Systems (ARAMIS)	4,871	1,318
		Swiss National Science Foundation (SNSF)	9,661	3,173
DEU	Germany	Federal Ministry of Education and Research (BMBF)[2]	55,541	38,637
		German Research Foundation (GEPRIS)[3]	257	-
FRA	France	National Research Agency (ANR)	5,891	5,166
GBR	United Kingdom	Gateway to Research (GtR)[4]	35,588	14,951
		National Institute for Health Research (NIHR)	1,771	1,872
JPN	Japan	Agency for Medical Research and Development (AMED)	5,393	2,711
		Database of Grants-in-Aid for Scientific Research (KAKEN)[5]	95,143	6,658
LVA	Latvia	Latvian Council of Science (NZDIS)	3,898	401
NOR	Norway	Research Council Norway (RCN)	4,768	2,224
SWE	Sweden	Swedish Current Research Information System (SWECRIS)[6]	11,403	4,006
USA	United States	Congressionally Directed Medical Research Programs (CDMRP)	2,716	2,801
		Department of Energy (DOE)	2,644	6,191
		National Institute of Health (NIH)[7]	207,854	118,693
		National Science Foundation (NSF)	37,283	17,715
		Federal Procurement Data System (FPDS)[8]	2,608	4,604
EC-EU	European Commission	Community Research and Development Information Service (CORDIS)	14,362	42,483
Total			607,098	287,579

Notes:

[1] The INVENT database in its composition also includes the FRIS database. However, for the purpose of this table, it excludes the information already provided directly from FRIS. The funding amount provided by BELSPO is an approximative number, obtained via extrapolation.

[2] BMBF also includes BMDV, BMEL, BMUV, BMWK.

[3] GEPRIS data identified by the German Research Foundation as connected to COVID-19. The project funding amounts have not been provided.

[4] GtR includes Innovate UK, ESRC, NERC, EPSRC, BBSRC, MRC, AHRC, STFC, NC3Rs, UKRI.

[5] JPN_KAKEN database is principally covering funding from the Japan Society for the Promotion of Science (JSPS).

[6] SWECRIS includes FBEES, IFAU, VR, Forte, Formas, Vinnova, SHLF, RJ, SWEA.

[7] NIH also includes CDC, FDA, AHRQ, VA.

[8] Federal procurement award data identified by US authorities as connected to COVID-19, covering multiple government agencies. This is based on the COVID-19 Contract Obligation Tracking Dashboard for all government-wide Emergency Acquisitions spending in support of the COVID-19 pandemic.

[9] FY2021 is partial because of data reporting structures and data availability across different countries. Specifically, FY2021 data are unavailable for BEL_INVENT, CAN_CIHR, CAN_NSERC, CAN_SSHRC, DEU_COVID-GEPRIS, JPN_AMED, and NOR_RCN.

Source: OECD analysis, Fundstat database, March 2023.

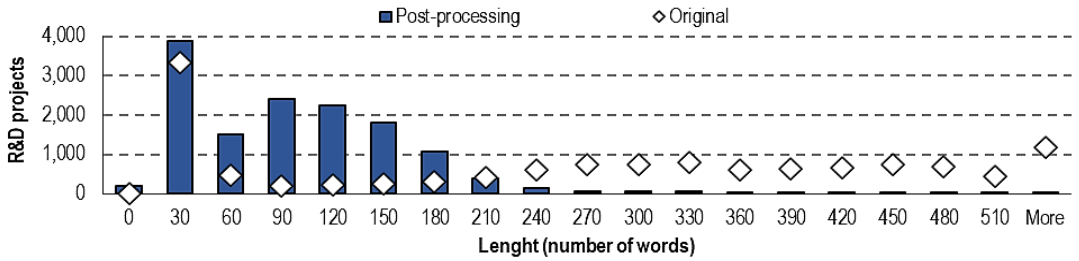
Annex B. Bias analysis and robustness checks

This study conducted several analyses to ensure the robustness of the results and to identify potential biases. This study reports on the process of processing the project's text description field (combined title and abstract), the sensitivity analysis on the number of meaningful words to keep for topic modelling and on handling multiple languages in a multilingual topic model, the degree of topic mixing, and the sense check verification of the cluster labels using ChatGPT.

