OXFORD

Briefings in Bioinformatics, 22(2), 2021, 812–822

doi: 10.1093/bib/bbaa418
Advance Access Publication Date: 18 January 2021
Method Review

Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview

Arianna Dagliati, Alberto Malovini, Valentina Tibollo and Riccardo Bellazzi

Corresponding author: Riccardo Bellazzi, Department of Electrical, Computer and Biomedical Engineering, University of Pavia, via Ferrata 5, Pavia 27100, Italy. Tel.: +39 0382 985720; Fax: +39 0382 985373; E-mail: riccardo.bellazzi@unipv.it

Abstract

The coronavirus disease 2019 (COVID-19) pandemic has clearly shown that major challenges and threats for humankind need to be addressed with global answers and shared decisions. Data and their analytics are crucial components of such decision-making activities. Rather interestingly, one of the most difficult aspects is reusing and sharing of accurate and detailed clinical data collected by Electronic Health Records (EHR), even if these data have a paramount importance. EHR data, in fact, are not only essential for supporting day-by-day activities, but also they can leverage research and support critical decisions about effectiveness of drugs and therapeutic strategies. In this paper, we will concentrate our attention on collaborative data infrastructures to support COVID-19 research and on the open issues of data sharing and data governance that COVID-19 had made emerge. Data interoperability, healthcare processes modelling and representation, shared procedures to deal with different data privacy regulations, and data stewardship and governance are seen as the most important aspects to boost collaborative research. Lessons learned from COVID-19 pandemic can be a strong element to improve international research and our future capability of dealing with fast developing emergencies and needs, which are likely to be more frequent in the future in our connected and intertwined world.

Key words: COVID-19 pandemic; Electronic Health Record; clinical research; international initiatives; data sharing

Introduction

The coronavirus disease 2019 (COVID-19) pandemic has clearly shown that major challenges and threats for humankind need to be addressed with global answers and shared decisions. Data and their analytics are crucial components of such decision-making activities. To this end, several initiatives have started to allow national and international sharing of different types of COVID-19 data, including molecular (from sequences to drug targets [https://www.covid19dataportal.org/] [1]), epidemiological (https://www.ecdc.europa.eu/en/covid-19-pandemic) and finally policies and intervention strategies data (https://www.coronanet-project.org/).

Rather interestingly, one of the most difficult aspects turned out to be reuse and sharing of accurate and detailed clinical data collected by Electronic Health Records (EHR), even if these data

Arianna Dagliati is a Research Fellow of the Department of Electrical, Computer and Biomedical Engineering of the University of Pavia. She holds a PhD and MSc in Biomedical Engineering and works in artificial intelligence in medicine and machine learning.
Alberto Malovini has an MSc in Biotechnologies and a PhD in Bioengineering and Bioinformatics. He supports research activity of ICS Maugeri hospital by the analysis of clinical data using advanced statistical and machine learning methods.
Valentina Tibollo has an MSc in Biomedical Engineering, and she works at ICS Maugeri. She coordinates the production and implementation of information technology platforms to support data collection and research activities in ICS Maugeri.
Riccardo Bellazzi, PhD, is Full Professor and Director of the Department of Electrical, Computer and Biomedical Engineering of the University of Pavia. Prof. Bellazzi also leads the Laboratory of Informatics and Systems Science of the ICS Maugeri hospital. He works in clinical research informatics, artificial intelligence in medicine, machine learning and telemedicine systems.
Submitted: 11 August 2020; **Received (in revised form):** 29 October 2020

have a paramount importance. EHR data, in fact, are not only essential for supporting day-by-day activities, but also they can leverage research and inform about effectiveness of drugs and therapeutic strategies. As a matter of fact, as also reported by Moore *et al*. [2], EHR and health information systems have a 2-fold nature, as they must provide flexible and robust point-of-care solutions and at the same time they can feed clinical research with invaluable real-world data.

Two recent retractions of papers published in leading medical journals [3, 4] witness how critical is the process of clinical data collection, curation and sharing.

The COVID-19 emergency has suddenly made visible what are the current problems and limitations, including old and new barriers, as well as the existing opportunities.

On the one hand, the pandemic has required a sudden physical reorganization of hospital wards and a full redesign of clinical workflows. Sometimes this has forced a redesign of parts of the hospital information systems, including the introduction of new terms and definitions, with a consequent delay in data collection and following analysis [5]. Moreover, the pandemic has pushed telemonitoring of COVID and non-COVID patients, suddenly putting into practice telemedicine solutions [6, 7]. On the other hand, the need for coordinated multicentre studies has been made clear. Although a number of consolidated international projects and consortia had the chance to show the potential of their approaches, the lack of standardization and interoperability has still emerged as one of major obstacles for efficient and effective data sharing. Moreover, strict privacy regulations and political concerns are slowing down international collaborations. Coordinating Institutional Review Board (IRB) processes among centres across the continents, for instance, is very complex and sometimes turns out to be an insurmountable challenge.

The COVID-19 pandemic triggered an unprecedented growth of collaborative efforts, deployment of analysis frameworks and an extraordinary generation of scientific literature, which has been largely made available as preprint: nearly 8000 papers have been uploaded on medRxiv (https://www.medrxiv.org/). Given not only the rapid evolution of the COVID-19 pandemic, but also the fast pace at which we are accumulating novel knowledge about the disease and developing tools to gain more information to infer more refined knowledge, conducting a systematic review appears to be unavailing.

PubMed and Scopus search using the query 'COVID-19' AND 'Electronic Health Record*' AND 'shar*' retrieved a relatively small number of manuscripts (Supplementary Material S1) pertinent to this overview. Thus, while including these evidences, we extend this overview on the basis of our knowledge of health informatics initiatives to collect international federated EHR data based on *de facto* standards for clinical data sharing such as Informatics for Integrating Biology and the Bedside (i2b2) (https://www.i2b2.org/) and Observational Health Data Sciences and Informatics (OHDSI) (https://ohdsi.org/).

