Toward an observatory of the evolution of COVID-19 Vaccines Trials through phylomemy reconstruction

Quentin Lobbé, David Chavalarias, Alexandre Delanoë, Gabriel Ferrand, Sarah Cohen-Boulakia, Philippe Ravaud, Isabelle Boutron

Toward an observatory of the evolution of COVID-19 Vaccines Trials through phylomemy reconstruction

Quentin Lobbé[a], David Chavalarias[a,1], Alexandre Delanoë[a], Gabriel Ferrand[b,c,d], Sarah Cohen-Boulakia[e], Philippe Ravaud[b,c,d], and Isabelle Boutron[b,c,d]

[a]CNRS, Complex Systems Institute of Paris Île-de-France; [b]Université de Paris, INSERM, INRAE, CNAM, CRESS, F-75004 Paris, France; [c]Centre d'Épidémiologie Clinique, AP-HP, Hôpital Hôtel-Dieu, F-75004 Paris, France; [d]Cochrane France, F-75004 Paris, France; [e]Université Paris-Saclay, France

This paper aims at reconstructing the evolution of all the available COVID-19 vaccines trials extracted from the *COVID-NMA database* by applying the *phylomemy reconstruction process*. We visualize the textual contents of 1,794 trials descriptions and explore their collective structure along with their semantic dynamics. We map the continuous progress of the main COVID-19 vaccine platforms from their early-stage trials in February 2020 to their most recent combinations driven by the rise of variants of concern, third dose issues and heterologous vaccinations. This paper brings insights for the global coordination between research teams especially in crisis situations such as the COVID-19 pandemic.

COVID-19 | vaccination | phylomemy | knowledge dynamics

Significance statement. *The COVID-19 pandemic has resulted in an unprecedented volume of publications that have generated an information overload for the medical community. One of today's challenges is to synthesize this overwhelming amount of information in order to improve coordination between the different research streams. Our paper thus proposes to apply a new method for reconstructing the evolution of knowledge and to visualize the collective structure and semantic dynamics of 1,794 COVID-19 vaccines trials descriptions.*

Over the past two years, the ongoing COVID-19 pandemic has impacted a wide number of human domains: from economy to education, from public health to politics. Among others, Science swung early on into action to find both a cure and an effective vaccine. This has resulted in an unprecedented volume of publications that have generated an information overload for the medical community. One of today's challenges is to synthesize this overwhelming amount of information about current COVID-19 research in order to improve coordination between the different research streams. Our paper thus proposes to address this issue by applying a new method for reconstructing the evolution of knowledge. We take as a case study the COVID-19 vaccines clinical trials from the *COVID-NMA database* and use the *phylomemy reconstruction process* (1). The *COVID-NMA database* stores the curated dataset of all the clinial trials available in the set of international primary and secondary trial registries* (2, 3) (see Materials). For the purpose of this study, the *COVID-NMA database* has been reduced to a pruned corpus called $\mathcal{D}_{vt}$ (see Pre-processing). We then combine the expertise of epidemiologists and *Complex Systems* researchers to interpret the resulting visualizations and reveal insights for upcoming COVID-19 research.

i.e., all trials registered in the International Clinical Trials Registry Platform (ICTRP), Clinicaltrials.gov and the EU clinical trials registry

