Abstract

This chapter will take a quick look at certain technological requirements associated with ‶precision,″ ‶personalized,″ or ‶patient-centered″ medicine. We will emphasize how the goals of precision medicine are advanced through bioimaging, and also how precision medicine may influence trial design. We use features of image-analysis and clinical trials to consider our overview of biomedical research methods initiated in Chapter 2. We also discuss biomedical text mining as a further dimension in research methodology, summarizing the CORD-19 corpus on Covid-19 related research as a case-study.

Chapter 3: Data Mining and Predictive Analytics for Cancer and Covid-19

1 Introduction

This chapter will touch on research methodology and clinical trial design in light of Patient-Centered paradigms and Precision Medicine. We will focus in particular on bioimaging/radiology, on trial architecture (particularly vis-à-vis Covid-19) and on text and data mining.

After examining how Precision Medicine affects requirements for bioimaging software, we will consider patient-centered priorities in the context of clinical trials. The goal of incorporating more granular patient-specific data within clinical observations in general translates over to clinical trials in particular, with trial-specific data models needing to incorporate patient-centered details at several stages, including recruitment of patients for trials, information management while trials are being conducted, and subsequent follow-up studies and/or analyses.

This chapter will conclude by examining text and data mining techniques oriented toward biomedical publications and data sets, on the premise that text and data mining represent a computationally sophisticated methodology in their own right. We will focus, as a case-study, on **CORD-19**, a large corpus of open-access Covid-related articles.

2 Precision Medicine and Bioimaging

Patient-centered research in the radiological context is centered on improving the precision of diagnostic-imaging techniques and corresponding clinical interventions. Indeed, the goal of contemporary radiology is not only to confirm diagnoses, but also to extract cues from medical images that suggest which course of treatment has the highest probability of favorable outcomes. A related goal is curating collections of diagnostic images so as to improve our ability to identify such diagnostic and prognostic cues, potentially using Machine Learning and/or Artificial Intelligence applied to large-scale image repositories.

The goal of building "searchable" image repositories has inspired projects such as the Semantic Dicom Ontology (**SEDI**)[1] and the **VISION** "structured reporting" system.[2] As explained in the context of **SEDI**:

> [If] a user has a CT scan, and wants to retrieve the [corresponding] radiation treatment plan ... he has to search for the RTSTRUCT object based on the specific CT scan, and from there search for the RTPLAN object based on the RTSTRUCT object. This is an inefficient operation because all RTSTRUCT [and] RTPLAN files for the patient need to be processed to find the correct treatment plan. [37, page 1]

[1] See https://bioportal.bioontology.org/ontologies/SEDI.

[2] See https://epos.myesr.org/esr/viewing/index.php?module=viewing_poster&task=&pi=155548.

Even relatively simple queries such as "display all patients with a bronchial carcinoma bigger than 50 cm^3" cannot be processed by **PACS** systems: "although there are various powerful clinical applications to process image data and image data series to create significant clinical analyses, none of these analytic results can be merged with the clinical data of a single patient."[3] These limitations partly reflect the logistics of how information is transferred between clinical institutions and radiology labs. In response, and in an effort to advance the science of diagnostic image-analysis, organizations such as the Radiological Society of North America (**RSNA**) have curated open-access data sets encompassing medical images as well as image-annotations (encoding feature vectors) that can serve as reference sets and test corpora for investigating analytic methods. Such repositories are designed to integrate data from multiple hospitals and multiple laboratories — bypassing the conventional data flows wherein radiological information is shared between clinicians and radiologists, but is not also merged into broad-spectrum corpora.

This renewed focus on patient outcomes has intriguing consequences for the scope and requirements of diagnostic-imaging software. In particular, the domain of radiological applications is no longer limited to **PACS** workstations where pathologists perform their diagnostic analysis, with the results transferred back to the referring institution (and subsequently available only through that institution's medical records, if at all). In the older, conventional workflow, radiographic images are requested by some medical institution for diagnostic purposes. Relevant information is therefore shared between two end-points: the institution which prescribes a diagnostic evaluation and the radiologist or laboratory which analyzes the resulting images. Building radiographic data repositories complicates this workflow because a third entity becomes involved — the organization responsible for aggregating images and analyses is generally distinct from both the prescribing institution and the radiologists themselves. As a result, both radiologists and prescribing institutions, upon participation in the formation of the target repository, must identify which image series and which patient data are proper candidates for the relevant repository.

For a concrete example, **RSNA** has announced the forthcoming publication of an open-access image repository devoted to Covid-19 (specifically the organization established two "task forces" in 2020; as of mid-2021 the overall project appears still to be under development, although several recent studies on Covid-19 in general have been published in **RSNA** journals).[4] This repository is being curated in collaboration with multiple European, Asian, and South American organizations so as to collect data from hospitals treating Covid-19 patients. Such a collaboration requires protocols both for data submission and for patient privacy and security.

2.1 The Basic Synthesis Between Bioimaging and Precision Medicine

Patient-centered research in the radiological context has several dimensions, including analysis of the rationales for diagnostic imaging in the first place: how well does image-based diagnostics actually

[3] See https://semantic-dicom.com/starting-page/.

[4] See https://www.rsna.org/covid-19.

correlate with improved patient outcomes? How can we quantify the experience of image-based testing itself (e.g., to identify factors such as cost, discomfort, and radiation danger which can rank some tests as less desirable than others, as one parameter to consider when deciding whether to prescribe imaging, and which modality)? These were among the questions addressed by the Patient-Centered Outcomes Research Institute's (**PCORI**'s) "Patient-Centered Research for Standards of Outcomes in Diagnostic Tests" (**PROD**) study [43], which also established a useful protocol for how medical institutions could provide reports on the experience and effectiveness of imaging from a patient's as well as clinician's perspective.

Assuming diagnostic imaging of a given modality *is* appropriate, an additional priority should then be to structure the diagnostic workflow — the image procurement, analysis, and data/metadata sharing protocols — to maximize the probability that subsequent clinical interventions are chosen which promote favorable outcomes on multiple patient-centered criteria, including quality of life and patient engagement. In the best case scenario, the goal of radiology is not only to confirm a diagnosis, but also to extract cues from medical images that suggest which course of treatment has the highest probability of favorable treatment outcomes.

While most radiologists and pathologists would probably agree with this assessment — that in the best-case scenario image-processing can yield predictive analytics which would sort patient populations into groups steered toward distinct treatments deemed most likely to be effective — there are technological and operational challenges to making patient-centered perspectives a central feature of diagnostic-imaging methodology. Effectively cross-referencing imaging and outcomes data requires integrating heterogeneous information obtained at different times and places. Some clinical data is associated with each patient at the time that a radiological (or other imaging) study is prescribed. The images themselves, and subsequent diagnostic reports, then provide layers of data that exist prior to the initiation of a course of treatment. Moreover, a rigorous data-integration protocol would need to incorporate information emerging *after* the treatment starts: descriptions and assessments of clinical outcomes, and, potentially, new data garnered by applying different image-analysis techniques. In brief, data-management protocols must be in effect before, during, and after image-analysis itself.

Ideally, image analysis can be powerful enough not only to identify a pathology, but to classify diagnoses into clusters based on recommended course of treatment. In order for analytic techniques to achieve this level of detail, however, it is necessary to arrive at a feedback loop where known clinical outcomes are associated with prior images, so that developers have those outcomes available as a further dimension of clinical data that may be statistically cross-referenced with image analysis.

The **PROD** study demonstrates that experiential factors should be evaluated — both in terms of testing itself (and its risks/costs) and in terms of post-diagnostic quality of life — as part of the data modeling treatment outcomes, and the comparative effectiveness of a selected course of treatments compared to alternative diagnostic methods and/or clinical interventions. From the perspective of standard data models, initiatives to cross-reference imaging and outcomes data include several Semantic Web ontologies, such as the Semantic **DICOM** Ontology (**SEDI**) and the **VISION** "structured reporting" system (both referenced earlier). The purpose of these ontologies is to standardize the terms through which radiographic procedures, analyses, and recommendations are described — more precisely or predictably than older technologies such as **DICOM** headers, **DICOM-RT**, and the **RADLEX** lexicon. By properly aligning image metadata spanning multiple patients, it is possible to create "searchable" image archives such that images can be selected or classified within a larger image collection, yielding image series or patient cohorts that can be studied through the lens of predictive modeling or patient-centered outcomes. Projects such as **SEDI** implement "semantic" **PACS** workstations where the space of known images is defined by a particular **PACS** system, but analogous techniques could be used to construct larger-scale image corpora as well, for research purposes, data mining, or as test-beds for code and algorithms. Patient-centered data points, such as those formulated via **PROD**, may then be incorporated as supplemental data.

2.2 Case Study: The Cancer Imaging Phenomics Toolkit

At the forefront of diagnostic imaging are new analytic techniques which statistically analyze image data, often detecting signals that are invisible to the naked eye, or at least processing groups of image-series on a scale beyond the reach of human radiologists. The proliferation of image-analysis methodologies places a new emphasis on *extensibility,* in which the capabilities of bioimaging software can be expanded in a decentralized, modular fashion.

A representative example of this new paradigm is **CAPTK** (Cancer Imaging Phenomics Toolkit), which provides a primary component supplying a centralized User Interface and taking responsibility for acquiring and loading radiographic/microscopy images, paired with multiple "peer" applications which can be launched from **CAPTK**'s main window, or alternatively used as standalone programs [10, especially section 4]. In particular, **CAPTK** provides an implementation (apparently the only **C++**-based implementation) of the Common Workflow Language (**CWL**), using this workflow model in conjunction with the **QT** Reflective Programming system to implement workflows connecting the central **CAPTK** application with its analytic extensions.[5] In effect, **CAPTK** achieves a workflow and messaging protocol for what they term "native," "standalone" applications, yielding an extensible architecture through which new image-analysis techniques can be integrated into an underlying **PACS** system (a detailed discussion of **CAPTK** module implementation is outside the scope of this chapter, but this book's supplemental materials include a more technical overview of **CAPTK**; for now we just use **CAPTK** as a case-study in modular design for bioimaging).

