Investigating the Scientific 'Infodemic' Phenomenon Related to the COVID-19 Pandemic

- a Position Paper from the IMIA Working Group on „Language and Meaning in BioMedicine"

László Balkányi[1], Lajos Lukács[2], Ronald Cornet[3]

[1] Medical Informatics Research and Development Center (MIRDC), Pannon University, Veszprém, Hungary
[2] DSS Consulting, Ltd. Budapest, Hungary
[3] Department of Medical Informatics, Amsterdam University Medical Center - University of Amsterdam, Amsterdam Public Health research institute, Amsterdam, The Netherlands

Summary

Objectives: The study aims at understanding the structural characteristics and content features of COVID-19 literature and public health data from the perspective of the 'Language and Meaning in Biomedicine' Working Group (LaMB WG) of IMIA. The LaMB WG has interest in conceptual characteristics, transparency, comparability, and reusability of medical information, both in science and practice.

Methods: A set of methods were used (i) investigating the overall speed and dynamics of COVID-19 publications; (ii) characterizing the concepts of COVID-19 (text mining, visualizing a semantic map of related concepts); (iii) assessing (re)usability and combinability of data sets and paper collections (as textual data sets), and checking if information is Findable, Accessible, Interoperable, and Reusable (FAIR). A further method tested practical usability of FAIR requirements by setting up a common data space of epidemiological, virus genetics and governmental public health measures' stringency data of various origin, where complex data points were visualized as scatter plots.

Results: Never before were that many papers and data sources dedicated to one pandemic. Worldwide research shows a plateau at ~ 2,200 papers per week – the dynamics of areas of studies being slightly different. Ratio of epidemic modelling is rather low (~1%). A few 'language and meaning' methods, such as using integrated terminologies, applying data and metadata standards for processing epidemiological and case-related clinical information and in general, principles of FAIR data handling could contribute to better results, such as improved interoperability and meaningful knowledge sharing in a virtuous cycle of continuous improvements.

Keywords

COVID-19 pandemic, review literature, metrics, FAIR principles, infodemic

Yearb Med Inform 2021:245-56
http://dx.doi.org/10.1055/s-0041-1726483

1 The COVID-19 Pandemic and Related Literature

Defining the 'What to study': The last decade has seen epidemics with world-wide significance (2009: H1N1, 2012: MERS, 2015: Zika) but they neither had a similar global impact on the world economy and everyday life nor led to a similar amount of related research (see Table 3. in *Results* for details). The COVID-19 pandemic generated a previously unseen intensity of scientific research and as a result, an unseen number of publications (see Figure 1. in Results) and related public health data (openly available, mostly epidemiological). We consider both textual publications and COVID-19 related data published by relevant public health authorities as raw material for further studies. We paraphrase this phenomenon to be called a **'scientific infodemic'** – narrowing the Merriam Webster general definition of infodemic [1]: '*…. is a blend of "information" and "epidemic" that typically refers to a rapid and far-reaching spread of both accurate and inaccurate information about something, such as a disease…*' to the more restricted world of scientific publications and available public health data. We think that using the paraphrase '*scientific infodemic*' versus calling the papers of the observed period simply the '*scientific literature*' emphasizes uncertainty, partly caused by the unusual number / ratio of publications that were only available as preprints, lacking peer review, but still made available under the time pressure of the need for controlling the COVID-19 pandemic. Not surprisingly, according to the Retraction Watch Database1 over 60 papers have been retracted just in 2020.

Explaining the 'Who': Authors of this paper partly constitute the leadership of the 'Language and Meaning in Biomedicine' Working Group of IMIA (LaMB WG) - having an interest in medical concept representation (see detailed history of the WG here [2]). Therefore, we were motivated to investigate conceptual characteristics, transparency, comparability, and reusability of the COVID-19 research material. Our WG background helped to realize that to gain understanding, both quantitative and qualitative aspects of the COVID-19 scientific infodemic should be investigated.

The background for 'Why' and 'How' to study COVID-19: Just reading, observing the research literature reveals peculiarities, such as divergent or even contradicting data on epidemiology, clinical features, and response to the various therapeutical interventions to COVID-19. Inconsistencies and contradicting information bits and pieces lead even to retraction of some papers (i.e. [3]). Some of these peculiar features (as e.g., speed and sheer volume) are judged positively, some others negatively by the research community. However, observations and reading leads only to anecdotal evidence. Systematic, analytic study of conceptual structures and that of quantitative content

[1] http://retractiondatabase.org/RetractionSearch.aspx#?ttl%3dcovid-19

characteristics might help understanding why the peculiarities occur, as well as the fight against the COVID-19 pandemic. Examples of quantitative studies in this paper are measuring *speed* and *amount* of accumulating scientific information, an example of analytic study that is investigating the (appreciated) *openness of data* related to studies of a new pathogen and the caused disease (see the *Methods* section for full details).

Let us note that countries and various international agencies share a remarkable amount of almost real-time, fine-grained epidemic and pathogen-related data. Countries and supranational organizations had and still have a chance to design and execute efficient public health, economic and social responses. However, the COVID-19 pandemic research and responses have brought home this potential only partially. As mentioned by Flack and Mitchell [4] "the response has been ambivalent, uneven and chaotic – we are fumbling in low light, but it's the low light of dawn". Science and the applied public health response are just part of the global activities related to COVID-19 – and not the only sources for decision making, neither for the public in general nor for the policy makers.

The objective of this study is to gain better understanding of the structure and content of the COVID-19 literature (both that of the textual and data resources). We also strive for better understanding and the clarification of the (possible) role of 'language and meaning' paradigms – such as homogenous semantics, conceptual interoperability, standardized data, role of ontologies, description logics and hybrid architectures, and possible role of knowledge representation. We think that all these objectives can be achieved by looking for measurable qualitative and quantitative characteristics of the COVID-19 research literature and the related open data resources. We also aim to check whether the above-mentioned impressions of unusual speed, amount and conceptual detail of the literature are correct. Once the peculiarities are characterized by these measurements, *the secondary objective* of this position paper is to check whether (and to what extent) 'language and meaning' paradigms, tools, methods could help to establish better transparency, adding reliability and more credibility of COVID-19 research results to the scientific community and beyond.

2 Methods

2.1 The Quantitative Aspect: Dynamics of COVID-19 Literature

For the first objective, to characterize quantitatively the features of Covid-19 literature, we check speed, dynamics (numbers and structure) of publications and show results of text mining of the research corpora, trying to find measurable semantic characteristics.

- ***Measuring the amount of publications related to COVID-19 and its dynamics***

To check the speed and dynamics, a snapshot of COVID-19 literature, as registered in PubMed, was taken on Aug. 25, 2020. We used a well-defined, pre-processed literature curation: numbers were retrieved by checking the 'COVID19 Article Collection' from LitCovid/ PubMed. In LitCovid the dynamics were checked and analyzed for the following subdomains: COVID-19 in *general*, its *(patho-)mechanism*, disease *transmission*, *diagnosis, treatment, prevention, specific case reports* and f*orecast* - modelling. Publication numbers were granulated by week. Our added value here was the evaluation of tendencies and comparison of various areas. In addition to that, overall COVID-19 numbers were compared by us to numbers of other similar events of the past decade by a simple, date limited search.

