https://medium.com/@docmusher/ontology-preludes-data-science-a-covid-19-use-case-cadfcc0b81cf

Ontology preludes Data Science: a COVID-19 use case.

Sven Van Poucke
May 19 · 17 min read

Sven Van Poucke, Brian Tvenstrup, Margot Vander Laenen, Werner Ceusters

1. Department of Anesthesia, Critical Care, Emergency Medicine and Pain

Therapy, Ziekenhuis Oost-Limburg, Genk, Belgium

2. Managing Director of Lindon Ventures, Tampa Bay FL, USA.

3. Department of Biomedical Informatics, University at Buffalo, Buffalo NY, USA.

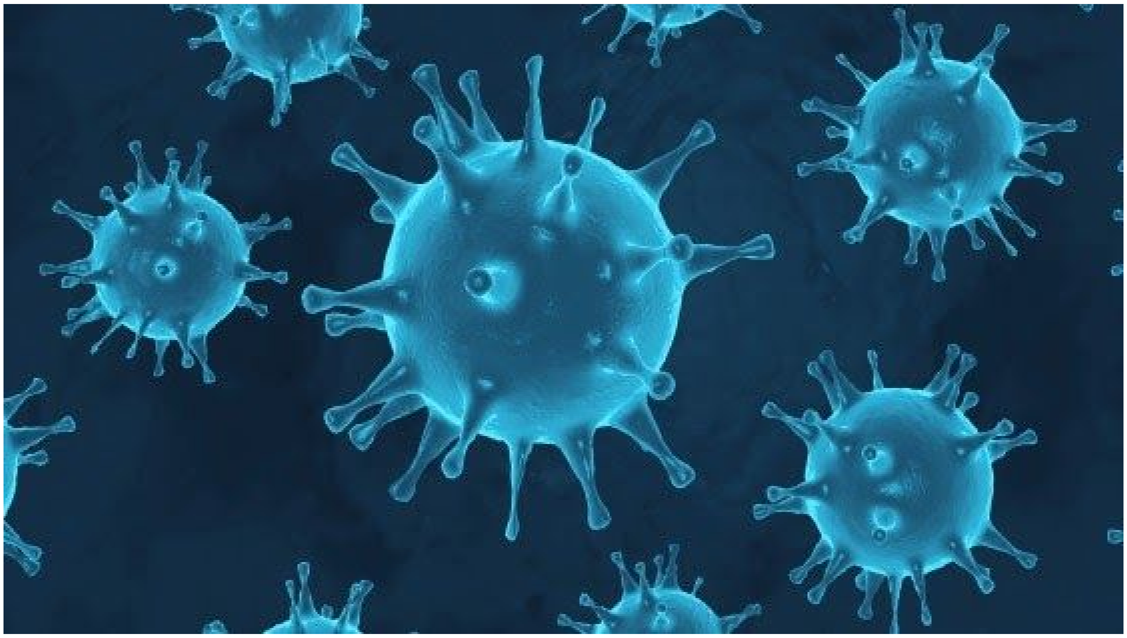
Correspondence to: Sven Van Poucke, MD, PhD. Department of Anesthesia, Critical Care, Emergency Medicine and Pain Therapy, Ziekenhuis Oost-Limburg, Genk, Belgium Email: svanpoucke@gmail.com.

Abstract and Keywords

The COVID-19 pandemic presents critical fulminations to global public health and the economy since it was identified in late December 2019 in China. Policy makers, pharmaceutical companies, health care providers, and patients, all are confronted with novel, complex, often conflictive problems demanding prompt action. There is a general perception that the preparedness and readiness was suboptimal or even lacking. In this paper we cover the actual problems and discuss the missed opportunities that were neglected from the start of the COVID-19 crisis up to this moment. In particular anything that was related to the distributed, published or taken for granted COVID-19 data (data science pet peeves), biomedical terminology and disease classifications (ontology laxness) is critically reviewed. In conclusion, a different level of professionalism and responsibility is required when dealing with these topics now and for any similar setting in the future.