Taking the **RSNA** Covid-19 repository as a case study for promoting research into post-diagnostic outcomes, this repository is possible because an international team of hospitals and institutions have agreed to pool radiological data relevant to SARS-CoV-2 infection according to a common protocol. Taking **CAPTK** as a case-study in

[5] No native **C** or **C++** libraries are described on the **CWL** website among the tools and parsers available for **CWL**, but **CAPTK** is mentioned on a corresponding discussion thread concerning **C++** libraries. It seems therefore that the **CAPTK** "utilities" repository provides the de-facto standard **C++** implementation of **CWL**, at least according to the **CWL** group themselves.

multi-modal image analysis, this system is likewise possible because analytic modules can be wrapped into a plugin mechanism which allows many different algorithms to be bundled into a common software platform. Of course, these two areas of data-sharing overlap: one mission of repositories such as the **RSNA**'s is to permit many different analyses to be performed on the common image assets. The results of these analyses then become additional information which enlarges the repository proportionately.

The **CaPTk** project serve as a useful case-study for the analytic convergence or cross-referencing between image-analysis and outcomes/patient-centered data because is extensible as part of its essential design (although in terms of large-scale adoption more conventional **PACS** clients may also be used simply because **CaPTk** has certain software-engineering innovations which make it an outlier from an implementational point of view). Analysis of the aforementioned **RSNA** Covid-19 images, for example, could be enacted via specialized **CaPTk** modules, which could in turn be provisioned with Patient Outcomes and Comparative Effectiveness Research (**CER**) capabilities. That is, modular design at the image-processing level can be leveraged to incorporate **CER** and predictive-analytic information sources (such as clinical records and immunological profiles) alongside image data proper (such as feature vectors calculated via Computer Vision algorithms).

A further level of integration between **CaPTk** and **CER** initiatives (again, staying with **CaPTk** as a representative example of bioimaging applications in general) can be achieved if one observes that clinical outcomes may be part of the analytic parameters used by **CaPTk** modules. As presently constituted, **CaPTk** analytic tools are focused on extracting quantitative (or quantifiable) features from image themselves, without considering additional patient-centered context. There is no technical limitation, however, which would prevent the **CaPTk** system from sharing more detailed clinical information with its modules, allowing these analytic components to cross-reference image features with clinical or patient information. This then raises general questions about sharing clinical data *as well as* information derived from bioimage analysis, bringing us to more general themes in lab/clinical data-sharing.

2.3 Multi-Application Networks in the Context of Scientific Research Data

Architecturally, the pattern of organization just described — semi-autonomous applications linked together (often by virtue of being common extensions to an overarching "core" software platform) — is analogous to the collection of software components that may share access to a data repository or a research-data corpus, include a corpus of medical/diagnostic images. The purpose of research data archives — particularly when they embrace contemporary open-access standards such as **FAIR** (Findable, Accessible, Interoperable, Reusable) [36] and the Research Object Protocol[6] — is to promote reuse and reproduction of published data and findings, such that multiple subsequent research projects could be based on data originally published to accompany one book or article. As a result, it is expected that numerous projects may overlap in their use of a common underlying data set, which potentially means a diversity of software components implementing a diversity of analytic techniques, each offering a unique perspective on the underlying data.

Implementing a robust research-data software framework involves integrating multiple scientific applications, but also (ideally) extending these applications with features specifically of interest to those conducting or reviewing research using published data sets and/or described in academic literature: for instance, capabilities to download data sets from open-access scientific portals; to parse microcitation formats; and to interoperate with document viewers. This review of data-publishing technology is relevant to radiology and to Patient-Centered Outcomes because it typifies the emerging ecosystem where scientific research and open-access data is being disseminated. The architecture employed by **CaPTk** is a useful example of how multiple autonomous, stand-alone, native applications can be federated into a decentralized but unified platform, logistically embodying the kinds of application networks appropriate for the technology supporting archives of research data (including diagnostic-imaging repositories).

Initiatives such as Research Objects and **FAIR** advocate for a technological infrastructure characterized by a diverse software ecosystem paired with open-access research data sets [3]. Although formats such as Research Objects have been standardized over the last decade, there has not been a comparable level of attention given to formalizing how multiple software applications should interoperate when manipulating overlapping data. The Common Workflow Language (**CWL**), which has been explicitly included in the Research Object model, documents one layer of inter-application messaging, including the encoding of parameters via command-line arguments (as mentioned earlier, **CaPTk** provides the most complete **C++** implementation of **CWL**, using it to pass initial data between modules). Serializing larger-scale data structures is of course a generic task of canonical encoding formats such as **JSON**, **XML**, **RDF**, and Protocol Buffers — not to mention text or binary resources serialized directly from runtime objects via, for instance, **QTextStream** and **QDataStream.** This means that some level of inter-application communications is enabled via **CWL**, and that essentially any computationally tractable data structure can be encoded via formats such as **XML**. These solutions, however, are sub-optimal: **XML** (as well as **JSON** and analogous formats) is limited because it takes additional development effort to compose the code that marshals data between runtime and serial formats. Similarly, although **CWL** can model information passed between applications, it provides only an indirect guide for programmers implementing each application's "operational semantics" — viz., the procedures which must be executed before and after the event wherein data is actually passed between endpoints.

In the context of **CaPTk**, for example, integrating peer modules with the **CaPTk** core application involves more than simply ensuring that these endpoints communicate via a standardized data-serialization format: the plugins must be *registered* with the core application, which affects the core in several areas, including the build/compile process and construction of the main **GUI** window. Modeling the interconnections between semi-autonomous modules, as **CaPTk** demonstrates, therefore requires more detail than simply

[6] see http://www.researchobject.org/scopes/

modeling their shared data encodings; it is furthermore necessary to represent all procedural and User Interface requirements in each component that may be affected by the others. Despite the standardization efforts that have been invested in Research Objects and the Common Workflow Language, we contend that this fully detailed protocol for multi-application interop has not yet been rigorously formalized.

Rigorous models of application networks among semi-autonomous components acquire an extra level of complexity precisely because of this intermediate status: protocol definitions have to specify both the functional interdependence and the operational autonomy of different parts of the application network. Although one application does not need detailed knowledge of the other's internal procedure signatures (which would break encapsulation), the functional interdependence between applications can accordingly be modeled by defining protocols which must be satisfied by procedure-sets internal to each end-point — the relevant information from an integrative standpoint is not the actual procedures involved, but confirmation that the relevant procedure sets adhere to the relevant multi-procedural protocol.[7]

While we have initially approached multi-application networking from the bioimaging perspective, this topic is equally applicable to multi-site trial architecture, the theme of next section.

3 Precision Medicine in Trial Design

Converting basic research to clinical practice directly benefiting patients — sometimes called *translational informatics*, a "research cycle, which involves the translation of knowledge and evidence [to] provision of evidence-based care in the clinical or public health settings" [26, page 2] — is sometimes represented as a two-stage process which first involves translating research to clinical trials, and then formulating point-of-care practices on the basis of trial results (*ibid.*). Data-sharing initiatives need to pay particular attention to the logistics of translational informatics in contexts where granular patient-specific information is important, such as immunoprofiling. Questions which should be addressed include (1) where data is to be hosted; (2) how participating institutions should submit data to a central repository; (3) how participating institutions and/or outside investigators should access previously-deposited data; (4) how to ensure anonymization of patient-specific records; (5) how to ensure that different labs used by different hospitals are utilizing compatible protocols, so that results from multiple labs/hospitals can be seamlessly merged in a shared data commons; (6) how to ensure proper alignment between software employed at different institutions; and (7) how to incorporate data curated within the context of a multi-institutional data-sharing initiative into scientific papers documenting research findings. Each of these areas of concern are technically demanding because of the complex and heterogeneous nature of immunological profiles.

As a case-study in clinical-trial software engineering, consider again the proposals in Shrestha *et al.* [35] which we reviewed last chapter. As these authors recommend, Covid-19 trials should be designed to focus on specific patient groups which are more likely to benefit from the interventions that form the basis of the relevant clinical trials. Moreover, toward the goal of applying precision medicine to Covid-19 clinical practice, it should be possible to construct a quantifying domain of patient-profile signals (antecedent to trial commencement) to quantify the statistical probability that a given treatment will have a favorable patient outcome in relation to all the prior data in a patient's profile. Since researchers assume that certain factors in a patient's profile will be statistically correlated with favorable outcomes in conjunction with specific treatment plans, part of the trial's purpose is to determine which parts of the patient profile are, in fact, statistically relevant.