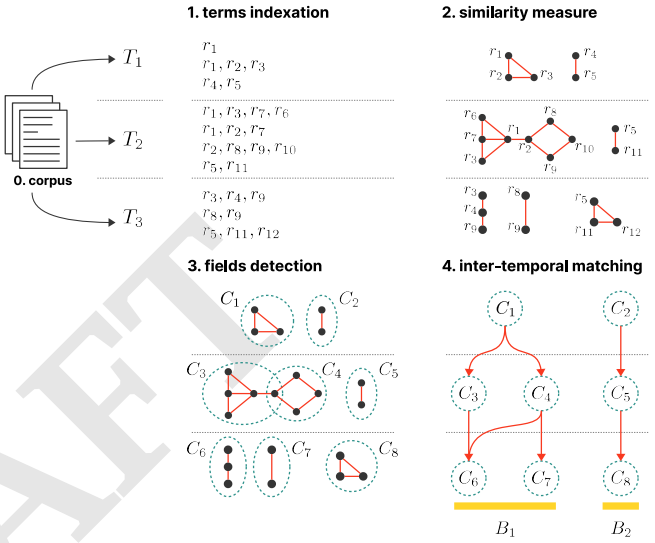

Fig. 1. The four operators of the phylomemy reconstruction process: 1. terms indexation, 2. similarity measures, 3. fields detection, 4. inter-temporal matching

The phylomemy reconstruction process

The phylomemy reconstruction process (1, 4) combines advanced text-mining methods, scientometrics and methods for the reconstruction of evolving complex networks in order to reconstruct the latent semantic structures of an unstructured – but timestamped – set of textual documents. Applied to a scientific corpus, it results in an inheritance network of research areas covered by all the collected publications. The phylomemy reconstruction process can be described as a combination of four subsequent operators of summarized by the Figure 1:

1. **Terms indexation.** By means of natural language processing (NLP) algorithms and human validations†, we first extract from an original corpus of documents (Figure 1.0) a core vocabulary as a list $\mathcal{L} = \{r_i \ | i \in \mathcal{I}\}$ of sets r_i of equivalent expressions called *roots* (Figure 1.1).

   In our case study, the corpus is a set of 1,794 trials descriptions. The *roots* are all the technical and equivalent names (including characteristics variations and any misspelling) given for a same vaccine. For instance, the technical

†NLP algorithms and human validations are handled by the free software *Gargantext* (5)

The authors declare no competing interests

expressions *"rad5"* and *"rad26"* were aggregated into *"gam-covid-vac"*[‡].

The corpus is then sliced into periods of interest $\mathcal{T}^* = \{T_i\}_{1 \leq i \leq K}$, $T_i \subset \mathcal{T}$ for which roots' co-occurrences are computed.

In our case study, we consider two weeks periods starting every monday from February 2020 to October 2021 and the output is a series of matrices of roots co-occurrences.

2. **Similarity measure.** Within each period of time and on the basis of its co-occurrences matrix, we estimate the semantic similarity between roots using the *confidence* measure (6). The completion of this task results in a temporal series of graphs of similarity (Figure 1.2).

3. **Fields detection.** For each period, a community detection algorithm – the *frequent item set* method (7) – is applied to detect subsets of densely connected roots within the graphs of similarity. Theses subsets C^T are called *fields* (Figure 1.3) and their aggregated root expressions describe consistent research topics that were explored at a given period.

   In our case study, the *fields* correspond to one or more descriptions of clinical trials sharing the same vaccine strategy. The output of this field detection step is a temporal series of clustering $\mathcal{C}^* = \{\mathcal{C}^T | T \in \mathcal{T}^*\}$ with $\mathcal{C}^T = \{C_j | j \in J^T\}$ and $C_j = \{r_i \, | r_i \in \mathcal{L}, i \in \mathcal{I}_j \subset \mathcal{I}\}$ computed over all the periods. It describes all the research directions explored from February 2020 to October 2021.

4. **Inter-temporal matching.** A temporal matching algorithm is then applied to identify meaningful kinship connections between fields from one period of time to another, *i.e.* fields that belong to the same research stream. We finally highlight the different research steams B_k over time and called them *branches of knowledge* (Figure 1.4).

The phylomemy reconstruction process makes it possible to draw the knowledge lineages at different resolutions through the tuning of a *level of observation* (1). The complexity of the resulting semantic landscape can range from a wide 'continent' to an 'archipelago' of specialized branches of knowledge.