Keywords:methodology; COVID-19; ontology; data science; bias

https://medium.com/@docmusher/ontology-preludes-data-science-a-covid-19-use-case-cadfcc0b81cf



-Introduction

As healthcare providers are overwhelmed digesting the current severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic, the opportunity to use data in order to optimize prevention and treatment on a short, mid or long-term notice, requires prompt action preventing any wisdom from being flushed down the toilet!

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an infectious disease caused by COVID-19 virus (1). Although the lungs are most commonly affected, the virus may also infect other organs. The infection can spread through the mucous membranes, from the nose down to the rectum. As the body's immune system shifts into a higher gear to battle the infection, the resulting inflammation may cause multiple organ failure.

More than at any time in the past, an awareness-raising campaign for sharing and communicating data, information and knowledge, based on data management best practices and a common vocabulary is needed (2). This includes an agreed-upon meaning for each term used in that vocabulary which needs to come equipped with built-in mechanisms for assessing and representing (dis)similarity of individuals — as we explain later, we don't mean just individual patients, but also their individual diseases, their individual symptoms, and so forth — described in these terms. It also requires adequate principles for the categorization of such individuals on the basis of features they share, implemented, for example, by means of formal properties restricted for appropriate domains, ranges, cardinality, and qualified as to whether they are transitive, reflexive, etc. At least, it is in everyone's interest to have semantic interoperability, where health care providers dealing with COVID-19 patients should tend to converge on standard terms and categorizations for their domain or to develop mappings from theirs to others. Adequate ontologies — in contrast to quick and dirty work that pops up each

https://medium.com/@docmusher/ontology-preludes-data-science-a-covid-19-use-case-cadfcc0b81cf

time major final resources are made available to address some crisis — would come in very handy here!

-***What is the actual problem and which missed opportunities are neglected during this COVID-19 crisis?***

Data science pet peeves

Although massive amounts of COVID-19 patient related data is produced, published, and embedded in spectacular dashboards geared towards the general public, the data quality is often as murky as the pandemic itself. In times where gigantic efforts are spent to develop, deploy and monitor predictive models, where resilience is key over accuracy, where overfitting models to the test set and "inaccurate" models delivering disastrous impacts on health management, curating data sets is of crucial importance (3).

From a data scientist perspective, here are some of the pet peeves from the COVID-19 reporting:

1. Results from different countries are continuously compared although counting and testing policies wildly differ from state to state and vary over time.

2. Using aggregate measures when there are substantial regional or sub regional variations.

https://medium.com/@docmusher/ontology-preludes-data-science-a-covid-19-use-case-cadfcc0b81cf

3. Data is visually represented on graphs without a clear definition of the x- or y-axis. Graphical material is sometimes published with no consistent y-axis scale (neither linear nor logarithmic but totally made up to illustrate the alleged dramatic nature of the context)

4. An ambiguity or inconsistency in terms like positive cases versus confirmed cases, is noticed.

5. There is a tendency for a general focus on comparison of metrics not adjusted for population size, density or age.

6. There have been low quality choices of visualization exhibits that humans are generally not good at interpreting such as area graphs and pie charts.

7. Invention of totally useless metrics then reported as though they are meaningful.

8. Lack of attention to uncertainty and variance in forecasts and predictions, particularly when such predictions are then used to make serious decisions for populations.

9. Reliance on outdated models or projections even after subsequent data shows they are unlikely to be accurate (e.g. failure to appropriately update models).

10. A general lack of consideration of precision versus recall and implications of false positives versus false negatives in test design and usage.