B.1. Length (number of words) of original and processing text fields

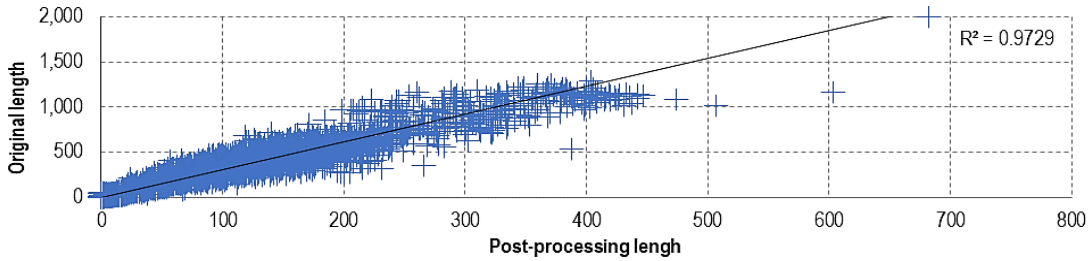
Figure B.1 shows the distribution of the R&D project text description's original and processed length. The text's length is defined as the number of words found in the combined title and abstract.

Figure B.1. Distribution of length (in number of words) in projects' combined title and abstract



Notes:
[1] Length includes the R&D project's combined title and abstract.
[2] The original length of the R&D project is the title and abstract sourced directly from the funding agency/data source.
[3] Text processing involves the removal of stopwords, removal of language-specific stopwords, removal of common phrases, removal of highest frequency non-informative words, lowercasing and lemmatising.
Source: OECD analysis of Fundstat database, March 2023.

Figure B.2. Relationship of original vs. post-processing length (in number of words)



Notes:
[1] Length includes the R&D project's combined title and abstract.
[2] The original length of the R&D project is the title and abstract sourced directly from the funding agency/data source.
[3] Text processing involves the removal of stopwords, removal of language-specific stopwords, removal of common phrases, removal of highest frequency non-informative words, lowercasing and lemmatising.
Source: OECD analysis of Fundstat database, March 2023.

The original length of R&D projects is cleaned and pre-processed by removing: (i) general stopwords, (ii) language-specific stopwords, (iii) common phrases, and (iv) highest frequency non-informative words. The text is also lowercased, and lemmatisation is applied. Figure B.2 shows the relationship between original length and post-processing length, which follows a linear pattern with an r^2 value of 0.97. Table B.1 shows examples of project's original and post-processing text.