Visualizing phylomemies. The structures highlighted by a phylomemy reconstruction process synthesize the complexity of the knowledge produced by a research community. In order to make this newly reconstructed knowledge actionable and explorable, a phylomemy can be visualized as a temporal network with time going by from top to bottom (8). Fields are represented by full circles and solid dark lines translate their kinship connections. *Emerging terms*[§] are displayed over the whole structure according to the combined coordinates of their period and fields of appearance. Term's size depends on their frequencies in the original corpus of trials. Branches are sorted from left to right so that closely related ones lie side by side. Interactive features can be used to reveal the entire fields' content, follow the dissemination of a given term throughout the phylomemy or simplify the *scale of description* of a selected branch.

[‡] The full list of roots is available at https://doi.org/10.7910/DVN/JTRI7A.

[§] Terms appearing for the first time in the phylomemy

Description of the resulting phylomemy

For our case study, we have used the corpus $\mathcal{D}_{vt}$ of 1,794 COVID-19 vaccines clinical trials (see Materials) in order to reconstruct the weekly evolution of the research on COVID-19 vaccines between February 2020 and October 2021. Key expressions are extracted from the original descriptions of the tested vaccines, grouped into roots and then into fields. The reconstructed fields thus embody a set of trials at a given period of time. We here choose a level of observation $\lambda = 0.5$ to shape quite precise branches. The resulting phylomemy (Figure 2) contains 175 roots and 550 fields distributed among 55 branches. The largest ones '*dna based vaccine*', '*non-COVID vaccines*', '*rna based vaccine*', '*non-replicating viral vector*', '*inactivated virus*' and '*protein sub-unit*' are highlighted by yellow shades. Shades of blue indicate the proportion of randomized clinical trials among the total number of trials on which the corresponding field has been reconstructed. The visualization of a phylomemy can also offer its user to interactively highlight some key information, as for example the research paths addressing vaccine boost issues, highlighted in red at the bottom of this figure.

Following the worldwide tracks of COVID-19 vaccines

General observations. After having explored and analyzed Figure 2 alongside epidemiologists, we noticed that the reconstructed phylomemy clearly retrieves five major COVID-19 vaccine platforms in the form of complete branches. These platforms include the classical vaccine platforms i.e., '*non-replicating viral vector*', '*inactivated virus*' and '*protein sub-unit*' as well as the next-generation vaccine platform i.e., '*dna based vaccines*' and '*rna based vaccines*'. The visualization shows the continuous development of each branch and the way some of them started to interact and eventually blended while others stopped. Interestingly, trials of '*rna based vaccines*' were registered very early in the course of the pandemic (February 2020) with trials evaluating the vaccine developed by Moderna TX (mRNA-1273) followed by the vaccine developed by Pfizer/BioNTech (BNT162b2) and sibling ones like BNT162b1 or BNT162b2sa that were not much longer tested (see Figure 2.a). The number of trials increased rapidly and interactions with other widely explored techniques were observed shortly afterwards: notably with the '*non-replicating viral vector*' family (ChAdOx1 – AstraZeneca – see Figure 2.b). The latest interaction involved the '*protein subunit*' branch in July 2021. In contrast, '*dna based vaccines*', with a first trial registered in April 2020, had a very limited number of trials planned and the whole branch stopped rapidly in 2020. Similarly, other platforms of '*replicating viral vector vaccine*', '*virus-like particle vaccine*' and '*live attenuated virus vaccine*' showed a very limited development.