Ontology laxness

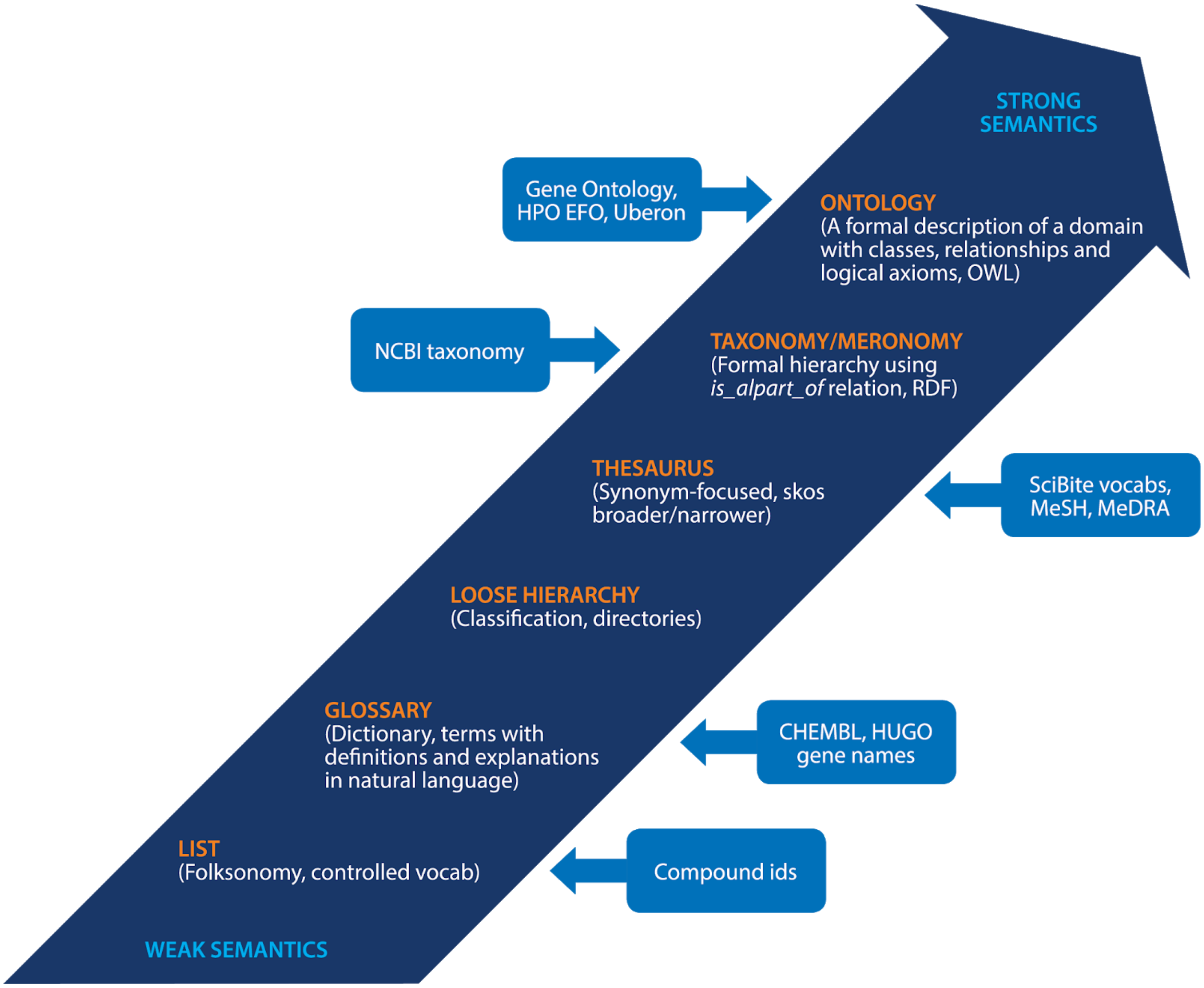
The purpose of biomedical terminology is to inventorize, and provide, adequate names for substances, qualities and processes employed in the biomedical domain both by practitioners and researchers. Biomedical terminology reflects not only the various subspecialties of biomedicine (roughly corresponding to specialized subdomains or dimensions of biomedical reality), but also the many purposes for which terminologies are developed.

It is often said that there is nothing that cannot be encountered in the domain of medicine. Deviations are everywhere. It is essentially impossible to describe disease manifestations without resorting not only to lists of associated signs and symptoms but also to the frequency distributions of the latter for each particular disease. In addition to the most common, prototypical form of the disease, there are many clinical variants in which some of the common manifestations are missing and other, less frequent, manifestations take their place.