In practical terms, then, setting up Covid-19 trials would involve defining patient-selection criteria and implementing systems to screen for patients who may be good candidates for different trials. This would require two steps: (1) constructing a format where trial criteria can be rigorously notated; and (2) implementing software at participating sites to search for trial candidates. This software would need to access, represent, and analyze patient-specific trial-eligibility factors that covering a broad spectrum of data types (sociodemographics, medical history, lab/image results, etc.). Proposing an automated recruitment engine for cancer trials, [29] argued (in 2014) that

> Many ... tools have been developed for accessing the institutional data warehouse to screen patients for clinical trials or for creating an alert system for physicians ... However, ... existing systems have limited access to the complete patient information, such as the latest laboratory test results, and are not integrated with the clinical systems used in routine patient care. Further, existing tools have limited support for structured entry of trial information, interactive user interfaces (UIs) that allow clinicians to review the matching results and re-execute the matching process with updates to patient records. [Although] there has been extensive research in creating formal, computable representation of eligibility trial specifications that can be used together with electronic representation of patient data in EHR systems [an] important challenge for computational representation of trial information is the lack of suitable interfaces for entering eligibility criteria. (page 2)

Ten years earlier, [20] advocated for trial models "focused on interoperability, using modern object-oriented techniques" (page 3):

> Object-oriented software ... is more reliable because fewer software code changes are required as one's needs evolve; and, therefore, more changes can be made without effectively changing the core structure of the application ... The difference between object-oriented software and traditional software is substantial. (page 6)

However, he concludes, "Very few CTMS products are object-oriented." Both [29] and [20], we should note, are not neutral observers, but writing to highlight features of software their own teams designed. Whatever the merits of their systems, these design

[7] Reviewing the source code and documentation for CAPTK confirms that multi-application messaging along these lines is implicitly adopted by CAPTK; *see* for example https://www.med.upenn.edu/cbica/assets/user-content/images/captk/2018_ISBI_CaPTk.0404.Part2.pdf, particularly the material starting on the 30th slide of that presentation.

principles have not apparently been incorporated into mainstream **CTMS** in intervening years.

Given the sheer scale of the SARS-CoV-2 pandemic, there are likely to be many candidates for almost any Covid-19 trial. However, methods for recruiting patients would need to be aligned across multiple institutions, at least in the case of multi-site trials. For example, in the context of antigen tests (measuring virus antibody levels) the US Centers for Disease Control recommends or has authorized a large list of assays performed by many different companies, using many different biochemical methods.[8] Because the data format resulting from immunoassays depends on the specific biochemical mechanism which (within each assay) yields quantitative data, a broad spectrum of antigen tests requires a diverse array of data formats which need to be integrated. As such, whenever Covid-19 immunoassays are considered as factors in immunological profiles for mapping patients to appropriate Covid-19 clinical trials, querying for good trial candidates means querying across a wide spectrum of structurally different data types that correspond to this broad array of antigen tests — specifically to the mechanisms through which laboratory instruments generate quantifiable data and to the computational procedures which process such data. Here is an example of why specialized trial software can be warranted: the heterogeneity of trial data, may call for integrative procedures implemented directly in programming languages such as **C++** (rather than query languages such as **SQL**). The complexity of trial criteria was a motive for the software-based **CTMS** systems we cited earlier; Covid-19 serves as a trenchant case-study supporting recommendations along those lines (*see also*, say, [17], [4], [33], [6], [38, especially pages 20-29]).

3.1 Customizing Clinical Trial Management Software

Once trials embrace heterogeneous data models (which require special-purpose software for accessing some of the trial data), Clinical Trial Management Systems (**CTMS**) requirements become more complex. In these situations, **CTMS**s may need to model and in some cases replicate complex computational workflows, such as those employed by Dearlove *et al.* for calculating SARS-CoV-2 genomic sequences from patients' blood samples. The **CTMS** software may also need to interoperate with domain-specific applications, as in bioimaging and image analyses, signal processing (e.g., for **EKG** analysis), Flow Cytometry, biochemical assays, genomic analysis, epidemiological modeling and so forth. If possible, such applications should be configured or extended to work with the clinical trial software. For instance, if a **DICOM** (Digital Imaging and Communications in Medicine) client is used to study an image derived from a specific trial — e.g., a radiological scan of a Covid-19 patient's lungs — the **DICOM** software could be provided with a plugin that would show trial information in a separate window, which could then be juxtaposed with the main-image view. In this context, software alignment means that all institutions participating in a trial could use the *same* plugins, so that the trial's central **CTMS** system could interoperate with special-purpose software in a consistent manner. This would also aid in establishing, as part of the trial design, protocols for depositing special-purpose data assets (such as **FCS** or **DICOM** files) alongside clinical data and **eCRF**s.

A further benefit of **CTMS** customization is that custom software adds flexibility for trial design. By definition, trials allow researchers to test biomedical hypothesis in a controlled manner. Trials are, therefore, defined around the premise that observational information resulting from the trial is empirically significant, revealing something new about what the trial was designed to investigate. For instance, a Covid-19 trial might assess how well patients with varying prior immunological profiles respond to monoclonal antibody (**mAB**) treatments. The relevant observations in this case derive from the subsequent course of the disease for each patient, as well as potential adverse reactions, but there are inherent complicating factors: were patients receiving other treatments as well? For patients who recover, how do we know that the antibodies expedited that recovery? How quickly was the recovery? And, did patients continue to suffer from Covid-related symptoms even when they were no longer infectious? Situational details specific to the trial — such as each patient's **mAB** dosage level, prior Covid-19 risk factors, or viral-load change over time — also belong in the trial's unique data models. Moreover, a comprehensive investigation could well incorporate both information about the patient's unique immunological profiles and the nature of the SARS-CoV-2 variant/strain found in the patient.

In addition, customizing trial-management software has the added potential benefit of greater flexibility for incorporating personalized/patient-centered data into the overall trial results. Since patients' reactions to interventions are difficult to anticipate in full specificity ahead of time (especially when subjective experience is taken into account) allowing data models to evolve during the course of a trial can help trial designs respect the experiential dimensions intrinsic to patient-centered paradigms of care.

3.1.1 Toward Fine-Grained Sociodemographic Models

Patient-centric data could likewise include sociodemographic information about the patient, supplemented by epidemiological metrics, such as contact tracing. Designers need to identify, for example, what dimensions of patients' immunological and sociodemographic profiles are likely to be consequential when analyzing treatment outcomes (indeed, biomedical research has been criticized in recent years for bias toward certain populations, e.g., white, middle-class non-seniors). This results in uncertainties as to how well trial results carry over to populations at large — that is, populations characterized by a heterogeneous mix of demographic factors. Researchers can mitigate these concerns by demonstrating sociodemographic diversity among trial participants. Those goals, along with more precise predictive analytics, could be advanced by adopting more detailed sociodemographic reporting standards [18], [16], [21], etc.

Demonstrating sociodemographic diversity, however, calls for transparency about how sociodemographic details are represented. The process of grouping patients into ethnic/racial and/or socioeconomic strata can be equivocal at times. For example, if a trial participant is a graduate student at the University of Chicago, should their socioeconomic status be assessed on the basis of their own income or that of their parents? If their zip code places them on the

[8] *See*, for example, https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/vitro-diagnostics-euas#individual-antigen.

South Side of Chicago, a region with both a prestigious campus and pockets of extreme poverty, should they be demographically classified alongside residents of that neighborhood? What about a varsity linebacker who appears to be in excellent health before a Covid-19 infection? Should his status as an athlete be taken to indicate being extremely fit prior to the disease, or might his background as a football player intimate a potential history of brain trauma which may compound his neurological damage due to Covid-19?

In short, sociodemographic data can be notoriously imprecise. It is well-known that patients of lower socioeconomic status have higher Covid-19 infection rates and mortality rates than patients of higher socioeconomic status, but this disparity may be explained by several causal factors: more infectious workplaces, inferior post-infection treatment, poorer state of health before the onset of Covid, and so forth. Teasing apart these factors demands fine-grained analysis of the individual patient's pre-Covid history; sociodemographic generalizations can exclude important details. For example, for purposes of analysis, in the case of a graduate student with middle-class parents but no health insurance of their own, how should we quantify their degree of access to health care? How well does geographic location serve as a proxy for socioeconomic status?

The case of a middle-class student living in a well-off near-campus corner of an otherwise impoverished neighborhood suggests that geospatial metrics (such as zip code or congressional district) are imperfect proxies for wealth; but other factors — such as air/water pollution or the risk of being the victim of a crime — may be statistically correlated among geographically proximate residents even if they have otherwise divergent sociological profiles. These examples illustrate that the value of sociodemographic data is proportionate to the level of statistical detail with which the data may be analyzed. Instead of a broad and vague designation, such as "low income," one may want to derive more detailed subclasses, incorporating information about patients' employment, access to health care, physical fitness, and so forth. Two patients with similar income levels, for instance, may have different levels of access to health care (depending on factors such as whether the patients have employer-based insurance or geographic proximity to healthcare facilities) or different levels of exposure to Covid-19 in the workplace, depending on the nature of their job. Even if a trial does not quantify granular sociodemographic assessments — such as patients' workplace conditions and access to health care, which might serve as a more accurate estimation of how socioeconomic status causally affects disease outcomes — subsequent researchers may determine ways to analyze or add on to data generated by a trial (e.g., through follow-up studies of enrolled patients), so as to make such sociodemographic granularity part of the quantifiable framework.