Although the main purpose of the overview was to underline the importance on international and collaborative informatics infrastructures (described in the section 'Clinical and epidemiological data collected from EHRs into standardized formats'), we further investigate initiatives to create open access data portals (in 'Open Data portals') leveraging on our knowledge of institutional schemes, such as the ones promoted by the European Union (EU). To the best of our knowledge, while several countries collected clinical data at a national level (https://www.england.nhs.uk/contact-us/privacy-notice/how-we-use-your-information/covid-19-response/coronavirus-covid-19-research-platform/, https://www.aihw.gov.au/covid-19), none of them made them accessible to non-government establishments for independent researchers.

As thoroughly described in [8, 9] and in [10], digital technologies for big data analytics, next-generation telecommunication networks and artificial intelligence might play a crucial role to tackle major problems related to the management and containment of the pandemic. Among these disruptive digital technologies, we explored those that might be used to gather and handle patients' data and that can potentially be integrated into health informatics systems: data lakes and blockchains. PubMed and Scopus queries and results including key term 'COVID-19', 'Data Lake*' and 'Blockchain*' are reported in the Supplementary Material S1.

Once we had introduced some of the available health informatics tools (i.e. collaborative data infrastructures, databases and digital technologies), which have the potential of accelerating our discoveries about the epidemiology, pathophysiology and healthcare system dynamics of COVID-19, we discuss open challenges of data sharing and data governance and foreseeable future directions to enhance and support the clinical research in the COVID-19 pandemic.

Available tools: collaborative infrastructures, institutes networks, shared databases and digital technologies

System interoperability, shared knowledge of structured processes, common standards and terminologies are the pillar for data sharing. The actual implementation of data-sharing systems can then follow different purposes; data can be gathered and released with different formats and data silos organized accordingly for specific aims.

As noted in [11], to date, healthcare professionals do not have access to robust data-sharing systems for large-scale, real-time analysis. This is an important limitation for epidemiological studies and to develop treatment protocols, especially given the absence of clinical trials and the obstacles for rapidly setting them up. Authors state that, 'when considering COVID-19, the insight we could gain from a pooled, publicly available dataset analysed by researchers in academic institutes and industry is invaluable and necessary'. We agree with this claim; furthermore, we underline the importance of deploying fast access to clear and shared policies, which should be promptly communicated and could be easily implemented into healthcare informatics systems.

Cosgriff *et al*. [11] envision 'a unifying multinational COVID-19 electronic health record waiting for global researchers to apply their methodological and domain expertise'. Although the adoption of internationally shared EHRs systems still seems a non-viable solution, especially due to data protection regulations, there are few promising initiatives. Some of these are based on experiences and infrastructures developed before the COVID-19 pandemic, which were aimed at gathering clinical and epidemiological data and at making these data available to international researchers for joint studies and meta-analyses.

In the following, and in Table 1, we report several of current national and international initiatives promoted by governments, academia and industry for providing shared informatics infrastructures and databases.

This overview is not meant to be exhaustive and we have confidence that many other collaborative projects will be initiated and released in the following months.

Table 1. Collaborative infrastructures

Name	Resource Link	Founders	Description	Accessibility	Format
4CE	https://covidclinical.net/	4 CE—Consortium for Clinical Characterization of COVID-19 by EHR	International consortium for EHR data-driven studies. The goal is to inform doctors, epidemiologists and public about COVID-19 patients with data acquired through the healthcare process	Free download of aggregated data provided strictly for research purposes	Files available in csv format
AWS data lake	https://aws.amazon.com/it/covid-19-data-lake/	Amazon AWS	Hosted on the AWS cloud, this curated data lake contains useful datasets such as COVID-19 case tracking data from The New York Times, COVID-19 testing data from the COVID Tracking Project, hospital bed availability from Definitive Healthcare, health survey data from the Delphi Research Group and research data from over 45 000 articles about COVID-19 and related coronaviruses from the Allen Institute for AI	It requires an Amazon AWS account	Unstructured data
C3.ai data lake	https://c3.ai/products/c3-ai-covid-19-data-lake/	C3.ai	Daily case reports, epidemiology line lists, genomic sequences of COVID-19 nucleotide and protein samples, collection of journal articles, repositories of clinical assets related to COVID-19 such as CT and X-ray lung images, patient test results, vaccine coverage, active therapeutics and clinical trials, data sources on movement trends during COVID-19 in different geographies around the world, collection of actions and policies taken by government and regulatory bodies to address COVID-19	Upon registration	Unstructured data
CORD-19	https://www.semanticscholar.org/cord19	The Semantic Scholar team—Allen Institute for AI	A free resource of >130 000 scholarly articles	Free download corpus	Annotated corpus
CORONANet	https://www.coronanet-project.org/index.html	NYU Abu Dhabi, Hochschule für Politik at the TU Munich, Yale University	The data yields detailed information on the level of government responding to the COVID-19 crisis with focus on specific actions taken and the geographical areas targeted by these measures	Free download	Files available in csv format
Coronavirus Disease Dashboard	https://covid19.who.int/	WHO	Trends over time and querying and retrieving information about epidemic summary statistics by country	Free download	Files available in csv format

(Continued)