In this context of high variability, it is not surprising that names and terms are crafted to represent not only the prototypical categories but also the many possible variants. Thus names are formed that include information identifying specific clinical variants, or information about associated lesions or injuries. The default assumption on the part of those working with, and on, terminologies, is that such specially crafted terms correspond to entities found in biomedical reality in just the same sense as do more straightforward terms such as ‘meningitis’ or ‘fever’.

https://medium.com/@docmusher/ontology-preludes-data-science-a-covid-19-use-case-cadfcc0b81cf

This assumption takes many forms, and on the weakest possible reading — unfortunately the one that is endorsed by biomedical terminology and disease classifications as currently practised — it posits that every term used in clinical practice or in biomedical research is ipso facto to be accepted as designating a corresponding 'concept' irrespective of whether the term denotes something truly existing on the side of entities in reality or some putative (clinical) idea or notion as correlate of the term with which it is associated (4). In what follows, however, we will provide evidence to the effect that this view is too simplistic as it conflates various dimensions along which terms convey information.


STRONG SEMANTICS
Gene Ontology, HPO EFO, Uberon
ONTOLOGY
(A formal description of a domain with classes, relationships and logical axioms, OWL)
NCBI taxonomy
TAXONOMY/MERONOMY
(Formal hierarchy using is_alpart_of relation, RDF)
THESAURUS
(Synonym-focused, skos broader/narrower)
SciBite vocabs, MeSH, MeDRA
LOOSE HIERARCHY
(Classification, directories)
GLOSSARY
(Dictionary, terms with definitions and explanations in natural language)
CHEMBL, HUGO gene names
LIST
(Folksonomy, controlled vocab)
Compound ids
WEAK SEMANTICS

The *first dimension* is concerned with whether terms truly denote something existing in reality, and if so, at what level of individuality. Important distinctions along this dimension are whether a term denotes 1. an *individual* aka *particular* such as denoted by the term 'Belgium', 2. a *type*, such as denoted by the terms 'human being', 'infection', and 'coronavirus', 3. a *class* such as a collection of human beings, and 4. nothing at all such as the term ‘miasma’ which once was thought to denote something assumed to be the cause of cholera.

The *second dimension* captures whether terms carve out entities using their natural bona fide boundaries or by means of fiat boundaries determined by human convention for practical purposes. It is along this dimension that, for example, terms of the third kind along the first dimension can be divided in terms that denote *genuine classes*, for example the collection of all individuals that are instances of some type (e.g. all human beings are instances of the type ‘human being’) versus those terms which are formulated merely in order to meet current practical requirements of coding and classification and which denote *ad hoc subclasses* of genuine classes, e.g. ‘human beings who died from a coronavirus infection in Belgium’, possibly abbreviated as ‘Belgium covid-19 victims’.

The *third dimension* allows us to recognize to what extent authors or users of certain terms judge what they convey to be faithful to reality. Such terms contain epistemological references reflecting detectability, modality, uncertainty, vagueness, and so forth. Epistemology-loaded terms are pervasive in biomedical vocabularies as the putative classes they attempt to denote often do not comply with sound classification principles. Such terms have been documented to cause problems in the evolution and alignment of terminologies (5). They contribute significantly to the complexity of text mining, one of the techniques used to automatically derive meaningful and actionable information from unstructured text data. If patient related data will ever be useful in the context of the COVID-19 pandemic for any scientific purpose, the use of text mining techniques able to identify the epistemic context in which a medical term appears is mandatory. *Negation*, for example, is widely used, as in ‘without evidence of infection’, and ‘infection has been ruled out’. Then there is the phenomenon of *hedging* which occurs when a clinician modifies or uses a diagnostic term in such a way that it questions the likelihood that the diagnosis is correct, an extreme, but nevertheless concrete example from a real patient record being ‘probably a possible COVID-19 infection’. Hedging is frequently used in both the biological literature and in clinical notes to denote uncertainty or speculation. It is important for text-mining applications to detect hedge cues and their scope; otherwise, uncertain events are incorrectly identified as factual events. However, due to the complexity of language, identifying hedge cues and their scope in a sentence is not a trivial task.