Repurposing non-COVID vaccines. As the development and approval of COVID-19 vaccines was expected to take time, researchers also explored repurposing non-COVID vaccines. Considering the lower severity of the disease in children and young adults, some researchers hypothesized the possible heterologous protective effect of these vaccines. Some evidence shows that live-attenuated vaccines such as Bacille Calmette–Guerin (BCG), Measles, Mumps, Rubella (MMR) can induce protective innate immunity, which could be central in controlling SARS-CoV-2 (9). While this hypothesis was appealing, it

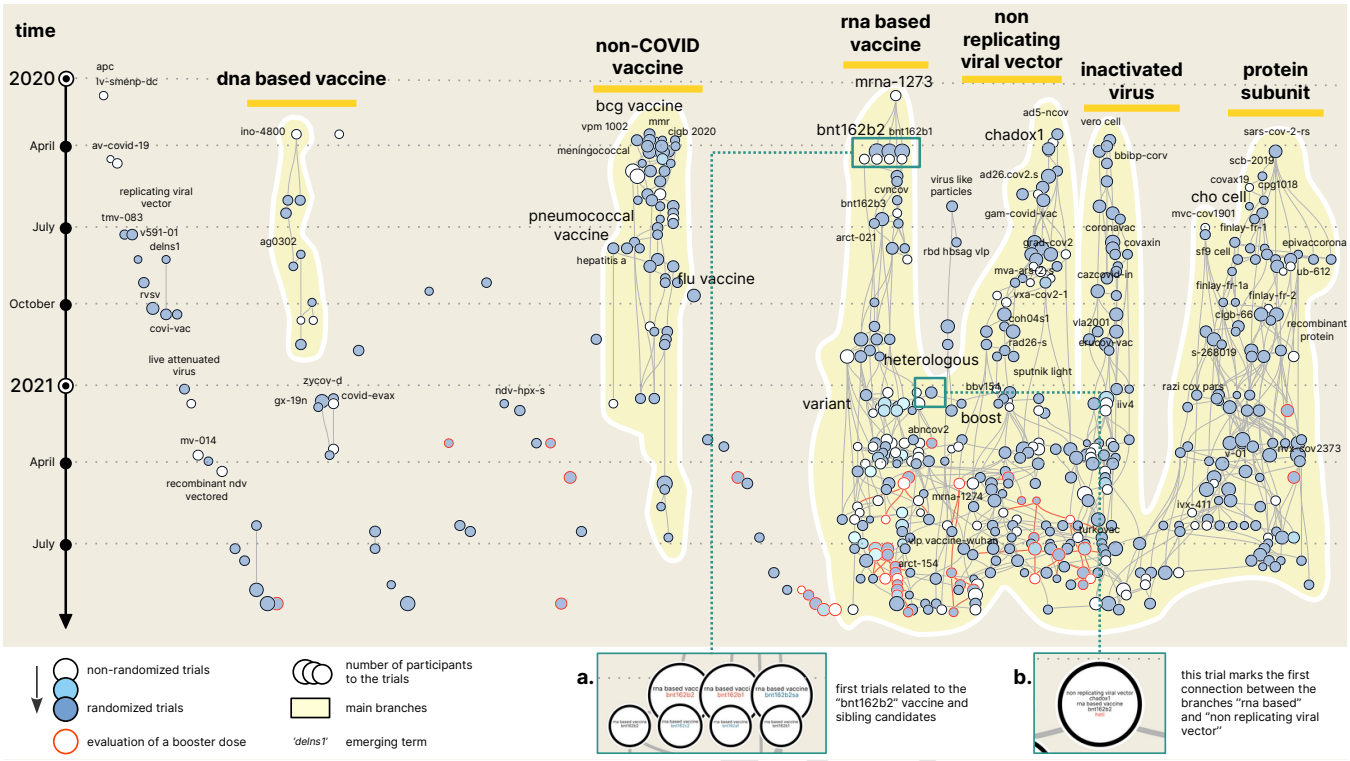

Fig. 2. Phylomemy of 1,794 COVID-19 vaccines trials recorded between February 2020 and October 2021 in the COVID-NMA database. Online and interactive version available at maps.gargantext.org/publications

did not seem to expand into a wider research domain. The branch of '*non-COVID vaccines*' appears and expands at the beginning of the pandemic but progressively decreases towards the end of 2020 as other more promising vaccines arose. Nevertheless, some researchers highlighted the need to adequately assess the use of non-COVID live-attenuated vaccines as they could potentially boost response in high-risk populations, be used in addition to COVID-vaccines to increase effectiveness and durability of their effect, or be used to protect people exposed to COVID-19 patients (9).