3.1.2 Measuring Cognitive and Neurological Effects

Similar interpretive issues of interpretation also apply to post-treatment observations. How should researchers decide which observations qualify as clinically significant consequences of Covid-19? We have seen that as the pandemic has unfolded, a fair number of cases have been cited in the professional literature describing Covid-19 patients who suffer certain cognitive/neurological effects, such as muscular fatigue or weakness, mental confusion or poor concentration (sometimes referred to as "brain fog"), or symptoms of Guillain–Barré syndrome spectrum. Almost certainly, some of these symptoms may be the product of cognitive/neurological effects due to SARS-CoV-2. At the same time, Covid-19 patients — even those who fight off the infection successfully or who test positive but remain asymptomatic — may find their lives so disrupted by the pandemic that this may (indirectly) cause cognitive and neurological problems. For instance, prolonged inactivity (for a typically active person), which commonly occurs as the result of a quarantine, may contribute to poor concentration and other diminished forms of mental acuity [11, say]. Given the fact that most people's lives during the pandemic are not "normal," it may be difficult to establish which symptoms experienced by a patient are actually biological effects of the disease itself or, alternatively, indirect consequences of lifestyle restrictions. This sort of ambiguity also applies to potential adverse side-effects of Covid-19 treatment. Rigorous design practice suggests, for instance, that parameters should be established framing the post-treatment and post-recovery window of time where patients' symptoms might be noted as potential effects of the disease or of the administered therapies themselves. How long after a treatment is administered should a patient's symptoms be considered potential side-effects of the therapy itself or, alternatively, the result of nagging uncertainties and a disrupted lifestyle imposed by the pandemic?

The fact that these questions have no predetermined answers indicates that trial designs need to anticipate a shifting clinical and information landscape ad the trial evolves. For example, because SARS-CoV-2 was initially believed to affect lung functioning primarily, the risk of long-term cognitive/neurological damage was not widely anticipated when considering treatment over the course of Covid-19 infection. Consequently, because tests of a cognitive/neurological/physiological/radiographic (except for basic lung scans, with respect to radiology) had not been a common facet of early Covid-19 trials/observational studies, there were no corresponding data structures included in those studies for capturing this full spectrum of neuropsychological and neurological data.

Systematically tracking Covid-19's cognitive/neurological deficits entails neurological, laboratory, neuropsychological and radiographic tests to understand the full extent of their cognitive and neurological impairments. For example, the use of **MRI**s when considering Covid-related ischemic stroke [23], [13], the use of electrophysiological tests, cerebrospinal fluid tests (**CSF**), or the **MRC** (Medical Research Council) Scale for muscle-strength test when considering Covid-related Guillain-Barré symptoms [27], or the use of neuropsychological tests, such as Trail Making Test (**TMT**), Sign Coding Test (**SCT**), Continuous Performance Test (**CPT**), and Digital Span Test (**DST**) — which measure a patient's executive abilities (letter and number recognition mental flexibility, visual scanning, and motor function) and sustained and selective attention, along with other cognitive and neurological functioning — when considering cognitive/neuropsychological impairments after a serious bout with Covid-19 [41].

These cognitive, neurological, physiological, laboratory, and radiographic data structures thereby become an integral part of the

information relevant to trial evaluation, because they document symptoms which are presumptively attributable to Covid-19. However, in practice, prior to clinicians having been alerted to the fact that Covid-19 may cause lingering cognitive/neurological damage, Covid-19 trials were not designed to incorporate neurological or cognitive data in a systematic manner. This scenario points to how trial data models can benefit from a built-in capacity to be redesigned while the trial is ongoing, so as to accommodate new and emerging information. On this basis it is reasonable to adopt for clinical trials malleable information models such as graph databases and/or Object-Oriented software components.

3.1.3 Aggregating Trial Data via Graph Models

A useful data-integration case-study is the University of Pennsylvania's Carnival project (which achieves data integration by adopting property-graph databases, illustrating the flexibility of graph models in a way that we can also apply to trial design). Carnival synthesizes heterogeneous biomedical data by translating information from disparate sources into a common property-graph representation and then querying this data with the Gremlin Virtual Machine. Gremlin is a "step-based" virtual machine where "steps" between potential focus elements in a property graph play the role of primitive processing instructions; querying and traversing property graphs involves executing a series of Gremlin steps. Most Gremlin implementations are based on the Java programming language and the Java Virtual Machine (**JVM**), so that queries themselves are written in a **JVM** language (Groovy, in the case of Carnival). The challenge for any database engine which employs a relatively complex data-representation strategy — such as a hypergraph, property-graph, tuple-store (a collection of records with varying numbers of fields) or a multi-dimensional (possibly sparse) array — is to efficiently map the high-level data structures manipulated within the database itself to the lower-level memory units which are stored to disk. Lower-level data structures are typically modeled via simpler database constructions such as key-value stores, memory cells, or relational tables, so there must be a translation pipeline between high-level structures (properties, hypernodes, hyperedges, and so forth) and low-level points (record cells, shared memory address, elements in key-value pairs, etc.)

Consider how patient profiles usually notate the medications each patient is taking — any patient (a node) could in principle be connected to any medication (another node); some patients may be taking *no* medications, others may take just *one*, and some may take *two or more*. Also, connections between patients and medications can be the basis of further details that emerge over time (and are registered in the graph) inasmuch as medications are prescribed to patients by a specific doctor at a specific time, in a specific dosage, in response to specific diagnostic tests, and so forth. In short, information can "fan out" from the patient-to-medication connection in a relatively free-form manner. In general, then, as a subset of overall patient profiles, information about medication evinces the structural features which are, in many contexts, optimally represented via free-form labeled graphs. Meanwhile, patient profiles may also consider medical history, which can be modeled as a graph with detailed logical and temporal inter-node connections. According to this representational strategy, each patient's history is a series of events and observations which are temporally ordered — it is possible to query or traverse the graph in a manner which takes before/after relations into account — and where there are also logical or causal connections defined between nodes. For instance, an edge might assert that a given medication was prescribed to a patient (an event) *because of* the results of a given lab test (an observation). In these examples, different sorts of clinical data — sociodemographic, pharmacological, medical-history — are modeled according to different sorts of graph structures (hypernodes, nonschematic labeled edges, temporalized graphs, and so forth).

Insofar as different formations within data-structures tend to be conveyed via different graph-database constructions, the partition of graph-database technology itself into competing (partially incompatible) systems (distinct architectures, query languages, query-evaluation strategies, and so on) can also be a hindrance to data-integration, even after adopting flexible graph-database technologies. This is one motivation for the hybrid graph models we discuss in later chapters, which attempt to recognize numerous structuring elements endemic to different flavors of graph systems, accommodated into a single hypergraph-based query framework.

3.2 Representing Trial Data via Object Models

As a concrete example of open-ended data-integration concerns being anticipated as a central element of trial design, consider how Object-Oriented models for Covid Phylogeny (which we briefly considered last chapter) could serve as a nexus for integrating SARS-CoV-2 phylogenetic data across multiple studies and healthcare systems. Such an Object Model may be extended in different ways for different clinical trials examining a range of Covid-19 treatments.

In the context of antibody regimens, for instance, scientists need to quantify how well the antibodies disable Covid-19 spike proteins directly and/or how well the antibodies block the virus's ability to attach to human cells. These measurements generate data which gauge a patient's immune response to Covid-19, whether innate ("naive") or boosted by the administered antibodies. Researchers have mined such data from different angles, including contrasting symptoms presented with different levels of severity [22], [42], biochemically describing the mechanisms of immune response so as to augment or simulate that response via monoclonal antibodies or similar antiviral therapies [40], [7], identifying factors explaining act-risk groups' greater susceptibility to severe cases [25], or comparisons of SARS-CoV-2 against other coronaviruses [1] (these citations are representative examples of work conducted by many research groups; they could be expanded with similar references we included in Chapter 2 when discussing immune repertoires, for example). A Covid-19 software ecosystem should then ensure that such immune-response data can be effectively parsed and integrated into Object Models describing how SARS-CoV-2 is evolving around the globe, so researchers (as much as logistically feasible) could track each patients' immune response in light of their particular acquired SARS-CoV-2 strain/variant and their personal immunological profile. The relevant information for geospatial and sociodemographically tagged studies would then include patients' prior immuno-profiling and risk assessment, immune response (naive or

affected by treatment) and clinical outcomes, as well as genetic tests for the viral variant. Cross-sectionally mapping such data geographically and along sociodemographic/socioeconomic contours could provide a holistic picture of the pandemic at a global scale.

Object Models intended to facilitate such globally-scoped data integration could then facilitate trials designed according to the protocol proposed by Shrestha *et al.*, discussed above, where each trial would study a preselected (non-random) cohort of patients for whom both pre-treatment immunoprofiling data and post-treatment outcomes data would be available, so as to compare multiple trials against one another. Object Models customized for each trial would generate a data framework through which the causal relations between patient profiles and treatment outcomes would be investigated. Customized trial software would, accordingly, provide a Reference Implementation demonstrating each trial-specific Object Model. If adopted for multiple trials conducted across multiple clinics/hospitals, trials' data models may help doctors better understand which aspects of patient profiles are particularly significant when matching Covid-19 patients to the most salutary treatment.

4 Text and Data Mining via CORD-19

The third methodological area we will concentrate on in this chapter is text and data mining, using analyses of large-scale document corpora and/or biomedical data sets to discover connections between research work which might be less evident to individuals reading publications in isolation. Text Mining and Natural Language Processing (**NLP**) is certainly one methodology which has been explored for personalized/precision medicine, on the theory that automated searches across biomedical publication archives may detect diagnostically or prognostically significant connections between individual patients' profiles/symptoms and prior studies which doctors might not discover otherwise. In general, text-mining has been developed as one branch of data-mining and Machine Learning in general applied to document corpora under the aegis of precision-medicine (*see* for instance [5], [32], [31]).

Biomedical text mining is also a good case-study in data-integration workflows, because typically the text-mining process requires synthesizing multiple **NLP** workflows and reading data from multiple input sources — [15], for example, is a good precision-medicine context example — including sentence-parses, datasets of biomedical nomenclature, domain-specific knowledge bases (for gene-sequences, cancer variants, genomic-proteomic or genomic-antigen assocations, and so forth), manual text annotations, etc.