- ***Characterizing the most important and emerging concepts describing the COVOD-19 pandemic***

We established and studied a corpus using a text mining tool (Voyant [6]). The above-mentioned LitCovid article collection was used. The titles of the listed ~ 45,000 papers were compiled, allowing authors of this paper to see the applied semantics of the research domain. The Voyant tool was used to visualize a semantic map of related concepts, using quasi-clusters of article title word (as labels for the concepts) frequencies. From a variety of analytic tools that Voyant has, in this case we used the T-distributed Stochastic Neighbour Embedding (t-SNE) scatter plotting. This is a non-linear dimensionality reduction algorithm, where visualizing t-SNE results allows a look at how the terms are arranged in a high-dimensional conceptual space.

2.2 A Qualitative Aspect of COVID-19 Data

For the qualitative study we collected and studied several existing independent data sets and paper collections (understood as textual data sets).

As Bakken remarks in her paper [7]: '*Regardless of the type of biomedical and health informatics research conducted (e.g., computational, randomized controlled trials, qualitative, mixed methods), transparency, reproducibility, and replicability are crucial to scientific rigor, open science, and advancing the knowledge base of our field and its application across practice domains'*. In agreement with that view, we focused on methods dealing with such aspects of information quality as transparency, reproducibility, and replicability.

As a first step, we assessed **FAIR-**ness [8] of these information sets – checking if they are **f**indable, **a**ccessible, **i**nteroperable, and **r**eusable. FAIR-fitting can be checked by a well-defined, easy-to-apply checklist, covering all criteria. The FAIR principles are embraced by many science communities as well as by the European Commission. Checking FAIR requirements allowed us to investigate how these information sources could be (re) used and combined. The details of FAIRness checking are explained in *Results*, in Table 5.

As a second step, the usability of FAIR requirements was challenged by a real test case. We performed a combined analysis - setting up a common data space [9], using independent data sources as detailed below.

- ***Finding examples of open access information sources of COVID 19***

COVID-19-related information (both textual and data collections) might be grouped in various types: related to the *spread of disease* on population level (field epidemiology and modelling); describing the *pathogen* (e.g. virus RNA sequences); related to *clinical manifestation* (case numbers, comorbidity epidemiology, diagnosis and therapy of the disease); and characterizing *public health and policy actions* (e.g. a complex stringency measure of various government measures against the spread of COVID-19 disease). Table 1 lists such information sources. The sources were identified by Google Scholar search, by screening global and local public health organization portals (e.g., WHO [10], ECDC [11]) and by Medisys [12]. The identified sources are used for further analysis [13-25].

Table 1 COVID-19 related, open access, numerical and textual information sources

Type	Name	Start Date	Size (at cut off date)
epidemiological data sources: This type contains five independent data sources, each with worldwide coverage collecting data from multiple sources and partially cross checking each other for COVID 19 pandemic. Two data sources are that of the relevant international organizations (WHO, ECDC), one is provided by an internationally acknowledged academic source, and the last two are independent data sources.	WHO Coronavirus Disease (COVID-19) Dashboard https://covid19.who.int	2020. 01.04.	08.10: 33,812 records
	ECDC COVID 19 Worldwide data https://www.ecdc.europa.eu/en/covid-19/data	2020. 12.31.	08.10: 35,150 records
	COVID-19 Dashboard at Johns Hopkins University (JHU) https://coronavirus.jhu.edu/map.html	2020. 01.22	08.10: 4*267= 1,068 records
	Our World in Data COVID-19 database https://github.com/owid/covid-19-data/tree/master/public/data	2019. 12.31.	08.12: 36,138 records
	Worldometer Coronavirus data https://www.worldometers.info/coronavirus/?zarsrc=130	2020. 01.22	08.12: 215*200= 43,000 records
virus genomics data sources: This type of data sources collects RNA sequences independently from each other, obviously having overlapping sequences from all around the world.	GISAID EpiCoV database https://www.epicov.org/epi3/frontend#5e7e53	2019. 12.04	08.13: 81,625 records
	NCBI SARS CoV-2 GenBank https://www.ncbi.nlm.nih.gov/genbank/	2020. 01.11	08.13: 16,046 records
	China National Center for Bioinformation 2019Novel Coronavirus Resource https://bigd.big.ac.cn/gwh/browse/virus/coronaviridae	2019. 12.30.	08.13: 98,324 records
	EMBL-EBI's COVID-19 Data Portal https://www.covid19dataportal.org/sequences?db=embl	2019. 01.28.	08.13: 17;730 records
public health & policy actions These data sources present data of the response of the societies around the globe. (e.g., stringency, economics etc)	CORONAVIRUS GOVERNMENT RESPONSE TRACKER https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker	2020. 01.01.	08.13: 46,996 records
	CoronaNet Research Project https://www.coronanet-project.org/index.html	2020. 03.03	08.13: 24.663 records
COVID-19 clinical data	WHO: Global COVID-19 Clinical Data Platform https://www.who.int/publications/i/item/WHO-2019-nCoV-Clinical_CRF-2020.4	2020.07.13	(closed data set)
textual data - research paper collections	CDC Library COVID-19 Research Articles Downloadable Database www.cdc.gov/library/researchguides/2019novelcoronavirus/researcharticles.html	2020. 03.20.	07.31: 77,424 records
	WHO COVID-19 - Global literature on coronavirus disease search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/	2020. 02.01	2020.07: 67,530 records
	NLM LitCovid www.ncbi.nlm.nih.gov/research/coronavirus/	2020. 01.13	08.25: 45,499 records
	CORD-19 - COVID-19 Open Research Dataset https://allenai.org/data/cord-19	2020. 03.13	08 26: 100,008 records

• ***Assessing fulfilment of FAIR requirements of information sources***

We checked the FAIR requirements by using the EUDAT Fair Data Checklist [26]. Each requirement is checked against a set of relevant EUDAT checklist conditions (detailed in *Results*). In order to evaluate these checklist conditions, all sources were approached, opened, and read or in case of data sources downloaded as well. If all the specific conditions of a given area (e.g., 'interoperability') are met, the requirement is fulfilled. If none of them are met, obviously, the requirement is not fulfilled. If some of the conditions are met, the requirement is 'partially' fulfilled.

• ***Testing FAIR-ness in practice: compiling epidemiological, virus genomics and stringency*** *(of government measures against the COVID-19 pandemic)* ***data***

As a test for practical interoperability, reusability and semantic consistency authors of this paper did also a compilation of data, using the above-mentioned data sources. We set up a

common data space [9] where epidemiological, virus genetics and governmental measures' stringency data are combined to characterize the COVID-19 pandemic. We used the 'Oxford Covid-19 Government Response' stringency data, which is a complex index, calculated form 19 indicators, organized into four groups [20]: C - containment and closure policies; E - economic policies; H - health system policies; and M - miscellaneous policies.