Heterologous vaccination. The branches interactions reflect the exploration of a new approach to vaccine implementation moving from homologous prime vaccination (i.e., injections of two doses of the same vaccine) to heterologous prime vaccination (i.e., injection of the first dose of a given vaccine and the second dose of another vaccine). This is clearly shown in Figure 3 with the assessment of the heterologous prime vaccination of '*rna based vaccine*' (BNT162b2-Pfizer/BioNTech) and '*non-replicating viral vector*' (ChAdOx1-AstraZeneka) in early 2021. This new approach was motivated by concerns about waning vaccine immunity, but also by practical considerations. Following concerns about the safety of the AstraZeneca ChAdOx1 vaccine, the EMA recommended giving a second dose Pfizer BNT162b2 vaccine to patients under the age of 55 years old who received one dose of ChAdOx1-S-nCoV-19. Furthermore, decision makers needed flexibility to overcome the issue of vaccine availabilities during the vaccine rollout. This new approach proved to be relevant and other associations were evaluated: '*non replicating viral vector*' and '*inactivated virus*' in June 2021 and later '*rna based vaccine*' and '*inactivated virus*' in September 2021.

Boosters. Phylomemies are essential in identifying shifts in research questions. While evidence of the beneficial effect of vaccines is mounting, research questions are moving toward exploring the effect of booster to overcome the waning of vaccine efficacy over time. Early in 2021, new trials assessing the impact of administrating a third dose (see Figure 2, red outline at the bottom) have been registered particularly for '*rna based vaccines*' and '*non-replicating viral vector*' (10). An important part of the research on boosters' effects is considering heterologous boosters.

Filters and upcoming research questions. By using additional data from the trials registries, we can filter the current phylomemy and thus push faceted observations to the fore or identify upcoming research questions.

Phylomemies also provides important information on research planning and reporting. As shown in Figure 2, most trials registered are randomized controlled trials. Early in the pandemic, non-randomized trials were primarily early phase trials while those registered in 2021 include both early phase trials exploring new vaccines and phase 4 trials assessing vaccines safety.

We can also explore the visualization to better understand how different countries participated in the overall research effort over time. For example, when filtering on the country (see maps.gargantext.org/phylo/vaccines/countries), we see that trials conducted in the USA explored all vaccine platforms and that first registered trials frequently involved a center in the USA, confirming their leading role in clinical research (*e.g.*, '*dna based vaccine*', '*rna based vaccine*', '*protein subunit*'). Other important trials characteristics such as funding sources can also be highlighted (see