Table 1. Continued

Name	Resource Link	Founders	Description	Accessibility	Format
COVID-19 research database	https://covid19researchdatabase.org/	Public–private consortium (Datavant, Health Care Cost Institute, Medidata, Mirador Analytics, Veradigm, Change Healthcare, Snowflake)	The database includes de-identified and limited datasets from medical and pharmacy claims data, EHR data, mortality data and consumer data. More information, including coverage, data dictionaries and update frequency is available on our knowledge base for approved researchers	It requires registration. The database can be accessed by academic, scientific and medical researchers conducting real-world data studies related to COVID-19. Although researchers may come from any sector, only non-profit, non-commercial projects related to COVID-19 or pandemics will be considered. All results must be made publicly available, preferably through peer-reviewed publications	–
EU Open Data Portal	https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data	European Centre for Disease Prevention and Control	The dataset contains the latest available public data on COVID-19, including a daily situation update, the epidemiological curve and the global geographical distribution (EU/EEA and the UK, worldwide). The updates come from EU/EEA countries through the Early Warning and Response System (EWRS), The European Surveillance System (TESSy), the World Health Organization (WHO) and email exchanges with other international stakeholders	Free download for daily situation update. Access to TESSy data for individuals nominated by the EU/EEA countries, European Commission, EU bodies, international organizations and other entities, following the TESSy nomination procedure. Access to aggregated published data is unrestricted	Files available in csv, xlsx, json, xml formats
European Data Portal	https://www.europeandataportal.eu/en/about/european-data-portal	European Union	Collection of datasets that are directly or indirectly related to COVID-19, organized in categories. It includes epidemiological data, patient-level and population-level data, data related to the impact on lifestyle (food price monitor, slaughtered bovine animals)	Freely accessible URL to the datasets resource	URL to the datasets resource
Microsoft data lake	https://azure.microsoft.com/en-us/services/open-datasets/catalog/covid-19-data-lake/	Microsoft	COVID-19-related datasets from various sources, covering testing and patient outcome tracking data, social distancing policy, hospital capacity, mobility, etc.	Free download	Files available in csv and json formats

(Continued)

Table 1. Continued

Name	Resource Link	Founders	Description	Accessibility	Format
National COVID Cohort Collaborative (N3C)	https://ncats.nih.gov/n3c/about	NCATS (CTSA, Clinical and Translational Science Awards) Program hubs, the National Center for Data to Health (CD2H)	It contains real-world data from patients who were tested for COVID-19 or whose symptoms are consistent with COVID-19, as well as data from individuals infected with pathogens such as SARS 1, MERS and H1N1, which can support comparative studies	Under an approved Data Use Agreement with NCATS, anyone can access N3C data after receiving approval for their Data Use Request. N3C users can include, but are not limited to, non-profit or not-for-profit organizations, federal, state and local health departments, researchers from industry and citizen scientists. Access is dependent on the level of data being requested, and IRB approval may be needed	OMOP CDM
OHDSI-CHARYBDIS	https://data.ohdsi.org/Covid19CharacterizationCharybdis/	The Observational Health Data Sciences and Informatics (OHDSI)	It contains baseline demographic, clinical characteristics, treatments and outcomes of interest among individuals tested for SARS-CoV-2 and/or diagnosed with COVID-19 overall and stratified by sex, age and specific comorbidities. As more data becomes available, it will include additional databases that are formatted to the OMOP-CDM. There are now 13 databases from three different continents and from seven European countries	The analytical code is available at https://github.com/ohdsi-studies/Covid19CharacterizationCharybdis.	OMOP CDM
UK Biobank	http://biobank.ndph.ox.ac.uk/ukb/exinfo.cgi?src=COVID19	University of Oxford	It contains data about the health and well-being of 500 000 volunteer participants. It is now receiving COVID-19 test data for UK Biobank participants in England and more frequent updates of data on deaths, inpatient hospital admissions, including intensive care and primary care data	Upon registration (approved UK Biobank researcher)	–

Clinical and epidemiological data collected from EHRs into standardized formats

The simplest approach to quickly release shared data is to rely on previous resources and integrate them with specific COVID-19 information. That is the approach followed by the UK Biobank (http://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=COVID19), which is receiving COVID-19 test data for UK Biobank participants and frequent updates on deaths, inpatient hospital admissions—including intensive care—and primary care data. The great advantage given by this kind of approach is to have already coupled clinical, omics and patients generated data in standardized formats and to already have in place both information infrastructures and data protection policies allowing to give access to patient-level data.

Among other initiatives that leverage on pre-existing communities and data-sharing networks, there is the one promoted by the OHDSI community (https://www.ohdsi.org/covid-19-updates/), namely the project CHARYBDIS (Characterizing Health Associated Risks, and Your Baseline Disease In SARS-COV-2) (https://data.ohdsi.org/Covid19CharacterizationCharybdis/). Data are collected from international general practice, hospitals and outpatient's specialist EHRs and organized following the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [12]. The project is aimed at describing baseline demographic, clinical characteristics, treatments and specific outcomes among individuals diagnosed with COVID-19, also using seasonal influenza infections as a benchmark. Preliminary results indicate the feasibility of exploiting these data for developing risk scores of hospital admission, intensive care unit admission and fatality [13] and for detailed characterization and phenotyping of hospitalized patients [14].

Within the 4CE consortium (https://covidclinical.net/), a large international community of researchers set up a network of 96 hospitals across five countries to answer some of clinical and epidemiological questions around COVID-19 through data harmonization, analytics and visualizations. Contributors utilized the i2b2 or OMOP platforms to map to a CDM. The 4CE initiative showed how an international consortium was able to rapidly harmonize and integrate EHR data, thanks to a strict collaboration among clinicians and bioinformatics who understood both technical characteristics and clinical relevance of data. Preliminary results [15] show laboratory trajectories abnormalities in several tests and discuss the importance of interoperability and data alignment.

International collaborations could certainly be facilitated by the creation of national cohorts that—as advisable—will be made accessible to international researchers. A National Institutes of Health effort, called the National COVID Cohort Collaborative (N3C) (https://ncats.nih.gov/n3c/about), aims to build a centralized US data resource [16]. Data are systematically collected from EHRs and include clinical, laboratory and diagnostic information. Harmonized data are mapped into an OMOP CDM and shared with participating partners via a cloud-based research environment. From the health informatics point of view, the most interesting and ambitious goal of N3C is to create an analytics platform on top of the data collected into a CDM, to enable collaborative analyses and that can be reused for other diseases in the future.