Source data were compiled to data points, each expressing a numerical value on a given week of a given country regarding (i) a mean of genetic divergence of virus RNA samples - as a kind of index of what strains were active then and there, (ii) a cumulated number of death/million inhabitants of the following two to five weeks (a measure of severity of the outbreak), and (iii) value of the complex stringency measure, indicating the strength of government response measures.

2.3 Determining the Possible Role of 'Language and Meaning' Paradigms, Tools in Improving COVID-19 Research

For the second objective, authors use references to earlier publications of the IMIA Language and Meaning in Biomedicine Working Group members – on how these 'language and meaning' related paradigms could and should be applied in case of COVID-19. The main paradigms are (a) the role of homogeneous semantics and inherent interoperability (terminologies) used in publications; (b) the need for standardized data in field and clinical epidemiology, enabling large-scale predictive data analysis – both in related papers and databases; (c) the role of ontologies, description logics and hybrid architectures; and (d) the role of knowledge representation - especially in studies related to artificial intelligence.

We scrutinized earlier LaMB WG authors' related publications for finding examples, answers to questions, as :

- which of these paradigms are relevant to improve research quality and mitigate inconsistencies in COVID-19 related data collection and interpretation;
- what 'language and meaning' methods tools can do for connecting different fields, as e.g. genetics and pathophysiology data of SARS CoV 2 and COVID-19; and we check;
- if 'language and meaning' methods are able to improve clinical response to COVID-19 (diagnostic and therapeutic issues);
- how 'language and meaning' tools in the broad sense can help to overcome the problem of transparency and comprehensibility caused by the sheer amount of research information related to COVID-19.

3 Results

3.1 Results on Quantitative Characteristics

- Amount and dynamics of COVID-19 literature

The results of the snapshot taken on 25. Aug 2020 - for all papers published since the beginning of the year, as listed in PubMed are reported in Figure 1.

At this date, the number of papers on COVID-19 in PubMed was 45,499. The dynamics (first papers appearing already in January 2020) shows a quick growth from March 2020 to mid-May 2020. It is interesting to see the ratio of various areas, shown in Table 2. These results are published (and regularly updated) by the LitCovid curated paper collection. The LitCovid curation process deals with subdomain classification as well. We added red tendency arrows to show how data are changing and the analysis below. These are not calculated trendlines, not based on data, just visual aids, to catch direction of change.

Dynamics of the various areas are shown on Figure 2. Note that the axes were distorted to make the scales visually comparable to each other. Red linear tendency arrows show the different speed and dynamics.

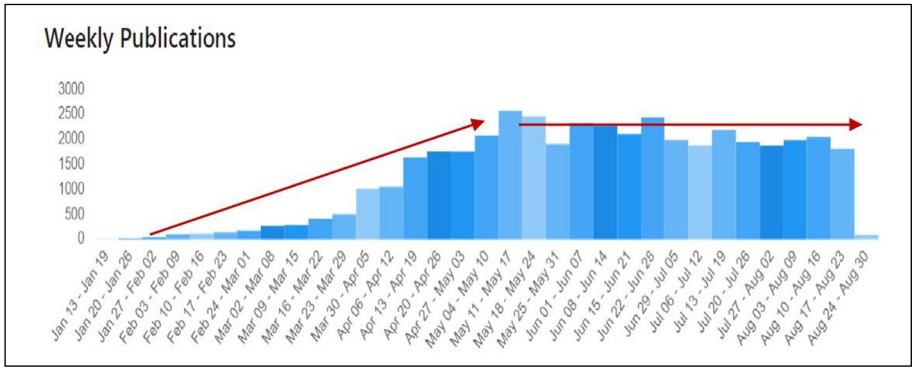

Fig. 1 COVID-19 related publications in PubMed (bars show actual publication numbers, while red tendency arrows are not data-based trends, they are just visual aids helping to see the basic direction of change of publication numbers (i.e. growing, steady, etc.). Shades of blue, generated by the LitCovid tool are just helping to discern bars, do not have specific meaning).

Table 2 Ratio of various COVID-19 related publications in PubMed, for the first nine months of the pandemic (percentage values rounded).

Area:	Ratio:
general information and news:	~ 4%
transmission characteristics and modes of covid-19 transmissions:	~ 3%
treatment strategies, therapeutic procedures, and vaccine development:	~23%
case reporting	~7%
mechanisms underlying cause(s) of covid-19 infections and transmission & possible drug mechanism of action:	~ 11%
diagnosis disease assessment through symptoms, test results, and radiological features:	~15%
prevention, control, response, and management strategies:	~36%
forecasting, modelling, and estimating the trend of covid-19 spread:	~1%