Table B.1. Examples of projects' original and post-processed length of text

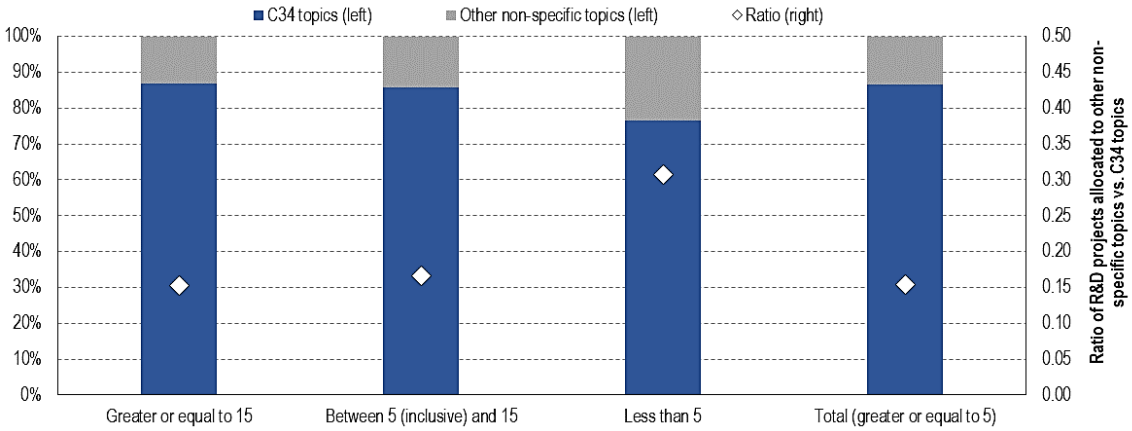
Country/ area	Funding agency/ database	Funding [USD million]	Original combined title and abstract (excerpt)	Number of words in original text	Post-processing combined title and abstract (excerpt)	Number of words in post-processed text
DEU	BMBF	428.33	Infektion. BNT-Covid-19-Vaccine - Beschleunigte Entwicklung und Bereitstellung eines mRNA-basierten COVID-19-Impfstoffs (BNT162)	11	infektion covid-19 vaccine beschleunigte entwicklung bereitstellung basierten impfstoffs	8
USA	NIH	393.95	HVTN 405/HPTN 1901 Characterizing SARS-CoV-2-specific immunity in convalescent individuals… HIV Vaccine Trials Network (HVTN), the collaboration of physician scientists at 64 clinical trial sites in 15 countries on 4 continents dedicated to developing globally effective vaccines for HIV, tuberculosis and now SARS-CoV-2. The HVTN has led HIV prevention science for over 20 years through robust phase 1 and 2 clinical development trials and currently has 2 vector based vaccines…	238	characterize sars-cov-2 immunity convalescent outline scientific agenda vaccine trial physician scientist clinical trial continent dedicate globally vaccine tuberculosis prevention science robust clinical development trial vector vaccine broadly neutralize monoclonal antibody undergo randomize efficacy trial...	67
EU	CORDIS	171.33	PROPOSAL FOR FUNDING RESEARCH DEVELOPMENT AND MANUFACTURING OF VACCINE COVID-19. Funding of research and innovation programmes for the development of novel SARS-CoV-2 vaccine against COVID-19 disease…	321	development manufacture vaccine covid-19 innovation development vaccine disease…	77
GBR	NIHR	30.00	RNA COVID-19 Vaccine Clinical Trial. A vaccine is critical to tackling coronavirus. The clinical and scientific communities… isolation, social distancing and testing can get the world through the current coronavirus problem, the only long-term solution to beating the disease will be finding a vaccinecollective effort of government, academia, industry and healthcare. We know that traditionally vaccine development can take years and we also know more fail than succeed. To accelerate development, government will have a key role to play in derisking projects, by funding their early stage R&D and clinical trialing, and in corralling industry to make sure we have the manufacturing capacity and effective supply chains to produce vaccines at scale, quickly...	470	rna covid-19 vaccine clinical trial vaccine tackle coronavirus clinical scientific increasingly whilst isolation distance coronavirus problem beat vaccine successful vaccine collective government academia healthcare traditionally vaccine development succeed accelerate development government derisking stage clinical trialling corral manufacture capacity supply chain produce vaccine...	133

Notes:
[1] Length includes the R&D project's combined title and abstract.
[2] The original length of the R&D project is the title and abstract sourced directly from the funding agency/data source.
[3] Text processing involves the removal of stopwords, removal of language-specific stopwords, removal of common phrases, removal of highest frequency non-informative words, lowercasing and lemmatising.
Source: OECD analysis of Fundstat database, March 2023.

B.2. Sensitivity analysis on the meaningful length

Figure B.3 shows the share of projects allocated to C34 topics and other non-specific topics by category of meaningful length. There are four categories for the number of meaningful words in R&D projects: (i) greater or equal to 15, (ii) between 5 and 15 (inclusive), (iii) less than 5, and (iv) total (greater or equal to 5) meaningful words. Categories (i), (ii) and (iv) seem to be similar, with approximately the same share of projects allocated to C34 topics and other non-specific topics. However, for projects with less than 5 meaningful words, the share of projects allocated to non-specific topics is significantly higher. Looking at the ratios of R&D project allocated to other non-specific topics vs. projects allocated to the C34 topics, the ratio of meaningful length less than 5 words appears to be more than double. Overall, the ratios suggest that R&D projects related to C34 topics with greater or equal to 5 meaningful words contain concise and relevant content and is meaningful for classification by the topic model. This also means that projects with less than 5 meaningful words can be regarded as limited in content with inference of any type of research activity or topic being difficult.