The *last, at least for our purposes, dimension* reflects the extent to which terms are likely to convey what they are intended to convey. That includes not only their understandability for the audience for which they are used, but also how fine-grained and appropriately they delineate what they intend to denote thereby taking into account how reality, our understanding thereof, and even language itself evolves over time (6). It can be argued, for example, that the term ‘smartwatch’ was years ago appropriate because of what these types of watches were capable to do, but not anymore appropriate as they are nowadays way more used for other purposes (e.g. heart rate monitoring, gps tracking) than as portable timepiece.

These four dimensions **do justice to three different fields of enquiry that are lumped together in the concept-oriented approach towards biomedical terminology**: *ontology* (1st and 2nd dimension), *epistemology* (3rd dimension) and *linguistic terminology* (4th dimension). It is therefore that each biomedical *terminology* should be linked to a biomedical *ontology* which clearly distinguishes the nature of represented entities in function of the 1st and 2nd dimension, and ideally to a *realism-based ontology*, i.e. an ontology which endorses the view that it is genuine classes and types that should directly reflect the categorization principles. These types must be represented in the ontology in such a way that they reflect the world structure by highlighting the resemblance between categories as maximizing the sum of all the common features within a category minus the sum of the measures of all of the distinctive features.

Realism-based Ontology as applied in biomedical informatics starts from the idea that types are invariants in reality and that it is types which are, or should be, captured in the general terms used in the textbooks of biological science (7). These types are instantiated by the particulars which are members of the genuine classes associated with these types. Ontology is the study of such types and of how instances of these types, for example organisms, relate to instances of other types, such as organism parts, qualities, functions, processes, diseases or symptoms. As an example, the *is_a* relation between two types such as ‘human being’ and ‘mammal’ reflects in a realism-based ontology the scientific law that for all instances of the first type (‘human being’) it is the case that at all times they are instances of the first type they are also at the same time instances of the second type (‘mammal’). As another example, the *part_of* relation between two types, when construed appropriately, reflects that for all instances of the first type (f.i ‘coronavirus’) it is the case that at all times they are instances of the first type they have as part some instance of the second type (f.i. ‘protein shell’).

Although there are some restrictions, there is no fixed limit to the number or nature of relationships that an instance of some type can hold with instances of other types or of the same type. A medical diagnosis, for instance, stands in an *aboutness* relation not only to a configuration formed by at least four components, but also to each of these components themselves: the patient, one or more conditions in the patient, the types that these conditions instantiate and the relationship in which the former components stand. When a diagnosis is inaccurate with respect to one or more of these components — most common is an erroneous assignment of the type of the condition — it fails at the level of compound expression but is still about the other components at the level of individual reference. (8)

-Case definition of COVID-19 infected patients

Case definitions are typically made for surveillance purposes and are dynamical in nature. From a general health policy perspective it can be comprehended that the risk assessment should be constantly updated in order to provide guidance for countries and authorities to respond to an outbreak.

We noticed that in China and other early affected regions, the case definition was initially narrow and that it became gradually broadened to allow detection of more cases as knowledge increased, particularly milder cases and those without epidemiological links to updated areas of presumed community transmission. These changes should be taken into account when making

https://medium.com/@docmusher/ontology-preludes-data-science-a-covid-19-use-case-cadfcc0b81cf

inferences on epidemic growth rates and doubling times, and therefore on the reproductive number, to avoid bias.

As such, case definitions are not intended to replace clinical or public health practitioner judgment in individual patient assessment and management.

The WHO introduced new codes for COVID-19 including clinical coding examples in the context of COVID-19 (9).

The new codes for COVID-19 consist of the following sections:

-New ICD-10 codes for COVID-19

-Clinical Coding of COVID-19 with ICD-10

-Mortality Coding of COVID-19 with ICD-10

-WHO COVID-19 Case definitions for Global Surveillance

The International Classification of Diseases (ICD-10-CM) evolved out of a terminology for compiling mortality and morbidity statistics but now constitutes a controlled vocabulary used by the insurance industry for reporting claims. As noticed, in many cases biomedical terms are crafted not only for naming the classes of entities found in biomedical reality (1st and 2nd dimension), but also to represent additional information (3rd and 4th dimension).