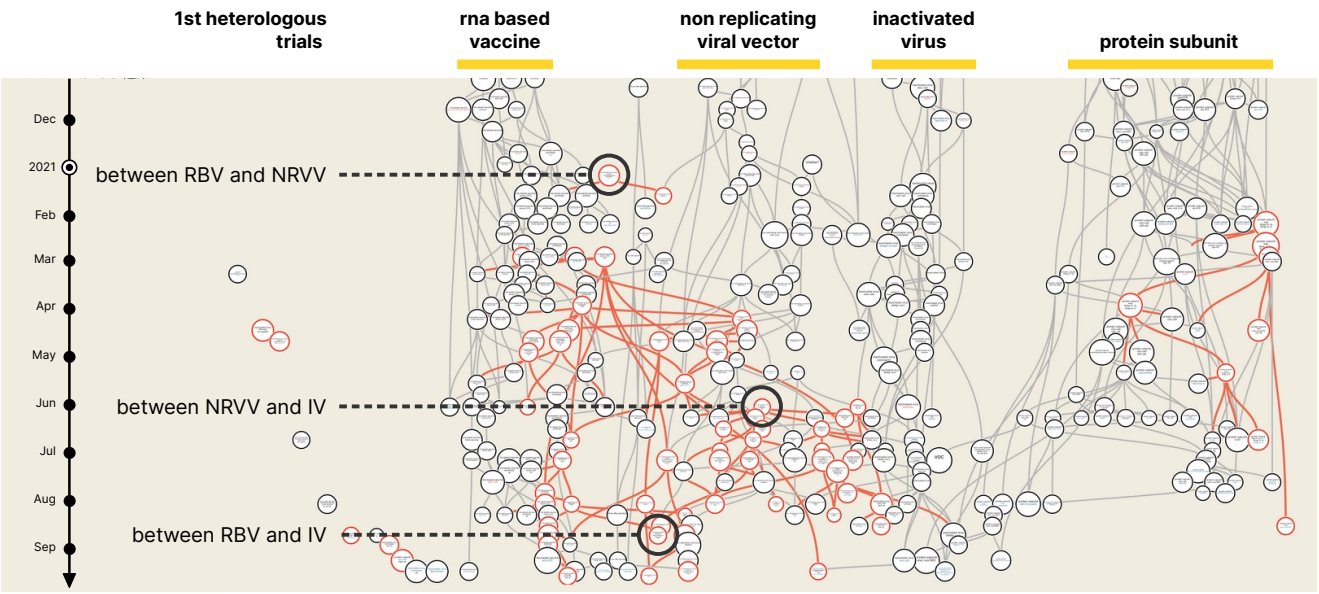

Fig. 3. A focus of Figure 2. In red are highlighted all the trials evaluating heterologous primary vaccination and heterologous booster. We circle the first heterologous trials involving different platforms.

maps.gargantext.org/phylo/vaccines/fundings).

Finally, we address the question of the publication of trial results (i.e., preprint or peer-reviewed articles). As shown in Figure 4, we currently have access to the results of a very limited number of planned trials. While most of the COVID vaccine trials registered in early 2020 are published, most of the non-COVID vaccine trials are still unpublished. Understanding whether these trials were actually conducted with unpublished results or were unable to recruit is an important issue.

Perspectives and insights for COVID-19 research

Global coordination between research teams is a key for accelerating innovation in Science, especially during crisis situations such as the COVID-19 pandemics. Reducing redundancies and providing heuristics to find new search paths as they arise can save time and lives (3). We claim that phylomemy reconstruction could be instrumental to guide trialists, funders and decision makers in biomedical research. In times of crises, it would enable them to better adapt to the evolution of the situation by following emerging research questions and identify less promising domains. It could also facilitate the identification of research gaps, research questions that may have been abandoned prematurely and redundancy in research. Our phylomemies could also be enriched with other data :

- data already recorded in the trials registries such as outcomes or participants characteristics which would allow exploring research conducted on vulnerable populations (children, pregnant women, immunocompromised patients, elderly etc.), trial results posted on the registries when possible;
- data that are not part of the registries but which should be added in pandemic times like the number of patients actually included in the trials;
- data that exists outside of the registries (publications, trials results, etc.) but for which a difficult work of data pruning and integration is required.

The addition of such information could be fulfilled through the collaborative and cumulative features of the *Gargantext* platform[¶]: the software used for computing the phylomemy reconstruction process. This generalization would increase the benefits of this approach tenfold. In a world where experts are increasingly specialized, it could draw attention to alternative solutions developed in other branches of science or to problems already encountered in research direction to be explored. It could also lead to new conceptual operations to be performed on a knowledge database, such as "give me all the branches of knowledge that are merging" or "suggest a promising combination of compounds to test". This could both accelerate research by making tangible the latent structure of innovation, and promote collaborations between teams that would not otherwise be interested in each other's work. Phylomemy reconstructions may thus become collective and reflective tools to foster the worldwide collective coordination between researchers. This revolution in clinical trial processing is within reach. Nevertheless, it would imply having access to high quality data on research planning and protocol.