Finally, the Secure COllective Research (SCOR) consortium has developed a secure infrastructure using advanced privacy and security technologies to support data federation, query and analytics in a distributed manner. SCOR is based on the Collective protection of medical data (MedCO) data analytics platform, which is implemented exploiting homomorphic encryption and secure multiparty methodology to ensure privacy and, at the same time, multicentric analysis. A total number of 24 centres around the world are involved in this project [17].

Open Data Portals

Data collected at patient level from EHRs have invaluable potential for accelerating knowledge discovery and to support clinical research and practice. Nevertheless, their value could be augmented, thanks to their integration with ancillary information derived from open datasets regarding national policies, surveillance plans, socio-economics factors and populations' characteristics that could influence both available treatments and outcomes. Therefore, these aspects should be evaluated as potential confounders. We reviewed some of the available data portals *ad hoc* created for COVID-19 and hosted by pre-existing platforms. Detailed information is reported in Table 1.

Open data best practices have been promoted and supported by the EU through several projects, among which the European Data Portal (https://www.europeandataportal.eu/en/about/european-data-portal), containing a wide range of metadata and datasets organized according to specific categories (including economy, public sectors, health, population and society). As a response to the pandemic, they have created a collection of datasets directly or indirectly related to COVID-19 with information on the COVID-19 research, epidemiology and healthcare systems. The EU Open Data Portal (https://data.europa.eu/euodp/en/home) gives access to open data published by EU institutions and bodies among the ones published by the European Centre for Disease Prevention and Control (https://www.ecdc.europa.eu/en/covid-19/data-collection). The EU Open Data Portal contains the latest available public data on COVID-19, including daily situation updates, epidemiological curves and global geographical distributions.

At a global level, World Health Organization (WHO) provides a Coronavirus Disease Dashboard (https://covid19.who.int/), which allows rapidly visualizing contagion trends over time and querying and retrieving information about epidemic summary statistics by country.

As previously mentioned, governments responses to the pandemic are important factors to consider and to possibly include in integrated analyses. The CoronaNet Research Project compiles a database on various fine-grained actions governments are taking to defeat the coronavirus, including travel bans and investments in the public health sector (https://www.coronanet-project.org/index.html). Thanks to the collaboration of political, social and public health science scholars, they have collected >10 000 international policy announcements [18]. Data, reports and analytics tools can be downloaded from their site.

Another key aspect to consider to get valuable insight to support the research on COVID-19 is the impressive amount of scientific literature produced in a very short time frame. The COVID-19 Open Research Dataset (CORD-19) (https://www.semanticscholar.org/cord19) provides a daily updated corpus of >130 000 scholarly articles about COVID-19. The goal is to support the application of natural language-processing approaches and help researchers to overcome the possible information overload.

Digital technologies: data lakes and blockchains

Given the rapidly evolving situation and the amount of heterogeneous data collected during the pandemic, data lakes seem another ample opportunity to explore, especially for the implementation of machine learning frameworks. Data lakes are storage repositories that hold a vast amount of raw data in its native format, including both structured and unstructured data. Such storage solutions can be rapidly set up in presence of adequate infrastructures to manage them; on the other hand, querying data lakes could be challenging due to their unstructured nature. Although it is important to underline industry efforts for releasing COVID-19 data lakes, see, for example, Amazon (https://aws.amazon.com/it/blogs/big-data/a-public-data-lake-for-analysis-of-covid-19-data/), Microsoft (https://azure.microsoft.com/en-us/services/open-datasets/catalog/covid-19-data-lake/) and C3.AI (https://c3.ai/products/c3-ai-covid-19-data-lake/), here we focus on specific matters related to the collection of medical data. As further discussed in the following sections, data governance is often associated with cumbersome processes that slow down data gathering efforts and affect data availability. In [19], the authors propose a data lake-based strategy for clinical data repositories to achieve legal interoperability for research purposes. In the COVID-19 pandemic, the proposed approach allows rapidly combining and making available legally compliant datasets, responding to the necessity to harmonize international legal requirements and supporting global collaborations in context where data governance heterogeneity might hinder rapid analyses and responses. Data lakes can provide fast solutions for big data collection. However, once not-harmonized data are collected, they can hardly provide meaningful insights and derive clinical value if not carefully manipulated. In [20], authors propose the use of a semantic approach based on automated ontology mapping and merging, where data lakes are used to interoperate ontologies for learning object retrieval and reuse from local to global ontology. In this way, data lakes were exploited to extract information from heterogeneous data sources and generate real-time statistics and reports.

Blockchains provide decentralized computational architectures and data management technology, so that actions on data (such as transactions) take place in a decentralized manner. One of their main features of interest is to provide security, anonymity and data integrity without the control of third-party organizations [21]. Blockchains have been identified as a key technology for fighting against COVID-19 in several case applications and service opportunities [22–24], such as contact tracing, supply chain management, online education, e-government and patient information sharing, where blockchains are seen as a possible solution to preserve privacy and quickly and accurately share clinical information.

As reported in [25], the use of blockchain can facilitate the creation of generalizable predictive systems in healthcare and contain infections. Authors conducted a rigorous analysis in opportunities and limits to blockchain-based adoption within the COVID-19 pandemic and conclude that blockchain, applied to the health sector, can offer effective opportunities to improve prevention activities, management of clinical risk, patient data and EHR data, and also the scientific research and the divulging of scientific knowledge. Fusco *et al*. [25] underline how the adoption of blockchain systems as a bridge to ensure cross-communication might overcome the issue of interoperability among different EHR systems and allow the rapid collection and sharing of healthcare data respecting privacy and security. Also pertinent to the focus of this overview, we want to highlight the opportunities of using blockchain for improving the exchange of health records, especially for enhancing patient-centric interoperability and user-centred medical research [26] and to ensure secure and effective EHR data sharing that allows personal medical data remaining in control of the patient [27].