Despite the perceived potential of patient-centered text mining, some scientists caution against overestimating the power of automated **NLP** platforms (*see* [14], for instance). These critiques are not rejecting text-mining in general, but rather observing limitations in existing document-encoding formats, which are derived from publishing technologies whose primary targets are human readers rather than machine-automated text processing. More systematic text representation and document annotation could alleviate the need for probabilistic **NLP** reasoning engines, making text-mining operations more precise and reliable.

As an example of the possibilities and challenges of text and data mining scenarios we will consider **CORD-19**, a collection of Covid-19-related research articles which was developed (starting in Spring 2020) in conjunction with a White House "call to action" to spur Covid-19 research. This White House initiative was described as a "call to action ... to develop new text and data mining techniques that can help the science community answer high-priority scientific questions related to COVID".[9] As raw data for this initiative, the US government helped spearhead a consortium of industry and academic institutions, headed by the Allen Institute for AI Research, who curated a "machine-readable Coronavirus literature collection" which includes article metadata and (in most cases) publication text for over 280,000 coronavirus research papers (as of mid-2021) [9], [39]. This corpus is paired with links to publisher portals (including Springer Nature, Wiley, Elsevier, the American Society for Microbiology, and the New England Journal of Medicine) providing full open access to Covid-19-related literature; these resources collectively constitute **CORD-19** (the "Covid-19 Open Research Dataset").

4.1 Overview of CORD-19

The **CORD-19** collection was formulated with the explicit goal of promoting both text mining and data mining solutions to advance coronavirus research. This means that **CORD-19** is intended to be used both as a document archive for text mining and as a repository for finding and obtaining coronavirus data for subsequent research. The White House announcement directly requests institutions to develop *additional* technologies which would help scientists and jurisdictions to take advantage of **CORD-19** as it was initially published. In short, **CORD-19** was released with the explicit anticipation that industry and academia would augment the underlying data by layering on additional software.

Despite the obvious benefit to researchers, the health-care community, and the public at large in publishers choosing to release a substantial quantity of Covid-19 related literature in Open-Access fashion, **CORD-19** is not without certain limitations, as acknowledged by the curators themselves:

> We have performed some data cleaning that is sufficient to fuel most text mining & NLP research efforts. But we do not intend to provide sufficient cleaning for this data to be usable for directly consuming (reading) papers about COVID-19 or coronaviruses. There will always be some amount of error, which will make CORD-19 more/less usable for certain applications than others. We leave it up to the user to make this determination ...[10]

Some problems stem from how the articles are encoded into an ostensibly (**JSON**-based) machine-readable format, while others are unavoidable limitations of **NLP** overall.

To be clear, the concerns we identify here reflect an informal survey of **CORD-19**; they are not opinions provided by researchers working directly with **CORD-19** or analyses discussed in peer-

[9] See https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset.
[10] *See* https://github.com/allenai/cord19

reviewed literature. With that caveat, however, We assert that certain issues deserve mention:

Transription Errors Transcription errors can cause the machine-readable text archive to misrepresent the structure and content of documents, hindering text-mining technology that targets the archive. In **CORD-19**, for instance, there are cases of scientific notation and terminology being improperly encoded. As a concrete example, "2'-C-ethynyl" is encoded in **CORD-19** as "2 0 -C-ethynyl", which could stymie text searches against the **CORD-19** corpus (see [12] for the human-readable publication where this error is observed; the corresponding index in the corpus is 9555f44156bc5f2c6ac191dda2fb651501a7bd7b.json).

Poorly Indexed Research Data Although **CORD-19** provides a structured representation of a large collection of research *papers*, there is no easy-to-use tool for finding research *data* through **CORD-19**.

Poorly Integrated Research Data The research data which *can* be accessed through **CORD-19** evinces a wide variety of technical fields and formats, with distinct software requirements; as a result, it is a difficult task to merge and integrate different data sets related to Covid-19. At present, **CORD-19** does not include any software tools or computer code that would facilitate data integration.

Disconnect Between Text Data and Publisher Portals Although most of the **CORD-19** manuscripts represent peer-reviewed literature which can be accessed through document portals (for instance, the National Center for Biotechnology Information website), the **CORD-19** archival schema does not represent these links (except indirectly via Document Object Identifiers). As such, there is no easy way for researchers to find and read publications which have been flagged by text-mining algorithms as being potentially of interest to them. Furthermore, there is no direct mechanism to enlarge the **CORD-19** corpus with papers newly added to publisher portals.

To clarify the final comment: the Allen Institute for AI, which curated **CORD-19**, encourages publishers to contribute new (or newly-available) articles to the corpus. However, integration with **CORD-19** (or, so it seems, and other domain/topic-specific portal) is not developed as a formal step in the publication workflow. In particular, publishers are not themselves generating machine-readable document infosets that can be integrated with the **CORD-19** schema (which, in turn, causes transcriptions errors and other problems as just outlined).

With respect to text mining, an immediate problem arises in **CORD-19**'s archive-construction methodology: especially, how the text was parsed from **PDF** files. This is a process which almost inevitably causes imprecise or inaccurate text representation, which can degrading the quality of the archive unless manual or automated corrections are made. In particular, the **CORD-19** library evinces transcription errors, as mentioned above (especially in relation to technical or scientific phrases and terminology); scientific notation in particular may be improperly encoded. Moreover, there is no semantic marking identifying that (say) the "2 0 -C-ethynyl" text segment has a specific technical meaning. These errors or limitations arise in part from unavoidable anomalies which occur when reading texts from **PDF** files rather than from machine-readable, structured formats such as **XML**.

It is also worth observing that the **JSON** format used for encoding **CORD-19** manuscripts presents some logistical challenges for any operations related to text-mining or to cross-referencing publications and data sets. In particular, **CORD-19** makes partial use of "standoff annotation"; specifically, document features such as citations and references are notated through character offsets into the paragraph where they appear. As a result, accurately reading these document elements requires synthesizing data points parsed from several distinct objects in the **JSON** code, which is only feasible given a client library built to interface with the **CORD-19** files in accord with their specific schema. Such a client library would implement convenience procedures to handle recurring tasks, such as obtaining the full bibliographic reference affixed to a given location in a manuscript.

With respect to *data* mining in the **CORD-19** context, the limitations in the currently available raw **CORD-19** data are even more pronounced than in the context of text mining. In particular, neither the article metadata nor the full open-access document collections have any mechanism for actually obtaining data published alongside research papers, or even identifying which papers have accompanying data in the first place. The Springer Nature collection which was originally one component within **CORD-19** illustrates the limitations of this relatively unstructured data-publishing approach (this following analysis will focus on Springer Nature, but the problems identified are no less pronounced on the other **CORD-19** portals — if anything, because Springer Nature allows readers to browse articles in **HTML** within the web portal directly, one can ascertain whether research data exists for an article without downloading and reading it; with other **CORD-19** resources it is actually harder to locate supplemental data when available). Initially, the Springer Nature portal encompassed 43 articles, of which 15 were accompanied by research data that could be separately downloaded (this number does not include papers that document research findings only indirectly, via tables or graphics printed inline with the text). Collectively these articles referenced over 30 distinct data sets (some papers were linked to multiple data sets), forming a data collection which could be a valuable resource for Covid-19 research — not only through the raw data made available but as a kernel around which new coronavirus data could accumulate. However, there is currently no mechanism to make this overall collection available as a single resource.[11]

This problem demonstrates, among other things, how document-metadata formats such as the Research Object protocol are limited in applying only to *single* articles. As a result, there is no commensurate protocol for publishing *groups* of articles which are tied to groups of data sets unified into an integral whole. Open-access Covid-19 papers also reveal limitations of exiting online document portals, especially with respect to how publications are linked to data sets. In particular, there is no clear indication that a given paper

[11] As **CORD-19** has evolved, the publisher-specific sections therein appear to be merged into portals such as Springer Nature directly, so our above comments based on isolating Springer Nature articles are probably more applicable to the original archive design than the current technology. However, insofar as the current portal simply defers to publisher-specific search features, we would argue that accessing Covid-19 data sets through **CORD-19** is if anything more difficult than before.

is associated with downloadable data; usually readers ascertain this information only by reading or scrolling down to a "supplemental materials" or "data availability" addendum near the end of the article. Moreover, because the Springer Nature portal (and similar publisher resources) aggregates papers from multiple sources, there is no consistent pattern for locating data sets; each journal or publisher has their own mechanism for alerting readers to the existence of open-access data and citing where they could be downloaded.

4.2 Data Integration within CORD-19

Aside from the issues which are likely to hinder text and data mining across **CORD-19**, the collective group of Covid-19 data sets also illustrates the limitations of information spaces pieced together from disconnected raw data files with little additional curation. The files included in this group of data sets encompass a wide array of file types and formats, including **FASTA** (which stands for Fast-All, a genomics format), **SRA** (Sequence Read Archive, for **DNA** sequencing), **PDB** (Protein Data Bank, representing the **3D** geometry of protein molecules), **MAP** (Electron Microscopy Map), **EPS** (Embedded Postscript), and **CSV** (comma-separated values). There are also tables represented in Microsoft Word or Excel formats. Although these various formats are reasonable for storing raw data, not all of them are actually machine-readable; in particular, the **EPS**, Word, and Excel files need manual processing in order to use the information they provide in a computational manner. A properly curated data collection would need to unify disparate sources into a common machine-readable representation (such as **XML**).

Going further, productive data curation should also aspire to *semantic* integration, unifying disparate sources into a common data model. For example, multiple spreadsheets among the Springer Nature Covid-19 data sets hold sociodemographic and epidemiological information relevant to modeling the spread of the disease. These different sources could certainly be integrated into a canonical social-epidemiology-based representational paradigm which recognizes the disparate data points which might be relevant for tracking the spread of Covid-19 (with the potential to unify data from many countries and jurisdictions).