Figure B.3. Sensitivity analysis on the length (number of meaningful words) of post-processed project text (combined title and abstract) by project topic classification



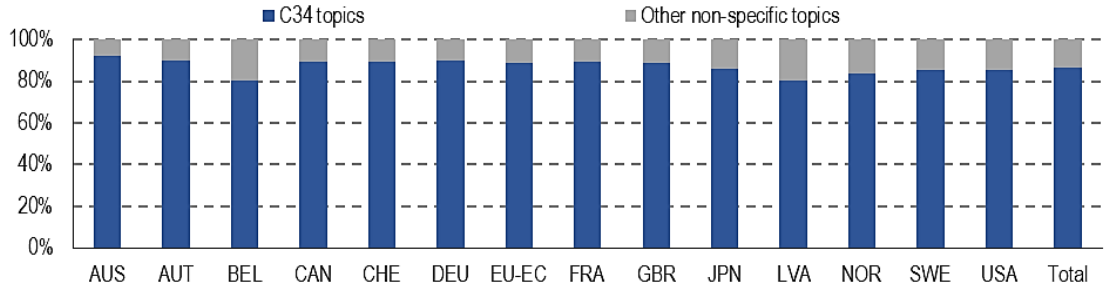
Notes:
[1] The quality of the topic model to cluster projects in topics, i.e., its topic 'specificity', seems to be affected by the length of the meaningful words in the project text (combined text and abstract).
[2] The topic model tends to classify projects with less than 5 meaningful words in other non-specific topics, giving an indication that the limited content is a potential barrier in inferring any research activity or topic (see Table 5).
Source: OECD analysis of Fundstat database, March 2023.

B.3. Sensitivity analysis on handling multiple languages

Since from Section 2.2.2, a multilingual embedding and topic model has been used to generate the topics, it is important to compare the degree of topic 'specificity' across the different countries and languages contributing to the COVID-19 funding database. The results in Figure B.4 indicate that the model performs similarly, with the share of projects in the residual 'non-specific' category being always less than 20% across countries with text data in different languages, apart from Belgium and Latvia. Figure B.5 presents the topic 'specificity' across the different languages. This is performed across two different language models: a machine-translated (MT) model in English and a multilingual (ML) model covering a total of 8 languages (English, Japanese, German, French, Latvian, Swedish, Dutch, and Norwegian). The MT model is developed with the same methodology as the ML model, with the difference that all titles and abstracts

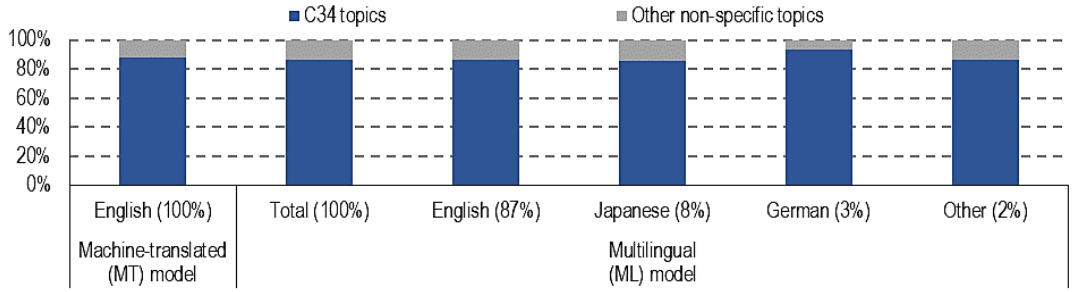
are translated in English using the latest Google Translate API. It is evident that the ML and MT model perform equally well, possibly since most projects are in English already, despite the translation error introduced by the MT model. Topic 'specificity' is better in the multilingual model for the German language.

Figure B.4. COVID-19 R&D projects by C34 and other non-specific topics for each country



Notes:
[1] The quality of the topic model to cluster projects in topics, i.e., its topic "specificity", seems to be unaffected by the differences across countries and writing style.
[2] Within funding agency/data sources in countries, projects in more than one language may exist.
Source: OECD analysis of Fundstat database, March 2023.

Figure B.5. COVID-19 R&D projects by C34 and non-specific topics by model type and language



Notes:
[1] The quality of the topic model to cluster projects in topics, i.e., its topic "specificity", seems to be unaffected by the differences across models (machine-translated vs. multilingual).
[2] The machine-translated model is trained on projects' text (combined title and abstract) translated in English using the Google Translate API.
[3] The percentages next to each language represent the number of R&D projects in that specific language in the data.
[4] Other languages include French, Latvian, Swedish, Dutch, and Norwegian.
Source: OECD analysis of Fundstat database, March 2023.