Challenges and directions

The COVID pandemic made clear that the health informatic community agrees and strongly demands for unified frameworks for sharing and exchanging digital epidemiological data and, accordingly with data protection regulations, facilitating the flow of information between health workers, stakeholders, policy makers and the public.

The demand for digital data sharing also raised some crucial discussion points.

Interoperability is key

Beyond obvious issues related to different international health systems and organizations, a substantial lack of cohesive data models in EHR and poor interoperability emerged during the pandemic, spotlighting a very well-known weakness of health information technology (IT) infrastructures [28].

In fact, while noteworthy efforts have been devoted towards syntactic and semantic interoperability over the last 50 years (http://www.hl7.org, https://loinc.org/, http://www.snomed.org/, https://www.npu-terminology.org/), the way in which individual-level data are collected and coded can be extremely different even between institutions within the same country.

Even if there is a general agreement that FAIR data principles (Findability, Accessibility, Interoperability and Reusability) should apply to EHR data, too, practice is often showing a disappointing reality [29, 30] .

Data sharing initiatives could be dramatically limited by such heterogeneity of data formats and standards. Therefore, data requires a long and painful pre-processing phase, which consists of variables mapping between coding standards and releases before being shared with others to contribute to multicentric studies.

Studies about COVID-19 pandemic are more than others affected by extreme heterogeneity in terms of data standardization. This is mostly due to the rapid spread and evolvement of the epidemic and to the limited time that research institutes had to organize data collection in a homogeneous way and to define and share vocabularies to represent a common 'core' set of new concepts.

To date, initiatives to improve the interoperability among systems for the diagnosis and treatment of COVID-19 are still relatively limited and being driven mostly by the collaborative efforts presented in the previous sections of our paper. Standards for data models have been exploited, such as i2b2 for the i2b2-Accrual to Clinical Trials (ACT) ontology, reported in (http://dbmi-ncats-test01.dbmi.pitt.edu/webclient/), supported by the 4CE and ACT consortia, OMOP within the ODHISI consortium and Clinical Data Interchange Standards Consortium (CDISC) standard for the data structure used in WHO data tools (https://www.cdisc.org/standards/therapeutic-areas/covid-19).

Furthermore, the HL7 international community indicates how (http://blog.hl7.org/hl7_fhir_applications_beging_to_support_better_response_to_covid-19) the COVID-19 pandemic is revealing possible novel application of the HL7's Fast Healthcare Interoperability Resource (FHIR) standard to share healthcare information and coordinate services. Some examples include situational awareness, patient questionnaire creation and communication via text and the definition of key data elements associated with the COVID-19 disease. Other very interesting examples regard the use of FHIR interoperability technology on top of the previously mentioned CORD open research dataset; a list of other CORD-19 semantic annotation projects is provided here (https://github.com/fhircat/CORD-19-on-FHIR/blob/master/README.md) to facilitate linkage with other biomedical datasets.

Another interesting initiative has been reported by Mishra *et al.* [31]. Authors adapted the FHIR-based architecture for infectious disease surveillance for sexually transmitted diseases to the general problem of outbreaks, showing the potential of this emerging technology in the COVID-19 scenario. Shared data models based on openEHR modelling, as the one presented in [32], should be more widely exploited for facilitating data exchange.

Let us finally note that interoperability mostly involves EHR vendors and providers. Effective data exchange will be possible only if different companies will work together to implement new strategies for systems communication, thus allowing patients data, currently trapped into single EHR silos, to become easily available to clinicians, researchers and patients themselves, for coordinated analyses and actions. *Ad hoc* informatics infrastructures to share EHR data would probably need to be jointly supported by governments, public and private entities to facilitate this process, as foresees for genomics in [33].

Processes awareness

EHR data are often gathered in a clinical processes-blind way. Although this is somehow manageable during routine practice, in emergency situations, this represents a critical bottleneck for assessing patients' risks and tailoring interventions. Difficulties in gathering EHR data into cohesive narratives imply partial views of patients' risk and the loss of essential timeline of health data [34]. For this reason, it would be always important to have a formal representation of the healthcare process underlying EHR to explicitly describe the context and assumptions of the specific implementation [35].

In addition to having the capability of sharing patients' careflows, EHR workflows should be flexible, embed clinical practice processes information and quickly adapt to sudden changes in clinical processes and guidelines.

An important step for building processes-aware systems is to learn processes in an objective way, possibly from the same EHR data. Analytics techniques such as process mining [36] allow for discovery, conformance and enhancement of processes, establishing a strong relation between the process model and the reality captured from data. If processes were learned in a systematic and structured way, they could be easily integrated into hospital information systems, thus accelerating the response to unexpected changes of clinical practice and ease EHR workflow adaptations to emergencies.

Process awareness is often fragmentary and incomplete, as the data from which it is gathered. This is an issue that health informaticians face for chronic diseases that have been studied—and for which data have been collected—for decades. To efficiently study the processes related to a novel disease with a global spread such as COVID-19, it is necessary to create evidence from a large amount of structured data collected in a very short time period but at an international level.

This is an unparalleled research opportunity to create shared systems for global scale analyses of clinical and processes data, especially if these repositories could be coupled with biological knowledge and omics data.

Data protection regulations

The global spreading of COVID-19 raised questions and challenges regarding privacy and security compliance. Among these, it is worth mentioning the presence of substantial differences in terms of data sharing and protection regulations among countries (e.g. Data Protection Regulation, General Data Protection Regulation [https://eur-lex.europa.eu/eli/reg/2016/679/oj] in Europe, Health Insurance Portability and Accountability Act [37] in United States) and the involvement of businesses that are not regulated by privacy rules, but that can collect information potentially useful to face COVID-19 pandemic. Beside these aspects, there are also potential cybersecurity issues that entities should account for.