This is not only an issue of data *representation* (viz., how data is physically laid out), but also of data types and computer code. According to the Research Object protocol, data sets should include a code base which provides convenient computational access to the published data. In the case of Covid-19, this entails creating a sociodemographic and epidemiological code library optimized for Covid-19 information, which would be the primary access point for researchers seeking to use the data which has been published in conjunction with the 43 manuscripts examined here that were aggregated on Springer Nature, along with any other coronavirus research which comes online. Similar comments apply not only to tabular data represented in spreadsheet or **CSV** form, but to more complex molecular or microscopy data that needs specialized scientific software to be properly visualized.

Considering the overall space of Covid-19 data, it is unavoidable that some files require special applications and cannot be directly merged with the overall collection. For instance, there is no coherent semantics for unifying Protein Data Bank files with sociodemographics and epidemiology; files of the former type have specific scientific uses and can only be understood by special-purpose software. Nevertheless, a downloadable Covid-19 archive can include source code for code libraries reading special formats such as **PDF** or **FCS** (as a case-study, this book's supplemental materials provides a build-environment for tools working with those file types, among others).

Earlier we advocated for Object-Orient models of Covid-19 variants will could be integrated with both chemical/molecular data and sociodemographic/epidemiological data. Different Covid-19 strains would then form a bridge linking these different sorts of information; researchers should be able to pass back and forth from molecular or genomic visualizations of Covid-19 to social-epidemiological charts and tables based on viral strains. Ideally, capabilities for this sort of interdisciplinary data integration would be provided by a Covid-19 archive as enhancements to applications that scientists would use to study the published data.

It is worth noting that a data-mining platform requires *machine-readable* open-access research data, which is a more stringent requirement than simply publishing data alongside which can be understood by domain-specific software. For example, radiological imaging can be a source of Covid-19 data insofar as patterns of lung scarring, such as "ground-glass opacity", is a leading indicator of the disease. Consequently, diagnostic images of Covid-19 patients are a relevant kind of content for inclusion in a Covid-19 data set (see [30] as a case-study). However, diagnostic images are not in themselves "machine readable." When medical imaging is used in a quantitative context (e.g., applying Machine Learning for diagnostic pathology), it is necessary to perform Image Analysis to convert the raw data (viz., in this case, radiological graphics) into quantitative aggregates (for instance by using image segmentation to demarcate geometric boundaries and then defining diagnostically relevant features, such as opacity, as a scalar field over the segments). In short, even after research data is openly published by article authors, it may be necessary to perform additional analysis on the data for it to be a full-fledged component of a machine-readable information space.[12]

Another concern in developing an integrated **CORD-19** data collection is that, logistically speaking, not all Covid-19 data is practical to reuse as a downloadable package. This is especially true for genomics; several of the aforementioned 43 coronavirus papers included data published via online data banks capable of hosting data sets that are too large for an ordinary computer. In these situations scientists formulate queries or analytic scripts that are sent remotely to the online repositories, so that researchers access the actual published data only indirectly. Nevertheless, access to these data sets can still be curated as part of a Covid-19 package; in particular, computer code can be provided which automates the process of networking with remote genomics archives through the accession numbers and file-formats which those archives recognize.

[12] This does not mean that diagnostic images (or other graphical data) should be excluded from data sets; only that computational reuse of such data will usually involve certain numeric processing, such as image segmentation. Insofar as this subsequent analysis is performed, the resulting data can be added as a supplement to the image data set itself.

As a final point on the topic of integrating disparate **CORD-19** research data, note that an overarching framework for indexing Covid-19 data sets would also facilitate the process of cross-referencing article text and research data. In particular, the annotation system employed for **CORD-19** could profitably be enhanced by a system of *microcitations* that apply to portions of manuscripts *as well as* data sets. In the publishing context, a microcitation is defined as a reference to a partially isolated fragment of a larger document, such as a table or figure illustration, or a sentence or paragraph defining a technical term, or (in mathematics) the statement/proof of a definition, axiom, or theorem. In data publishing, "data citations" are unique references to data sets in their entirety or to smaller parts of data sets. A data microcitation is then a fine-grained reference into a data set: for example, "the precise data records actually used in a study" (as defined by the Federation of Earth Science Information Partners; see [24]), one column in a spreadsheet, or one statistical parameter in a quantitative analysis.

Ideally, the text-mining and data-mining notions of microcitation should be combined into a unified framework. In particular, text-based searches against the **CORD-19** corpus should also try to find matches in the data sets accompanying articles within the corpus. As a concrete example, a concept such as "expiratory flow" appears in **CORD-19** both as a table column in research data and as a medical concept discussed in research papers; a unified microcitation framework should therefore map *expiratory flow* as a keyphrase to both textual locations and data set parameters. Similarly, a concept such as *2′-C-ethynyl* (mentioned earlier in the context of transcription errors) should be identified both as a phrase in article texts and as a molecular component present within compounds whose scientific properties are investigated through **CORD-19** research data, so that a search for this concept can trigger both publication and data-set matches. Implementing this kind of unified search mechanism requires that data sets be *annotated* with techniques similar to those used for marking up Natural Language techniques.

Considering the inter-disciplinary nature of Covid-19 research, it is unavoidable that different scientists will need different sorts of specialized software to analyze the kinds of information relevant to their research. For instance, the computational techniques applicable to diagnosing coronavirus infection are scientifically very different from those used for genomic or epidemiological studies of the disease; pathologists would not in general use the same software as for genomics/bioinformatics, or virology/epidemiology. In short, even while scientists start with a common pool of raw data, they will need to analyze this data through a diverse set of supplemental computational tools, which will vary not only across disciplines but also in terms of the software and laboratory facilities available to different researchers through their institutions.

Next chapter we will argue that biomedical research corpora have some similarities to "data lakes" and similar large-scale, heterogeneous information systems maintained by hospitals and other clinical and/or research institutions. These dynamics arguably extend beyond research *data* to include document archives such as **CORD-19** as well. Such archives (with **CORD-19** as a case in point) would be centralized in the sense of employing a single curation, accession, and data-management protocol, but would branch out into many distinct research areas, corresponding to data being consumed and studied through a wide range of software products and ecosystems. Common functionality for basic data-acquisition capabilities would then need to be shared among a spectrum of software components which are otherwise variegated in terms of the data formats and computational resources they can provide or recognize. Ideally, this mixture of feature alignment and diversification would be anticipated in the design of document corpora and corresponding protocols for accessing data sets associated with included publications, where applicable.

The prior paragraphs have highlighted limitations of data sets published in conjunction with coronavirus articles made available as open-access resources on Springer Nature (and, by extension, **CORD-19**). The central point here is to argue for a distinct data-curation stage in the publication process, with data curators playing a role distinct from that of both authors and editors.[13] Moreover, the discussion has hopefully highlighted problems with current data-sharing paradigms, even those such as the Research Object and **FAIR** initiatives which are explicitly devoted to improving how open-access data sets are published. **CORD-19** exposes several lacunae in the Research Object protocol, for example, which point to the need for a more detailed extension of this protocol. In particular, an enhanced protocol should encompass:

1. A canonical framework for archiving collections of data sets, not only single data sets (and not only groups of data sets published with a single research paper). For example, all data sets published alongside the 43 Springer Nature articles could be unified into a single collection.
2. A code base accompanying data-set collections designed to help research unify the information provided. Curating the overall collection would involve pooling disparate data into common representation, and implementing computer code which deserializes and processes the unified data accordingly. For instance, **CSV**, **EPS**, and Microsoft Word/Excel tables could be migrated to **XML**, **JSON**, or a more complex common format. Customized computer code could then be implemented specifically to parse and merge the information present in single data sets within the overall collection. This implementation would reciprocate the Research Object goal of unifying code and data, but, again, at the level of an aggregate of research projects rather than a single Research Object.
3. A unified data-set collection should be self-contained as much as possible, and should be built around a foundation where advanced computing capabilities are available in a transparent, standalone fashion, without requiring tools outside the collection itself. One way to achieve this is by gravitating toward components that can provide features such as scripting and data persistence through components that can be shared in pure source-code fashion, such as the WhiteDB database engine [28] and the AngelScript scripting language.[14]
4. A unified data-set collection should also provide prototyping and remote-access tools to interface with web-based information spaces that host data sets too large to be individually downloaded. Ideally,

[13] The point here is not to critique the work of individual authors; curating data sets according to exacting scientific standards demands a vein of expertise that typically lies outside researchers' disciplinary scope. The point is rather that publishers should recognize data curation as a distinct process and skill-set complementary to both writing and editing research works.

[14] *See* `http://www.angelcode.com/angelscript`.

these would include simulations of remote services, which would help scientists understand the design of the remote archives and how to interface with them.

5. Finally, a unified research portal could influence the design of web portals where associated texts are published (*see* [8], [2], [34], [19], for instance). Ideally, it would be easy for readers to identify which articles have supplemental data files and to download those files if desired. Moreover, a microcitation framework could be made available for textual links between publication content and data sets — for instance, a plot or diagram illustrating statistical or functional distributions should link to the portion of the data set from which that quantitative data is derived.

This discussion has used the Covid-19 crisis as a lens through which to examine data-publishing limitations in general. Such limitations are not specific to coronavirus in particular. However, the nearly unprecedented urgency of this epidemic reveals how both the scientific and publishing industries are still struggling to develop technologies and practices which keep pace with the intersecting needs of systematic research and public policy. An optimistic projection is that the crisis will spur momentum toward a more sophisticated data-sharing paradigm — perhaps a generalization of the Research Object protocol toward data-set collections.