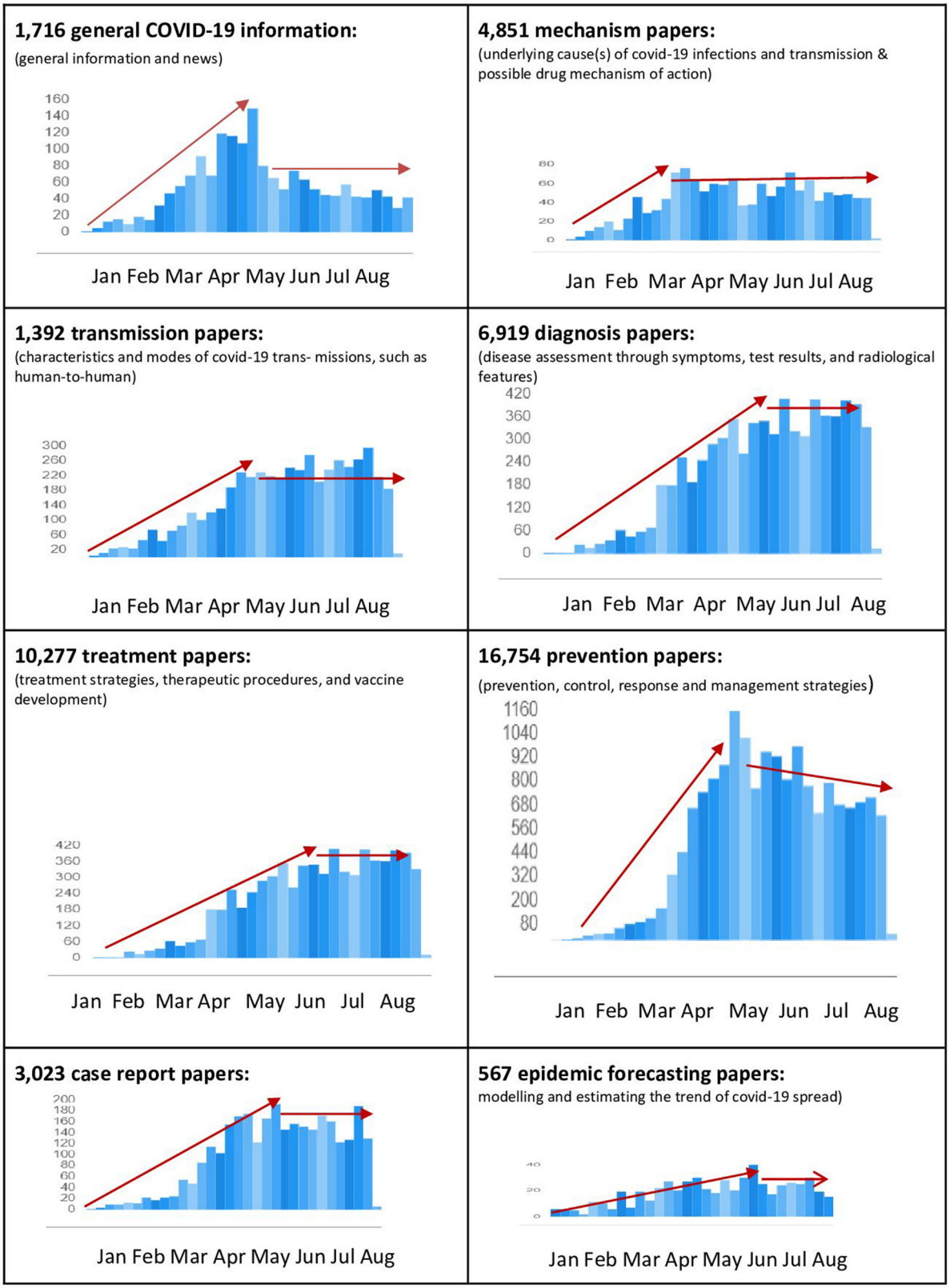

Fig. 2 Dynamics of various areas of scientific publications in the first 8 month of the COVID-19 pandemic (bars show actual publication numbers, red tendency arrows are not data-based trends, they are just visual aids helping to see the basic direction of change of publication numbers (i.e. growing, steady, etc.)

The arrows help to realize that in almost all areas there is a kind of "saturation" effect, but research reaches a maximum "throughput" on different levels (i.e., number of publications). - that is most probably mirroring the available scientific capacities.

It is also important to note the difference in the sheer number of COVID-19 publications related to the number of publications produced at other similar events in the past decade (for other outbreaks we covered two years to collect all publications related to the actual outbreak). We applied a simple, date limited PubMed search, see the results and the search strings in Table 3.

- ***Results for characterizing the most important and emerging concepts describing the COVID-19 pandemic***

A T-distributed Stochastic Neighbour Embedding (t-SNE) scatter plot of article title words frequencies – generated by using the appropriate Voyant tool - resulted in a concept map, shown in Figure 3. Coloring of dots shows the (quasi-)clustering, while dot sizes represent rates of word occurrences. Concepts in the blue area (like: *patient, treatment, respiratory*) show papers focusing mostly on clinical aspects, the purple area concepts (like: *emerging, covid, health, epidemic*) depict mostly epidemiology aspects, while the green area reaches out to response and other broad aspects of society status during COVID-19 (concepts: *implications, lessons, perspective, etc.*)

Table 3 Numbers of outbreak-related publications (worldwide) of the last decade in PubMed

Outbreaks	Period	No. of papers in PubMed
H1N1*	2009-2010	5,611
MERS*	2012-2013	1,034
ZIKA*	2015-2016	1,975
COVID-19*	2020	45,499**

*these labels used also as character strings for PubMed search **cut off date: 2020.08.25

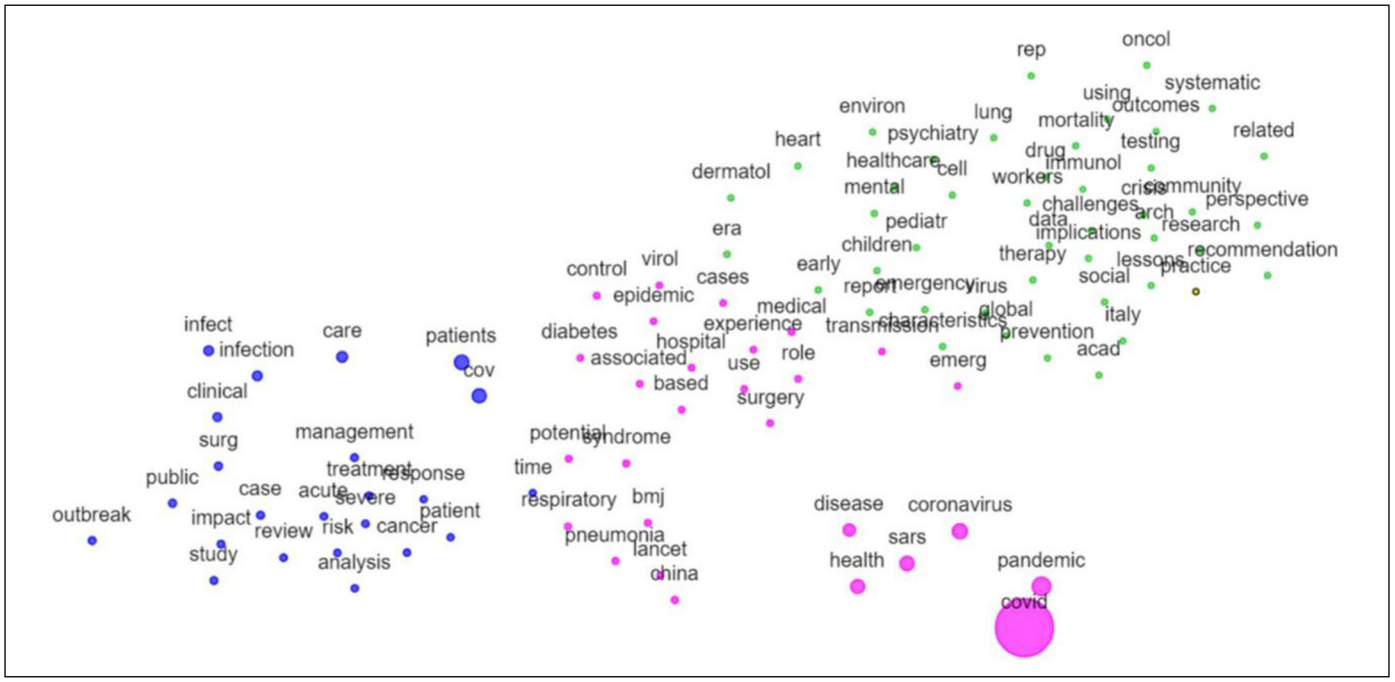

Fig. 3 (t-SNE) scatter plotting (by Voyant) - of title word rate frequencies, revealing conceptual areas of COVID-19 literature. Numerical values of the x- and y-axes are omitted, as the pattern of the terms (auto-colored clusters) is what is informative here. Blue concepts represent papers focusing mostly on clinical aspects, papers of purple concepts depict mostly epidemiology aspects, green concepts papers reach out to response and other broad aspects of society status during COVID-19.