As a consequence, health organizations are exceedingly risk-averse towards data sharing, even if regulations permit such activities. Therefore, some European groups [38] acknowledged 'an ethical obligation to use the research exemption clause of the European General Data Protection Regulation during the COVID-19 pandemic' for sharing digital health data for research purposes and support global collaborative health research efforts.

Research studies involving human subjects require IRB approval by local institutions and this process follows internal procedures. IRB protocol preparation and approval processes typically differ between institutions even within the same country. The key differences rely on documentation structure and contents, bureaucratic steps needed to envision the documentation, time required by the institution to process and to approve it, and number and nature of changes requested by the IRB. Thus, such differences represent a limitation for multicentric studies since each participating centre must deal with the IRB approval procedure independently according to internal regulations, causing a lack of synchronization between institutions and possible delays in starting research activities [39].

IRB approval is often tailored to specific scientific questions, but these can evolve rapidly, especially when dealing with new research challenges like the COVID-19 pandemic. Therefore, there is the need to be able to rapidly update IRB approvals based on the new requirements. For example, the possibility to obtain approvals for new specific tasks proposed in the context of the starting research activity could facilitate and speed up the whole process.

Data governance: keeping track of shared data

As a consequence of the spread of research initiatives about COVID-19 pandemic, research institutions are often involved in multiple activities by sharing data about local patients between different centres like the 4CE consortium (https://covidclinical.net/) and N3C partnership (https://ncats.nih.gov/n3c/about) or by starting independent internal research projects. Data from the same patients are typically shared between multiple studies, increasing the probability of non-independence of findings deriving from apparently unrelated research initiatives. Also, datasets are often generated incrementally and shared or analysed according to temporal 'releases' of increasing size, depending on the cumulative number of new cases collected.

Once published, results (summary statistics, odds ratios, ...,) from multiple scientific studies addressing the same question are often combined through meta-analysis, 'a quantitative, formal, epidemiological study design used to systematically assess previous research studies to derive conclusions about that body of research' [40]. By combining results, meta-analysis allows increasing the statistical power of the analyses planned. In this context, the potential non-independence of the findings reported in the studies considered could represent a heavy limitation and researchers should account for this possible bias when performing meta-analyses.

Starting from these observations, there is clear evidence about the need for *ad hoc* systems able to keep track of the different data releases generated by centres and shared between independent research initiatives. By keeping track of the data versioning and sharing history, researchers could have a more precise overview about the data used in different analyses, thus allowing to discriminate between independent and non-independent studies reducing potential methodological bias.

Rather interestingly, the importance of data governance and data stewardship in data reuse has been strongly advocated by the International Medical Informatics Association (IMIA) some years ago [41, 42]. IMIA proposed the data steward as a key actor to 'convey a fiduciary (or trust) level of responsibility toward the data'. Moreover, 'data governance is the process by which responsibilities of stewardship are conceptualized and carried out'. We can conclude that key responsibilities of the data steward are not only to comply with privacy regulations, but also to keep track of the different ways in which the same data sources are utilized and the evidence that is based on these data is extracted.

Conclusions

The COVID-19 pandemic needs multi-institutional data sharing strategies able to deal with manifold challenges that society is facing. Among them, the availability of reliable clinical data can further boost understanding of the disease, deepening insights on its time-varying nature, investigating the impact of different therapeutic strategies and finally informed decision-making.

From the researchers' perspective, the possibility to integrate EHR-derived information about patients' disease condition, treatments, interventions, clinical exams with other data sources is of paramount importance for a deeper comprehension of the COVID-19 disease mechanism and severity manifestation. Overmyer *et al.* [43] adopted a multiomic approach by quantifying thousands of different biomolecules from patients with and without COVID-19 in relation to their disease severity and outcomes. The integration of multiomics data showed good performances in predicting COVID-19 severity, allowing also to highlight informative features. A web-based tool (covid-omics.app) allows the scientific community to further explore the generated data. The Severe COVID-19 genome-wide association study (GWAS) Group [44] performed a genome-wide case–control association study on severe COVID-19 patients and controls and identified a cluster of genes representing a genetic susceptibility locus in patients with respiratory failure. Shen *et al.* [45] performed proteomic and metabolomic analysis of COVID-19 sera and identified differentially expressed factors correlating with disease severity and evidenced dysregulation of multiple immune and metabolic components in clinically severe patients. Taken together, these results confirm the relevance of an integrated approach to the COVID-19 disease characterization.

There are a number of important issues that need to be addressed to achieve such ambitious goals. First, the level of interoperability, both syntactic and semantic ones, of EHR is still far too low. Even if noteworthy efforts have been made over the last 50 years, ontologies, terms, languages and criteria of EHR design are still not properly standardized. This is not related to the unavailability of adequate solutions [46, 47], but rather to the combination of an underestimation of interoperability needs in the design of EHR, sometimes due to locking-in policies of software vendors [48] and of an insufficient capability of describing and formally representing healthcare processes and their information needs. The latter aspect is of fundamental importance to achieve interoperability and data interpretation in a comprehensive manner, considering the very nature of clinical data and their operational contexts. It is important to mention that the COVID-19 emergency has made this problem even more difficult, since hospitals and healthcare providers had to suddenly change their organization and careflows. As reported in our paper, other threats to successful data sharing are represented by different privacy regulations in different world regions, as well as by data protectionism that has recently emerged [49]. However, new and more flexible informed consent, as well as a more proactive and receptive role of IRB committees, are supporting international data sharing initiatives even in this complex scenario. Finally, data sharing implies also the introduction of stricter control on the process, not only to comply with privacy regulations, but also to avoid uncontrolled use of data for deriving evidence. To this end, a promising strategy is related to the introduction of a data steward that may apply data governance policies to support institutions overseeing data sharing not only from a legal viewpoint, but also from a holistic perspective for the benefit of individuals and of the society. Data stewards may provide also support to the more general aim of data quality, which currently is left to the policies of the single research networks.