4.3 Reviewing the CORD-19 Document Model

In order to discuss the possibilities and limitations of **CORD-19** (and potentially other document corpora with a similar design) it is worth examining how **CORD-19** encodes textual data in greater detail. This discussion has ramifications outside of **CORD-19** itself, insofar as **CORD-19** hopefully points to gaps in current publishing technologies. These gaps need to be addressed if publishers are to curate open-access corpora which truly leverage the digital and interactive technology available to us with modern software.

The basis of **CORD-19**'s infrastructure is a **JSON** scheme which describes the document hierarchy of research articles encoded within the corpus. Apart from metadata (consisting of basic details such as document title and authors' names) and bibliographic entries, all document content according to this schema is divided into paragraphs (implicitly the documents are divided into sections as well, but sections are notated as properties of the paragraphs they contain, not as a separate level in the hierarchy). Each paragraph encoding contains an underlying string vector (a stream of characters) and, separate and apart from that, character "spans" which point to references (such as Named Entities), citations, and equations. This indicates that the **CORD-19** encoding uses "standoff annotation," where any content modifying the interpretation assigned to portions of the main text is notated with a series of data structures described apart from the main text itself.

Standoff text-encoding systems may be contrasted with **XML** or **HTML**, where "tags" are mixed with character data. For example, consider a span of text which quotes from another document: in **HTML**, the special status of the quoted text may be marked by surrounding the text with **<quote>** start and end tags. Syntactically, this markup system has the effect that tags and text are seen side-by-side: any content governed by the **<quote>** (i.e., the text of the quote itself) is printed immediately after the begin-tag, and the quotation ends when the last character is followed by an end-tag (i.e., **</quote>**).

Apart from such syntactic details, the distinction between tag-based markup and standoff annotation determines the "semantics" of the document, insofar as tags form a document hierarchy. Continuing the **<quote>** example, the text-span inside the quote tags is represented as a *child element* of the quote, whereas the quote itself may be a child element of a larger-scale entity (such as a paragraph). In effect, the paragraph *contains* a quote, and the quote *contains* a string of characters. Such nested levels of containment provide the structure through which hierarchical documents (formats such as **XML** and **HTML**) are interpreted.

To see the contrast with *standoff* annotation, if one were to describe a document using a standoff annotation system, the notation that a particular span of characters belongs to a quotation would not be marked-up amidst the characters themselves. Instead, the quotation-designation would use numeric indices to declare that the character at a certain position in the main text begins a quotation, and some later character in the text ends that quotation. When serializing documents with standoff annotation, all the characters in a document are typically represented as one character-stream, and any notation describing markup applied to spans within that character stream is asserted afterward, using indexes into the stream to demarcate element boundaries.

The **JSON** schema used for **CORD-19** is not entirely standoff, because there is a document hierarchy (for example, a publication's abstract is modeled as a sibling element to the main body text, so abstracts and the main text represent an intermediate hierarchical level, contained within the overall document and containing individual paragraphs). However, **CORD-19** uses in effect a standoff-annotation system for each paragraph, so there is no hierarchical level smaller than paragraphs themselves, except implicitly; after the text (viz., the character stream) there are subsequent notations of spans within the paragraph (each span description is considered a child of the paragraph itself, as is the paragraph text).

This arrangement has consequences for text mining algorithms, which may be strengths or weaknesses in different contexts. One consequence is that the raw text is all grouped together in one place — algorithms do not have to tie together child nodes of disparate **XML** elements to derive a beginning-to-end sequence of the text belonging to any paragraph. Instead, it is simply necessary to read all data in the "text" field of the relevant "paragraph" object. The character-sequence in this text may contain words and sentences, but potentially other strings of symbols (such as chemical formulae) which are not explicitly marked. This may or may not be desirable. It could potentially complicate **NLP** tasks, because the language-processing components will be fed not only sequences of English words but also, sometimes interspersed among ordinary words, technical symbol-sequences such as **"2'-C-ethynyl"** (an example used earlier in this chapter). Standoff annotations may or may not be effective in marking the boundaries of such extra-lexical sequences; certainly we cannot rely on Named Entity detectors to properly identify and demarcate the boundaries of all uses of technical terminology or special symbols (again, the limitations of

automated annotation are discussed earlier in this chapter).

In discussing standoff annotation it is also worth considering how the text of **CORD-19** publications was obtained. According to **CORD-19** documentation, most full-text transcriptions in the corpus were obtained from **PDF** files, via a pipeline using **TEI** (Text Encoding Initiative) **XML** as an intermediate representation. Necessarily, then the encoded text is only an approximate representation of the original:

> To provide accessible and canonical structured full text, we parse content from PDFs and associated paper documents. The full text is presented in a JSON schema designed to preserve most relevant paper structures such as paragraph breaks, section headers, and inline references and citations. ... We recognize that converting between PDF or XML to JSON is lossy. However, the benefits of a standard structured format, and the ability to reuse and share annotations made on top of that format have been critical to the success of CORD-19. ... Though we have made the structured full text of many scientific papers available to researchers through CORD-19, a number of challenges prevent easy application of NLP and text mining techniques to these papers. First, the primary distribution format of scientific papers — PDF — is not amenable to text processing. The PDF file format is designed to share electronic documents rendered faithfully for reading and printing, not for automated analysis of document content. Paper content (text, images, bibliography) and metadata extracted from PDF are imperfect and require significant cleaning before they can be used for analysis. Second, there is a clear need for more scientific content to be made easily accessible to researchers. Though many publishers have generously made COVID-19 papers available during this time, there are still bottlenecks to information access. ... Lastly, there is no standard format for representing paper metadata. Existing schemas like ... JATS[,] Crossref [or] Dublin Core have been adopted as representations for paper metadata. However, there are issues with these standards; they can be too coarse-grained to capture all necessary paper metadata elements, or lack a strict schema. ... Without solutions to the above problems, NLP on COVID-19 research and scientific research in general will remain difficult. [39, page 6]

As an example of these **NLP** issues, consider the challenge of demarcating all named entities, particularly technical character-sequences (such as chemical formulae) which are not ordinary lexemes. Whether or not authors explicitly mark up such sequences (they may well do so in that formulae or equations are often typeset differently than normal text) this markup is not preserved in **PDF** versions of articles. As the authors of the last-cited article point out, many (roughly 38%) of papers in **CORD-19** are also available in the **JATS** (Journal Article Tag Suite) format, which is a more precise text encoding than **PDF**. However, even in this context **JATS** does not compel authors to explicitly notate textual entities such as special terms or character-sequences — in fact **JATS** does not truly have an obvious structure or set of alternative structures for identifying what would normally be considered annotation-worthy text spans or named entities; the closest correlates are probably the generic **<kwd>** (keyword) and **<abbrev>** (abbreviation) tags as well as discipline-specific options such as **<chem-struct>** (for chemical structures) and **<disp-formula>** (for mathematical expressions). In short, building a corpus such as **CORD-19** for rigorous text-mining is made more difficult because authors and publishers do not publish texts in formats which are optimized for text mining in the first place; the acknowledged limitations of **CORD-19** reflect problems of industry practice, not programming lacunae that could be alleviated with more sophisticated **NLP** algorithms.

Having acknowledged these limitations, a discussion of document corpora could then reasonably pivot from the empirical goal of curating useful text archives from currently published text to examining how more sophisticated corpora may be published in the future. It is reasonable, for example, to propose that full-text publications be released *both* in reader-friendly **PDF** form *and* in machine-readable forms such as **JATS**. This is not just an abstract proposal; indeed, the text of this very book has been prepared using a novel document-generation system which creates both machine-readable structured text and **PDF** output, moreover with cross-referencing between them; notably, the positions of discursively important textual markers, such as sentence boundaries, are mapped to **PDF** screen coordinates (the code library for the book includes document-generation code as well as the data set of coordinate positions generated as part of the book's publication workflow). In particular, it is reasonable for authors and editors to manually introduce textual annotations for content such as named entities, keywords, important technical terms, and other content which should be targeted by **NLP** engines separate and apart from ordinary lexemes with their conventional natural-language semantics. Typically such specialized terms/lexemes would be marked up in any case because they may require distinct fonts or styling than their surroundings. It is also reasonable to manually define sentence boundaries via simple rules (e.g. two following spaces mark the end of a sentence; a single space, such as that following an abbreviation, indicates situations where a character such as a period, which could potentially mark the end of a sentence, is actually playing a different discursive role).

By following simple rules of document content-entry and lexicography, certain **NLP** tasks, such as sentence-boundary and Named Entity recognition, can be optimized — eliminating the need for probabilistic algorithms and relying instead on much less sophisticated, but more accurate, markup-parsing logic. If sentence boundaries and Named Entities are explicitly annotated in machine-readable text encodings, then extracting these features is not really an issue of "Natural Language Processing" as such. On the other hand, **AI**-driven analysis of document corpora would still require **NLP** for other aspects of parsing documents; it is unreasonable to expect authors, for instance, to manually notate sentence parse-graphs. This then suggests the question of where the boundary lies between discursive structures which might reasonably be left to authors or editors to manually notate (e.g. sentence boundaries) and those which in practice could only be obtained via **NLP** (such as part-of-speech tags). Related to this question is how best to model **NLP** structures, such as the trees or graphs representing the syntax of natural-language sentences. We will consider this question in subsequent chapters in the context of Conceptual Space Theory.