3.2 Results of Investigating Qualitative Aspects

- **Metadata issues as a crucial element of research transparency**

It is important to note that (practical) data transparency and reusability should be judged at least on two levels: data syntax and data semantics. (Pragmatics, the third level is obviously not a transparency or reusability issue.) Regarding syntax, the below listed data sources are all transparent and reusable, provided either as 'csv' or 'xml/json' formats. However, the semantic layer is much less transparent - that makes data comparison and compilations problematic. Items in Table 4 show that the same conceptual entities (like country names) are labelled and coded differently (e.g., in the epidemiological sources).

- **FAIR requirements of information sources**

Checking FAIR-ness of these information sets provides us a structured, well-proven approach to overcome the complexity of investigating information sources. Table 5 below shows that most of the sources only partially meet the FAIR requirements - forecasting some difficulties for COVID-19 meta-analysis studies in the future.

This analysis shows that possibly the most critical requirement is interoperability. The sources usually fail the "controlled vocabularies, keywords, thesauri or ontologies are used where possible" requirement, and some also lack standard metadata formats.

- **Result of testing FAIRness in practice: compiling epidemiological, genomics and governmental measures' stringency data**

Figure 4 demonstrates the use of the FAIR-ness assessment. We use this visualization here as an illustration of using / applying FAIR principles in practice. Data sources, complying to FAIR requirements were used. In this complex figure, COVID-19 *epidemiological* data are combined with the COVID-19 related *stringency* of government measures data and SARS-CoV 2 virus *genetic divergence* - all from different, independent sources. Not surprisingly all the actual steps (as finding data sets, downloading, data cleansing, importing to spreadsheet, processing the data, and building meaningful visualizations) proved to be doable tasks. On Figure 4 each composite data point reflects values of a given week of a country. The x-axis presents an '*epidemic severity*' index, showing the cumulated new deaths/million people following 2-5 weeks of the actual week. The y-axis presents the *public health stringency governmental measures* of a country, while the size variations of the dots indicate the mean genetic divergence of the viral strains present in the country on that week. The colors of the dots show a K-means clustering along the timeline, using the Kmc tool [27] allowing us to watch out for the influence of progress of the COVID-19 pandemic over the timeline.

In this paper we analyzed the way to establish such data compilations - and not the figure information itself. Particularly, we emphasize the (FAIRrelated) problems of compiling these data. Regarding *findability* and *accessibility* all the sources were of equal high value, even though they do not fully comply to the formal requirements as described in the EUDAT FAIR requirements checklist [26}. For each of them (Our World in Data, Worldometer, Nextstrain – GISAID, EpiCoV, Oxford Government Response Tracker) the data were findable and accessible, metadata were well-defined and rich, but regarding a persistent ID for each data element, only the genetic data fulfilled this requirement. However, regarding *interoperability,* we discovered some problems. For instance, basic and crucial data element such as labels and standard abbreviations (e.g., ISO codes) for countries were differing in the various sources (e.g., in case of the United Kingdom or Macedonia). Also, none of them named the source of the country list properly, hampering *reusability* on the long run. In the same way, normalizing the data (i.e., for population size) was not based on transparent data sources. Lack of explicit usage of permalink type global unique identifiers is also an issue for long term *reusability.*

3.3 'Language and Meaning' Paradigms

- ***Overcoming inconsistencies in*** COVID-19 related data collection and descriptive, textual interpretations in papers. In [28], one of the main conclusions was, that benefits of the integrated terminologies in terms of homogenous semantics and inherent interoperability outweigh the complexity added to the system. This statement is relevant in case of COVID-19 that proved to be a disease with a broad set of various clinical manifestations. Indeed, various possible pathomechanism pathways were and are investigated. Using homogenous semantics by integrated terminologies both in related papers and in related data bases could have prevented inconsistent presentation of signs and symptoms, of progress of disease. For example, controversial anecdotal reports of using various hydroxychloroquine or chloroquine compounds were incomparable for various reasons, among them the incomplete and inconsistent terminologies.
- ***Connecting different fields of research.*** In this case of methods, as e.g., in genetics and pathophysiology data of SARS-CoV-2 and COVID-19, applying FAIR principles consistently across the available information sources would be of great value. In [29] Jacobsen et al. declare that, by intent, the 15 guiding principles for FAIR do not dictate specific technological implementation. It is noted that this has also resulted in inconsistent interpretations that carry the risk of leading to incompatible implementation. It is also concluded that while the FAIR principles are formulated on a high level for true interoperability, we need to support convergence in implementation choices that are widely accessible and (re)-usable. Our own findings support this as well.
- ***Improving clinical response to COVID-19*** *(diagnostic and therapeutic issues).* Here, Schulz et al. [30] guide us: they explain that interpretation of clinical data is highly dependent on contexts, data is often un- or semi-structured, and it is difficult to repurpose even standardized data, e. g. for clinical epidemiology, data analysis, or decision support. However, it is emphasized that data interoperability gained attention due the value of large-scale predictive data analysis.
- ***Overcoming the problem of transparency and comprehensibility caused by the sheer amount of research information related to COVID-19.*** A broader approach of '*language and meaning*'