It is finally important to remark that EHR and hospital data are not the only source of clinical data that are important to support understanding of disease evolutions determinants. It is now clear that in the challenge of managing the pandemic, the more the primary and secondary care have been coordinated among them and with social care organizations, the more the control has been effective [50]. The IT infrastructure in place to support such coordination is able to collect data that can be extremely useful for research purposes. For example, in the UW initiative, primary and secondary care have been strongly linked for the benefit of a better local management of the pandemic [51]. The collection of pre-hospital data has been demonstrated to be an important source of information for clinical research, as shown in [52]. More work on such integration can be extremely valuable to improve quality of data interpretation.

Lessons learned from COVID-19 pandemic can be a strong element to improve international research and our future capability of dealing with fast developing emergencies and needs, which are likely to be more frequent in our connected and intertwined world.

Key Points

- The coronavirus disease 2019 (COVID-19) pandemic has clearly shown that data and their analytics are crucial for handling this world-wide emergency.
- An important but difficult aspect is related to sharing of accurate and detailed clinical data collected by Electronic Health Records (EHR).
- EHR data are not only essential for supporting day-by-day activities, but also they can leverage research and support critical decisions about effectiveness of drugs and therapeutic strategies.

- In this paper, we will concentrate our attention on the current efforts related to collaborative data infrastructures to support COVID-19 research and on the open issues related to data sharing and data governance that COVID-19 had made emerge.
- Data interoperability, healthcare processes modelling, different data privacy regulations, and data stewardship and governance are seen as the most important aspects to boost collaborative research.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Acknowledgements

The authors gratefully acknowledge Luigia Scudeller (IRCCS Policlinico di Milano), Antonio Bellasi (Hospital Papa Giovanni XXIII, Bergamo) and John H. Holmes (University of Pennsylvania, Philadelphia) for the insightful comments and discussions. The authors also thank the 4CE consortium for giving us the opportunity to participate in this important international initiative.

References

1. Forster P, Forster L, Renfrew C, *et al*. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* 2020;**117**(17):9241–3.
2. Moore JH, Barnett I, Boland MR, *et al*. Ideas for how informaticians can get involved with COVID-19 research. *BioData Min* 2020;**13**:3.
3. Mehra MR, Desai SS, Kuy S, *et al*. Retraction: cardiovascular disease, drug therapy, and mortality in COVID-19. *N Engl J Med* 2020;**382**(26):2582.
4. Mehra MR, Desai SS, Ruschitzka F, *et al*. Retraction-hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* 2020;**395**(10240):1820. doi: 10.1016/S0140-6736(20)31324-6. Epub 2020 Jun 5.
5. Edelman LS, McConnell ES, Kennerly SM, *et al*. Mitigating the effects of a pandemic: facilitating improved nursing home care delivery through technology. *JMIR Aging* 2020;**3**(1):e20110. doi: 10.2196/20110.
6. Ford D, Harvey JB, McElligott J, *et al*. Leveraging health system telehealth and informatics infrastructure to create a continuum of services for COVID-19 screening, testing, and treatment. *J Am Med Inform Assoc* 2020;**27**(12):1871–7.
7. Judson TJ, Odisho AY, Neinstein AB, *et al*. Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. *J Am Med Inform Assoc* 2020;**27**(6):860–6.
8. Ting DSW, Carin L, Dzau V, *et al*. Digital technology and COVID-19. *Nat Med* 2020;**26**(4):459–61.
9. Vaishya R, Haleem A, Vaish A, *et al*. Emerging technologies to combat the COVID-19 pandemic. *J Clin Exp Hepatol* 2020;**10**(4):409–11.
10. Chamola V, Hassija V, Gupta V, *et al*. A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact. *IEEE Access* 2020;**8**:90225–65.
11. Cosgriff CV, Ebner DK, Celi LA. Data sharing in the era of COVID-19. *Lancet Digit Health* 2020;**2**(5):e224. doi: 10.1016/S2589-7500(20)30082-0. Epub 2020 Apr 28.
12. Stang PE, Ryan PB, Racoosin JA, *et al*. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med* 2010;**153**(9):600–6.
13. Williams RD, Markus AF, Yang C, *et al*. Seek COVER: development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. *medRxiv* 2020. https://www.medrxiv.org/content/10.1101/2020.05.26.20112649v4.full.
14. Burn E, You SC, Sena AG, *et al*. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. *Nat Commun* 2020;**11**(1):5009.
15. Brat GA, Weber GM, Gehlenborg N, *et al*. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit Med* 2020;**3**:109.
16. Rubin R. NIH launches platform to serve as depository for COVID-19 medical data. *JAMA* 2020;**324**(4):326.
17. Raisaro JL, Marino F, Troncoso-Pastoriza J, *et al*. SCOR: a secure international informatics infrastructure to investigate COVID-19. *J Am Med Inform Assoc* 2020;**27**(11):1721–6.
18. Cheng C, Barceló J, Hartnett AS, *et al*. COVID-19 Government Response Event Dataset (CoronaNet v.1.0). *Nat Hum Behav* 2020;**4**(7):756–68.
19. Thorogood A. Policy-aware data lakes: a flexible approach to achieve legal interoperability for global research collaborations. *J Law Biosci* 2020;**7**(1):lsaa065. doi: 10.1093/jlb/lsaa065.
20. Kachaoui J, Larioui J, Belangour A. Towards an ontology proposal model in data lake for real-time COVID-19 cases prevention. *Int J online Biomed Eng* 2020;**16**(9):123.
21. Zheng Z, Xie S, Dai H, *et al*. An overview of blockchain technology: architecture, consensus, and future trends. In: *IEEE International Congress on Big Data (BigData Congress)*, Honolulu, HI, 2017. p. 557–64.
22. Kalla A, Hewa T, Mishra RA, *et al*. The role of blockchain to fight against COVID-19. *IEEE Eng Manag Rev* 2020;**48**(3):85–96.
23. Chang MC, Park D. How can blockchain help people in the event of pandemics such as the COVID-19? *J Med Syst* 2020;**44**(5):102.
24. Shubina V, Holcer S, Gould M, *et al*. Survey of decentralized solutions with mobile devices for user location tracking, proximity detection, and contact tracing in the COVID-19 era. *Data* 2020;**5**(4):87. doi: 10.3390/data5040087.
25. Fusco A, Dicuonzo G, Dell'Atti V, *et al*. Blockchain in healthcare: insights on COVID-19. *Int J Environ Res Public Health* 2020;**17**(19):7167 Internet.
26. Rghioui A. Managing patient medical record using blockchain in developing countries: challenges and security issues. In: *2020 IEEE International conference of Moroccan Geomatics (Morgeo)*, Casablanca, Morocco, **2020**. p. 1–6.
27. Christodoulou K, Christodoulou P, Zinonos Z, *et al*. Health information exchange with blockchain amid COVID-19-like pandemics. In: *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, Marina del Rey, CA, USA, **2020** p. 412–7.
28. Shull JG. Digital health and the state of interoperable electronic health records. *JMIR Med Inform* 2019;**7**(4):e12712.
29. Prados-Suarez B, Molina C, Peña-Yañez C. Providing an integrated access to EHR using electronic health records aggregators. *Stud Health Technol Inform* 2020;**270**:402–6.
30. Leventhal JC, Cummins JA, Schwartz PH, *et al*. Designing a system for patients controlling providers' access to their