Both categories, U07.1 (COVID19, virus identified) and U07.2 (COVID19, virus not identified) are suitable for cause of death coding. Similarly, new codes were created for ICD-11. COVID-19 is reported on a death certificate as any other cause of death, and rules for selection of the single underlying cause are the same as for influenza (COVID-19 not due to anything else). It is hallucinating that for recording on a death certificate, no special guidance is

given. The respiratory infection may evolve to pneumonia that may evolve to respiratory failure and other consequences. Potentially contributing comorbidity (immune system problem, chronic diseases…) is reported in part 2, and other aspects (perinatal, maternal…) in frame B, in line with the rules for recording. A manual plausibility check is recommended for certificates where COVID-19 is reported, in particular for certificates where COVID-19 was reported but not selected as the single underlying cause of death.

https://medium.com/@docmusher/ontology-preludes-data-science-a-covid-19-use-case-cadfcc0b81cf

-Discussion

Assuming that clinicians deliver state of the art care by providing clinically everything that is required and possible, in order to achieve the best outcomes for their patients which includes having a clear and detailed clinical picture of each individual patient they see, then the current bottlenecks for good clinical practice based on available figures and used terminology of COVID-19 patients potentially are:

1. Their accurate clinical view based on their skills and expertise and the knowledge currently available together allow them to propose the best course of action in line with that knowledge (which may later turn out not to have been the best, but was the best at the time of decision making). As an example, the impact that COVID-19 is associated with a unique type of blood clotting disorder that is primarily focused within the lungs and which undoubtedly contributes to the high levels of mortality being seen in patients with COVID-19 was initially not recognized or at least underestimated.

2. Executing the proposed line of action is hampered diagnostically by limited availability (in number and time it takes to complete) of diagnostic testing with variable or unknown false positives and negatives, etc as well as, therapeutically (as demonstrated by bed limitations, ventilators, …), but in the end, certain patients die while other recover. In both groups, there are patients with and without COVID-19 but a major problem is that the absolute truth remains for many cases an absolute uncertainty.

3. From 1. and 2., a wealth of detailed data could be derived, but was not exploited to the extent we would like for several reasons: clinicians don’t have the time because of patient overload,

https://medium.com/@docmusher/ontology-preludes-data-science-a-covid-19-use-case-cadfcc0b81cf

the tools to do so accurately are not available or clinicians are unaware of their existence. Timely consulting data scientists, ontology experts is only common practice in the top facilities around the world.

4. The sadness of the reality is that whatever detail is recorded, gets lost in coding. On top of that, rules are vague, perhaps not always correctly applied, coding is done by coding clerks that don't have the time to ask clinicians for the missing information and even if they would, clinicians are too busy doing clinical hands-on work, … and this is the case for both the clinical coding as the death certificate administration.

5. Population health centers receive only the codes, not the details. They communicate their findings to politicians, but do not take into account baselines from previous years, seasonal variances, …

6.The superficial and binary flow for assumed non-COVID-19 patients and probably COVID- 19 positive patients lead to situations where asymptomatic patients afterwards diagnosed as COVID-19 positive were mixed with healthy patients or where patients with e.g. a dry throat related to extreme hyperglycemia and polyuria were exposed to COVID-19 positive patients after entering the COVID-19 flow at the emergency department based on the fact that a sore throat (which could be part of the COVID-19 disease) dictated to follow this flow.

7. It is confusing when the ICD-10 tables indicate 'no symptoms', if in this situation, 'no respiratory symptoms' is meant? However, there is now enough evidence that the new virus also attacks other organs, so it would not make any sense to only account for respiratory symptoms. Additionally, CT scans followed by COVID-19 PCR tests demonstrated that patients with predominantly abdominal symptoms and no or minimal respiratory symptoms ultimately tested positive for COVID-19. These patients failed to be categorized as suspected based on the WHO COVID-19 Case definitions for Global Surveillance.