Our case study focuses on a single disease, but this approach is fully generic and we call for a worldwide observatory for monitoring the dynamics of clinical trials. As it scales up, our approach could be implemented for any disease or research field.

Materials

Our data set has been collected and curated by the combined effort of epidemiologists, data integration and complex systems researchers.

[¶] Gargantext is a free software. See https://iscpif.fr/projects/gargantext

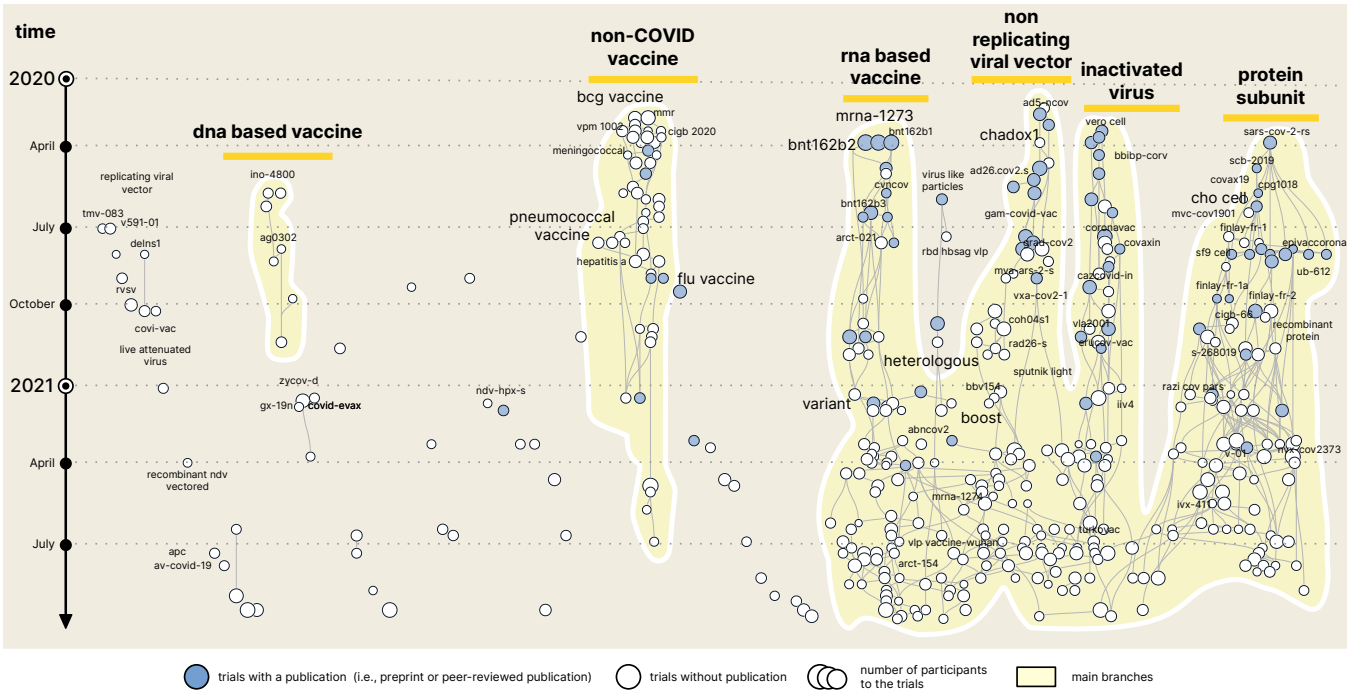

Fig. 4. Phylomemy of the randomized only COVID-19 vaccines trials. In blue, we highlight all the trials with an associated publication (i.e., preprint or peer-reviewed articles).