electronic health records: organizational and technical challenges. *J Gen Intern Med* 2015;**30**(Suppl 1):S17–24.

31. Mishra NK, Duke J, Lenert L, *et al*. Public health reporting and outbreak response: synergies with evolving clinical standards for interoperability. *J Am Med Inform Assoc* 2020;**27**(7):1136–8.
32. Li M, Leslie H, Qi B, *et al*. Development of an openEHR template for COVID-19 based on clinical guidelines. *J Med Internet Res* 2020;**22**(6):e20239. doi: 10.2196/20239.
33. Kaye J, Terry SF, Juengst E, *et al*. Including all voices in international data-sharing governance. *Hum Genomics* 2018;**12**(1):13.
34. Hripcsak G. Physics of the medical record: handling time in health record studies. In: Holmes J, Bellazzi R, Sacchi L, *et al*. (eds). *Artificial Intelligence in Medicine 15th Conference on Artificial Intelligence in Medicine, AIME 2015*, Pavia, Italy, Springer International Publishing, 2015, 3–6.
35. Maldonado JA, Marcos M, Fernández-Breis JT, *et al*. CLINIK-LINKS: a platform for the design and execution of clinical data transformation and reasoning workflows. *Comput Methods Programs Biomed* 2020;**197**:105616.
36. Van Der Aalst WMP. Process mining. *Process Min* 2011;**5**: 301–17.
37. Atchinson BK, Fox DM. The politics of the Health Insurance Portability and Accountability Act. *Health Aff (Millwood)* 1997;**16**(3):146–50.
38. McLennan S, Celi LA, Buyx A. COVID-19: putting the General Data Protection Regulation to the test. *JMIR Public Health Surveill* 2020;**6**(2):e19279. doi: 10.2196/19279.
39. Hall DE, Hanusa BH, Stone RA, *et al*. Time required for Institutional Review Board review at one veterans affairs medical center. *JAMA Surg* 2015;**150**(2):103–9.
40. Haidich AB. Meta-analysis in medical research. *Hippokratia* 2010;**14**(Suppl 1):29–37.
41. Geissbuhler A, Safran C, Buchan I, *et al*. Trustworthy reuse of health data: a transnational perspective. *Int J Med Inform* 2013;**82**:1–9.
42. Hripcsak G, Bloomrosen M, FlatelyBrennan P, *et al*. Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 Health Policy Meeting. *J Am Med Inform Assoc* 2014;**21**(2):204–11.
43. Overmyer KA, Shishkova E, Miller IJ, *et al*. Large-scale multi-omic analysis of COVID-19 severity. *Cell Syst* 2020;**S2405–4712**(20):30371–9. Epub ahead of print.
44. Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F, *et al*. Genomewide association study of severe Covid-19 with respiratory failure. *N Engl J Med* 2020;**383**(16): 1522–34.
45. Shen B, Yi X, Sun Y, *et al*. Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell* 2020;**182**(1): 59–72.
46. Akhlaq A, McKinstry B, Sheikh A. The characteristics and capabilities of the available open source health information technologies supporting healthcare: a scoping review protocol. *J Innov Health Inform* 2018;**25**(4):230–8.
47. Fechner M, Brix TJ, Hardt T, *et al*. Evaluation of openEHR repositories regarding standard compliance. *Stud Health Technol Inform* 2020;**270**:592–6.
48. Koppel R, Lehmann CU. Implications of an emerging EHR monoculture for hospitals and healthcare systems. *J Am Med Inform Assoc* 2015;**22**(2):465–71.
49. Lancieri FM. Digital protectionism? Antitrust, data protection, and the EU/US transatlantic rift. *J Antitrust Enforce* 2018;**7**(1):27–53.
50. Park S, Elliott J, Berlin A, *et al*. Strengthening the UK primary care response to COVID-19. *BMJ* 2020;**370**:m3691. doi: 10.1136/bmj.m3691.
51. Grange ES, Neil EJ, Stoffel M, *et al*. Responding to COVID-19: the UW medicine information technology services experience. *Appl Clin Inform* 2020;**11**(2):265–75.
52. Fernandez AR, Crowe RP, Bourn S, *et al*. COVID-19 preliminary case series: characteristics of EMS encounters with linked hospital diagnoses. *Prehosp Emerg Care* 2020;**25**(1):16–27. Epub 2020 Jul 31.