B.5. Sense-check of cluster labels using ChatGPT

Table B.2 shows the tests conducted using ChatGPT as a sense check for the cluster labels (OpenAI, 2023[47]). As a reminder from Table 9, the OECD-STI team labelled the C8 cluster topics by reviewing: (i) the top 10 salient words in each cluster; (ii) the complementarity of these clusters with the WHO priority topics using a combination of the machine classification and the human expert classification; and (iii) the top funded R&D projects per cluster.

Three tests are conducted, where ChatGPT is prompt to propose labels for the clusters having the following information: (i) the cluster's top 5 salient words (Table 9) and the top 10 salient words of the topics belonging into the cluster (Table C.2); (ii) the data origin, i.e. analysis of the title and abstract in R&D funding projects, in addition to the above information; and (iii) complementary information such as

information on WHO priority topics or the text from the top 10 funded R&D projects in the cluster, in addition to all of the above information in the previous two steps. From Table B.2, the COVID-19 topic modelling and ChatGPT proposed labels that are similar. Difference is evident in cluster B, when there are multiple cross-horizontal projects, with ChatGPT proposing the terminology 'infrastructure', which captures a wider aspect of 'platforms and capabilities'. It is also clear that there is a fine balance between the amount of information provided to ChatGPT, so that when the prompt includes too little or too much information, the labels tend to be more generic for distinct clusters.

Table B.2. Sense check test of the cluster labels using ChatGPT

Cluster	COVID-19 Topic modelling label	ChatGPT proposed label	Information provided to ChatGPT in the prompt			
			Cluster top 5 salient words	Cluster topics top 10 salient words	Data origin (R&D funding projects)	Additional information provided
A	Coronavirus understanding, therapeutics, and vaccine development	Vaccine and Diagnostics Development for COVID-19	x	x		
		COVID-19 Treatments and Diagnostics Research	x	x	x	
		COVID-19 Treatment, Diagnostics, and Vaccines Research and Development	x	x	x	WHO topics
B	Platforms and capabilities	Medical and Management Solutions for COVID-19	x	x		
		COVID-19 Healthcare Management R&D	x	x	x	
		COVID-19 Research and Development Infrastructure	x	x	x	Text of the top 10 funded R&D cluster projects
		COVID-19 Research & Development Strategies	x	x	x	WHO topics
C	Epidemiology and social interventions	COVID-19 Economic Impact, Policy Interventions and Epidemiology	x	x		
		COVID-19 Policy, Behavioral and Epidemiological Implications	x	x	x	
		Economic, Social, and Scientific Aspects of the COVID-19 Pandemic	x	x	x	WHO topics
D	Digital access and online education	Impact of COVID-19 on Education	x	x		
		Digital Remote Education During COVID-19	x	x	x	
		COVID-19 Educational and Research Strategies	x	x	x	WHO topics
E	Cancer (screening and treatment)	Cancer and COVID-19 Research and Interventions	x	x		
		Disparities in Cancer Screening and Intervention for COVID-19	x	x	x	
		COVID-19 Research and Development	x	x	x	WHO topics
F	Public health and other groups at risk	Vulnerable Populations and COVID-19 Interventions	x	x		
		COVID-19 Interventions for Vulnerable Populations	x	x	x	
		Research and Development in Relation to COVID-19	x	x	x	WHO topics
G	Mental health and addictions	Mental Health Effects of COVID-19	x	x		
		COVID-19 Mental Health Effects Research	x	x	x	
		Mental Health Implications of COVID-19	x	x	x	WHO topics
H	Environmental detection, transmission, and protection	COVID-19 Prevention, Detection and Spread	x	x		
		COVID-19 Prevention Equipment and Transmission Research	x	x	x	
		COVID-19 Prevention and Control Research and Development	x	x	x	WHO topics

Notes:
[1] The test is conducted as a sense check of the topic modelling cluster labels using ChatGPT version 3.5.
[2] The OECD-STI team labelled the C8 cluster topics by reviewing: (i) the top 10 salient words in each cluster; (ii) the complementarity of these clusters with the WHO priority topics using a combination of the machine classification and the human expert classification; and (iii) the top funded R&D projects per cluster.
Source: OECD analysis of Fundstat database, March 2023.