The COVID-NMA database. The COVID-NMA project is an international initiative aimed at providing a living mapping and a living systematic review of all trials assessing treatments and preventive interventions for COVID-19 (2, 3). The development of the COVID-NMA database relies on a full methodology designed to generate and make available a complete, comprehensive, integrated, non-redundant and carefully annotated data sets on clinical trials. We automatically extract data from clinical registries on a weekly basis and provide assistance to epidemiologists on the curation and annotation process. Raw data is extracted from the EU clinical trials register, from the ClinicalTrial registry managed by the U.S. National Library of Medicine, from the IRCT registry and from the WHO International Clinical Trials Registry Platform (ICTRP) – an international registry that assembles information on clinical trials registered in 17 primary registries to identify new trial assessing COVID-19 vaccine and update of previously registered trial records. Data are extracted from registries, annotated by epidemiologists, then stored and made available through the COVID-NMA database[‖].

Pre-processing the database. We have pre-processed[**] the COVID-NMA database before using it for the phylomemy reconstruction to filter the 1,794 descriptions related to vaccines trials. The trials records have been first aggregated by publication week. Then, we have merged the sections '*pharmacological treatment*', '*treatment type*' and '*treatment name*' together to shape the trial descriptions. These descriptions have also been enriched with extra-information such as trial phases, funding, involved countries or associated publications. The resulting corpus $\mathcal{D}_{vt}$ has latter been collectively and collaboratively curated by epidemiologists thanks to the free software *Gargantext* (5). There, these experts have extracted and validated a core vocabulary as a list of 175 root terms.

Data Availability. The original COVID-NMA database can be downloaded at covid-nma.com. The reconstructed phylomemy is available for live explorations at maps.gargantext.org/phylo/vaccines/publications and downloadable at https://doi.org/10.7910/DVN/JTRI7A.

1. D Chavalarias, Q Lobbé, A Delanoë, Draw me science – multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies. *Scientometrics* (2021).
2. I Boutron, et al., The COVID-NMA Project: Building an Evidence Ecosystem for the COVID-19 Pandemic. *Ann Intern Med* **173**, 1015–1017 (2020).
3. VT Nguyen, et al., Research response to coronavirus disease 2019 needed better coordination and collaboration: a living mapping of registered trials. *J Clin Epidemiol* **130**, 107–116 (2021).
4. D Chavalarias, JP Cointet, Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PloS one* **8**, e54847 (2013) 00000 bibtex: chavalariasPhylomemetic2013.
5. A Delanoë, D Chavalarias, Mining the digital society - Gargantext, a macroscope for collaborative analysis and exploration of textual corpora. (forthcoming 2021).
6. G Dias, R Mukelov, G Cleuziou, Mapping general-specific noun relationships to wordnet hypernym/hyponym relations in *International Conference on Knowledge Engineering and Knowledge Management*. (Springer), pp. 198–212 (2008).
7. T Uno, M Kiyomi, H Arimura, , et al., Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets in *Fimi*. Vol. 126, (2004).
8. Q Lobbé, A Delanoë, D Chavalarias, Exploring, browsing and interacting with multi-level and multi-scale dynamics of knowledge. *Inf. Vis.*, 14738716211044829 (2021).
9. K Chumakov, et al., Old vaccines for new infections: Exploiting innate immunity to control covid-19 and prevent future pandemics. *Proc. Natl. Acad. Sci.* **118** (2021).
10. PR Krause, et al., Considerations in boosting COVID-19 vaccine immune responses. *Lancet* **398**, 1377–1380 (2021).

[‖] We here note that international trials registries can be post-updated by research teams, e.g. for post-adding a related publication. Future versions of the phylomemies presented in this paper might thus be slightly different from the current ones. A promising way to get around this issue would be to archive every modifications of the original registries and then choose the version we want to integrate in the phylomemies.

[**] The pre-processing script can be downloaded at https://doi.org/10.7910/DVN/JTRI7A