Table 4 Metadata element labels in various COVID-19 data sources

Type	Name (short version)	Metadata labels
Epi data	WHO CoV Dashboard	Date_reported, Country_code, Country WHO_region, New_cases, Cumulative_cases, New_deaths, Cumulative_deaths
	ECDC COVID 19 Worldwide data	dateRep, day, month, year, cases, deaths, CountriesAndTerritories, geoId, countryterritoryCode, popData2019, continentExp, Cumulative_number_for_14_days_of_COVID-19_cases_per_100000
	Johns Hopkins University (JHU) dashboard	FIPS, Admin2, Province_State, Country_Region, Last_Update, Lat, Long_, Confirmed, Deaths, Recovered, Active, Combined_Key, Incidence_Rate, Case-Fatality_Ratio
	Our World in Data COVID-19 database	iso_code, continent, location, date, total_cases, new_cases, total_deaths, new_deaths, total_cases_per_million, new_cases_per_million, total_deaths_per_million, new_deaths_per_million, new_tests, total_tests, total_tests_per_thousand, new_tests_per_thousand, new_tests_smoothed, new_tests_smoothed_per_thousand, tests_per_case, positive_ratetests_units, stringency_index, population, population_density, median_age, aged_65_older, aged_70_older, gdp_per_capita, extreme_poverty, cardiovasc_death_rate, diabetes_prevalence, female_smokers, male_smokers, handwashing_facilities, hospital_beds_per_thousand life_expectancy
	Worldometer Coronavirus data	country_other, total_cases, new_cases, total_death, new_death, total_recovered, active_cases, serious_critical, total_cases_per_million, death_per_million, total_tests, tests_per_million, population
virus genomics data	GISAID EpiCoV database	sequence + meta data: Virus name, Accession ID, Collection date, Location, Host, Additional location information, Gender, Patient age, Patient status, Passage, Specimen, Additional host information, Lineage, Clade
	NCBI SARS CoV-2 GenBank	sequence + meta data: LOCUS, RNA, DEFINITION, ACCESSION, VERSION, KEYWORDS, SOURCE, ORGANISM, REFERENCE, AUTHORS, TITLE, JOURNAL, COMMENT, FEATURES, source, /organisms, /mol_type, /isolate, /human, /CHN, /host, /db_xref, /country, /collection_date, gene, CDS, /gene, /codon, /product, /protein_id, /translation, mat_peptide,/codon_start
	CNC 2019 Novel Coronavirus Resource	Virus Strain, Name, Accession ID, Data Source, Related ID, Nuc.Completeness, Sequence Length,Sequence Quality, Quality Assessment, Host, Sample Collection Date, Location, Originating Lab, Submission Date, Submitting Lab, Create Time, Last Update Time
	EMBL-EBI'sCOVID-19 Data Portal	Accession, Collection date, Country, Host, Strain, Isolate, Location, Mol. type, Taxonomy
public health & policy actions	CORONAVIRUS GOVERNMENT RESPONSE TRACKER	CountryName; CountryCode; RegionName; RegionCode; Date; C1_School closing; C2_Workplace closing; C3_Cancel public event; C4_Restrictions on gatherings; C5_Close public transport, C6_Stay at home requirements; C7_Restrictions on internal movement, C8_International travel controls, E1_Income support, E2_Debt/contract relief, E3_Fiscal measures, E4_International support, H1_Public information campaigns, H2_Testing policy, H3_Contact tracing, H4_Emergency investment in healthcare, H5_Investment in vaccines, M1_Wildcard, ConfirmedCases, ConfirmedDeaths, StringencyIndex, StringencyIndexForDisplay, StringencyLegacyInde, StringencyLegacyIndexForDisplay, GovernmentResponseIndex, GovernmentResponseIndexForDisplay, ContainmentHealthIndex, ContainmentHealthIndexForDisplay, EconomicSupportInde,x EconomicSupportIndexForDisplay
	CoronaNet Research Project	record_id, policy_id, entry_type, correct_type, update_type, update_level, description, date_announced, date_start, date_end, country, ISO_A3, ISO_A2,, init_country_level, domestic_policy, province, city, type, type_sub_cat, type_text, school_status, target_country, target_geog_level, target_region, target_province, target_city, target_other, target_who_what, target_direction, travel_mechanism, compliance, enforcer, index_high_est, index_med_est, index_low_est, index_country_rank, link, date_updated, recorded_date
COVID-19 clinical data	WHO: Global COVID-19 Clinical Data Platform	MODULE 1: 1a. CLINICAL INCLUSION CRITERIA; 1b. DEMOGRAPHICS; 1c. DATE OF ONSET AND ADMISSION VITAL SIGNS(first available data at presentation/admission); 1d. COMORBIDITIES (existing at admission) (Unk = Unknown); 1e. PRE-ADMISSION AND CHRONIC MEDICATION Were any of the following taken within 14 days of admission; 1f. SIGNS AND SYMPTOMS ON ADMISSION (Unk = Unknown); 1g. MEDICATION On the day of admission, did the patient receive any of the following:; 1h. SUPPORTIVE CARE On the day of admission, did the patient receive any of the following:; 1i. LABORATORY RESULTS ON ADMISSION (*record units if different from those listed) MODULE 2. Daily follow up during hospital stay (daily or as frequent as possible based on feasibility) 2a. VITAL SIGNS (record most abnormal value between 00:00 to24:00) 2b. DAILY CLINICAL FEATURES (Unk = Unknown); 2c. LABORATORY RESULTS (*record units if different from those listed); 2d. MEDICATION At any time during this 24-hour hospital day, did the patient receive:; 2e. SUPPORTIVE CARE At any time during this 24-hour hospital day, did the patient receive: MODULE 3. Complete at discharge/death; 3a. DIAGNOSTIC/PATHOGEN TESTING; 3b. COMPLICATIONS At any time during hospitalization, did the patient experience; 3c. MEDICATION While hospitalized or at discharge, were any of the following administered: ;3d. SUPPORTIVE CARE At any time during hospitalization, did the patient receive/undergo:
paper collections	CDC Library COVID-19 Research	Date Added, Author, Title, Abstract, Year, Journal/Publisher, Volume, Issue, PagesAccession Number, DOI, URL, Name of Database, Database Provider, Language, Keywords
	WHO COVID-19 Global literature	ID, Title, Authors, Source, Journal, Database Type, Language, Publication year, Descriptor(s), Publication Country, Fulltext URL, Abstract, Entry Date, Volume number, Issue number, DOI
	NLM LitCovid	(RIS:) TY, AN, TI, JO, A1, AB, DO, KW, PY, ER
	CORD-19	cord_uid, sha, source_x, title, doi, pmcid, pubmed_id, license, abstract, publish_time, authors, journal, mag_id, who_covidence_id, arxiv_id, pdf_json_files, url, s2_id

Table 5 FAIR-ness of COVID-19 information sources

Type	Name	Findable	Accessible	Interoperable	Reusable
epidemiological data sources:	WHO Coronavirus Disease (COVID-19) Dashboard	partial	yes	partial	yes
	ECDC COVID 19 Worldwide data	partial	yes	partial	yes
	COVID-19 Dashboard by CSSE at Johns Hopkins University (JHU)	partial	yes	partial	yes
	Our World in Data COVID-19 database	partial	yes	partial	partial
	Worldometer Coronavirus data	partial	yes	partial	partial
virus genomics data sources:	GISAID EpiCoV database	yes	partial	partial	yes
	NCBI SARS CoV-2 GenBank	yes	yes	partial	yes
	China National Center for Bioinformation	yes	partial	partial	partial
	EMBL-EBI's COVID-19 Data Portal	yes	yes	partial	yes
databases of public health policy actions	Oxford Stringency Index	yes	yes	partial	partial
	CoronaNet Research Project	partial	yes	partial	partial
paper collections (textual databases)	CDC Library COVID-19 Research	partial	yes	partial	yes
	WHO COVID-19 Global literature	partial	yes	partial	yes
	NLM LitCovid	yes	yes	yes	yes
	CORD-19	partial	yes	partial	yes

Findable: It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier. Checks: (i) A persistent identifier is assigned to your data, (ii) There are rich metadata, describing your data, (iii) The metadata are online in a searchable resource e.g. a catalogue or data repository, (iv)The metadata record specifies the persistent identifier.