References

1 Teresa Aydillo, *et al.*, "Immunological Imprinting of the Antibody Response in COVID-19 Patients". *Nature Communications*, Volume 12 (2021). https://www.nature.com/articles/s41467-021-23977-1

2 Alessia Bardi, *et al.*, "Enhanced Publication Management Systems: A systemic approach towards modern scientific communication". Dissertation, University of Pisa, 2016. https://core.ac.uk/download/pdf/79621484.pdf

3 Khalid Belhajjame, *et al.*, "Workflow-Centric Research Objects: First class citizens in scholarly discourse". https://users.ox.ac.uk/~oerc0033/preprints/sepublica2012.pdf

4 Crenguta Bogdan, *et al.*, "Towards a Clinical Trial Ontology Using a Concern-Oriented-Approach". *IRPPS Working Papers*, Volume 10 (2006). https://www.movetothecloud.it/irpps/e-pub/index.php/wp/article/view/10/78

5 Kevin Bretonnel Cohen and Lawrence E. Hunter, "Text Mining for Translational Bioinformatics". *PLOS Computational Biology*, Volume 9, Issue 4 (2013). https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003044

6 Steve Canham and Christian Ohmann, "A Metadata Schema for Data Objects in Clinical Research". *Trials*, Volume 17 (2016). https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-016-1686-5

7 Yao-Qing Chen, *et al.*, "Rapid Isolation and Immune Profiling of SARS-CoV-2 Specific Memory B Cell in Convalescent COVID-19 Patients via LIBRA-seq". *Signal Transduction and Targeted Therapy*, Volume 6 (2021). https://www.nature.com/articles/s41392-021-00610-7

8 Helena Cousijn, *et al.*, "A Data Citation Roadmap for Scientific Publishers". *Nature Scientific Data*, Volume 5 (2018). https://www.nature.com/articles/sdata2018259.pdf

9 "COVID-19 Open Research Dataset (CORD-19)". 2020. Version 2020-03-13. Retrieved from https://pages.semanticscholar.org/coronavirus-research. Accessed 2020-03-20. doi:10.5281/zenodo.3715506 https://pages.semanticscholar.org/coronavirus-research

10 Christos Davatzikos, *et al.*, "Cancer Imaging Phenomics Toolkit: Quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome". *Journal of Medical Imaging*, Volume 5, Number 1 (2018). https://www.nmr.mgh.harvard.edu/~you2/publications/Davatzikos18_CaPTk.pdf

11 Sonia Difrancesco, *et al.*, "Sociodemographic, Health and Lifestyle, Sampling, and Mental Health Determinants of 24-Hour Motor Activity Patterns: Observational Study". *Journal of Medical Internet Research*, Volume 23, Number 2 (2021). https://www.jmir.org/2021/2/e20700

12 Ludëk Eyer, *et al.*, "Nucleoside analogs as a rich source of antiviral agents active against arthropod-borne flaviviruses". https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5890575

13 Johanna T Fifi and J Mocco, "COVID-19 related stroke in young individuals". *The Lancet*, Volume 19 (2020). https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(20)30272-6/fulltext

14 Holger Fröhlich, *et al.*, "From hype to reality: data science enabling personalized medicine". *BMC Medicine*, Volume 16 (2018). https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-018-1122-7

15 Obi L. Griffith, *et al.*, "Text-mining clinically relevant cancer biomarkers for curation into the CIViC database". *Genome Medicine*, Volume 11 (2019). https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0686-y

16 Maritt Kirst, *et al.*, "Sociodemographic Data Collection for Health Equity Measurement: A mixed methods study examining public opinions". *International Journal for Equity in Health*, Volume 12 (2013). https://equityhealthj.biomedcentral.com/articles/10.1186/1475-9276-12-75

17 Li-Min Liu, *et al.*, "A New Software Development Methodology for Clinical Trial Systems". *Advances in Software Engineering*, Volume 2013. https://www.hindawi.com/journals/ase/2013/796505

18 Aaron M. Orkin, *et al.*, "Reporting of Sociodemographic Variables in Randomized Clinical Trials". *JAMA Network Open*, Voume 4, Number 6 (2021). https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2780555

19 Barbara McGillivray, *et al.*, "The Citation Advantage of Linking Publications to Research Data". *PLOS One*, 2020. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230416

20 John S. McIlwain, "Clinical Trial Management Systems (CTMS) System Selection Considerations". https://velos.com/wp-content/uploads/Clinical-Trial-Management-Systems-System-Selection-Considerations-Whitepaper-1.pdf

21 Shelby Meyer, *et al.*, "Sociodemographic Diversity in Cancer clinical Trials: New findings on the effect of race and ethnicity". *Contemporary Clinical Trials Communications*, Voume 21 (2021). https://www.sciencedirect.com/science/article/pii/S245186542100020X

22 Zachary Montague, *et al.*, "Dynamics of B Cell Repertoires and Emergence of Cross-Reactive Responses in Patients with Different Severities of COVID-19". *Cell Reports*, Voume 35 (2021). https://www.sciencedirect.com/science/article/pii/S2211124721005180

23 Ross W Paterson, *et al.*, "The Emerging Spectrum of COVID-19 Neurology: Clinical, radiological and laboratory findings", *Brain*, Volume 143, Number 10 (2020), pages 3104-3120. https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awaa240/5868408

24 Mark A. Parsons and Ruth Duerr, "Data Identifiers, Versioning, and Micro-citation". https://www.thelancet.com/action/showPdf?pii=S1473-3099%2820%2930086-4

25 Lisa Paschold, *et al.*, "SARS-CoV-2-specific Antibody Rearrangements in Prepandemic Immune Repertoires of Risk Cohorts and Patients with COVID-19", *Journal of Clinical Investigation*, 2020. https://www.jci.org/articles/view/142966

26 Philip R. O. Payne, "Translational Informatics: Enabling high-throughput research paradigms". *Physiology Genomics*, Volume 39 (2009), pages 131-140. https://journals.physiology.org/doi/pdf/10.1152/physiolgenomics.00050.2009

27 Samir Abu-Rumeileh, *et al.*, "Guillain-Barré Syndrome Spectrum Associated with COVID-19: An up-to-date systematic review of 73 cases", *Journal of Neurology*, Volume 268, Number 4 (2020), pages 1133-1170. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7445716

28 Enar Reilent, "Whiteboard Architecture for the Multi-agent Sensor Systems". https://www.thelancet.com/action/showPdf?pii=S1473-3099%2820%2930086-4

29 Sayta S Sahoo, *et al.*, "Trial prospector: matching patients with cancer research studies using an automated and scalable approach". *Cancer Informatics*, Volume 13 (2014), pages 157-166. https://pubmed.ncbi.nlm.nih.gov/25506198

30 Heshui Shi, *et al.*, "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study". https://www.thelancet.com/action/showPdf?pii=S1473-3099%2820%2930086-4

31 Michael Simmons, *et al.*, "Text Mining for Precision Medicine: Bringing structure to EHRs and biomedical literature to understand genes and health". *Advances in Experimental Medicine and Biology*, 2016, pages 139-166. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5931382

32 Ayush Singhal, *et al.*, "Text Mining for Precision Medicine: Automating disease-mutation relationship extraction from biomedical literature". *Journal of the American Medical Informatics Association*, Volume 23, Issue 4 (2016), pages 766-772. https://academic.oup.com/jamia/article/23/4/766/2201020

33 Michael Souillard, *et al.*, "A Flexible Framework for Managing Temporal Clinical Trial Data". http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.75.9238

34 Markus Suhr, *et al.*, "menoci: Lightweight Extensible Web Portal enabling FAIR Data Management for Biomedical Research Projects". *BMC Bioinformatics*, Volume 21 (2020). https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03928-1

35 Gentle Sunder Shrestha, *et al.*, "Precision Medicine for COVID-19: A call for better clinical trials". *Critical Care*, Volume 24 (2020). https://ccforum.biomedcentral.com/articles/10.1186/s13054-020-03002-5

36 Alina Trifan and José Luís Oliveira, "FAIRness in Biomedical Data Discovery", *12th International Conference on Health Informatics*, Proceedings (2019), pages 159-166. https://www.scitepress.org/Papers/2019/75764/75764.pdf

37 Johan Van Soest, *et al.*, "Towards a semantic PACS: Using Semantic Web Technology to Represent Imaging Data". *Studies in Health Technology and Informatics*, Volume 205 (2014), pages 166-179. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5119276

38 Kris Verlaenen, "Middleware for Advanced Service Configuration: a Policy-Based Approach". Dissertation, Leuven, 2008. https://www.cs.kuleuven.be/publicaties/doctoraten/cw/CW2008_07.pdf

39 Lucy Lu Wang, *et al.*, "CORD-19: The Covid-19 Open Research Dataset". *ArXiv*, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251955

40 Pingping Wang, *et al.*, "Comprehensive Analysis of TCR Repertoire in COVID-19 Using Single Cell Sequencing". *Genomics*, Volume 113, Issue 2 (2021), pages 456-462. https://www.sciencedirect.com/science/article/abs/pii/S0888754320320838

41 Hetong Zhou, *et al.*, "The Landscape of Cognitive Function in Recovered COVID-19 Patients". *Journal of Psychiatric Research*, Volume 128 (2020), pages 98-102. https://www.sciencedirect.com/science/article/abs/pii/S0022395620308542

42 Yonggang Zhou, *et al.*, "Profiling of the Immune Repertoire in COVID-19 Patients with Mild, Severe, Convalescent, or Retesting-Positive Status". *Journal of Autoimmunity*, Volume 118 (2021). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7837046

43 Monica L. Zigman Suchsland, *et al.*, "Patient-Centered Outcomes Related to Imaging Testing in US Primary Care". *Journal of the American College of Radiology*, Volume 16, Number 2 (2019), pages 156-163. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7050575