Annex C. Complementary material on topic modelling of COVID-19 R&D projects

Table C.1 summarizes examples of COVID-19 R&D projects assigned to the high-level topic clusters (C8), based on an analysis of the Fundstat database as of March 2023 (Table 9 and Table 10). The table includes the country, database, project title (excerpt), abstract (excerpt), and USD million for each project. For instance, one project is the development of a SARS-CoV-2 vaccine based on mRNA technology in Germany, with a funding of USD 224.24 million under the cluster A.

Table C.1. Examples of COVID-19 R&D projects assigned to the high-level topic clusters (C8)

C8 topic cluster	Country	Database	Project title (excerpt)	Abstract (excerpt)	USD million
A. Coronavirus understanding, therapeutics and vaccine development	DEU	BMBF	Infection	ACE-mR-CoV - Development, testing and production of a SARS-CoV-2 vaccine based on mRNA technology	224.24
B. Platforms and capabilities	EU	CORDIS	European Clinical Research Alliance on Infectious Diseases (ECRAID-Base)	… As a European clinical research network ECRAID-Base will generate rigorous evidence to improve diagnosis, prevention, and treatment of infections…	37.58
C. Epidemiology and social interventions	GBR	GtR_MRC	COVID-19 Modelling Consortium: quantitative epidemiological predictions…	… mathematical and statistical modelling have been used to provide estimates…about the impact of interventions… epidemiological statisticians…	3.95
D. Digital access and online education	GBR	GtR_EPSRC	COVID-19 Transmission Risk Assessment Case Studies – education...	…with the recent increased awareness of airborne transmission of Covid-19… need to monitor the situation and to provide guidance on ventilation…	3.18
E. Cancer (Screening and treatment)	USA	NIH	Frederick National Laboratory for Cancer Research (FNLCR) Center for SARS-CoV-2…	… immune response to COVID-19… expanding the national testing capacity for SARS-CoV-2 antibodies, and … understanding of immune responses…	85.67
F. Public healthcare and other groups at risk	GBR	NIHR	Characterization, determinants, mechanisms, and consequences of long-term effects of COVID-19…	… long-COVID, including how best to diagnose, risk factors, health, and economic consequences, is poor, limiting efforts to help people…	13.19
G. Mental health and addictions	USA	NIH	Impact of the COVID-19 pandemic on patient outcomes…treatment for unhealthy alcohol use in vulnerable patients…	…significant adverse impact on vulnerable populations with serious comorbid medical conditions… It is critical to understand how to effectively manage these patients…	0.71
H. Environmental detection, transmission, and protection	AUS	NHMRC	BREATHE - mitigating airborne threats to health	…this research will improve preparedness, reduce health impacts of airborne threats, inform worker and occupant safety, building design…	1.88

Notes:
[1] Analysis limited to the retained COVID-19 R&D projects.
[2] Illustrative examples based on the automatic classification procedure to top level C8 clusters.
[3] The title and abstract have been shortened into excerpts to fit in the table.
Source: OECD analysis of Fundstat database, March 2023.

Table C.2 shows the topics labels for each of the clusters in Table 9, with the top 10 salient words for each of the topics. For example, Cluster A includes 7 topics, of which topic 1 has the following salient words: protein, human, spike, bind, coronavirus, immunity, therapeutic, antiviral, development, structure.