Moreover, for a clinician, the ICD-10 tables are a bit puzzling and require to be explained. They don't seem to be designed for clinical use, but for coders who have to derive codes from what clinicians documented in the medical record. It is the sentence 'COVID-19 documented as cause of death' in the middle section of the first table that suggests this. The problem is lack of information about the rules clinicians are using to determine that COVID-19 is the cause of death. If a cardiac patient comes to the ER with chest pain, they do a COVID-19 PCR test which turns out to be positive, and the patient dies of an AMI, why on Earth would COVID-19 be the cause of death? Outside the pandemic, nobody bothered to take virus tests under such circumstances. What if they did and found, say, herpes. Would then herpes be the cause of death?

Also, with respect to comorbidities, we assume the same rules as always should be applied, but are they applied when clinicians are buried under the work?

https://medium.com/@docmusher/ontology-preludes-data-science-a-covid-19-use-case-cadfcc0b81cf

-Conclusions:

The COVID-19 pandemic initiated a global health crisis with an unknown impact on different levels of society on a short, mid or long-term notice. Policy makers, pharmaceutical companies, health care providers, and patients, all are confronted with novel, complex, often conflictive problems demanding prompt action. There is a general perception that the preparedness and readiness was suboptimal or even lacking. In this paper we covered the actual problems and discussed the missed opportunities that were neglected from the start of the COVID-19 crisis up to this moment, in particular anything that was related to the distributed, published or taken for granted COVID-19 data, biomedical terminology and disease classifications was critically reviewed. In conclusion, a different level of professionalism and responsibility is required when dealing with these topics now and for in any similar setting in the future (10).

Acknowledgments:

Part of the research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number KL2TR001413 to the University at Buffalo. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflicts of Interest:

All authors have completed the ICMJE uniform disclosure form. The authors have no conflicts of interest to declare.

Ethical Statement:

The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Open Access Statement:

https://medium.com/@docmusher/ontology-preludes-data-science-a-covid-19-use-case-cadfcc0b81cf

-References

1. Ahn DG, Shin HJ, Kim MH, et al. Current Status of Epidemiology, Diagnosis, Therapeutics, and Vaccines for Novel Coronavirus Disease 2019 (COVID-19). J Microbiol Biotechnol. 2020; 30(3):313–324.
2. Caulfield T. Pseudoscience and COVID-19 — we've had enough already. Nature.2020. doi: 10.1038/d41586–020–01266-z. [Epub ahead of print]
3. Johns Hopkins Coronavirus Resource Center. Baltimore, Maryland, USA. [updated 2020; cited 2020 May 08] [Internet] Available from: https://coronavirus.jhu.edu/
4. Smith B. Beyond concepts: Ontology as reality representation. In Achille C. Varzi & Laure Vieu (eds.), Formal Ontology in Information Systems (FOIS). 2004 pp. 1–12.
5. Bodenreider O, Smith B, Burgun A. The Ontology-Epistemology Divide: A Case Study in Medical Terminology. Form Ontol Inf Syst. 2004;2004:185–195.
6. Ceusters W, Smith B. A Realism-Based Approach to the Evolution of Biomedical Ontologies. Proceedings of AMIA 2006, Washington DC, 2006;:121–125.
7. Smith B, Ceusters W. Ontological Realism as a Methodology for Coordinated Evolution of Scientific Ontologies. Applied Ontology, 2010;5(3–4):139–188.
8. Hogan WR, Ceusters W. Diagnosis, misdiagnosis, lucky guess, hearsay, and more:an ontological analysis. J Biomed Semantics. 2016;7(1):54.
9. Emergency use ICD codes for COVID-19 disease outbreak. World Health Organization. Geneve, Switzerland. [updated 2020; cited 2020 May 08] [Internet] Available from: http://www9.who.int/classifications/icd/covid19/en/
10. Van Poucke S, Zhang Z, Schmitz M, et al. Predictive Analysis in Critically Ill Patients: Using a Visual Open Data Analysis Platform. PLOS ONE. 2016;11(1): e0145791