Accessible: It should be possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate. FAIR does not mean that data need to be open! There should be metadata, even if the data are not accessible. Checks: (i) Following the persistent ID will take you to the data or associated metadata, (ii) The protocol by which data can be retrieved follows recognized standards e.g. http, (iii) The access procedure includes authentication and authorization steps, if necessary, (iv) Metadata are accessible, wherever possible, even if the data are not.

Interoperable: Data and metadata should conform to recognized formats and standards to allow them to be combined and exchanged.
checks: (i) Data is provided in commonly understood and preferably open formats, (ii)The metadata provided follows relevant standards, (iii) Controlled vocabularies, keywords, thesauri or ontologies are used where possible, (iv) Qualified references and links are provided to other related data.

Reusable: Lots of documentation is needed to support data interpretation and reuse. The data should conform to community norms and be clearly licensed, so others know what kinds of reuse are permitted. Checks: (i) The data are accurate and well described with many relevant attributes, (ii) The data have a clear and accessible data usage license, (ii) It is clear how, why and by whom the data have been created and processed, (iii) The data and metadata meet relevant domain standards.

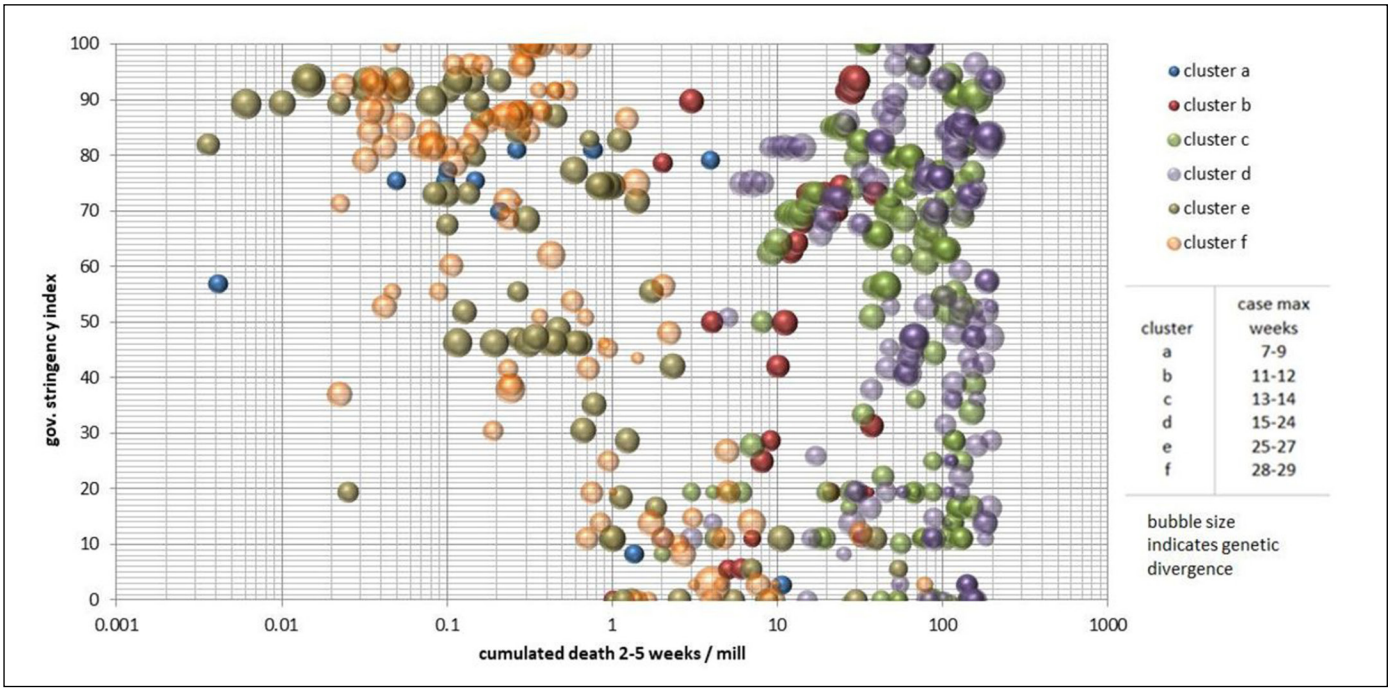

Fig. 4 COVID-19 pandemic severity versus stringency of governmental public health measures - showing also genetic diversity of countries, clustered along the timeline of maximal case numbers

tools might apply. It is impossible to critically follow a certain area if the number of papers is above 200-300 per day. Applying artificial intelligence tools to such order of magnitude of research information seems to be the way to overcome this. Authors of this paper in a previous position paper of the LamB WG [31] concluded that, for artificial intelligence in medicine, "… *neural networks, robotics, and machine learning are the research areas with the largest number of indexed publications … Knowledge representation publications … expose the highest interplay … (of various fields) … The notion of knowledge representation might play both a historical and foundational role in the various areas, providing a common cognitive layer, a still needed context, even for domains such as machine learning, neural nets, fuzzy logic, and robotics …*". Applying tools and methods of these specific areas might be the proper response to overcome the above-mentioned difficulties in cooperation with bioinformatics and network medicine.

4 Discussion

4.1 Quantitative Characteristics

Regarding the lack of further growth in literature numbers since May on Figure 1. – authors of this paper suggest that this shows a kind of scientific community "bandwidth" or capacity of scientific publication channels. This ceiling seemingly appears at roughly 2,000 – 2,200 papers per week (see also the red tendency arrows). The relatively low number of forecasting - modelling papers probably shows that so far there are significantly less scientific resources available for this important area compared to other aspects of the COVID-19 pandemic. This could mean as well that the human capacities for forecasting and modelling resources were full-time involved in daily operational tasks or there is insufficient data to develop performant models. In this case there will be an increase of modelling papers in the future. A third reason for this could be the lack of enough validated, controlled data.

Another disputable aspect is whether the relative growth of COVID-19 related literature, perceived as unprecedented and impressive by numbers, as shown in in Table 3, does match the general tendency of growth of scientific publications in general or outweigh it. This should be further investigated.

4.2 Qualitative Aspects

FAIR-ness testing: While getting the results for practical usability testing, we came across some issues of compiling epidemiological, genomics and government measures' stringency data, as described above in the *Results* section. In addition to the problems described there, we have to think about some possible interesting cognitive dissonances: i.e. is the 'number of infections' data or metadata of an outbreak? Obviously, for an index-based disease surveillance database this number is 'data', while for an outbreak event database the same number is a descriptive metadata of a given outbreak. Similarly, an aggregate of cases per country could be considered metadata if you consider the patient-level as data, but it is data if your study object is "country".

Could '*language and meaning'* methods improve clinical response in connection to field epidemiology for COVID-19 (diagnostic and therapeutic issues)? Authors of this paper agree with [30] that difficulties in interpreting COVID-19 pandemics literature highlight a need of data standards for making clinical data interoperable and shareable in a virtuous cycle of continuous improvement for field epidemiology as well. We also support the need for the application of the eStandards methodology - aligning reusable interoperability components, specifications, and tools.