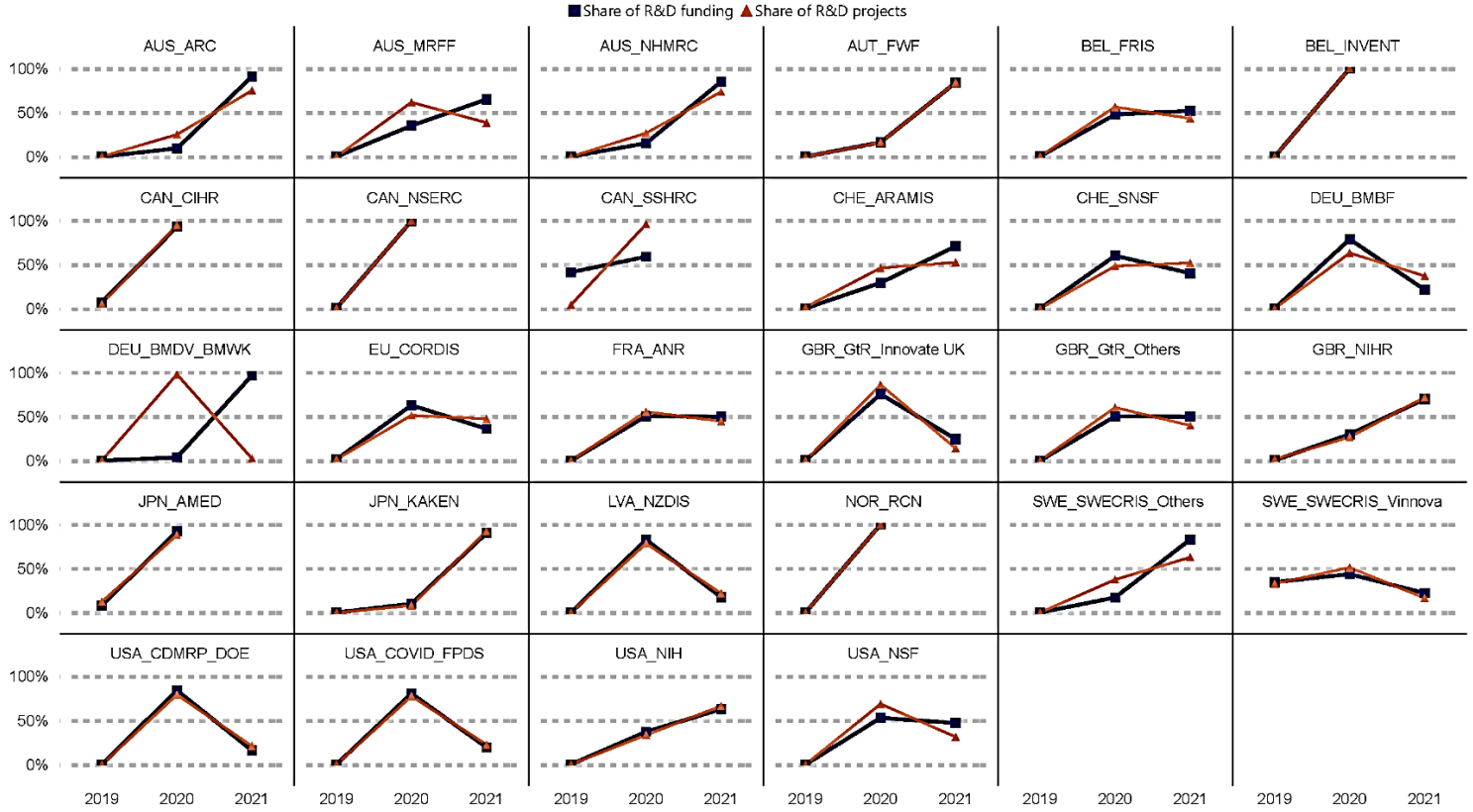
Figure C.1 shows the COVID-19 R&D projects and funding for the years 2019, 2020, and 2021 (partial information) for each funding agency/data source.

Table C.2. Top 10 salient words for the C34 topics in COVID-19 R&D funding projects

C8 cluster labels	C34 topics	Top 10 salient words for each C34 topic cluster, listed left to right by degree of within-cluster importance
A	1	protein, human, spike, bind, coronavirus, immunity, therapeutic, antiviral, development, structure
	4	detection, device, virus, antibody, saliva, sensitivity, diagnostics, development, antibody, oxygen
	6	inflammatory, acute, clinical, cytokine, tissue, syndrome, immune, blood, injury, mechanism
	10	immune, model, mouse, human, tissue, inflammatory, viral, innate, coronavirus, macrophage
	11	sequence, human, genome, wildlife, variant, viral, transmission, model, analysis, colony
	16	brain, blood, kidney, cognitive, symptom, alzheimer, injury, clinical, factor, expression
	18	variant, sequence, factor, genetics, biomarkers, clinical, human, expression, disorder, viral
B	9	trial, treatment, train, development, drug, management, implementation, review, investigator, staff
	15	exercise, intervention, digital, triage, medication, nurse, decision, access, train, deliver
	29	investigation, informatics, analysis, biomedical, protocol, partner, management, implementation, infrastructure, access
	30	treatment, healthcare, management, access, quality, delivery, visit, management, staff, measure
C	2	digital, market, manufacture, platform, contact, company, carbon, customer, sustainable, innovation
	3	family, intervention, policy, psychology, wellbeing, interview, youth, survey, psychological, factor
	5	model, spread, policy, behavior, epidemiology, decision, distance, mathematical, intervention, economic
	7	policy, decision, scientific, evidence, resilience, government, economic, local, analysis, interview
	12	train, virtual, education, platform, online, professional, workshop, environment, online, video
	22	slavery, cultural, local, business, policy, business, economic, researcher, interview, development
	23	tourism, business, behavior, model, decision, policy, manager, economic, transportation, condition
D	13	online, education, computer, remote, digital, science, practice, school, computer, virtual
	21	teacher, learn, education, online, remote, pupil, graduate, science, instruction, environment
	27	learn, remote, development, online, wellbeing, skill, provision, grade, young, closure
E	17	cancer, screen, tobacco, oncology, prevention, woman, disparity, trial, rural, delay
	33	tumor, mutation, genome, inhibitor, protein, metabolism, pathway, checkpoint, neoantigen, signal
F	14	pregnant, postpartum, stress, breastfeed, exposure, postpartum, woman, cohort, mental
	19	nurse, intervention, veteran, healthcare, measures, virus, health, train, management, area
	25	diabetes, weight, factor, heart, chronic, intervention, death, disparity, trial, hospitalization
	26	vaccination, intervention, communication, vulnerable, acceptance, minority, decision, pediatric, attitude, disparity
	31	opioid, treatment, emergency, healthcare, disorder, policy, intervention, access, respect, provider
	32	elderly, dementia, veteran, anxiety, measures, labor, society, activity, nursing, life
	34	vaccination, disparity, child, intervention, vulnerable, trial, county, staff, white, clinical
G	28	stress, substance, treatment, mental, alcoholic, treatment, harm, examine, consumption, young
H	8	mask, surface, filter, droplet, transmission, coat, material, water, waste, equipment
	20	water, surveillance, surface, spread, bioaerosol, material, plant, environmental, disinfection, coronavirus
	24	surface, coat, surveillance, disinfection, contact, occupational, handle, fibrosis, material, human

Notes:
[1] Analysis limited to the retained COVID-19 R&D projects.
[2] The breakdown of the C34 is based on the topics generated by the topic model. C34 reflects the 34 topics produced by the topic model, which are labelled according to the top 10 salient words. These C34 topics can be found at the left of the hierarchical dendrogram of Figure 6, with the C8 clusters emerging at a higher distance to the right. The C34 topics have not been labelled but rather they are presented with their top 10 salient words.
Source: OECD analysis of Fundstat database, March 2023.

Figure C.1. COVID-19 R&D projects and funding per year by funding agency/data source



Notes:

[1] Analysis limited to 10,472 COVID-19 R&D projects in the Fundstat database with known funding information. Biases due to the data coverage and the topic model development may distort the estimate of support breakdown by funding agency/data source. Sorted by country followed by agency/data source alphabetical order. FY2021 is partial because of data reporting structures and data availability across different countries. Specifically, FY2021 data are unavailable for BEL_INVENT, CAN_CIHR, CAN_NSERC, CAN_SSHRC, DEU_COVID-GEPRIS, JPN_AMED, and NOR_RCN.

[2] The analysis does not include the DEU_COVID-GEPRIS database, as there is no funding information available for those projects.

[3] GBR_Other includes ESRC, NERC, EPSRC, BBSRC, MRC, AHRC, STFC, NC3Rs, and UKRI.

[4] SWE_SWECRIS_Other includes FBEES, IFAU, VR, Forte, Formas, SHLF, RJ, and SWEA.

Source: OECD analysis of Fundstat database, March 2023.