Limitations: In this paper we focus only on the scientific part of the 'infodemic' phenomenon – we do not deal with the general media infodemic. Specifically, due to scope and mandate limits, we use one (albeit outstanding) source, called 'COVID19 Article Collection' of PubMed - from 'LitCovid', that is "*a curated literature hub for tracking up-to-date scientific information about the 2019 novel Coronavirus. It is* (declared to be) *the most comprehensive resource on the subject, providing a central access to …. a growing (number of) relevant articles in PubMed. The articles are updated daily and are further categorized by different research topics and geographic locations for improved access.*" [5].

In addition to that we do not offer detailed quality analysis of the information in the studied literature as this is out of the scope of the present study, however we recognize the need for such further investigations. A possibly promising comparison of the contents of the relevant bioRxiv/medRxiv COVID selection with LitCovid pre-prints is also missing from this study.

5 Conclusions

Summing up characteristics and some peculiarities of COVID-19 literature it is remarkable that never were that many papers dedicated to a pandemic. A certain "saturation" (~ 2,200-2,300 papers per week) might show either the upper limit of scientific capacities around the globe, or that of the scientific publication "bandwidth". As various on-line channels were and are available for scientific publication even without peer review, we think that the number of papers stabilizing at that level are an indicator of what the science community can produce. Never before were such amount of various open access database contents available appearing in such a short period of time.

At the same time, the vast potential of using these data was not fully brought home, and quality is often dubious. We argue that 'language and meaning' related methods and paradigms would contribute to better results. Specifically, using integrated terminologies in terms of homogenous semantics would lead to better, easier detection of inter-connectedness of results of various studies. Applying data and metadata standards for processing epidemiological and case related clinical information would lead to better comparable data, coming from various sources - as the need for data normalization, validation, cleaning would need less efforts. In general, principles of FAIR data handling would further enable machine and technical level interoperability and meaningful knowledge sharing in a virtuous cycle of continuous improvements.

References

1. Editors of Merriam-Webster. Words We're Watching: 'Infodemic' [cited 2020 Aug 10]. Available from: https://www.merriam-webster.com/words-at-play/words-were-watching-infodemic-meaning
2. Balkanyi L, Cornet R. History and charter of IMIA Working Group 'Language and Meaning in Biomedicine' , earlier called 'Medical Concept Representation' (Version v. 1.0.0.). Zenodo; 2019 Aug 22. Available from: http://doi.org/10.5281/zenodo.3374148
3. Mehra MR, Desai SS, Kuy SR, Henry TD, Patel AN. Retraction: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med 2020 Jun 25;382(26):2582.
4. Flack J, Mitchell M. Complex Systems Science Allows Us to See New Paths Forward [cited 2020 Aug 21]. Aeon. Available from: https://aeon.co/essays/complex-systems-science-allows-us-to-see-new-paths-forward
5. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. Nature 2020;579(7798):193. Available from: https://www.ncbi.nlm.nih.gov/research/coronavirus/
6. Sinclair S, Rockwell G. Voyant Tools [cited 2020 Aug 20]. Available from: https://voyant-tools.org/?view=ScatterPlot&corpus=a8dc9118215c367fe859cf811f49c68
7. Bakken S. The journey to transparency, reproducibility, and replicability. J Am Med Inform Assoc 2019;26:185-7.
8. Wilkinson M, Dumontier M, Aalbersberg I, Appleton G, Axton M, Baak A, et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016; 3:160018.. Available from: https://doi.org/10.1038/sdata.2016.18
9. Balkanyi L, Lukacs L, Dorkó B. Data set, combining epidemiological, genetics, and government stringency data of COVID-19 pandemic. Zenodo Data Set, DOI: 10.5281/zenodo.4152999, October 29, 2020
10. [cited: 2020 Aug 10]. Available from https://covid19.who.int/
11. [cited: 2020 Aug 10]. Available from https://www.ecdc.europa.eu/en/covid-19/data
12. [cited: 2020 Aug 10]. Available from https://medisys.newsbrief.eu/medisys/homeedition/en/home.html
13. [cited: 2020 Aug 10]. Available from https://coronavirus.jhu.edu/map.html
14. [cited: 2020 Aug 12]. Available from https://github.com/owid/covid-19-data/tree/master/public/data
15. [cited: 2020 Aug 12]. Available from https://www.worldometers.info/coronavirus/?zarsrc=130
16. [cited: 2020 Aug 13]. Available from https://www.epicov.org/epi3/frontend#f873d
17. [cited: 2020 Aug 13]. Available from https://www.ncbi.nlm.nih.gov/genbank/
18. [cited: 2020 Aug 13]. Available from https://bigd.big.ac.cn/gwh/browse/virus/coronaviridae
19. [cited: 2020 Aug 13]. Available from https://www.covid19dataportal.org/sequences?db=embl
20. [cited: 2020 Aug 13]. Available from https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker
21. [cited: 2020 Aug 13]. Available from https://www.coronanet-project.org/index.html
22. [cited: 2020 Aug 13]. Available from https://www.who.int/publications/i/item/WHO-2019-nCoV-Clinical_CRF-2020.4
23. [cited: 2020 Aug 13]. Available from https://www.cdc.gov/library/researchguides/2019novelcoronavirus/researcharticles.html
24. [cited: 2020 Jul 10]. Available from https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/?output=site&lang=en&from=0&sort=DATAENTRY_ASC&format=summary&count=20&fb=&page=1&skfp=&index=tw&q=
25. Wang LL, Lo K, Chandrasekhar Y, Reas R, Reas R, Yang J, Eide D, et al. CORD-19: The COVID-19 Open Research Dataset. ArXiv 2020 Apr 22;arXiv:2004.10706v2. [Preprint].
26. Jones S, Grootveld M. How FAIR are your data? Zenodo, 10.5281/zenodo.3405141, Nov. 2017. Available from: https://doi.org/10.5281/zenodo.3405141
27. Rauh O. Kmc and kmc User Guide. kmc version 001; June 2013 [cited 2020 Aug 13]. Available from: https://www.orauh.de/software/kmc-clustering-tool/
28. Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. Yearb Med Inform 2018

Aug;27(1):129–39.
29. Jacobsen A, Azevedo RM, Juty N, Batista D, Coles S, Cornet R, et al. FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2020; 2(1-2):10-29.
30. Schulz S, Chronaki C: Chapter 3: Standards in Healthcare Data, In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science; 2019. p. 19. Available from: https://doi.org/10.1007/978-3-319-99713-1_3
31. Balkanyi L, Cornet R. The Interplay of Knowledge Representation with Various Fields of Artificial Intelligence in Medicine. Yearb Med Inform 2019 Aug;28(1):27-34.

Correspondence to:
László Balkányi
Tel: +3620983022
E-mail: laszlo@balkanyi.hu